A Head Pose-free Approach for Appearance-based Gaze Estimation

Feng Lu
Takahiro Okabe
Yusuke Sugano
Yoichi Sato
{lufeng,takahiro,sugano,ysato}@iis.u-tokyo.ac.jp

Institute of Industrial Science
the University of Tokyo
Tokyo, Japan

Abstract

To infer human gaze from eye appearance, various methods have been proposed. However, most of them assume a fixed head pose because allowing free head motion adds 6 degrees of freedom to the problem and requires a prohibitively large number of training samples. In this paper, we aim at solving the appearance-based gaze estimation problem under free head motion without significantly increasing the cost of training. The idea is to decompose the problem into subproblems, including initial estimation under fixed head pose and subsequent compensations for estimation biases caused by head rotation and eye appearance distortion. Then each subproblem is solved by either learning-based method or geometric-based calculation. Specifically, the gaze estimation bias caused by eye appearance distortion is learnt effectively from a 5-seconds video clip. Extensive experiments were conducted to verify the effectiveness of the proposed approach.

1 Introduction

Gaze intuitively plays an essential role in representing human attention, feeling, and desire *et al*. [13]. Therefore, research into human gaze tracking has attracted much attention in recent years. Commercial systems have already been used in specific areas such as market research, driver/pilot training, and helping people with disabilities. However, these systems require expensive and cumbersome hardware, which stops them from being used in consumer applications. With the development of computer vision technology, it is hoped that gaze will be able to be estimated via much fewer devices, or even a single camera.

According to recent surveys [4, 8], there exist two main categories of computer vision-based methods, namely feature-based and appearance-based methods. Feature-based methods extract small scale features from eye images, such as corneal infrared reflections, pupil centre [14], and iris contour [16]. These features are used along with 3-D eye models to determine the gaze direction independently of head pose. Beymer and Flickner [2] proposed generating and detecting corneal reflections via stereo pan-tilt units equipped with zoom-in cameras and infrared LEDs. Also, two additional wide range stereo cameras are used for eye position tracking. Similar methods were also introduced by Brolly and Mulligan [3], Nagamatsu *et al*. [9], and Zhu and Ji [20]. Villanueva and Cabeza [15] suggested reducing the number of cameras while using more infrared LEDs for geometric calculation. Yoo and

© 2011. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

BMVC 2011 http://dx.doi.org/10.5244/C.25.126

Chung [19] proposed a novel method based on cross-ratio that avoids the explicit computation of the 3-D positions of the eye, cameras, and screen. Kang *et al*. [5] further improved this method by considering the differences between individual eye parameters.

Disadvantages of feature-based methods mainly include 1) to extract small eye features via high resolution infrared imaging, special cameras/lights are always required that are not robust enough in uncontrolled environments, and 2) the accuracy of geometric-based calculation depends heavily on system calibrations that are often too difficult for ordinary users.

On the other hand, appearance-based methods work with only a single webcam under natural light and regard the entire eye image as a high-dimensional input. Baluja and Pomerleau [1] proposed a neural network trained by 2000 labelled training samples. Xu *et al*. [18] also used a similar method. Tan *et al*. [12] proposed to utilize the local linearity of the eye appearance manifold and collected 252 training samples for interpolation. Williams *et al*. [17] introduced a semi-supervised method based on Gaussian Process regression to reduce the number of labelled training samples. Recently, Sugano *et al*. [11] proposed obtaining training samples via automatically generated saliency maps from a video clip to make the user unaware of the calibration. Lu *et al*. [6] introduced adaptive linear regression to further reduce the number of training samples for high accuracy gaze inferring.

The limitation of these methods lies in that they all assume a fixed head pose. To our knowledge, one exception was proposed by Sugano *et al*. [10]. However, its estimation accuracy is low (around 4°) even after obtaining up to 1000 training samples.

1.1 Motivation

We focus on the problem of appearance-based human gaze estimation under free head motion using a single webcam. This problem is high-dimensional because the head motion has 6 degrees of freedom. Therefore directly solving the problem requires a prohibitively large number of training samples. To effectively solve this problem while significantly reduce the training cost, we propose a novel approach with characteristics as follows:

1. A decomposition scheme is introduced to decouple the original problem into subproblems, namely initial estimation and subsequent compensations.

2. Geometric priors are introduced in appearance-based estimation. Specifically, the combination of 3-D geometric-based and learning-based methods reduces the number of required training samples.

3. The gaze estimation bias caused by eye appearance distortion is learnt effectively using training samples obtained from a 5-seconds video clip.

The rest of the paper is organized as follows. Sec. 2 overviews the proposed approach and explains the decomposition scheme. Sec. 3 describes the proposed methods in detail. Sec. 4 shows the experimental results and Sec. 5 concludes the paper.

