Pattern Recognition 79 (2018) 216–227

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

3D gaze estimation without explicit personal calibration

Kang Wang*, Qiang Ji

Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, 110 Eighth Street, Troy, NY 12180, USA

ARTICLE INFO

Article history:
Received 26 July 2017
Revised 24 January 2018
Accepted 28 January 2018
Available online 31 January 2018

Keywords:
Gaze estimation
Implicit calibration
Natural constraints
Human computer interaction

ABSTRACT

Model-based 3D gaze estimation represents a dominant technique for eye gaze estimation. It allows free head movement and gives good estimation accuracy. But it requires a personal calibration, which may significantly limit its practical utility. Various techniques have been proposed to replace intrusive and subject-unfriendly calibration methods. In this paper, we introduce a new implicit calibration method that takes advantage of four natural constraints during eye gaze tracking. The first constraint is based on two complementary gaze estimation methods. The underlying assumption is that different gaze estimation methods, though based on different principles and mechanisms, ideally predict exactly the same gaze point at the same time. The second constraint is inspired by the well-known center prior principle, it is assumed that most fixations are concentrated on the center of the screen with natural viewing scenarios. The third constraint arises from the fact that for console based eye tracking, human's attention/gaze are always within the screen region. The final constraint comes from eye anatomy, where the value of eye parameters must be within certain regions. The four constraints are integrated jointly and help formulate the implicit calibration as a constrained unsupervised regression problem, which can be effectively solved through the proposed iterative hard EM algorithm. Experiments on two everyday interactions Web-browsing and Video-watching demonstrate the effectiveness of the proposed implicit calibration method.

1. Introduction

Eye gaze tracking is to track human's attention or predict where human looks in real time. Eye gaze tracking technology has been applied to various fields. In Human Computer-Interaction, eye gaze can replace traditional actions using mouse clicks and make interactions fast, fun and totally natural. For instance, with an eye gaze tracking system, we can zoom where we look at, text scrolls as we read, etc. This is much faster and natural than the traditional input. Eye tracking data can also help user-behavior study [1], medical research [2,3], understand human's cognitive process [4] etc.

Various techniques have been proposed to perform eye gaze tracking. Gaze estimation methods can be divided into model-based methods and regression-based methods. Model-based methods [5–11] build a 3D eye model according to the anatomy of human eyes/faces. Leveraging on the geometric relations among different facial and eye features (facial landmarks, cornea, pupil, etc), 3D gaze direction can be computed. Along with this direction, different feature extraction approaches have been proposed [12–15]. 3D model-based methods mimic the human vision system and compute the exact gaze direction as the human brain does. Known for their accuracy and ability to handle head movement, 3D model-based methods are being widely used nowadays in many commercial eye trackers. Since model-based methods require the knowledge of human eyes and related parameters, a personal-calibration is necessary to achieve good accuracy. However, personal calibration requires explicit collaboration from the user, which makes eye tracking system unfriendly to use and degrades the user experience.

Regression-based methods leverage on powerful learning techniques and assume mappings from eye appearance/features to gaze positions/directions. Compared to model-based approaches, they avoid modeling the complex eyeball structure and only require collecting a large amount of data. Regression-based methods can be further divided into feature-based and appearance-based methods. Feature-based regression methods [16–19] learn the mapping function from eye features to gaze positions/directions. Typical eye features include pupil-glint vectors, pupil-eye corner vectors, cross-ratios among glints, etc. Appearance-based regression methods [20–23] learn a mapping function from eye appearances to gaze positions/directions. The learning algorithms range from traditional support vector regression, random forest to most recent deep learning techniques. However, regression-based methods typically suffer from head movement issues without using extra data

* Corresponding author.
E-mail addresses: wangk10@rpi.edu (K. Wang), qji@ecse.rpi.edu (Q. Ji).

https://doi.org/10.1016/j.patcog.2018.01.031
0031-3203/© 2018 Elsevier Ltd. All rights reserved.

to compensate the movement. Besides, learning algorithms also require a large amount of data to learn a good mapping function. We suggest readers refer to [24] for more detailed discussion on different eye gaze tracking approaches.

