Estimating 3D Gaze Directions Using Unlabeled Eye images via Synthetic Iris Appearance Fitting

Feng Lu Member, IEEE, Yue Gao, Senior Member, IEEE, Xiaowu Chen, Senior Member, IEEE

Abstract—Estimating 3D human eye gaze by capturing a single eye image without active illumination is challenging. Although the elliptical iris shape provides a useful cue, existing methods face difficulties in ellipse fitting due to unreliable iris contour detection. These methods may fail frequently especially with low resolution eye images. In this paper, we propose a Synthetic Iris Appearance Fitting (SIAF) method that is model-driven to compute 3D gaze direction from iris shape. Instead of fitting an ellipse based on exactly detected iris contour, our method first synthesizes a set of physically possible iris appearances and then optimizes inside this synthetic space to find the best solution to explain the captured eye image. In this way, the solution is highly constrained and guaranteed to be physically feasible. In addition, the proposed advanced image analysis techniques also help the SIAF method be robust to the unreliable iris contour detection. Furthermore, with multiple eye images, we propose a SIAF-joint method that can further reduce the gaze error by half, and it also resolves the binary ambiguity which is inevitable in conventional methods based on simple ellipse fitting.

Index Terms—3D human gaze, gaze estimation, iris fitting, unlabeled eye images

I. INTRODUCTION

HUMAN gaze tracking is important for intelligent systems to understand user behavior and conduct effective interactions, due to the fact that more than 80% percent of sensory information is received by human eyes, and more than 90% sensory information is processed by human visual systems. Therefore, human eye gaze becomes an essential cue to reflect how we interact with the outside world. For instance, eye gaze has shown a strong relation with the visual saliency of the real world [1], [2]. Recently, with the fast development of techniques, such as virtual reality (VR) and human-computer interaction (HCI), gaze tracking has been found very useful in more and more practical applications [3], [4], [5], [6].

On the other hand, despite the recent research progress in the field of computer vision, estimating human gaze directions from eye appearance only is still challenging [7], [8]. The most widely employed methods are usually based on active illumination and infrared imaging, which require additional hardware and controlled environment. Some other methods [9], [10] use a single camera to capture eye images under passive illumination. However, they rely on sufficient training samples captured beforehand to learn the gaze regression. Recently, a few works use other types of prior knowledge to relax the assumption on training. They use depth information or visual saliency map of the scene to reduce the training cost [11], [12]. However, the newly introduced assumptions, i.e., depth and saliency, are not always available in practice.

In this paper, we consider estimating 3D human eye gaze purely from unlabeled eye images captured without active illumination. This problem is obviously challenging since we do not assume any of the above information required by previous methods; while for the same reason, the resulting technique will be highly practical. To achieve this goal, considering a single eye image as the input, the only cue to infer eye gaze direction is the shape of the iris contour. In particular, by assuming circular iris contour, we can analyze its 2D projection in the image, which always takes the shape of an ellipse, to recover its 3D orientation in real world. This is the common idea of iris-shape-based methods [13], [14]. However, existing methods along this direction face a major problem that detecting the tiny iris contour is always unreliable in practice. Since ellipse fitting is very sensitive to iris contour’s accuracy, existing methods have to use high resolution eye images [15] or wearable cameras [16] to deal with individuality in eye shape, the existence of eyelids and complex corneal reflections. This again increases the requirements on hardware and reduces the system practicability in some scenarios.

We propose a novel iris-appearance-based method, namely Synthetic Iris Appearance Fitting (SIAF), to compute gaze direction from common eye images. Our focus is on the robustness of the iris shape analysis, in the case that eye images are captured from a remote camera at common resolutions; no other system requirements are required. To this end, we propose a model-driven method that first synthesizes a set of virtual iris appearances to construct the solution space, and then searches the best solution to explain the really captured iris appearance. Because our synthetic data are constrained by physically possible iris shape and size, the solution is ensured to be physically valid, even if the captured eye image is noisy. This is essentially different from previous methods which fit the detected iris contour to an ellipse in an unconstrained way and thus are highly sensitive to image noise.

Our solution also benefits from the proposed advanced image analyses, including polar coordinate analysis, image symmetry, image gradient map and robust distance measurements. Therefore, outliers in iris contour detection can be
further tolerated and the proposed SIAF is possible to work with low resolution images robustly.

Furthermore, we propose a SIAF-joint method in the case that a set of aligned eye images are available. This can be satisfied if we take multiple shots of the eye incessantly before the head pose changes. Based on their individual gaze estimation results using SIAF, the SIAF-joint method performs a joint optimization to refine these results. The key is to assume the temporal consistency of their eyeball center positions, which helps further reduce the solution space. Experimental results show that the proposed SIAF-joint method reduces the average gaze error by half. It also resolves the binary ambiguity which is inevitable in determining an ellipse’s orientation.

Overall, the contributions of this work include: 1) an eye model-driven method SIAF based on synthetic iris data to compute 3D gaze direction from an unlabeled eye image; 2) techniques for SIAF to perform robust iris appearance analysis and fitting; 3) a SIAF-joint method to further refine the gaze results using multiple eye images and reduce the error by half.

