Supervised Gaze Bias Correction for Gaze Coding in Interactions

Rémy Siegfried, Jean-Marc Odobez
IDIAP Research Institute, Martigny, Switzerland
École Polytechnique Fédérale de lausanne, Lausanne, Switzerland

Understanding the role of gaze in conversations and social interactions or exploiting it for HRI applications is an ongoing research subject. In these contexts, vision based eye trackers are preferred as they are non-invasive and allow people to behave more naturally. In particular, appearance based methods (ABM) are very promising, as they can perform online gaze estimation and have the potential to be head pose and person invariant, accommodate more situations as well as user mobility and the resulting low resolution images. However, they may also suffer from a lack of robustness when several of these challenges are jointly present. In this work, we address gaze coding in human-human interactions, and present a simple method based on a few manually annotated frames that is able to much reduce the error of a head pose invariant ABM method, as shown on a dataset of 6 interactions.

Keywords: appearance model, attention, bias correction, eye tracking, gaze, usability

Introduction

Eye movement analysis can provide rich information about a person's attention (Velichkovsky, Domhoefer1, Pannasch, & Unema, 2000; Ba & Odobez, 2006). In particular, gaze is a non-verbal communication cue playing a major role in human interaction (Gatica-Perez, Vinciarelli, & Odobez, 2014), and its accurate perception is a key factor of social interaction.

There are different methods to track eye movements (Chennamma & Yuan, 2013), but we focus on appearance based methods, as they allow people to behave more naturally by avoiding invasive devices and tolerating low resolution images (Funes Mora & Odobez, 2013).

In the online gaze extraction method presented in (Funes-Mora & Odobez, 2016), the frontal face image is extracted from the 3D mesh of the head, knowing the head pose. It allows to obtain eye images in the ideal position and to train a head pose and person invariant appearance model, at the price of a loss of accuracy. This is a minor drawback if we focus on attention extraction applications, which does not require an accurate point of gaze (Majaranta & Bulling, 2014).

Our goal is to apply this method to challenging data composed of real social interactions. In this paper, we show the difficulties of the above method for this scenation and present a supervised correction method to improve its robustness.

The 2017 COGAIN Symposium: Communication by Gaze Interaction
Wuppertal, Germany. August 19th and 21st, 2017

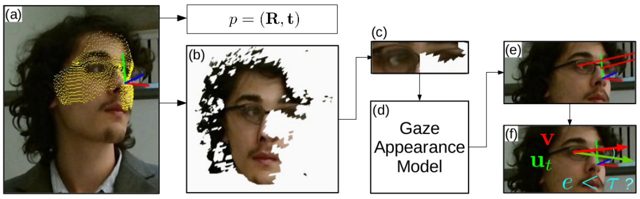

Figure 1. Gaze and attention framework (Funes-Mora & Odobez, 2016). a) The 3DMM mesh is robustly fitted to the RGB-D data to estimate the head pose $\mathbf{p}$. b) The frontal face image is computed by rotating and projecting the textured mesh. c-d) The gaze angles $\mathbf{g}$ are mapped from the eye images. e) The gaze direction $\mathbf{v}$ is computed from $\mathbf{g}$ and $\mathbf{p}$. f) The angle error e between the vector pointing to the target $\mathbf{u}_t$ and the gaze vector $\mathbf{v}$ is compared to a threshold τ.

Methods

We use a similar approach that the one presented in (Funes-Mora & Odobez, 2016), whose main steps are described in the Figure 1. It is based on the data provided by a RGBD camera and estimates (1) the head pose $\mathbf{p} = (\mathbf{R}, \mathbf{t})$, where $\mathbf{R}$ and $\mathbf{t}$ are the rotation matrix and translation vector to pass from the camera coordinate system to the head coordinate system, and (2) the gaze angles $\mathbf{g} = (\phi, \theta)$, where ϕ is the yaw and θ the elevation, which can be transformed into a gaze vector $\mathbf{v}(\mathbf{g})$.

Then, an attention decision is taken by comparing the vector pointing to a visual target $\mathbf{u}_t$ and the gaze vector $\mathbf{v}$:

$$\mathrm{e} = \arccos\left(\mathbf{u}_\mathrm{t} \cdot \mathbf{v}(\mathbf{g})\right) < \tau, \tag{1}$$

with e the angular distance and τ the decision threshold.

Thus, if e < τ, we consider that the person is looking at the visual target. This threshold takes into account both the noise of the method and the fact that a visual target is not a single point in space (the face of a person for example).

Gaze Bias Correction

This method performed well in the work of Funes Mora et al. (Funes-Mora & Odobez, 2016), but does not present similar results on our dataset (see Table 1). Our hypothesis is (1) that the eye positions are estimated from the theoretical eye position on the 3DMM mesh, which suffers some errors, and (2) that the training set of the gaze appearance model does not represent well the specific eye shape of everyone.