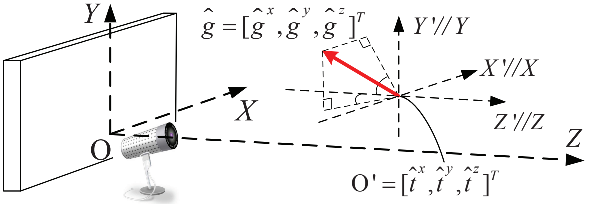
2 Overview of the approach

2.1 Problem statement

Table 1 defines some important notations. The generalized appearance-based gaze estimation problem can be formulated as using training data $\mathcal{T}$ to map the eye appearance feature $\hat{\boldsymbol{e}}$ to

Notation	Description
$\boldsymbol{e} \in \mathbb{R}^m$	Eye appearance feature vector extracted from an eye image
$\boldsymbol{r} = [r^x, r^y, r^z]^{\mathrm{T}} \in \mathbb{R}^3$	3-D head rotation vector[1]
$\boldsymbol{t} = [t^x, t^y, t^z]^{\mathrm{T}} \in \mathbb{R}^3$	3-D head translation vector
$\boldsymbol{g} = [g^x, g^y, g^z]^{\mathrm{T}} \in \mathbb{R}^3$	Unit vector for gaze direction under world coordinate system
$\{\hat{\boldsymbol{e}}, \hat{\boldsymbol{r}}, \hat{\boldsymbol{t}}, \hat{\boldsymbol{g}}\}$	Data for test input
$\mathcal{T}^e = \{\boldsymbol{e}_i \mid i = 1, \cdots, n\}$	Collection of appearance features of training samples
$\mathcal{T}^r = \{\boldsymbol{r}_i \mid i = 1, \cdots, n\}$	Collection of head rotations of training samples
$\mathcal{T}^t = \{\boldsymbol{t}_i \mid i = 1, \cdots, n\}$	Collection of head translations of training samples
$\mathcal{T}^g = \{\boldsymbol{g}_i \mid i = 1, \cdots, n\}$	Collection of gaze directions of training samples
$\mathcal{T} = \{\mathcal{T}^e, \mathcal{T}^r, \mathcal{T}^t, \mathcal{T}^g\}$	Dataset including all training samples
$\boldsymbol{r}_0, \boldsymbol{t}_0$	Constant values of fixed head rotation and translation
$\mathcal{T}_0 = \{\mathcal{T}_0^e, \mathcal{T}_0^r, \mathcal{T}_0^t, \mathcal{T}_0^g\}$	Subset of $\mathcal{T}$ consisting of training samples whose $\boldsymbol{r}_j = \boldsymbol{r}_0$ and $\boldsymbol{t}_j = \boldsymbol{t}_0$

Table 1: Definitions of notations used in this paper.



Y
$\hat{g} = [\hat{g}^x, \hat{g}^y, \hat{g}^z]^T$
Y'//Y
X'//X
X
Z'//Z
O
Z
$O' = [\hat{t}^x, \hat{t}^y, \hat{t}^z]^T$

Figure 1: Gaze direction unit vector $\hat{\boldsymbol{g}} = [\hat{g}^x, \hat{g}^y, \hat{g}^z]^{\mathrm{T}}$ under the world coordinate system.

the gaze direction unit vector $\hat{\boldsymbol{g}}$ under head pose $(\hat{\boldsymbol{r}}, \hat{\boldsymbol{t}})$:

$$\hat{\boldsymbol{g}} = \mathcal{M}(\hat{\boldsymbol{e}}, \hat{\boldsymbol{r}}, \hat{\boldsymbol{t}} \mid \mathcal{T}) \tag{1}$$

Typically, conventional feature-based methods assume a fixed head pose. Thus they are actually focused on a simplified version of the problem:

$$\hat{\boldsymbol{g}}\text{[2]} = \mathcal{M}_{\boldsymbol{r}_0, \boldsymbol{t}_0}(\hat{\boldsymbol{e}} \mid \mathcal{T}_0^e, \mathcal{T}_0^g) \tag{2}$$

while in this paper, we solve the original problem in Eq. (1) for the gaze direction vector $\hat{\boldsymbol{g}} = [\hat{g}^x, \hat{g}^y, \hat{g}^z]^{\mathrm{T}}$ under the world coordinate system (WCS), as shown in Fig. 1.

2.2 Proposed decomposition approach

The problem in Eq. (1) is about mapping eye appearance features to gaze direction vectors. This problem can be solved directly by collecting enough training samples under variant

[1] In our implementation, values of r^x and r^y are calculated as the angles made by the projections of the face normal $\boldsymbol{n}$ and the Z' axis in the planes $Y'O'Z'$ and $Z'O'X'$ (similar to the angles in Fig. 1), while r^z is the rotation angle around $\boldsymbol{n}$ (yaw).

[2] In practice, conventional methods usually estimate the 2-D gaze position on the screen instead of 3-D gaze direction vector for convenience because under a fixed head pose, their values directly correspond to each other.