Both model-based and regression-based methods require personal calibration. The calibration procedure requires explicit collaboration from users, which may not be applicable for certain applications (eye tracking for babies). For users who are capable of collaboration, the procedure is intrusive and degrades the user experience. To eliminate explicit personal calibration, we propose to better utilize information during natural human-computer interactions. Despite the importance of the information, it is typically ignored during eye gaze tracking. The information can be obtained in the backend while subjects naturally operate on their computers, making eye gaze tracking more fun and friendly. Specifically, we formulate four constraints from the information. The first one is the complementary gaze constraint, which is inspired by the binocular-constraint introduced in [25]. Binocular constraint states that gaze positions estimated from two eyes should be exactly the same. But the use of binocular constraint limits head movement since it requires both eyes in the view of the camera. Differently, we assume two gaze estimation methods predict exactly the same gaze point at the same time. The two methods are based on different principles and mechanisms but are complementary to each other. The two methods we choose are the 3D model-based method and the feature-based regression method. The second constraint comes from the well-known center prior principle. It is assumed that most gaze fixations are concentrated near the screen center while users watch videos. Third, it is assumed human's attention/gaze are always within the screen region for a period of time. Finally, from human eye anatomy, personal eye parameters must have reasonable values. These four constraints are integrated into a constrained unsupervised learning problem, which can be effectively solved through the proposed iterative hard-EM algorithm.

Compared with existing work on reducing/eliminating explicit personal calibration, the proposed method makes following novel contributions:

- A non-intrusive and user-friendly eye gaze tracking system is proposed.
- Personal eye parameters can be implicitly calibrated with natural constraints.
- Propose the hard-EM algorithm to solve the constrained unsupervised regression problem.
- The proposed method achieves comparable gaze estimation accuracy with state-of-the-art implicit calibration methods, while is less restricted and can be applied to a wider range of practical applications.

2. Related work

Much work has been done to reduce/eliminate explicit personal calibration for model-based methods. Guestrin and Eizenman [7] proposed a 1-point calibration method with two cameras and four IR lights. By exploiting eye geometry knowledge, their system only has two unknown personal parameters. Thus 1 reference point which gives two equations is sufficient to solve the two eye parameters. However, their method still requires explicit collaboration from users. Model and Eizenman [25] proposed solve the two personal eye parameters with the help of binocular constraint. They assume the gaze directions from two eyes intersect on the same gaze point on the display device. However, their method is limited in applications with larger displays and cannot produce good results for general usage with small displays. Maio et al. [22] proposed to alleviate the problem for ordinary display (36 cm × 28.7 cm) by introducing additional generic person-independent constraints to the framework. However, binocular-constraint based methods limit head movement since two eyes are required to be captured by the camera and the experimental settings are rather complex.

Chen and Ji [26] proposed to eliminate explicit calibration with the help of saliency map. A Bayesian network is built to represent the probabilistic relationship among optical axis, visual axis, and eye parameters. User's attention is assumed to be captured by the saliency map, from which eye parameters can be estimated by maximizing the posterior given the observed optical axes. Later, Chen and Ji [27] extended the work to use general Gaussian distribution (center prior) as their prior model to alleviate issues from saliency map. However, the use of center prior limits potential applications that exist strong center-biased gaze patterns like watching videos/images. Besides, their algorithm also takes a longer time to converge. Recently Wang et al. [28] proposed to leverage on the fixation map learned from a deep model. By minimizing the KL divergence between the user gaze distribution (a function of eye parameters) and underlying fixation map, they are able to recover the eye parameters. Despite the advantage of using fixation map instead of saliency map, the method also requires explicit user collaboration to look at saliency content and the computation of fixation map takes time.

Alnajar et al. [29] proposed to leverage on human gaze pattern to eliminate personal calibration. It is assumed that different subjects tend to have similar gaze patterns on the same stimuli. Therefore off-line learned gaze patterns can be used to estimate the regression coefficients for a new subject. However, the underlying assumption remains too strong and unrealistic. Different types, content of the stimulus may result in different gaze patterns for different subjects, therefore the proposed method may not be applied to real-world applications. Lu et al. [20] proposed another similar idea related to gaze pattern. By exploring the 2D gaze manifold, they are able to recover the relative gaze positions purely from 2D eye appearance. The recovered uncalibrated gaze pattern can be mapped to true gaze positions with task-dependent domain knowledge. However, their method cannot handle head movement, the 2D gaze manifold assumption does not hold anymore when head motion exists. Sugano et al. [30] proposed to implicitly collect training data using mouse clicks. They assumed that users would unconsciously look at the cursor when they click the mouse. Eye appearance and gaze position pairs are implicitly collected and used to learn the regression parameters. However, besides cases that users might not look at the cursor when they click the mouse, a lot of applications/interactions only require a few mouse clicks (watching videos, reading articles on the website, etc). The calibration algorithm might require a much longer time to converge in such scenarios. Later, Sugano et al. [31] introduced a visual saliency-based calibration framework. Similar to their previous work, by assuming users look at salient region/objects of video frames, they were able to implicitly collect appearance/gaze position pairs and use them to compute regression parameters. However, it also suffers the common issues for saliency-based applications. Pfeuffer et al. [32] proposed to make calibration more flexible and less tedious by using moving targets. The correlation between eye movement and target trajectory is explored to perform implicit calibration. However, the assumption might not hold in practice and subjects need to focus on the moving target for a period of time.

