Robust Unsupervised Gaze Calibration using Conversation and Manipulation Attention Priors

RÉMY SIEGFRIED, Idiap Research Institute and EPFL, Switzerland
JEAN-MARC ODOBEZ, Idiap Research Institute, Switzerland and EPFL, Switzerland

Gaze estimation is a difficult task, even for humans. However, as humans, we are good at understanding a situation and exploiting it to guess the expected visual focus of attention (VFOA) of people, and we usually use this information to retrieve people's gaze. In this paper, we propose to leverage such situation-based expectation about people's VFOA to collect weakly labeled gaze samples and perform person-specific calibration of gaze estimators in an unsupervised and online way. In this context, our contributions are the following: i) we show how task contextual attention priors can be used to gather reference gaze samples, which is a cumbersome process otherwise; ii) we propose a robust estimation framework to exploit these weak labels for the estimation of the calibration model parameters; iii) we demonstrate the applicability of this approach on two Human-Human and Human-Robot interaction settings, namely conversation and manipulation. Experiments on three datasets validate our approach, providing insights on the priors effectiveness and on the impact of different calibration models, in particular the usefulness of taking head pose into account.

CCS Concepts: • **Computing methodologies** → *Activity recognition and understanding*; **Tracking**; • **Human-centered computing** → **Human computer interaction (HCI)**; Collaborative and social computing.

Additional Key Words and Phrases: Gaze estimation, visual focus of attention, remote sensor, RGB-D camera, conversation, manipulation, unsupervised calibration, online calibration.

ACM Reference Format:
Rémy Siegfried and Jean-Marc Odobez. 2022. Robust Unsupervised Gaze Calibration using Conversation and Manipulation Attention Priors. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 1, Article 20 (January 2022), 27 pages. https://doi.org/10.1145/3472622

1 INTRODUCTION

Gaze is a visual cue that provides rich information about people, ranging from measuring their attention to interpreting their intention and mind. Thus, the capacity to estimate gaze has a great potential in many applications related to human-human interaction (HHI) analysis [2], psychological studies [30], medical diagnosis [11], or human-robot/computer interaction (HRI, HCI) [15] among others. However, despite recent improvements in sensors and methods, the precise tracking of people's gaze remains difficult without using intrusive sensors like wearable eye trackers or constraining people's movement like in screen-based settings. This is the case in applications where users behaving naturally are recorded by remote sensors, which results in limited frame rate and, more importantly, low eye image resolution and large head pose variability [43]. In this work, our goal is to improve gaze and VFOA estimation by building user-specific models, investigating

Authors' addresses: Rémy Siegfried, Idiap Research Institute, Martigny , EPFL, Lausanne, Switzerland, remy.siegfried@idiap.ch; Jean-Marc Odobez, Idiap Research Institute, Martigny, Switzerland , EPFL, Lausanne, Switzerland, odobez@idiap.ch.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
1551-6857/2022/1-ART20 $15.00
https://doi.org/10.1145/3472622

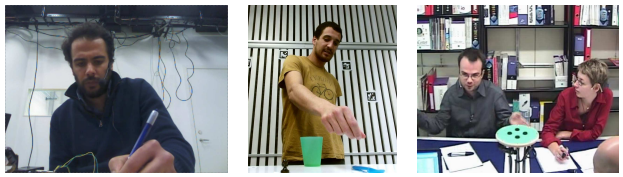


Fig. 1. Example of people's gaze during a task. One can reasonably guess what they are looking at, even if it is difficult to see eyes (the word he is writing, the object that will be grasped, the speaking person).)

weak conversation or manipulation attention prior along with robust estimation to perform online and unsupervised gaze calibration.

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 [16] and to overestimate it when people look at a target that is far from us [24]. Furthermore, we are biased by features like head orientation (Wollaston effect) [16] and nose direction [19]. However, while we encounter difficulties at estimating gaze direction (a continuous value), we are good at exploiting context information to estimate people's visual focus of attention (VFOA), i.e. "what they are looking at" (a categorical value), as illustrated in Fig. 1. Indeed, we can 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. Systems recording users exhibiting unconstrained behaviour in natural environments must rely on remote sensors. In these cases, besides the core challenge of handling the large variation in eye characteristics across people, eye images usually have a low resolution (e.g. 60x70 pixels with a Kinect v2 at 1 meter), people present more dynamic gaze behaviours as they actively use their head and gaze to look around, so that eye appearance is affected by head movements, gestures, and facial expressions. In addition, as collecting data for each setup is inconvenient and challenging, gaze estimation models are usually trained on separate datasets which may lead to mismatched conditions and hence performance loss. Overall, this makes gaze estimation with remote sensors more difficult compared to relatively stable and controlled situations like screen-based setups.

To overcome these difficulties, gaze estimation models are usually adapted to new users and settings through calibration, which highly increases their performances. However, 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.

The human visual attention is usually described as driven by two mechanisms: *top-down* attention which is conscious and related to the task, and *bottom-up* attention which is unconscious and related to the saliency of our perception field [13]. As systems interacting with people in natural conditions using remote sensors rarely have access to people's field of view, this 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 [12];
- driving: people often look at the future path [20];
- web browsing: the gaze is strongly related to the cursor position [4].

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 [2, 40], but in this work, we also propose a method to exploit them as calibration points to improve gaze estimation.

1.2 Approach summary and contributions

Our approach for unsupervised user-specific gaze calibration is summarized in Fig. 5. In essence, it comprises two main steps. The first one corresponds to our aim and consists in 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.
Paper plan. This paper is structured as follows. First, we discuss related works in the field of gaze estimation, gaze calibration, and VFOA estimation (Sec. 2). Second, we introduce the datasets we used for our experiments (Sec. 3) to illustrate our goal and then present our method (Sec. 4). Third, we provide our experimental results on weak VFOA labeling (Sec. 5) as well as offline, and online calibration (Sec. 6 and 7). Finally, we discuss the limitations of this work and future work (Sec. 8).