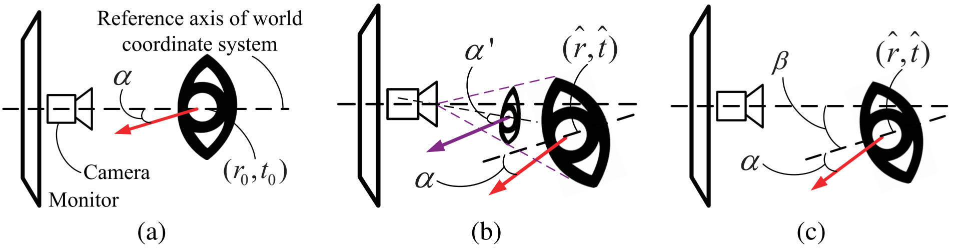


Figure 2: 2-D illustration of relationship between gaze direction and head pose. (a) Under a fixed head pose $(\boldsymbol{r}_0,\boldsymbol{t}_0)$, gaze direction α can be estimated from appearance by Eq. (2). (b) To obtain α under another head pose $(\hat{\boldsymbol{r}},\hat{\boldsymbol{t}})$, the estimated α' by Eq. (2) should be corrected because of captured eye appearance distortion. (c) Under head pose $(\hat{\boldsymbol{r}},\hat{\boldsymbol{t}})$, gaze direction under WCS should be further compensated for head rotation β.

head poses in $\mathcal{T}$ for regression, as proposed by Sugano *et al.* [10]. However, as data in $\mathcal{T}$ has 6 degrees of freedom for head poses, even when a large number (*e.g.* 10^3) of training samples are obtained, the accuracy is still insufficient.

We propose first solving the problem in Eq. (2) by assuming a fixed head pose $(\boldsymbol{r}_0,\boldsymbol{t}_0)$ as shown in Fig. 2(a) and then compensating for the estimation bias by taking into account the true head pose $(\hat{\boldsymbol{r}},\hat{\boldsymbol{t}})$. The bias under WCS mainly depends on two factors: 1) the estimation error caused by eye appearance distortion (see α' and α in Fig. 2(b)) in accordance with specific capture direction; and 2) the eye orientation variation in accordance with head rotation (see β in Fig. 2(c)). In fact, the problem in Eq. (1) is decomposed into:

$$\hat{\boldsymbol{g}} \simeq \mathcal{M}_{\boldsymbol{r}_0,\boldsymbol{t}_0}(\hat{\boldsymbol{e}}|\mathcal{T}_0^e,\mathcal{T}_0^g) \otimes \mathcal{C}^D_{\boldsymbol{r}_0,\boldsymbol{t}_0}(\hat{\boldsymbol{r}},\hat{\boldsymbol{t}}|\mathcal{T}) \otimes \mathcal{C}^R_{\boldsymbol{r}_0}(\hat{\boldsymbol{r}}) \tag{3}$$

where the operator '$\otimes$' indicates the manipulation in gaze direction vector via a series of specified rotations, and $\otimes\mathcal{C}^D_{\boldsymbol{r}_0,\boldsymbol{t}_0}(\hat{\boldsymbol{r}},\hat{\boldsymbol{t}}|\mathcal{T})$ and $\otimes\mathcal{C}^R_{\boldsymbol{r}_0}(\hat{\boldsymbol{r}})$ denote the compensations for eye appearance distortion and head rotation. Similar to Sugano *et al.* [10], we obtain the required head rotation and translation values via a computer vision-based head tracker.

2.3 Gaze estimation procedures

We implemented a head pose-free gaze tracking system based on the proposed approach. Only single camera was used. In general, the estimation includes the following steps.

Obtaining training data. Training data $\mathcal{T}=\{\mathcal{T}^e,\mathcal{T}^r,\mathcal{T}^t,\mathcal{T}^g\}$ are obtained via calibration. The user is asked to sit in front of the screen and gaze at certain positions on the screen (*i.e.* calibration points). A single camera is used to capture the user's appearances. Then, the inner eye corners are detected using edge maps and serve as landmark points for rectangular eye region alignment and extraction. Finally, these extracted eye regions are rescaled and raster-scanned into eye appearance features $\{\boldsymbol{e}_i\}$. The head poses $\{\boldsymbol{r}_i\}$ and $\{\boldsymbol{t}_i\}$ are calculated from the raw data provided by a vision-based head pose tracker [7]. The gaze positions $\{\boldsymbol{x}_i\}$ on the screen are saved to calculate the gaze direction vectors $\{\boldsymbol{g}_i\}$. Specifically, training samples in $\mathcal{T}_0$ are collected under a fixed head pose $(\boldsymbol{r}_0,\boldsymbol{t}_0)$, whereas the others are obtained from a short video clip that is introduced later in Sec. 3.3.

Gaze estimation. Any test data $\{\hat{\boldsymbol{e}},\hat{\boldsymbol{r}},\hat{\boldsymbol{t}}\}$ are obtained similarly to the training data. With the training data $\mathcal{T}$, gaze direction vector $\hat{\boldsymbol{g}}$ is estimated from $\{\hat{\boldsymbol{e}},\hat{\boldsymbol{r}},\hat{\boldsymbol{t}}\}$ by Eq. (3). Each procedure in Eq. (3) is introduced in detail in the following sections.

3 Proposed methods

The decomposition-based approach for head pose-free gaze estimation was introduced in Sec. 2.2. In this section, we explain each step for solving the decomposed problem in Eq. (3).

