Geometric Generative Gaze Estimation (G^{3}E) for Remote RGB-D Cameras

Kenneth Alberto Funes Mora Jean-Marc Odobez
Idiap Research Institute, CH-1920, Martigny, Switzerland
École Polytechnique Fédéral de Lausanne, CH-1015, Lausanne, Switzerland
{kfunes, odobez}@idiap.ch

Abstract

We propose a head pose invariant gaze estimation model for distant RGB-D cameras. It relies on a geometric understanding of the 3D gaze action and generation of eye images. By introducing a semantic segmentation of the eye region within a generative process, the model (i) avoids the critical feature tracking of geometrical approaches requiring high resolution images; (ii) decouples the person dependent geometry from the ambient conditions, allowing adaptation to different conditions without retraining. Priors in the generative framework are adequate for training from few samples. In addition, the model is capable of gaze extrapolation allowing for less restrictive training schemes. Comparisons with state of the art methods validate these properties which make our method highly valuable for addressing many diverse tasks in sociology, HRI and HCI.

1. Introduction

As a display of attention and interest, gaze is a fundamental cue in understanding people activities, behaviors, and state of mind, and plays an important role in many applications and research fields. In psychology and sociology, gaze information helps to infer inner states of people or their intention, and to better understand the interaction between individuals. In particular, gaze plays a major role in the communication process, like for showing attention to the speaker or indicating who is addressed, which makes gaze highly relevant for Human Robotics Interaction (HRI).

In another direction, in Human Computer Interfaces (HCI) gaze information coordinated with other user inputs can lead to the development of intuitive systems beneficial for instance for people with limited body mobility.

For these reasons, computer vision based gaze estimation has been studied for over 3 decades [6]. Many solutions have been proposed. Some achieve very high accuracy but require expensive and specialized hardware, like infrared setups or wearable sensors. A solution based on consumer hardware is needed.

To minimize intrusion and accommodate user's movement, remote cameras with wide field of view are preferred but lead to the challenge of low resolution imaging.

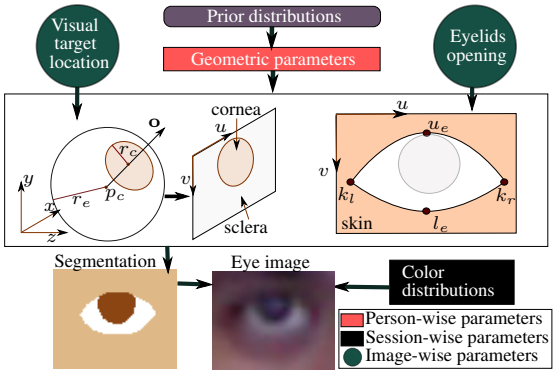


Figure 1: Method overview. The probabilistic generative process links the gazing at a visual target and eyelids movements with a semantic segmentation of the eye region and the resulting eye image observation. This process also depends on (and decouples) user specific parameters describing the eye and eyelid geometry, and ambient/session specific parameters (color distributions).

Appearance based methods, which learn a direct mapping between the eye image to the gaze parameters, avoid local features tracking which is problematic under low resolution conditions. However, they either require large sets of training data to handle eye image variations due to person's specific appearance, head pose, scale, illumination and eyelids movements when learning a general mapping, or require (less) per session training data resulting in overfitting to the person and conditions used during the training phase.

In this paper we propose a head-pose invariant gaze estimation method. It relies on an appearance generative process that model head-pose rectified eye images recovered thanks to the use of consumer RGB-D cameras.

The process is illustrated and briefly explained in Fig. 1, and has several advantages. Thanks to the use of an explicit geometric gaze model, it handles head pose and gaze direction in a unified framework, making it appropriate to reason in the 3D space and extrapolating to gaze directions not seen in the training data, which is useful for instance in inferring attention towards objects or people in HRI rather than only interpolating screen positions. The use of semantic regions (eyelids, cornea, sclera) allows to decouple the

gazing process and user geometry from the ambient conditions (color appearance), while avoiding the critical feature (iris/pupil) tracking problem of standard geometric methods. Overall, the method is able to span a large variety of people and conditions, allowing to easily adapt the model (to the user, viewing conditions) from a few training samples and use it to estimate gaze on unseen data.

Paper structure: Section 2 discuss related works. The RGB-D approach for head-free gaze estimation is described in Section 3. Our gaze model is detailed in Section 4, followed by the inference scheme in Section 5. Section 6 presents experiments and Section 7 concludes this paper.

2. Related Work

The recent survey by Hansen [6] provides a comprehensive overview on computer vision methods for gaze estimation. The most accurate techniques rely on eye geometry and on pupil center-corneal reflections detection under IR illumination [4]. They can accommodate head pose variation when using multiple light sources, but need specialized and usually costly IR hardware.

Natural light based methods have also been studied. Many proposals extract geometric models of the eyes, which can be an ellipse fitted to the pupil/iris [16], or complex shapes incorporating the eyelids [18]. Moriyama et al. [11] proposed a generative approach to segment an eye image into many detailed semantic regions (more than 10), but the actual gaze estimation was not investigated. Ishikawa et al. [7] built a geometric model of the eyeball and optical axis from image data, but relied on facial feature tracking and iris ellipse fitting. Closer to our work, Yamazoe et al. [17] proposed to fit such a model from ad-hoc segmentations of the eye images obtained through simple thresholding. However, at test time, they used the iris center derived from a fitted ellipse to infer gaze. Thus, all these methods require high contrast or high resolution images. In addition, further techniques are needed to infer the actual line of sight (LoS), specially for reasoning in 3D scenarios.