2 RELATED WORKS

In the following, we review gaze estimation, gaze calibration, and their use for VFOA estimation.

2.1 Gaze estimation

The difficulty of gaze estimation was traditionally addressed by using expensive specialized hardware or highly controlled scenarios [10]. Recently, progress in technology (e.g. sensors) coupled with machine learning techniques have opened the way to gaze estimation using remote sensors. For instance, the authors of [8] proposed to use the 3D information from an RGB-D sensor to frontalize the eye images and estimate the gaze using an appearance-based model, i.e. by learning a direct mapping from an eye image to its corresponding gaze angles. Lately, such appearance-based gaze estimation gained popularity with the growing number of gaze estimation dataset, like Columbia Gaze [45], UT Multiview [46], EYEDIAP [6] or MPIIGaze [52], allowing to take advantage of recent advances in machine learning [9, 46] and particularly deep learning methods [23, 34, 35, 48, 53], leading to better performances. Appearance-based methods are now considered more effective than their model-based counterparts [51]. For instance, it was shown that an adapted VGG 16 model that takes normalized eye images in input beats state of the art methods even in cross-dataset

settings [53]. Gaze estimation in videos was also addressed, e.g. by using a recurrent CNN that incorporates eye images, face image, and facial landmarks [34].

Nevertheless, one main challenge is the performance drop in cross-dataset conditions, i.e. when a gaze estimation 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 10° on cross-dataset evaluation conducted on existing gaze datasets, using for example an adapted VGG 16 model [53] or an hourglass network with model-based gaze estimation [35]. Such performances are far below those achieved in single dataset cross-subject situations (3-5 degrees), showing that it is still difficult to train models that generalize well to other setups.

2.2 Gaze estimation calibration

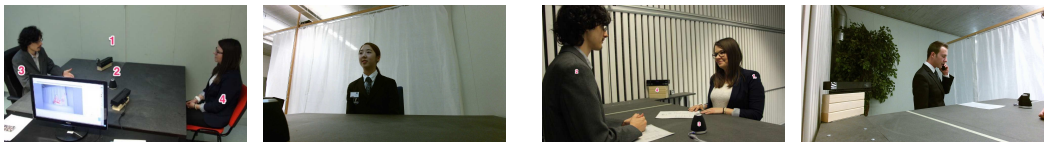
To further improve results, appearance-based methods use calibration to adapt a learned generic gaze model to the specificities of the current setup (domain adaptation), or of users (personalization) [51], as people exhibit a large range of variations in eye characteristics like for instance the difference between the optical and visual axis which can not be measured from image only.

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 approach is to learn a regression from the gaze output (or intermediate features) to a better gaze estimation [17]. In [23], it was shown that using a simple correction model, like linear regression, can already significantly improve the results. Another approach is to fine-tune the gaze estimation model itself, by retraining part of it [25]. Finally, some 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 [22], or the gaze model can be directly trained to estimate the difference in gaze between a new eye image and a reference one [23].

However, while improving gaze calibration for appearance-based methods recently grew in interest [5, 22, 23, 50], less attention was paid to how to collect the required calibration samples. Traditional screen-based gaze trackers rely on a 5 or 9 points calibration procedure consisting of asking users to look at some pre-defined points [31]. This has rather poor usability [27] and further researches tried to overcome this limitation by exploring pursuit calibration [36, 39] which automatically detects when the user looks at a moving target. This approach is less tedious and more flexible [36], 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 performed task, to infer highly probable gaze targets and use them as calibration points. For example, it was shown possible to calibrate gaze estimation methods using gaze and mouse operations coordination [47], communication conventions (e.g. addressee looking at the speaking person) [44], the knowledge of the subject's field of view (e.g. looking at one's smartphone) [28] or dwell-time during gaze typing [37]. 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 [37, 47] and head-mounted devices [28]. In our work, we investigate whether such calibration paradigm can hold in the case of 3D remote sensors setups, which present more variability in terms of head pose and gaze, by extending our work [44] to other calibration models, priors, and settings like object manipulations.



(a) *Interviews* setup (b) *Desk* setup

Fig. 2. UBImpressed dataset. Setup view (left) and example of recording (right) for both setups.

2.3 VFOA estimation

Our goal is to improve gaze estimation via calibration in settings and tasks where the attention of users to the environments is monitored by the system, which assumes that the scene itself is monitored, i.e. that the position of people and object of interest are tracked by the system.

In such cases, earlier works used head pose as a proxy for gaze because of the lack of accurate gaze trackers. They relied on models of gaze behaviors from the head pose, and GMM or Dynamical Bayesian Networks [26, 40] as inference mechanism, potentially taking into account context information [2, 40] or modeling the joint VFOA of all participants [2, 26]. Indeed, the head pose is easier to estimate and even current methods provide more stable head pose than gaze estimates.

Nevertheless, with the recent improvement of gaze estimation, it is possible to rely on simpler frame-based geometrical models [7]. Indeed, even if gaze estimation tends to be less stable and accurate than head pose estimation, the latter can be a poor proxy in certain cases as head movements are not always present at gaze shift, typically during aversion. In those cases, VFOA estimation from gaze is more accurate, even if the gaze signal itself can be noisy and biased.

3 DATASETS

To frame our study, we introduce here the datasets used in our experiments (annotation statistics in Appendix A.1). As we are interested in settings where subjects are recorded while performing natural action in conversation and object manipulation settings, we selected three datasets that present different situations: dyadic interactions, four-party meetings, and object manipulation. These datasets have similar settings, as they are recorded with RGB-D sensors using one camera per subject in the scene (plus an additional camera for the manipulated objects when it is relevant). This allows tracking people and objects and thus to build a 3D scene representation.

