I see what you see: Point of Gaze Estimation from Corneal Images

Christian Nitschke*, Atsushi Nakazawa† and Toyoaki Nishida‡
*Graduate School of Informatics, Kyoto University, Yoshida-Honn machi, Sakyo-ku, Kyoto, 606–8501 Japan
Email: {christian.nitschke, nakazawa.atsushi, nishida}@i.kyoto-u.ac.jp, Web: http://ii.iit.i.kyoto-u.ac.jp/~nitschke/
†PRESTO, Japan Science and Technology Agency (JST), 4–1–8 Honcho, Kawaguchi, Saitama, 332–0012 Japan

Abstract—Eye-gaze tracking (EGT) is an important problem with a long history and various applications. However, state-of-the-art geometric vision-based techniques still suffer from major limitations, especially (1) the requirement for calibration of a static relationship between eye camera and scene, and (2) a parallax error that occurs when the depth of the scene varies. This paper introduces a novel concept for EGT that overcomes these limitations using corneal imaging. Based on the observation that the cornea reflects the surrounding scene over a wide field of view, it is shown how to extract that information and determine the point of gaze (PoG) directly in an eye image. To realize this, a closed-form solution is developed to obtain the gaze-reflection point (GRP), where light from the PoG reflects at the corneal surface into a camera. This includes compensation for the individual offset between optical and visual axis. Quantitative and qualitative evaluation shows that the strategy achieves considerable accuracy and successfully supports depth-varying environments. The novel approach provides important practical advantages, including reduced intrusiveness and complexity, and support for flexible dynamic setups, non-planar scenes and outdoor application.

I. INTRODUCTION

Eye-gaze tracking (EGT) refers to the problem of detecting and tracking a person’s eye movements, gaze direction and point of gaze (PoG) in a scene, which has applications in various fields, such as visual systems, psychology, medicine, life sciences, engineering and marketing [1]. A considerable amount of research has been devoted to develop capable and usable EGT systems resulting in a large number of techniques for different scenarios [2]. These days, most popular and usable are geometric vision-based techniques that estimate the pose of a 3D eye model relative to a scene model or camera. The two major system configurations are mobile head-mounted systems [3], [4] and static remote systems [5], [6].

Problems of existing methods. In order to obtain the PoG in the scene, existing methods suffer from two major drawbacks: First, they require some form of system calibration to determine the relationship between eye camera and scene. In mobile head-mounted systems, where the relationship is dynamic, calibration determines the pose of the eye camera w.r.t. to a scene camera that points in the direction the person is facing. In remote systems, where the relationship is static, calibration determines the pose of the eye camera w.r.t. to either scene camera(s) or a 3D scene model. There exist mainly two approaches (TABLE I): (1) gaze-mapping calibration to establish an implicit mapping between gaze direction and PoG, and (2) geometric calibration to determine the explicit relation between eye camera and scene. In any case, calibration requires expert knowledge, effort and specific hardware. Moreover, this imposes a static geometry, which largely reduces flexibility and requires re-calibration when the configuration is changed. Second, due to the fundamental concepts in calculating the PoG, systems potentially suffer from a parallax issue (Fig. 1, top). Although, head-mounted systems are designed for mobile use, the common gaze-mapping calibration assumes a planar surface at a certain depth, which invalidates the mapping when the user moves away. To reduce parallax effects, remote systems assume planar scenes, provide manual parallax compensation [4], or use multiple scene cameras [5], [6], however, at the expense of a complex setup with additional pose calibration. In any other case, parallax errors occur under depth-varying conditions.

Proposed approach. This paper introduces a novel concept
for PoG estimation that overcomes the described limitations using corneal imaging techniques, based on the observation that the cornea reflects the light from the surrounding scene [7], [8]. Nishino and Nayar [7] first formalize the cornea–camera catadioptric image formation and apply the extracted environment map for the computation of scene panoramas and face relighting. Other applications include recovering a point spread function for deblurring corneal images [9], identifying composite images from inconsistent corneal reflections [10], and reconstructing a planar computer screen from corneal reflections of a moving person [11], [12]. The proposed approach estimates the PoG directly in an eye image (Fig. 1, bottom/right). Therefore, we introduce the concept and a closed-form solution to calculate the gaze-reflection point (GRP), where light from the PoG reflects at the corneal surface into an eye-observing camera. This comprises a solution for the forward projection problem with a spherical mirror, where the scene point is located at unknown distance, not covered by the recent method of Agrawal et al. [13]. A simple one-point calibration further increases accuracy to the level of state-of-the-art EGT by compensating for the individual offset between optical and visual axis. The approach combines four major advantages: (1) It does not require a scene camera or model. (2) It does not require an eye camera–scene calibration and, thus, naturally (3) supports flexible dynamic systems and (4) non-planar complex scenes.