To avoid features tracking, there has been an increased interest on appearance based methods [1, 3, 8, 10, 14] that learn a direct mapping from the eye image to gaze parameters. Baluja and Pomerlau [1] trained a neural network but required thousands of training samples, while Williams et al. relied on semi-supervised Gaussian Process Regression (GPR) [15]. Sugano et al. [14] proposed taking user-computer interaction as training data. More recently, Lu et al. [8] proposed adaptive linear regression (ALR) which is based on sparse image reconstruction. They report high accuracy, even using low-resolution test images, but these images were artificially created from the same experimental session. Also, the method required a fixed head pose.

To remove this last constraint, Lu et al. subsequently proposed a GPR-based pose correcting scheme on top of their fixed head pose model [10]. In another direction, Lu et al. [9] proposed to synthesize eye images as seen from different viewpoints given eye images from a single head pose and a few images from different head poses. More conveniently, Funes and Odobez [3] leveraged on RGB-D cameras to directly handle eye appearance variation by generating frontal looking eye images used as input to ALR.

Altogether, however, appearance based methods suffer from generalization problems. Either they require large amounts of training data [1, 14, 15] to handle variations due to eye shape, pose, illumination conditions, or they are trained from session dependent samples [3, 8, 10] to be used for interpolation. In both cases, the absence of an explicit geometric model make them rather inappropriate for adaptation to users or ambient conditions, or extrapolation, which is problematic when training from a few points on a screen and estimating gaze for different head poses.

Our generative approach has several advantages with respect to the aforementioned methods. Thanks to the use of a color-based semantic segmentation approach, it is suitable for low resolution imaging as compared to traditional geometric based methods, and decouples ambient conditions from the user specific geometry.

The model's geometric prior makes it appropriate for training from a few samples and extrapolating to other conditions. This is valuable for reasoning in a 3D environment, a desired property in psychology, sociology and HRI while still appropriate for HCI applications.

3. Gaze estimation from RGB-D cameras

This section summarizes the gaze estimation method from RGB-D cameras. To acquire head-pose invariance we followed a similar procedure to [3]. In an offline step a 3D face mesh (template) is built for the user by fitting a 3D Morphable model (3DMM) [13] to depth data. Then, in an online stage, the following steps are executed:

- The 3D head pose $\mathbf{p}_t$ is obtained by fitting, frame-by-frame, the personalized 3D mesh template to depth data using iterative closest points (ICP), resulting for frame t in the 3D head rotation and translation $\mathbf{p}_t = \{\mathbf{R}_t, \mathbf{t}_t\}$.
- Assuming a calibrated RGB-D setup, the RGB-D frame is transformed to a textured 3D mesh. We then re-render the texture, lying on the 3D data surface, using the inverse head pose parameters $\mathbf{p}_t^{-1} = \{\mathbf{R}_t^\top, -\mathbf{R}_t^\top \mathbf{t}_t\}$. This results in facial images as if the head was static and in front of the camera. The 3DMM defines a priori the eyes location referred to the head coordinate system. This position is used to crop eye images from the frontal looking facial texture, resulting in pose-rectified eye images.
- The gaze direction is estimated from the pose-rectified eye images using our proposed method (cf. Section 4).
- The gaze direction is transformed back to the world coordinate system, according to the head pose.

4. Geometric generative gaze model

Table 1: List of symbols related to our model

Symbol	Description
$I; (u, v)$	Image *index* and pixel *coordinates*
p_c	Eyeball rotation center
$\kappa = (\phi_\kappa, \theta_\kappa)$	Visual axis deviation
d	Nodal point distance from p_c
$\mathbf{a} := (\kappa, d)$	*Axial* parameters
r_e, r_c	Eyeball and cornea radii
$k_l = (k_{lu}, k_{lv})$	Left eye corner in image coordinates
$k_r = (k_{ru}, k_{rv})$	Right eye corner in image coordinates
$k_{lr} = (k_l, k_r)$	Left and right eye corners
$\mathbf{s} := (r_e, r_c, k_l, k_r)$	*Structure* parameters
p	Visual target 3D position
$\mathbf{o} = (\phi, \theta)$	Optical axis orientation
u_e, l_e	Upper and lower eyelid opening
$\mathbf{m} := (\mathbf{o}, u_e, l_e)$	*Movement* parameters
Λ_l	Class l color distribution parameters
c	Observed color at pixel u, v
$\lambda \in \{0, 1\}$	Occlusion state for pixel u, v

4.1. Approach overview

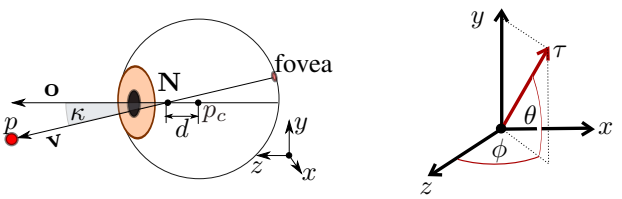
The proposed approach is summarized as a block diagram in Fig. 1. Before describing the method, notice that all measures can be referred to a coordinate system fixed to the head (with the z axis directed towards the head front) due to the procedure described in Section 3. This makes it possible to deal with head fixed quantities. In addition, there is no scale ambiguity in the pose-rectified eye images as depth information provides the pixel size in meters.

The model is characterized by user specific parameters $\mathcal{U} = \{p_c, r_e, r_c, \kappa, d, k_{lr}\}$, which define the fixed eye geometry (all notations are defined in Table 1), and image specific parameters $\mathbf{m} = \{\mathbf{o}, u_e, l_e\}$ related to the actual gaze activity: what is the person's eye orientation (characterized by the optical axis $\mathbf{o}$) and how are the eyelids open (u_e, l_e).

