PERSON INDEPENDENT 3D GAZE ESTIMATION FROM REMOTE RGB-D CAMERAS

Kenneth Alberto Funes Mora and Jean-Marc Odobez

Idiap Research Institute, CH-1920, Martigny, Switzerland
École Polytechnique Fédérale de Lausanne, CH-1015, Lausanne, Switzerland

ABSTRACT

We address the problem of person independent 3D gaze estimation using a remote, low resolution, RGB-D camera. The approach relies on a sparse technique to reconstruct normalized eye test images from a gaze appearance model (a set of eye image/gaze pairs) and infer their gaze accordingly. In this context, the paper makes three contributions: (i) unlike most previous approaches, we exploit the coupling (and constraints) between both eyes to infer their gaze jointly; (ii) we show that a generic gaze appearance model built from the aggregation of person-specific models can be used to handle unseen users and compensate for appearance variations across people, since a test user eyes' appearance will be reconstructed from similar users within the generic model. (iii) we propose an automatic model selection method that leads to comparable performance with a reduced computational load.

Index Terms— 3D gaze estimation, appearance based methods, sparse reconstruction, person-independence.

1. INTRODUCTION

Gaze is recognized as an important attentional and non-verbal communication cue, for which it has gained increased attention from fields such as psychology, sociology and robotics, to mention a few. Thus, gaze estimation has been studied for over 3 decades[1] and solutions have emerged from the Human-Computer Interface (HCI) field and led to highly accurate systems. Such systems are often expensive, and are either invasive (eg using head-mounted cameras [2]) or based on specialized hardware like infrared (IR) light sources and sensors [3]. Moreover, they often require the user to cooperate and perform certain actions like a calibration session in which the system trains model parameters from gaze observations to specific points in a screen.

In many applications, however, user cooperation is limited, or even, nonexistent. In such situations, a remote camera with enough field of view to accomodate the person mobility is desired. This leads to the challenge of low resolution imaging and the extreme difficulty to track local eye features. Thus, an alternative to feature based approaches [1] is needed.

Appearance based methods have gained increased attention recently [2, 4, 5, 6, 7, 8]. Thanks to the learning of a direct mapping from the eye image, or a holistic description of it, to the gaze parameters, they do not need to explicitely track features like pupils, iris or glints [1], and have shown potential for gaze estimation under low-resolution imaging [5, 6].

There have been many proposals to learn this image to gaze mapping. Baluja and Pomerlau [4] trained a neural network but required a few thousands of training samples (2000). Williams et al. proposed to use Gaussian Process Regression [9] in a semi-supervised manner, to reduce the needed samples. Sugano et al. [8] proposed an incremental approach, taking user-computer interaction as training data. While this is a valuable strategy, it is not always applicable. Feng et al. [6] proposed Adaptive Linear Regression (ALR) that interpolates the gaze parameters of a test sample from fewer training examples. High accuracy was obtained, even using low-resolution test images, but experiments were conducted using a chin-rest, ie. assuming a single head pose. They subsequently proposed to learn the mapping for a fixed head pose, and then correcting, using Gaussian Process Regression, the gaze direction due to head pose variations [7]. Alternatively, Funes and Odobez [5] proposed to use an RGB-D camera to correct the eye appearance variation due to head pose. This method generated frontal looking eye images, which were then processed using the ALR method. The estimated gaze is then corrected according to the estimated head pose.

Nevertheless all these methods were trained and tested on the same person. In this paper we address the more challenging problem of gaze estimation for an unseen person, from low resolution ($\sim 15\times10$ pixels per eye) and remote cameras. These properties are highly desired in many applications.

We build upon our previous work [5] and propose to create a generic gaze model by aggregating gaze appearance models from different people. Using a sparse reconstruction we obtain a soft selection of models in the training set, such that they are more appropriate to the test person, reducing the error. In order to achieve robustness we also add constrains which encode the coupled movements of the left and right eye and we show that this improves the resulting estimation under these adverse conditions.