3.1 Dyadic interactions - UBImpressed dataset

Data. The UBImpressed dataset [29] consists of 330 short dyadic interactions (five to ten minutes) in which a participant interacts with an actor in two different scenarios. 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. 2. From a gaze perspective, this dataset presents a situation where a unique important target is present (the other person). The short interaction and the formal setup favor mutual gazes, so it is interesting to study people's behaviours 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. In this formal interaction, people exhibit constrained behaviors. In the *Desk* scenario, a receptionist 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.

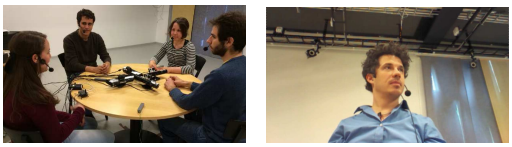


Fig. 3. KTH-Idiap dataset. Whole setup (left) and example of recording (right).

Fig. 4. ManiGaze dataset. Recording views of the subject (left) and table (right) cameras.

Annotations. We annotated 4 *Interviews* and 4 *Desk* sessions (16 videos in total). In each video, we annotated the first minute of the video and 5 additional segments of 10 seconds, separated by 50 seconds of not annotated video (so 6 segments in total). 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. Ignoring blinking frames, we ended up with 42'000 annotated frames with 52% of them having the "gazing" label (hence 48% with the aversion label). Moreover, the sound was synchronously recorded by a microphone array that automatically detects the beginning and end of each participant utterances, thus indicating who is talking in each video frame.

3.2 Four-party meetings - KTH-Idiap dataset

Data. This dataset [32] consists of five one-hour four-party meetings in which three students present their projects to an interviewer who leads the discussion. 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 generate a high variability of head gestures, facial expressions, and gaze patterns. 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 seen in Fig. 3.

Annotations. VFOA was manually annotated on the first minute and on nine additional 30 s segments spread on the entire video, to catch different situations like monologues, dialogues, attentive listening, moments of aversion, etc. Each frame was annotated whether the participant was blinking, looking at another person (left, facing, right), aversion. Overall, it represents 110 minutes of annotation (180'000 frames excluding blinking) in which 19% are aversion and 81% are looking at another person. Finally, as participants were wearing lapel microphones, the utterance information was also estimated using sound intensity [32].

3.3 Object manipulation - ManiGaze dataset

Data. The ManiGaze dataset [43] consists of 16 sessions where a user performs different looking and manipulation tasks in front of a Baxter robot, illustrating data that could be acquired in an HRI setup, with two cameras being embedded on the robot. The first one, a Kinect v2 recorded the user, and the second one, an Intel RealSense D435, the table from above (see Fig. 4). In our experiments, we focus on three out of the four scenarios of this dataset:

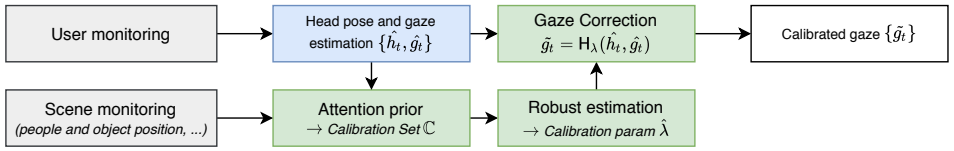


Fig. 5. Calibration framework. An initial head pose and gaze estimations $\{\hat{\mathbf{h}}_t, \hat{\mathbf{g}}_t\}$ are extracted from the video of the user for each frame t (blue block). In parallel, the calibration set $\mathbb{C}$ is automatically built relying on the attention prior and the scene 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 $\widetilde{\mathbf{g}}_t$.

- *Markers on the Table Targets (MT)*: the user looks in turn to 14 markers present on the table in front of him in the order given by the robot.
- *End-effector Targets (ET)*: the user looks at the robot'arms that are moving in the 3D space, which results in higher variability of head poses and gaze directions. The robot regularly stops its arm for a few seconds in different positions (37), where the user's VFOA is annotated.
- *Object Manipulation (OM)*: the user is asked to behave freely while performing pick-and-place actions ordered by the robot.

Annotations. The dataset provides automatic VFOA annotation for the MT (5391 frames in total) and ET scenarios (5894 in total). In this dataset, aversions were not annotated, so all annotated frames correspond to the user looking to a known object.

In addition, to study context-based priors based on eye-hand coordination (see below), we manually annotated the moments where the user is grasping or releasing an object during the OM session. More precisely, we annotated the first and last frames of contact between the user's hand and the object, for a total of 22 annotations per video (11 pick-and-place actions).

4 METHOD

4.1 Approach overview

Our overall approach is shown in Fig. 5. 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 $\hat{\mathbf{g}}_t = (\hat{\phi}_{\mathrm{g},t}\hat{\theta}_{\mathrm{g},t}) = \mathrm{G}(I_{RGBD})$ 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 $\hat{\mathbf{h}}_t = (\hat{\phi}_{\mathrm{h},t}\hat{\theta}_{\mathrm{h},t})$. The gaze estimate is updated using a calibration function H according to:

$$\widetilde{\mathbf{g}}_t = \mathrm{H}_\lambda(\hat{\mathbf{g}}_t, \hat{\mathbf{h}}_t). \tag{1}$$

As can be seen, the head pose is exploited in the correction as gaze errors might depend on it (more in Sec. 4.5). 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.

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

$$\mathbb{C} = \{(t_i, j_i), i = 1 \dots N_{\mathbb{C}}\}, \tag{2}$$

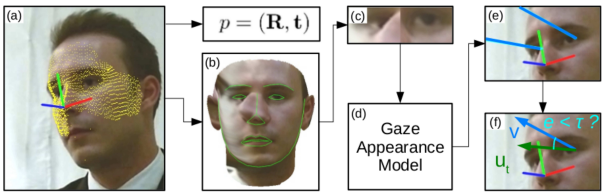


Fig. 6. Framework. a) Head pose estimation [49]. b) Face image frontalization. c-d) Mapping of eye images to gaze angles **g**. e) Computation of gaze direction **v** from **g** and **p**. f) Extraction of the angle error e between the vector x_j pointing to the target and the gaze vector **v**.