3.1 Estimation under fixed head pose by $\mathcal{M}_{\boldsymbol{r}_0,\boldsymbol{t}_0}(\hat{\boldsymbol{e}}|\mathcal{T}_0^e,\mathcal{T}_0^g)$

The training data $\mathcal{T}_0 = \{\mathcal{T}_0^e,\mathcal{T}_0^r,\mathcal{T}_0^t,\mathcal{T}_0^g\}$ obtained under a fixed head pose $(\boldsymbol{r}_0,\boldsymbol{t}_0)$ are used. We obtain the training samples sparsely, meaning that gaze positions are selected with large intervals on the screen to avoid a tedious calibration stage. Let m-D vector $\boldsymbol{e}_j \in \mathcal{T}_0^e$ denote the eye appearance feature generated from j-th eye image and $\boldsymbol{g}_j \in \mathcal{T}_0^g$ denote the corresponding gaze direction vector, where $j = 1,\cdots,n_0$. The head pose is fixed at $(\boldsymbol{r}_0,\boldsymbol{t}_0)$ and thus not considered. We seek a mapping $\boldsymbol{e}_j \mapsto \boldsymbol{g}_j$ from the m-D feature space to the 3-D gaze direction vector space.

It has proven that in such cases, interpolation methods using pre-selected local training samples are effective [10, 12]. Unlike them, we propose directly solving the problem using all the samples in $\mathcal{T}_0^e$ and $\mathcal{T}_0^g$:

$$\mathcal{M}_{\boldsymbol{r}_0,\boldsymbol{t}_0}(\hat{\boldsymbol{e}}|\mathcal{T}_0^e,\mathcal{T}_0^g): \quad \hat{\boldsymbol{g}} = \sum_{j=1}^{n_0} w_j \boldsymbol{g}_j \quad \text{subject to} \quad \{w_j\} = \arg\min \|\hat{\boldsymbol{e}} - \sum_{j=1}^{n_0} w_j \boldsymbol{e}_j\|^2 \tag{4}$$

It has not been mentioned by the previous methods that under the condition of sparse sampling and $m \gg n_0$, solving Eq. (4) automatically selects a small number of local training samples with weights $w_j > 0$. Therefore, it becomes unnecessary to pre-select the 'local samples'. We demonstrate in Sec. 4 that estimation by Eq. (4) achieves high accuracy.

3.2 Compensation for head rotation by $\mathcal{C}^R_{\boldsymbol{r}_0}(\hat{\boldsymbol{r}})$

In this step, we ignore the eye appearance distortion and only focus on compensation for head rotation. For a test sample, we initially estimate the gaze direction vector $\hat{\boldsymbol{g}}^0$ by assuming head rotation $\boldsymbol{r}_0$ and then apply a series of rotations to the head coordinate system so that $\boldsymbol{r}_0 \Rightarrow \hat{\boldsymbol{r}}$, which simultaneity rotates $\hat{\boldsymbol{g}}^0$ to the final result $\hat{\boldsymbol{g}}$ under $\boldsymbol{r}$. This procedure is used to compensate for head rotation and denoted as

$$\hat{\boldsymbol{g}} = \hat{\boldsymbol{g}}^0 \otimes \mathcal{C}^R_{\boldsymbol{r}_0}(\hat{\boldsymbol{r}}) = \mathcal{R}(\hat{\boldsymbol{g}}^0,\boldsymbol{r}_0,\hat{\boldsymbol{r}}) \tag{5}$$

where the function $\boldsymbol{a} = \mathcal{R}(\boldsymbol{a}_0,\boldsymbol{r}_0,\boldsymbol{r})$ finds the local coordinate system rotations starting from $\boldsymbol{r}_0$ to $\boldsymbol{r}$ and computes $\boldsymbol{a}$ from the initial vector $\boldsymbol{a}_0$ simultaneity by using the same rotations. The calculation is provided in the Appendix. A.

3.3 Learning $\mathcal{C}^D_{\boldsymbol{r}_0,\boldsymbol{t}_0}(\hat{\boldsymbol{r}},\hat{\boldsymbol{t}}|\mathcal{T})$ from a short video clip taken with varying head poses

While the eye orientation varies relatively to the camera, distortion exists in the captured eye image. In the eye coordinate system (ECS), this orientation is depicted by the capture direction that is calculated by a vector pointing to the camera centre. In this section we investigate the relationship between the changes of capture directions and the biases of gaze estimations caused by eye appearance distortions under ECS.

The capture direction unit vectors are denoted as $\boldsymbol{v}^c \in \mathbb{R}^3$ and $\boldsymbol{v}^{c,0} \in \mathbb{R}^3$ under head poses $(\boldsymbol{r},\boldsymbol{t})$ and $(\boldsymbol{r}_0,\boldsymbol{t}_0)$. Then the capture direction variation is $\Delta\boldsymbol{v}^c = \boldsymbol{v}^c - \boldsymbol{v}^{c,0}$. Also, the

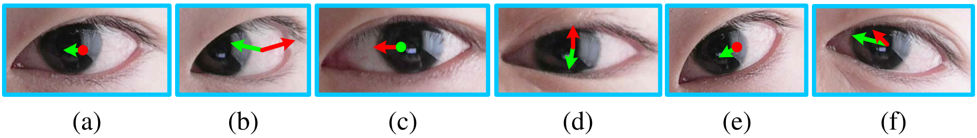


Figure 3: Eye images from video clip. Green/red arrows indicate capture direction vectors/eye (face) normals. Note that in (a) and (e), capture directions are similar under ECS, thus their appearance distortions and gaze direction biases are also similar.