As shown in Fig. 1, given these parameters, an eye and eyelid configuration can be specified, from which a semantic segmentation of the eye region can be generated. The generative process then further combines this segmentation with session dependent color model distributions, parametrized by $\{\Lambda_l\}_{l=1..3}$, to produce eye color-images.

Our probabilistic model is thus able to compute the likelihood of such eye images, which constitute our observations. Hence, during a training phase, user parameters can be learned by maximizing the likelihood of gaze annotated training samples, while at test time, the image optimization leads to the actual estimation of $\mathbf{m}$, and thus the LoS.

In the following, we describe more precisely the different elements of our model: the eye geometric model, the parametric segmentation function, the definition of the likelihood, and our generative model.



(a) (b)

Figure 2: a) Eye geometry with optical ($\mathbf{o}$) and visual ($\mathbf{v}$) axis definition. b) spherical parametrization of an axis "τ".

4.2. Eye geometric model

Fig. 2a illustrates the geometric eye model we use [6]. The process of gazing a visual target $p \in \mathbb{R}^3$ consists of rotating the eyeball around the point $p_c \in \mathbb{R}^3$ such that the *visual axis* ($\mathbf{v}$) intersects p. The visual axis is the line connecting the fovea (the point of highest visual acuity in the retina) and the nodal point $\mathbf{N}$. It differs from the *optical axis* ($\mathbf{o}$), which is the line connecting the center of rotation p_c and the pupil center. We parametrize these axis by two angles (as in Fig. 2b). As the eye is a rigid body, the angular difference between these axis is fixed and can be represented by the person dependent angles $\kappa = (\phi_\kappa, \theta_\kappa)$ [1]:

$$\mathbf{v} = \mathbf{o} + \kappa \tag{1}$$

Thus, implicitly, if the "axial" parameters $\mathbf{a} := (\kappa, d)$ are known, then the eye rotation ($\mathbf{o}$) can be defined as a function of the position of p. We denote this process as:

$$\mathbf{o}(p) = (f_\phi(p; \kappa, d, p_c), f_\theta(p; \kappa, d, p_c)) \tag{2}$$

4.3. Parametric segmentation function

An eye image is segmented into three regions: the cornea[2], sclera and skin. The central rectangle in Fig. 1 shows our parametric segmentation: assuming that the user eye geometric parameters $\mathcal{U}$ are known, then a given eye orientation $\mathbf{o}$ define a cornea-sclera segmentation, obtained as the orthogonal projection of the 3D cornea contour into the xy plane, followed by a transformation to image coordinates uv. This is possible due to the eye image rectification procedure described in Sec. 3 which provides a mapping of the 3D data into the rectified eye image coordinates.

To define the segmentation of the skin region, we rely on a set of parameters characterizing the eyelids structure (eye corners k_l and k_r) and another set controlling the eyelids opening. We take a simple approach, shown in the right part of Fig. 1 where the upper and lower eyelids are quadratic bezier curves sharing the eyelids corners k_l and k_r.

The vertical position of the inner control points are denoted as u_e and l_e. They define the eyelids opening, and thus, the skin segmentation. The skin class overrides the sclera and cornea regions in the overall segmentation.

[1] This representation ignores eye torsion. Even though it is known that the eyes rotate according to Listing's and Donder's laws, this simplification was shown to have little impact on gaze estimation [5].

[2] We define here "cornea" as the region composed of the pupil and iris.

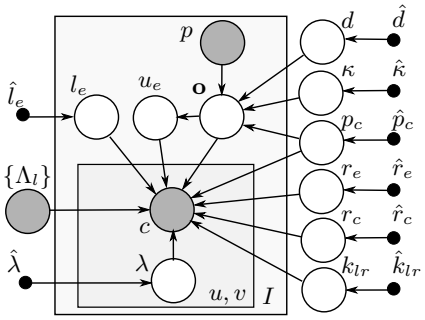


Figure 3: Graphical representation of the geometric generative gaze model. The symbols are described in Table 1.

Given this procedure, we define the *segmentation function* given in Eq. 3. Notice this is also a function of the parameters which define the structure of the eyes and the current movement of the eye and eyelids.

$$S_{u,v}(l; \mathbf{m}, \mathbf{s}, p_c) = \begin{cases} 0 & \text{if pixel } (u,v) \notin \text{class } l \\ 1 & \text{if pixel } (u,v) \in \text{class } l \end{cases} \quad (3)$$

4.4. Image likelihood and outlier modeling

So far we used 3 classes to define the eye image segmentation regions. Here we introduce a fourth class for pixel outliers, denoted by the variable $\lambda = \{0, 1\}$ where 1 indicates that the pixel is an outlier. This is intended to address missing data, occlusions and specular reflections.

Our observation data is an eye image I. Its likelihood given the parameters is defined as $p(I|.) = \prod_{u,v} p_{u,v}(c|.)$ which assume that all pixels are independent observations given the parameters. To model the likelihood of individual pixels, we define the color distribution associated to a class l as $p(c|\Lambda_l)$, a 2 component GMM in the RGB space.

For outliers we assume an equal probability of observing any color, such that $p(c|\lambda = 1) = \epsilon$. The likelihood of a color pixel is then simply defined as the likelihood given its class (either an outlier, or one of the 3 eye region classes), which can be written in condensed form as:

$$p_{u,v}(c|\lambda, \mathbf{m}, \mathbf{s}, p_c, \{\Lambda_l\}_l) = \epsilon^{\lambda} \left[\prod_l p(c|\Lambda_l)^{S_{u,v}(l;\mathbf{m},\mathbf{s},p_c)} \right]^{1-\lambda}$$

4.5. Generative model

The graphical model of our geometric generative gaze estimation (G^{3}E) approach is shown in Fig. 3. It is a stochastic extension of the full process of gazing up to the generation of eye images, under which every geometric parameter is defined as a random variable.

