Towards the Use of Social Interaction Conventions as Prior for Gaze Model Adaptation

Rémy Siegfried
remy.siegfried@idiap.ch
Idiap Research Institute
Martigny, Switzerland
École Polytechnique Fédéral de Lausanne
Switzerland

Yu Yu
yyu@idiap.ch
Idiap Research Institute
Martigny, Switzerland
École Polytechnique Fédéral de Lausanne
Switzerland

Jean-Marc Odobez
odobez@idiap.ch
Idiap Research Institute
Martigny, Switzerland
École Polytechnique Fédéral de Lausanne
Switzerland

ABSTRACT

Gaze is an important non-verbal cue involved in many facets of social interactions like communication, attentiveness or attitudes. Nevertheless, extracting gaze directions visually and remotely usually suffers large errors because of low resolution images, inaccurate eye cropping, or large eye shape variations across the population, amongst others. This paper hypothesizes that these challenges can be addressed by exploiting multimodal social cues for gaze model adaptation on top of an head-pose independent 3D gaze estimation framework. First, a robust eye cropping refinement is achieved by combining a semantic face model with eye landmark detections. Investigations on whether temporal smoothing can overcome instantaneous refinement limitations is conducted. Secondly, to study whether social interaction convention could be used as priors for adaptation, we exploited the speaking status and head pose constraints to derive soft gaze labels and infer person-specific gaze bias using robust statistics. Experimental results on gaze coding in natural interactions from two different settings demonstrate that the two steps of our gaze adaptation method contribute to reduce gaze errors by a large margin over the baseline and can be generalized to several identities in challenging scenarios.

CCS CONCEPTS

• **Computing methodologies** → **Tracking**; *Activity recognition and understanding*;

KEYWORDS

Gaze estimation, Appearance based model, Bias correction, Person-invariance, RGB-D cameras.

ICMI'17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11...$15.00
https://doi.org/10.1145/3136755.3136793

ACM Reference Format:
Rémy Siegfried, Yu Yu, and Jean-Marc Odobez. 2017. Towards the Use of Social Interaction Conventions as Prior for Gaze Model Adaptation. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3136755.3136793

1 INTRODUCTION

Gaze and eye movements provide rich information about a person's attention and how he/she is attending the world [3, 9, 27], and are non-verbal cues playing a major role in human-human interactions (HHI) [8, 12]. They are highly involved in communication for floor control management or signaling addressees, and their occurrences are also depending on higher level constructs, like thought process (eg it can be characteristic of the cognitive load) or personality traits (introversion, dominance), or autism. Thus, sensing gaze and modeling gaze and attention behaviors is necessary for a large range of applications, ranging from human interaction analysis to human-computer interaction (HCI) or human-robot interactions (HRI). It can be exploited for predicting end of turns and next speakers [10], which could be useful for improving dialog fluency in HCI/HRI. Or, analysing and synthesizing gaze behaviors can improve HRI by both having robots better understand human intentions and making their actions more natural in a social context, allowing them to appropriately answer humans [18] or improve the perception toward the robot [1].

In this paper, we address gaze estimation for HHI and HRI. This is a particularly challenging task where sensing conditions are usually quite different than in screen-gazing applications: higher pose variability (people do not face the sensor), lower (eye) image resolution since the sensor need to accomodate potentially larger user mobility or more people, larger illumination variations, unknown user and absence of user cooperation (no calibration data). Some of these challenges are illustrated in Fig. 1.

Related Work. There exist many systems and methods to track eye movements [4]. If we exclude electro-oculography, sceleral search coil and infrared (IR) oculography which are invasive and most suited for specific application, like sensing eye movements while sleeping, the most recent ones rely on computer vision by extracting gaze from an image, i.e.

Figure 1: Data considered in this paper: interview (top) and registration (bottom) scenarios [19].

without direct contact with the user. They can be classified in two main categories: geometric based methods (GBM) and appearance based methods (ABM) [17].

GBM approaches are the most precise ones and work by extracing local face and eye features and by mapping them into gaze cues. Features either rely on IR illumination (corneal glints), or on visual features (often the pupil center [11]). However, these methods require high resolution images, limiting user mobility, can only handle limited head pose variations, and are mainly targetting screen-gazing applications. As an alternative to sparse feature extraction, [29] introduced a dense 3D morphable model of the eye region, but the method still required high resolution data, and outliers with extreme error could be produced.

By directly learning a mapping from the eye image to the gaze parameters and avoiding feature tracking, ABM methods are more appropriate to handle lower image resolution and to be applied in HRI and HHI. Most ABM methods in this domain either assume a static head pose and train a user specific gaze appearance model, or they rely on some form of head pose dependent image rectification to crop the eye image in a canonical reference frame [7, 14–16, 26]. For instance [16] handles head movement by learning a head pose bias due to deformations created by an eye image rectification towards a user-specific reference pose, but the method needs to be calibrated for each new user. Authors in [7] alleviate this need by leveraging precise head pose estimation with a 3D Morphable Model (3DMM) facial mesh fitted onlineto compute a frontal face (and eye) image allowing to train a canonical appearance gaze model. Still, on their dataset, the pose and person invariant model (ie handling an unknown user) achieves an error of 6-12 degrees depending on the task, compared to 2-6 degrees for a person specific gaze model.

Very recent works started to exploit Deep Neural Networks (DNN) or CNN (Convolutional NN) to regress gaze from eye appearance directly. Zhang et al. [31, 32] collected a large dataset of eye images under diverse illumination conditions used for training a gaze estimation CNN, showing promising cross-dataset generalization capacity. Krafka et al. [13] had exploited a similar idea, collecting 2.5M frames from 1474 subjects to learn fixation points in a smartphone or tablet. Differently from [31], they trained and fused through fully connected layers parallel CNNs for each eye, the entire face, and the facial image location. Limitations include no access to 3D gaze information, as required for social scene analysis, and the use of high resolution eye images. To alleviate the cost of data collection, several works have proposed synthesized datasets for appearance based learning [25, 28].

