2018 IEEE International Conference on Robotics and Automation (ICRA)
May 21-25, 2018, Brisbane, Australia

Real-time 3D Glint Detection in Remote Eye Tracking Based on Bayesian Inference

David Geisler[1], Dieter Fox[2] and Enkelejda Kasneci[1]

Abstract— **As human gaze provides information on our cognitive states, actions, and intentions, gaze-based interaction has the potential to enable a fluent and natural human-robot collaboration. In this work, we focus on reliable gaze estimation in remote eye tracking based on calibration-free methods. Although these methods work well in controlled settings, they fail when illumination conditions change or other objects induce noise. We propose a novel, adaptive method based on a probabilistic model, which reliably detects glints from stereo images and evaluate our method using a data set that contains different challenges with regarding to light and reflections.**

I. Introduction

Eye tracking holds enormous potential for real-time recognition of our behavior, actions, and intentions, and can therefore be used to improve the interaction between robots and humans in a much more natural way [26]. For instance, an autonomous robot or a self-driving car should detect whether a pedestrian is aware of the moving robot or whether the car's driver is observing the road ahead; or a manipulator needs to know which object the person is looking at. In fact, since the way we explore our environment (i.e. the selection of fixation targets), is closely linked to higher cognitive processes, gaze direction provides fine-grained information about the focus of the attention [11]. Characteristic patterns in eye movements can be used to draw conclusions about the current activity of a subject, the degree of task difficulty, or assess expertise [4], [20], [21]. Simultaneously, eye contact and gaze gestures provide important clues in natural human communication [1]. Especially in the interaction with autonomous agents, information derived from gaze patterns of the user could be used to adapt the behavior of these agents more efficiently and naturally: users may use their gaze as an advanced input device to control applications just by looking on it or as a combination with conventional input methods such as voice control, where the gaze preselects the affected component in e. g. the navigation system or speedometer to setup a new route or maximum velocity, respectively [12], [23].

Video-based eye tracking is characterized by two approaches. Head-mounted devices are attached to the subject's head and record the pupil position relative to their head position. This positioning comes with the advantage that the cameras are located close to the eyes and with a fixed axial offset, which allows for accurate and highly available gaze estimation [14]. In contrast to head-mounted devices, remote eye-tracking systems record the subject out of a stationary position and estimate a subject's gaze vector in a 3D scene. Such systems usually deal with a low resolution of the eye, which make accurate gaze estimation more challenging [13]. Additionally, the working area is limited by the camera perspective, which restricts the user mobility. Conversely, remote eye tracking has the advantage of being non-invasive, which is an imperative requirement for many real-world applications since they are more comfortable to use and work without additional effort for the user.

An unsolved problem in remote eye tracking is the mapping of the detected pupil position on the eye video to a point of regard (POR) in the scene. A common way to cope with this challenge for head-mounted eye tracking is to regress a function that maps the pupil position to a POR on a camera scene through a calibration step. Despite recent advances to make calibration faster [25], recalibration is needed if, for example, the device slips. Using remote eye tracking, the estimate can be achieved through appearance-based methods, which are still restricted to accuracies around six degrees [34]. The only calibration-free and accurate approaches are model-based, which usually require the detection of the eyes, the eye location in the 3D space, and an eye model where the eye center is used as the gaze vector origin. Additionally, the pupil center must be located within the eye region, which is used as an intersection point of the gaze vector. Therefore, a robust eye model is a key component for calibration-free eye tracking. Parametrizing the eye model in a robust way remains, however, an open challenge that directly affects the applicability of model-based gaze estimation methods. For example, Swirski et al. describe a method for estimating the eyeball and cornea as two intersecting spheres based on the detected pupil contour [30]. Their method, however, requires a high resolution of the pupil in the image and a reliable contour detection [30]. Other approaches use characteristic reflections of light sources or objects caused by the structure of the human eye, which are known as Purkinje images. The most pronounced Purkinje image is caused from the surface of the cornea and is referred to as glint. Some approaches use these Purkinje images to determine reference points of the internal eye structure, which are then further used to estimate the eye pose in space. The gaze direction can be then derived without an explicit detection of the pupil boundary, but requires a very accurate localization of reflections, which is usually

[1] Department of Perception Engineering, University of Tuebingen, Germany. {david.geisler,enkelejda.kasneci}@uni-tuebingen.de
[2] Department of Computer Science & Engineering, University of Washington, Seattle, USA, fox@cs.washington.edu

978-1-5386-3081-5/18/$31.00 ©2018 IEEE

not feasible in the wild [5]–[7]. Such approaches are not applicable to remote eye tracking due to low resolution and low contrast of the eye image compared to head-mounted eye tracking. Therefore, various approaches adapt the eye model based on a combination of detected facial landmarks and glints, e.g., [31], [35].

Most glint-based approaches, work well for controlled (laboratory) settings and assume a fixed number of static light sources. Under these constraints, it is possible to extract glints by investigating histograms [17] or applying edge detectors [27], [28], but they fail as soon as these laboratory conditions are no longer present. Strong light gradients, varying lighting, and reflections, as found especially in driving scenarios, are very likely to interfere with glint detection. In this paper, we address the aforementioned challenges and provide a method for robust glint detection in remote eye tracking based on a probabilistic method. In the next section, we first propose a robust filter to extract glint candidates from stereo images inspired by FAST-like features and perform a stereo matching to select the most probable glint pair for each eye. These points can then be used in a subsequent step to adapt an eye model. Additionally, we introduce a new dataset for glint detections in Section III and compared our approach with state-of-the-art glint detectors in Section IV.

II. Methods

Our glint detection method consists of two parts: (i) first, glint candidates are extracted from a stereo input image by a spatial peak detection. We call this step FAST2 since it was inspired by the FAST feature detection. Then, (ii) after detecting potential glint candidates, we use stereo matching and triangulation to filter the most likely glint pair in 3D by applying an adaptive probabilistic model. The second filter step allows to adjust the filters in the first step less restrictively, which on one hand leads to a larger amount of extracted glint candidates due to reflections, other light sources or other noise, but also considers weakly pronounced glints. This way, we are able to detect glints even under very challenging illumination conditions. In addition, the probabilistic model provides a suitable confidence measure, which enables us to detect the lack of a glint. Figure 1 outlines the proposed algorithmic pipeline consisting of six successive steps, whereby the steps ①-③ address the extraction of glint candidates, whereas the steps ④-⑥ implement the stereo matching and the probabilistic model. In the following, we describe these steps in detail.