Let us denote by $x \sim \mathcal{N}(\mu_x, \sigma_x)$ a random variable x being drawn from a Gaussian distribution with mean μ_x and standard deviation σ_x, and the "hat" $\hat{(.)}$ notation to represent the hyperparameters of a prior distribution, e.g. $\hat{d} := (\hat{\mu}_d, \hat{\sigma}_d)$. The generative process shown in Fig. 3 can be described as follows:

- Draw the eyeball rotation center p_c:
  - $p_c \sim \big(\mathcal{N}(\hat{\mu}_{p_{cx}}, \hat{\sigma}_{p_{cx}}), \mathcal{N}(\hat{\mu}_{p_{cy}}, \hat{\sigma}_{p_{cy}}), \mathcal{N}(\hat{\mu}_{p_{cz}}, \hat{\sigma}_{p_{cz}})\big)$
- Draw axial parameters $\mathbf{a} := (\kappa, d)$:
  - $\kappa \sim \big(\mathcal{N}(\hat{\mu}_{\phi_\kappa}, \hat{\sigma}_{\phi_\kappa}), \mathcal{N}(\hat{\mu}_{\theta_\kappa}, \hat{\sigma}_{\theta_\kappa})\big)$
  - $d \sim \mathcal{N}(\hat{\mu}_d, \hat{\sigma}_d)$
- Draw "structure" parameters $\mathbf{s} := (r_e, r_c, k_l, k_r)$:
  - $r_e \sim \mathcal{N}(\hat{\mu}_{r_e}, \hat{\sigma}_{r_e})$
  - $r_c \sim \mathcal{N}(\hat{\mu}_{r_c}, \hat{\sigma}_{r_c})$
  - $k_l \sim \big(\mathcal{N}(\hat{\mu}_{k_{lu}}, \hat{\sigma}_{k_{lu}}), \mathcal{N}(\hat{\mu}_{k_{lv}}, \hat{\sigma}_{k_{lv}})\big)$
  - $k_r \sim \big(\mathcal{N}(\hat{\mu}_{k_{ru}}, \hat{\sigma}_{k_{ru}}), \mathcal{N}(\hat{\mu}_{k_{rv}}, \hat{\sigma}_{k_{rv}})\big)$
- For each image $I = 1, \ldots, N$:
  - Draw the visual target $p \sim uniform$
  - Draw movement parameters $\mathbf{m} := (\mathbf{o}, u_e, l_e)$:
    - $\mathbf{o} \sim \big(\mathcal{N}(f_\phi(p; \mathbf{a}, p_c), \hat{\sigma}_\mathbf{o}), \mathcal{N}(f_\theta(p; \mathbf{a}, p_c), \hat{\sigma}_\mathbf{o})\big)$
    - $u_e \sim \mathcal{N}(a_u\theta + b_u, \hat{\sigma}_{u_e})$
    - $l_e \sim \mathcal{N}(\hat{\mu}_{l_e}, \hat{\sigma}_{l_e})$
  - For each $(u, v) = [1, \ldots, width], [1, \ldots, height]$:
    - Draw outlier or not indicator $\lambda \sim Bernoulli(\hat{\lambda})$
    - Draw pixel color $c \sim p_{u,v}(c|\lambda, \mathbf{m}, \mathbf{s}, p_c, \{\Lambda_l\}_l)$

It is important to make a few remarks about this model:

- *Upper eyelid opening*. The upper eyelid is correlated with the elevation angle of the eye by means of a linear Gaussian model. This encodes the effect of the upper eyelid following the vertical orientation of the eye.
- *Eye rotation*. A stochastic extension of Eq. 2 was defined to allow uncertainty in the target position or eye fixation.
- *Stochastic segmentation*. Under this model the segmentation becomes a stochastic process. Drawing a sample from the geometric parameters or the movement parameters $\mathbf{m}$ is equivalent to "drawing a segmentation".
- *Prior distributions and hyperparameters*. Prior distributions have a semantic and/or anatomical interpretation. Therefore the hyperparameters are fixed to values that can be found in the literature (e.g. $r_e \approx 12mm$) or are a consequence of the pose-rectification processing described in Section 3 (e.g. it is known where the eye corners are expected to be from the eye image cropping).
- *Color distributions*. In this paper, the color model parameters $\{\Lambda_l\}$ are defined as observed. In practice, we acquire color samples from a single image to estimate them. Automatic color model learning is left for future work. Notice that decoupled color modeling is an important advantage of G^{3}E. It allows for adaptation to different illumination and contrast conditions, without re-estimating the geometric parameters.

5. Model inference

There are two inference goals for our model. i) *Training phase*: from a set of pairs of image samples and visual target locations we aim to infer the person dependent geometry. ii) *Test phase*: given an input image we infer the eye rotation $\mathbf{o}$ and eyelids opening leveraging on the previous training.

The inferred $\mathbf{o}$ is used to estimate the direction of the visual axis (cf. Eq. 1). We resorted to Variational Bayes (VB) as an approximate inference method to address the complexity of our model. We summarize the main points below and details are provided in the supplementary material.