The authors gratefully acknowledge the financial support from the Swiss National Science Foundation (Project: FNS-203, TRACOME). www.snf.ch and from the HUMAVIPS project, funded by the European Commission Seventh Framework Programme, Theme Cognitive Systems and Robotics, Grant agreement no. 247525.

2. PROPOSED METHOD

In this section we first describe our method overview, followed by a description of the proposed coupling constrains and gaze estimation from appearance for an unseen person.

2.1. Method overview

The overall gaze estimation procedure follows closely our previous work [5], which we briefly describe here: a 3D mesh of the user's face is built offline for each individual. To this end we use a non-rigid iterative closest points algorithm [10] to find the weights of a 3D Morphable Model (3DMM) [11] that fit best a few snapshots of the individual's face. Then, in an online stage, the following steps are executed:

a) Using the personalized 3D mesh (template) we use a frame-by-frame (online) iterative closest point (ICP) method to fit the template's 3D pose to the depth data. In this manner we obtain, for frame t, the head's pose $\mathbf{p}_t = \{\mathbf{R}_t, \mathbf{t}_t\}$, composed of a 3D rotation and translation.

b) Assuming a calibrated RGB-D setup, the RGB-D frame can be transformed to a textured 3D mesh. We then re-render the texture, lying in the 3D surface, using the inverse of the head pose parameters, i.e. $\mathbf{p}_t^{-1} = \{\mathbf{R}_t^\top, -\mathbf{R}_t^\top \mathbf{t}_t\}$. Using this procedure we remove the eye appearance variation due to head pose, as we obtain facial images as if the head was static and in front of the camera. As we use a 3DMM to generate the 3D template, semantic information, such as the eyes location, is predefined and propagated during tracking. Therefore we know, frame-by-frame, the position of the eyes, which is used to crop eye images from the frontal looking facial texture.

c) Using these frontal-looking eye images we apply the appearance based gaze estimation algorithm to retrieve the gaze direction. This will be the main focus of this paper.

d) The final step is to transform the gaze direction back to the world coordinate system. This transformation is given by the head pose parameters.

In the following sections we describe in detail step c.

2.2. Appearance based gaze estimation

The goal of the gaze estimation algorithm is to obtain the gaze parameters $\hat{\mathbf{g}}$ given a test image $\hat{\mathbf{I}}$. To that end we take a sparse image reconstruction approach. Assuming a training set of eye images, with associated gaze directions $\{(\mathbf{I}_i, \mathbf{g}_i)\}$, we aim to infer the weights w_i which reconstruct best the test image $\hat{\mathbf{I}}$ from the linear combination of $\{\mathbf{I}_i\}$.

In addition we allow only a few w_i's to be different than zero. Ideally this would retrieve only the samples which are close in the gaze space to the test image.

As we assume the eye images are frontal, due to our rectification procedure, we represent the gaze direction by the angles $\mathbf{g} = (\phi, \theta)$, c.f. Fig. 1a. Here θ is the gaze elevation, and ϕ is the gaze yaw. Given this we can estimate the test gaze direction as $\hat{\mathbf{g}} = \sum_i w_i \mathbf{g}_i = (\mathbf{g}_\phi^\top \mathbf{w}, \mathbf{g}_\theta^\top \mathbf{w})$, where $\mathbf{w}$ is the column vector composed of the w_i's, and $\mathbf{g}_\phi$ and $\mathbf{g}_\theta$ represent the column vectors of concatenated ϕ_i's and θ_i's respectively.

