Head Pose-Free Appearance-Based Gaze Sensing via Eye Image Synthesis

Feng Lu Yusuke Sugano Takahiro Okabe Yoichi Sato
Institute of Industrial Science, the University of Tokyo, Japan
{*lufeng, sugano, takahiro, ysato*}*@iis.u-tokyo.ac.jp*

Abstract

This paper addresses the problem of estimating human gaze from eye appearance under free head motion. Allowing head motion remains challenging because eye appearance changes significantly for different head poses, and thus new head poses require new training images. To avoid repetitive training, we propose to produce synthetic training images for varying head poses. First, we model pixel displacements between head-moving eye images as 1D pixel flows, and then produce such flows to synthesize new training images from the original training images captured under a fixed default head pose. Specifically, we produce all the required 1D flows by using only four additionally captured images. Our method was successfully tested with extensive experiments to demonstrate its effectiveness.

1. Introduction

Estimating human gaze is useful for many applications such as human-computer interaction, medical treatment, and marketing research. For years, it has attracted much research interest, especially with the rapid development of computer vision technology.

Computer vision-based methods can be categorized as either feature/model-based or appearance-based [4]. The former methods extract small features such as infrared corneal reflections, pupil center [11], and iris contour [12] from high resolution eye images to fit specific eyeball models. However, they usually require dedicated hardware typically with infrared illumination.

On the other hand, appearance-based methods use an entire eye image as a high-dimensional input, and therefore require only a single video camera under uncontrolled lighting condition. Baluja and Pomerleau [2] proposed neural networks trained by thousands of training samples. Tan *et al*. [10] proposed a method based on local linear mapping between eye image manifold and gaze space using 252 training samples. To reduce the number of training samples, Williams *et al*. [13] developed a semi-supervised method to accept unlabeled training samples. Recently, Sugano *et al*. [8] proposed a novel method that automatically collects labeled samples by utilizing saliency prior from a video clip. Lu *et al*. [6] introduced an efficient adaptive regression method that uses significantly fewer training samples to guarantee accurate estimation.

While these methods work well with a fixed head pose, their performance degrades greatly when a user's head is not stationary. The reason is that head motion deforms the input eye image so drastically that it can differ significantly from original training images even if they all correspond to the same gaze direction.

Few appearance-based methods have been reported to deal with this problem. Sugano *et al*. [9] proposed recollecting training images for each cluster of new head poses, which results in a long-term training. Lu *et al*. [5] suggested initiating estimation with the original training images and then compensating for the bias via regression. However, nearly 100 additional training images for different head poses are needed for regression.

In this paper, we allow head motion in appearance-based gaze estimation by producing synthetic training images. Like in the conventional methods, original gaze training images are first captured only under a fixed head pose. Then for any unseen head poses, their training images are *synthesized* rather than physically captured. This is done by using a 1D pixel displacement model constructed from only four additionally captured images. Using the synthetic training images, gaze estimation can be done in a conventional way.

In terms of image-based rendering, the synthesis in our method belongs to the category where the geometric model is used implicitly (Shum *et al*. [7]). Meanwhile, our method is distinctive in that 1) we rectify the input images using 3D head pose information while conventional methods usually need to recover the fundamental matrix, and 2) instead of extracting or assigning correspondence for sparse feature points, we design techniques to pursuit dense pixel flow for accurate synthesis.

2. Proposed method

Fixed-head pose methods [10, 6, 8] use training images $\{\boldsymbol{I}_{\boldsymbol{\Gamma}_0,m}\}$ captured under a fixed head pose $\boldsymbol{\Gamma}_0$ to estimate gaze direction of an input eye image $\hat{\boldsymbol{I}}_{\boldsymbol{\Gamma}_0}$, where m indicates different gaze directions due to eye ball rotation. However, $\{\boldsymbol{I}_{\boldsymbol{\Gamma}_0,m}\}$ cannot estimate for input image $\hat{\boldsymbol{I}}_{\boldsymbol{\Gamma}}$ from a different head pose $\boldsymbol{\Gamma}$, because eye appearance changes greatly with head motion.

This paper proposes a novel method to synthesize training images $\{\boldsymbol{I}_{\boldsymbol{\Gamma},m}\}$ for any unseen head pose $\boldsymbol{\Gamma}$ based on the original training images $\{\boldsymbol{I}_{\boldsymbol{\Gamma}_0,m}\}$ and the corresponding 1D pixel displacement flow (shorted as "1D flow" or "flow") $\mathbf{u}_{\boldsymbol{\Gamma}}$. To do this, only four additional eye images $\{\boldsymbol{I}_{\boldsymbol{\Gamma}_i}|i = 1\dots4\}$ captured under four reference head poses $\{\boldsymbol{\Gamma}_1 \dots \boldsymbol{\Gamma}_4\}$ are required. Algorithm 1 overviews the method.