II. RELATED WORK

There are many previous techniques proposed for human gaze estimation. Although a few of them only consider head poses [17], most accurate methods are designed to analyze eye images. According to recent surveys [7], [8], they can be roughly categorized into two types: model-based and appearance-based. However, a number of recent works exploit both model-based and appearance-based priors to compute human gaze directions. Therefore, we alternatively introduce another way for categorization. That is, we consider any gaze estimation method as either an active method or a passive method.

An active method generates certain features on the eyeball surface actively and uses them to compute eyeball orientations and gaze directions geometrically. The most commonly used eye feature is near infrared (NIR) corneal reflections and pupil reflections [18], [19], [20]. To produce such reflections, active NIR imaging is necessary, which requires multiple NIR light sources and cameras with calibrated positions. Then, the NIR features are used to fit certain 3D eyeball models [19], [21], [22] or cross ratio models [18], [23], [24] to compute the eyeball orientations and gaze directions. Such systems are usually expensive on the market, and their requirements on additional hardware, e.g., IR lights, stereo cameras, and pan-tilt/high-definition cameras [25], [26], [27] prevent their usage with common devices such as a mobile phone or tablet.

Recently, methods based on depth sensor, e.g., Kinect device, have been proposed [28], [11], [29]. The depth sensor measures depth information of the scene, which is human face geometry in their cases, in an active way, and therefore these methods are also categorized into active methods.

On the other hand, a passive method only captures and uses natural eye images without active illumination, and in most cases, only one ordinary camera is required. For gaze estimation, except some feature-based methods [30], most conventional methods simply learn a regression between the image vector and gaze position through training. Early systems reported by Baluja and Pomerleau [31] and by Xu et al. [32] were based on neural networks which were trained by using several thousands of training samples. Alternatively, Tan et al. [9] proposed a local linear interpolation-based method to exploit the relationship between the eye image space and screen coordinates. To train their system, the number of required training samples was reduced to around two hundreds. In order to use fewer training samples, Williams et al. [33] proposed a semi-supervised method that can use both labeled and unlabeled samples for training. Lu et al. [34] introduced an adaptive regression framework based on sparse optimization to use even fewer training samples while it can also solve other problems in gaze estimation [10]. Sugano et al. [12] extracted visual saliency from video clips to automatically prepare training samples. Lu et al. [35] introduce patch-based features to reduce person-dependence in training. While the above methods assume a fixed head pose, there are also methods that work for different head poses. However, they need much more training samples to allow head motion [36], [37].

In order to eventually avoid any training before gaze estimation, a few methods were proposed to analyze the eye appearance in a more analytical way. The key is that human gaze direction is uniquely determined by eyeball orientation, which can be seen from the iris orientation/position. By this means, Yamazoe et al. [38] and Heyman et al. [39] proposed gaze tracking techniques by recovering the relative position between iris and eyeball center. Their methods need accurate face tracking/head modeling to obtain multiple 3D face landmarks such as eye corners and (invisible) eyeball centers, which rely on more information than 2D eye appearance. The idea of face landmark tracking was also used by Ishikawa et al. [40] to propose an AAM-based method. Some other methods [13], [14] fit an ellipse to iris contour, and back project the ellipse into a circle in the 3D space to obtain its orientation. These methods are advantageous since they only use 2D eye images and do not need the whole face geometry, however accurate detection of iris contours is usually unreliable due to image noise. To overcome such a difficulty, high resolution eye images [15] and wearable cameras [16] can be used. Overall, computing gaze direction only from eye images is still challenging.

III. IRIS APPEARANCE AND 3D GAZE DIRECTION

A. Human eyeball structure

The human eye is incredibly complex, and it is among the most complex organs existed. The average human eye only weights about 7 grams, but it contains many structures such as cornea, iris, lens, retina and so on. For human gaze estimation from eye appearance, many previous works need to use precise 3D eye models to accurately detect reflection points on different layers, i.e., cornea and pupil; while in our method, we only use a very simple 3D eye model. As shown in Fig. 1, in our model, we consider a spherical eyeball, with a circular iris contour on its surface. The gaze direction is assumed to pass through both the eyeball center and the iris center. In this simple model, the only parameters to know are the diameters of the eyeball and iris contour.
Based on the 3D eyeball model in Fig. 1, it is possible to infer eye gaze direction from the captured eye image. Since the gaze direction passes through both the eyeball center and iris center, it coincides with the surface normal of the circular region defined by the iris contour. Therefore, the problem is about estimating the 3D surface normal, \( \mathbf{g} \), or, the orientation, of iris circle form 2D image.

To do this, one can easily prove that when changing the orientation of the iris circle in 3D, the resulting 2D projection will be an ellipse with variant shapes. In particular, if the iris circle ideally faces the 2D image plane, the resulting 2D shape is also a circle; if the iris circle is rotated, the projection will become an ellipse that shrinks in its short axis. In this way, by examining the shape of the captured 2D iris contour, it is possible to recover its orientation in 3D.

