Towards a low cost and high speed mobile eye tracker

Frank H. Borsato
Federal University of Technology
Paraná, Brazil
frankhelbert@utfpr.edu.br

Carlos H. Morimoto
University of São Paulo
São Paulo, Brazil
hitoshi@ime.usp.br

ABSTRACT

Despite recent developments in eye tracking technology, mobile eye trackers (ET) are still expensive devices limited to a few hundred samples per second. High speed ETs (closer to 1 KHz) can provide improved flexibility for data filtering and more reliable event detection. To address these challenges, we present the Stroboscopic Catadioptric Eye Tracking (SCET) system, a novel approach for mobile ET based on rolling shutter cameras and stroboscopic structured infrared lighting. SCET proposes a geometric model where the cornea acts as a spherical mirror in a catadioptric system, changing the projection as it moves. Calibration methods for the geometry of the system and for the gaze estimation are presented. Instead of tracking common eye features, such as the pupil center, we track multiple glints on the cornea. By carefully adjusting the camera exposure and the lighting period, we show how one image frame can be divided into several bands to increase the temporal resolution of the gaze estimates. We assess the model in a simulated environment and also describe a prototype implementation that demonstrates the feasibility of SCET, which we envision as a step further in the direction of a mobile, robust, affordable, and high-speed eye tracker.

CCS CONCEPTS

• **Applied computing** → **Imaging**; • **Computing methodologies** → *Tracking*; • **Human-centered computing** → Interaction techniques;

KEYWORDS

mobile eye-tracking, catadioptric system, stroboscopic lighting, rolling shutter

ACM Reference Format:
Frank H. Borsato and Carlos H. Morimoto. 2019. Towards a low cost and high speed mobile eye tracker. In *2019 Symposium on Eye Tracking Research and Applications (ETRA '19), June 25–28, 2019, Denver, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3314111.3319841

1 INTRODUCTION

Gaze interaction has gained increased attention as a result from the latest technological developments in mobile and pervasive attention-aware systems and interfaces [Majaranta and Bulling 2014]. Recent advancements in mobile eye-tracking technology allow us to investigate eye movements during natural behavior [Bulling and Gellersen 2010; Eivazi et al. 2018]. The development of mobile eye trackers (ET) is still an active research topic that faces several challenges however. Commercial systems are still expensive for general use (more than USD $ 1 K) and even more for high speed ETs, that might help in designing improved gaze-based interfaces by offering better data filtering and more reliable event detection. Due to the easy setup and good accuracy, most mobile ETs are based on video and employ feature-based gaze estimation methods [Fuhl et al. 2018; Kassner et al. 2014; Tobii AB 2018].

Feature-based gaze estimation methods exploit local features such as contours, eye corners, and reflections from the eye image and can be broadly divided into two categories, interpolation-based methods, which map image features to gaze coordinates, and model-based (geometric) methods [Hansen and Ji 2010]. Geometric models of gaze estimation typically rely on metric information and thus require camera calibration and a geometric model of the eye, camera and light sources [Dierkes et al. 2018; Morimoto et al. 2002; Newman et al. 2000; Wang et al. 2005].

An alternative that falls into this last category is to model the eye-camera geometry as a catadioptric imaging system [Nitschke et al. 2013]. Catadioptric systems combine mirrors with cameras and can either have a single viewpoint (like a perspective camera) [Baker and Nayar 1999] or multiple viewpoints, referred to as non-central catadioptric systems. In non-central systems, the optical rays coming from the camera and reflected by the mirror surface do not intersect into a unique point [Swaminathan et al. 2006]. For the purpose of this work, we assume the cornea to be spherical and, thus, a non-central system.

As feature-based gaze estimation methods use images of the eye, their temporal resolution is generally limited by the camera being used. The use of low-cost cameras, typically with a low frame-rate, lead to sampling-related errors, which in turn affect the detection of fixations and saccades [Andersson et al. 2010]. Additionally, in low-cost cameras each frame line is exposed a little shifted in time by a technique known as rolling shutter [Grundmann et al. 2012]. Its overlapping behavior and time delay between each row exposure may introduce spatial distortions on moving objects, such as the eye [QImaging 2014]. To improve the image quality and allow frame synchronization, stroboscopic lighting can be successfully employed [Borsato et al. 2015; Borsato and Morimoto 2017, 2018].

In this paper we propose a novel approach for mobile ET that exploits the reflective properties of the cornea, the rolling shutter technology, and stroboscopic structured infrared lighting. Particularly, we devise a geometric model where the cornea acts as a spherical mirror in a catadioptric system, changing the projection as it moves. Calibration methods for the geometry of the system

ETRA '19, June 25–28, 2019, Denver, CO, USA

ACM ISBN 978-1-4503-6709-7/19/06…$15.00
https://doi.org/10.1145/3314111.3319841

and for the gaze estimation are presented. By carefully adjusting the camera exposure and the lighting period, we show that multiple samples can be captured within a frame period. Instead of tracking common eye landmarks, such as the pupil, we detect the reflection from multiple point light sources arranged close to the eyes.

The contributions of this work are four-fold: 1. We devise a geometric model where the eye acts as a mirror in a catadioptric system. We disclose how to calibrate the model and how to estimate the gaze thereafter; 2. We present how stroboscopic lighting can be used in non-synchronized setups to take multiple partial snapshots within a frame; 3. We assess the model with synthetic data, including simulated noise and translation of the apparatus with respect to the eye/face. 4. We also propose a prototype built as a proof-of-concept based on a Raspberry Pi platform, an Arduino board, and a small full-hd camera, all assembled using 3D-printed parts. The prototype can take high resolution images of the cornea illuminated by stroboscopic light.

The rest of the paper is organized as follows. The next section describes the proposed system, including the calibration methods and capture model under stroboscopic lighting. Section 3 presents simulated results, followed by Section 4, that describes our current prototype that we use to study the banding effects. Section 5 presents a brief discussion of our findings and Section 6 concludes the paper.

2 STROBOSCOPIC CATADIOPTRIC EYE-TRACKING (SCET)

The proposed stroboscopic catadioptric eye-tracking system exploits the combination of widely used rolling shutter cameras and a high-speed stroboscopic lighting to take multiple snapshots of the corneal surface during a frame interval. This virtually increases the temporal resolution of the system, as one camera frame can provide information from multiple time instants. Instead of detecting eye landmarks, such as the pupil center or iris contour, we track specular reflections on the cornea generated by the lighting.