Motivation. When handling unknow users, the above methods face several limitations. First, to crop the eye image in a canonical reference frame, most method rely on an eye alignment step. This alignment either rely on a semantic model (eg when fitting faces with known landmarks). [5, 7], or more often on explicit eye landmark (usually eye corners) localization [16, 22], even when using DNN models [13, 31]. However, such localization is usually ill-defined (the location of the visual 'eye corners' changes depending on the opening of the eyelids) and difficult and inaccurate (even for humans) given the low resolution encountered in HRI/HHI settings. Moreover, given the low resolution, a few pixel shifts in localization quickly result in high gaze errors. Since such localization might be user specific, this may result in a systematic bias, which also explains partially why individual gaze models work better. Secondly, while user-specific models are usually better, user adaptation often rely on a manual calibration requiring user cooperation or manual data processing [5]. Unsupervised methods based on visual saliency have been considered, but they are mainly restricted to screen gazing application [24].

Approach and Contribution. In this paper, we address gaze and attention estimation in human communication contexts. We rely on the head-pose independent gaze estimation framework of [7] and address the eye alignment and user adaptation issues through the following contributions.

Landmarks. we investigate the use of landmarks for gaze correction, a standard method used in most papers. More precisely, we show that frame-based alignment, although desirable, has limitations when dealing with non high-resolution images or people not only looking to the front, and investigate whether temporal averaging can overcome them.

Social cues for automatic gaze correction. It is well known that human follow conventions for smooth interactions and improve communication, as discussed earlier. For instance people nod, use audio backchannels, or look at speakers to show their attentiveness [20]. Our goal is thus to investigate whether such conventions can be used as prior (i.e. soft labels) for user gaze model adaptation. In our case, this will be achieved through the automatic selection of frames with high probability of looking at a target (a speaker), and by using these frames to correct user-specific gaze biases.

Experiments on 16 persons from 8 interactions in two challenging settings (see Fig. 1) demonstrate the rational of our approach. In the following, we first introduce the overall gaze and attention framework in Section 2, then present our contributions (Section 3) and experiments (Section 4), before discussing and concluding the paper.

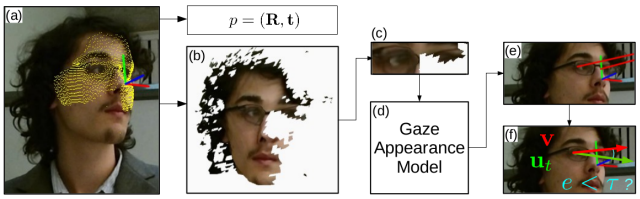


Figure 2: Gaze and attention framework [7]. a) The 3DMM mesh (adapted online to the specific user) is robustly fitted to the RGB-D data to estimate the head pose $\mathbf{p} = (\mathbf{R}, \mathbf{t})$. b) The frontal face image is computed by rotating and projecting the textured mesh (depth + rgb image). c) An eye cropping $\mathcal{C}$ region is defined to align the eye images from the frontal face into a canonical frame. d) The gaze angles g in the HCS are estimated from the eye images. e) The gaze g and head pose p are used to estimate the gaze direction v in the 3D space. f) The angle error e between the vector pointing to the target $\mathbf{u}_t$ and the gaze vector v of the eye closest to the camera is compared to a threshold τ to decide whether the person is looking at the target.

2 HEAD-POSE INDEPENDENT GAZE ESTIMATION FRAMEWORK

We use a method similar to the one in [7], whose main elements are summarized in this section. It takes as input the data provided by several RGBD cameras (eg two cameras in our setups, Fig. 1), which can be seen as textured 3D meshes once both RGB and depth camera are calibrated [5], computes the head pose and gaze of subjects, and decides whether people look at targets (other persons). We assume a calibrated set-up, so that observations (positions, vectors) can be expressed either in world coordinate system (**WCS**), the camera coordinate systems, or the head coordinate system of people (**HCS**) once their head pose is estimated.

Fig. 2 presents the framework. Note that all steps can be performed on-line, without previous knowledge of the person. We provide below further details about them.

Head Pose. It is extracted by fitting the Basel Face 3D morphable model (3DMM) mesh [21] to the depth data, using a variant of the iterative closest point (ICP) method and further depth and visual processing to obtain an accurate head pose even in case of difficult poses which are common in our setup [30]. Note that the mesh is adapted online to the specific user using a multi-instance fitting approach.

Frontal Face. The textured 3D mesh is rotated using the head pose parameters to obtain a frontal representation of the head. The texture can further be projected into into a 2D plan, resulting in a frontal face image independent of the initial head pose. The image will only suffer from some local deformations caused by the rotation and white spaces denoting an absence of data due to eventual self occlusions or lack of depth data (eg sometimes around frames of glasses).

Eye Alignment and Cropping. To estimate the gaze in the frontalized image, we crop the eye region. In the baseline approach, the cropping $\mathcal{C}^b$ is defined by relying on the theoretical positions of the eyes and eye corners which are known on the 3DMM and whose projections are used to crop canonical images of 75x60 pixels for both the left and right eye.

Gaze Estimation. The eye image is then used to estimate the gaze direction using an appearance gaze model. Following [7], we rely on Support Vector Regression (SVR) applied to multi-level Histogram-of-Gradients (HoG) features derived from the eye images. The model is trained on the Eyediap dataset [6]. It estimates for each eye the gaze direction $\mathbf{g} = (\phi, \theta)$ (in **HCS**, i.e. implicitly for a frontal face) defined by their yaw (ϕ) and tilt (θ) angles. These angles $\mathbf{g}$ can be mapped into a corresponding gaze direction unitary vector though a transform denoted Φ, i.e. we have $\mathbf{v} = \Phi(\mathbf{g})$.