We propose to rectify the gaze by estimating the bias introduced by those weaknesses. Considering a frame where the subject is looking at the visual target, the components of the gaze ϕ and θ become the actual error of the method. The correction can then be computed by taking the mean $\mathbf{b}$ of those errors on n frames and subtracting it from the gaze $\mathbf{g} = (\phi, \theta)$ on the whole video to compensate the bias. Ultimately, the error becomes $e = \arccos(\mathbf{u}_t \cdot \mathbf{v}(\mathbf{g} - \mathbf{b}))$ For now, the information of when the subject is looking at the other person comes from the ground truth (i.e. manual annotations).

Data used for experiment

We test our correction method on 12 videos of the UBImpressed dataset, taking 8 videos of the interview scenario and 4 videos of the desk scenario.

- the "interviews" scenario shows an applicant in front of an interviewer in a formal situation where both persons are sitting in front of each other.
- the "desk" scenario shows a receptionist that answer to the question of a client. This scenario is more challenging, as both persons are standing and talk to each other more naturally, increasing the variety of head movements and gaze behaviors.

We annotated 2800 frames per videos, writing down when a person is looking at the other or not. To compute the error, we used the middle eye point of the other person as visual target, whose 3D position is known using the same head pose tracker. Studying the effect of the number of frames n on the resulting angular error revealed that increasing it over 20 does not improve significantly the gaze estimation.

Results

The Table 1 presents the angular error before (n = 0) and after a correction with n = 20. It shows also the classification accuracy, which is basically the number of correctly annotated frames over the total number of annotations.

Looking at the results without correction, one can see big angular errors and low accuracy that increases with the threshold. However, increasing the threshold reduces also the resolution of the attention decision.

Table 1
Mean angular error and classification accuracy

Conditions			Mean	Mean classification accuracy			
Metric	Scenario	n	angular error	τ=5	τ=10	τ=15	τ=20
Mean	Interviews	0	21.04	0.49	0.53	0.61	0.63
Mean	Interviews	20	8.84	0.59	0.72	0.67	0.62
Mean	Desk	0	22.29	0.56	0.59	0.64	0.71
Mean	Desk	20	13.09	0.63	0.71	0.73	0.70
Std dev	Interviews	0	10.04	0.15	0.16	0.20	0.18
Std dev	Interviews	20	2.78	0.12	0.10	0.10	0.16
Std dev	Desk	0	8.55	0.02	0.03	0.07	0.12
Std dev	Desk	20	4.04	0.02	0.06	0.11	0.13

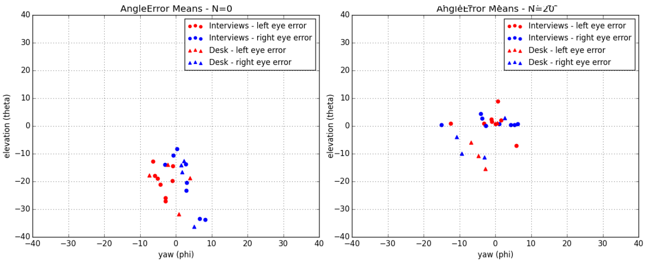

Figure 2. Angles differences in degrees for each person

After correction, the angular error decreased in both scenarios, but more on the "interviews" one. The best threshold is not the same for both scenarios, but fixing a threshold at τ=10° would give an average of 72% accuracy, which is much better than the method without correction.

The Figure 2 represents the mean error of each video in a (ϕ, θ) representation, before and after the correction. It validates that the bias decreased and shows that the angular errors are more uniform across video. Thus, the correction enables to compensate the specificities of the different subject, i.e. provides a person-specific calibration to the method.

Conclusions

In this work, we present a method that successfully reduces the error of the gaze extraction, despite a challenging dataset where subjects are not staying still. We reduced the error to 10.26° on average, retrieving the error range claimed in (Funes-Mora & Odobez, 2016)

Future work will consist to investigate other methods to improve the eye positions estimation and ways to perform the correction on-line, like using the speech recordings to guess when the subject looks at the other person.

Acknowledgements.
This research has been supported by the "Valais-Wallis Ambition" program of the Idiap research institute, and the European Union Horizon 2020 research and innovation programme (grant agreement no. 688147, MuMMER project. The authors declare that there is no conflict of interest regarding the publication of this paper.

References

Ba, S., & Odobez, J. (2006, May). A study on visual focus of attention recognition from head pose in a meeting room. In *Proc. workshop on machine learning for multimodal interaction (mlmi).* Washington DC.

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

Funes Mora, K. A., & Odobez, J.-M. (2013, sep). Person Independent 3D Gaze Estimation From Remote RGB-D Cameras. In *International conference on image processing.* IEEE.

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

Gatica-Perez, D., Vinciarelli, A., & Odobez, J.-M. (2014). Nonverbal behavior analysis. In *Multimodal interactive systems management* (pp. 165–187). Lausanne, CH: EPFL Press.

Majaranta, P., & Bulling, A. (2014). Eye tracking and eye-based humanâĂŞcomputer interaction. In *Advances in physiological computing* (pp. 334–341). London, UK: Springer.

Velichkovsky, B., Domhoefer1, S., Pannasch, S., & Unema, P. (2000). Visual fixations and level of attentional processing. In *Proceedings of the 2000 symposium on eye tracking research and applications* (pp. 79–85). New York, USA: ACM.