As part of our method, we manipulate the camera exposure such that, for each time instant, only a limited range of the frame lines gets illuminated (inside what we call a band), which contributes to the robustness and computational efficiency. In order to have information from the cornea at any frame band, we assume the illumination is provided by an extended source or by multiple point sources at known locations. Such illumination must be structured to produce reflections all over the corneal surface.

The cornea and the camera can be considered as a catadioptric sensor system [Nitschke et al. 2013], which requires the geometric setup, projection model, and calibration process to be fully defined [Barone et al. 2018]. The projection model of a catadioptric system like ours can be described by the forward projection (FP), which states where a 3D scene point will be projected in 2D pixel coordinates. In our system, the FP allow us to estimate the position of reflected light sources in the image plane, given a particular position and orientation of the cornea. To solve the FP problem, we exploit the analytical closed-form solution presented in [Barone et al. 2018].

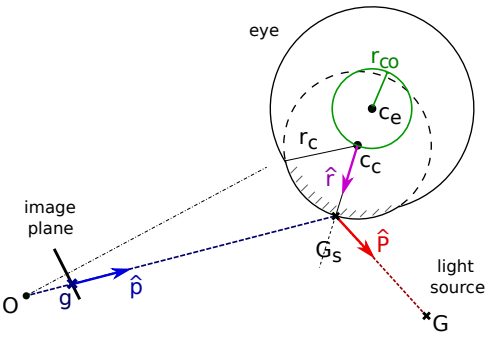

Figure 1: Catadioptric eye-tracking geometric model. Planar view defined by the camera center O, light source G, and corneal center c_c. (For simplicity, the eye center c_e lies on this plane as well). The corneal surface acts as a moving spherical mirror in the model.

In the next section, we describe the proposed geometric model and present how the gaze direction can be estimated from a calibrated system. In Section 2.3 we thoroughly explain the rolling shutter imaging model under multiple strobes of light. Lastly, in Section 2.4, we describe the system geometric calibration in detail.

2.1 Geometric eye-tracking model

Our eye-tracking geometric model is composed of three main components: a camera pointing to the eye, the eye, and the illumination source, as shown in Figure 1. The camera center O is at the origin of the coordinate system. The light source G is reflected at the corneal surface at point G_s and projected at the camera image plane at point g. For simplicity, the eyeball is represented as a sphere rotating around its center (c_e) and the cornea as a spherical cap centered at c_c, as in [Böhme et al. 2008].

The cornea is off-centered with respect to the eye so that a rotation of the eyeball induces both rotation and translation of the cornea. We assume the transformation that takes the gaze from one direction to another can be expressed by a rigid roto-translation of the cornea that can be measured by the translation induced in the image plane (the translation of point g in the 2D image plane) of known point light sources (represented by G in our model) when reflected by the corneal surface. As we assume the cornea as a section of a sphere, we can further simplify the corneal motion model to keep just the translational component.

If we assume the camera to follow a pinhole model and know both the 3D coordinates of point G and the 2D coordinates of point g in the image plane, the unknowns become the corneal center c_c and radius r_c. This problem can be considered as a forward projection, in which neither direction $\hat{P}$ nor $\hat{p}$ are known. The method presented in [Barone et al. 2018] computes the forward projection solution in an analytical closed-form, the problem is reduced to a 4th-order polynomial by determining the reflection point on the mirror surface (cornea) through the intersection between a sphere and an ellipse, as described in [Glaeser 1999].

The calibration of this geometric model, presented in detail in Section 2.4, involves estimating the relative position of G, the center of the eye c_e and the radius of the orbit followed by the cornea r_{co} (in green in Figure 1). The light source G represents any arbitrary

number of point sources. This number does not change the complexity of the problem assuming all the lights are part of the same rigid structure, as only position and rotation of such structure is needed.

2.2 Estimating the gaze direction

Gaze estimation can be considered as an optimization process which adjusts virtual corneal parameters in order to minimize the distance between estimated light reflections with the ones detected in the image. It can also be described by a mapping between 3D corneal center coordinates and gaze angles (or targets at a fixed distance). This second approach simplifies the calibration of the system, as we do not need to make assumptions about the sphericity of the orbit followed by the cornea as the eye gazes. Both methods however, require the relative position and orientation of the lighting.

2.2.1 Gaze as corneal rotation - geometric gaze estimation. The re-projection of the lighting on the camera image plane is univocally determined by the forward projection model, given the center and radius of the cornea. Assuming the eyeball center c_e and corneal orbit radius r_{co} are known, an optimization process can then be defined, with the optimization parameters represented by the actual cornea center in spherical coordinates (two parameters, with origin at c_e). The target function to be minimized is represented by the sum of the squared distances between the re-projected point light sources and the detected ones.

The initial estimation of the corneal center is obtained by assuming the eye at the primary position. Subsequent computations use the previous estimated position as initialization. A non-linear least square optimization procedure is used, exploiting the Trust-Region-Reflective algorithm [Moré and Sorensen 1983], which allows to set upper and lower bounds for the parameters.

2.2.2 Gaze as corneal position - polynomial gaze estimation. Another way to estimate the gaze is by mapping measurements of eye features, such as the pupil or the pupil-glint vector, to gaze using some function. Both linear and second-order polynomials are commonly employed [Cerrolaza et al. 2008; Cherif et al. 2002; Morimoto et al. 2000; Rosengren et al. 2018].

While we compute measurements in the image plane, the forward projection allows us to estimate the corneal position in space. The polynomial used to map between corneal center and screen coordinates is defined as:

$$\begin{aligned} s_x &= a_0 + a_1x + a_2y + a_3z + a_4xy + a_5xz + a_6yz + a_7xyz, \\ s_y &= b_0 + b_1x + b_2y + b_3z + b_4xy + b_5xz + b_6yz + b_7xyz, \end{aligned} \tag{1}$$

where (s_x, s_y) are screen coordinates and (x, y, z) is the corneal center. Each polynomial have 8 unknowns that can be computed using a 9-point calibration procedure. We also assessed a second-order polynomial using a 25-point calibration pattern.