We show this in Fig. 2, where the iris contour in a captured 2D eye image is approximated by an ellipse. Then, the 2D projection of the 3D normal vector should coincide with its short axis. In addition, let \( r_x \) and \( r_y \) denote the lengths of the long axis and short axis; the normal vector's elevation angle \( \theta \) w.r.t. the image plane is computed by

\[
\theta = \arccos \frac{r_y}{r_x},
\]

(1)

In this section, we propose a robust method to compute 3D eye gaze direction from the 2D iris appearance in a single eye image. The key idea is to develop a model-driven method; it synthesizes virtual iris appearances for many possible gaze directions, and then chooses the best one to explain the real eye image. This method is advantageous over conventional ellipse fitting-based methods for two reasons. First, the detected iris contour is no longer the only cue since we propose and use advanced image analysis techniques. Second, the model-driven method constrains the possible eye parameters and gaze directions and thus ensures physically feasible solutions.

A. Iris contour data synthesis

Our method is model-driven in that it synthesizes virtual iris appearances for possible gaze directions before gaze estimation. To this end, we need to model the eyeball in 3D and also obtain the camera configuration. For the 3D eyeball modeling, as introduced in Sec. III-A, we use a simple model comprises a pure sphere (eyeball) and a circle (iris). The diameters of them are denoted by \( D_e \) and \( D_i \) and set to 25 mm and 12 mm according to [41]. These values correspond to a typical human eyeball while our method can tolerate their variations due to individuality.

By using this simple eyeball model, we generate different gaze directions via eyeball rotations. As shown in Fig. 3, let \( g \in \mathbb{R}^{3 \times 1} \) denote a 3D gaze vector, and let \( \mathbf{e}(g) \in \mathbb{R}^{3 \times 1} \),
c(g) ∈ R^{3×1} and \{p_i(g) ∈ R^{3×1}\} denote the corresponding eyeball center, iris center and points on the iris contour. Then, if we fix e(g), c(g) and \{p_i(g)\} can be analytically determined by 3D gaze direction g as follows:

\[
c(g) = \begin{bmatrix}
1 & 0 & 0 \\
0 & \cos \beta & -\sin \beta \\
0 & \sin \beta & \cos \beta
\end{bmatrix}
\begin{bmatrix}
\cos \alpha & 0 & \sin \alpha \\
0 & 1 & 0 \\
-\sin \alpha & 0 & \cos \alpha
\end{bmatrix}
\begin{bmatrix}
\gamma \\
\sin \gamma \\
\cos \gamma
\end{bmatrix}
\]

(3)

\[
p_i(g) = \begin{bmatrix}
1 & 0 & 0 \\
0 & \cos \beta & -\sin \beta \\
0 & \sin \beta & \cos \beta
\end{bmatrix}
\begin{bmatrix}
\cos \alpha & 0 & \sin \alpha \\
0 & 1 & 0 \\
-\sin \alpha & 0 & \cos \alpha
\end{bmatrix}
\begin{bmatrix}
D_i/2 \\
D_i/2 \cdot \sin \gamma \\
(D_i^2/4 - D_i^2/4 - 2)^{-1}
\end{bmatrix}
\]

(4)

where \( g = [g_x, g_y, g_z]^T \), \( \alpha = \arctan(g_x/g_z) \), \( \beta = \arctan(g_y/g_z) \), and \( \gamma \) is the azimuth angle of the i-th iris contour point. By using Eq. (3) and Eq. (4), we can synthesize a set of possible 3D eye appearances, which are described by their 3D iris positions c(g_n), \{p_i(g_n)\} and eyeball centers e(g_n) for a set of given gaze directions \( \{g_n\} \). To choose the set of \( \{g_n\} \), we sample in the gaze direction space uniformly as following:

\[
g_n = \begin{bmatrix}
\sin u \\
\sin v \\
(1 - \sin^2 u - \sin^2 v)^{-1}
\end{bmatrix}
\]

(5)

s.t. \( u, v \in -60^\circ \sim 60^\circ, (\sin^2 u + \sin^2 v)^{-2} < \sin 45^\circ \), where we sample over angles \( u \) and \( v \) in an interval of 5°, and limit the maximum gaze bias by 45°. This method generates a dataset to cover possible gaze directions in practice, and then we use them to produce iris appearance data via Eq. (3) and Eq. (4).

The next step is to convert the obtained 3D iris data into pixel positions in the image plane. To do this, a straightforward way is to use the camera formula as following:

\[
z \begin{bmatrix}
p \\
p_1
\end{bmatrix} = K \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
p' \\
p_1
\end{bmatrix},
\]

(6)

where \( P \) is a 3D point (can be any of the \( e(g) \), \( c(g) \) and \( p_i(g) \) above) and \( P' \) is the corresponding pixel position in the image. Besides, \( K \) is the intrinsic matrix of the camera. By using Eq. (6), we can convert \( e(g) \), \( c(g) \) and \( p_i(g) \) into 2D image pixel positions which we will eventually use for gaze estimation.

Note that the above synthesis requires to know the camera intrinsic matrix \( K \) and the 3D eyeball center position \( e(g) \). In practice, \( K \) can be obtained via camera calibration beforehand, and \( e(g) \) can be estimated by using a single camera-based face tracker [42]. However, there are cases where we do not have the particular data of \( K \) and \( e(g) \), e.g., when we only have a single eye image captured elsewhere. In such cases, we can alternatively use an approximation method to convert 3D iris points to 2D. First, we set \( e(g) \) to 0 since we do not know it. Then, after computing via Eq. (3) and Eq. (4), we apply a simple scaling transformation

\[
P = s \begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0
\end{bmatrix} P',
\]

(7)

where the scale factor \( s \) is determined by first roughly estimating the iris diameter in the eye image and then dividing it by \( D_i \).