① *Stereo Capture:* Since the proposed probabilistic model operates on glints in a 3D coordinate system, we use a stereo camera to retrieve depth information. Our setup consists of two planar aligned cameras with a fixed vertical and horizontal offset. The subject's head should be facing to the cameras, so as both glints are visible.

② *FAST2:* We propose a novel filter to highlight glints in an image based on the assumption that areas in the image corresponding to glints are significantly brighter than their closer neighborhood. We define a glint as a group of related pixels that are at least brighter than a threshold t_{fast2} compared to all surrounding pixels in a certain radius r_{fast}. We take the difference of the maximal intensity in the surrounding pixels $e_0, \ldots, e_n$ to the considered pixel c_{xy} and normalize them by the average intensity $\hat{c}_{xy}$ around c_{xy} as shown in Equation (1). The resulting response k_{xy} is then thresholded by t_{fast2}, which leads to the sparse response map $m_{xy} \in M$. The threshold t_{fast2} controls the sensitivity of the glint detection (smaller t_{fast2} results in a more sensitive glint detection). We regulate t_{fast2} in an adaptive way, by decreasing it if less than a certain number of glints were detected or no two corresponding glints could be selected in the last iteration in step ⑥. Accordingly, t_{fast2} is increased if too many glints were found and the glint detection was successful. The radius r_{fast2} controls the spatiality of the glint detection. It should be larger than the maximal expected glint radius, but as small as possible.

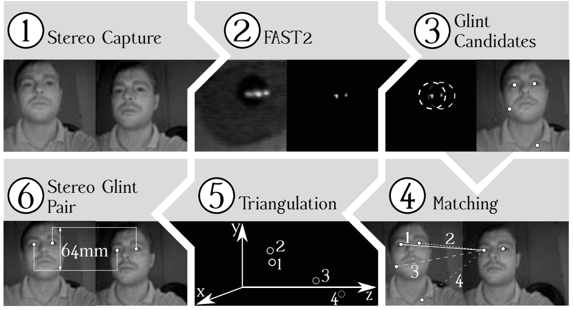
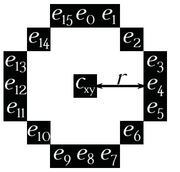


Fig. 1: Workflow of our algorithm. The upper row shows the extraction of glint candidate from input images (①) captured from a stereo camera. In step ②, we apply our FAST2 feature extraction to emphasize the glints in a response map. Finally, the highest peaks are selected. The lower row shows the stereo matching of the extracted glint pairs ④, their mapping into 3D space ⑤, and the selection of the most probable glint pair based on a probabilistic model ⑥.



$$k_{xy} = \frac{c_{xy} - \max(e_0, \ldots, e_n)}{\hat{c}_{xy}}$$
$$m_{xy} = \begin{cases} k_{xy} & t_{\text{fast2}} \leq k_{xy} \\ 0 & t_{\text{fast2}} > k_{xy} \end{cases}, \tag{1}$$

where $c_{xy} \in I$ is the raw and $\hat{c}_{xy} \in \hat{I}$ is the average filtered input image. $k_{xy} \in K$ contains the dense, and $m_{xy} \in M$ the sparse response map of the feature extraction.

③ *Glint Candidates:* After extracting the above features we sort the pixels by their response and consider the n highest as glint candidates. Taking into account that glints may spread over more than one pixel (depending on r_{fast2}), we suppress new glint candidates by weighting the distance to all already found candidates exponentially based on the following condition:

$$g_i = \begin{cases} m_i & \text{if } t_{\text{dist}} < \sum\limits_{j<i}^{j=0} e^{-\frac{\|m_i - g_j\|^2}{2\sigma^2}} \text{ and } m_i \in R \\ \emptyset & \text{otherwise} \end{cases}, \quad (2)$$

where $g_i \in G$ is the i-th glint candidate and $m_i \in M$ the i-th element of the sorted pixels. t_{dist} and σ are two competing parameters to adjust the minimal allowed distance of two glints. We suggest to set $\sigma = \sqrt{0.5} \cdot r$ and choose $t \leq e^{-1}$ to ensure a minimal distance of r of two glint candidates. R is the region of interest where the head of the subject is expected. Without further knowledge, we choose R to cover the whole input image.

④ *Matching:* After extracting a set of glint candidates $g_{\text{left } l} \in G_{\text{left}}$ and $g_{\text{right } k} \in G_{\text{right}}$, in the left and right frame, respectively, the following step matches the two sets by their maximum likelihood $\mathcal{L}_{\text{match } l,k}$, taking into account the similarity of the patches around the glints as well as geometric constraints derived from the camera setup. For this purpose, we define in the following, a series of likelihood functions that model the requirements as non-normalized exponential distributions.

The similarity likelihood $\mathcal{L}_{\text{BRIEF } l,k}$ of two glints is defined by the distance of the BRIEF descriptor for a patch around the glint candidate of extracted from the left respectively right frame.

$$v_{\text{BRIEF } l,k} = \delta(f_{\text{BRIEF}}(g_{\text{left } l}), f_{\text{BRIEF}}(g_{\text{right } k})), \quad (3)$$

$$\mathcal{L}_{\text{BRIEF } l,k} = e^{-\frac{v^2_{\text{BRIEF } l,k}}{2 \cdot \sigma^2_{\text{BRIEF}}}}, \quad (4)$$

where $\delta(f_{\text{BRIEF}}(g_{\text{left } l}), f_{\text{BRIEF}}(g_{\text{right } k}))$ is the Haussdorf distance of the two BRIEF features f_{BRIEF} of the glints $g_{\text{left } l}$ and $g_{\text{right } k}$. The scaling parameter σ_{BRIEF} defines the tolerance of deviation in the BRIEF features.