2.3 Imaging model under stroboscopic lighting

Our imaging model assumes a rolling shutter camera that exposes each line slightly shifted in time [Borsato and Morimoto 2017; Bradley et al. 2009]. Here we exploit this property to obtain exposures from different time instants within one frame period, therefore increasing the number of gaze estimate samples. The stroboscopic

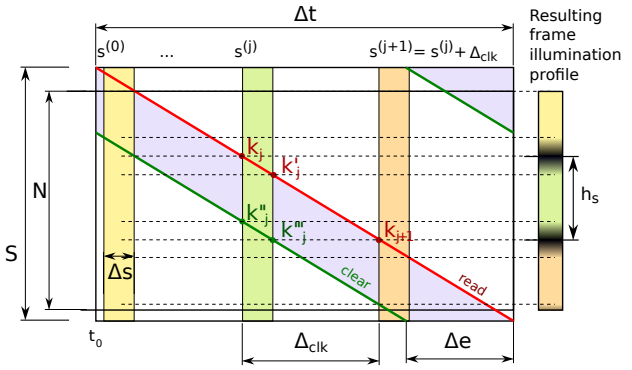

Figure 2: Rolling shutter imaging model using stroboscopic lighting. Multiple (3) pulses shown.

lighting creates a short, virtual exposure proportional to the duration of the pulse [Borsato and Morimoto 2017]. The general idea is to trigger the lighting several times during a frame period, creating the corresponding number of exposures slightly shifted in time. In this section, we discuss the conditions that must be met for the technique to work and what are the expected outcomes.

Figure 2 shows the rolling shutter imaging model using multiple strobes, where S is the total number of sensor scanlines and N is the effective number of transferred (or visible) lines. The difference between S and N defines the invisible scanline range. Δt defines the frame period and Δe the exposure. The lighting pulse duration is given by Δs and the period by Δ_{clk}. Note that in Figure 2, the frame is the result of several snapshots taken by stroboscopic light pulses (strobes) triggered at different instants, $s^{(0)}$, ..., $s^{(j)}$ and $s^{(j+1)}$. The symbols k_j and k_j''' represent the first and last scanlines lit by the strobe j, which defines what we call a band in the image, whose height is denoted by h_s.

The interval defined by the pair (k_j, k_j') and (k_j'', k_j''') corresponds to lines partially illuminated. We want to minimize them, as the lack of contrast might impair their use. The band position (k_j) in the frame depends on the moment when the strobe j is triggered $(s^{(j)})$ with respect to the start of the frame readout, given by t_0.

The banding characteristics depend on the readout and exposure times. Consider t_0 to be the readout time instant of the topmost scanline in a given frame, and let $R(y)$ to be the readout time of an arbitrary scanline y, that can be computed as (from [Bradley et al. 2009]):

$$R(y) = t_0 + \frac{y}{S} \cdot \Delta t, \tag{2}$$

then the time when the scanline y starts to be exposed can be estimated by (from [Bradley et al. 2009]):

$$E(y) = R(y) - \Delta e. \tag{3}$$

Let k_j be the first scanline subject to the light pulse j in a given frame, $R(k_j) = s^{(j)}$. Thus, using (2), we obtain

$$k_j = S \cdot \frac{s^{(j)} - t_0}{\Delta t}, \quad \text{with } s^{(j)} = s^{(0)} + j \cdot \Delta_{clk}, \tag{4}$$

where $s^{(j)}$ denotes the instant strobe j is triggered.

The number of partially illuminated lines on the strobe onset can be estimated by noting that $R(k_j) + \Delta s = R(k_j')$. Accordingly,

replacing $R(k_j)$ by its definition in (2), we obtain

$$k'_j - k_j = S \cdot \frac{\Delta s}{\Delta t}. \tag{5}$$

Note that the number of partially illuminated lines is proportional to the strobe duration Δs, as S and Δt are constant for a given resolution and frame rate [Borsato and Morimoto 2017, 2018].

Using the same reasoning, we can estimate $k''_j - k'_j$ by noting that $R(k'_j) - \Delta s = E(k''_j)$, and hence

$$k''_j - k'_j = S \cdot \frac{\Delta e - \Delta s}{\Delta t}, \tag{6}$$

where $k''_j - k'_j$ defines the range of usable lines in the frame lit by strobe j. The band height can be estimated as $h_s = (k''_j - k'_j) + 2 \cdot (k'_j - k_j)$, which results in

$$h_s = S \cdot \frac{\Delta e + \Delta s}{\Delta t}. \tag{7}$$

Lastly, we can estimate the lighting period Δ_{clk} by noting that $E(k'''_j) - \Delta s = R(k_{j+1}) - \Delta_{clk}$, and accordingly

$$\Delta_{clk} = \Delta e + \Delta s. \tag{8}$$

Note we can improve the utilization of the frame by enforcing $k_{j+1} = k''_j$, which eliminates the term Δs from both (7) and (8). This will eliminate the dark areas at the band boundaries (as seen in Figure 2), but an overlap proportional to Δs will persist.

The camera exposure plays an important role to the banding effect. Figure 2 was drawn to highlight its interaction with the strobe duration and period. Note in (7) that the band height depends on the camera exposure, and in (8) we define Δ_{clk} such that the end of a band coincides with the beginning of the next.

Note that the banding can be obtained in any camera, as synchronization can be achieved either by hardware or by software, as described in [Borsato and Morimoto 2017]. The number of lines for a given resolution (S) can be computed by software using the estimation technique described in [Borsato and Morimoto 2018].

2.4 Geometric calibration

In a traditional non-central catadioptric system, the calibration computes the mirror's radius and center with respect to the camera reference frame [Barone et al. 2018]. The major difference to our method is that in our setup the mirror moves with respect to the camera, as the eye moves. The movement however is constrained, as the cornea is tied to the eyeball. We assume that the eyeball has a spherical shape and rotates around its center. Therefore, besides the corneal radius (r_c), the calibration must solve for the eyeball center (c_e) and the orbital radius (r_{co}) followed by the cornea.

For the geometric calibration, a single pulse is used to capture sharp images of the eye, without banding. To estimate c_e and r_{co}, first observe that any two gaze directions $\overrightarrow{d_i}$ and $\overrightarrow{d_j}$ defines a triangle with vertices at c_e, and corneal centers $c_{c,i}$ and $c_{c,j}$. Assuming $c_{c,i}$ and $c_{c,j}$ are known, we can compute r_{co} applying the law-of-sines using the angle between $\overrightarrow{d_i}$ and $\overrightarrow{d_j}$. To estimate the eye center c_e we use least squares over several gaze directions to minimize the sum of the squared differences between a mean r_{co} and $\|c_{c,i} - c_e\|$, for every i. As a baseline method, we solve for both c_e and r_{co} using a least-squares sphere fitting as well [Ahn et al. 2001].

The problem of estimating the corneal center for a given gaze direction is similar to the calibration of a catadioptric system, which is traditionally performed by acquiring a planar chessboard reflected in the mirror from several different positions and orientations [Barone et al. 2018; Mei and Rives 2007; Scaramuzza et al. 2006].