In summary, existing methods still require certain level collaboration from users or make strong assumptions about where the users look at during experiments, or have limited practical utilities. Differently, the proposed method does not assume specific interaction scenarios or special content on the screen, while collecting information silently to estimate personal eye parameters. The pro-

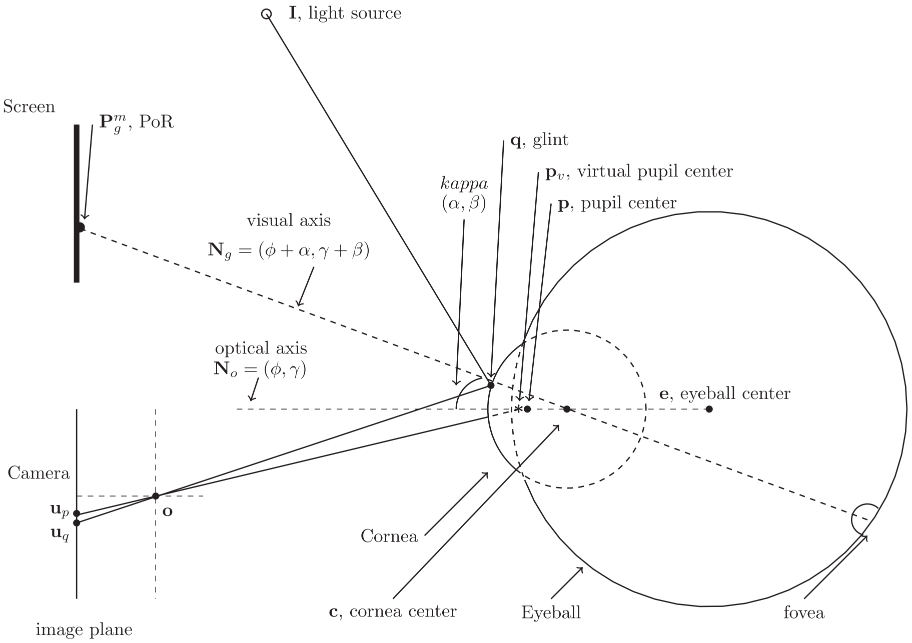


Fig. 1. 3D eye model and gaze estimation.

posed method hence enables non-intrusive and user-friendly eye gaze tracking, while achieving comparable gaze estimation accuracy compared to explicit calibration methods.

3. Model-based gaze estimation

Model-based methods estimate the 3D gaze directions or the 2D point of regard (PoR) by leveraging on the 3D geometric eye model as shown in Fig. 1. Eye is modeled as two spheres, the eyeball sphere, and the cornea sphere. The two spheres rotate together around eyeball center to see different directions. Gaze direction is defined as the visual axis that passes through fovea and cornea center. PoR is defined as the intersection of visual axis and the screen plane. Optical axis is defined as the line connects eyeball center **e**, cornea center **c** and pupil center **p**. For 3D model-based methods, the eye is typically illuminated with IR lights **I** as shown in Fig. 1. A ray comes from the light source **I** and reflects on a point on cornea surface so that the reflected ray passes through the camera nodal point **o**. The reflection point is defined as glint **q**, and the reflection ray intersects with the camera image plane and produces glint image $\mathbf{u}_q$. Similarly, a ray comes from pupil center **p** and refracts at cornea surface, the refracted ray passes through camera nodal point and intersects with the image plane at pupil image $\mathbf{u}_p$. Virtual pupil $\mathbf{p}_v$ is typically assumed to be the intersection point of the extension of refraction ray and the optical axis [9].

