Appearance-Based Gaze Estimation via Uncalibrated Gaze Pattern Recovery

Feng Lu, *Member, IEEE*, Xiaowu Chen, *Senior Member, IEEE*, and Yoichi Sato, *Senior Member, IEEE*

***Abstract*—Aiming at reducing the restrictions due to person/scene dependence, we deliver a novel method that solves appearance-based gaze estimation in a novel fashion. First, we introduce and solve an "uncalibrated gaze pattern" solely from eye images independent of the person and scene. The gaze pattern recovers gaze movements up to only scaling and translation ambiguities, via nonlinear dimension reduction and pixel motion analysis, while no training/calibration is needed. This is new in the literature and enables novel applications. Second, our method allows simple calibrations to align the gaze pattern to any gaze target. This is much simpler than conventional calibrations which rely on sufficient training data to compute person and scene-specific nonlinear gaze mappings. Through various evaluations, we show that 1) the proposed uncalibrated gaze pattern has novel and broad capabilities; 2) the proposed calibration is simple and efficient, and can be even omitted in some scenarios; 3) quantitative evaluations produce promising results under various conditions.**

***Index Terms*—gaze estimation, uncalibrated gaze pattern, dimension reduction**

I. Introduction

Eye gaze provides crucial cues for vision-based intelligent systems to understand human visual attention, emotion, feeling and so on [1], [2], [3]. In the field of computer vision, various techniques for human gaze estimation have been proposed to support applications such as human-computer interaction, saliency detection, ego-centric activity analysis, and video surveillance [4], [5], [6].

As surveyed in [7], computer vision-based gaze estimation methods can be categorized into either model-based or appearance-based. Appearance-based methods only need a single camera, but they face an essential problem of burdensome person-specific calibration to collect enough training samples. In general, such a calibration stage needs to be done every time before the user can start tracking eye gaze. This is not only time-consuming (several minutes per session), but also annoying since users have to repeat the strict procedure by fixating on each calibration points on the screen at the right time. Therefore, although a careful calibration can collect enough training data to ensure a good accuracy, in many scenarios like video surveillance and medical treatment, such a user-participation-based calibration is impossible.

Techniques have been proposed to avoid explicit calibration. For instance, saliency-based methods [8], [9] let users watch known images/videos and assume that the users' fixations are driven by the saliency distribution on the screen. Another method [10] assumes that mouse clicks always coincide with user fixation points on the screen. Because these methods rely on various assumptions as above, they can work well in their designated scenarios, but may fail in many other applications where users do not watch known images/videos or use mouse clicks, or their eye gaze movements do not follow external stimulus in a passive way. Overall, these methods' availability is still highly dependent on specific scenes where they are designed.

Aiming at recovering human gaze automatically in a less person/scene-dependent manner to allow practical applications such as video surveillance, patient treatment and cognitive analysis, two basic principles need to be carefully considered in advance. First, the input should be just eye images/videos captured by a single camera, so that they can be transmitted and processed uniformly by most common devices. Second, calibration should be avoided or at least without requiring users to repeatedly fixate on dozens of predefined screen markers to capture eye images for training.

In order to meet these requirements, regarding the former issue, an appearance-based method is more suitable to use since it only needs a monocular camera to capture eye appearances. Regarding the latter issue, in order to reduce the calibration burden, the most important issue is about how to avoid troublesome user-system synchronization – both the user and the system have to wait for seconds before capture each sample to ensure that the user has enough time to fixate on the correct gaze position. Repeat this for dozens of times is obviously time-consuming and it makes high demands on user attention, participation and system design. The need of making the calibration simple, fast and even user-unaware for many applications motivates our research in this paper.

Accordingly, the key of this work is to deliver a less person/scene-dependent approach by avoiding person/scene-specific calibration/training. Given that there are exponentially increasing gaze tracking researches in recent years, we first list our method's contributions as bellow:

1) Our method works with common eye images captured by a monocular camera in unspecified scenes. In contrast, previous methods have to capture data in their specific scenes with certain hardware/settings.

2) We propose and solve an *uncalibrated gaze pattern*,

Correspondence should be addressed to Feng Lu and Xiaowu Chen

This work was supported in part by the Joint Funds of NSFC-CARFC (U1533129), the NSFC (61602020, 61532003, and 61325011), and the Fundamental Research Funds for the Central Universities

Feng Lu and Xiaowu Chen are with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. Feng Lu is also with the International Research Institute for Multidisciplinary Science, Beihang University, Beijing, 100191, China. E-mail: {lufeng,chen}@buaa.edu.cn

Yoichi Sato is with the Institute of Industrial Science, the University of Tokyo, Tokyo 153-8505, Japan. E-mail: ysato@iis.u-tokyo.ac.jp

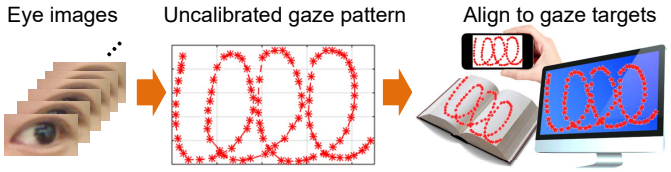

Fig. 1: We recover uncalibrated gaze pattern solely from eye images regardless of the scenario/individuality. The gaze pattern solves gaze positions up to only scaling and translation ambiguities. A calibration can be done by aligning the gaze pattern to various gaze targets through scaling and translation.

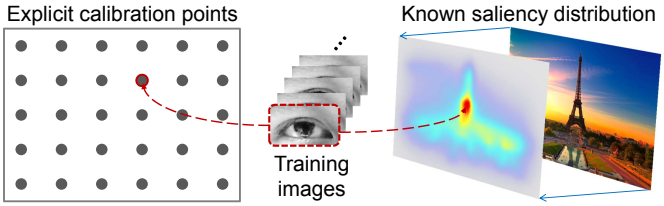

Fig. 2: Previous methods assume explicit calibration (left) or known saliency distributions (right) on the screen to capture training images.

which predicts gaze movements up to only scaling and translation ambiguities (Fig. 1). The method is completely *unsupervised* and *person/scene-independent*, requiring no calibration/training. This is the first time in the literature and it enables new applications.

3) We can align the uncalibrated gaze pattern to real-world gaze positions, e.g., screen coordinates (Fig. 1), via a simple calibration. Compared to conventional methods, our calibration is flexible and task non-interrupting; it can be orders of magnitude faster than conventional calibrations and requires minimum or even zero user participation.

These characteristics make our method fundamentally different from others. They also have practical significance. For instance, the proposed uncalibrated gaze pattern meets the requirements in many applications, e.g., cognitive analysis and video surveillance, where conventional calibration based on user-system collaboration is hard or impossible.

A. Related Works

Many computer vision-based gaze estimators have been proposed, which can be divided into two major categories: model-based and appearance-based [7], [11]. Model-based methods assume 2D or 3D eye models [12] or other geometric models. These parametric models are computed by using near infrared (IR) corneal reflections [13], [12], [14], [15], pupil centers [16] and iris contours [17], [18], [19] detected from eye images. These methods are widely used in laboratory environments due to their high accuracy. However, they typically require hardware that may comprises infrared lights, stereo/high-definition cameras and RBG-D cameras [18], [19], which may not be supported by many common devices.

Appearance-based methods, however, only require a single camera to capture user eye appearances. They typically need a calibration stage to collect training data, including eye images and the corresponding gaze positions, to learn a gaze mapping function. Baluja and Pomerleau [20] collected thousands of training images, while Tan et al. [21] used hundreds of training images to learn a regression. Williams et al. [22] introduced a semi-supervised Gaussian regression method, and Lu et al. [23] introduced an adaptive regression method to reduce the number of training images and also handle other related issues [24]. Sugano et al. [8] developed a technique to auto-calibrate gaze positions by leveraging visual saliency information from videos; Chen et al. [25] and Alnajar et al. [9] also used similar cues in gaze estimation. More recently, Schneider et al. [26] proposed to use dimension reduction and align the resulting 2D embeddings. Different from our method, it requires training data covering the same gaze region in the same scene, and thus the method is still training-based and scene-dependent.