Variational Bayes. Let $\mathbf{X}$ denote the observed data and $\mathbf{Z}$ to be the latent variables to infer. In VB the posterior $p(\mathbf{Z}|\mathbf{X})$, which might not be possible to estimate analytically, is approximated by some proposal distribution $q(\mathbf{Z})$.

This leads to the definition of the *variational lower bound* $\mathcal{L}(q)$, a functional whose maximization with respect to q is equivalent to a minimization of the Kullback-Leibler divergence between $q(\mathbf{Z})$ and $p(\mathbf{Z}|\mathbf{X})$ [2]. The optimal $q^*(\mathbf{Z})$ is then used as a substitute of the posterior.

Proposal distribution. We define $q(\mathbf{Z})$ with the following parametric form:

$$
\begin{aligned}
q(\mathbf{Z}) = &\mathcal{N}(\mu_d, \sigma_d)\mathcal{N}(\mu_{\phi_\kappa}, \sigma_{\phi_\kappa})\mathcal{N}(\mu_{\theta_\kappa}, \sigma_{\theta_\kappa})\mathcal{N}(\mu_{r_e}, \sigma_{r_e}) \\
&\mathcal{N}(\mu_{r_c}, \sigma_{r_c})\mathcal{N}(\mu_{p_{cx}}, \sigma_{p_{cx}})\mathcal{N}(\mu_{p_{cy}}, \sigma_{p_{cy}})\mathcal{N}(\mu_{p_{cz}}, \sigma_{p_{cz}}) \\
&\mathcal{N}(\mu_{k_{lu}}, \sigma_{k_{lu}})\mathcal{N}(\mu_{k_{lv}}, \sigma_{k_{lv}})\mathcal{N}(\mu_{k_{ru}}, \sigma_{k_{ru}})\mathcal{N}(\mu_{k_{rv}}, \sigma_{k_{rv}}) \\
&\prod_I \left[\mathcal{N}(\mu_\phi, \sigma_\phi)\mathcal{N}(\mu_\theta, \sigma_\theta)\mathcal{N}(\mu_{u_e}, \sigma_{u_e})\mathcal{N}(\mu_{l_e}, \sigma_{l_e}) \prod_{u,v} q(\lambda)\right],
\end{aligned}
$$

where we omit the image and pixel indices to avoid clutter.

Every continuous random variable has been defined as a univariate Gaussian. The motivation for this $q(\mathbf{Z})$ is that it is possible to compute the derivatives of $\mathcal{L}(q)$ with respect to the Gaussian parameters. Following [12] we compute the derivatives using Monte Carlo expectations[3] to address the complex relations in our model (cf. Eq. 2 and Eq. 3).

A factorized $q(\mathbf{Z})$ also allows to optimize $\mathcal{L}(q)$ in an iterative fashion, where one factor is optimized at the time, leading to an increase of $\mathcal{L}(q)$ until global convergence.

The only non continuous variable is λ. It can be shown that the optimal $q(\lambda)$ is a Bernoulli distribution with $P(\lambda = 1) = \omega$, where ω is given by

$$
\omega = \frac{\hat{\lambda}}{(1-\hat{\lambda})\frac{1}{\epsilon}\prod_l p(c|\Lambda_l)^{\mathbb{E}_{\mathbf{m},\mathbf{s},p_c}[S_{u,v}(l;\mathbf{m},\mathbf{s},p_c)]} + \hat{\lambda}} \tag{4}
$$

Notice that $\mathbb{E}_{\mathbf{m},\mathbf{s},p_c}\left[S_{u,v}(l;\mathbf{m},\mathbf{s},p_c)\right]$ can be interpreted as the *expected segmentation* of an image. According to Eq. 4 an outlier is considered as a color observation which is either unlikely for any class, or that is likely for a given class but is spatially incoherent w.r.t. the geometric model.

Efficient group factor optimization. We can optimize $\mathcal{L}$ efficiently by defining Jacobians over groups of variables (e.g. $\mathbf{J_a} = [\frac{\partial\mathcal{L}}{\partial\mu_{\phi_\kappa}}, \frac{\partial\mathcal{L}}{\partial\sigma_{\phi_\kappa}}, \frac{\partial\mathcal{L}}{\partial\mu_{\theta_\kappa}}, \frac{\partial\mathcal{L}}{\partial\sigma_{\theta_\kappa}}, \frac{\partial\mathcal{L}}{\partial\mu_d}, \frac{\partial\mathcal{L}}{\partial\sigma_d}]^\top$). This is efficient in terms of derivatives computation, as we found that their Monte Carlo expectations require group sampling rather than univariate sampling, due to complex dependencies in Eq. 2 and Eq. 3.

[3] All expectations are defined with respect to $q(\mathbf{Z})$

Gradient ascent is then used to find the optimal Gaussian parameters of the corresponding factor of $q(\mathbf{Z})$ (e.g. $q(\mathbf{a})$).

Inference algorithms.
Training. Our overall inference method is given in Algorithm 1. This method finds the person-specific geometry from a set of eye images and their corresponding p.
Test phase (Gaze inference). At test time, the geometry is fixed and we only optimize w.r.t. the test image's $q(\mathbf{m})$ and outliers in an iterative fashion. In this case the visual target location p is unknown; as we assume a uniform prior over p, its influence on $p(\mathbf{o}|.)$ becomes uninformative. The inferred mode of $q^*(\mathbf{Z})$ can then be used to derive the MAP visual axis, leading to the 3D line of sight for the given image.

Algorithm 1 Geometric generative gaze model inference.