In practice, we use two IR light sources. By exploring the reflection property of two IR lights and refraction property of the pupil, we are able to compute the 3D cornea center **c** and 3D virtual pupil center $\mathbf{p}_v$ given the pupil image $\mathbf{u}_p$ and glint image $\mathbf{u}_q$, where $\mathbf{u}_p$ and $\mathbf{u}_q$ can be efficiently detected on the captured eye images. Please refer to [8] for a more detailed theory about 3D geometric eye model. Optical axis is thus estimated by: $\mathbf{N}_o = \mathbf{p}_v - \mathbf{c}/||\mathbf{p}_v - \mathbf{c}||$. Various studies show that the distance between eyeball center **e** and cornea center **c** is a subject-dependent constant value K[1]. Thus we can estimate the position of the eyeball center as:

$$\mathbf{e} = \mathbf{c} - K\mathbf{N}_o. \tag{1}$$

The angle between optical axis and visual axis is also a constant subject-dependent value. We use $\theta = [\alpha, \beta]$ to represent the angles. Notice $\mathbf{N}_o$ is a unit vector in 3-dimensional space, which can be represented by two angles ϕ and γ as follows:

$$\mathbf{N}_o(\phi, \gamma) = \begin{pmatrix} \cos(\phi)\sin(\gamma) \\ \sin(\phi) \\ -\cos(\phi)\sin(\gamma) \end{pmatrix}. \tag{2}$$

Since ϕ and γ can be computed directly given observation **c** and $\mathbf{p}_v$, we call $\{\mathbf{c}, \phi, \gamma\}$ the observations from one frame of data in later sections. Visual axis $\mathbf{N}_g$ is then computed by adding $[\alpha, \beta]$ to optical axis: $\mathbf{N}_g = \mathbf{N}_o(\phi + \alpha, \gamma + \beta)$.

The point $[x, y, z]^T$ on the screen satisfies the surface equation $f(x, y, z) = 0$, which can be estimated by a one-time off-line display-camera calibration. Without loss of generality, we assume the screen is a plane and satisfies: $f(x, y, z) = z = 0$. Intersecting visual axis with the screen plane, we can obtain the 2D PoR:

$$\mathbf{P}_g^m = \mathbf{h}(\mathbf{c}, \phi, \gamma; \theta) = \begin{pmatrix} \mathbf{c}[1] - \mathbf{c}[3]\tan(\gamma + \beta) \\ \mathbf{c}[2] - \mathbf{c}[3]\frac{\tan(\phi+\alpha)}{\cos(\gamma+\beta)} \end{pmatrix}, \tag{3}$$

where $\mathbf{c}[i]$ represents the ith element of cornea center **c.** $\mathbf{h}(\cdot)$ denotes the function to compute PoR given observations $\{\mathbf{c}, \phi, \gamma\}$ and subject-dependent eye parameters θ.

In traditional explicit calibration methods, users are required to look at N pre-defined points ($\mathbf{g}_i$, i = 1,..., N) on the screen plane. Eye parameters θ can be estimated by solving the supervised regression problem:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} (\mathbf{g}_i - \mathbf{h}(\mathbf{c}_i, \phi_i, \gamma_i; \theta))^2. \tag{4}$$

4. Model-based gaze estimation with implicit personal calibration

Explicit personal calibration requires user collaboration. For some applications, this procedure can be cumbersome and degrades the user experience. In this work, we propose to remove the explicit personal calibration by leveraging on four natural constraints during eye gaze tracking.

[1] For this research, we assume K is a constant for different subjects.

4.1. Complementary gaze constraint

The first constraint arises from the fact that 3D model-based method and the feature-based method are complementary to each other. In traditional 2D feature-based gaze estimation methods, it is assumed a linear relationship between the pupil-glint vector (δ_x, δ_y) and the PoR (x, y) on the screen plane. Pupil-glint vector is defined as the difference between pupil image $\mathbf{u}_p$ and glint image $\mathbf{u}_q$: $(\begin{smallmatrix}\delta_x\\ \delta_y\end{smallmatrix}) = \mathbf{u}_p - \mathbf{u}_q$. To map pupil-glint vector to PoR, we can construct the following equations:

$$x = a_x\delta x + b_x\delta y + c_x\,, \tag{5}$$

$$y = a_y\delta x + b_y\delta y + c_y\,, \tag{6}$$

where (a_x, b_x, c_x, a_y, b_y, c_y) are the regression parameters, which can be efficiently learned from training data through a one-time calibration. However, because the pupil-glint vector contains only 2D information on image coordinates, the learned regression parameters are only valid for a fixed head pose. Once we move our head, we need to learn another set of regression parameters. We consider eliminating the influence of head movement by adding 3D information. Notice eyeball center is the rotation center of the eye, thus its position $\mathbf{e}$ can well represent the position of the head. The position of eyeball center $\mathbf{e} = (e_x, e_y, e_z)^T$ can be computed through Eq. (1).