Considering the planar alignment of the cameras and the fixed distance between them, the angle between g_{left} and the corresponding g_{right} can be deduced. Taking into account any imperfect distortion or alignment of the cameras, we also describe the geometric constraints as an unnormalized distribution to accommodate minor deviations:

$$v_{\text{angle } l,k} = g_{\text{right } k} \underset{xy}{\sphericalangle} g_{\text{left } l}, \quad (5)$$

$$\mathcal{L}_{\text{angle } l,k} = e^{-\frac{(C_1 \underset{xy}{\sphericalangle} C_0 - v_{\text{angle } l,k})^2}{2 \cdot \sigma^2_{\text{angle}}}}, \quad (6)$$

where C_1 and C_0 are matrices representing the intrinsic parameters of the corresponding cameras (see Figure 2). $\mathcal{L}_{\text{angle } l,k}$ provides the likelihood scaled by σ_{angle} of the angle between $g_{\text{left } l}$ and $g_{\text{right } k}$. Since $\mathcal{L}_{\text{angle } l,k}$ was designed to compensate small errors in the alignment of the cameras, calibration, or other noise, we suggest to choose $\sigma_{\text{angle}} \propto$ RMS Error of the intrinsic and extrinsic calibration. Besides the angle of the detected glint candidates in the stereo image, we also limit the minimal and maximal depth of the glint regarding the working distance of w_{min} up to w_{max}[1]:

$$v_{\text{max}} = P_1 w_{\text{min}}^{\top} - P_0 w_{\text{min}}^{\top}, \quad v_{\text{min}} = P_1 w_{\text{max}}^{\top} - P_0 w_{\text{max}}^{\top}, \quad (7)$$

where P_0 and P_1 are projection matrices of the both cameras. The vectors v_{min} and v_{max} describe the minimal and maximal pixel distance on the projection plane of both glints $g_{\text{left } l}$ and $g_{\text{right } k}$ inside the working distance. As above, we define the constraint as a likelihood function:

$$v_{\text{dist } l,k} = g_{\text{right } k} - g_{\text{left } l}, \quad (8)$$

$$\mathcal{L}_{\text{dist } l,k} = e^{-\frac{\|(0.5 \cdot v_{\text{min}} + 0.5 \cdot v_{\text{max}}) - v_{\text{dist } l,k}\|^2}{2 \cdot \sigma^2_{\text{dist}}}}, \quad (9)$$

whereby the scaling parameter σ_{dist} should be selected in such a way that v_{min} and v_{max} result in a likelihood above 0.5. Please note, we assume that one pixel on the sensor has the same scale in width and height. If that does not hold, v_{dist} needs to be scaled properly.

The final matching likelihood $\mathcal{L}_{\text{match } l,k}$ of two glint candidates $g_{\text{left } l}$ and $g_{\text{right } k}$ is defined as following:

$$\mathcal{L}_{\text{match } l,k} = \mathcal{L}_{\text{BRIEF } l,k} \cdot \mathcal{L}_{\text{angle } l,k} \cdot \mathcal{L}_{\text{dist } l,k}. \quad (10)$$

⑤ *Triangulation:* Using triangulation, two corresponding glint image points can be mapped from the respective camera to 3D. In theory, it suffices to calculate the intersection of the two rays derived from the pinhole models of both cameras to obtain the projected point in 3D. Indeed, noise in the intrinsic and extrinsic calibration as well as sampling lead to the fact that the derived rays probably will not intersect. The challenge, then, is to find a 3D glint point which describes the detected corresponding glint image points optimal. There are several approaches to define the optimality and how to calculate it [15], [22]. The most common and intuitive approach that we also apply to our problem, is to solve the linear equation system $P_0 \cdot g_{\text{scene}} = g_{\text{left}}$ and $P_1 \cdot g_{\text{scene}} = g_{\text{right}}$ for a least squares error [16].

⑥ *Stereo Glint Pair:* After triangulating the detected glint candidates from the image points to 3D scene points, we select the most probable pair of glints representative to the left and right eye. Similar to step ④, we define likelihood functions to model the probability of a glint pair based on geometric constraints. For that, we consider the interpupillary distance $v_{\text{IPD } u,v}$ which is spanned by the two considered glint candidates $g_{\text{scene } u}$ and $g_{\text{scene } v}$, as well as the rotations $v_{\text{yaw } u,v}$ and $v_{\text{roll } u,v}$, and the distance $v_{\text{remote } u,v}$ relative to the camera system as shown in Figure 3.

The mean interpupilar distance (IPD) on humans is around $63\,\text{mm}$, with the vast majority of adults having IPDs in the range $50\,\text{mm}$-$75\,\text{mm}$ [9]. Since IPD and the distance of the glints are both depending on the distance between the eyeballs, we assume that the distance of the glints $g_{\text{scene } u}$ and $g_{\text{scene } v}$ should be similar to the IPD:

$$\mathcal{L}_{\text{IPD } u,v} = e^{-\frac{(63\,\text{mm} - \|g_{\text{scene } u} - g_{\text{scene } v}\|)^2}{2 \cdot \sigma^2_{\text{IPD}}}}, \quad (11)$$

[1] In our instance: $w_{\text{min}} = (0, 0, 800)$, $w_{\text{max}} = (0, 0, 1200)$

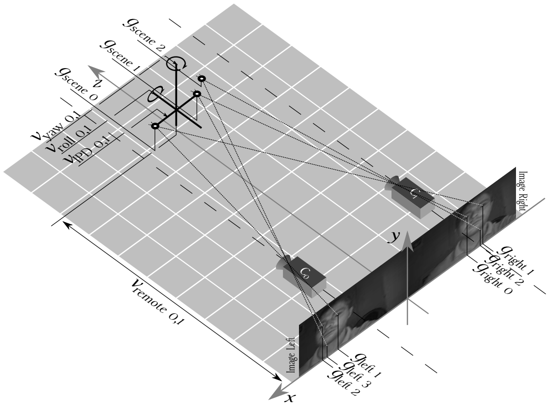


Fig. 2: After the triangulation of the matched glints in step ⑤, the pairwise likelihood $\mathcal{L}_{\text{pair}}$ of the 3D glint candidates is calculated according to Eq. (16) using the metrics v_{yaw} (vertical head rotation), v_{roll} (lateral head inclination), v_{IPD} (pupil distance) and v_{remote} (distance to the camera system).

where $\mathcal{L}_{\text{IPD }u,v}$ is the likelihood of the glint pair $g_{\text{scene }u}$ and $g_{\text{scene }v}$. We choose $\sigma_{\text{IPD}} \approx 11.04\,\text{mm}$ to keep $\mathcal{L}_{\text{IPD }u,v} \geq 0.5$ for the range $50\,\text{mm}$-$75\,\text{mm}$ for the IPD.