Calibration becomes a severer problem if user head pose changes. In such a case, old training data cannot be used. Sugano et al. [10] continually collected new training images for new head poses. Lu et al. proposed methods [27], [28] that collect a set of additional training images to work with old training images for new head poses. Overall, additional calibration and input data are needed.

II. Motivation and Overview

Before going into the technical details, we give the motivation and an overview of the proposed approach and discuss its two components, namely uncalibrated gaze pattern recovery and flexible calibration.

A. Motivation

Most accurate gaze estimators require certain types of calibration to capture eye images as training data before every session. The capture can be explicit, e.g., the user intentionally gaze at each calibration point and trigger a capture (Fig. 2 (left)), or implicit, e.g., the user gaze is assumed to follow known saliency distributions from specific images/videos [8], [25], [9] (Fig. 2 (right)). These methods share similar limitations: 1) sufficient training data (eye image and gaze position pairs) should be obtained before estimation, 2) complex person-specific gaze mappings have to be built, and 3) estimation can only work within the calibrated region in the same scene.

Unlike them, we propose to recover an "uncalibrated gaze pattern" by using only a set of eye images without requiring any training data, and thus it poses no assumption about specific person/scene/gaze region. The gaze pattern can also be aligned to any target to obtain similar output with calibration-based method, and it is extremely simple and can be done afterwards without additional input. Therefore, our method is flexible and much less person/scene-dependent (Fig. 1), as detailed below.

B. Uncalibrated Gaze Pattern

As shown in Fig. 1, given a set of eye images, we can estimate an uncalibrated gaze pattern, i.e., relative gaze positions/movements solely from these eye images. The gaze

TABLE I: Comparison on (explicit/implicit/auto) calibrations for appearance-based methods

Method	Image capture time	Style and timing	Gaze positions	Problem to solve
Ours	< 0.1s per point	Flexible, task non-interruptive	Free on the gaze target	Simple scaling & translation
Explicit calib. [21], [22], [23]	$2 \sim 5$s per point[1]	Full participation before use	Pre-determined	Nonlinear gaze mapping
Saliency-based [8], [9]	Entire watching	Watch known images/videos	Visual saliency-driven	Nonlinear gaze mapping

pattern is called 'uncalibrated' since it is computed without calibration/training, and thus it does not have scene-specific coordinates; meanwhile, gaze movements are determined up to only scaling and translation ambiguities, and thus it reflects the shape of scene-independent gaze trajectory. In fact, we humans can also identify the eye movement pattern from eye images without knowing the person or scenario (Fig. 1). This shows a broad range of potential applications.

The uncalibrated gaze pattern distinguishes our method from most others. Meanwhile, we find that [9] and [26] also exploit relationships between eye images and uncalibrated gaze positions. However, our method is unique in that 1) their methods need pre-collected training data from different users, and should be applied with the same system in the same scene. On the contrary, our method is completely training-free and *person/scene-independent*. 2) their obtained uncalibrated gaze positions are less accurate and contain more ambiguities, which require additional training data or saliency data to improve. In contrast, our technique directly outputs robust and standalone gaze patterns via technique in Section III. No additional input or scene prior is needed.

C. Simple and Flexible Calibration

As a major challenge in appearance-based methods, calibration is necessary before every session. It takes time and interrupts user tasks. In contrast, our calibration via alignment can be orders of magnitude faster than conventional calibrations, while it has minimum effect on the user task.

Table I shows comparisons from several aspects. We also summarize the important points as bellow.

1) *Time duration*: our method does not capture particular calibration image. It just captures free eye images during the task in real time. Therefore, capturing each eye image costs less than 0.1s. In contrast, conventional calibrations require a separate calibration stage before use, which takes several minutes[1] to capture all training samples.

2) *Task interruption*: our method requires no special actions of user/system and thus does not interrupt the task; while conventional methods require full collaborations of both user and system[1] during calibration before the task.

3) *Scenarios*: we have no particular assumption; while conventional calibrations rely on specific scene/settings and severely restrict the gaze range during task.

4) *Problem to solve*: our calibration only needs to align the previously solved gaze pattern to the gaze target via simple scaling and translation, which is an easy 2D linear problem; while conventional calibrations capture sufficient data to train nonlinear and person-specific gaze mappings only valid for the current scene.

Overall, our calibration is flexible and fast; it assumes nothing special about the scene, and even can be done after the task. This is impossible for previous methods.

[1]Time and effort is needed for synchronization, i.e., showing a marker on the screen, waiting for user fixation and capturing an image, etc.

III. Uncalibrated Gaze Pattern Recovery

With a roughly stable head pose during a certain period, we recover uncalibrated gaze pattern sole from eye images.

A. Eye Image Preprocessing

Like most appearance-based methods, we expect aligned eye images with the same size. However, in most cases, the captured eye regions always vary in position due to slight head motion. Therefore, given an input image sequence, eye image alignment and cropping should be done first.

We adopt eye corner-based alignment and cropping. In particular, we first use a face tracking algorithm, e.g., [29], to roughly detect eye regions in the input image sequence. These eye regions are cropped out and converted into gray-scale. Then, we apply the edge filter to the first eye image. By combining the edge map and the eye corner position provided by the face tracker, we can determine an eye corner position in the first eye image, and then an eye corner template T can be extracted from the eye image at this position. In other words, T is a small image patch of the eye corner area from the first eye image.

For each following eye image E, we find its eye corner region by matching its appearance to the template T. In particular, we solve for the optimal pixel position (u, v) so that the image patch extracted at (u, v) from E best matches T in terms of their appearances. Besides, note that eye corner appearance varies obviously due to different degrees of eye open. Therefore, we propose to use a vertically scalable version of the template T. The problem can be then formulated as following

$$(\hat{u}, \hat{v}, \hat{\tau}) = \underset{u,v,\tau}{\operatorname{argmin}} \frac{\|T \circ \tau - E\left(u, v, \operatorname{size}\left(T \circ \tau\right)\right)\|_2}{\text{Number of pixels in } T \circ \tau}, \quad (1)$$

where $T \circ \tau$ scales image template T in the vertical direction, and $E\left(u, v, \operatorname{size}\left(T \circ \tau\right)\right)$ denotes the image area in E at position (u, v) with the same size as $T \circ \tau$. This problem can be efficiently solved by steepest descend methods and the resulting $\hat{u}$, $\hat{v}$ and $\hat{\tau}$ determine the eye corner position in E. Based on the obtained eye corner positions, eye images can be precisely aligned and cropped out with a fixed size.

Note that the above method focuses on eye image translation rather than rotation and scaling. The reason is that we assume a roughly stable head pose, and thus the image deformation due to slight head motion majorly causes eye image translation.

Besides eye image alignment, illumination change is also a common problem for appearance-based gaze estimation. In our case, since we mainly capture eye images with a roughly fixed head pose in a short time period, the illumination is likely to stay unchanged. However, one can still add an explicit process, e.g., using local sensitive histograms [30], to handle illumination effect.