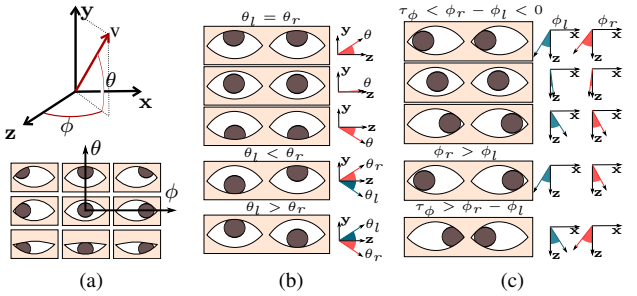

Fig. 1: a) Gaze direction parameterization. b-c) Coupled eye gazing constrains on b) gaze elevation c) gaze yaw. For both cases we show the 3 examples following the constrains and 2 examples which break the constrains.

The use of the same weights to estimate the gaze parameters is justified by locality, as they implicitly represent a linear mapping between $\hat{\mathbf{g}}$ and $\hat{\mathbf{I}}$. This method was initially proposed by Feng et al. [6], which they called Adaptive Linear Regression (ALR). In our previous work we proposed its application to head pose invariant 3D gaze estimation [5].

Rather than using the raw eye image $\mathbf{I}$ we build a descriptor $\mathbf{e}$ from it. This procedure is described in the following section. Therefore, given the training examples $\{(\mathbf{I}_i, \mathbf{g}_i)\}$, we call *gaze appearance model* $\mathcal{A}$ the set $\{(\mathbf{e}_i, \mathbf{g}_i)\}$. During test, for an image $\hat{\mathbf{I}}$, with descriptor $\hat{\mathbf{e}}$, we will infer the gaze direction $\hat{\mathbf{g}}$. Let E be the matrix whose column i correspond to $\mathbf{e}_i$, and ϵ a tolerance parameter. Finding the optimal $\mathbf{w}$ is formulated as a sparse reconstruction problem, by minimizing the L1 norm of $\mathbf{w}$, as shown in Eq. 1.

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} ||\mathbf{w}||_1 \quad s.t. \quad ||E\mathbf{w} - \hat{\mathbf{e}}||_2 < \epsilon \qquad (1)$$

Eye image representation: Instead of using directly an eye image $\mathbf{I}$, we compute a descriptor $\mathbf{e}$ as follows: we convert the image to gray-scale, then we normalize the intensity values by setting their mean to 125 and standard deviation equal to 30 (given that the pixel intensity is initially in the range $[0, 255]$. This is done to acquire robustness to global illumination changes. After normalization the image is divided into a grid of m rows and n columns. At each bin $j = (m, n)$, the sum of pixel intensity S_j is computed. The descriptor $\mathbf{e}$ is then given by the concatenated S_j values, and normalized such that $\sum_j e^j = 1$. We used $m = 3$ and $n = 5$.

2.3. Coupled eyes gazing constrains

The method previously described is applicable to the left ("l") and right ("r") eye to obtain their gaze directions separately $(\hat{\mathbf{g}}_l, \hat{\mathbf{g}}_r) = ((\phi_l, \theta_l), (\phi_r, \theta_r))$, as done in previous works. However, both eyes work jointly to fixate at a specific point in a 3D space. Here we extend ALR to build upon this fact.

As a first observation, if the eyes are horizontally aligned, then their gaze elevation should be the same, i.e. $\theta_l = \theta_r$ as shown in Fig. 1b. This allows us to represent the gaze elevation, for both eyes, as a single parameter θ.

The second observation is that $\phi_r < \phi_l$ is always fulfilled such that a *single* 3D point is observed. Equality occurs when gazing a point at an infinite distance. We can also limit the closest a 3D point is expected to be by setting $\phi_r - \phi_l > \tau_\phi$. Where τ_ϕ is a constant. This is observed in Fig. 1c.