Attention Decision. Thanks to the 3D approach and since we know the eye locations and the target location (a person) in the 3D space, deciding whether a person looks at a target can simply be done by comparing the gaze direction to the direction $\mathbf{u}_t$ associated to looking at the target, as shown in Fig. 2f. More precisely, as gaze information, we rely on the measure obtained from the eye that is the closest to the camera chosen because it is usually the most visible and thus less prone to occlusions and deformations from the rotation compared to the other eye, resulting in more precise and stable estimations. Then, to compare the gaze direction $\mathbf{v}$ of this eye with the target direction $\mathbf{u}_t$, we compute their angle difference and compare it to a threshold according to:

$$e = \arccos\left(\Phi(\mathbf{g} + \mathbf{b}) \cdot \mathbf{u}_t\right) < \tau. \quad (1)$$

Thus, if $e < \tau$, one consider that the subject is looking to the considered visual target. In the above, $\mathbf{b}$ is a gaze bias which can be added to the gaze estimate (see next section) and which is set to 0 in the baseline. To set the threshold τ, we must account for both uncertainties in the gaze direction estimates and the fact that visual targets are usually not a single point in space (e.g. the face of a person). A value of $\tau = 10°$ is usually fine, and corresponds roughly to a distance of 35cm at a distance of 2m, which is a typical distance between people in our setting.

3 ONLINE ADAPTATION APPROACH

While the baseline approach provides an interesting head-pose independent 3D gaze estimation framework, it suffers from some limitations which can adversely affect the outputs, as discussed in the introduction. First, the 3DMM model fitted online to the individual may not represent well her/his actual head shape, either because the eigenshapes in the BFM model are not rich enough, or due to an inaccurate fitting (eg due to the fact that the person is only seen from a 45°to profile side during most of the video). As a result, the baseline eye cropping $\mathcal{C}^b$ may not be accurate, as illustrated in Fig. 3, resulting in erroneous gaze estimates.

Secondly, the person eye shapes might not be well represented in the gaze estimation training data, and this may result in noisy estimates biased towards some direction. In

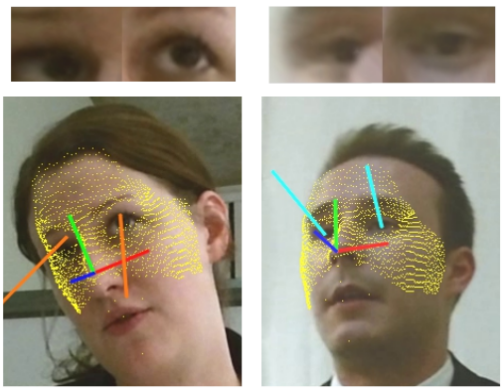


Figure 3: Baseline alignment issues. Yellow dots represent the 3DMM mesh, the red-green-blue coordinate system represents the head pose, and the blue (looking at other person) and orange (not looking) arrays represent the estimated gaze direction originating from the 3D eye ball centers. The cropped frontalized eye images are shown on the top of each result. Right: correct outcome. Left: due to a slightly inaccurate (user) 3DMM fit, cropped eyes are lower and to the right, resulting in an under-estimated gaze elevation θ.

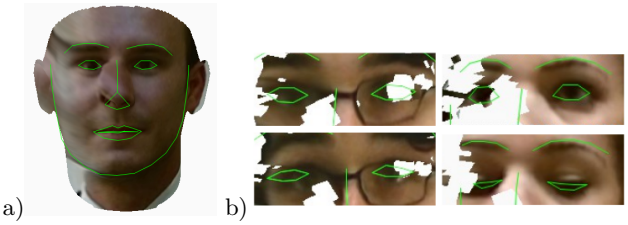


Figure 4: a) Dlib's facial landmarks extraction applied on the frontal face image. b) Zoomed sampled results illustrating erroneous detections (left column) and variability in eye corner localization in function of the eyelid opening (right column) as the person look to the front (top), to the side (bottom left) or down (bottom right image).

addition the two effects could be combined. Below, we describes the two methods proposed to address these issues: the use of landmarks, and the estimation of a gaze bias based on social interaction priors and robust statistics.

3.1 Eye Cropping Alignment Using Landmarks

Since the semantic information from the 3DMM mesh might not be enough to obtain a precise eye localization and cropping, we also resort to landmark detection. Note that [7] evaluated different alignment methods, but they were not improving much the results, maybe because experiments did not involve much cross-dataset (or setting) situations. In addition, the best performing method ('SICPA') required labeled frames for user adaptation.

To detect landmarks, we relied on the Dlib[1] library. We applied the 68 facial landmarks extractor to the rectified frontal face image. This led to much more robust and stable results, esp. in the desk scenario where due to the much less frontal poses, landmark detection completely failed when dealing with the original images (more than 20% of the time).

Alignment Procedure. Given an estimate of the eye corner locations, we simply compute the alignment translation that minimizes the discrepancy between the aligned corners (after translation) and their expected location in the canonical frame which was used to train the appearance model.

Eye Corner Estimates. Two approaches were tested.

• *Frame alignment Lm.* In most gaze estimation system, the eye corner positions detected at each frame are used to perform the alignment, and we evaluated this approach. Such strategy might be desirable as it can not only correct for an overal cropping bias, but also for per frame mis-alignment due for instance to inaccurate pose estimation. However, while it might be working well for people looking to the front (i.e with dominantly open eyes), high eye image resolution, and near frontal pose, as is typical of screen looking applications, the situation is much worse in HRI or HHI where one or several of these characteristic is not met. This is illustrated in Fig. 4. Considering the low resolution, missing pixels due to occlusions, it is sometimes difficult to correctly place the landmarks, like in the left bottom parts of the examples displayed in Fig. 4. Also, a major problem arises when the subject is looking down. In that case, the eyelids are close to each other and the detected 'visual' eye corners do not match their counterparts in other situations.