B. Eye Appearance Manifold Construction

As preliminarily shown in [23], eye images captured for a fixed head pose constitute a manifold in m-D space, where m is the number of pixels in an eye image. Since human eyeball movement only has two degrees of freedom, an ideal eye image manifold should have an intrinsic dimensionality close to two, and if we examine a local region in the manifold, it should have a similar 2D structure with the corresponding 2D gaze positions. Therefore, conventional appearance-based methods [10], [21], [28] assume local interpolation consistency between eye image manifold and 2D gaze position space, and perform gaze estimation by

$$\hat{\mathbf{x}} = \sum_i w_i \mathbf{x}_i \quad \text{s.t.} \quad \hat{\mathbf{e}} = \sum_i w_i \mathbf{e}_i, \tag{2}$$

where $\hat{\mathbf{e}}$ and $\hat{\mathbf{x}}$ denote a test image vector and its corresponding gaze position, $\{\mathbf{e}_i\} \in \mathbb{R}^m$ and $\{\mathbf{x}_i\} \in \mathbb{R}^2$ denote eye image vectors and gaze positions of local training samples that are similar to the test sample, and $\{w_i\}$ are interpolation weights. In other words, the local Euclidean structures of an eye image manifold region and its corresponding gaze position area are similar:

$$\|\mathbf{e}_i - \mathbf{e}_j\|_2 \propto \|\mathbf{x}_i - \mathbf{x}_j\|_2. \tag{3}$$

Relationship in Equation (3) holds true only within a local region as assumed above, meaning that a pair of eye images should be very similar with each other, otherwise their Euclidean distances are no longer proportional to the gaze movements, and thus they cannot recover gaze positions linearly. Therefore, conventional methods based on Euclidean measurements of eye appearance differences cannot handle large eye gaze variation directly, and they have to collect dense calibration/training samples to help solve the problem inside each local region.

Experimental results are shown in Fig. 3 (left) to visualize the local linearity and global nonlinearity problem more intuitively. Obviously, linear relation between the eye image distances and the corresponding gaze movements cannot hold true globally. To solve this problem, our idea is to extend the linear region by introducing the Geodesic distances of eye images. A geodesic distance measures the shortest path between two nodes in a graph, and is computed by adding up small Euclidean distances of neighboring nodes along that path [31]. For instance, given a set of eye images $\{\}I_1, \cdots, I_{22}\}$ with 22 smoothly changed gaze directions, as shown in Fig. 3 (right), the Geodesic distance between I_1 and I_{22} is computed by $\|I_1 - I_2\| + \cdots + \|I_{21} - I_{22}\|$; by comparison, traditional Euclidean distance only computes $\|I_1 - I_{22}\|$. By plotting all results in Fig. 3 (right), we see a strong global linear relation between the Geodesic distance of eye images and the corresponding gaze movement.

The resulting large-scale linearity shows a good potential to recover global gaze movements without requiring dense training data. In the following, we propose two steps, namely, computing pair-wise Geodesic distances and embedding samples into 2D space, to compute the relative gaze positions based on the observed global linearity[1].

Pair-wise Geodesic distance We first compute local Euclidean distances for all eye image pairs by

$$d(\mathbf{e}_i, \mathbf{e}_j) = \begin{cases} \|\mathbf{e}_i - \mathbf{e}_j\|_2 & \text{if } \|\mathbf{e}_i - \mathbf{e}_j\|_2 < \varepsilon_i \\ +\infty & \text{otherwise,} \end{cases} \tag{4}$$

where ε_i is a threshold to only keep local distance. In our experiments, we set it to the n-th shortest distance from $\mathbf{e}_i$ to all other $\mathbf{e}_j$, where n is set to a quarter of total images' number. Then the geodesic distance $d_G(\mathbf{e}_p, \mathbf{e}_q)$ between any two eye images $\mathbf{e}_p$ and $\mathbf{e}_q$ can be computed using $\{d(\mathbf{e}_i, \mathbf{e}_j)\}$ as

$$d_G(\mathbf{e}_p, \mathbf{e}_q) = D_{sp}(\{d(\mathbf{e}_i, \mathbf{e}_j)\}), \tag{5}$$

where function $D_{sp}(\cdot)$ runs the Dijkstra's algorithm to find the shortest paths between $\mathbf{e}_p$ and $\mathbf{e}_q$ in the graph defined by pair-wise distances $\{d(\mathbf{e}_i, \mathbf{e}_j)\}$.

Embedding into 2D space The computed $\{d_G(\mathbf{e}_p, \mathbf{e}_q)\}$ is a set of Geodesic distances measuring the appearance differences of eye images. According to our analysis and Fig. 3, these distances are supposed to be proportional to gaze movements, i.e., $d_G(\mathbf{e}_p, \mathbf{e}_q) \propto \|\mathbf{x}_p - \mathbf{x}_q\|_2$ holds true for any p and q even when they do not indicate similar samples. In other words, the distance set $\{d_G(\mathbf{e}_p, \mathbf{e}_q)\}$ can be a good substitute for the unknown gaze movements, i.e., $\{\|\mathbf{x}_p - \mathbf{x}_q\|_2\}$.

Meanwhile, we consider how to convert gaze movements $\{\|\mathbf{x}_p - \mathbf{x}_q\|_2\}$ into certain gaze positions. Without loss of generality, we denote any gaze positions by 2D coordinates. Apart from the reason that we usually deal with 2D screen gaze positions, even if considering 3D free gaze vectors, they still have the degrees of freedom (DoF) of 2, meaning that they can be expressed in 2D. In fact, as shown in Fig. 1, we can effectively express free gaze motions by 2D gaze trajectories on a virtual plane.

Consequently, we should solve for a set of points $\{\mathbf{z}_p\} \in \mathbb{R}^2$ in the 2D space under the condition that their pair-wise distances are preserved from the gaze movements $\{\|\mathbf{x}_p - \mathbf{x}_q\|_2\}$. Now, recall that $d_G(\mathbf{e}_p, \mathbf{e}_q) \propto \|\mathbf{x}_p - \mathbf{x}_q\|_2$, the problem can then be formulated as

$$\{\mathbf{z}_p\} = \underset{\{\mathbf{z}_p\}}{\operatorname{argmin}} \sum_{p,q} \left[(\mathbf{z}_p - \mathbf{z}_q)^2 - d_G^2(\mathbf{e}_p, \mathbf{e}_q) \right]^2. \tag{6}$$

This problem can be converted into an eigenvalue problem and then solved by methods as proposed in [32]. The solution gives 2D gaze coordinates $\{\mathbf{z}_p\}$.

[1]Although the global linearity is better preserved by using the Geodesic distance, it may still fail if the data are too sparse.

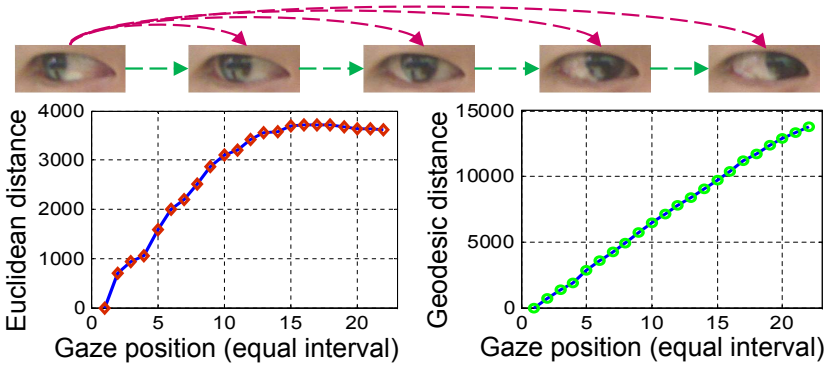

Fig. 3: Experimental results of the linear relation between eye image distance and gaze movement. Eye images are captured with eye ball rotation around 5° between each. When eyeball movement becomes large, Euclidean distance is no longer proportional to gaze movement, while Geodesic distance achieves good linearity.