Algorithm 1 Overview of the proposed method

- Rectify all the images (Section 2.1)
- Estimate flows $\{\mathbf{u}_{\boldsymbol{\Gamma}_i}\}$ for $\{\boldsymbol{I}_{\boldsymbol{\Gamma}_i}\}$ (Section 2.2)

while input image $\hat{\boldsymbol{I}}_{\boldsymbol{\Gamma}}$ under unseen head pose $\boldsymbol{\Gamma}$ **do**

- Produce $\mathbf{u}_{\boldsymbol{\Gamma}}$ by using $\{\mathbf{u}_{\boldsymbol{\Gamma}_i}\}$ (Section 2.1)
- Synthesize $\{\boldsymbol{I}_{\boldsymbol{\Gamma},m}\}$ using $\mathbf{u}_{\boldsymbol{\Gamma}}$ and $\{\boldsymbol{I}_{\boldsymbol{\Gamma}_0,m}\}$ (Section 2.3)
- Estimate gaze direction of $\hat{\boldsymbol{I}}_{\boldsymbol{\Gamma}}$ using $\{\boldsymbol{I}_{\boldsymbol{\Gamma},m}\}$

end while

2.1. 1D pixel displacement model

This section proposes a 1D pixel displacement model to handle eye appearance variation due to head motion. The goal is that, for an unseen head pose, its corresponding eye appearance can be synthesized using a correctly obtained 1D flow, and such flow can be produced only from certain reference flows.

To do this, the key is to regard head motion as camera motion. To be specific, head-moving images captured by a fixed camera are considered to be captured by multi-view cameras with a fixed head pose (Figure 1(a)). The camera positions can be calculated from head pose parameters obtained by a head pose tracker. Because these camera positions are unconstrained, for further treatment, we move all the cameras parallel and onto the same camera plane. As a result, all the images lie on an image plane parallel to the camera plane (Figure 1(b)). Note that this process deforms the images while we can handle such deformation via projection transformation.

For any 3D point $\mathbf{P}$ on the eye surface, denote the camera positions for different head poses as $\{\mathbf{C}_i\}$ on the camera plane, and the recorded pixels as $\{\mathbf{p}_i\}$ on the image plane, as shown in Figure 1(b). Note the image

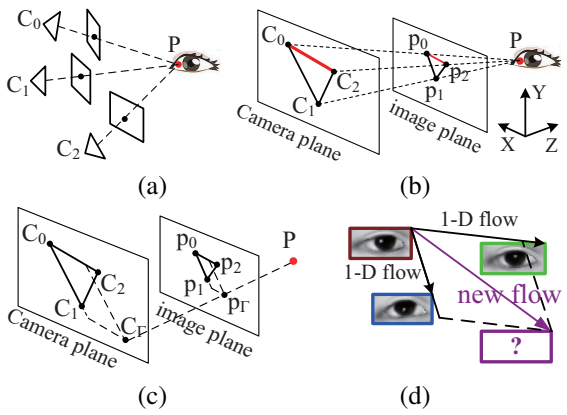

Figure 1. (a) Head motion is regarded as multi-camera capturing. (b) The cameras are parallelized and moved to the camera plane while images go to the parallel image plane. (c,d) New flow is produced for unseen head pose for image synthesis.

plane and camera plane are parallel, therefore any polygons formed by $\{\mathbf{C}_i\}$ and $\{\mathbf{p}_i\}$ are always *similar*. This leads to two important conclusions:

(1) Pixel displacement $\mathbf{p}_i - \mathbf{p}_j$ for any 3D point $\mathbf{P}$ due to head motion $i \Rightarrow j$ is parallel to camera motion $\mathbf{C}_i - \mathbf{C}_j$, meaning that the pixel flow has identical direction at all pixels, which we call "1D flow".

(2) Pixel displacement $\mathbf{p}_i - \mathbf{p}_j$ is also proportional to $\mathbf{C}_i - \mathbf{C}_j$ in length, and thus if we know the camera position of any unseen head pose, we can easily compute its pixel flow using known samples.