• *Running average alignment AvgLm.* To handle the above noise issues, we also evaluated a simple method averaging the frame-level translation correction over the last 30 frames. This appeared to remove most of the observed erroneous estimates and to produce better and more consistent stabilized eye sequences, which should be important for the next step.

3.2 Robust Gaze Bias Estimation From Interaction Prior Soft Labels

Even if landmark alignment provides some normalization of the input images, there are several factors that may still result in gaze estimation errors. In particular, eye and iris sizes, or eye shape variations may result in an unusual eye corner estimate, or a biased gaze estimate (e.g. a different correlation between eyelid opening and gaze direction, as such correlation play an important role when looking down) or both. Below we propose a simple adaptation method to correct such errors, assuming that they are consistent, i.e. they result in a systematic bias.

If the algorithm would know where the subject is looking, it could compute a gaze error and perform some on-line calibration. In our case, as part of the error seems to be a constant bias, even a few frames should be enough to compute it and compensate for it. In this view, the problem can then be separated in two aspects: (1) selecting N frames with high

[1]Dlib C++ library: http://dlib.net/

probability of looking at a known target and (2) computing the bias based on those frames.

Interaction Prior Based Frame Selection. In dyadic or multiparty interactions, other people are the most predominant targets and we would like to select frames with high probability of looking at any of the participant. In absence of any knowledge about the specific interaction, we resort to the conversational dynamics and the fact that people usually follow some (implicit) rules when behaving in a social context. In multiparty interaction, such a rule is that people tend to look at the speaker (eg to show attentiveness or look for facial expressions) [12]. For instance, in [2], it is shown that in 4-party meetings involving slide presentations, people look 45% of the time at other people. However, there are 5 to 8 times more chances to look at a speaking person than at a non speaking one. Thus, *the speaking status of a partner* can be used to gather frames where the subject under study is looking with high probability at that partner.

Secondly, we use *the head pose information* to further increase the precision of the selection, that is, the chances that frames correspond to actual looking event (and exclude for instance moments looking at table, slide). More precisely, our pose-compatibility strategy simply verifies that the subject can physically look at the person. We thus remove frames for which the angular distance between the head direction and the direction to the visual target is more than 30°as this would lead to a really uncomfortable gaze behavior [23].

Finally, in our data, valid frames according to the above criteria are collected during the first minute of the interaction, out of which up to N are randomly selected to define the frame sample set $\mathcal{F}$.

Robust Bias Estimation. As the frameset $\mathcal{F}$ is expected to contain frames with wrong labels, we resort to a robust estimator, the Least Median of Squares (LMedS) to estimate the bias $\mathbf{b} = (\mathbf{b}_\phi, \mathbf{b}_\theta)$. More precisely, denoting by o_i the angular difference between the estimated gaze and the angular value of the direction to the visual target, we optimize

$$\mathbf{b}^{med} = \min_{\mathbf{b}} \underset{i \in \mathcal{F}}{\text{Median}}\ r_i^2, \quad \text{with } r_i = \|o_i - \mathbf{b}\|. \tag{2}$$

As there is no closed form solution, the method resorts by setting $\mathbf{b}$ to each value of o_i and ranking the residuals r_i to measure the median. The value for which this median is minimum is our estimate $\mathbf{b}^{med}$. To improve the estimation efficiency of the LMedS estimator, we use as our final bias estimate $\hat{\mathbf{b}}$ the mean of the $0.5N$ errors o_i which were closest to $\mathbf{b}^{med}$. Thus, with this estimator, the implicit hypothesis is that subjects look at their partner half the time when this one speaks. Finally, the estimated bias $\hat{\mathbf{b}}$ can be used to derive the attention as given by Eq. 1.

4 EXPERIMENTS

In this Section, we first present our experimental protocol before discussing the results and approaches.

4.1 Study Case and Experimental Protocol

Dataset. We perform our experiments on the UBImpressed dataset [19], which involves students from an hosteling school participating in two different dyadic interaction scenarios, whose setup (samples images) are illustrated in Fig. 1.

Interview scenario. The applicant and the interviewer are sitting in front of each other at a distance of around 2 meters. It represents a formal type of social interaction, with constrained behaviors and rather frontal faces.

Desk scenario. In this role simulation, the students plays a receptionist having to deal with the complains of a difficult client. Persons are standing and talk to each other in more open and animated fashion. There are also moments when the receptionist use the phone or discuss a bill and hotel rates on the desk with the client. Due to the setup (see the typical viewpoint in Fig. 1) and scenario, this creates a larger diversity of challenging body and head movements as well as gazing behaviours.

The videos are acquired with two Kinect 2 sensors at 30 fps. The camera is set about one meter away from the persons and are not totally in front of them. For each session, two cameras are used, each one recording a different person. The Fig. 1 presents typical images of the dataset. For our experiments, we worked with 8 interactions (16 videos in total): 4 interviews and 4 desk scenarios.

Ground Truth (GT) Annotations. Our goal is to automatically detect if the subject is looking at the other person or not using the classification process described in Sections 2 and 3, with the target being the other person represented by the middle point between her two eyes. To compute a correction bias and assess methods, we annotated whether the subject is looking at the partner or not (binary annotation) for each frame (i.e. every 33ms) of several segments of the videos. We annotated the first minute (ignoring the first 10 seconds of each video), and then 10 seconds every minute over 5 minutes, allowing to see how the error evolves over time under more head pose and gaze variations. For the 16 videos, it gives a total of over 42000 annotated frames (2643 annotations on average per video, some videos being shorter than others), with 52% of the frames with the "gazing" label (looking at the other person) and 48% with the "not gazing" one. The frames where a person is blinking or where the class is ambiguous were excluded from evaluation, but not from the automatic frame selection methods for computing the bias.