We formalize these observations by estimating the left and right gaze directions jointly by solving the following constrained optimization problem:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} ||\mathbf{w}||_1 \quad s.t. \quad ||E\mathbf{w} - \hat{\mathbf{e}}||_2 < \epsilon \tag{2}$$
$$\mathbf{g}_{\theta_l}^\top \mathbf{w}_l - \mathbf{g}_{\theta_r}^\top \mathbf{w}_r = 0$$
$$\tau_\phi < \mathbf{g}_{\phi_r}^\top \mathbf{w}_r - \mathbf{g}_{\phi_l}^\top \mathbf{w}_l < 0$$

Where we redefine $\mathbf{w} = [\mathbf{w}_l^\top, \mathbf{w}_r^\top]^\top$, $\hat{e} = [\hat{e}_l^\top, \hat{e}_r^\top]^\top$ and E is the following block matrix:

$$E = \begin{bmatrix} E_l & \mathbf{0} \\ \mathbf{0} & E_r \end{bmatrix} \tag{3}$$

Given that $\mathbf{0}$ is a zero filling matrix. Eq. 2 can be solved as a Second Order Cone Programming (SOCP) problem. We call this method coupled adaptive linear regression (CALR).

2.4. Gaze appearance model

In the previous sections it is assumed that training samples are given to create a gaze appearance model $\mathcal{A} = \{(\mathbf{e}_i, g_i)\}$. Here we discuss different possibilities to obtain $\mathcal{A}$.

We will call *own* model when estimating gaze for a person k using its model $\mathcal{A}_k$, provided by a training phase where samples $\{(\mathbf{e}_i^k, g_i^k)\}$ were collected.

However, for some applications, a training phase is not always available. One solution, which we call *cross* model, is to use $\mathcal{A}_{j|j\neq k}$, i.e. the model built for another person j. However the selection of which model to use has to be done carefully, as the appearance variation between two people might be large, as can be seen in Fig. 2.

Another solution, which we call *generic* gaze appearance model, is to aggregate the person specific models available as $\mathcal{A}_\mathcal{G} = \cup\{\mathcal{A}_j\}_{j\neq k}$. This approach allows to interpolate appearance across people, in addition to across gaze directions. However, it is computationally expensive as the pool of samples becomes large, together with their associated weights.

Therefore, we propose what we call *selected* model. It is based on the observation that, for the generic model estimation, we can compute the weight given to each person j as $W_j = \sum_{i|i\in\mathcal{A}_j} |w_i|$. If the test image radically differs from the samples of person j, then it is expected for W_j to tend to zero due to sparsity. Whereas, if the eyes appearance between two people is similar, then W_j will be high. By accumulating the W_j's through time, we can rank the models according to their relevance to the test person. Thus we create the selected model $\mathcal{A}_\mathcal{S}$ as the subset of models with the highest weights.

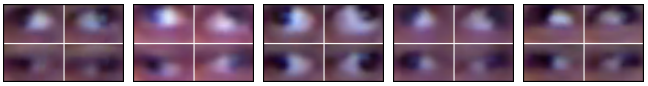

Fig. 2: Eye image samples taken from the gaze appearance models of 5 different participants (left eye).

3. EXPERIMENTS

3.1. Data collection

We collected data from 5 different people using our method proposed in [5]. We used a Kinect sensor from Microsoft as RGB-D camera. The participant was recorded following with the gaze a 3cm ball moving between the person and the camera. The head pose and the ball are tracked automatically. Using the head pose, and the position of the visual target, we compute the ground-truth gaze direction.

For each person we recorded two sessions; in the first one the participant was asked to keep the head pose fixed and facing the camera. We call this session *frontal*. The first half was used to collect the samples for the gaze appearance model. In this paper it is composed of 42 samples approximately uniformly distributed in the gaze space. In a range of $[-50°, 50°]$ for ϕ and $[-40°, 40°]$ for θ. The second half of the frontal session was used for evaluation. In the second session, which we call *free*, the person was requested to do (challenging) head movements while gazing the visual target.

3.2. Gaze estimation experiments

We conducted different experiments to validate our method. In general they differ according to the used formulation for gaze estimation (ALR or CALR), the evaluation session (frontal or free), and the used appearance model.