initially estimated gaze direction vector from a distorted eye image is denoted as $\boldsymbol{g}^{d,0} = \mathcal{M}_{\boldsymbol{r}_0,\boldsymbol{t}_0}(\boldsymbol{e}|\mathcal{T}_0^e,\mathcal{T}_0^g)$ and the ground truth as $\boldsymbol{g}^0 = \boldsymbol{g} \otimes (\mathcal{C}_{\boldsymbol{r}_0}^R(\boldsymbol{r}))^{-1} = \mathcal{R}(\boldsymbol{g},\boldsymbol{r},\boldsymbol{r}_0)$, where $\boldsymbol{g}$ is the true gaze direction vector under WCS. Then the gaze direction bias is represented by a 2-D rotation $\Delta\boldsymbol{\phi} = [\Delta\phi^x, \Delta\phi^y]$ that rotates $\boldsymbol{g}^{d,0}$ to $\boldsymbol{g}^0$. The calculations of $\boldsymbol{v}^c$ and $\Delta\boldsymbol{\phi}$ are shown in the Appendix. B and C. We predict the gaze direction bias of any test sample via the mapping of $\Delta\boldsymbol{v}^c \mapsto \Delta\boldsymbol{\phi}$, which is learnt by regression.

Training the regression needs adequate training samples with different $\Delta\boldsymbol{v}^c$. Note that there is no requirement of specified gaze positions or head poses for the training samples. Thus we propose an unconventional calibration process that captures a short video clip while the user is gazing at a *fixed but arbitrarily assigned position* on the screen and moving his/her head (just rotating is effective). As there is no change of gaze positions and the user's head motion is free, the procedure can be done within several seconds while obtaining sufficient training samples. Therefore, a tedious calibration is avoided. Fig. 3 shows examples of eye images from a captured video clip and visualizes their camera directions under ECS.

For every obtained training sample $\{\boldsymbol{e}_i, \boldsymbol{r}_i, \boldsymbol{t}_i, \boldsymbol{g}_i\}$, we calculate $\Delta\boldsymbol{\phi}_i$ and $\Delta\boldsymbol{v}_i^c$ as described above, and then the regression is performed on the basis of a Gaussian Process (GP) model. Note that $\{\Delta\boldsymbol{\phi}_i\} \in \mathbb{R}^2$ has two degrees of freedom, so we utilize two 1-D regressions. If the first dimension $\{\Delta\phi_i^x\}$ is taken as an example, the regression function is denoted as follows:

$$\Delta\phi_i^x = f_x(\Delta\boldsymbol{v}_i^c) \sim \mathcal{GP}(m(\Delta\boldsymbol{v}_i^c), k_\omega(\Delta\boldsymbol{v}_i^c, \Delta\boldsymbol{v}_j^c)) \tag{6}$$

where the mean function and covariance function are defined by

$$m(\Delta\boldsymbol{v}_i^c) = 0, \quad k_\omega(\Delta\boldsymbol{v}_i^c, \Delta\boldsymbol{v}_j^c) = \kappa \exp(-\|\Delta\boldsymbol{v}_i^c - \Delta\boldsymbol{v}_j^c\|^2/2l^2) + \sigma^2\delta_{ij} \tag{7}$$

where σ^2 comes from the observation noise. The training procedure uses the above obtained training data $\boldsymbol{y} = [\Delta\phi_1^x, \cdots, \Delta\phi_i^x, \cdots, \Delta\phi_n^x]^{\mathrm{T}}$ and $V = [\Delta\boldsymbol{v}_1^c, \cdots, \Delta\boldsymbol{v}_i^c, \cdots, \Delta\boldsymbol{v}_n^c]^{\mathrm{T}}$ to optimize the hyperparameters $\omega = \{\kappa, l, \sigma^2\}$ by minimizing the log marginal likelihood function

$$\log p(\boldsymbol{y}|V,\omega) = -\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}(K_\omega(V,V) + \sigma^2 I)^{-1}\boldsymbol{y} - \frac{1}{2}\log|K_\omega(V,V) + \sigma^2 I| - \frac{n}{2}\log 2\pi \tag{8}$$

where $K_\omega(V,V)$ is the covariance matrix whose element in (i,j) is $k_\omega(\Delta\boldsymbol{v}_i^c, \Delta\boldsymbol{v}_j^c)$. With these hyperparameters, the predicted $\Delta\hat{\phi}^x$ from $\Delta\hat{\boldsymbol{v}}^c$ of a test sample is given by

$$\Delta\hat{\phi}^x = K_\omega(\Delta\hat{\boldsymbol{v}}^c, V)(K_\omega(V,V) + \sigma^2 I)^{-1}\boldsymbol{y} \tag{9}$$

$$\mathrm{cov}(\Delta\hat{\phi}^x) = 1 - K_\omega(\Delta\hat{\boldsymbol{v}}^c, V)(K_\omega(V,V) + \sigma^2 I)^{-1}K_\omega(V, \Delta\hat{\boldsymbol{v}}^c) \tag{10}$$

After regression for both $\Delta\hat{\phi}^x$ and $\Delta\hat{\phi}^y$, the bias caused by appearance distortion can be compensated for by $\hat{\boldsymbol{g}}^0 = \hat{\boldsymbol{g}}^{d,0} \otimes \mathcal{C}_{\boldsymbol{r}_0,\boldsymbol{t}_0}^D(\hat{\boldsymbol{r}}, \hat{\boldsymbol{t}}|\mathcal{T})$, which rotates $\hat{\boldsymbol{g}}^{d,0}$ by $\Delta\hat{\phi}^x$ and $\Delta\hat{\phi}^y$.