Moreover, for each interaction, the sound was synchronously recorded by a microphone array automatically detecting in a robust and precise fashion the beginning and end of the utterance segments of each person. Thus, we know for each gaze annotated frame the looking and speaking status.

Performance Measures. To compare methods, we use two metrics: the average gaze angular error and the classification accuracy. The gaze angular error is simply the angle difference e introduced in Section 2. That is, for frames for which the subject is looking to the partner according to the GT, the angle between the gaze vector and the direction to the middle

of the partner's eyes is our estimation error. Although people might look at other parts of the face we use the middle of the eyes for simplicity, as it is more or less at the center of the head, and people tend to look at the others' eyes during interactions. The average angular error is thus the average of the per-subject mean angular error computed over all frames annotated as 'gazing' after the first minute. Then, as final score, we computed the average of these mean angular errors.

The classification accuracy is measured at the frame level, i.e. as the percentage of gazing/not gazing frames classified correctly, using as threshold $\tau = 10$. The reported accuracy is the mean of per subject mean accuracy.

Evaluated Methods. The *Baseline* is the results obtained from the framework presented in Section 2 without modification. We then compare methods relying on either of the eye alignment methods presented in Section 3.1, denoted *Lm* (alignment based on per-frame detected eye corner locations) and *AvgLm* (alignment based on a running average over 1s of these frame-based locations), the latter allowing to find a stable point much less affected by blinks and short looks down.

For the bias correction presented in Section 3.2 we evaluated the following aspects:

- **Speech based** (*Spk*) frame selection with exclusion of high head pose - target direction discrepancy, as described in Section 3.2, collected over the first minute of the video.
- **Ground truth** (*GT*) frame selection, where frames for bias computation are randomly selected among the ones that are manually annotated as "gazing" in the first minute of the video. It allows to see how the method can work in presence of an 'oracle', and the maximum gain we could obtain.
- **Mean** (*Mean*) bias estimation: computing the bias as the mean of the error on N of the selected frames. In practice, we found that N higher than 20 only marginally improve the gaze estimation, so we used N =20 in experiments.
- **Median** (*Med*) bias estimation. As described in Section 3.2, the bias is computed as the mean of the error over the $\frac{N}{2}$ frames resulting from applying the LMedS estimator on N frames selected using the *GT* or *Spk* criteria. For this method, we used $N = 40$, so that the number of frames remaining after exclusion of the 50% outliers thanks to the LMedS is 20, the same number used in the *Mean* method.

4.2 Quantitative Results

Results - Eye Alignment. The Tab. 1 presents the results for the different methods (different combination of processing), for each scenario (desk and interview) and altogether.

As can be seen, the *Baseline* has difficulties to handle our data, with more than 23°errors. The landmarks alignment improves the performances significantly, making it a better method than estimating the eye position from the 3DMM information. One can notice that averaging the landmarks over time results in a better alignment and gaze estimation accuracy (with 1°of error lower than *Lm*), but this does not bring a better accuracy overall. However, the performance is more regular across scenarios and *AvgLm* performs better in the Desk scenario where more extreme head poses are present.

Table 1: Mean angular error (in degree) and classification accuracy (in percent) depending on the scenario

	Interviews		Desk		Overall	
Method	error	accuracy	error	accuracy	error	accuracy
Baseline	21.04	0.53	26.19	0.59	23.62	0.56
Lm	10.41	0.72	15.24	0.62	12.82	0.67
AvgLm	9.79	0.67	13.72	0.66	11.76	0.67
GT-Mean	7.42	0.75	10.31	0.75	8.86	0.75
GT-Med	7.53	0.75	10.45	0.74	8.99	0.74
Spk-Mean	8.49	0.66	10.59	0.74	9.54	0.70
Spk-Med	9.58	0.68	10.25	0.75	9.91	0.72
Lm + GT-Mean	5.61	0.85	8.85	0.82	7.23	0.84
Lm + GT-Med	5.67	0.84	8.94	0.82	7.30	0.83
Lm + Spk-Mean	7.92	0.73	9.26	0.80	8.59	0.77
Lm + Spk-Med	6.25	0.82	9.03	0.81	7.64	0.82
AvgLm + GT-Mean	6.08	0.82	9.34	0.80	7.71	0.81
AvgLm + GT-Med	6.44	0.82	10.26	0.78	8.35	0.80
AvgLm + Spk-Mean	8.39	0.72	10.65	0.74	9.45	0.73
AvgLm + Spk-Med	6.66	0.82	9.49	0.79	8.07	0.80

Results - Bias Correction. Applying the bias correction alone also improves the results even more than the eye alignment. This shows that the errors are not random, but really results from some user specific cropping or eye shape that generate a coherent bias (see also below and Fig. 7).