Set initial $q(\mathbf{Z})$ from the prior distribution parameters.
repeat
- Optimize $\mathcal{L}$ w.r.t. eye corners and all eyelids opening: $q(k_{lu})q(k_{lv})q(k_{ru})q(k_{rv})\prod_I q(u_e^I)q(l_e^I)$
- Optimize $\mathcal{L}$ w.r.t. eyeball geometry and orientation: $q(r_e)q(r_i)q(p_{cx})q(p_{cy})\prod_I q(\mathbf{o}^I)$
- Optimize $\mathcal{L}$ w.r.t. axial parameters and eyeball depth: $q(\mathbf{a})q(p_{cz})$
- Update outliers $q(\lambda_{u,v}^I)$ for all pixels using Eq. 4

until Convergence
Return $q^*(\mathbf{Z})$

6. Experiments

To validate our model, we first studied its behavior using synthetic data. We then compared it against representative geometric and appearance approaches on real data to validate its advantages and added properties.

6.1. Experiments on synthetic data

To validate our method we created synthetic data using the generative process described in Section 4.5. Examples are shown in Fig. 4; their resolution in pixel is 55×40.

Synthetic data allows us to study the inference scheme and the observability of the gaze model parameters by comparing the parameters inferred by G^{3}E to their true values.

The left plot of Fig. 5 shows the parameter estimation errors as a function of the number of training samples, where each parameter is inferred separately while the other parameters are set to their true values. We can conclude the following: i) almost all parameters can be well estimated, and

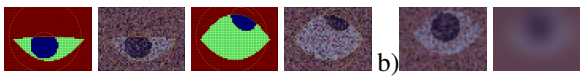


Figure 4: a) Synthetic data samples (drawn segmentation and the generated image from color sampling). b) Sample (left) smoothed by a gaussian filter of $\sigma = 3.0mm$ (right).

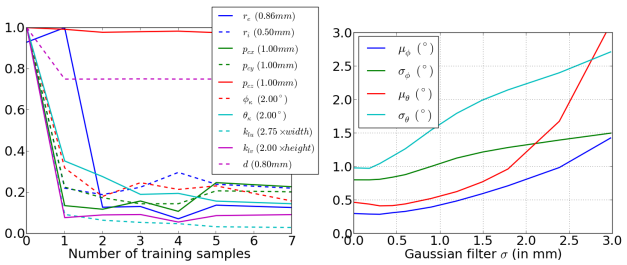


Figure 5: Left. Parameter estimation error vs. number of eye training samples. The y axis scale is given in the legend of each parameter. Right. Mean and standard deviation (derived from the infered $q(\mathbf{o})$) of the gaze estimates $\mathbf{o} := (\phi, \theta)$, vs. the standard deviation of the Gaussian blurring filter ($1mm = 1.68pixels$). For the gaze means, the deviation from their true values is plotted. For each experiments, averages over 500 runs are reported.

this requires only a few samples; ii) d and the eyeball depth p_{cz} are difficult to infer, due to their small impact on $\mathbf{o}$. Nevertheless, this means that their impact on gaze estimation is small. iii) The visual axis κ angle parameters, which are important for accurate gaze estimation but are often neglected, are well constrained by the image likelihood and the known object position p, and can thus be inferred.

The right plot of Fig. 5 shows a similar experiment: we evaluate the gaze estimation accuracy (given true geometric parameters) as a function of image resolution simulated through blurring (see Fig. 4). Notice the high robustness w.r.t. resolution due to the optimization of a global image likelihood measure. We also show the estimated variances, which correctly reflect the uncertainty of the gaze estimates.

These results suggest that given proper training data, our approach can potentially have highly accurate gaze estimation ($< 2^\circ$ error) under poor sensing conditions.

6.2. Real data collection

We collected RGB-D data using a Kinect sensor. To collect ground-truth, we asked participants to gaze at a visual target while either keeping their head fixed, or in second phase, asking them to rotate the head (observing $\pm 30^\circ$ pose ranges for head yaw and elevation) while gazing. Each phase lasted a few minutes. As target, we used either a point displayed on a $24''$ flat screen or a moving floating target located between the participant and camera. The screen was calibrated with the camera so that 2D screen coordinates are interpreted as a 3D point. The floating target was automatically tracked to find its 3D position.

The distance of the participant to the screen and sensor was $\approx 85cm$. In the experiments using the floating target, people distance to the sensor ranged between 1m and 1.5m, resulting in eye image sizes between 20×14 and 13×9. In our experiment, the pose-corrected image were upsampled to a fix resolution of 55×40 pixels with known pixel size ($0.595mm/pixel$). As performance measure, we used the angular gaze error, defined as the angle between the estimated line of sight (LoS) and the vector pointing from the LoS's origin to the (known) visual target's 3D position (p).

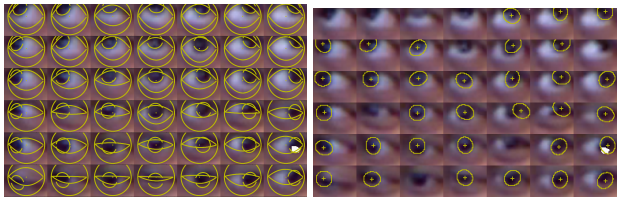


Figure 6: Left: Geometric fitting given by G3E. Right: Ellipse fitting given by the Starburst algorithm on training data collected using the floating target.

6.3. G^{3}E inference and geometric methods

We illustrate the inference process output for the training samples shown in Fig. 6. The result of the training can be visualized using the mode of $q^*(\mathbf{Z})$ (MAP estimate) and by overlaying the contours of the associated segmentation and eyeball structure, as shown in Fig. 6 (left).

Our method follows properly the position of the eyelids and eye orientation despite the low resolution and the sometimes unclear boundaries between eye regions.