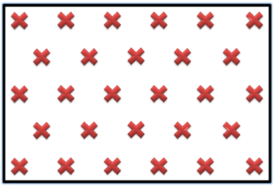


Figure 4: Gaze positions on screen for training samples.

Method	Error	Training samples
Proposed	**0.85°**	**33**
S^3GP+edge+filter [17]	0.83°	16 labelled and 75 unlabelled
Tan *et al.* [12]	0.5°	252
Baluja *et al.* [1]	1.5°	2000
Xu *et al.* [18]	1.5°	3000

Table 2: Comparison of estimation accuracy under fixed head pose.

4 Experimental verification

The performance of the proposed method was evaluated via extensive experiments. A system was built upon a desktop PC with a 22-inch LCD monitor and a VGA resolution webcam, which are about 50cm from the user. The estimation errors were measured in degrees. The entire assessment includes three stages: 1) evaluation of gaze estimation accuracy under fixed head pose; 2) verification of eye appearance distortion compensation via a short video clip; and 3) the overall assessment of estimation accuracy under free head motion.

4.1 Evaluation of estimation accuracy under fixed head pose

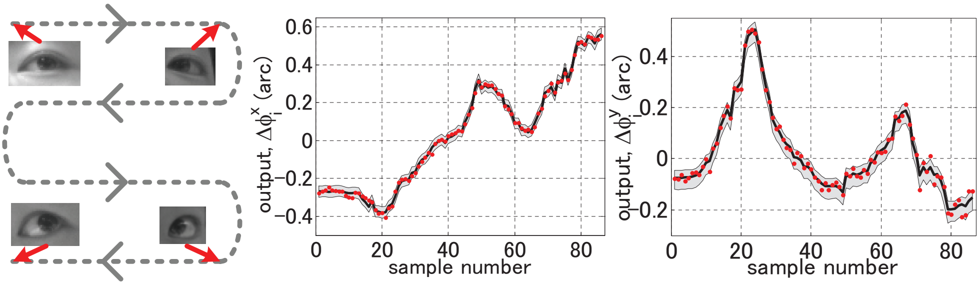
We first focused on the conventional problem in Eq. (2), which assumes a fixed head pose. The user was requested to gaze at each point displayed on the screen as shown in Fig. 4, while the eye appearance and other data were collected for training samples. Then test samples were obtained similarly. Finally, gaze directions of the test samples were estimated by the training samples using the method introduced in Sec. 3.1.

Table 2 compares the estimation accuracy of the proposed method with those of existing appearance-based methods. Our method obviously achieves a good trade-off between easy calibration (it requires only 33 training samples) and high precision.

4.2 Verification of eye appearance distortion compensation

We examined the ability of the method proposed in Sec. 3.3 to compensate for eye appearance distortion. The training data were obtained from a 5-seconds video clip of the eye appearances recorded while the user was gazing at the same position on the screen and rotating his head as shown in Fig. 5(a). The range of the corresponding capture angles is given in Table 3. The regression that maps the capture direction variation to gaze direction estimation bias was obtained using the method introduced in Sec. 3.3. Fig. 5(b) plots the regression curve with error bars that indicate a 90% confidence interval.

To verify the effectiveness of the proposed compensation technique, leave-one-out experiments were conducted. Each training sample was selected as a test sample for once while the other samples were used to train the regression. Then the estimation bias of the test sample was obtained via regression and then used in compensation. Fig. 6 plots the estimation errors, and Table 4 gives their averages. These results demonstrate that the proposed regression-based method effectively compensates for eye appearance in gaze estimation.



(a) (b)

Figure 5: Regression for appearance distortion compensation. (a) Distorted eye images captured under head motion and fixed gaze position from short video clip as training samples. (b) Regression results for $\Delta\phi_i^x$ and $\Delta\phi_i^y$. The shaded region shows 90% confidence interval.

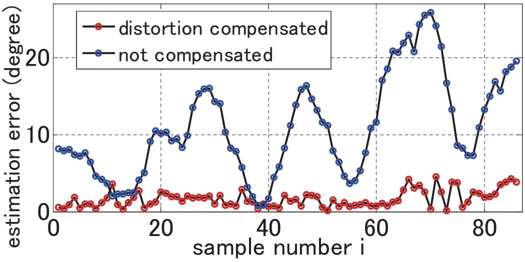


Figure 6: Results of leave-one-out experiments with/without appearance distortion compensation.