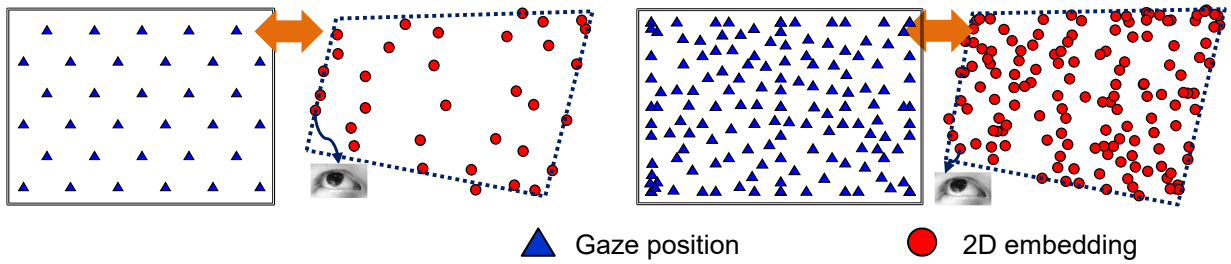

Fig. 4: True gaze positions and dimension reduction results. Their structures show rotation/flip/distortion ambiguities.

Examples of 2D embedding results are given in Fig. 4, where we show two cases with different numbers of eye images. As expected, the results show similar structures with the ground truths of 2D gaze positions, which confirms the proposed global linearity and the gaze recovery method. This further infers that eye images themselves already carry essential information to recover the global gaze pattern without assuming training data.

However, note that the 2D embedding results are not our final estimates, because there exist ambiguities, i.e., flip, rotation, translation and scaling. For instance, the 2D embeddings in Fig. 4 are rotated by nearly 90° compared to the real gaze positions (there is also a little keystone distortion). This is because Equation (6) only preserves distance which is invariant agains flip and rotation; moreover, $d_G(\mathbf{e}_p, \mathbf{e}_q) \propto \|\mathbf{x}_p - \mathbf{x}_q\|_2$ also involves an implicit and unknown scaling factor. These ambiguities are resolved in the following sections to produce the final gaze pattern.

In summary, using non-linear distance of eye images to preserve large-scale linearity in gaze computation is the key of this work. As shown in Fig. 3, it helps recover large eye gaze movements more accurately than using conventional linear methods, e.g., principle component projection [9].

C. Gaze Pattern Regularization

The 2D embedding result in Section III-B contains inherent ambiguities including rotation, flip and even a little distortion. In this section we propose an optimization method to remove them. The idea is to use motion cues extracted from eye images to resolve these ambiguities.

The eye images are first processed by inverting their pixel values so that the iris regions have large pixel intensities. Then, we compute the gravity center $\mathbf{g} = [g_x, g_y]^{\mathrm{T}}$ of each image (denoted as $I(u, v)$) by

$$\begin{aligned} g_x &= \frac{\Sigma_u \Sigma_v I(u,v) \cdot u}{\Sigma_{u,v} I(u,v)}, \\ g_y &= \frac{\Sigma_u \Sigma_v I(u,v) \cdot v}{\Sigma_{u,v} I(u,v)}. \end{aligned} \tag{7}$$

Intuitively, image gravity center position (g_x, g_y) is highly related to iris position in the eye image. Although it cannot be accurate to directly predict iris position from (g_x, g_y), its motion $(\Delta g_x, \Delta g_y)$ across eye images does reveal the iris movement. Especially, it conveys *directions* of short-range eyeball/iris movements, as shown in Fig. 5. For instance, a vertical gravity center movement indicates a upward/downward eyeball movement. Therefore, we use this partial information to resolve the ambiguities, i.e., rotation, flip and distortion, in the 2D embedding results.

First, we use matrix $\boldsymbol{A} = [\eta, 1, 1] \cdot \boldsymbol{R} + \boldsymbol{H}$ to model the transformation from the 2D embedding results to a correct gaze pattern, where $\eta = \pm 1$ denoting a flip operation, and

$$\begin{aligned} \boldsymbol{R} &= \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}, \\ \boldsymbol{H} &= \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix}, \end{aligned} \tag{8}$$

for rotation and perspective transformations. Note that the

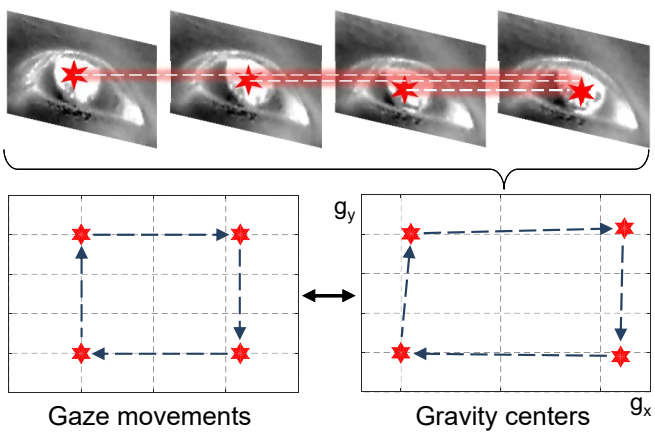

Fig. 5: Gravity centers of inverted eye images show similar movements with their corresponding eye gazes.

rotation transformation $\boldsymbol{R}$ is also a special case of perspective transformation. However, $\boldsymbol{R}$ and $\boldsymbol{H}$ are defined separately here since we will solve their corresponding subproblems separately later.

In order to compute the unknown transformation, we collect a set of gravity center pairs $\{(\mathbf{g}_p, \mathbf{g}_q)\}$ where the motion between any $\mathbf{g}_p$ and $\mathbf{g}_q$ is small to ensure the data reliability. By assuming the same direction of $\mathbf{g}_p \rightarrow \mathbf{g}_q$ and the correctly transformed $\mathbf{z}_p \rightarrow \mathbf{z}_q$, we can write down the following formulation

$$\min_{\boldsymbol{A}} \sum_{p,q} \cos^{-1} \frac{(f(\boldsymbol{A}, \mathbf{z}_p) - f(\boldsymbol{A}, \mathbf{z}_q)) \cdot (\mathbf{g}_p - \mathbf{g}_q)}{|f(\boldsymbol{A}, \mathbf{z}_p) - f(\boldsymbol{A}, \mathbf{z}_q)| \, |\mathbf{g}_p - \mathbf{g}_q|}, \quad (9)$$

where the arccosine of two normalized vectors' inner product measures their difference in direction, and the function $f(\mathbf{A}, \mathbf{z})$ performs homogeneous transformation to transform any point $\mathbf{z}$ into its new position according to the transformation matrix $\mathbf{A}$. Therefore, minimizing Equation (9) ensures that image gravity center movements and gaze movements always follow the same direction.

Solving Equation (9) is nontrivial due to its high complexity. More importantly, although transformations w.r.t. η and $\boldsymbol{R}$ can preserve distances as we expect, the non-distance-preserving transformation $\boldsymbol{H}$ may affect the recovered structure excessively and should be suppressed. Therefore, we first solve a subproblem without considering $\boldsymbol{H}$

$$\min_{\boldsymbol{R},\eta} \sum_{p,q} \cos^{-1} \frac{[\eta, 1]^{\mathrm{T}} \cdot \boldsymbol{R}_{2\times 2} \cdot (\mathbf{z}_p - \mathbf{z}_q) \cdot (\mathbf{g}_p - \mathbf{g}_q)}{|\mathbf{z}_p - \mathbf{z}_q| \, |\mathbf{g}_p - \mathbf{g}_q|}. \quad (10)$$

Since the only unknowns are the rotation angle θ and $\eta = \pm 1$, Equation (10) can be easily solved. Then, by applying the known transformations with θ and η, the 2D embedding coordinates become $\{\hat{\mathbf{z}}_p = [\eta, 1]^{\mathrm{T}} \cdot \boldsymbol{R}_{2\times 2} \cdot \mathbf{z}_p\}$.