In practice, we capture four reference eye images $\{\boldsymbol{I}_{\boldsymbol{\Gamma}_i}|i = 1\dots4\}$ without eye ball rotation under four reference head poses $\{\boldsymbol{\Gamma}_i|i = 1\cdots4\}$ other than the default head pose $\boldsymbol{\Gamma}_0$. Their corresponding flows $\{\mathbf{u}_{\boldsymbol{\Gamma}_i}\}$ with respect to the default eye image under $\boldsymbol{\Gamma}_0$ can be estimated as described later in Section 2.2. Now our goal is to find the unknown 1D flow $\mathbf{u}_{\boldsymbol{\Gamma}}$ for an unseen head pose $\boldsymbol{\Gamma}$. Without lose of generality, let $\mathbf{C}_0$ and $\{\mathbf{C}_i\}$ denote the corresponding camera positions of $\boldsymbol{\Gamma}_0$ and $\{\boldsymbol{\Gamma}_i\}$ in the camera plane, and assume that the camera position $\mathbf{C}_{\boldsymbol{\Gamma}}$ of the unseen head pose $\boldsymbol{\Gamma}$ is near $\mathbf{C}_1$ and $\mathbf{C}_2$. We solve the following equation for λ_1 and λ_2:

$$\mathbf{C}_{\boldsymbol{\Gamma}} - \mathbf{C}_0 = \lambda_1(\mathbf{C}_1 - \mathbf{C}_0) + \lambda_2(\mathbf{C}_2 - \mathbf{C}_0) \in \mathbb{R}^3 \quad (1)$$

where $\lambda_1(\mathbf{C}_1 - \mathbf{C}_0)$ and $\lambda_2(\mathbf{C}_2 - \mathbf{C}_0)$ are considered the two sides of a parallelogram on the camera plane and $\mathbf{C}_{\boldsymbol{\Gamma}} - \mathbf{C}_0$ is actually the diagonal. Thus a *similar* parallelogram exists in the image plane to produce $\mathbf{u}_{\boldsymbol{\Gamma}}$

$$\mathbf{u}_{\boldsymbol{\Gamma}} = \lambda_1\mathbf{u}_{\boldsymbol{\Gamma}_1} + \lambda_2\mathbf{u}_{\boldsymbol{\Gamma}_2} \in \mathbb{R}^2 \quad (2)$$

for all the eye surface points. An illustration is shown in Figure 1(c,d). To summarize, we can produce pixel flows for any unseen head poses by using two of the four reference flows $\{\mathbf{u}_{\mathbf{\Gamma}_i}\}$, where $\{\mathbf{u}_{\mathbf{\Gamma}_i}\}$ can be estimated from only four additional images captured under four reference head poses, as described in the next section.

2.2. Reference flow estimation

This section describes how to extract the four reference 1D flows $\{\mathbf{u}_{\mathbf{\Gamma}_i}\}$. First of all, to capture the reference eye images $\{\boldsymbol{I}_{\mathbf{\Gamma}_i}\}$ under four reference head poses $\{\mathbf{\Gamma}_1 \dots \mathbf{\Gamma}_4\}$, the user just needs to turn his/her head upward/downward/leftward/rightward without eye ball rotation. Then for each $\boldsymbol{I}_{\mathbf{\Gamma}_i}$, we compute its flow $\mathbf{u}_{\mathbf{\Gamma}_i}$ with regard to the eye image in $\{\boldsymbol{I}_{\mathbf{\Gamma}_0,m}\}$ which has the same eye ball orientation. Denote these two images as $\boldsymbol{I}$ and $\boldsymbol{J}$, and the flow as $\mathbf{u}$. Remember that we already know the direction of $\mathbf{u}$ from camera displacement. Therefore we rewrite $\mathbf{u}$ with inclination angle θ:

$$\mathbf{u} = [u(\mathbf{x})\cos\theta, u(\mathbf{x})\sin\theta]^{\mathrm{T}} \tag{3}$$