We can consider the illuminators as a rigid group of point light sources, that can be assumed as the known calibration pattern. The corneal center estimation can then be computed by an optimization process on the following unknowns: c_c and r_c (with respect to camera center, four parameters), and rotation and translation of the lighting (three parameters for rotation and three for translation).

The target function to be minimized is represented by the sum of the squared distances between the re-projected point light sources and the detected ones. The forward projection assumes that the camera intrinsic parameters are known [Zhang 2000].

Again, we exploit the Trust-Region-Reflective algorithm [Moré and Sorensen 1983] in a non-linear least square optimization. The initial estimation, upper, and lower bounds for the corneal radius r_c are taken from [Read et al. 2006]. Whereas the expected position, orientation and bounds for the lighting are taken from our apparatus' 3D model considering the available degrees of freedom.

2.5 Robustness to camera and lighting translations

As shown in the Section 3.3, our method is very sensitive to translations of the camera and lighting with respect to the eye (geometric calibration drifts). To compensate such movements one could track features from the camera image. However, as we assume the exposure will be very short to enable banding, detecting eye features other than the reflections would not be reliable.

Instead, while in practice the translation of the camera along the z axis might differ from the one of the lighting, we assume they are equal. Remember that our model assumes that the camera and lighting are part of the same rigid structure. Thus, the relative orientation and position of the lights do not change with respect to the camera. Additionally, we assume that, to keep the camera and lighting in place, a glasses-like frame that sits on the nose is used.

The relative position of the camera-lighting with respect to the face is estimated as a function of the eye distance defined by the value d_a. The gaze calibration considers the estimated corneal center (c_c) and the relative distance (d_a) as independent variables with the subtended gaze position (s_x, s_y) as a dependent variable. A second-order, 15 term polynomial with four variables is evaluated as shown in Section 3.3.1.

3 EVALUATION USING SIMULATED DATA

In this section we assess both the geometric calibration of the system and the estimated gaze accuracy. Then, we assess the losses introduced by camera and lighting translations to the overall accuracy. Albeit the stroboscopic lighting was not directly evaluated in this section, we considered the accuracy taking only a subset of the frames in order to simulate the proposed banding.

The simulation of the geometrical properties of our model were performed using the framework presented in [Böhme et al. 2008]. We simulated a single camera remote ET setup using a monitor to

present stimuli. The monitor was set at 70 cm from the eye, with a screen size of 48 × 30 cm. Figure 3 summarizes the simulated settings.

System component	Coordinates
Screen	$x = -0.240 \ldots 0.240$ $y = -0.200 \ldots 0.100$ $z = -0.700$
Camera	$c = (0.012, -0.008, -0.030)$ Pointing to the eye apex
Eye	$c_e = (0, 0, 0)$, $r_c = 7.8 \cdot 10^{-3}$
Lights	$G = (-0.010, 0.020, -0.015)$

Figure 3: Simulated geometrical setup of system components (coordinates in meters). The coordinate system is right-handed, with the eye in the x-y-plane and the z-axis pointing against the screen.

The framework provides a parameter called feature position error (e_f), a random perturbation to each relevant point in the projected image to simulate the combined effects of finite image resolution, residual errors after camera calibration, and inaccuracies in the image analysis step [Böhme et al. 2008]. Unless otherwise stated, we use a feature position perturbed by an uniformly distributed error with a maximum magnitude of 0.5 pixel on both coordinate axes ($e_f = 0.5$). Note that the only features we use are the projected light reflections. The simulated sensor resolution is 640 × 480 pixels, with a camera focal length of 1300 pixels.

The lighting is composed of 40 discrete omni-directional point sources, arranged along a 90° circular section with radius 33 mm. The lights form a rigid body that is rotated 45° with respect to the horizontal plane x-z. In Figure 3, G denotes the first light source in the array, which is the pivot point for rotations. At the xz plane, the light is at 39.1° in the visual field, while the camera is at 21.8°. This setup is compatible with a glasses-like apparatus imaging the left eye with the camera close to the nose. Neither the lights nor the camera occlude the simulated monitor at the given distance. Figure 4 is a graphical representation of this setup.

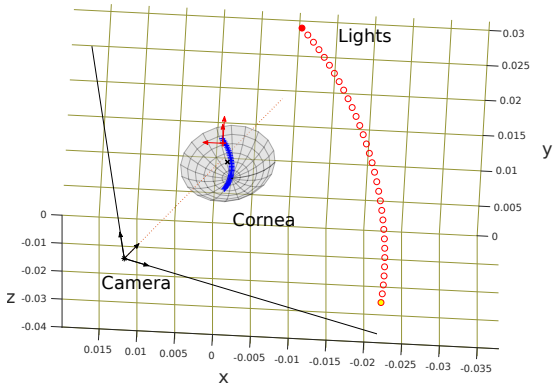

Figure 4: Graphical representation of the simulated setup. Eye-centered coordinate system shown.

First, the system is calibrated using the method described in Section 2.4, with the eye gazing a single point on the screen. The lighting is composed of a planar grid of omni-directional sources, as described in the next section. A typical calibration procedure is performed, with the eye gazing at either 9 or 25 points distributed throughout the screen area, depending on the method employed.

To compute the accuracy, we use 360 points arranged in a regular 24 × 15 grid. The accuracy is defined as the average gaze estimation error over all 360 point locations. Error in gaze estimation is the difference between the actual point position and the estimated gaze position in degrees of visual angle.

3.1 System geometric calibration

As described in Section 2.4, and particularly for the geometric gaze estimation, the calibration must solve for the eyeball center (c_e) and the orbital radius (r_{co}) followed by the cornea. We assessed the estimated values for different feature error magnitudes, varying e_f from 0 to 0.5 pixel in 0.05 steps. For each e_f, we performed 50 calibration trials. Each trial involves gazing each one of the 25 calibration targets, estimating the corneal center (c_c) and radius (r_c). With the 25 corneal centers, we fitted a sphere (baseline) and our method to estimate both c_e and r_{co}. Figure 5a), b), and c) presents the results.

Notice that the sphere fitting performs well with precise reflection locations ($e_f = 0$). However, as the subtended solid angle by gazing the calibration targets is small, the fitting degrades quickly with the increase of e_f. Our method is also affected by noise, but at a lower pace. With respect to the corneal radius, the noise affects the standard deviation with only a slight influence on the averaged value $r_c = 7.78$ mm (SD=0.053 for $e_f = 0.5$).