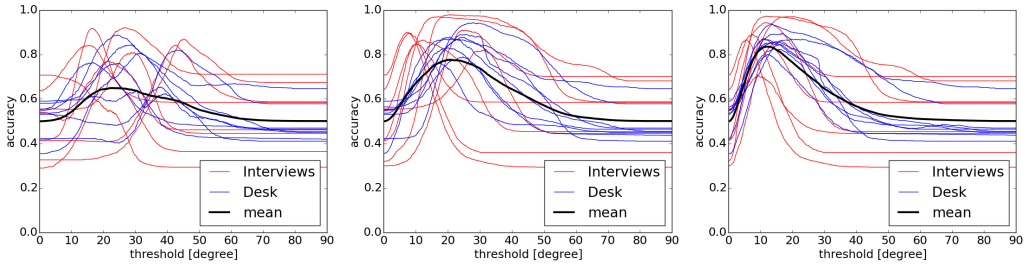
Results - Alignment and Bias Correction. Interestingly, we can note that both eye alignment and bias correction improve the results, but that their combination works even better, showing their complementarity. It can be explained by the difficulty to accurately place landmarks and specific eye shapes which affect the gaze and which needs to be corrected afterwards. Furthermore, we can see that in general, this reduction in error from the bias correction does not apply mainly to the short time after the correction, but last throughout the interactions, as shown by Fig. 5.

Also, surprisingly, after applying bias correction, higher results are reached when *Lm* was used rather than *AvgLm*. Our hypothesis is that despite being less efficient in itself to correct the error, *Lm* provides in average more stable eye images to the gaze estimator, making it easier to correct the remaining bias afterwards.

One can notice that the *Med* brings nothing compared to *Mean* when using GT frames, which was to be expected since samples are all valid. However, it becomes extremely useful when using other clues (like speech in our case) to sample the frames where people look at partners with high probability since erroneous guesses are expected and can badly affect the bias computation.

Finally, it is not surprising to see the *GT-Mean* (with *Lm*) correction obtaining the bests results, but the *Spk-Med* one is not far behind, giving a good hope for on-line experiments.

Impact of the Threshold. Fig. 6 presents the classification



(a) *Baseline* (b) *Lm* (c) *Lm + Spk- Med*

Figure 6: Classification accuracy for different thresholds (bold black line: mean accuracy)

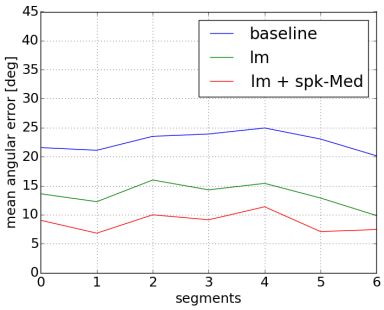


Figure 5: Error evolution over time. Average gaze error for each segments, where each segment happens x minutes after the first interaction minute where the bias is computed.

accuracy on each video depending on the chosen threshold. In the *Baseline* condition, no threshold is really suitable for all videos, as all peaks are at different places. It is highlighted by the mean accuracy over all videos, which is rather flat and only reaches 63% accuracy. It means that an adaptation of the threshold for each subject is needed to obtain the best results, which might be quite hard, and in any case just denote the erroneous gaze predictions.

Fig. 6b,c) show the same plots after *Lm* and *Lm + Spk-Med*. Clearly, the peaks tend to gather and an optimal threshold appears. In the *Lm + Spk-Med* case, all accuracies for a 10°threshold are between 70% and 95%, which is far better than the *Baseline*. It shows that the alignment and correction do not only improve performance, but also make the gaze estimation quality more uniform across videos.

Bias Analysis. Fig. 7 presents the mean angular error obtain on each video in term of (ϕ, θ). For the *Baseline*, the errors have a general tendency to the bottom, but the variance is high. After applying the *Lm* alignment, errors are closer to zero, the range of error is similar on both axis, the desk scenario is more challenging. It is consistent with previous results, showing that landmarks improve gaze estimation but that enhancement is still possible.

After the *Lm + Spk-Med* correction, mean angular errors are packed around zero and do not exceed 10°on both axis. This shows that the bias was well estimated for each subject.

These plots give also some hints on how this two step correction work: landmarks alignment acts as a normalization, making the eye images more uniform before passing them to the appearance model. Then, the bias correction is able to more efficiently compensate errors made by the model and provides some adaptation to each subject.

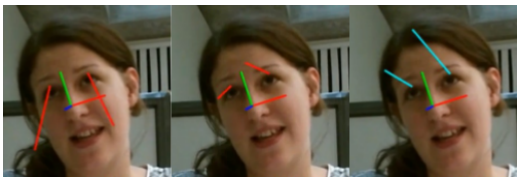


Figure 8: Qualitative comparison between three methods. From left to right: *Baseline*, *Lm*, *Lm + Spk-Med*. Gaze arrays: blue if classified as "gazing", red otherwise.

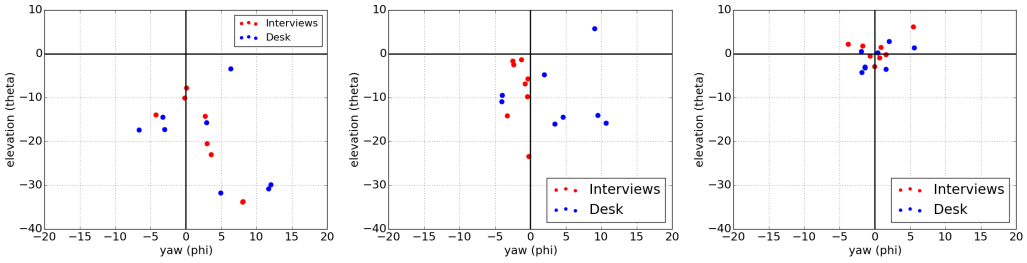
4.3 Qualitative Results

Fig. 8 presents a typical comparison between three methods on the same frame. More example are shown in the supplementary material video. In our dataset, one generally notice that *Lm* lacks of robustness when the head is not aligned with the gaze direction and more importantly when the face has occlusions like for large head poses. Results are also dependend on subject. In a few cases, *Lm* gives similar or even higher accuracy than *Lm + Spk-Med*, but in other, the *Lm* perform can be bad, reaching even an accuracy as low as 40% in one case. Moreover, as previously mentioned, consistency across subjects and situation is a critical feature to build automatic gaze estimation systems, which is achieved by our system.