**Related approaches.** Nishino and Nayar [7] determine an approximate PoG using a retinal projection method that depends on further eyeball parameters, does not include a visual axis offset, and is, thus, inherently less accurate. Also, there is no investigation conducted for the gaze-estimation task. In [14] we describe an active system to illuminate a scene by coded IR-light that allows robust matching of eye and scene images at the PoG. However, this involves complex hardware and restricts to controlled environments. Having equal geometric foundation, the current paper describes a more general PoG estimation framework, and additionally provides an overview on eye tracking and pose estimation. The absence of illumination supports any eye image, allows texture extraction from scene images, and enables outdoor application under high brightness. The reduced hardware allows compact and non-intrusive setups. This is ideal for wearable cameras and augmented reality head-mounted displays (AR-HMDs) like Google Glass [15].

**II. GEOMETRIC EYE MODEL**

Figure 2(a) shows a cross-section of the human eye and its geometric model. The eyeball consists of two main segments: the anterior and the posterior, which can be approximated by two overlapping spheres. The corneal limbus is the surface shape discontinuity at the intersection between corneal and eyeball sphere, where the cornea dissolves into the sclera. It coincides with the iris contour and has a radius \( r_L \) of approximately 5.6 mm [16]. The cornea is modeled as a spherical cap that is cut off from the corneal sphere by the limbus plane. We use this model to estimate the 3D pose of the eye relative to the camera and the location of the GRP in the image.

**III. EYE POSE ESTIMATION**

As there exists a large number of methods, we want to provide an overview and discussion of different strategies [8]. Eye pose estimation requires two tasks: *image processing* to track the eye and its features in an image [2], and *geometric modeling* to estimate the pose of a 3D eye model. We distinguish passive methods that work on any image and active-light methods that require additional controlled illumination.

**A. Active-light methods**

Active light methods are developed for robust automatic EGT and require a complex hardware setup. The pupil-center–corneal-reflection (PCCR) technique is largely covered in....

---

**TABLE I**

**ISSUES IN EYE CAMERA–SCENE CALIBRATION.**

<table>
<thead>
<tr>
<th>Method</th>
<th>Gaze-mapping calibration</th>
<th>Geometric calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaze markers</td>
<td>Gaze marker (attached or displayed) + interaction (tedious)</td>
<td>Interaction (difficult + tedious)</td>
</tr>
<tr>
<td>Calibration hardware (object + mirror)</td>
<td>n/a</td>
<td>Interaction (difficult) + low accuracy</td>
</tr>
<tr>
<td>Manual setting</td>
<td>n/a</td>
<td>Inflexible rigid setup</td>
</tr>
<tr>
<td>Factory setting</td>
<td>n/a</td>
<td></td>
</tr>
</tbody>
</table>

**Scene representation**

<table>
<thead>
<tr>
<th>Camera</th>
<th>Model (not for mobile systems)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–N</td>
<td>Parallax error when deviating from calibration plane</td>
</tr>
<tr>
<td></td>
<td>Requires 3D scene model + commonly assumed planar (with parallax error when deviating from plane)</td>
</tr>
</tbody>
</table>

Fig. 2. (a) Cross section of the human eye and a corresponding geometric eye model. (b) Image projection of the eye.
Although, there exists individual
\( \tau = (0 \cdot n \cdot \sin \theta) / (\cos \theta \cdot \sin \theta) \)

and normal vector \( \mathbf{n} \) as in (2)
\( \mathbf{C} = (\cos \tau, \sin \tau)^T \), \( \mathbf{n} = (\cos \theta, \sin \theta)^T \).