By using either Eq. (6) or Eq. (7), we obtain 2D pixel positions of synthetic 3D iris data, including iris center \( e(g_n) \), iris contour points \( \{p_i(g_n)\} \) and eyeball center \( e(g_n) \), for all possible gaze directions \( \{g_n\} \). Hereafter, we still use these notations to indicate the 2D pixel positions for convenience.

B. Eye image pre-processing

Besides iris data synthesis, we also pre-process the captured eye images before gaze estimation. The goal is to obtain the robust iris contour information from the image. Some previous works, e.g., [13], also involve image processing; while our method introduces additional steps for better robustness.

Given an input eye image (Fig. 4(a)), we first do histogram modification to enhance the contrast between the brightness regions, e.g., sclera, and the dark regions, e.g., iris and cornea (Fig. 4(b)). This is done by

\[
I_k' = \begin{cases}
I_k, & \text{if } I_k < \text{median}(I_k), \\
\text{median}(I_k), & \text{otherwise},
\end{cases}
\]

and then

\[
I_k'' = 255 \times \frac{I_k' - \min_k(I_k')}{\max_k(I_k') - \min_k(I_k')}.
\]

Fig. 3: Iris data synthesis for all possible gaze directions \( \{g_n\} \). By fixing the 3D eyeball center \( e(g_n) \), the 3D iris center \( c(g_n) \) and iris contour points \( \{p_i(g_n)\} \) can be geometrically computed and projected onto the 2D image plane.

Fig. 4: Eye image pre-processing. (a) original image; (b) histogram modification; (c) darkest regions; (d) iris region after dilation; (e) orientation mask; (f) canny edge map; (g) iris contour candidates; (h) extracted iris contour.
where $I_k$ and $I_k'$ are values of an image pixel before and after histogram modification.

In the modified image, we then find the darkest regions by using a threshold level of 0.8 (Fig. 4(c)). We select the largest region from the results and enlarge it by using the image dilation operation. The final region (Fig. 4(d)) is considered to be the mask of the iris region and its center position is denoted as $o$. At the same time, we compute an orientation mask (Fig. 4(e)) from the histogram-modified image by

$$M_k = \begin{cases} 1, & \text{if } \text{angle(} \text{grad}(I_k), \text{vec}(o \rightarrow \text{position}(k)) \text{)} < 30^\circ, \\ 0, & \text{otherwise}, \end{cases}$$

meaning that the mask covers potential iris contour pixels, based on the observation that the pixel gradient direction (from dark iris region to bright sclera region) on the iris contour should point outward away from iris center. Therefore, for a pixel $k$, if the image gradient direction on $k$ and the direction from $o$ to the pixel position of $k$ is similar, we consider it a potential iris contour pixel. Note that in Fig. 4(e) we only keep the mask in two fan-shaped regions in order to avoid the affect by eyelids.

Next, we compute the canny edge map (Fig. 4(f)) which contains the iris contour and many other edges. After masking this edge map by using the iris region mask (Fig. 4(d)) and the orientation mask (Fig. 4(e)), we obtain the iris contour candidates (Fig. 4(g)). Note that there may exist isolated pixels due to incomplete masking, and thus we remove small pieces of edges with only a few pixels. The resulting final iris contour can be seen in (Fig. 4(h)).

C. Iris appearance fitting and the cost function

In this section, we use the synthetic iris appearance data (Sec. IV-A) and the pre-processed eye image and iris contour (Sec. IV-B) to determine the unknown 3D gaze direction. In particular, we optimize over different solutions in the synthetic space, containing combinations of eyeball positions and eyeball orientations (gaze) used for synthesis, to find the one that best explains the captured eye image and iris appearance. Note that the synthetic and real iris may be a little different in their sizes due to individuality and camera calibration. Therefore, we propose robust image analysis techniques to handle such a difference as shown in Fig. 5. In particular, the proposed fitting function comprises two terms: one for radial symmetry and the other for iris contour fidelity.

Radial symmetry: we propose radial symmetry on pixel values near iris contour. It is based on the observation that the iris region is radially symmetric, and thus if we examine the pixel values across the iris contour from the dark iris region to the bright sclera region, they should vary in a similar pattern at different iris contour positions. Therefore, as shown in Fig. 5, f if the iris position is accurate, pixel value variation perpendicular to the synthetic iris contour at each point $p_i(g_n)$ should be the same.

To measure the radial symmetry, we compute the Polar Gradient Map (PGM) of the eye image region near the synthetic iris contour. In particular, we process the image under the polar coordinate system. The horizontal axis of the PGM indicates the angular coordinate and the vertical axis indicates the radial coordinate. Values in the PGM is collected as follows: first, for each synthetic iris point $p_i(g_n)$, we compute the azimuth angle of vector $c(g_n) \rightarrow p_i(g_n)$. This angular value determines a column index in PGM. Second, along the direction of $c(g_n) \rightarrow p_i(g_n)$, we continuously sample a set of pixel values from inside to outside of the synthetic iris contour. These pixel values are filled in the previous column of the PGM. Note that in order to avoid any affect by eyelids, we do not collect data for any $p_i(g_n)$ that lies in the top or bottom fan-shaped region as shown in the left plots of Fig. 5. After filling all columns of PGM, we compute its vertical gradient map as the final result.