To handle the rest perspective distortion, we solve

$$\min_{\boldsymbol{H}} \sum_{p,q} \cos^{-1} \frac{(f(\boldsymbol{H}, \hat{\mathbf{z}}_p) - f(\boldsymbol{H}, \hat{\mathbf{z}}_q)) \cdot (\mathbf{g}_p - \mathbf{g}_q)}{|f(\boldsymbol{H}, \hat{\mathbf{z}}_p) - f(\boldsymbol{H}, \hat{\mathbf{z}}_q)| \, |\mathbf{g}_p - \mathbf{g}_q|} + \lambda \|\boldsymbol{H} - \boldsymbol{I}_3\|_F, \quad (11)$$

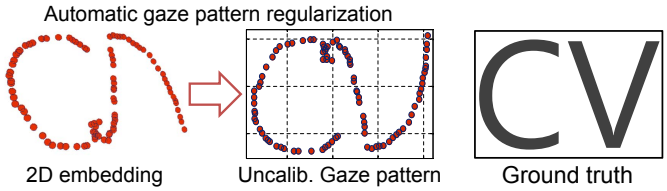

Fig. 6: We obtain the uncalibrated gaze pattern from 2D embeddings by using the proposed gaze regularization. The result shows a similar spatial distribution with the gaze positions.

where $\boldsymbol{I}_3$ is the identity matrix, $\lambda\|\boldsymbol{H} - \boldsymbol{I}_3\|_F$ controls the strength of the perspective distortion and λ is set to be the number of pairs in $\{(\mathbf{g}_p, \mathbf{g}_q)\}$. This problem can be solved by respectively updating each of the eight unknown elements of $\boldsymbol{H}$. Finally, the output gaze pattern comprises points $\{\tilde{\mathbf{z}}_p = f(\boldsymbol{H}, \hat{\mathbf{z}}_p)\}$.

An example of the uncalibrated gaze pattern is shown in Fig. 6. It is highly identical to the real eye gaze pattern. Note that the uncalibrated gaze pattern is computed from only eye images; no calibration or training procedure/data is needed.

IV. Calibration

By now, the uncalibrated gaze pattern, denoted as $\{\tilde{\mathbf{z}}_p\}$, already recovers relative gaze movements in the 2D virtual plane. This is sufficient for many applications to do gaze pattern analysis. However, in some other cases we need real-world gaze coordinates in the scene. This requires resolving the scene-dependent ambiguities, i.e., scaling and translation, which are the only ones left from the last section.

Disambiguation can be done by aligning the gaze pattern to the real-world gaze target via global scaling and translations. For instance, if consider the most common case where the gaze target is a monitor screen, the alignment can be

$$\underset{s_x, s_y, t_x, t_y}{\operatorname{argmax}} \; s_x \cdot s_y, \quad \text{s.t.} \quad \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix} \tilde{\mathbf{z}}_p + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \in \mathbb{S}, \quad (12)$$

where s_x,s_y, t_x and t_y are the only parameters for scaling and translations, and $\mathbb{S}$ indicates the screen area. At the core of Equation (12) is the assumption that the contour points of the gaze pattern reach the screen boundaries, while other points need no consideration since their relative positions have already been fixed in the gaze pattern accurately. As a result, the alignment can be extremely easy to do. In practice, a user may choose to use a little gaze control to glance through the screen boundaries to satisfy the above assumption. This process neither costs time nor interrupts the task. These collected boundary point can also help detect outliers, e.g.fixations outside the screen area, during the task.

Equation (12) is just an example of simple and flexible calibration. It is totally unrestricted for users to modify Equation (12) or develop many feasible calibration methods based on their own assumptions to align the gaze pattern. Because there are only four simple parameters to solve, the problem will always be easy to do.

V. Evaluations

In this section we test our method under various settings and scenes. We show how to perform gaze estimation in conventional scenarios, as well as how to use the uncalibrated gaze pattern alone without calibration. We conduct quantitative and qualitative experiments to demonstrate the efficiency and broad capabilities of our method.

A. Accuracy in Common User-Screen Scenario

The user-screen scenario is commonly assumed in previous researches for quantitative evaluation. In such cases, because ground truths of gaze positions on the screen are available, quantitative evaluations becomes possible.

We implement our system with a desktop PC and a 22-inch LCD monitor. A common web camera is located under the monitor and used to capture user appearances at a resolution of 1280×1024. During our experiments, a typical user eye-screen distance is 600 mm.

Twelve users participate in our experiments. They are asked to fixate at each point shown on the screen. These points appear in a random order and roughly cover the screen area. User face appearances are captured by the web camera during each eye fixation. These images are used as test samples for quantitative evaluation. *No training data* are captured.

Analyses on the experimental data are done with non-optimized Matlab codes. It takes around one second to process $100+$ test samples for each user. Estimation errors are measured by angular gaze differences

$$error \approx \arctan\left(\frac{\|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2}{D}\right), \tag{13}$$

where D is the eye-screen distance, $\hat{\mathbf{x}}_i$ is the estimated screen gaze position and $\mathbf{x}_i$ is its ground truth.

Moreover, our experiments are done in two stages. First, a chinrest is used to ensure a fixed user head pose. Then, the user keeps a natural head pose without the chinrest.

Fixed head pose In this stage, we collect data for all users. We also change the gaze positions' density on the screen to include 140, 90, 66, 33 and 18 points each time.

Uncalibrated gaze patterns are computed by using the captured test eye images of each user, and then aligned to the screen coordinates. The alignment is done by using Equation (12) because the gaze positions in our data extend to the screen boundary areas. In practice, as explained in Section II-C, a user can also use a quick gaze control to glance over screen boundaries to help the alignment while this does not cost extra time or interrupt the task.

Quantitative results for all users are summarized in Table II. In average, we obtain an accuracy of $2.57°$ when using all 140 images. This is a reasonably good accuracy even compared with fully-calibrated methods. Furthermore, Table II also shows results obtained using fewer images. We observe that reducing the sample number to 66 still assures the accuracy. However, a smaller number than 33 can significantly increase the gaze error by 20%. In the final case, an average error of $3.40°$ is obtained with only 18 input images.

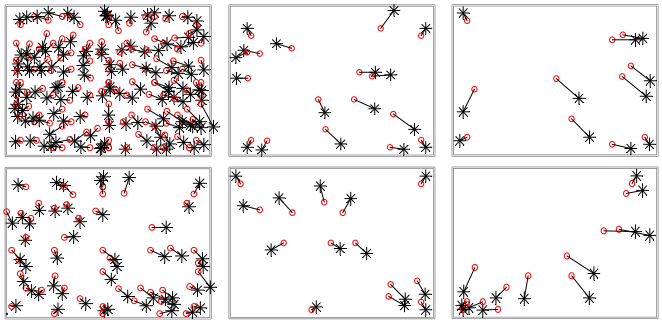

Fig. 7: Examples of gaze estimation results on the screen. Circles indicate groundtruth gaze positions and stars are our estimates.

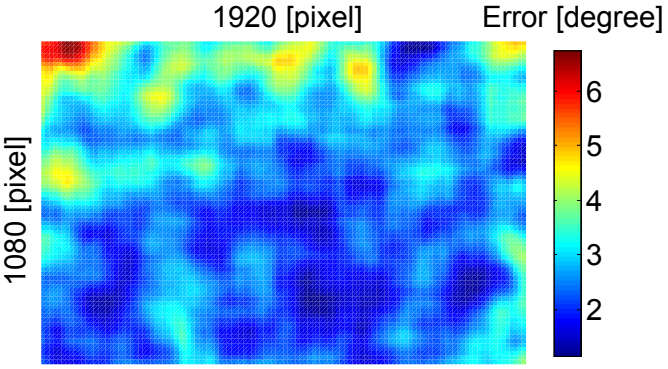

Fig. 8: Average on-screen error map of all subjects.