Gaze apperance models evaluation. Table 1 reports the obtained angular errors according to the selection of the gaze appearance model, described in Section 2.4. In addition we provide results when the gaze direction is assumed to be that of the head pose, i.e. $(\phi, \theta) = (0, 0)$. We call this *Head* model. From the large errors obtained in the Head model evaluation we observe the large variation within the data.

In the case of the cross model, for each participant there were 4 models, 1 per each other participant. Here we report the obtained average error between these 4 evaluations.

As can be seen, the best results are obtained when using the own model, as expected. The worse case is when cross model estimation is used, due to large eye appearance variation between participants. Nevertheless, the results are improved when aggregating different models into a single

Table 1: Effect on mean angular gaze error (°) due to changing the gaze appearance model used for evaluation. We used CALR as gaze estimation method.

		Test person					
Model	Session	1	2	3	4	5	Avg
Head	Frontal	28.2	29.8	28.5	28.6	34.0	29.8
	Free	26.1	21.9	21.9	24.8	20.1	23.0
Own	Frontal	8.5	5.9	6.8	9.0	7.6	7.6
	Free	16.7	9.7	17.3	11.9	8.8	12.9
Cross	Frontal	13.7	12.5	15.4	13.1	15.1	13.9
	Free	22.0	15.9	20.5	16.9	14.6	18.0
Generic	Frontal	13.9	9.4	11.9	9.6	10.2	11.0
	Free	14.1	12.5	16.6	15.5	10.6	13.9
Selected	Frontal	12.8	9.3	11.4	9.9	11.1	10.9
	Free	28.4	12.3	17.6	13.7	12.0	16.8

generic model. We observe that in average the results become comparable to the own model in the free session, where the conditions are less ideal.

Model selection. In Section 2.4 we assumed that the sparse reconstruction weights provide a mean for model selection for an unseen person. To validate this hypothesis we present detailed results of the cross model estimation in Table 2. Simultaneously we present the weights assigned to each separate person during the generic model estimation in Fig. 3. We can observe a correlation between these weights and the resulting error given by the cross model estimation, showing that the weights automatically rank the models according to how appropriate they are for estimation on an unseen person.

We selected the 2 models with the highest weights to create the selected model, as shown in Fig. 3. The results are shown in Table 1. As can be seen, there is clear decrease of the error with respect to hand-picking a single model (cross model) and the results are comparable to the generic model case, except for one case in the free head pose session.

Coupled eyes constrains. In Table 3 we compare CALR with ALR. We include a session using the own model (for ALR this is equivalent to [5]) but focus on the generic model case. As can be seen, for most of the cases, there is a reduction of the error when using CALR over ALR. We found this to be consistent, even for cross and selected models evaluations.

Table 2: Mean angular gaze error (°) for cross model gaze estimation. Evaluated using CALR in the frontal session.

		Used model				
		1	2	3	4	5
Test person	1	-	14.4	13.3	14.6	12.5
	2	14.3	-	13.0	14.7	7.8
	3	10.4	17.5	-	19.4	14.1
	4	16.3	11.5	15.4	-	9.4
	5	17.8	13.7	15.9	13.2	-

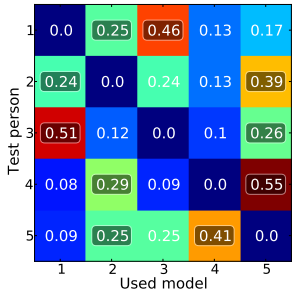

Fig. 3: Weight associated to each of the participants in the generic model. Evaluated in the frontal session, using CALR. The 2 selected models are shown within a bounding box.

Table 3: Mean angular error (°) for each of the participants using either ALR or CALR. Evaluated on the specified session (frontal or free) and model (own or generic).