Examples of PGMs are shown in Fig. 5, where a good fit leads to a PGM with large intensities in only a few rows. These rows indicate similar brightness changes across different iris contour points. A bad fit cannot show such a consistency and has scattered large intensities in the PGM. This allows us to distinguish good fits from bad fits. In particular, we compute the following concentration function to measure the intensity distribution in PGM:

$$RS = \frac{\sigma_1}{\sum(\sigma_i)}, \text{ s.t. } PGM = U \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \end{bmatrix} V_*,$$

where $PGM$ is a matrix and $\{\delta_i\}$ are its singular values. A larger RS score in Eq. (11) suggests a better fit. Eq. (11) is computed based on comprehensive information in the PGM rather than a few edge pixels used by conventional methods and thus it can work robustly even with low resolution images.

Iris contour fidelity: Sec. IV-B describes how to extract iris contour from eye image edge map. Therefore we are able to examine whether these edge points $\{q_j\}$ and the synthetic contour points $p_i(g_n)$ can match, i.e., we check the contour fidelity. As shown in Fig. 5 (bottom-left), all synthetic iris contour points $\{p_i(g_n)\}$ should have the same distance from the extracted iris contour.

A straightforward way to check the fidelity is to compute the average distance between every edge point $q_j$ on the extracted contour and its nearest synthetic contour point in $\{p_i(g_n)\}$, while this raises questions that 1) what if the synthetic and extracted irises are different in size and what if there are falsely extracted contour edge points? To handle these problems, we propose the following criterion:

$$ICF = \sum_j 1 - \sum_j \delta_j,$$

where

$$\delta_j = \begin{cases} 1, & \text{if } \min_i \left( \left\| q_j - p_i(g_n) \right\|_2 \right) \\ \min_i \left( \left\| q_j - p_i(g_n) \right\|_2 + \Delta \right), & \text{otherwise}, \end{cases}$$

and $\Delta$ is a constant value to tolerate distance variation. In our experiment we set $\Delta = 2$ pixels. Eq. (12) counts the number of edge points which are farther from the synthetic iris contour.
than other ones. A perfect iris fitting makes all edge points {\(q_j\)} have the same distance from the synthetic iris contour (ICF score equals to 0), while a wrong iris shape/position cannot achieve this. Therefore, the larger the ICF score is, the worse the fit is. Moreover, any false iris edge points like those in the rightmost plots of Fig. 5 always has a constant \(\delta_j = 0\) so it does not affect the ICF score during optimization. The criterion in Eq. (12) is robust since it does not rely on exact point matching which may be unreliable in practice.

**Final optimization:** As explained above, the final optimization should consider both Eq. (11) and Eq. (12). In particular, we maximize the optimization score to determine the optimal gaze direction \(\hat{g}\) and a pixel bias \(b\) as following:

\[
\{\hat{g}, \hat{b}\} = \arg \max_{g \in \{g_n\}, b} RS(g, b) - \lambda \cdot ICF(g, b),
\]

where \(\lambda\) (we set to 0.01) balances the influences of the two terms, and \(b\) is a bias value measured in pixels that tells the output iris center position based on its initial value. Note that in Eq. (11) and Eq. (12), although we do not explicitly introduce \(b\), it is straightforward to add such a bias to the synthetic iris data beforehand.

We solve Eq. (14) by combining a pyramid search and a steepest descend search strategy. The search step is gradually decreased to search over continuous eyeball positions and eyeball orientations contained in the synthetic data space, as described in Algorithm 1. The final output is the optimal gaze direction \(\hat{g}\) that best explains the captured eye image.

**Algorithm 1 3D gaze direction by using SIAF**

1. Initialize synthetic data \(\{g_n, p_j(g_n), c(g_n)\}\); pre-process the eye image.
2. for decrease the search step for \(n\) do
   1. Fix the current \(g_n\) and set \(b = 0\)
   2. \(\text{Score} \leftarrow (RS - \lambda \cdot ICF)\)
   3. while not converged do
      1. Update \(b\) by steepest descend
      2. Update \(\text{Score}\) by \(b\)
   4. end while
   5. \(\hat{g} \leftarrow g_n\) if \(\max(\text{Score})\) is updated
3. end for
4. Output optimal \(\hat{g}\)

**A. Recovery of eyeball center**

The key to handle multiple aligned eye images is to do a joint refinement. In particular, based on their initial estimates individually obtained in Sec. IV, we further determine their shared eyeball center position, and then use it as a cue to refine each gaze direction.

In particular, by solving Eq. (14), we obtain the optimal gaze direction \(\hat{g}\) and the iris center position (from image bias \(b\)) that best match the input eye image to one instance of the synthetic eyeball/iris data in Sec. IV-A. In other words, we now know which case in Fig. 3) best explains the input image. Notice that every case in Fig. 3) is fully accessible, and therefore we can directly look up its corresponding relative 3D eyeball center position and its 2D image projection, denoted by \(e(g)\).