From these results, we conclude that in a common user-screen scenario, around 100 eye images for distributed gaze positions allow our method to achieve its best performance. With fewer images, the accuracy drops but it is still reasonable. In fact, our method achieves $3.40°$ with only 18 test images, which are even fewer than the number of images required in the training stage of conventional methods.

Moreover, to evaluate the proposed alignment, we introduce a 'Lower Bound (L.B.)' method. It has four additional samples, which correspond to the four known screen corner positions, to help align the gaze pattern to the screen ideally. Table II shows the results, where the errors of our method approximate these lower bound errors well, and the gaps are especially small when the input images are adequate.

We also study individual cases regarding gaze pattern alignment. Different sample numbers and distributions are examined including some extreme cases. Representative results are visualized in Fig. 7. Since the proposed uncalibrated gaze patterns are determined up to only global scaling and translation ambiguities, they can be easily aligned to the screen region even if they do not spread over the entire screen.

Finally, we compute and show averaged error map of all subjects in Fig. 8. It is clear that large errors tend to appear in the boundary areas of the screen.

On dimension reduction-based methods The proposed method is closely related to the dimension reduction techniques. For the sake of comparison, we conduct experiments by using two other typical dimension reduction methods: PCA

TABLE II: Average gaze estimation errors by using different number of input image. Results of other dimension reduction-based methods, i.e., PCA and LLE, are also compared.

Images	Proposed method														PCA [9]	LLE [33]
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Avg.	L.b.	Avg.	Avg.
140	2.20°	1.87°	2.36°	2.94°	2.65°	2.93°	2.67°	1.92°	2.64°	2.81°	2.88°	3.01°	2.57°	2.49°	4.48°	3.45°
90	1.97°	1.90°	2.23°	3.09°	3.26°	3.13°	2.67°	1.88°	2.48°	2.89°	3.88°	2.76°	2.68°	2.37°	4.40°	3.39°
66	2.22°	1.99°	2.17°	2.76°	2.81°	2.80°	2.72°	1.91°	2.67°	2.92°	2.98°	2.73°	2.56°	2.22°	4.20°	3.30°
33	2.34°	2.40°	3.03°	3.06°	3.87°	2.51°	2.78°	1.75°	2.70°	2.75°	4.70°	3.23°	2.93°	2.36°	4.96°	3.38°
18	2.34°	2.50°	3.62°	3.62°	4.18°	3.54°	2.76°	2.35°	3.21°	3.32°	5.69°	3.74°	3.40°	2.71°	4.43°	3.71°

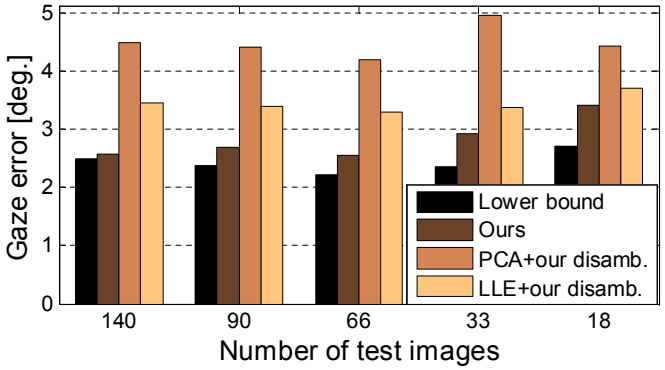

Fig. 9: Average on-screen gaze errors with different number of input eye images by using different methods.

TABLE III: Comparison with previous methods on the same dataset.

Method	Error	Calibration
Ours	$2.57 \pm 0.38°$	Align to screen, $\approx$ **1**s
ALR [23]	$0.65 \pm 0.42°$	Stop user task, and
HOG&SVR [34]	$0.99 \pm 0.56°$	display & capture, and
CSLBP&GPR [35]	$1.37 \pm 0.80°$	33 training images, and
PCA&GPR [36]	$1.22 \pm 0.63°$	3s$\times$33$\approx$ **100**s
Saliency-based [8], [9]	$3 \sim 5°$[1]	Watch full video with known saliency
Dimension reduction-based [26]	$3 \sim 6°$[1]	Training for multiple users in the same scene

as in [9] and LLE [33]. They are widely used in previously related researches for linear/nonlinear dimension reduction and can be directly compared with ours. Note that we allow them to use our technique in Section III-C to resolve the rotation/flip ambiguities so that the comparison can be focused on dimension reduction itself. Results can be seen in Table III, and we also plot their averages in Fig. 9. Our method clearly outperforms the other two methods under exactly the same setting. This is because our method better extracts the large scale linearity in the nonlinear gaze space as explained in Section III-B.

Comparison with other typical methods We test the recent appearance-based methods on our dataset. Because they require training-based calibration, we choose 33 images with evenly distributed gaze points as training samples and use the rests as test samples. We only test fixed head pose methods because they produce the best accuracy. Average results are given in Table III. In terms of accuracy, our method generally cannot beat the conventional calibration-based methods but they are still comparable. However, in terms of calibration, our method shows merits in that it only uses test data as input; no extra training data is needed and there is no explicit calibration stage before every session. While the conventional calibration-based methods share all those limitations listed in Table III, which may be unbearable in various practical applications.

Table III also show methods based on saliency and dimension reduction. The saliency-based methods are scenario-dependent, and images/videos with known saliency maps are assumed to control the user's gaze movement. The method [26] assumes multiple users' data in the same scene with the same gaze distribution. Overall, they introduce additional assumptions about the scene. In contrast, our method only captures and uses eye images of a single person in an unspecified scene.

Regarding the computational time, solving for the gaze pattern with N points has a complexity of $O(N^2)$ since our method computes pair-wise distances of N images. In particular, for 140 points it costs $<$ 1s with a Matlab implementation, while aligning it to the screen by solving Equation (12) costs less than 0.1s. It shows a good computational efficiency.

Natural head pose with slight motion Head motion is another major problem in appearance-based gaze estimation. Some previous methods require more training samples [28], [27], [10], while we suggest a different solution.

In this experiment we do not use a chinrest. Each user keeps a comfortable head pose during a time period and gazes at 110 screen points, while their eye images are captured. Unconscious head motions exist which are relatively small. Examples of slight head motions are shown in Fig. 10. Typical head movements are ranged $10 \sim 30$ mm.

Quantitative results are provided in Table IV, where our method achieves an average accuracy of 2.37°. Compared to the fixed head pose results in Table II, the accuracy remains high. To understand the reason, we modify our method by

[1]We use their reported errors since our database does not contain the saliency information or special training data they need.

Fig. 10: Examples of slight head motions in our experiments. Green boxes are shown to help notice the motion of eye regions.

TABLE IV: Gaze estimation errors under slight head motion

	Subj. 1	Subj. 2	Subj. 3	Subj. 4	Subj. 5	Subj. 6	Subj. 7	Subj. 8	Avg.
Aligned	2.44°	1.65°	2.37°	3.04°	2.16°	3.03°	2.13°	2.13°	$2.37 \pm 1.25^\circ$
Not aligned	19.60°	7.21°	8.47°	9.64°	18.59°	10.56°	11.55°	16.56°	$12.77 \pm 5.71^\circ$

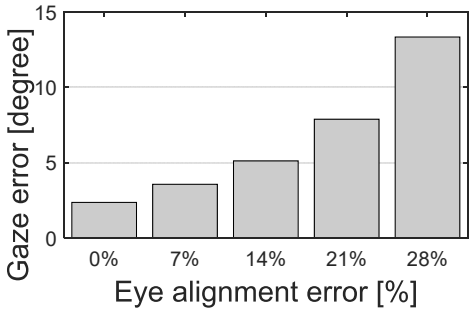

Fig. 11: Relationship between eye image alignment error and the corresponding gaze estimaiton error.

cropping eye images at fixed pixel coordinates without eye alignment (Section III-A), and the gaze estimation errors increase drastically as shown in Table IV. These results demonstrate that slight head motion affects gaze error mainly because it introduces pixel biases to the captured eye images; meanwhile, our method can handle slight head motion well because of the effective eye image alignment technique.