In addition to the IPD, we also include the rotation as a constraint for valid glint pairs. We assume that it is unlikely to get valid glints with a vertical head rotation (yaw) of more than a certain degree (i.e., $\pm 30°$). Of course, we can not rule out that there is still head poses existing with more than $\pm 30°$ vertical rotation which produces visible glints on both eyes in the stereo images. Therefore, we assume a decreasing probability for such glint pairs:

$$v_{\text{yaw }u,v} = g_{\text{scene }v} \underset{zx}{\sphericalangle} g_{\text{scene }u}, \tag{12}$$

$$\mathcal{L}_{\text{yaw }u,v} = e^{-\frac{{v_{\text{yaw }u,v}}^2}{2\cdot\sigma^2_{\text{yaw}}}}, \tag{13}$$

where $\mathcal{L}_{\text{yaw }u,v}$ is the likelihood of the glint pair $g_{\text{scene }u}$ and $g_{\text{scene }v}$ with the vertical rotation $v_{\text{yaw }u,v}$. We suggest to choose $\sigma_{\text{yaw}} \approx 4.25°$ to keep $\mathcal{L}_{\text{yaw}} \geq 0.5$ for vertical head rotations less than $\pm 35°$. Simultaneously to the vertical head rotation we constrain the lateral head inclination to $\pm 30°$ and define analogously $\mathcal{L}_{\text{roll }u,v}$.

Finally, we determine the distance between the glint pair and the camera system to ensure that only glint pairs within the defined working range w_{min} - w_{max} are selected:

$$v_{\text{remote }u,v} = 0.5 \cdot g_{\text{scene }v} + 0.5 \cdot g_{\text{scene }u}, \tag{14}$$

$$\mathcal{L}_{\text{remote }u,v} = e^{-\frac{\|0.5\cdot w_{\text{min}}+0.5\cdot w_{\text{max}}-v_{\text{remote }u,v}\|^2}{2\cdot\sigma^2_{\text{remote}}}}, \tag{15}$$

where $\mathcal{L}_{\text{remote }u,v}$ represents the likelihood of the distance of the glint pair $(g_{\text{scene }v}, g_{\text{scene }u})$ regarding the working distance. We suggest choosing the scaling parameter σ_{remote} in such way that a likelihood of at least 0.5 can be obtained at the border of the working area. The overall likelihood $\mathcal{L}_{\text{pair }u,v}$ for a certain glint pair is defined as the product of all single metric likelihoods:

$$\begin{aligned}\mathcal{L}_{\text{pair }u_0,u_1} =& \mathcal{L}_{\text{match }u_0,u_1} \cdot \mathcal{L}_{\text{match }v_0,v_1} \cdot \mathcal{L}_{\text{IPD}} \cdot \\ & \mathcal{L}_{\text{yaw }u,v} \cdot \mathcal{L}_{\text{roll }u,v} \cdot \mathcal{L}_{\text{remote }u,v}\end{aligned} \tag{16}$$

where $\mathcal{L}_{\text{match }u_0,u_1}$ and $\mathcal{L}_{\text{match }v_0,v_1}$ are the matching likelihood of the two 2D glint candidates of the triangulated glint $g_{\text{scene }u}$ and $g_{\text{scene }v}$. The glint pairs with a likelihood higher than a certain threshold t (we usually use $t = 0.5$) are reported as valid detected glints:

$$d_{u,v} = \begin{cases} 0 & \text{if } \mathcal{L}_{\text{pair }u,v} < t \\ 1 & \text{otherwise} \end{cases}, \tag{17}$$

with $0 < t < 1$. $d_{u,v}$ indicates if the corresponding combination of glint candidates were classified as valid (1) or not (0).

a) Confidence measure: $c_{u,v}$ indicates the certainty of a classification for a glint pair:

$$c_{u,v} = \begin{cases} \mathcal{L}_{\text{pair }u,v} & \text{if } \mathcal{L}_{\text{pair }u,v} > t \\ 1 - \mathcal{L}_{\text{pair }u,v} & \text{otherwise} \end{cases}. \tag{18}$$

Considering that the detection $d_{u,v} \in 0, 1$ of a glint pair in the stereo image is Bernoulli distributed, and the conjugated likelihood $\mathcal{L}_{\text{pair }u,v}$ is part of the exponential family, we can conclude that the prior probability for the correct classification can be modeled using a Beta distribution [8]:

$$P(\mathcal{L}_{\text{pair }u,v}) = \frac{{\mathcal{L}_{\text{pair }u,v}}^{\alpha-1} \cdot (1 - \mathcal{L}_{\text{pair }u,v})^{\beta-1}}{\mathrm{B}(\alpha, \beta)}, \tag{19}$$

where α and β control the expected correct and incorrect prediction rate, respectively. Given prior knowledge of the accuracy of the algorithm for a specific setup, it can be integrated into an advanced confidence measure $\hat{c}_{u,v}$:

$$\hat{c}_{u,v} = \begin{cases} P(\mathcal{L}_{\text{pair }u,v} < t) & \text{if } \mathcal{L}_{\text{pair }u,v} > t \\ P(\mathcal{L}_{\text{pair }u,v} > t) & \text{otherwise} \end{cases}. \tag{20}$$

b) Adaptive parameters: The behavior of the presented classifier is strongly related to the choice of the parameters of each single likelihood function. In the following, we will adapt these parameters in an online fashion in order to adapt the classifier to the subject and compensate for possible calibration inaccuracies. Since the likelihood functions considered here are proportional to a normal distribution (without the constant normalization), we can transform the parameters of the likelihood function in exponential distributions and use Bayesian inference to update them after each iteration. Therefore, we redefine the likelihood $\mathcal{L}_{\text{pair }u,v}$ as the marginal likelihood:

$$\mathcal{L}_{\text{pair }u,v} = P(X_i|\eta_{i-1}) = \int_\theta P(X_i|\theta) \cdot P(\theta|\eta_{i-1}) d\theta, \tag{21}$$

where X_i is the considered set of glint candidates $\{g_{\text{left }u_0}, g_{\text{right }u_1}, g_{\text{left }v_0}, g_{\text{right }v_1}\}$ extracted in iteration i. $P(X_i|\theta)$ is the product of the likelihoods as previously stated in eq. (16) using the parameters θ. η_{i-1} is the set of the hyperparameters of the parameter distributions gain from the last iteration. The Bayes' theorem allows us to update the parameter distribution by the observed data as follows [19]:

$$P(\theta|\eta_i) \propto P(X_i|\theta) \cdot P(\theta|\eta_{i-1}) = P(\theta|\eta_0) \cdot \prod_{j=1}^{i} P(X_j|\theta), \tag{22}$$