5 DISCUSSION

The *Baseline* performance are very low compared to those of the paper describing the method [7]. However we are facing a more challenging situation: the setup and illumination conditions are different, subjects are not facing the camera and behave freely in the frame of the given scenarios. Nevertheless, after having applied our method, t he angular error are comparable with [7], and are rather small if we consider the difficulty of the task.

There are two main limitations to the proposed approach. The first one is that although we make the assumption that the bias is time, head pose and target independent, the variability in our data does not allow to fully test the validity of this hypothesis: the interviews are rather constrained, with a relatively constant head pose-target configuration. The desk



(a) *Baseline* **(b)** *Lm* **(c)** *Lm + Spk-Med*

Figure 7: Per person average gaze error (in degrees) for each person in (ϕ, θ) representation.

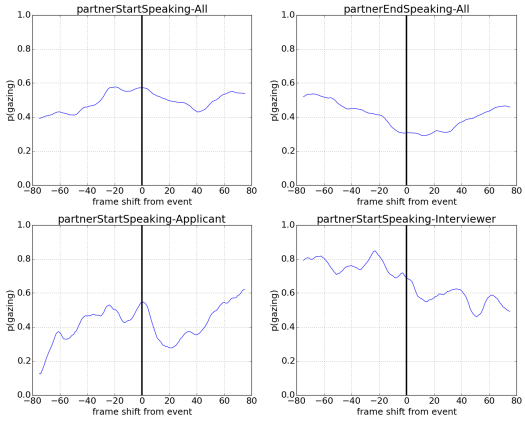


Figure 9: Probability that a subject looks at her partner before and after she ends (top right) or starts (three others) to speak. Statistics taken for all person (top), applicants (bottom left) and interviewers (bottom right)

situation already offers much more variability, but further validation in other scenarios, eg in multiparty situations with people looking at different sides, are needed to get a full (empirical) proof. More complex models taking into account these elements might be needed and more effective at reducing the gaze angular error, e.g. by computing another type of correction like an affine transform in function of the pose, or by using an online selection/adaptation scheme.

The second limitation is the need for frames where a subject looks at a known person (ie. with its location). Currently, the speaking status is used as a relatively high prior of being looked at. However, this assumption may not hold in presence of artefacts: listeners tend to look more at the artefacts/object than at the speaker. Currently, such situations are handled by the frame selection removal relying on non-compatible head pose for looking at the person. Other interaction settings, and especially multi-party situations, may also result in a weaker prior of looking at the speaker.

Using social interaction conventions as weak prior for adaptation may thus require more complex and subtle situation-based models. For instance, we started to study the relation between gaze and speaking turns, since it has been shown that people tend to look at the current or next speakers near a turn to potentially grab or acknowledge a floor change. We first computed the average probability that a subject looks at his partner when this one speaks, obtaining a result of 57% justifying the validity of our approach. We also computed the frame-based temporal evolution of the probability that a subject looks at his partner near speaker turn events (start or end of speaking turn). Some results are shown in Fig. 9. As can be seen, no clear trend for all people emerge, as the gaze probability seems to depend on the role of the person (compare applicants and interviewers). Nevertheless, these patterns show some tendencies which could be exploited as prior for adaptation if more strongly validated and combined within a more complex adaptation scheme with other multimodal cues like head pose, body pose, head gestures (nods), the gaze itself, and situational information (dialog act).

6 CONCLUSION

We presented methods to improve gaze estimation based on appearance based approached, methods that do not require a person-specific calibration phase, is robust to head movements and enables gaze estimation in a large field of view.

Our contributions consist of improving the eye cropping step using facial landmarks, and removing person and interaction specific errors by automatically estimating a gaze bias, relying on speaking turns information and head pose compatibility for looking at the person target, providing clues on whether the subject is looking at the speaker. These clues provide weak labels on the gazing direction of the subject, allowing to compute and compensate a gaze estimation bias.

The final multimodal method reaches a mean angular error of 7.64°and a mean classification accuracy of 82% for the 16 studied subjects, in two different scenarios. As it does not need any calibration, it is applicable on-line, providing an automatic adaptation to persons.

Future work will consist to test our method on different datasets, with different scenario including for example more people in the discussion.

ACKNOWLEDGMENTS

This research has been supported by the European Union Horizon 2020 research and innovation programme (grant agreement no. 688147, MuMMER project) and by the UBIMPRESSED project of the Sinergia interdisciplinary program of the Swiss National Science Foundation (SNSF).

REFERENCES

[1] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational Gaze Aversion for Humanlike Robots. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, New York, USA, 25–32.

[2] S Ba and J-M. Odobez. 2008. Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. Las-Vegas.

[3] Robert Bixler, Nathaniel Blanchard, Luke Garrison, and Sidney D'Mello. 2015. Automatic Detection of Mind Wandering During Reading Using Gaze and Physiology. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. ACM, New York, NY, USA, 299–306. DOI:https://doi.org/10.1145/2818346.2820742

[4] H.R. Chennamma and Xiaohui Yuan. 2013. A Survey on Eye-Gaze Tracking Techniques. *Indian Journal of Computer Science and Engineering* 4 (2013), 388–393.

[5] K Funes, L Nguyen, D Gatica-Perez, and J.-M. Odobez. 2013. A Semi-Automated System for Accurate Gaze Coding in Natural Dyadic Interactions. In *Proc. Int. Conf. on Multimodal Interfaces (ICMI), Sydney*.

[6] K Funes and J.-M. Odobez. 2014. EYEDIAP: A Database for the Development and Evaluation of Gaze Estimation Algorithms from RGB and RGB-D Cameras. In *Proc. of the Eye Tracking Research and Application (ETRA) conference*.