We also assessed the estimation of the lighting position and orientation. To increase the camera sensor coverage, we used 36 point light sources organized in a 6 × 6 planar grid with 33 mm side. The rigid roto-translation that takes this grid to the lighting used by the gaze estimation is known from calibration. Again, we performed 50 trials per e_f. The results are presented in Figure 5d). To facilitate the interpretation, for each trial, we computed the individual light positions given the position and orientation estimated for the grid, and then computed the mean difference to the ground truth locations.

3.2 Polynomial x geometric gaze calibration

We proposed two methods to estimate the gaze: geometric and polynomial gaze estimation. In geometrical estimation, the gaze is solved as an optimization problem which adjusts the rotation of the eyeball to minimize the difference between the re-projected light sources and the detected ones. In the polynomial estimation, there is an optimization problem as well, which solves for the corneal center. The coordinates along with the coefficients calculated during calibration are then plugged into the polynomial to obtain the gazed screen coordinates.

The geometric gaze estimation performs very well for ground truth c_e, r_{co}, and G, with an accuracy of 0.022° ($SD = 0.012$, $e_f = 0.5$). But the accuracy degrades to 4.62° ($SD = 0.72$) when using our method and to 17.76° ($SD = 6.71$) with the sphere fitting corneal orbit estimation (both with the estimated G). The gaze error with the polynomial (1) with 9-point calibration is 0.70° ($SD = 0.54$), and 0.31° ($SD = 0.24$) with 25 points. The error with the second-order polynomial is 0.30° ($SD = 0.17$). When we use the estimated lighting,

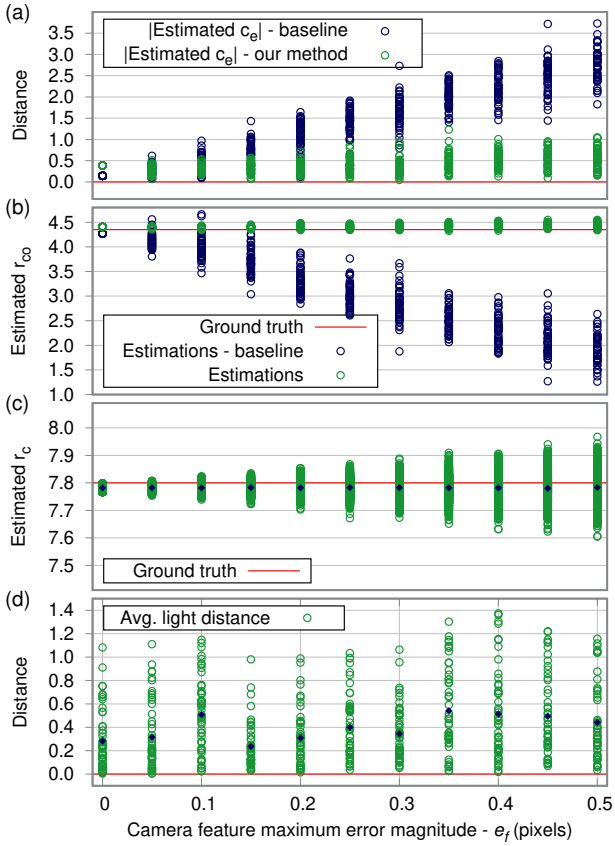

Figure 5: System geometric calibration results for varying feature error magnitudes. Results in millimeters. a) Eye center estimation. b) Corneal orbit radius estimation. c) Corneal radius estimation. d) Average distance between estimated and ground truth lighting positions.

the accuracy with the polynomial (1) with 9-point calibration drops to 1.27° ($SD = 0.74$). With 25-point calibration, both polynomials are little affected, with 0.32° ($SD = 0.12$) and 0.32° ($SD = 0.11$) for the second-order polynomial. Figure 6 shows the distribution of the error in gaze estimation throughout the screen for each calibration method, with the lighting calibrated as described in the previous section. It is worth noting that for the polynomial gaze estimation, only the corneal radius and the lighting position and orientation are necessary, which reduces the source of errors in the gaze estimation.

3.3 Camera and lighting translation

In this section we evaluate how the movements of the camera and lights affect the accuracy after calibration. For this experiment, we assume the setup of Figure 3, and translations of the camera along the plane $c \pm (\frac{w}{2}, 0, \frac{d}{2})$, and then, along the plane $c \pm (0, \frac{h}{2}, \frac{d}{2})$. Where w, h, and d are equal 1 mm.

The experiment is performed in iterations that begin with a calibration with the nominal c and G, followed by the accuracy assessment with perturbed values. The change in position from one iteration to the next is 0.1 mm, and the same translation is applied for both the camera and lights. As we are testing the effects of translations, ground truth values for the eye center (c_e) and corneal orbit radius (r_{co}) are employed. Figure 7 presents the results.

3.3.1 Compensating for translations. We also assessed our method with an additional input (d_a) related to the relative distance of the camera and lighting to the eye. This input can be implemented with a simple infrared distance sensor arranged close to the forehead. We assume the output of such a sensor is linear with the distance, and accordingly, can be modeled as $d_a = c \cdot t + v$, where t is the translation along the z axis, c is a sensor transduction constant, and v is the quantization noise. For the experiments in this section, we assume that $c = 10^3$ and v comes from a zero mean uniform distribution with standard deviation of $8 \cdot 10^{-3}$ (equivalent to 3 LSB in 10 bit ADC of a unity signal).

The accuracy assessment described previously was performed again (for the 2nd order polynomial), this time with 27-point calibration, i.e., the 9-point calibration at three different z coordinates (0, -0.5 mm and -1 mm). Figure 8 presents the results.

3.4 Banding effect in accuracy and robustness

To increase the temporal resolution of our eye-tracking technique, we proposed to reduce the exposure of a rolling shutter camera under stroboscopic light. As a consequence, we have a limited number of lines available to estimate the gaze. In this section, we assess the effect of different banding heights in the accuracy and robustness.

For this experiment, we took only the gaze calibration method we judge to have the best performance, the 2nd order polynomial with 25-point calibration. To estimate the gaze, we considered the light reflections within a central band with varying sizes. As we reduce the band height, less data is available to estimate the gaze (signal-to-noise ratio drops), but the temporal resolution is increased. Figure 9 presents the averaged results over 10 trials for our simulated setup, and for a modified setup where we duplicated the lighting, the new one rotated 64° with respect to plane x-z. The robustness is defined as the percentage of gaze estimations with angular error smaller than 5°, with the accuracy computed over this subset only.