As a qualitative comparison, we tested the Starburst algorithm [16] on the same data. This approach is representative of the geometric paradigm, which relies on fitting an ellipse to the cornea from the voting of thresholded gradients, estimated along rays from an initial estimation of the cornea region center. As initialization we set the true center value. Despite this, and parameters tuning, we obtained the results shown in Fig. 6. The low recall and unaccurate estimation demonstrate the important difficulties of ellipse fitting, which is a critical step for many geometric gaze estimation methods, e.g. [7]. Notice that our approach does not have this limitation as it avoids local feature computations.

6.4. Appearance based methods

We compared our approach to Funes and Odobez's method [3]. To our knowledge, this is the only method using RGB-D data, for which the eye images pose-rectification was proposed. This method in turn uses Adaptive Linear Regression (ALR) for gaze estimation from the low-resolution pose-rectified images. ALR is a state-of-the-art appearance based method proposed by Lu et al. [8].

By comparing to [3] we indirectly compare to [8] within the pose-rectified eye images context. In the following, when referring to ALR, we implicitly refer to [3]. Our intention is to contrast to the appearance based paradigm. We now describe the result of experiments designed to raise awareness of the limitations of appearance based methods.

Number of training samples. As concluded in Section 6.1 our model is adequate for training from few data. We validated this on real data, and contrasted to ALR, as shown in

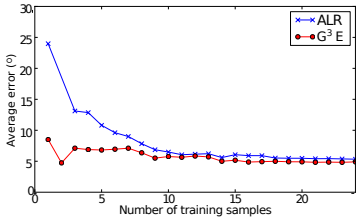


Figure 7: Average gaze error as a function of the number of training samples. Computed on test data from a participant gazing at a floating target with a fixed frontal pose.

Fig. 7, which presents a typical error curve obtained for a participant in functions of the number of training samples.

As shown in Fig. 7, ALR need to cover densely the gaze space with the training samples in order to achieve lower errors. This is a limitation of appearance based methods.

Gaze extrapolation. As our method is based on a explicit eye model, we argue it can extrapolate to gaze directions outside the training set. To illustrate this, we conducted an experiment where we used the same 49 samples restricted to gaze yaw and elevation angles within the range $[-15°, 15°]$ to train the ALR and G^{3}E models.

Fig. 8 shows the gaze tracking results on a test sequence, where our claim is validated. ALR, as any interpolation based method, is not able to estimate gaze outside the range of directions used for training, thus causing the saturations observed in Fig. 8. This is not a limitation of our method.

Gaze estimation across different sessions. In this experiment we collected data with the floating target in two different sessions (A and B) performed 6 months apart, each for two participants. Across sessions there is a drastic change in the illumination and distance to the camera (see Fig. 9).

We then learned an ALR and a G^{3}E model using the data from session A. For the G^{3}E approach, applying directly the learned model -including the color distributions from session A- on session B results in large errors ($40.2°$ and $38.2°$ for the respective participants). This is due to the obvious color mismatch between the two conditions. However, we can easily leverage on the important property of our model which is the decoupling of the ambient conditions from the person-specific geometry. By learning the color distributions of session B using color samples picked from a single eye image in session B, we quickly obtain an adapted model that result in very good performance, as shown in Table 2.

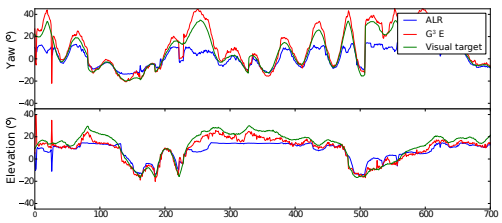


Figure 8: Estimated eye rotation (°) on a test sequence, with training samples restricted to the $[-15°, 15°]$ range.

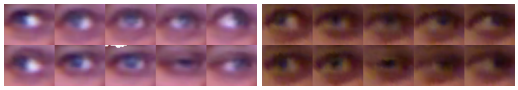


Figure 9: Eye image samples across different sessions. Participant 1 in session A (left) and session B (right).

Table 2: Mean angular error (°) when training the model on session A and estimating gaze on session B. See text.

Method	Participant 1	Participant 2
ALR	21.7	25.0
G^{3}E	8.0	5.5

On the other hand, given the lack of geometrical model, ALR does not offer much flexibility for adaptation. Even if it relies on normalized features that should be robust to global illumination variation within the eye image [8], results in Table 2 shows that the Session A model is not appropriate for Session B, demonstrating that session changes go beyond simple illumination and contrast corrections.

The automatic learning of color distributions for G^{3}E is left for our future work, but this experiment already validates its potential for cross-session adaptation.

6.5. Screen gazing evaluation

We evaluated the performance of our method for the screen target estimation task, where we considered both the fixed and moving head pose case for the five participants.

Notice that, due to the proximity to the depth sensor, there is regularly missing depth data, which affects the pose-rectified eye image, as it is visible in Fig. 10.

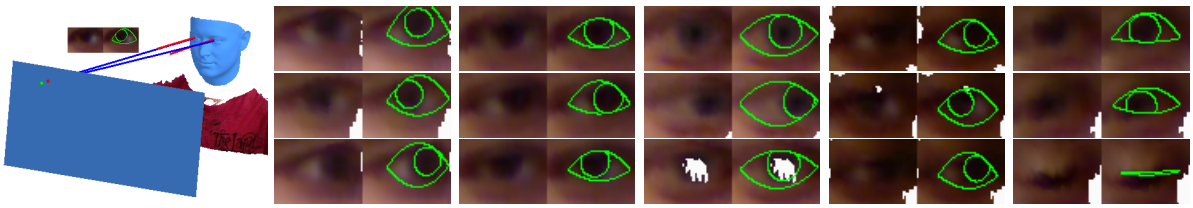
In our method, we address this problem by forcing the pixels to be outliers (i.e. setting $\omega \sim 1$ for missing pixels). However, as ALR does not provide a straightforward way to handle missing data, we do not report ALR results here.