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:

$$\mathbb{F} = \left\{ (\hat{\mathbf{g}}_t, \mathbf{g}_t), \text{ with } \mathbf{g}_t = \mathsf{V}^{-1}\left(\frac{x_{j,t} - e_t}{\|x_{j,t} - e_t\|}\right), \forall (t, j) \in \mathbb{C} \right\} \tag{3}$$

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 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 $\widetilde{\mathbf{g}}$ and the calibration gaze **g**:

$$r_t(\lambda) = \widetilde{\mathbf{g}}_t - \mathbf{g}_t = \mathsf{H}_\lambda(\hat{\mathbf{g}}_t, \hat{\mathbf{h}}_t) - \mathbf{g}_t. \tag{4}$$

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:

- Gaze estimation G from RGB-D images (Sec. 4.2);
- Gaze calibration set $\mathbb{C}$ collection using different weak VFOA labeling schemes (Sec. 4.4);
- Calibration function H_λ (Sec. 4.5);
- Robust estimation of the calibration parameters λ (Sec. 4.6);
- Online estimation of λ (sec. 4.7).

4.2 Gaze and VFOA estimation

We used the same framework as [9], which is summarized in Fig. 6, using more recent methods for head pose and gaze estimation.

Head pose estimation. We relied on the Headfusion method [49] to get the head pose, defined as the rigid transform $\mathbf{p} = (\mathbf{R}, \mathbf{t})$ between the head and the camera coordinate systems. Relying on the automatic fitting of both a 3D Morphable Model of the face and a 3D raw representation of the head, the usage of depth information makes the method more robust to large head poses variations that might occur in recordings of people behaving freely in natural interactions settings, compared to 2D landmarks based methods. Note that the head pose angles $\hat{\mathbf{h}}$ expressed in the coordinate system attached to the camera can be directly computed from **R**.

Eye images extraction and normalization. Using the head pose, the textured mesh obtained from RGB and depth image is rotated to get a frontal image of the face [9], in which a facial landmark detector [14] is applied to get the position of the eyes and crop the eye images (36x60

pixels, centered on the middle point between the eye corners). It allows normalizing the size and appearance of the eye images, making it easier to learn an appearance-based model. This is similar to the perspective warping normalization [46], but leverage the availability of depth data.

Gaze estimation. We selected the GazeNet [53] architecture for its performances on state-of-the-art datasets in cross-subject and cross-dataset settings. It is based on a *VGG16* architecture and takes head pose angles into account. We trained it to take a normalized eye image as input and to output the corresponding gaze yaw and pitch angles. We validated our implementations on the Columbia Gaze, UT Multiview, and Eyediap (floating target, mobile pose) datasets and obtained state-of-the-art results (respectively 2.88°, 3.63°, and 6.3° mean angular error). We used 10 epochs, a learning rate of 0.001 which is divided by 2 every 2 epochs, and a batch size of 128.

As the datasets used in this paper do not contain enough gaze-annotated frames and diversity of target positions, we trained GazeNet with an external dataset. We used Eyediap as it provides depth data, allowing the use of a similar eye image extraction and normalization, reducing the gap between the training and testing domains. We trained the network on all Eyediap subjects of the floating target setting with mobile head pose and used the resulting model for all our experiments.

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

$$\kappa_{j,t} = \arccos\left(\mathsf{V}(\widetilde{\mathbf{g}}_t) \cdot \frac{(x_{t,j} - e_t)}{||(x_{t,j} - e_t)||}\right). \tag{5}$$

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_T. 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:

$$f_t = \begin{cases} j_{\min} & \text{if } j_{\min} = \text{argmin}_j(\kappa_{j,t}) \text{ and } \kappa_{j_{\min},t} < \kappa_T, \\ aversion & \text{otherwise.} \end{cases} \tag{6}$$

To set the threshold κ_T, 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 $\kappa_T = 10°$, which is the angular size of an object of 35cm located at 2m away from the camera.

4.3 Target 3D positions

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. 4.1, 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 Head Coordinate System. 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.

4.4 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. 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.

4.4.1 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 [3]. In particular, studies on social interactions showed that listeners are likely to look at the person being listened to (88% of the time) [1]. 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 [2]. Also, speakers are often looking at their addressee, i.e. the target of their speech (77% of the time) [1]. 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

$$\mathbb{C}_{conv} = \left\{(t, j) \mid j \in \mathbb{S}_t \text{ and } |\mathbb{S}_t| = 1\right\} \tag{7}$$

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 [33]). 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.

4.4.2 Manipulation prior. It has been shown that vision is intimately bound up with the control of purposeful actions [18]. During object manipulation, gaze typically reaches the object before any movement of the hands has started [1] 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 [12]. The reliability of this behavior makes it a good candidate for unsupervised gaze calibration [42].

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 [12], where it was shown that people look at the origin and destination positions of the 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.

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

$$\mathbb{C}_{mani} = \left\{(t, j) | t' \in \mathbb{M}_j \text{ and } t \in \mathcal{W}_{t'}\right\}. \tag{8}$$

4.4.3 Physical constraints. The proposed method estimates the head pose with high accuracy (2° to 5° errors as reported in [49]), 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 [21]. However, it was reported that looking further than 30° on the side is already uncomfortable [45]. 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:

$$\mathbb{C}_{pc} = \left\{(t, j) \mid |\phi_{j,t} - \phi_{hp,t}| < \tau_\phi, |\theta_{j,t} - \theta_{hp,t}| < \tau_\theta\right\}. \tag{9}$$

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 paper, 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 [41], 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.

4.5 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:

$$\mathrm{H}_\lambda(\mathbf{h}, \mathbf{g}) = A\mathbf{g} + B\mathbf{h} + c, \tag{10}$$

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

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.