Rotation	angle range
around x	$-13.90^\circ \sim 30.17^\circ$
around y	$-30.65^\circ \sim 38.62^\circ$

Table 3: Angle ranges of capture directions $\{\boldsymbol{v}_i^c\}$ for all training samples from video clip.

Compensation	Average error
With	1.65°
Without	10.85°

Table 4: Average results with/without appearance distortion compensation.

4.3 Overall assessment of estimation accuracy under free head motion

The gaze estimation efficacy under free head motion is evaluated. Experiments are done with 4 subjects, three of whom are non-experienced users of any gaze tracker. Training samples are first collected as introduced before. Then test samples are obtained for the experiments. Table 5 shows the head motion ranges covered by the test samples from subject S1, which are sufficiently large for a 22-inch screen user. Fig. 7 illustrates curves of estimation errors with/without the proposed compensation methods, which demonstrate that only if the compensations are fully applied, the estimation becomes accurate. Table 6 gives all the estimation errors and also compares our results to those of the method by Sugano *et al.* [10], which is one of the very few previously known head pose-free appearance-based methods. The proposed method obviously achieves higher accuracy and requires much less calibration effort. In fact, the average estimation accuracy of 2.38° is comparable to the feature-based methods [3, 9, 15, 19, 20], which commonly report accuracies of $1 \sim 3^\circ$ by utilizing complex devices such as infrared/stereo cameras/lights and pan-tilt units.

5 Conclusion and discussion

We have presented a novel appearance-based gaze estimation approach that allows free head motion. The high-dimensional original problem is decomposed into subproblems.

Type	Range
x-translation	$-30.91mm \sim 107.89mm$
y-translation	$10.18mm \sim 50.14mm$
z-translation	$533.78mm \sim 599.58mm$
x-rotation	$3.89^\circ \sim 19.39^\circ$
y-rotation	$-13.41^\circ \sim 12.21^\circ$
z-rotation	$-8.51^\circ \sim 3.49^\circ$

Table 5: Head motion ranges of test samples from subject S1.

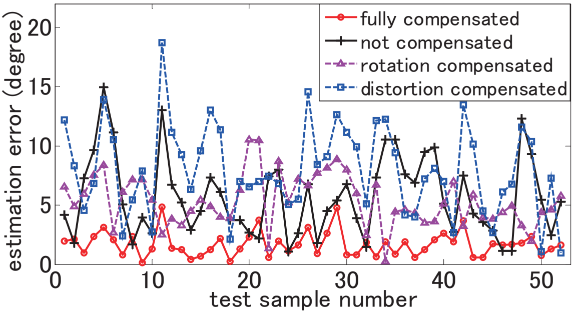


Figure 7: Final results of gaze estimation under free head motion for subject S1. Comparisons are provided with/without proposed compensations.

Subject	Full comp.	Dist. comp.	Rot. comp.	No comp.	Training samples
S1	1.70°	7.97°	5.34°	5.68°	33 training samples and 5-seconds video clip
S2	2.49°	4.28°	5.44°	4.09°	
S3	2.74°	6.01°	3.81°	7.86°	
S4	2.57°	4.14°	2.94°	6.31°	
Average	**2.38°**	5.60°	4.38°	5.99°	
Sugano *et al.* [10]	$4^\circ \sim 5^\circ$				$\approx 10^3$

Table 6: Estimation accuracy under free head motion. S2-S4 are non-experienced users.

Then initial estimation and subsequent compensations are done by either learning-based or geometric-based methods. Experimental results demonstrate two major benefits: 1) high estimation accuracy is achieved, and 2) the number of training samples is significantly reduced.

To our knowledge, the proposed method is the most accurate appearance-based method under free head motion and is comparable to the feature-based methods. On the other hand, difficulty still exists in aligning and extracting the deformed eye images under different head poses, which is the major problem we plan to solve in the future.

Appendix: Calculations of $\boldsymbol{a} = \mathcal{R}(\boldsymbol{a}_0, \boldsymbol{r}_0, \boldsymbol{r})$, $\boldsymbol{v}^c$, and $\Delta\phi$

A. Let the initial vector $\boldsymbol{a}_0$ be rotated along with the local coordinate system by $\boldsymbol{r}_0 \Rightarrow [0,0,r_0^z]^{\mathrm{T}} \Rightarrow [0,0,r^z]^{\mathrm{T}} \Rightarrow \boldsymbol{r}$, then we have

$$\boldsymbol{a} = \mathcal{R}(\boldsymbol{a}_0, \boldsymbol{r}_0, \boldsymbol{r})$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_2^x & -\sin\theta_2^x \\ 0 & \sin\theta_2^x & \cos\theta_2^x \end{bmatrix} \begin{bmatrix} \cos\theta_2^y & 0 & \sin\theta_2^y \\ 0 & 1 & 0 \\ -\sin\theta_2^y & 0 & \cos\theta_2^y \end{bmatrix} \begin{bmatrix} \cos\theta_{12}^z & -\sin\theta_{12}^z & 0 \\ \sin\theta_{12}^z & \cos\theta_{12}^z & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta_1^y & 0 & \sin\theta_1^y \\ 0 & 1 & 0 \\ -\sin\theta_1^y & 0 & \cos\theta_1^y \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_1^x & -\sin\theta_1^x \\ 0 & \sin\theta_1^x & \cos\theta_1^x \end{bmatrix} \boldsymbol{a}_0 \quad (11)$$

where $\theta_1^x = -r_0^x$, $\theta_1^y = -\arctan(\tan r_0^y \cdot \cos r_0^x)$, $\theta_{12}^z = r^z - r_0^z$, $\theta_2^y = \arctan(\tan r^y \cdot \cos r^x)$, and $\theta_2^x = r^x$.