Results are summarized in Table 3. Given the quality of the input data and that head pose variation was within a range of $\pm 30°$ for yaw and elevation, the performance are highly promising. To illustrate this, we provide in Fig. 10 an example of the setup together with qualitative segmentation results for the 5 participants.

They show that our method has a good behavior at test time, although we observe on the bottom right a problematic situation for our approach which is extreme gazing down, where the cornea region gets heavily occluded by the eyelid.

Table 3: Gaze angular median error (°) for people looking at screen targets.

	Participant					
Head pose	1	2	3	4	5	Avg
Fixed	2.9	2.7	3.1	2.5	5.9	3.4
Moving	9.3	5.5	3.6	4.6	8.6	6.3



(a) (b)

Figure 10: Screen gaze estimation task. a) RGB-D frame. The user's facial 3D template is rendered with the estimated head pose. The blue lines and green dot on the screen are the ground truth. The red lines correspond to the estimated lines of sight and the red dot is the screen intersection (for the left eye). Video results are provided in the supplementary material. b) Test left eye images ($\approx$ 20 pixels eye width prior to pose rectification). Each column is for a different participant. White pixels are missing data. Contours represent the mode of $q^*(\mathbf{Z})$.

7. Conclusions

We have proposed a novel method for head pose invariant gaze estimation from RGB-D cameras. We call it geometric generative gaze estimation (G^3E). It is based on a geometric understanding of the 3D gaze action and generation of eye images, formalized as a generative process.

We developed an inference technique, based on Variational Bayes, to find the person specific geometric parameters from training data (i.e. gaze annotated eye images).

We have shown that our method has many advantages with respect to previous approaches. Due to priors on the geometric parameters it is adequate for training from a few samples. Our model is able to extrapolate outside the training data, unlike previous approaches based on appearance. It has also proven adequate for low resolution imaging. Finally, our model correctly decouples the person specific geometry from the observed pixel values, which are dependent on the ambient conditions. This is adequate for adaptation and estimating gaze in different situations. These advantages were validated using both synthetic and real data.

We believe our method has an important potential for gaze estimation in many different scenarios, making it relevant for HCI, HRI, sociology and psychology.

Acknowledgments The authors gratefully acknowledge the financial support from the Swiss National Science Foundation (Project: 200021_130152, TRACOME) www.snf.ch.

References

[1] S. Baluja and D. Pomerleau. Non-Intrusive Gaze Tracking Using Artificial Neural Networks. Technical report, CMU, 1994. 2

[2] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, Oct. 2007. 5

[3] K. A. Funes Mora and J.-M. Odobez. Gaze estimation from multimodal Kinect data. In *Computer Vision and Pattern Recognition Workshops*, pages 25–30, June 2012. 2, 6

[4] E. D. Guestrin and M. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *Trans. on bio-medical engineering*, June 2006. 2

[5] E. D. Guestrin and M. Eizenman. Listing's and Donders' laws and the estimation of the point-of-gaze. In *Symp. on Eye Tracking Research & Applications*, Austin, TX, 2010. 3

[6] D. W. Hansen and Q. Ji. In the eye of the beholder: a survey of models for eyes and gaze. *IEEE trans. on pattern analysis and machine intelligence*, 32(3):478–500, Mar. 2010. 1, 2, 3

[7] T. Ishikawa, S. Baker, I. Matthews, and T. Kanade. Passive Driver Gaze Tracking with Active Appearance Models. In *Proc. World Congress on Intelligent Transportation Systems*, pages 1–12, Oct. 2004. 2, 6

[8] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Inferring human gaze from appearance via adaptive linear regression. In *Int. Conf. on Computer Vision*, Barcelona, Nov. 2011. 2, 6, 7

[9] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Head pose-free appearance-based gaze sensing via eye image synthesis. In *Int. Conf. on Pattern Recognition*, Nov. 2012. 2

[10] F. Lu, O. Takahiro, Y. Sugano, and Y. Sato. A Head Pose-free Approach for Appearance-based Gaze Estimation. In *Proc. of the British Machine Vision Conference*, 2011. 2

[11] T. Moriyama and J. Cohn. Meticulously detailed eye model and its application to analysis of facial image. *Int. Conf. on Systems, Man and Cybernetics*, 1:629–634, 2004. 2

[12] M. Opper and C. Archambeau. The variational gaussian approximation revisited. *Neural Comput.*, Mar. 2009. 5

[13] 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 Advanced Video and Signal based Surveillance*, Genova, Italy, 2009. IEEE. 2

[14] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike. An incremental learning method for unconstrained gaze estimation. In *ECCV*, pages 656–667. Springer, 2008. 2

[15] O. Williams, A. Blake, and R. Cipolla. Sparse and semi-supervised visual mapping with the S3GP. In *Computer Vision and Pattern Recognition*, pages 230–237, 2006. 2

[16] D. Winfield and D. Parkhurst. Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. *Proc. of Computer Vision and Pattern Recognition Workshops*, 3:79–79. 2, 6

[17] H. Yamazoe, A. Utsumi, T. Yonezawa, and S. Abe. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. In *Symp. on Eye Tracking Research & Applications*, New York, 2008. 2

[18] A. L. Yuille, P. W. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8(2):99–111, Aug. 1992. 2