We then solve for the corneal angle \( \theta \) that defines the GRP, using the known eye gaze angle \( \tau \) as in
\[ \theta = \arctan((1 - \sin \tau) / \cos \tau) = \tau / 2. \]
The distance between the limbus center and the GRP in the local orthogonal image plane is defined as
\[
|\mathbf{I}_S - \mathbf{I}_L| = r_C \sin \theta - d_{\mathbf{LC}} \sin \tau.
\]
Assuming the average depth plane of the weak perspective projection to be located at limbus center \( \mathbf{L} \), with scale-factor \( s = \tau_{\text{max}} / r_L \), and center and semi-minor axis of the limbus ellipse, \( \mathbf{I}_L \) and \( \mathbf{v}_{sm} \), the location of the GRP \( \mathbf{I}_S \) is given as in
\[
\mathbf{I}_S = \mathbf{I}_L + s \cdot \mathbf{v}_{sm} |\mathbf{I}_S - \mathbf{I}_L|.
\]

B. Calibration and application of individual parameters

Many studies report that the optical axis of the human eye differs by an individual offset from the visual axis, which is the true gaze direction [2]. This difference can be described by static offset angles in a facial coordinate system, where \( x_{\text{face}}\) is the facial frontal direction, \( z_{\text{face}} \) is the direction from left to right eye, and \( y_{\text{face}} \) is orthogonal to \( x_{\text{face}} \) and \( z_{\text{face}} \). The offset angles around \( x_{\text{face}} \) (tilt) and \( y_{\text{face}} \) (pan) are commonly within 1.5 \( \sim \) 3 and 4 \( \sim \) 5 degrees. However, a one-time personal calibration is needed to obtain accurate values for a particular user [34].

We introduce a one-point calibration technique to obtain these parameters through the analysis of a corneal reflection image. In this process, a user gazes a calibration point in the scene. We calculate the GRP and manually find the reflection of the calibration point (CRP), in the corneal image. We then introduce a coordinate system \( \mathbf{R}_e = [x_e, y_e, z_e] \) at the limbus center, where \( x_e \) and \( y_e \) correspond to the 3D vectors for the semi-major and -minor axes of the projected limbus. Figure 4(2) illustrates the cross section at the semi-minor axis in the \( y_ez_e \)-plane. In this plane, the angle \( \gamma \) between the optical axis and the CRP is described as in
\[
\gamma = \tau - \arcsin(d_{\mathbf{LC}} \cdot \cos \tau + |\mathbf{I}_L - \mathbf{I}_T| / r_C),
\]
and the vector \( \mathbf{v}_{sm} \) is
\[
|\mathbf{I}_L - \mathbf{I}_T| = \mathbf{v}_{sm} \cdot (\mathbf{i}_T - \mathbf{i}_L) / s,
\]
where \( \mathbf{I}_T \) and \( \mathbf{i}_T \) are the locations of the CRP and its projection in the corneal image. Considering a surface reflection at the CRP, then the offset angle \( \delta x_e \) between the optical and the visual axes in this plane is defined as in
\[
2 \cdot (\gamma + \delta x_e) = \tau + \delta x_e, \quad \delta x_e = \tau - 2 \cdot \gamma.
\]
Similarly, the angle \( \delta y_e \) in the \( x_ez_e \)-plane is defined as in
\[
\delta y_e = 2 \cdot \arcsin(\mathbf{v}_{im} \cdot (\mathbf{I}_T - \mathbf{i}_L)) / s \cdot r_C.
\]
where \( \mathbf{v}_{im} \) is the semi-major axis of the limbus ellipse.

When the camera locates frontal to the face, we can assume the camera \( x \) axis to be parallel to \( x_{\text{face}} \) and transform the offset angle into the camera coordinate system as in
\[
\delta x = \arctan\left(\frac{e^T \mathbf{R}_e \mathbf{u}}{e^T \mathbf{R}_e \mathbf{u}}\right), \quad \delta y = \arctan\left(\frac{e^T \mathbf{R}_e \mathbf{u}}{e^T \mathbf{R}_e \mathbf{u}}\right),
\]
where \( \mathbf{u} = (\sin \delta y \sin \delta x, \sqrt{1 - \sin^2 \delta x - \sin^2 \delta y}) \), and \( e_{x} = (1, 0, 0)^T \), \( e_{y} = (0, 1, 0)^T \), \( e_{z} = (0, 0, 1)^T \). Once the transformed offset angles are obtained, the corrected GRP (GRP*) can be estimated using the corrected eye coordinate system \( \mathbf{R}_e^* = \mathbf{R}_e^*(\delta x) \mathbf{R}_e^*(\delta y) \mathbf{R}_e \) that is rotated by \( (\delta x, \delta y) \) from the original eye coordinate system \( \mathbf{R}_e \), where \( \mathbf{R}_e \) and \( \mathbf{R}_e^* \) are the rotation matrices around the \( x \) and \( y \) axes. Using \( \mathbf{R}_e^* \), we obtain the corrected angle \( \tau^* \) and semi-minor axis of the projected limbus \( \mathbf{v}_{sm}^* \) as in
\[
\tau^* = \arccos(e^T \mathbf{R}_e^* e_x), \quad \mathbf{v}_{sm}^* = \left(\frac{e_x^T}{e_y^T}ight) \mathbf{R}_e^* (\mathbf{R}_e^*)^T \mathbf{R}_e^* e_x.
\]
Applying \( \tau^* \), \( \mathbf{v}_{sm}^* \), and equations (3) and (4), we obtain the position of the GRP* that reflects the light from the true visual axis of the eye.