Hereafter we use u to stand for $u(\mathbf{x})$. On the other hand, real images contain cropping misalignments which also introduce pixel displacements. Let $\mathbf{h}$ denote one component of the misalignment which is orthogonal to $\mathbf{u}$, while the other component is added to $\mathbf{u}$. Note that this cropping misalignment is unique for all pixels, therefore $\mathbf{h} = [-h\sin\theta, h\cos\theta]^{\mathrm{T}}$. Then we minimize the optical flow function following Brox *et al*. [3]:

$$E(u,h) = E_{Data}(u,h) + \alpha E_{Smooth}(u) \tag{4}$$

where $E_{Data}(u,h)$ ensures the synthesis accuracy and $E_{Smooth}(u)$ controls the smoothness of u. Let $\{\boldsymbol{I}^c\}$ and $\{\boldsymbol{J}^c\}$ be the different channels of image $\boldsymbol{I}$ and $\boldsymbol{J}$, then

$$\begin{aligned} E_{Data}(u,h) &= \int_{\Omega}\sum_{c}\Psi([\boldsymbol{J}^c(\mathbf{x}+\mathbf{u}+\mathbf{h}) - \boldsymbol{I}^c(\mathbf{x})]^2)\mathbf{dx} \\ E_{Smooth}(u) &= \int_{\Omega}\Psi([\nabla u]^2)\mathbf{dx} \end{aligned} \tag{5}$$

where Ω indicates the image domain containing all pixels, and $\Psi(s^2) = \sqrt{s^2+\varepsilon^2}$ ($\varepsilon = 0.001$) is a robust function approximating an l^1-minimization of s. This minimization problem can be solved by an iterative method similar to that proposed by Brox *et al*. [3].

2.3. Training image synthesis for gaze sensing

This section shows how to synthesize training images $\{\boldsymbol{I}_{\mathbf{\Gamma},m}\}$ for any unseen head pose $\mathbf{\Gamma}$ based on the original $\{\boldsymbol{I}_{\mathbf{\Gamma}_0,m}\}$ and the corresponding pixel displacement flow $\mathbf{u}_{\mathbf{\Gamma}}$. Let vector $\mathbf{x}$ be the image pixel positions, then

$$\boldsymbol{I}_{\mathbf{\Gamma},m}(\mathbf{x}+\mathbf{u}_{\mathbf{\Gamma}}) = \boldsymbol{I}_{\mathbf{\Gamma}_0,m}(\mathbf{x}) \tag{6}$$

meaning that $\boldsymbol{I}_{\mathbf{\Gamma},m}$ is warped by moving the pixels of $\boldsymbol{I}_{\mathbf{\Gamma}_0,m}$ from $\mathbf{x}$ to $\mathbf{x}+\mathbf{u}_{\mathbf{\Gamma}}$. The basic assumption to perform such synthesis is that the pixel displacement flow $\mathbf{u}_{\mathbf{\Gamma}}$ shall not be affected much by gaze variation (eyeball rotation), meaning that for any m, $\boldsymbol{I}_{\mathbf{\Gamma},m}$ can be synthesized from $\boldsymbol{I}_{\mathbf{\Gamma}_0,m}$ using the same $\mathbf{u}_{\mathbf{\Gamma}}$.

Synthesizing a set of training images for each input image with an arbitrary head pose is prohibitively expensive. Therefore, training images $\{\boldsymbol{I}_{\mathbf{\Gamma}_n,m}\}$ are synthesized for a set of selected anchor head poses $\{\mathbf{\Gamma}_n\}$ in advance. In estimation, for an unseen head pose $\mathbf{\Gamma}$ we can directly acquire the pre-stored $\{\boldsymbol{I}_{\mathbf{\Gamma}_n,m}\}$ where $\mathbf{\Gamma}_n$ is closest to $\mathbf{\Gamma}$. It is enough to pre-synthesize training images for less than a hundred anchor head poses to achieve high accuracy and avoid online synthesis.

3. Experimental evaluation

Evaluation is presented in this section. We first examined our proposed eye appearance synthesis, then evaluated how accurate the gaze estimation can be achieved using the synthetic training images.

Our system was built on a desktop PC with a 22-inch LCD monitor and a VGA webcam. The users sat in front of the monitor (about 60 cm away) and let the camera capture their appearances and track their head poses by a vision-based head pose tracker [1].

The training process consisted of two steps. First, the users tried to keep a fixed head pose (without using chinrest) and focused their gaze on each of the 33 training points shown on the screen in turn to capture the original training images. Second, they moved their heads vertically/horizontally to capture four reference images. During this period they had to gaze at a moving point on the screen shown in accordance with their head orientation to avoid eyeball rotation. In the test process, the users were allowed to move their head freely and gaze at any position on the screen, and collected test samples by mouse click on the gaze positions. Typical head motion range in the experiments was 25° in rotation or 30 cm in translation approximately.

3.1. Eye image synthesis

We first estimate reference 1D pixel flows for four reference eye images under different head poses. Figure 2 gives examples of the estimated 1D flows and the synthetic eye images with ground truths. The well-synthesized images show the efficacy of the 1D flows.