[7] Kenneth A. Funes-Mora and Jean-Marc Odobez. 2016. Gaze Estimation in the 3D Space Using RGB-D Sensors. *International Journal of Computer Vision* 118 (2016), 194–216.

[8] Daniel Gatica-Perez, Alessandro Vinciarelli, and Jean-Marc Odobez. 2014. *Nonverbal Behavior Analysis*. EPFL Press, Lausanne, CH, 165–187.

[9] Rui Hiraoka, Hiroki Tanaka, Sakriani Sakti, Graham Neubig, and Satoshi Nakamura. 2016. Personalized Unknown Word Detection in Non-native Language Reading Using Eye Gaze. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016)*. ACM, New York, NY, USA, 66–70. DOI:https://doi.org/10.1145/2993148.2993167

[10] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2016. Prediction of Who Will Be the Next Speaker and When Using Gaze Behavior in Multiparty Meetings. *ACM Trans. Interact. Intell. Syst.* 6, 1 (2016), 4:1—-4:31. DOI:https://doi.org/10.1145/2757284

[11] L. Jianfeng and L. Shigang. 2014. Eye-Model-Based Gaze Estimation by RGB-D Camera. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 606–610.

[12] A Kendon. 1967. Some functions of gaze direction in social interaction. *Acta Psychologica* 26 (1967), 22–63. DOI:https://doi.org/10.1016/0001-6918(67)90005-4

[13] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye Tracking for Everyone. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] Nanxiang Li and Carlos Busso. 2013. Evaluating the Robustness of an Appearance-based Gaze Estimation Method for Multimodal Interfaces. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI '13)*. ACM, New York, NY, USA, 91–98. DOI:https://doi.org/10.1145/2522848.2522876

[15] Nanxiang Li and Carlos Busso. 2014. User Independent Gaze Estimation by Exploiting Similarity Measures in the Eye Pair Appearance Eigenspace. In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI '14)*. ACM, New York, NY, USA, 335–338. DOI:https://doi.org/10.1145/2663204.2663250

[16] Feng Lu, Okabe Takahiro, Yusuke Sugano, and Yoichi Sato. 2014. Learning gaze biases with head motion for head pose-free gaze estimation. *Image and Vision Computing* 32 (2014), 169–179.

[17] Päivi Majaranta and Andreas Bulling. 2014. *Eye Tracking and Eye-Based Human–Computer Interaction*. Springer, London, UK, 334–341.

[18] AJung Moon, Daniel Troniak, Brian Gleeson, Matthew Pan, Minhua Zheng, Benjamin Blumer, Karon MacLean, and Elizabeth Croft. 2014. Meet Me where I'm Gazing: How Shared Attention Gaze Affects Human-Robot Handover Timing. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, New York, USA, 334–341.

[19] 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 *ICMI*. Tokyo, Japan.

[20] C Oertel, K Funes, J Gustafson, and J.-M. Odobez. 2015. Deciphering the Silent Participant: On the Use of Audio-Visual Cues for the Classification of Listener Categories in Group Discussions. In *Int. Conf. on Multimodal Interactions (ICMI)*.

[21] Pascal Paysan, Reinard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009. A 3D Face Model for Pose and Illumination Invariant Face Recognition. In *2009 IEEE International Conference on Advanced Video and Signal Based Surveillance*. Columbus, Ohio,USA, 296–301.

[22] Timo Schneider, Boris Schauerte, and Rainer Stiefelhagen. 2014. Manifold alignment for person independent appearance-based gaze estimation. In *Proceedings - International Conference on Pattern Recognition*. 1167–1172. DOI:https://doi.org/10.1109/ICPR.2014.210

[23] Brian Smith, Qi Yin, Steven Feiner, and Shree Nayar. 2013. Gaze Locking: Passive Eye Contact Detection for Human–Object Interaction. In *Proceedings of the 2013 symposium on User Interface Software and Technology*. ACM, New York, USA, 271–280.

[24] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2013. Appearance-Based Gaze Estimation Using Visual Saliency. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 2 (2013), 329–341. DOI:https://doi.org/10.1109/TPAMI.2012.101

[25] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2014. Learning-by-Synthesis for Appearance-Based 3D Gaze Estimation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. 1821–1828. DOI:https://doi.org/10.1109/CVPR.2014.235

[26] Roberto Valenti, nicu Sebe, and Theo Gevers. 2012. Combining Head Pose and Eye Location Information for Gaze Estimation. *IEEE Transactions on Image Processing* 21, 2 (2012), 802–815.

[27] Boris Velichkovsky, Sascha Domhoefer1, Sebastian Pannasch, and Pieter Unema. 2000. Visual Fixations and Level of Attentional Processing. In *Proceedings of the 2000 symposium on Eye tracking research and applications*. ACM, New York, USA, 79–85.

[28] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. 2015. Rendering of Eyes for Eye-Shape Registration and Gaze Estimation. In *Proc. of the IEEE International Conference on Computer Vision (ICCV 2015)* (2015-12-12).

[29] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. 2016. A 3D Morphable Model of the Eye Region. In *Proceedings of the 37th Annual Conference of the European Association for Computer Graphics: Posters (EG '16)*. Eurographics Association, Goslar Germany, Germany, 35–36. DOI:https://doi.org/10.2312/egp.20161054

[30] Yu Yu, Kenneth Funes-Mora, and Jean-Marc Odobez. 2017. Robust and Accurate 3D Head Pose Estimation through 3DMM and Online Head Model Reconstruction. In *2017 12th IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 711–718.

[31] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based Gaze Estimation in the Wild. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. 4511–4520.

[32] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2016. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. *CoRR* abs/1611.08860 (2016). http://arxiv.org/abs/1611.08860