4.5.1 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:

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \text{ and } \lambda = \{c_\phi, c_\theta\}. \tag{11}$$

4.5.2 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:

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}, B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \text{ and } \lambda = \{a_1, \ldots, a_4, c_\phi, c_\theta\}. \tag{12}$$

4.5.3 Linear model with head pose. Gaze shifts usually involve both head pose and gaze-in-the-eye coordinated motion [40]. 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:

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}, B = \begin{bmatrix} b_1 & b_2 \\ b_3 & b_4 \end{bmatrix}, \lambda = \{a_1, \ldots, a_4, b_1, \ldots, b_4, c_\phi, c_\theta\}. \tag{13}$$

4.6 Robust estimation

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 [44]. 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 in Appendix A.2 describes the whole process.

Least Median Square (LMedS) estimation. It is defined as the parameters optimizing the median of the residuals (defined in Equ. 4):

$$\lambda_{\text{Med}} = \underset{\lambda}{\text{argmin}}\, \mathsf{E}_{Med}(\lambda, \mathbb{C}) \text{ with } \mathsf{E}_{Med}(\lambda, \mathbb{C}) = med\left\{\|r_i(\lambda)\|^2\right\}. \tag{14}$$

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 [38]. 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. Fig. 11 in Appendix A.4 shows an example of outliers filtering.

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

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:

$$\hat{\lambda} = (X^T X + \Lambda)^{-1}(X^T Y + \Lambda\mu_0), \tag{15}$$

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. 4.5.3 for which parameter estimation can be conducted separately for the ϕ and θ axes, we have when estimating the yaw parameters $\lambda_\phi = \{a_{11}, a_{12}, b_{11}, b_{12}, c_\phi\}$:[1]

$$X = \begin{bmatrix} \hat{\mathbf{g}}_{\phi,1} & \hat{\mathbf{g}}_{\theta,1} & \hat{\mathbf{h}}_{\phi,1} & \hat{\mathbf{h}}_{\theta,1} & 1 \\ \dots & \dots & \dots & \dots & \dots \\ \hat{\mathbf{g}}_{\phi,N} & \hat{\mathbf{g}}_{\theta,N} & \hat{\mathbf{h}}_{\phi,N} & \hat{\mathbf{h}}_{\theta,N} & 1 \end{bmatrix}, Y = \begin{bmatrix} \mathbf{g}_{\phi,1} \\ \dots \\ \mathbf{g}_{\phi,N} \end{bmatrix}. \tag{16}$$

4.7 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 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:

$$p\left((t,j) \in \mathbb{C}_k\right) = \left(\frac{N_{\mathbb{C}}^{max} - 1}{N_{\mathbb{C}}^{max}}\right)^{N_{upd}(t,k)}, \tag{17}$$

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.

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

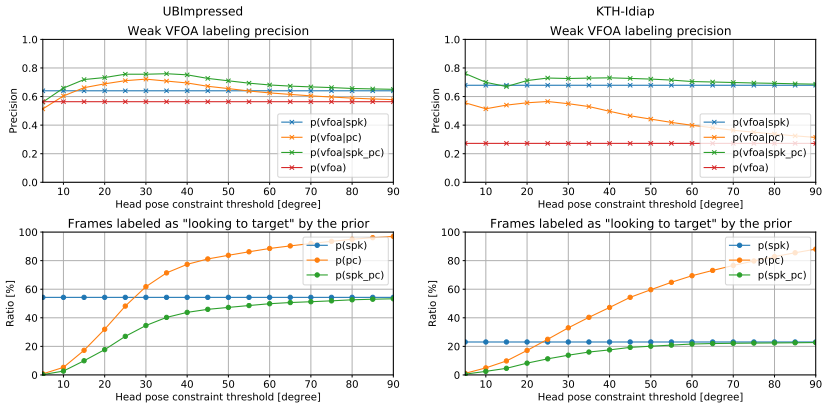


Fig. 7. Conversation prior statistics for each dataset. Top: weak VFOA labeling precision, i.e. probability that a subject looks at a given target (*vfoa*) during a *spk* event (the target is speaking) and/or a *pc* event (physical constraints are satisfied) for different values of τ_ϕ and τ_θ. Bottom: 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.

4.8 Implementation details

In all experiments, we set the VFOA threshold (see Equ. 6) to $\kappa_T = 10°$ for VFOA estimation. The iterations number for the LMedS was set to 500, and physical constraints thresholds τ_ϕ and τ_θ (see Equ. 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:

$$w_j = \frac{1}{J \cdot N_{\mathbb{C}}^j}, \tag{18}$$

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".

Our implementation is available on the Idiap's GitHub: https://github.com/idiap.

5 WEAK LABELING EVALUATION

We evaluate in this section the capacity of the VFOA priors at collecting good calibration sets using the annotated data in each dataset. 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.

5.1 Conversation prior

Fig. 7 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

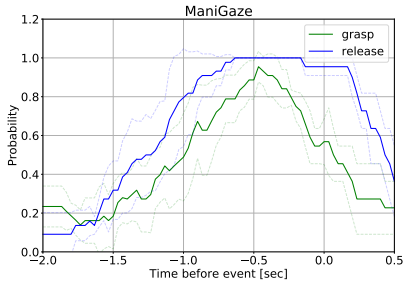


Fig. 8. Manipulation prior statistics. Probability that the subject looks at the initial (for grasps) or final (for releases) object position during a pick-and-place action in a time window around the grasp/release frame, computed from 4 ManiGaze dataset's subjects (OM session). Full line: mean. Dashed line: standard deviation.

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. 7, 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.

5.2 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. 8 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 in [12]. 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. (see Appendix A.3, Fig. 10) 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 in [12] as a prior to collect calibration samples, as described in Sec. 4.4.2.

6 OFFLINE CALIBRATION EXPERIMENTS

First, we compare the gaze estimation performances using the *Offline* calibration protocol (see Sec. 4.7), i.e. using all the available data for both calibration and evaluation.