3.5 Discussion

The geometric calibration of the system, particularly for the determination of the eye center (c_e) and corneal orbit radius (r_{co}), did not perform well for an accurate gaze estimation. For the polynomial gaze estimation on the other hand, that require only the lighting position and orientation, the calibration results were satisfactory, with only a minimal difference in the gaze accuracy when compared to results obtained with the ground truth geometry.

The geometric gaze calibration performs well for a perfectly calibrated system, and on top of assumptions that would not be completely satisfied in practice, such as the spheric orbit followed by the cornea. The polynomial method showed a good overall performance, with average gaze estimation errors as small as 0.5°, requiring only the relative lighting position.

Our method requires system geometric calibration and is very sensitive to changes. The accuracy drops a few degrees with less than half a millimeter of deviation from the calibrated setup. The translation compensation method worked effectively in diminishing the effects of calibration deviations.

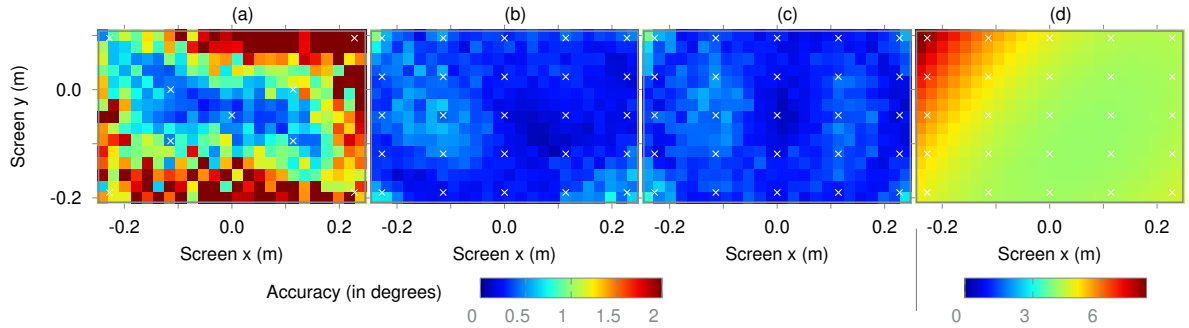

Figure 6: Distribution of the error in gaze estimation throughout the screen area for a given calibration method. a) Polynomial (1) with 9-point calibration. b) Polynomial (1), 25-point calibration. c) Second-order polynomial. Gaze computed by estimating the rotation of the eyeball, 25-point calibration. The white crosses represent the calibration points.

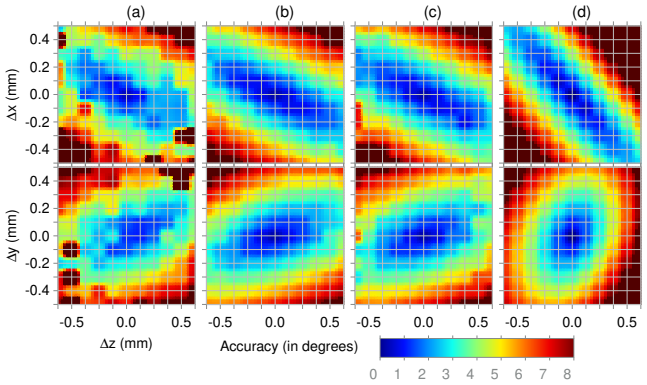

Figure 7: Accuracy in gaze estimation with the camera and lighting translating in the x-z plane (top) and y-z plane (bottom) after calibration. a) Polynomial (1) with 9 point calibration. b) Polynomial (1), 25 points. c) Second-order polynomial, 25 points. d) Geometric gaze estimation, 25 points.

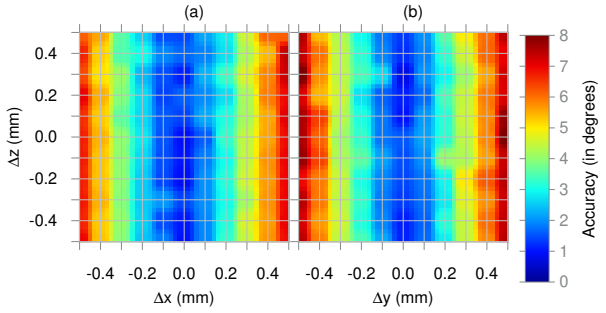

Figure 8: Accuracy in gaze estimation with translation compensation. a) Camera and lighting translating in the x-z plane. b) Camera and lighting translating in the y-z plane.

The band height also impacts the system accuracy and robustness, and is inversely related to the temporal resolution. The inclusion of more light sources improved the results considerably. However, it is important to note that our experiments considered a band at the image center. In practice however, bands close to the top and bottom of the frame might be affected by the eyelid and eyelashes, reducing the robustness for such bands.

Bands	1	2	4	8	12	16	20
Temp. resol. (Hz)	60	120	240	480	720	960	1200
Accuracy (°)	0.32	0.56	1.18	1.31	1.26	1.26	1.34
Std. dev.	0.11	0.58	1.36	1.89	2.09	2.42	2.51
Robustness (%)	100	87	62	30	22	18	15
Accuracy* (°)	0.22	0.29	0.45	0.84	1.17	1.50	1.73
Std. dev.*	0.18	0.26	0.41	0.72	0.93	1.14	1.20
Robustness* (%)	100	100	100	100	97	93	83

Figure 9: Temporal resolution, accuracy, and robustness, versus number of bands in the image for a camera at 60 Hz. * Doubled lighting.

4 SCET PROTOTYPING

The simulated environment allowed us to assess the effects of several factors in the performance of the proposed method, such as different gaze calibration alternatives and the impact of translations to the accuracy. However, the banding formation, an important contribution of our work could not be tested accordingly. In this section we describe the apparatus under construction to capture images under stroboscopic lighting in a setup very similar to the simulated one.

4.1 Apparatus

Our prototype is composed of an imaging and a lighting component. The imaging system is composed of a camera module and a computer platform that controls the camera and processes the images. The camera module uses the OV5647 CMOS chip [OmniVision Technologies, Inc. 2009], a 5 megapixel color image sensor capable of 90 Hz at VGA resolution and 30 Hz at 1080p. The computer board was a Raspberry Pi Model 3B+.

The lighting is composed of a single high efficiency near IR LED from Osram [OSRAM Opto Semiconductors 2014], pointing towards a thin 3D printed circular-section mirror. The mirror is 2 *mm* wide, with a 90° circular section of radius 33 *mm*, and covered with 35 turns of a 0.8 *mm* wire to work as point sources. The LED is triggered by an Arduino board responsible for the stroboscopic activation. The lighting is configured to pulse 16 times during a frame period, resulting in 16 snapshots (bands) per frame. The lighting is kept on for 100 μs, resulting in a 10% duty cycle. The

Figure 10: Prototype employed in the banding assessment.