		Test person					
Conditions	Method	1	2	3	4	5	Avg
Frontal Own	ALR	9.0	6.1	7.4	10.5	8.0	8.2
	CALR	8.5	5.9	6.8	9.0	7.6	7.6
Frontal Generic	ALR	16.8	9.8	12.5	9.7	10.0	11.8
	CALR	13.9	9.4	11.9	9.6	10.2	11.0
Free Generic	ALR	17.2	13.3	17.2	13.0	11.7	14.5
	CALR	14.1	12.5	16.6	15.5	10.6	13.9

4. CONCLUSION

We have presented a method to estimate the human gaze direction in a 3D space from remote, low resolution, RGB-D cameras. Furthermore we address the problem of gaze estimation for an unseen person. We stress that these conditions are highly challenging and less restrictive. Therefore we aim to enable many applications which otherwise can't be solved.

By using RGB-D cameras we alleviate eye appearance variation due to head pose. Then given the low-resolution eye image we reconstruct it from a sparse set of samples within a gaze appearance model. The reconstruction weights are used to combine the gaze vectors, associated to these training samples, to interpolate the test gaze direction.

We have proposed a way to incorporate the notion of coupled eye movements (left and right) in the form of constrains to the sparse reconstruction problem. We have shown that it reduces the gaze estimation error in most cases.

We proposed different alternatives to estimate the gaze of an unseen person. We showed that an aggregation of person specific models can perform properly, as the test user appearance can be interpolated from the users in the training set. We have shown that naturally the sparse reconstruction tends to select models which explain better the samples of the current test person. This allows us to propose an automatic model selection mechanism such that performance is comparable to the full model at lower computational cost.

5. REFERENCES

[1] Dan Witzner Hansen and Qiang Ji, "In the eye of the beholder: a survey of models for eyes and gaze.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 3, pp. 478–500, Mar. 2010.

[2] Basilio Noris, J.B. Keller, and Aude Billard, "A wearable gaze tracking system for children in unconstrained environments," *Computer Vision and Image Understanding*, pp. 1–27, 2010.

[3] Elias Daniel Guestrin and Moshe Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections.," *IEEE transactions on biomedical engineering*, vol. 53, no. 6, pp. 1124–33, June 2006.

[4] Shumeet Baluja and Dean Pomerleau, "Non-Intrusive Gaze Tracking Using Artificial Neural Networks," Tech. Rep., Pittsburgh, PA, USA, 1994.

[5] Kenneth Funes and Jean-Marc Odobez, "Gaze Estimation From Multimodal Kinect Data," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, June 2012, pp. 25–30.

[6] Lu Feng, Yusuke Sugano, Okabe Takahiro, and Yoichi Sato, "Inferring Human Gaze from Appearance via Adaptive Linear Regression," in *ICCV: International Conference on Computer Vision*, Barcelona, Spain, 2011.

[7] Lu Feng, Okabe Takahiro, Yusuke Sugano, and Yoichi Sato, "A Head Pose-free Approach for Appearance-based Gaze Estimation," in *Proceedings of the British Machine Vision Conference*, 2011.

[8] Yusuke Sugano, Yasuyuki Matsushita, Yoichi Sato, and Hideki Koike, "An incremental learning method for unconstrained gaze estimation," in *ECCV*. 2008, pp. 656–667, Springer.

[9] Oliver Williams, Andrew Blake, and Roberto Cipolla, "Sparse and semi-supervised visual mapping with the S3GP," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, 2006, pp. 230–237.

[10] Brian Amberg, Reinhard Knothe, and Thomas Vetter, "Expression invariant 3D face recognition with a Morphable Model," in *2008 8th IEEE International Conference on Automatic Face and Gesture Recognition*. Sept. 2008, pp. 1–6, IEEE.

[11] P Paysan, R Knothe, B Amberg, S Romdhani, and T Vetter, "A 3D Face Model for Pose and Illumination Invariant Face Recognition," in *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments*, Genova, Italy, 2009, IEEE.