where $P(\theta|\eta_0)$ is the probability of the parameters θ given the initial hyperparameters η_0, and $P(X_j|\theta)$ is the probability of the glint pair X_j given the parameters θ. This requires that the complete data series $X_0 \dots X_i$ seen up to now must be stored. Since this is not practicable for large data series, we estimate the hyperparameter after each iteration anew:

$$P(\theta|\eta_i) = P(\theta|\eta_{i-1} + \delta\eta_i), \tag{23}$$

where $\delta\eta_i$ is the set of gradients between the hyperparameters η_i and η_{i-1}. All the likelihood functions $L_{\text{match}} \dots L_{\text{pair}}$ can be converted to the form:

$$\mathcal{L} = P(X_i|\{\mu, \sigma^2\} \in \theta) = e^{-\frac{(\mu - X_i)^2}{2 \cdot \sigma^2}}, \tag{24}$$

which corresponds to a non-standardized normal distribution. The likehood of the parameters μ and σ are described by the distributions based on the hyperparameters η. We approximate the gradients $\delta\eta$ very roughly by the distance of the observed data and estimated mean:

$$\begin{aligned} \delta\hat{\eta}_{\mu \text{ mean } i} &\approx \eta_{\mu \text{ mean } i-1} - X_i, \\ \delta\hat{\eta}_{\mu \text{ variance } i} &\approx \delta\hat{\eta}_{\mu \text{ mean } i}^{\;2}, \\ \delta\hat{\eta}_{\sigma \text{ mean } i} &\approx \eta_{\sigma \text{ mean } i-1} - (\eta_{\mu \text{ mean } i-1} - X_i)^2, \\ \delta\hat{\eta}_{\sigma \text{ variance } i} &\approx \delta\hat{\eta}^2_{\sigma \text{ mean } i}, \end{aligned} \tag{25}$$

and normalize them by a weight function h (e.g., tanh) and the marginal likelihood:

$$\delta\eta_i = h(\delta\hat{\eta}_i) \cdot P(X_i|\eta_{i-1}). \tag{26}$$

Finally, we estimate η_i using Stochastic Gradient Descent (SGD) to minimize $\delta\hat{\eta}$ [3]. This way, the parameters of the model are continuously fitted to the observed glints, which works fine for parameters modeling physical circumstances, such as the IPD. However, we need to be careful when adapting parameters that model temporal circumstances, such as the distance of glints to the camera system. Otherwise, it may happen that after sometime of recording and adaption, the model is overfitted and not able to detect the glints anymore (e.g., if the subject moves around). Therefore, we limited the hyperparameters representing the variance to a lower bound. In addition, we generalize our model when the confidence level falls below a critical value $t_{\hat{c}}$, so that it is possible to adapt to larger changes after a few iterations:

$$\delta\eta_i = \begin{cases} h(\delta\hat{\eta}_i) \cdot P(X_i|\eta_{i-1}) & \text{if } \hat{c} > t_{\hat{c}} \\ h(\eta_0 - \eta_{i-1}) & \text{otherwise} \end{cases}. \tag{27}$$

The number of extracted glint candidates per frame and the resulting number of possible combinations of glints has a considerable impact on the runtime. Therefore, we adapt the threshold t_{fast2} with the gradient $\delta\hat{\eta}_{\text{fast2}}$ as the difference between the number of detected glints and the number of expected glints. In addition, we adapt the region of interest (ROI, defined by a center point and the width and height of a rectangle: $R = (x, y, w, h)$) to reduce the search area for glint candidates. The gradient of the ROI ($\delta\hat{\eta}_{\text{ROI}}$) is calculated as the difference between the ROI from the last iteration and a ROI defined around the detected glint with a fixed padding.

Function	Parameter	Hyper-parameter η	Initial η_0	Description
M_{left}		$\eta_{t_{\text{fast2}}}$ left	0	Threshold for FAST2 response on the left and right input image (1).
M_{right}		$\eta_{t_{\text{fast2}}}$ right	0	
G_{left}	R_{left}	η_{R_x} left	$\frac{I_{\text{width left}}}{2}$	Region of interest in which valid glint candidates are expected (2).
		η_{R_y} left	$\frac{I_{\text{height left}}}{2}$	
		η_{R_w} left	$I_{\text{width left}}$	
		η_{R_h} left	$I_{\text{height left}}$	
G_{right}	R_{left}	η_{R_x} right	$\frac{I_{\text{width right}}}{2}$	
		η_{R_y} right	$\frac{I_{\text{height right}}}{2}$	
		η_{R_w} right	$I_{\text{width right}}$	
		η_{R_h} right	$I_{\text{height right}}$	
L_{BRIEF}	μ_{BRIEF}	η_μ mean BRIEF	0	BRIEF similarity likelihood of a glint candidate from the left and the right stereo image (4).
		η_μ variance BRIEF	1	
	σ_{BRIEF}	η_σ mean BRIEF	1	
		η_σ variance BRIEF	1	
L_{angle}	μ_{angle}	η_μ mean angle	$C_1 \lhd C_0$	Likelihood of the angle between two matched 2D glint candidates (6).
		η_μ variance angle	$1\,(^\circ)^2$	
	σ_{angle}	η_σ mean angle	$\propto$ RMS Error	
		η_σ variance angle	$1\,(^\circ)^2$	
L_{dist}	μ_{dist}	η_μ mean dist	$\frac{v_{\text{max}} + v_{\text{min}}}{2}$	Likelihood of the distance between two matched 2D glint candidates (9).
		η_μ variance angle	$1\,\text{px}^2$	
	σ_{dist}	η_σ mean dist	$\frac{\lVert v_{\text{max}} - v_{\text{min}} \rVert^2}{\log(256)}$	
		η_σ variance dist	$1\,\text{px}^2$	
L_{IPD}	μ_{IPD}	η_μ mean IPD	$63\,\text{mm}$	Likelihood of the distance between a 3D glint pair (11).
		η_μ variance IPD	$1\,\text{mm}^2$	
	σ_{IPD}	η_σ mean IPD	$11.04\,\text{mm}$	
		η_σ variance IPD	$1\,\text{mm}^2$	
L_{yaw}	μ_{yaw}	η_μ mean yaw	0°	Likelihood of the vertical rotation of a 3D glint pair (13).
		η_μ variance yaw	$1\,(^\circ)^2$	
	σ_{yaw}	η_σ mean yaw	4.25°	
		η_σ variance yaw	$1\,(^\circ)^2$	
L_{roll}	μ_{roll}	η_μ mean roll	0°	Likelihood of the lateral inclination of a 3D glint pair (13).
		η_μ variance roll	$1\,(^\circ)^2$	
	σ_{roll}	η_σ mean roll	4.25°	
		η_σ variance roll	$1\,(^\circ)^2$	
L_{remote}	μ_{remote}	η_μ mean remote	$1000\,\text{mm}$	Likelihood of the distance from the 3D glint pair to the camera system (15).
		η_μ variance remote	$1\,\text{mm}^2$	
	σ_{remote}	η_σ mean remote	$167.86\,\text{mm}$	
		η_σ variance remote	$1\,\text{mm}^2$	

TABLE I: Overview of all adapted parameters η and the suggested initial value η_0.