Under the assumption that all the eye images are well aligned, we can consider that they have the same eyeball center \(\hat{e}\) since eyeball rotation does not change its center position. Therefore, the problem is how to determine a unique \(\hat{e}\) for all images. Assume we have totally \(M\) eye images with estimated eyeball centers \(\{e(g_m), m = 1, \cdots, M\}\). A straightforward way is to take their average:

\[
\hat{e} = \frac{1}{M} \sum_{m=1}^{M} e(g_m).
\]

However, for better accuracy, we would expect to only use a
Following the previous section, assume we have obtained the unique eyeball center $\hat{e}$, we then use it as a prior to refine the 3D gaze estimation for each eye image of the same person. In particular, we still use the proposed criterions including radial symmetry and iris contour fidelity defined by Eq. (11) and Eq. (12). However, because we fix the eyeball center position $\hat{e}$, the solution space is highly reduced and therefore we do not need to optimize over all possible 3D gaze directions and image biases. We explain this with more details as following.

**Possible gaze direction:** for each eye image, note that we already have its initial 3D gaze estimation result $\hat{g}$ by solving Eq. (14). Based on this initial guess, we can set the search range by adding $\pm 15^\circ$ to both the X and Y directions. Then, within this range, we search for the final gaze direction $\tilde{g}$.

**Image bias:** in Eq. (14), image bias $b$ is a free variable to be optimized. However, now we have fixed the eyeball center position $\hat{e}$ in the image, and thus the image bias is already included and fixed. As a result, as shown in Fig. 3, the synthetic iris center position $c(g)$ and iris contour points $p_i(g)$ can be fully determined by the gaze direction $g$.

**Final optimization:** following the above analyses, we propose a 3D gaze direction refinement method based on the known eyeball center position $\hat{e}$. The refinement is done for each subject independently. We search for the optimal gaze direction $\tilde{g}$ within a $15^\circ$ range on the basis of the initial gaze direction $\hat{g}$. The optimization is done by maximizing the score similar to that in Eq. (14). In particular, we solve the following problem:

$$
\tilde{g} = \arg \max_{g \in \{\hat{g}\}} \left( RS(g, \hat{e}) - \lambda \cdot ICF(g, \hat{e}) \right),
$$

s.t. \( \arccos(g \cdot \tilde{g}) < 15^\circ \),

by using a simple pyramid search for the optimal $\tilde{g}$.

The resulting 3D gaze direction $\tilde{g}$ is the final output of our method. Note that by using the eyeball center position $\hat{e}$, we also resolve the binary ambiguity described in Sec. III-B, which is inevitable when determining the orientation of an ellipse only from its shape. Examples are given by experimental results in the next section.

\footnote{Such a range is generally enough according to our reported SIAF results.}
TABLE I: Gaze ranges in our dataset for experiments.

<table>
<thead>
<tr>
<th></th>
<th>Horizontal</th>
<th>Vertical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angular range</td>
<td>$-19.0^\circ \sim 27.3^\circ$</td>
<td>$-29.0^\circ \sim -0.5^\circ$</td>
</tr>
<tr>
<td>Screen pixel range</td>
<td>$\approx 500 \text{ mm}$</td>
<td>$\approx 300 \text{ mm}$</td>
</tr>
</tbody>
</table>

VI. NUMERICAL EVALUATION

We quantitatively examine the proposed method via extensive experiments. In order to allow direct comparisons with previous methods and also ensure repeatability, our experiments are conducted on a public dataset proposed in [36].

A. Experimental data

We use the gaze data proposed in [36]. There are totally 50 different subjects included in the dataset. For each of the subject, eye images are captured from eight different viewpoints, and there are 160 different gaze points on a screen. Therefore, the total number of samples is $50 \times 8 \times 160 = 64000$. These gaze data are captured in a common user-screen scenario and in each session the user keeps a natural but fixed head pose.

The dataset contains a very large amount of data to support regression-based methods; it collects sufficient data by varying multiple eye appearance factors to allow regressions for unknown eye region appearances. However, this is not required in our experiments since our method belongs to the model-based category which just cares the iris shape. Therefore, we extract a subset we need. In particular, we choose 12 subjects. We also choose the eye images captured by a camera under the screen because other capture angles in the dataset were set too large for the sake of multiview synthesis.

Overall, our used dataset contains $12 \times 160 = 1920$ eye images. Some statistics for the dataset are shown in Table I. These numbers are typical and sufficient for a user-screen scenario.

Note that when testing the SIAF-joint method, since it assumes temporally stable head poses between occasional head motions, our current test dataset is somewhat imperfect since it does not contain head motion in the same session. Therefore, during head motion, we should refer to the results of the basic SIAF method instead of the SIAF-joint method. This should not influence the evaluation much since in common tasks gaze motion happens more frequently than head motion. On the other hand, although there are gaze tracking datasets containing head motion, e.g., Eyedip [43], they are designed to quickly traverse in different head poses without considering head motion frequency, and therefore not suitable here. Developing datasets to meet our assumption can be a potential future work.