6.1 Experimental protocol

6.1.1 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

Table 1. 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

metrics, namely the gaze angular error and the VFOA accuracy:

$$angErr = \frac{1}{T}\sum_{t=1}^{T} \arccos(\mathsf{V}(\mathbf{g}_t) \cdot \mathsf{V}(\widetilde{\mathbf{g}}_t)), \qquad vfoaAcc = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\widetilde{f}_t = f_t}, \tag{19}$$

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 Section 4.3) 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).

6.1.2 Evaluated models. We considered several calibration models (see Sec. 4.5) 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.

6.1.3 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.

6.2 Results using Conversation prior

Tab. 1 presents the average of the angular error and VFOA accuracy across subjects in the conversation datasets for different models and calibration types.

6.2.1 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. 2 and 3). 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.

6.2.2 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.

6.2.3 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 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 (see Sec. 6.2.5 as well).

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.

6.2.4 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. 9 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

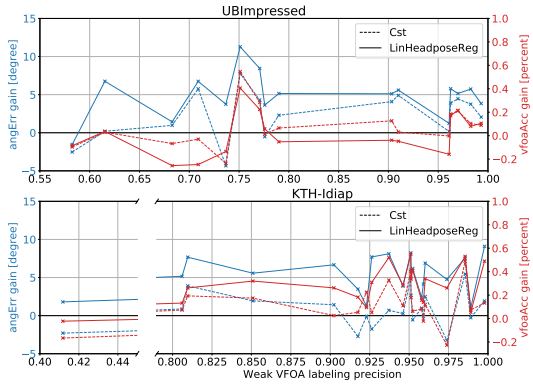


Fig. 9. Performance gain after calibration over the weak VFOA labeling precision for each subject using the **Cst** (dashed line) and **LinHeadGazeReg** (plain line) models.

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.

6.2.5 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.

6.3 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. 2 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 2. Mean angular errors (in degrees) and mean VFOA accuracy (in percent) across subjects for the ManiGaze dataset with calibration based on the entire session (*Offline*). Here we compare the raw gaze estimation without calibration (**Baseline**), the supervised calibration using the same session VFOA annotations (intra-session), the supervised calibration using VFOA annotations of the other session (cross-session), and the 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

6.3.1 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). The camera captures subjects from a very unusual point of view [43], 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°).

6.3.2 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

6.3.3 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,

Table 3. 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

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.

7 ONLINE CALIBRATION EXPERIMENTS

In this section, we evaluate the *Online* calibration method proposed in Sec. 4.7 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.

7.1 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.

7.2 Results.

Tab. 3 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.

8 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, especially when the calibration parameters are adapted over time. 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.

One limitation of the work is the simplicity of the proposed weak VFOA estimation method for collecting calibration data. A potential improvement could be to use some weighting and/or sampling strategies, by exploiting the probability of looking at a target. Furthermore, learning better VFOA prior models would be interesting, for instance by using more cues (e.g. gaze of a speaker). Also, the proposed VFOA inference is 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 and HRI, 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 will include the investigation of 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 [50]. It would be interesting to see if these approaches could benefit from the proposed 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. 5.2). 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.

9 CONCLUSION

In this work, 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.

ACKNOWLEDGMENTS

This research has been supported by the EU Horizon 2020 research and innovation program (grant no. 688147, MuMMER project), and uses the KTH-Idiap database made available by KTH, Sweden.

REFERENCES

[1] Henny Admoni and Brian Scassellati. 2017. Social Eye Gaze in Human-Robot Interaction: A Review. *Journal of Human-Robot Interaction* 6, 1 (2017), 25–53.
[2] Sileye Ba and Jean-Marc Odobez. 2011. Multiperson Visual Focus of Attention from Head Pose and Meeting Contextual Cues. *Transactions on Pattern Analysis and Machine Intelligence* 33, 1 (2011).
[3] Dan Bohus and Eric Horvitz. 2010. *Computational Models for Multiparty Turn Taking.* Technical Report. Microsoft Research Technical Report MSR-TR 2010-115.
[4] Mon Chu Chen, John R Anderson, and Myeong Ho Sohn. 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.
[5] Zhaokang Chen and Bertram Shi. 2020. GEDDnet: A Network for Gaze Estimation with Dilation and Decomposition. *arXiv preprint arXiv:2001.09284* (2020).
[6] Kenneth Funes, Florent Monay, and Jean-Marc Odobez. 2014. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *ACM Symposium on Eye Tracking Research and Applications.*
[7] Kenneth Funes, Laurent Nguyen, Daniel Gatica-Perez, and Jean-Marc Odobez. 2013. A Semi-Automated System for Accurate Gaze Coding in Natural Dyadic Interactions. In *International Conference on Multimodal Interaction.* ACM.
[8] Kenneth Funes and Jean-Marc Odobez. 2012. Gaze Estimation From Multimodal Kinect Data. In *Conference in Computer Vision and Pattern Recognition, Workshop on Gesture Recognition.*
[9] Kenneth Funes and Jean-Marc Odobez. 2016. Gaze Estimation in the 3D Space Using RGB-D Sensors. *International Journal of Computer Vision* 118 (2016), 194–216.
[10] Elias Daniel Guestrin and Moshe Eizenman. 2006. General theory of remote gaze estimation using the pupil center andcorneal reflections. *Transactions on bio-medical engi-neering* 53, 6 (2006).
[11] Linda Isaac, Janna Vrijsen, Mike Rinck, Anne Speckens, and Eni Becker. 2014. Shorter gaze duration for happy faces in current but not remitted depression: Evidence from eye movements. *Psychiatry Research* 218, 1-2 (2014), 79–86.
[12] Roland Johansson, Goran Westling, Anders Backstrom, and Randall Flanagan. 2001. Eye-Hand Coordination in Object Manipulation. *Journal of Neuroscience* 21, 17 (2001), 6917–6932.
[13] Fumi Katsuki and Christos Constantinidis. 2014. Bottom-up and top-down attention: different processes and overlapping neural systems. *The Neuroscientist* 20, 5 (2014), 509–521.
[14] Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.
[15] David Klotz, Johannes Wienke, Julia Peltason, Britta Wrede, Sebastian Wrede, Vasil Khalidov, and Jean-Marc Odobez. 2011. Engagement-based Multi-party Dialog with a Humanoid Robot. *Proceedings of the SIGDIAL 2011 Conference.*
[16] Nathan Kluttz, Brandon Mayes, Roger West, and Dave Kerby. 2009. The effect of head turn on the perception of gaze. *Vision Research* 49, 15 (2009), 1979–1993.
[17] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye tracking for everyone. In *Proc. of the conference on computer vision and pattern recognition.* IEEE.
[18] Michael Land. 2009. Vision, eye movements, and natural behavior. *Visual Neuroscience* 26, 1 (2009), 51–62.
[19] Stephen Langton, Helen Honeyman, and Emma Tessler. 2004. The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception & psychophysics* 66, 5 (2004), 752–771.
[20] Otto Lappi, Esko Lehtonen, Jami Pekkanen, and Teemu Itkonen. 2013. Beyond the tangent point: gaze targets in naturalistic driving. *Journal of vision* 13, 13 (2013), 11–11.