In term of large head pose changes, which can be detected by significant changes in images, the uncalibrated gaze pattern should be re-computed and aligned to the screen to maintain a high accuracy. As explained before, this process is very fast without interrupting the current task, which shows advantages over conventional methods. In other words, gaze recovery for different head poses is independently processed. For this reason, we do not provide extra results; readers can simply check the results in Table IV which are independently computed for different head poses of different users as a reference.

On eye alignment accuracy Our method relies on accurate eye alignment and it may fail otherwise. To understand how sensitive our method is to eye alignment, we provide additional results using the misaligned eye image data from the previous experiments. Different degrees of misalignment (added to our alignment results) are grouped by pixel shifting percentage, and the corresponding gaze errors are shown in Fig. 11. It can be concluded that: 1) achieving good eye image alignment is clearly important to the final accuracy, and 2) our alignment ("0%" in Fig. 11) is effective.

B. Evaluation on the Usage of Uncalibrated Gaze Pattern

The previous section evaluates our method in conventional ways to demonstrate its advantages, i.e., user/scene-independence and minimum calibration burden. While in this section, we show that it is even possible to use the proposed uncalibrated pattern alone without any calibration, which is a novel task that has not been reported before.

For this purpose, we use some simple but representative gaze patterns. In particular, we design seven basic gaze patterns as shown in Fig. 12. Note that we do not include complex gaze patterns with dense gaze positions for the following reasons. 1) Dense gaze positions have been tested in Section V-A, showing a good accuracy. 2) More importantly, simple gaze patterns are more likely to be used as meaningful symbols in real scenarios.

For each user, eye images are captured when he/she moves eye gaze following the pattern shown on the screen. The number of captured eye images has a range from 15 to 35. We then compute the uncalibrated gaze pattern, denoted as $\{\tilde{\mathbf{z}}_p\}$, as described in Section III-B and Section III-C. Examples of the recovery results are shown in Fig. 12, under the ground truths, which are visually correct. However, computing their quantitative errors is not straightforward because these results are not converted into screen coordinates. Therefore, we resize $\{\tilde{\mathbf{z}}_p\}$ and its ground truth to the same size by fitting their bounding boxes into a unit square. Then the matching cost for the l-th gaze pattern can be computed by

$$c_l = \sum\nolimits_p \|\tilde{\mathbf{z}}_p - \mathbf{w}_{m(p)}\|^2, \tag{14}$$

$$m(p) = \underset{m(p)=1,\cdots,N}{\operatorname{argmin}} \|\tilde{\mathbf{z}}_p - \mathbf{w}_{m(p)}\|^2, \tag{15}$$

where $\{\mathbf{w}_n | n = 1, \cdots, N\}$ are the dense and ordered point positions from the ground truth of the l-th gaze pattern.

The matching costs $\{c_l\}$ can be sorted and the best matching can be found with the minimum c_l. Consequently, we check the matching correctness. Experiments are done for seven patterns and thirteen users. The average matching accuracy is 81.32%, as shown in Fig. 13, which confirms the correctness of gaze pattern recovery. In practice, one can further improve

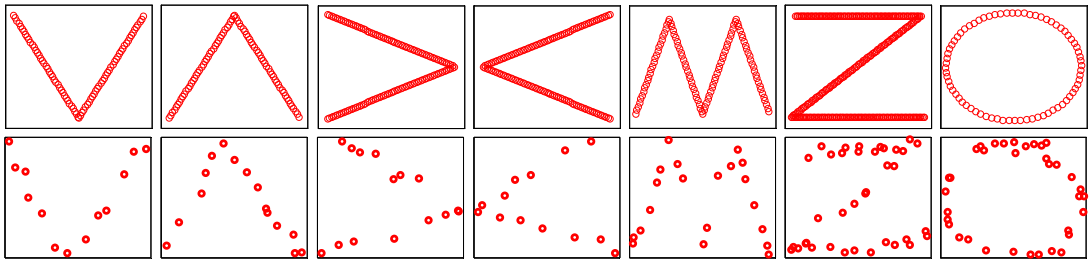

Fig. 12: Seven gaze patterns in our experiment: ground truths (top row) and examples of our recovery results using only captured test eye images (bottom row).

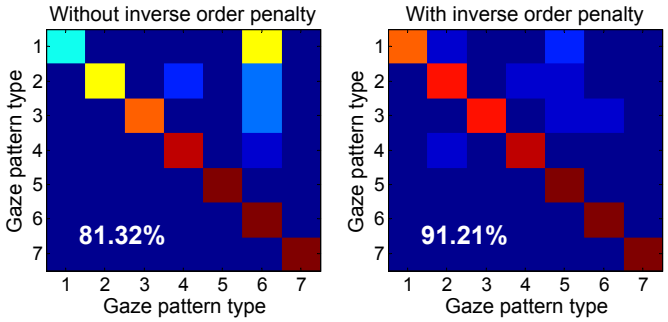

Fig. 13: Gaze pattern recognition results. Confusion matrices show accuracy with/without using inverse order penalty.

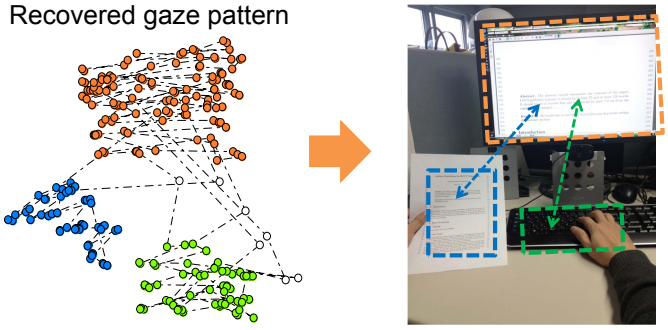

Fig. 14: Preliminary example: uncalibrated gaze pattern on multiple objects can be aligned to the real scene via post-calibration.

the accuracy by, e.g., assuming the gaze order:

$$\hat{c}_l = \sum\nolimits_p \|\tilde{\mathbf{z}}_p - \mathbf{w}_{m(p)}\|^2 + \alpha \sum\nolimits_p \max(0, m(p) - m(p+1)), \tag{16}$$

where the second term introduces inverse order penalty. By using $\hat{c}_l$ in gaze pattern matching, the accuracy raises to 91.21%, as shown in Fig. 13. It indicates that our method can successfully recover uncalibrated gaze patterns by preserving their unique structures and has a good potential for future practical use.

VI. Discussion and Conclusion

This paper proposes a novel method, whose performance has been carefully evaluated from different aspects. Besides, the proposed method shows multiple new features that can motivate novel and promising applications. Although they are not our focus in this paper, we briefly mention some potential applications here as discussions.

1) Uncalibrated gaze patterns without training: may be used as input methods in the future, eg, gaze gesture for human computer interaction. Besides, gaze pattern, rather than exact gaze positions, provides important cues in human cognitive researches [37][2]. In such cases, the proposed uncalibrated gaze pattern already provides sufficient information and it avoids conventional calibration.

2) Post/offline-calibration: it removes the traditional assumption that calibration should be done online before task. For instance, for reading analysis, we can simply recover the reader's gaze scanning pattern first, and then align it to the (book/screen) content for analysis in an offline stage. This avoids the burdensome calibration stage for the reader [38]. Besides, post-calibration can allow more flexible scenes like that in Fig. 14, which may be unknown before the task.