Table I provides an overview of all optimized hyperparameters η and a recommendation for the initial state η_0.

c) Implementation: The process presented here has been integrated into Caffe [18]. We added some layers to the Caffe framework to implement the steps ②-⑥. While in the forward propagation, we extract the glint candidates and calculate their likelihoods. Then, we use the back propagation to calculate the gradient $\delta\eta$ and then optimize the hyperparameters η with the SGD solver. In addition to the layers for implementing the steps ②-⑥, we also created layers for the confidence (also used as loss), evaluation, debugging, and writing journal files.

d) Combination with existing approaches: The seperation of the individual steps into different Caffe layers allows them to be easily replaced by other approaches. In addition to the feature extraction outlined in step ② for highlighting glint candidates, we also provide various methods which we have found in the literature. We modified some of the methods slightly in order to optimize their behaviour by adapting their parameters in the back propagation.

(Adaptive) Threshold: The majority of existing approaches require suitable lighting conditions and can therefore extract the glint positions based on fixed thresholds [29], [33], [36]. In other approaches, adaptive thresholds can be found, which take into account the average intensity in the surrounding area and are, therefore, more robust to varying illumination levels [32]. We implemented the adaptive threshold as following:

$$m_{xy} = \begin{cases} c_{xy} & \text{if } c_{xy} - \hat{c}_{xy} > t_{\text{adaptive threshold}} \\ 0 & \text{otherwise} \end{cases}, \tag{28}$$

where $c_{x,y} \in I$ is the intensity, and $\hat{c}_{x,y} \in \hat{I}$ is the mean intensity of the input image at position (x, y). The threshold $t_{\text{adaptive threshold}} \in \eta$ is the minimum difference that must result in the intensity with respect to the mean intensity to not be suppressed. For $\hat{c}_{xy} = 0$ this equals the use of a fixed threshold.

Ebisawa: Ebisawa et. al. use first an adaptive threshold as in the above Equation (28) to extract temporary glint positions. The position of the glints is determined then more precisely by calculating a gravitational center of the glint candidates within a certain window [10].

Canny, LoG, Sobel: Sharma et. al. evaluated the Laplacian of Gaussian (LoG), Sobel and Canny edge detector with regard to their applicability to glint detection [27]:

$$m_{xy} = \begin{cases} c_{xy} & \text{if } \mathrm{edge}_{xy}(I, t_{\text{edge sensitivty}}) \\ 0 & \text{otherwise} \end{cases}, \tag{29}$$

where $c_{x,y} \in I$ is the intensity of the input image. $t_{\text{edge sensitivty}} \in \eta$ is the threshold of the magnitude, in order to be counted as an edge (for Canny it is the upper Threshold of the hysteresis).

In order to weight glints which are located in a relatively poorly illuminated region more strongly, we extended all the extractors by a spatial normalization of the output. This process increases the detection rate of weakly pronounced glints significantly, but also increases the noise.

III. Dataset

We collected a dataset consisting of 6993 labeled IR stereo images (1152 × 1536 pixel at 10 Hz). The images were recorded by two planar arranged cameras with a vertical and horizontal offset to each other of 150mm and 20mm. Overall, the dataset provides 7 sequences (DS01, . . . , DS07), where each sequence consists of 999 stereo frames (99.9 s) collected from four different subjects. All sequences, except for DS0, contain at least one of four different types of lighting interference, which makes glint detection on these images very challenging. Figure 4 and 3 provide an example of each challenge and an overview on which of those are present in the respective sequences.

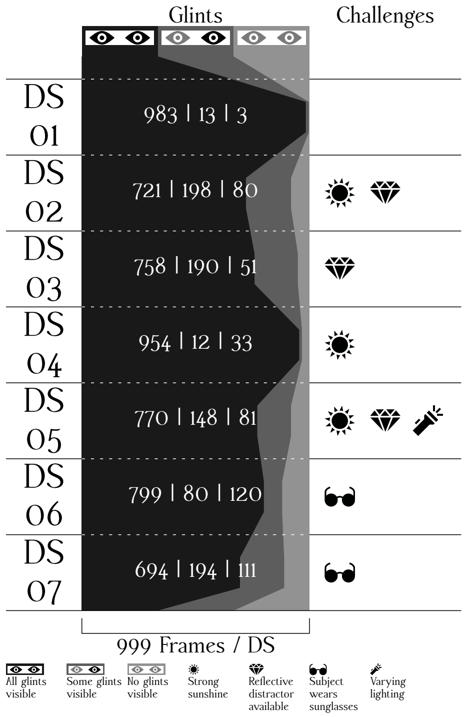


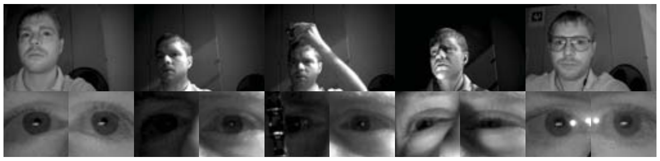
Fig. 3: Challenges addressed by our dataset

The sequence DS01 is considered the most easiest one since it presents a favorable setting for glint detection. In contrast, the sequences DS02, DS04, and DS05 contain strong illumination gradients due to sunshine, which additionally causes reflections on hair, jewelry, buttons, and other reflecting parts in the scene. In DS02, DS03, and DS05 additional reflecting objects were placed in the scene, which would cause additional disturbances and possibly even cover the actual glints. In the sequences DS06 and DS07, the subjects wear sunglasses. This causes additional reflection around the eye, which sometimes overlay the original glint. DS05 provides the most challenging lighting conditions since a strong infrared illumination source was moving around the subject.