[21] Won June Lee, Ji Hong Kim, Yong Un Shin, Sunjin Hwang, and Han Woong Lim. 2019. Differences in eye movement range based on age and gaze direction. *Eye* 33, 7 (2019), 1145–1151.

[22] Erik Linden, Jonas Sjostrand, and Alexandre Proutiere. 2019. Learning to Personalize in Appearance-Based Gaze Tracking. In *Proceedings of the International Conference on Computer Vision Workshops*.

[23] Gang Liu, Yu Yu, and Jean-Marc Odobez. 2020. A Differential Approach for Gaze Estimation. *Transaction on Pattern Analysis and Machine Intelligence* (2020).

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

[25] David Masko. 2017. Calibration in eye tracking using transfer learning.

[26] Benoit Masse, Sileye Ba, and Radu Horaud. 2018. Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction. *Transactions on Pattern Analysis and Machine Intelligence* 40, 11 (2018), 2711–2724.

[27] Carlos Morimoto and Marcio Mimica. 2005. Eye gaze tracking techniques for interactive applications. *Computer vision and image understanding* 98, 1 (2005), 4–24.

[28] Philipp Muller, Daniel Buschek, Michael Xuelin Huang, and Andreas Bulling. 2019. Reducing Calibration Drift in Mobile Eye Trackers by Exploiting Mobile Phone Usage. In *Proc. of the ACM Symp. on Eye Tracking Research & Applications*.

[29] Skanda Muralidhar, Laurent Son Nguyen, Denise Frauendorfer, Jean-Marc Odobez, Marianne Schmid Mast, and Daniel Gatica-Perez. 2016. Training on the Job: Behavioral Analysis of Job Interviews in Hospitality. In *Proceedings of the International Conference on Multimodal Interactions*.

[30] Skanda Muralidhar, Remy Siegfried, Jean-Marc Odobez, and Daniel Gatica-Perez. 2018. Facing Employers and Customers: What Do Gaze and Expressions Tell About Soft Skills?. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia*. ACM, 121–126.

[31] Marcus Nystrom, Richard Andersson, Kenneth Holmqvist, and Joost van de Weijer. 2013. The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods* 45, 1 (2013), 272–288.

[32] Catharine Oertel, Kenneth A Funes, Samira Sheikhi, Jean-Marc Odobez, and Joakim Gustafson. 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*. 27–32.

[33] Catharine Oertel, Marcin Wlodarczak, Jens Edlund, Petra Wagner, and Joakim Gustafson. 2013. Gaze patterns in turn-taking. In *Annual Conference of the International Speech Communication Association*.

[34] Cristina Palmero, Javier Selva, Mohammad Ali Bagheri, and Sergio Escalera. 2018. Recurrent CNN for 3D Gaze Estimation using Appearance and Shape Cues. In *Proceedings of the 29th British Machine Vision Conference*.

[35] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. 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*.

[36] Ken Pfeuffer, Melodie Vidal, Jayson Turner, Andreas Bulling, and Hans Gellersen. 2013. Pursuit calibration: Making gaze calibration less tedious and more flexible. *Proceedings of the Symposium on User Interface Software and Technology*.

[37] Jimin Pi and Bertram E. Shi. 2019. Task-embedded Online Eye-tracker Calibration for Improving Robustness to Head Motion. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*.

[38] Peter Rousseeuw. 1984. Least median of squares regression. *Journal of the American statistical association* 79, 388 (1984), 871–880.

[39] Thiago Santini, Wolfgang Fuhl, and Enkelejda Kasneci. 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*. ACM, 2594–2605.

[40] Samira Sheikhi and Jean-Marc Odobez. 2015. Combining dynamic head pose-gaze mapping with the robot conversational state for attention recognition in human-robot interactions. *Pattern Recognition Letters* 66 (2015), 81–90.

[41] Ludwig Sidenmark and Hans Gellersen. 2019. Eye, head and torso coordination during gaze shifts in virtual reality. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 1 (2019), 1–40.

[42] Ludwig Sidenmark and Anders Lundström. 2019. Gaze behaviour on interacted objects during hand interaction in virtual reality for eye tracking calibration. In *11th ACM Symposium on Eye Tracking Research & Applications*.

[43] Remy Siegfried, Bozorgmehr Aminian, and Jean-Marc Odobez. 2020. ManiGaze: a Dataset for Evaluating Remote Gaze Estimator in Object Manipulation Situations. In *Proc. of the ACM Symp. on Eye Tracking Research and Applications*.

[44] Remy Siegfried, Yu Yu, and Jean-Marc Odobez. 2017. Towards the use of social interaction conventions as prior for gaze model adaptation. In *Proceedings of the International Conference on Multimodal Interaction*. ACM, 154–162.