B. The capture direction unit vector $\boldsymbol{v}^c$ under ECS is determined by both head translation and head rotation. It can be geometrically computed by

$$\boldsymbol{v}^c = \mathcal{R}([-t_i^x, -t_i^y, -t_i^z]^{\mathrm{T}} / (t_i^{x2} + t_i^{y2} + t_i^{z2})^{\frac{1}{2}}, \boldsymbol{r}, [0,0,0]^{\mathrm{T}}) \tag{12}$$

C. The bias $\Delta\boldsymbol{\phi} = [\Delta\phi^x, \Delta\phi^y]$ rotates $\boldsymbol{g}^{d,0}$ to $\boldsymbol{g}^0$, thus can be obtained by solving

$$\boldsymbol{g}^0 = \begin{bmatrix} \cos\Delta\phi^y & 0 & \sin\Delta\phi^y \\ 0 & 1 & 0 \\ -\sin\Delta\phi^y & 0 & \cos\Delta\phi^y \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\Delta\phi^x & -\sin\Delta\phi^x \\ 0 & \sin\Delta\phi^x & \cos\Delta\phi^x \end{bmatrix} \boldsymbol{g}^{d,0} \tag{13}$$

whose solution is

$$\Delta\phi_i^x = \arctan(g^{0,y} / - g^{0,z}) - \arctan(g^{d,0,y} / - g^{d,0,z}) \tag{14}$$

$$\Delta\phi^y = \arctan(g_i^{0,x} / g^{0,z}) + \arctan(g^{d,0,x} / (1 - (g^{d,0,x})^2 - (g^{0,y})^2)^{\frac{1}{2}}) \tag{15}$$

References

[1] S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. In *Proceedings of Advances in Neural Information Processing Systems*, volume 6, pages 753–760, 1994.

[2] D. Beymer and M. Flickner. Eye gaze tracking using an active stereo head. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, pages 451–458, 2003.

[3] X.L.C. Brolly and J.B. Mulligan. Implicit Calibration of a Remote Gaze Tracker. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW 2004)*, page 134, 2004.

[4] D.W. 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*, 32(3): 478–500, 2010.

[5] J.J. Kang, M. Eizenman, E.D. Guestrin, and E. Eizenman. Investigation of the Cross-Ratios method for Point-of-Gaze estimation. *IEEE Transactions on Biomedical Engineering*, 55(9):2293–2302, 2008.

[6] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Inferring human gaze from appearance via adaptive linear regression. In *Proceedings of the 13th IEEE International Conference on Computer Vision (ICCV 2011)*, 2011.

[7] S. Machines. faceAPI. http://www.seeingmachines.com/product/faceapi/.

[8] C.H. Morimoto and M.R.M. Mimica. Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, 98(1):4–24, 2005.

[9] T. Nagamatsu, J. Kamahara, and N. Tanaka. 3D gaze tracking with easy calibration using stereo cameras for robot and human communication. In *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2008)*, pages 59–64, 2008.

[10] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike. An incremental learning method for unconstrained gaze estimation. In *Proceedings of the 10th European Conference on Computer Vision (ECCV 2008)*, pages 656–667, 2008.

[11] Y. Sugano, Y. Matsushita, and Y. Sato. Calibration-free gaze sensing using saliency maps. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, pages 2667–2674, 2010.

[12] Kar-Han Tan, D.J. Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. In *Proceedings of the 6th IEEE Workshop on Applications of Computer Vision (WACV 2002)*, pages 191–195, 2002.

[13] G.D.M. Underwood. *Cognitive processes in eye guidance*. Oxford University Press, USA, 2005.

[14] R. Valenti and T. Gevers. Accurate eye center location and tracking using isophote curvature. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pages 1–8, 2008.

[15] A. Villanueva and R. Cabeza. A novel gaze estimation system with one calibration point. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 38 (4):1123–1138, 2008.

[16] J.G. Wang, E. Sung, and R. Venkateswarlu. Eye gaze estimation from a single image of one eye. In *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV 2009)*, pages 136–143, 2003.

[17] O. Williams, A. Blake, and R. Cipolla. Sparse and semi-supervised visual mapping with the S^{3}GP. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, pages 230–237, 2006.

[18] L. Q Xu, D. Machin, and P. Sheppard. A novel approach to real-time non-intrusive gaze finding. In *Proceedings of British Machine Vision Conference (BMVC 1998)*, pages 428–437, 1998.

[19] D. H Yoo and M. J Chung. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *Computer Vision and Image Understanding*, 98(1):25–51, 2005.

[20] Z. Zhu and Q. Ji. Novel eye gaze tracking techniques under natural head movement. *IEEE Transactions on Biomedical Engineering*, 54(12):2246–2260, 2007.