B. Gaze estimation accuracy

In this section, we show gaze estimation accuracy for the proposed SIAF and SIAF-joint methods. Besides, we also provide results by using the following two methods for comparison:

Appearance-based:

An appearance-based method [10]. This type of methods require training samples. However, assuming exact person-specific training samples for each test subject is too strong in this comparison since other methods are unsupervised; therefore, we alternatively prepare 99 training samples from 3 other subjects not in the test dataset and use them as the shared training data. This strategy is quite similar to another recent work [36] in order to avoid online training. In this way, this method has been modified to avoid person-dependent training and represent the state of the art; however, the person-independent training data collected beforehand is still an additional requirement compared to our method.

EyeTab:

This method [15] additionally detect features such as eyelids to extract iris contours more precisely. It is also calibration-free. However, high resolution eye images and small eye-camera distances are necessary. Otherwise, like will be seen, the eyelid/iris detection may frequently fail and there is a chance that the system cannot give any output.

Ellipse-fitting:

This method fits an ellipse to the detected iris contour edge (Sec. IV-B) by using lease squares criterions, and recovers 3D gaze directions by examining the shape of the ellipse (Sec. III-B). In the implementation we use the proposed image processing techniques to achieve robust iris detections. Note that there is a binary ambiguity in the solution as explained before, and we choose the gaze direction pointing upwards rather than the other one. This method can represent a category of methods based on ellipse fitting [13], [14]

Gaze estimation results and comparison can be seen in Table II. The proposed SIAF-joint method achieves the best accuracy, followed by the appearance-based method. However, note that the appearance-based method needs training, which greatly limits its practicality since performing active training cost much time and efforts; while the other methods are unsupervised. The EyeTab method has large errors and failure rates mainly because the test data was captured only at a common resolution; it is more suitable to work with high resolution eye images and small eye-camera distance. Besides, the SIAF and ellipse-fitting methods achieve similar accuracy. Both of them in fact benefit from the nicely extracted iris edge by using our pre-processing in Sec. IV-B. Without good iris edge extraction, the ellipse-fitting method may perform extremely poor as shown later in the case study. Finally, the SIAF-joint method achieves highest accuracy with multiple aligned eye images while it does not require any training data. A more intuitive comparison on their average accuracy is given in Fig. 7

C. Study on individual cases

After comparing the average accuracies, we conduct case studies to show representative and more intuitive results. All the results are shown in Fig. 8. They are obtained by using three methods, i.e., ellipse-fitting, SIAF and SIAF-joint, and

\footnote{In particular, here ‘appearance-based’ refers to ‘regression-based’.}

\footnote{The appearance-based method may achieve higher accuracy under the condition that it has more training samples. However, requirement on training is also considered a major drawback of the appearance-based methods.}
TABLE II: Comparison on gaze estimation error. Results are obtained by four different methods for twelve subjects.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avg. err. [°]</td>
<td>Std [°]</td>
<td>Avg. err. [°]</td>
<td>Std [°]</td>
<td>Avg. err. [°]</td>
</tr>
<tr>
<td>S1</td>
<td>10.17</td>
<td>2.15</td>
<td>17.11</td>
<td>10.63</td>
<td>14.89</td>
</tr>
<tr>
<td>S3</td>
<td>5.27</td>
<td>2.60</td>
<td>18.11</td>
<td>10.14</td>
<td>13.60</td>
</tr>
<tr>
<td>S4</td>
<td>4.33</td>
<td>2.20</td>
<td>18.58</td>
<td>9.88</td>
<td>10.84</td>
</tr>
<tr>
<td>S5</td>
<td>14.13</td>
<td>2.31</td>
<td>18.74</td>
<td>9.82</td>
<td>17.08</td>
</tr>
<tr>
<td>S6</td>
<td>6.81</td>
<td>4.26</td>
<td>20.83</td>
<td>11.98</td>
<td>16.00</td>
</tr>
<tr>
<td>S7</td>
<td>15.40</td>
<td>3.12</td>
<td>16.21</td>
<td>9.97</td>
<td>23.07</td>
</tr>
<tr>
<td>S8</td>
<td>6.73</td>
<td>4.37</td>
<td>22.50</td>
<td>9.40</td>
<td>25.04</td>
</tr>
<tr>
<td>S9</td>
<td>8.53</td>
<td>2.08</td>
<td>18.56</td>
<td>12.02</td>
<td>10.96</td>
</tr>
<tr>
<td>S10</td>
<td>14.45</td>
<td>3.08</td>
<td>18.86</td>
<td>8.75</td>
<td>10.87</td>
</tr>
<tr>
<td>S11</td>
<td>17.03</td>
<td>6.42</td>
<td>21.44</td>
<td>10.78</td>
<td>13.38</td>
</tr>
<tr>
<td>Average</td>
<td>10.73</td>
<td>3.57</td>
<td>19.13</td>
<td>10.44</td>
<td>15.51</td>
</tr>
</tbody>
</table>

Notes

- Use person-independent training data
- Average no-output rate: 13.02%
- Based on the proposed iris edge detection
- Independent computing for each image
- Joint optimization

![Fig. 7: Comparison on average gaze estimation error.](image)

The errors in pixel are shown in Table III.