The intrinsic parameters were estimated using Matlab's camera calibration toolbox and a 10 × 8 checkerboard. We labeled all visible glints in the stereo images and triangulated those, thus we obtained the glint position in 2D image coordinates as well as in a 3D stereo camera coordinates. We will provide the dataset with the labels and calibration for download under: ftp://peg-public:peg-public@messor.informatik.uni-tuebingen.de/glintdb.zip.

IV. Evaluation

We evaluated our glint detection approach on the hand-labeled data set introduced above. More specifically, we



Good illumination ☀ Strong sunshine 💎 Distractors 🔦 Illumination 🕶 Sunglasses

Fig. 4: Our dataset comes with various challenges. The first frame shows an example of favorable conditions for glint detection. The next image contains a huge light gradient caused by lateral sun radiation. Thereby, the intensity of the right glint is suppressed. In the third frame, an additional distractor in front of the right eye causes additional reflections next to the glint. The 4th example shows the impact of a further varying IR light source. In the last frame the subject wears sunglasses. IR is not blocked by the glasses so the glints are well visible, but causes additional reflexions close to the eye.

	OpenFace		FAST2		Haar Cascade	
	Accuracy	F_1 score	Accuracy	F_1 score	Accuracy	F_1 score
DS01	0.862	0.926	**0.989**	**0.995**	0.446	0.617
DS02	**0.642**	**0.782**	0.609	0.757	0.050	0.094
DS03	0.586	0.739	**0.834**	**0.910**	0.499	0.666
DS04	**0.987**	**0.994**	0.887	0.940	0.026	0.051
DS05	0.697	0.822	**0.809**	**0.894**	0.086	0.158
DS06	**0.946**	**0.972**	0.901	0.948	0.064	0.120
DS07	**0.724**	**0.840**	0.716	0.835	0.019	0.037
AVG	0.777	0.869	**0.821**	**0.897**	0.170	0.256
All	0.790	0.883	**0.833**	**0.909**	0.177	0.301

Accuracy F_1 score

TABLE II: Evaluation results of our approach used as an eye detector compared to the two state of the art methods OpenFace and Haar Cascade.

compared the performance of various glint extractors from the state-of-the-art with our proposed FAST2 method (Figure 5). To the best of our knowledge, there are no comparable methods for selecting glint candidates in stereo images. Therefore, we compared our approach with two state-of-the-art methods for the detection of faces and eyes (Table II).

a) Glint Extractors: Figure 5 shows the evaluation of our glint detection using different methods from the state-of-the-art (Adaptive Threshold, Canny, Ebisawa, LoG Sobel) and our proposed FAST2 for feature extraction. With the exception of the Canny Edge Detector, all methods achieve predominantly good results. Our proposed extractor FAST2 is characterized by an overall very low false positive rate, which can be traced back to the implicit limitation of the size of possible glints (r). The dataset DS01 was detected almost without errors. The challenging illumination conditions present in the dataset DS02, DS04 and DS05 does not seem to interfere with the performance of our algorithm. For DS07, the sunglasses worn by the subject induces additional reflections along the frame and on the lenses. As a result, a large number of glint candidates were detected around the eye, thus the actual glint was suppressed due to the distance condition to the previously selected glints in step ③, which is finally reflected in a higher false positive rate.

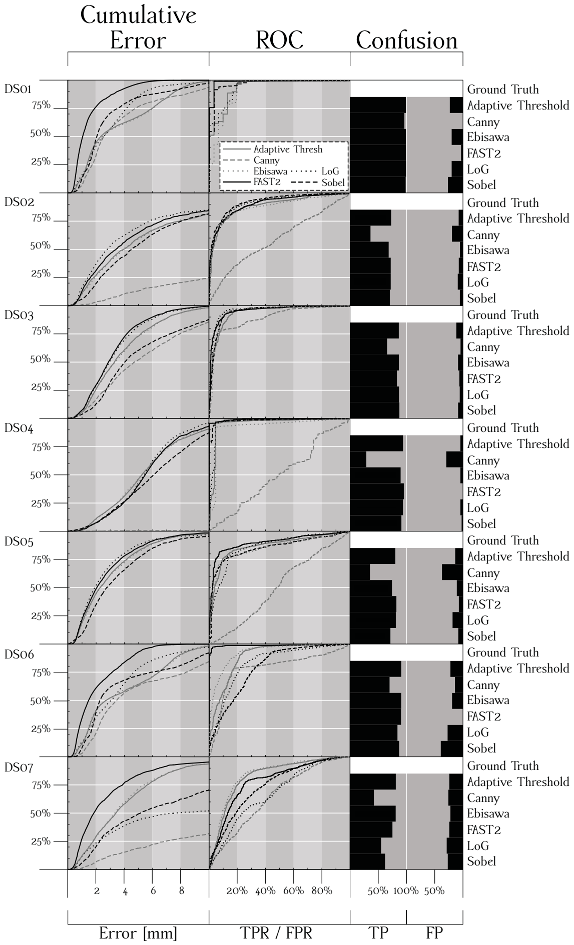


Fig. 5: Evaluation results for the different feature extractors used in FAST2. For glint selection, we used an adaptation rate of 0.15 and a back propagation after each forward step. The first column shows the detection rate on the vertical axis relative to a maximum permissible error in the horizontal axis. The second column contains the Receiver Operating Characteristic (ROC). The last column lists the relative proportion of correctly and incorrectly classified glints for $t = 0.5$.

b) Eye Detection: Table II shows results of a comparison between our method based on the FAST2 feature extraction and OpenFace [2] (for facial recognition) and a cross validated Haar Cascade. We applied OpenFace and the Haar Cascade separately on both stereo images and determined the center of the eye. Next, we triangulated these points and compared to the labeled points from our dataset. Since the labeled glint points do not necessarily coincide with the center of the eye, we allowed an error up to $2\,\mathrm{cm}$. In addition, we removed all images from the dataset where the eyes could be seen but no glints were present and therefore, not labeled. Our presented method achieves a performance similar to that of OpenFace and partly outperforms it. The performance of OpenFace collapses, especially on the dataset DS03 and DS05. For these datasets, the tracking in OpenFace partially interrupts; for DS03 the head of the subject is not always completely visible in both stereo images, whereas for DS05 the performance decreases due to varying illumination. On average, we achieve an accuracy of $82.1\,\%$ and $83.3\,\%$ on the whole dataset. The Haar Cascade shows very poor performance, possibly due to the unique lighting conditions in each dataset and the susceptibility of Haar-Like features to inconsistent lighting [24]. In addition, the detected bounding boxes are often inaccurate, thus after triangulation, the maximum permissible error of $2\,\mathrm{cm}$ is exceeded. Only two data sets, namely DS01 and DS03, which contained neither challenging lighting nor sunglasses, were partially detected correctly.