[45] Brian Smith, Qi Yin, Steven Feiner, and Shree Nayar. 2013. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the ACM symposium on User interface software and technology*. ACM.

[46] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2014. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 1821–1828.

[47] Yusuke Sugano, Yasuyuki Matsushita, Yoichi Sato, and Hideki Koike. 2015. Appearance-based gaze estimation with online calibration from mouse operations. *Transactions on Human-Machine Systems* 45, 6 (2015), 750–760.

[48] Kang Wang, Rui Zhao, Hui Su, and Qiang Ji. 2019. Generalizing eye tracking with bayesian adversarial learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 11907–11916.
[49] Yu Yu, Kenneth Funes, and Jean-Marc Odobez. 2018. HeadFusion: 360 degree Head Pose Tracking Combining 3D Morphable Model and 3D Reconstruction. *Transactions on Pattern Analysis and Machine Intelligence* 40, 11 (2018).
[50] Yu Yu, Gang Liu, and Jean-Marc Odobez. 2019. Improving Few-Shot User-Specific Gaze Adaptation via Gaze Redirection Synthesis. In *Conference on Computer Vision and Pattern Recognition*.
[51] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2019. Evaluation of Appearance-Based Methods and Implications for Gaze-Based Applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM.
[52] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the conference on computer vision and pattern recognition*. 4511–4520.
[53] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *Transactions on Pattern Analysis and Machine Intelligence* 41, 1 (2017), 162–175.

A ONLINE APPENDIX

A.1 Annotation statistics summary

Table 4. 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

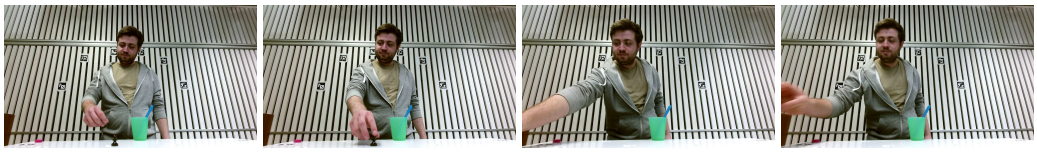
A.2 Robust calibration parameter algorithm

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 for cst, n=3 for lin, n=5 for 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$(\mathbb{C}'')$

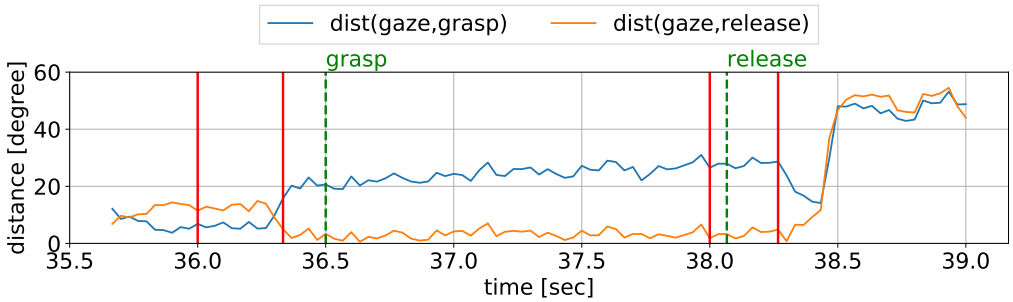
A.3 Example of gaze behavior during pick-and-place actions

Fig. 10 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 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.



a) Kinect2 recording at moments of interest.

b) Intel RealSense recording at moments of interest.

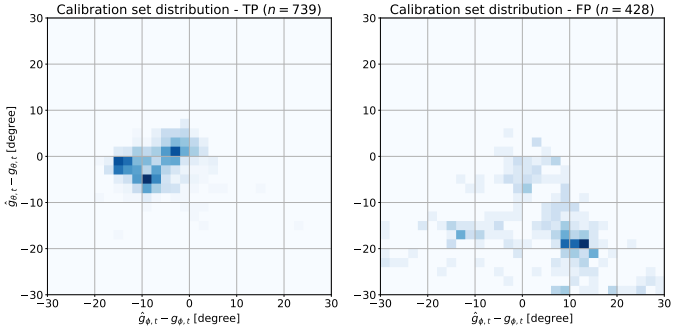


c) Cosine distance between gaze and grasping/releasing locations directions in the user's Head Coordinate System (HCS).

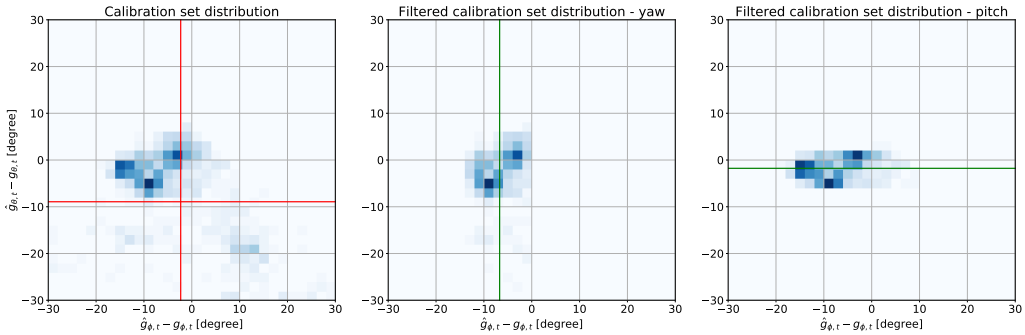
Fig. 10. Qualitative (a,b) and quantitative (c) gaze behaviour during pick-and-place actions. Red vertical lines in the plot indicate when Kinect2 and Intel RealSense pictures were taken. Green vertical dashed lines in the plot indicate when actions occur (object grasp and release).

A.4 Example of calibration set filtering using LMedS

Fig. 11 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 the Fig. 11a, and affects the calibration parameters estimation, especially in yaw, as seen in the Fig. 11b. 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.



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.

Fig. 11. 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).