3) Fast calibration: if re-calibration is frequently required, e.g., when using a mobile phone with frequent head pose changes, our method can allow continuous estimations.

The above features are not seen in conventional appearance-based methods and will enable novel applications. As for the limitation, a large head pose change will increase the error of our method greatly in a similar manner to conventional methods. Although a quick re-start can solve this problem and it is significantly faster than re-calibrating a conventional method, further researches should be conducted for a better solution to handle large head motions.

In conclusion, we solve the appearance-based gaze estimation problem in a less person/scene-dependent manner. We first recover an uncalibrated gaze pattern solely from eye images, and then align it to various gaze targets via a simple calibration. Since the uncalibrated gaze pattern determines the relative gaze positions, the rest calibration can be simple and flexible. Our method produces promising results and shows broad capabilities in potential applications. Future works include further improving its performance and designing practical applications based on the proposed techniques.

[2]Readers may refer to Figure 3 in [37] to justify this statement.

References

[1] M. Liang and X. Hu, "Predicting eye fixations with higher-level visual features," *IEEE Transactions on Image Processing*, vol. 24, no. 3, pp. 1178–1189, 2015.

[2] U. Rajashekar, I. van der Linde, A. Bovik, and L. Cormack, "Gaffe: A gaze-attentive fixation finding engine," *IEEE Transactions on Image Processing*, vol. 17, no. 4, pp. 564–573, 2008.
[3] X. Sun, H. Yao, R. Ji, and X.-M. Liu, "Toward statistical modeling of saccadic eye-movement and visual saliency," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4649–4662, 2014.
[4] Q. Cheng, D. Agrafiotis, A. Achim, and D. Bull, "Gaze location prediction for broadcast football video," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 4918–4929, 2013.
[5] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *ECCV*, 2012, pp. 314–327.
[6] B. Benfold and I. Reid, "Unsupervised learning of a scene-specific coarse gaze estimator," in *ICCV*, 2011, pp. 2344–2351.
[7] D. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Trans. on PAMI*, vol. 32, no. 3, pp. 478–500, 2010.
[8] Y. Sugano, Y. Matsushita, and Y. Sato, "Appearance-based gaze estimation using visual saliency," *PAMI*, vol. 35, no. 2, pp. 329–341, 2013.
[9] F. Alnajar, T. Gevers, R. Valenti, and S. Ghebreab, "Calibration-free gaze estimation using human gaze patterns," in *ICCV*, 2013.
[10] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike, "Appearance-based gaze estimation with online calibration from mouse operations," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 6, pp. 750–760, 2015.
[11] C. Morimoto and M. Mimica, "Eye gaze tracking techniques for interactive applications," *CVIU*, vol. 98, no. 1, pp. 4–24, 2005.
[12] E. Guestrin and M. Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections," *IEEE Trans. on Biomedical Engineering*, vol. 53, no. 6, pp. 1124–1133, 2006.
[13] D. Beymer and M. Flickner, "Eye gaze tracking using an active stereo head," in *CVPR*, 2003, pp. 451–458.
[14] Z. Zhu and Q. Ji, "Novel eye gaze tracking techniques under natural head movement," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 12, pp. 2246 –2260, 2007.
[15] A. Nakazawa and C. Nitschke, "Point of gaze estimation through corneal surface reflection in an active illumination environment," in *ECCV*, 2012, pp. 159–172.
[16] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Trans. on IP*, vol. 21, pp. 802–815, 2012.
[17] J. Wang, E. Sung, and R. Venkateswarlu, "Eye gaze estimation from a single image of one eye," in *ICCV*, 2003, pp. 136–143.
[18] K. A. Funes Mora and J.-M. Odobez, "Geometric generative gaze estimation (g3e) for remote rgb-d cameras," in *IEEE Computer Vision and Pattern Recognition Conference*, 2014, pp. 1773–1780.
[19] X. Xiong, Q. Cai, Z. Liu, and Z. Zhang, "Eye gaze tracking using an rgbd camera: A comparison with an rgb solution," in *The 4th International Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction (PETMEI 2014)*, 2014.
[20] S. Baluja and D. Pomerleau, "Non-intrusive gaze tracking using artificial neural networks," in *NIPS*, 1994, pp. 753–760.
[21] K. Tan, D. Kriegman, and N. Ahuja, "Appearance-based eye gaze estimation," in *WACV*, 2002, pp. 191–195.
[22] O. Williams, A. Blake, and R. Cipolla, "Sparse and semi-supervised visual mapping with the S^3GP," in *CVPR*, 2006, pp. 230–237.
[23] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Inferring human gaze from appearance via adaptive linear regression," in *ICCV*, 2011.
[24] ——, "Adaptive linear regression for appearance-based gaze estimation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 10, pp. 2033–2046, 2014.
[25] J. Chen and Q. Ji, "Probabilistic gaze estimation without active personal calibration," in *CVPR*, 2011, pp. 609–616.
[26] T. Schneider, B. Schauerte, and R. Stiefelhagen, "Manifold alignment for person independent appearance-based gaze estimation," in *International Conference on Pattern Recognition (ICPR)*, 2014, pp. 1167–1172.
[27] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Head pose-free appearance-based gaze sensing via eye image synthesis," in *ICPR*, 2012.
[28] F. Lu, T. Okabe, Y. Sugano, and Y. Sato, "Learning gaze biases with head motion for head pose-free gaze estimation," *Image Vision Comput.*, vol. 32, no. 3, pp. 169–179, 2014.
[29] S. Machines, "faceAPI," http://www.seeingmachines.com.
[30] S. He, Q. Yang, R. W. Lau, J. Wang, and M.-H. Yang, "Visual tracking via locality sensitive histograms," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2427–2434.
[31] J. Tenenbaum, V. De Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
[32] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. Academic Press, 1979.
[33] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[34] F. Martinez, A. Carbone, and E. Pissaloux, "Gaze estimation using local features and non-linear regression," in *ICIP*, 2012, pp. 1961–1964.
[35] K. Liang, Y. Chahir, M. Molina, C. Tijus, and F. Jouen, "Appearance-based gaze tracking with spectral clustering and semi-supervised gaussian process regression," in *ETSA*, 2013, pp. 17–23.
[36] B. Noris, K. Benmachiche, and A. Billard, "Calibration-free eye gaze direction detection with gaussian processes," in *VISAPP*, 2008, pp. 611–616.
[37] K. Kunze, Y. Utsumi, Y. Shiga, K. Kise, and A. Bulling, "I know what you are reading: recognition of document types using mobile eye tracking," in *Proceedings of 17th annual international symposium on International symposium on wearable computers*. ACM, 2013, pp. 113–116.
[38] K. Kunze, H. Kawaichi, K. Yoshimura, and K. Kise, "The wordometer–estimating the number of words read using document image retrieval and mobile eye tracking," in *ICDAR*. IEEE, 2013, pp. 25–29.

Feng Lu received the B.S. and M.S. degrees in automation from Tsinghua University, in 2007 and 2010, and the Ph.D. degree in information science and technology from The University of Tokyo, in 2013. 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 research interests including human gaze analysis, 3D shape recovery, and reflectance analysis.

Xiaowu Chen 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.

Yoichi Sato received the BS degree from the University of Tokyo in 1990, and the MS and PhD degrees in robotics from the School of Computer Science, Carnegie Mellon University, in 1993 and 1997, respectively. He is a professor at the Institute of Industrial Science, the University of Tokyo. His research interests include physics-based vision, reflectance analysis, and gaze and gesture analysis. He served/is serving in several conference organization and journal editorial roles including TPAMI and IJCV.