With the reference flows, new flows can be produced to synthesize the eye images for unseen head poses. Examples of synthetic eye images for unseen head poses are shown in Figure 3 with ground truths. The synthetic

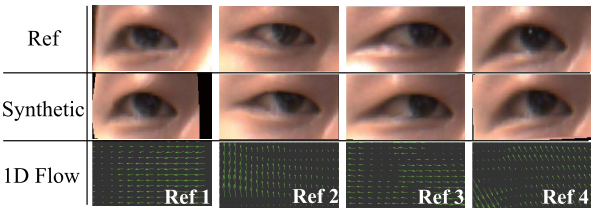

Figure 2. Four reference images: ground truths, synthetic results and 1D flows.

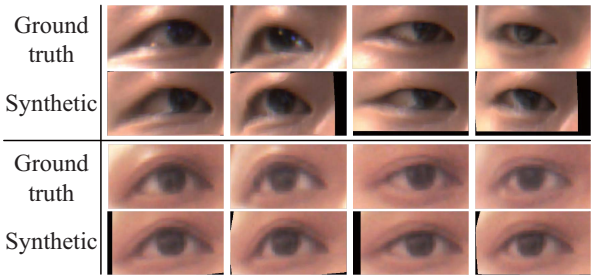

Figure 3. Representative results of eye image synthesis for unseen head poses.

images are obviously satisfactory, therefore we can synthesize training images for gaze estimation efficiently.

3.2. Gaze estimation results under head motion

Gaze estimation using the synthetic training images was assessed. Experiments were done for five subjects. For each subject, we collected more than one hundred samples combining head motion and gaze variation. For comparison, results of the recent method by Lu *et al*. [5] are also presented in Table 1(top) based on the same dataset. Our method clearly achieved higher accuracy when only four additional training images were used to allow head motion. Moreover, comparisons with existing head pose-free appearance-based methods with respect to their reported accuracies are shown in Table 1(bottom). Our method outperforms others since it requires much less training effort to handle head motion with the highest accuracy. Overall, the average accuracy of 2.24° is quite acceptable for common applications.

4. Conclusion

This paper proposes a novel method to allow head motion in appearance-based gaze estimation via eye image synthesis. By capturing only four additional images besides the original training images, our method synthesizes new training images for any unseen head poses to estimate gaze direction with high accuracy. We believe that the proposed technique can be also useful in other applications. Our future work may include increasing the robustness of the method by considering effects such as lighting changes and highlights.

Subject	Proposed	Lu *et al*. [5]	Training samples
S1	2.25°	2.93°	33 under default head pose and 4 under reference head poses
S2	2.22°	2.29°	
S3	2.07°	2.68°	
S4	2.36°	3.81°	
S5	2.30°	2.59°	
Average	**2.24°**	2.86°	

Method	Estimation error	Training samples
Proposed	**2.24°**	33+4
Lu *et al*. [5]	2.38°	33+video($\approx 10^2$)
Sugano *et al*. [9]	4° ~ 5°	$\approx 10^3$

Table 1. Comparison of estimation errors.

References

[1] faceAPI. http://www.seeingmachines.com.

[2] S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. In *NIPS*, pages 753–760, 1994.

[3] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, pages 25–36, 2004.

[4] D. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Trans. on PAMI*, 32(3):478–500, 2010.

[5] F. Lu, T. Okabe, Y. Sugano, and Y. Sato. A head pose-free approach for appearance-based gaze estimation. In *BMVC*, 2011.

[6] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Inferring human gaze from appearance via adaptive linear regression. In *ICCV*, 2011.

[7] H. Shum, S. Kang, and S. Chan. Survey of image-based representations and compression techniques. *IEEE Trans. on CSVT*, 13(11):1020–1037, 2003.

[8] Y. Sugano, Y. Matsushita, and Y. Sato. Calibration-free gaze sensing using saliency maps. In *CVPR*, pages 2667–2674, 2010.

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

[10] K. Tan, D. Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. In *WACV*, pages 191–195, 2002.

[11] R. Valenti and T. Gevers. Accurate eye center location and tracking using isophote curvature. In *CVPR*, pages 1–8, 2008.

[12] J. Wang, E. Sung, and R. Venkateswarlu. Eye gaze estimation from a single image of one eye. In *ICCV*, pages 136–143, 2003.

[13] O. Williams, A. Blake, and R. Cipolla. Sparse and semi-supervised visual mapping with the S^3GP. In *CVPR*, pages 230–237, 2006.