Among all cases, (a) and (b) are considered normal ones, where all methods obtain similar and accurate results; (c), (d) and (e) are more challenging cases, where the iris edge detection and ellipse-fitting method may fail, and even the SIAF method can produce large errors. The reason is that in these cases, the iris contour regions are heavily occluded by eyelids. On the other hand, the SIAF-joint method still produces reliable results. Finally, (f) shows a special case, where the gaze direction is in fact pointing downward while only the SIAF-joint method can recover this. The reason is that the iris elliptical shape in the case (f) can correspond to two gaze directions, i.e., an upward one and a downward one. Without knowing the eyeball center information, it is unlikely to distinguish them.

**D. On eyeball center estimation**

In this section, we experimentally show the reason why SIAF-joint outperforms SIAF. In particular, we find that the eyeball center estimation is very important for accurate gaze estimation in our method. To demonstrate this, we compare the estimated eyeball center positions with the ground truth. The errors in pixel are shown in Table III.

These results clearly show that the SIAF-joint method outperforms the SIAF method for all subjects by a large margin. This is easy to understand since the SIAF-joint method computes the eyeball center position by knowing all the individual estimates from SIAF, and it has the ability to pick up reliable estimates out of others. As a result, the average error reduces from around 9 pixels to only 2 pixels. This is satisfactory considering the typical eye image size of 150×75 in our dataset.

Fig. 9 shows some intuitive examples of individual eyeball center estimation for four subjects. More importantly, we examine the relationship between the eyeball center error and the final gaze estimation error. From the top row of Fig. 9, it is clear that large eyeball center errors cause large gaze errors, and vice versa. From the bottom figures, we see more directly strong positive relations between them. These results demonstrate that for the SIAF method, the eyeball center position is the key to control the gaze estimation accuracy. If we are able to estimate eyeball center position accurately, the gaze error can be reduced to a small number depending on individuality. Further removing this error requires an active calibration stage to compensate the visual axis offset as mentioned in Sec. III-B, while in our experiments it remains in the final results.

We use another plot, Fig. 10, to show how average gaze error of SIAF-joint varies with average eyeball center error for each of the twelve subjects. From these results, we can make two conclusions. First, the positive relation between gaze error and eyeball center error can still be seen in their averages. This suggests further reducing average gaze error by considering better eyeball center estimation methods for those

---

3Ground truth of the eyeball center is not available in the dataset. It is obtained by examining the test data with smallest gaze estimation errors.
Fig. 8: Comparison between three methods on individual cases. Estimated iris contours, 3D gaze directions and PGMs are shown for each case.

TABLE III: Estimation errors of eyeball center position in pixel.

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
<th>S8</th>
<th>S9</th>
<th>S10</th>
<th>S11</th>
<th>S12</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIAF-joint</td>
<td>2.43</td>
<td>1.32</td>
<td>2.50</td>
<td>1.80</td>
<td>1.86</td>
<td>0.98</td>
<td>1.39</td>
<td>1.98</td>
<td>2.17</td>
<td>4.40</td>
<td>4.51</td>
<td>2.63</td>
<td>2.33</td>
</tr>
</tbody>
</table>

Fig. 10: Average gaze errors and eyeball center errors for all subjects by using SIAF-joint.

‘difficult’ subjects like S10 and S11. Second, such average accuracies for gaze estimation are already promising for many practical applications, considering that the proposed method is unsupervised and it requires no special hardware.

VII. CONCLUSION AND DISCUSSION

In this paper, we propose to estimate 3D human eye gaze from a single eye image without active illumination. The key idea is to analyze the shape of the elliptical iris contour. Existing methods may face challenges in ellipse fitting because perfect iris contour detection is difficult, especially with low resolution eye images. To solve this problem, we propose a model-driven Synthetic Iris Appearance Fitting (SIAF) method. It synthesizes physically possible iris appearances and then optimizes over this synthetic space to find the best solution to explain the captured image. It also benefits from advanced image analysis techniques to guarantee a robust solution. Furthermore, when multiple eye images are available, a SIAF-joint method is introduced to further improve the gaze estimation accuracy, by assuming the consistency of eyeball center position under a temporarily stable head pose. Compared to many previous methods, our method achieves reasonable accuracies while it does not require any special hardware, training input or prior information. Future work will consider further exploiting the eyeball center’s temporal consistency in the SIAF-joint method to allow it work for frequently changed head poses.

ACKNOWLEDGEMENTS

This work was partially supported by the Joint Funds of NSFC-CARFC (U1533129), NSFC (61325011,61532003), and Fundamental Research Funds for the Central Universities.

REFERENCES

Fig. 9: Relation between eyeball center error and gaze error. Results for four subjects are shown in the four columns. It is clear that the gaze estimation error is linearly related to the eyeball center error.


---

**Feng Lu** (M’16) received the B.S. and M.S. degrees in automation from Tsinghua University, in 2007 and 2010, respectively, and the Ph.D. degree in information science and technology from The University of Tokyo, in 2013. After working with the Institute of Industrial Science, the University of Tokyo, he joined the Beihang University, China, in 2015. His research interests cover various topics in computer vision and human computer interaction, including human gaze analysis, 3D shape recovery, and reflectance analysis.

---

**Yue Gao** (SM’14) received the B.S. degree from the Harbin Institute of Technology, Harbin, China, and the M.E. and Ph.D. degrees from Tsinghua University, Beijing, China.

---

**Xiaowu Chen** (SM’15) received the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2001. He is currently a Professor with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. His current research interests include virtual reality, computer graphics, and computer vision.