V. Conclusions

We presented an adaptive glint detection approach for remote eye-tracking in the wild based on a probabilistic model. Our evaluation on more that 10.000 hand-labeled data showed that our algorithms significantly improves the state-of-the-art and can cope with various challenges arising in real-world settings, such as varying illumination conditions, or reflections. This way, we have provided a strong basis for the parametrization of eye models in remote eye tracking. In future work, we will integrate facial landmarks into the probabilistic model, which will serve as a basis for a later eye model adaption and gaze prediction. This approach provides robust gaze estimation in open ended settings, providing an additional stream of information that can greatly improve a robot's ability to interact with people in collaborative task settings.

References

[1] M. Argyle. Non-verbal communication in human social interaction. 1972.

[2] T. Baltrusaitis, P. Robinson, and L.-P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 354–361, 2013.

[3] L. Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

[4] C. Braunagel, E. Kasneci, W. Stolzmann, and W. Rosenstiel. Driver-activity recognition in the context of conditionally autonomous driving. In *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, pages 1652–1657. IEEE, 2015.

[5] M. R. Clark. A two-dimensional purkinje eye tracker. *Behavior Research Methods*, 7(2):215–219, 1975.

[6] T. N. Cornsweet and H. D. Crane. Accurate two-dimensional eye tracker using first and fourth purkinje images. *JOSA*, 63(8):921–928, 1973.

[7] H. D. Crane and C. M. Steele. Generation-v dual-purkinje-image eyetracker. *Applied Optics*, 24(4):527–537, 1985.

[8] P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *The Annals of statistics*, pages 269–281, 1979.

[9] N. A. Dodgson et al. Variation and extrema of human interpupillary distance. In *Proceedings of SPIE*, volume 5291, pages 36–46, 2004.

[10] Y. Ebisawa, S. Tsukahara, and D. Ishima. Detection of feature points in video-based eye-gaze detection. In *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*, volume 2, pages 1764–1765. IEEE, 2002.

[11] N. J. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, 24(6):581–604, 2000.

[12] M. M. Farid and F. D. Murtagh. Eye-movements and voice as interface modalities to computer systems. *Opto-Ireland*, pages 115–125, 2002.

[13] W. Fuhl, D. Geisler, T. Santini, and E. Kasneci. Evaluation of state-of-the-art pupil detection algorithms on remote eye images. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct publication – PETMEI 2016*, 09 2016.

[14] W. Fuhl, M. Tonsen, A. Bulling, and E. Kasneci. Pupil detection for head-mounted eye tracking in the wild: An evaluation of the state of the art. *Springer Machine Vision and Applications*, pages 1–14, 2016.

[15] R. Hartley and P. Sturm. Triangulation. In *Computer analysis of images and patterns*, pages 190–197. Springer, 1995.

[16] R. Hartley and A. Zisserman. Multiple view geometry in computer vision second edition. *Cambridge University Press*, 2000.

[17] C. Hennessey, B. Noureddin, and P. Lawrence. A single camera eye-gaze tracking system with free head motion. In *Proceedings of the 2006 symposium on Eye tracking research & applications*, pages 87–94. ACM, 2006.

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[19] J. Joyce. Bayes' theorem. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

[20] T. Kübler, S. Eivazi, and E. Kasneci. Automated visual scanpath analysis reveals the expertise level of micro-neurosurgeons. In *MICCAI Workshop on Interventional Microscopy*, 2015.

[21] T. C. Kübler, C. Rothe, U. Schiefer, W. Rosenstiel, and E. Kasneci. Subsmatch 2.0: Scanpath comparison and classification based on subsequence frequencies. *Behavior research methods*, 49(3):1048–1064, 2017.

[22] P. Lindstrom. Triangulation made easy. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1554–1561. IEEE, 2010.

[23] Y. Matsumotot, T. Ino, and T. Ogsawara. Development of intelligent wheelchair system with face and gaze based interface. In *Robot and Human Interactive Communication, 2001. Proceedings. 10th IEEE International Workshop on*, pages 262–267. IEEE, 2001.

[24] M. Rezaei and R. Klette. Adaptive haar-like classifier for eye status detection under non-ideal lighting conditions. In *Proceedings of the 27th Conference on Image and Vision Computing New Zealand*, pages 521–526. ACM, 2012.

[25] T. Santini, W. Fuhl, and E. Kasneci. Calibme: Fast and unsupervised eye tracker calibration for gaze-based pervasive human-computer interaction. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2594–2605. ACM, 2017.

[26] B. Scassellati. Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. In *Computation for metaphors, analogy, and agents*, pages 176–195. Springer, 1999.

[27] A. Sharma and P. Abrol. Glint detection and evaluation using edge detectors. *International Journal of Scientific and Technical Advancements (IJSTA)*, 1(3):319–323, 2015.

[28] A. Sharma and P. Abrol. Direction estimation model for gaze controlled systems. *Journal of Eye Movement Research*, 9(6), 2016.

[29] M. Stengel, S. Grogorick, M. Eisemann, E. Eisemann, and M. A. Magnor. An affordable solution for binocular eye tracking and calibration in head-mounted displays. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 15–24. ACM, 2015.

[30] L. Swirski and N. Dodgson. A fully-automatic, temporal approach to single camera, glint-free 3d eye model fitting. *Proc. PETMEI*, 2013.

[31] E. Wood and A. Bulling. Eyetab: Model-based gaze estimation on unmodified tablet computers. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 207–210. ACM, 2014.

[32] X. Yang, J. Sun, J. Liu, J. Chu, W. Liu, and Y. Gao. A gaze tracking scheme for eye-based intelligent control. In *Intelligent Control and Automation (WCICA), 2010 8th World Congress on*, pages 50–55. IEEE, 2010.

[33] D. H. Yoo and M. J. Chung. Non-intrusive eye gaze estimation without knowledge of eye pose. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 785–790. IEEE, 2004.

[34] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2015.

[35] X. Zhou, H. Cai, Y. Li, and H. Liu. Two-eye model-based gaze estimation from a kinect sensor. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 1646–1653. IEEE, 2017.

[36] Z. Zhu and Q. Ji. Eye and gaze tracking for interactive graphic display. *Machine Vision and Applications*, 15(3):139–148, 2004.