3D models were made freely available and can be downloaded at thingiverse.com[1]. Figure 10 shows a picture of the prototype.

4.2 Rolling shutter under stroboscopic lighting

To test if the prototype is able to create bands, its illumination was configured to pulse 16 times per frame, with pulse duration set to 100μs. The camera was set to capture VGA images at 60 Hz. The measured frame period was 16.636 ms and the lighting configured with a period of 1039.75 μs. The camera exposure was set while observing the banding effect. The dark regions between adjacent bands almost vanish with an exposure of 1 ms, which is consistent with the developed theory. A sample image is shown in Figure 11. Note that in the current prototype, the lighting is not ideally positioned so the reflections do not cover the whole cornea. Note also that the number of visible bands is smaller than 16, as N is smaller than S in practice due to the invisible scanlines.

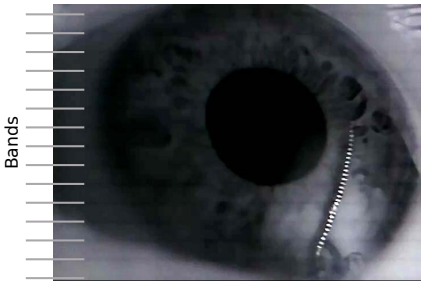

Figure 11: Image capture subject to stroboscopic lighting.

5 DISCUSSION

SCET can potentially increase temporal resolution in gaze estimation by exposing each camera frame to multiple stroboscopic light pulses, enabling us to improve the accuracy during fast eye movements, even using low frame-rate cameras. However, there is a trade-off. As the temporal resolution increases, the data available in each band for estimation reduces, affecting the system accuracy.

The configuration of the lights are also important to the SCET performance. The lights in the prototype were arranged in a very simple way to demonstrate banding. As Figure 9 indicates, more elaborate arrangements might be proposed to improve accuracy and robustness, and to facilitate the matching of each reflected light to its corresponding source.

Another key aspect of our method is the sensitivity to changes in the calibrated geometry. Nonetheless, the proposed approach to incorporate a relative measure of the apparatus distance to the face produced good results, and showed to generalize well for changes in depth not seen during calibration.

When compared to a typical active lighting dark-pupil mobile ET, our method only requires an additional circuit to pulse the lights. The proposed prototype can be constructed for less than $60 USD and the parts can be easily 3D-printed and assembled.

5.1 Limitations and future work

SCET enables higher temporal resolution at the cost of increased latency. The gaze data is only output after the current frame image is processed, which means that the latency can be as long as the period of one frame plus the processing time. Temporal gaps are also possible due to missing data, for example when reflections are obstructed by lashes or during periods when hidden scanlines are being sampled. Interpolation methods can be employed in such cases [Han et al. 2013].

In population, corneas exhibit a prolate elliptical shape [Douthwaite et al. 1999; Read et al. 2006]. Distortions in corneal shape from an unknown individual are not handled by our simple geometric eye model, that assumes the cornea to be spherical.

Our evaluation using real images is in an early stage. We plan to perform new experiments with the prototype to quantify the impact of individual characteristics in the performance of our method, and also to assess how the translation compensation method performs with real data.

Finally, we have not provided in this paper a method to match projected reflections with the actual light sources. While this task is trivial in the simulated environment, it can be quite difficult with real images. We are currently designing a grid of point source illuminators to generate enough glints in one band to estimate the cornea center. The grid would be detected in the whole frame, and the distribution of glints in each band will be used to extend the temporal support of the gaze estimates.

6 CONCLUSION

In this work we presented SCET - a stroboscopic catadioptric eye-tracking system which exploits the minimization of the forward re-projection error of light sources reflected at the cornea to estimate its position, and thereafter, the gaze direction. The temporal resolution is increased by exploiting the rolling shutter mechanism present in most low-cost cameras, associated with a high-speed stroboscopic lighting that allow us to capture multiple bands of corneal reflections temporally offset within a single frame. Therefore, our technique potentially allows the use of low-cost low-frame-rate cameras to obtain high (or increased) speed gaze sampling data.

Our results from simulations using simple lighting patterns demonstrate the feasibility of SCET. We have also presented a working prototype to study the banding effects and we are currently designing an illumination pattern to build a fully functional eye tracker.

ACKNOWLEDGMENTS

This work has been supported by the São Paulo Research Foundation (FAPESP), grant #2016/10148-3. We would also like to thank Paolo Neri for his valuable assistance with catadioptric systems.

[1] https://www.thingiverse.com/thing:3503629

REFERENCES

Sung Joon Ahn, Wolfgang Rauh, and Hans-Jürgen Warnecke. 2001. Least-squares orthogonal distances fitting of circle, sphere, ellipse, hyperbola, and parabola. *Pattern Recognition* 34, 12 (2001), 2283–2303.

Richard Andersson, Marcus Nyström, and Kenneth Holmqvist. 2010. Sampling frequency and eye-tracking measures: how speed affects durations, latencies, and more. *Journal of Eye Movement Research* 3, 3 (2010).

Simon Baker and Shree K Nayar. 1999. A theory of single-viewpoint catadioptric image formation. *International Journal of Computer Vision* 35, 2 (1999), 175–196.

Sandro Barone, Marina Carulli, Paolo Neri, Alessandro Paoli, and Armando Viviano Razionale. 2018. An omnidirectional vision sensor based on a spherical mirror catadioptric system. *Sensors* 18, 2 (2018), 408.

Martin Böhme, Michael Dorr, Mathis Graw, Thomas Martinetz, and Erhardt Barth. 2008. A Software Framework for Simulating Eye Trackers. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications (ETRA '08)*. ACM, New York, NY, USA, 251–258. https://doi.org/10.1145/1344471.1344529

Frank Borsato, Fernando Aluani, and Carlos Morimoto. 2015. A Fast and Accurate Eye Tracker Using Stroboscopic Differential Lighting. In *2015 IEEE ICCVW*. 110–118. https://doi.org/10.1109/ICCVW.2015.72

F. H. Borsato and C. H. Morimoto. 2017. Building Structured Lighting Applications Using Low-Cost Cameras. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. 15–22. https://doi.org/10.1109/SIBGRAPI.2017.9

F. H. Borsato and C. H. Morimoto. 2018. Asynchronous stroboscopic structured lighting image processing using low-cost cameras. In *2018 31th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. http://urlib.net/rep/8JMKD3MGPAW/3RNU8QS?ibiurl.backgroundlanguage=en

Derek Bradley, Bradley Atcheson, Ivo Ihrke, and Wolfgang Heidrich. 2009. Synchronization and rolling shutter compensation for consumer video camera arrays. In *CVPR Workshops 2009*. IEEE, 1–8.