V. EXPERIMENTS

A. Quantitative evaluation

For comparison, we implemented a common mapping-based PCCR method, using an active-light dark pupil technique [14].

Setup. Evaluation is performed with five subjects in an indoor environment under normal lighting. To evaluate the effects of depth variation, the subjects are seated at 1900 and 3000 mm in front of a wall with 20 ground-truth (GT) gaze markers attached (Fig. 5). Gaze-mapping calibration for PCCR is obtained at 1900 mm, from four eye images, where a subject gazes the outer corner markers. Detecting pupil center and glint from an off-axis IR-LED allows to calculate an affine matrix that maps the pupil center–glint vector to the corresponding PoG in scene coordinates. The subjects are asked to gaze each marker, where for the right eye, we calculate the angular error between GT marker direction and estimated visual axis, and the distance error between GT marker and PoG location on the wall. For the proposed method, the distance error is obtained from the angular error in the spherical panorama of
The backprojected corneal image. Vice versa, for the PCCR method, the angular error is obtained from the distance error of the estimated PoG using the calibrated mapping.

**Results and discussion.** TABLE II shows the mean angular (visual axis) errors. The performance of PCCR decreases when the depth changes (condition 2) from the one at which calibration was performed (condition 1), due to an increasing parallax error for the calibrated mapping. On the contrary, the proposed approach relies on an uncalibrated direct calculation in the corneal image, which performs accurately and stably, with an error of less than 1 deg under varying depth conditions\(^2\).

Figure 6 shows PoG estimation results and distance errors at conditions 1 and 2, where two effects can be observed: (1) The error increases under condition 2, as it scales with distance between eye and scene. (2) The error increases with gaze angle from the central direction (900, 600, z), due to an increasing scene distance (as in (1)) and errors in eye pose estimation. The latter relate to skin occlusion by eye corners (horizontal) and eye lids (vertical), which is larger towards the bottom.

**B. Qualitative evaluation**

To measure the performance in more realistic and complex scenarios, we perform a qualitative evaluation for several indoor and outdoor scenes with different subjects. Figure 7 shows scenes and results. Although, the cornea is not a perfect mirror and suffers from several issues, such as low reflectivity, contamination with iris texture, occlusion from eyelid, eyelashes and nose, and distortions from an aspherical shape, these effects decrease towards the apex. Thus, the re-projected images from the central area, including the GRP, show rich scene details (Fig. 7(e)). The method works well under challenging bright outdoor lighting conditions. In fact, this even increases the quality of iris contour tracking and obtained corneal images.

**VI. CONCLUSION**

The paper introduced the first EGT approach that estimates the PoG w.r.t. the reflected scene in an eye image. The two main contributions comprise (1) the concept of direct PoG estimation, and (2) a closed-form solution for the GRP, including compensation for the individual offset between optical
and visual axis. The approach overcomes two major limitations of existing methods, namely, (1) a calibration of the relationship between eye camera and scene, and (2) a parallax error that occurs under varying scene depth. Practical advantages include reduced intrusiveness and complexity, and the support for flexible dynamic setups, non-planar and bright outdoor scenes. This allows for various applications, including (1) EGT with infants, elderly and disabled, (2) wearable EGT with ubiquitous cameras and AR-HMDs [15], and (3) remote EGT, surveillance and forensics. To further increase image quality, one may (1) improve the corneal image through contrast enhancement, super-resolution [36], eye lash and iris removal [37], or (2) map the PoG into a high quality scene image, using passive correspondence matching between corneal and scene image.

ACKNOWLEDGMENT

This work was supported by the JST Precursory Research for Embryonic Science and Technology (PRESTO) program, Japan.

REFERENCES