Andreas Bulling and Hans Gellersen. 2010. Toward mobile eye-based human-computer interaction. *IEEE Pervasive Computing* 4 (2010), 8–12.

Juan J. Cerrolaza, Arantxa Villanueva, and Rafael Cabeza. 2008. Taxonomic Study of Polynomial Regressions Applied to the Calibration of Video-oculographic Systems. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications (ETRA '08)*. ACM, New York, NY, USA, 259–266. https://doi.org/10.1145/1344471.1344530

Z Ramdane Cherif, A Nait-Ali, JF Motsch, and MO Krebs. 2002. An adaptive calibration of an infrared light device used for gaze tracking. In *Instrumentation and Measurement Technology Conference, 2002. IMTC/2002. Proceedings of the 19th IEEE*, Vol. 2. IEEE, 1029–1033.

Kai Dierkes, Moritz Kassner, and Andreas Bulling. 2018. A Novel Approach to Single Camera, Glint-free 3D Eye Model Fitting Including Corneal Refraction. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA '18)*. ACM, New York, NY, USA, Article 9, 9 pages. https://doi.org/10.1145/3204493.3204525

W. A. Douthwaite, T. Hough, K. Edwards, and H. Notay. 1999. The EyeSys videokeratoscopic assessment of apical radius and p-value in the normal human cornea. *Ophthalmic and Physiological Optics* 19, 6 (1999), 467–474. https://doi.org/10.1046/j.1475-1313.1999.00462.x

Shaharam Eivazi, Thomas C. Kübler, Thiago Santini, and Enkelejda Kasneci. 2018. An Inconspicuous and Modular Head-mounted Eye Tracker. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA '18)*. ACM, New York, NY, USA, Article 106, 2 pages. https://doi.org/10.1145/3204493.3208345

Wolfgang Fuhl, David Geisler, Thiago Santini, Tobias Appel, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2018. CBF: Circular Binary Features for Robust and Real-time Pupil Center Detection. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA '18)*. ACM, New York, NY, USA, Article 8, 6 pages. https://doi.org/10.1145/3204493.3204559

Georg Glaeser. 1999. Reflections on spheres and cylinders of revolution. *Journal for Geometry and Graphics* 3, 2 (1999), 121–139.

Matthias Grundmann, Vivek Kwatra, Daniel Castro, and Irfan Essa. 2012. Calibration-free rolling shutter removal. In *Computational Photography (ICCP), 2012 IEEE International Conference on*. IEEE, 1–8.

Peng Han, Daniel R Saunders, Russell L Woods, and Gang Luo. 2013. Trajectory prediction of saccadic eye movements using a compressed exponential model. *Journal of vision* 13, 8 (2013), 27–27.

Dan Witzner Hansen and Qiang Ji. 2010. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence* 32, 3 (2010), 478–500.

Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. (April 2014). arXiv:cs-cv/1405.0006 http://arxiv.org/abs/1405.0006

Päivi Majaranta and Andreas Bulling. 2014. Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*. Springer, 39–65.

Christopher Mei and Patrick Rives. 2007. Single view point omnidirectional camera calibration from planar grids. In *Robotics and Automation, 2007 IEEE International Conference on*. IEEE, 3945–3950.

Jorge J Moré and Danny C Sorensen. 1983. Computing a trust region step. *SIAM J. Sci. Statist. Comput.* 4, 3 (1983), 553–572.

C. H. Morimoto, A. Amir, and M. Flickner. 2002. Detecting eye position and gaze from a single camera and 2 light sources. In *Object recognition supported by user interaction for service robots*, Vol. 4. 314–317 vol.4. https://doi.org/10.1109/ICPR.2002.1047459

Carlos Hitoshi Morimoto, Dave Koons, Arnon Amir, and Myron Flickner. 2000. Pupil detection and tracking using multiple light sources. *Image and vision computing* 18, 4 (2000), 331–335.

Rhys Newman, Yoshio Matsumoto, Sebastien Rougeaux, and Alexander Zelinsky. 2000. Real-time stereo tracking for head pose and gaze estimation. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. IEEE, 122–128.

Christian Nitschke, Atsushi Nakazawa, and Haruo Takemura. 2013. Corneal imaging revisited: An overview of corneal reflection analysis and applications. *IPSJ Transactions on Computer Vision and Applications* 5 (2013), 1–18.

OmniVision Technologies, Inc. 2009. *OV5647 1/4 color CMOS QSXGA (5 megapixel) image sensor with OmniBSI technology*.

OSRAM Opto Semiconductors. 2014. *SFH 4715 OSLON Black Series (850 nm)*. http://www.osram-os.com/Graphics/XPic5/00100752_0.pdf

QImaging. 2014. *Rolling Shutter vs. Global Shutter*. Technical Report. 9 pages. https://www.qimaging.com/ccdorscmos/pdfs/RollingvsGlobalShutter.pdf

Scott A Read, Michael J Collins, Leo G Carney, and Ross J Franklin. 2006. The topography of the central and peripheral cornea. *Investigative ophthalmology & visual science* 47, 4 (2006), 1404–1415.

William Rosengren, Marcus Nystöm, Björn Hammar, and Martin Stridh. 2018. Suitability of Calibration Polynomials for Eye-tracking Data with Simulated Fixation Inaccuracies. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA '18)*. ACM, New York, NY, USA, Article 66, 5 pages. https://doi.org/10.1145/3204493.3204586

Davide Scaramuzza, Agostino Martinelli, and Roland Siegwart. 2006. A toolbox for easily calibrating omnidirectional cameras. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*. IEEE, 5695–5701.

Rahul Swaminathan, Michael D Grossberg, and Shree K Nayar. 2006. Non-single viewpoint catadioptric cameras: Geometry and analysis. *International Journal of Computer Vision* 66, 3 (2006), 211–229.

Tobii AB. 2018. *Tobii Pro Glasses 2*. https://www.tobiipro.com/siteassets/tobii-pro/product-descriptions/tobii-pro-glasses-2-product-description.pdf/?v=1.95

Jian-Gang Wang, Eric Sung, and Ronda Venkateswarlu. 2005. Estimating the eye gaze from one eye. *Computer Vision and Image Understanding* 98, 1 (2005), 83–103.

Zhengyou Zhang. 2000. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence* 22 (2000).