Adding eyeball center position $\mathbf{e}$ as an additional feature to Eqs. (5) and (6) gives rise to the new gaze estimation equations:

$$\mathbf{P}_g^f = \mathbf{A}\delta, \tag{7}$$

where $\mathbf{A} = (\begin{smallmatrix}a_x, b_x, c_x, d_x, f_x, g_x\\ a_y, b_y, c_y, d_y, f_y, g_y\end{smallmatrix})$ represents the regression parameters, and $\delta = (\delta x, \delta y, e_x, e_y, e_z, 1)^T$ represents the augmented feature vector.

The 3D model-based method (Eq. (3)) and feature-based method (Eq. (7)), though based on different principles and mechanisms, should produce close enough PoRs. 3D model-based method provides the domain knowledge about eye anatomy, and mimic the human vision system via a 3D geometric eye model. Therefore eye gaze generated from vision system can be accurately approximated by the estimated gaze from 3D eye model. Feature-based method considers gaze estimation from the learning point of view. It requires enough training data to learn a regression function from eye features to eye gaze. Thus the feature-based method is data-driven, while the 3D model-based method is eye anatomy-driven. These two methods complement each other and inspire us to come up with the following complementary gaze constraint:

$$||\mathbf{P}^m_{g,i} - \mathbf{P}^f_{g,i}|| = ||\mathbf{h}(\mathbf{c}_i, \phi_i, \gamma_i; \theta) - \mathbf{A}\delta_i|| \le \epsilon_1^+ \quad \forall i, \tag{8}$$

where $\{\mathbf{c}_i, \phi_i, \gamma_i, \delta_i\}, i = 1, \ldots, N$ are the collected observations for 3D model-based and feature-based methods, ϵ_1^+ is a small positive constant. $h(\cdot)$ is defined in Eq. (3) and $\mathbf{A}$ is defined in Eq. (7).

4.2. Center prior constraint

Center prior states that when users perform natural viewing tasks, like watching videos/images, most of the fixations are concentrated on the center region of the screen. Center prior constraint has been widely used in saliency map estimation [33]. Chen and Ji [27] also uses center prior as the prior gaze distribution to help implicit personal calibration. Therefore we incorporate the weak center prior constraint in our framework, specifically, we require that:

$$||\frac{1}{N}\sum_{i=1}^{N}\mathbf{P}^m_{g,i} - \begin{pmatrix}\frac{W}{2}\\ \frac{H}{2}\end{pmatrix}|| \le \epsilon_2^+, \tag{9}$$

where ϵ_2^+ is a small positive constant, W and H represent the width and height of the screen (W and H use the same unit as cornea center $\mathbf{c}$, Eg. millimeter). This constraint states that the mean of all gaze points $\mathbf{P}^m_{g,i}$ is close to the screen center.

4.3. Display boundary constraint

For natural interactions with a monitor/screen, it is assumed for a continuous period of time, human's attention is always on the screen region. For example, when a subject watches a video clip or browses the website for some time, all the PoRs or PoRs from a consecutive time segment fall within the screen region. This inspires us to impose hard constraints that all PoRs lie within the display region. Therefore PoR must satisfy: $(\begin{smallmatrix}0\\ 0\end{smallmatrix}) \le \mathbf{P}^m_g = (\begin{smallmatrix}x\\ y\end{smallmatrix}) \le (\begin{smallmatrix}W\\ H\end{smallmatrix})$, where W and H represents the width and height of the display.

4.4. Angular constraint

From human eyeball anatomy, we know kappa $\theta = [\alpha, \beta]$ cannot take arbitrary values. According to [8], the average horizontal angle between optical and visual axis α typical is $+5^\circ$ for the left eye and -5° for the right eye. And the average vertical angle between the visual and optical axis β is typically 1.5°. Therefore we can impose hard constraint to force θ in a reasonable region S_a:

$$\theta \in S_a = \left\{\begin{pmatrix}\alpha\\ \beta\end{pmatrix} \in R^2 : \alpha_l \le \alpha \le \alpha_h, \beta_l \le \beta \le \beta_h\right\}, \tag{10}$$

where α_l and α_h are the lower and upper bounds for α, while β_l and β_h are the lower and upper bounds for β. In the experiments, we set $\alpha_l = \beta_l = -8$ and $\alpha_h = \beta_h = 8$ in degrees.

4.5. Implicit personal calibration with constrained unsupervised regression

Explicit personal calibration can be considered as a supervised regression problem. Given gaze data $\mathbf{X} = \{\mathbf{c}_i, \phi_i, \gamma_i\}$ and gaze labels $\mathbf{y} = \mathbf{g}_i$ (i = 1,...,N), our goal is to solve the supervised regression problem as in Eq. (4). Therefore explicit user collaboration is required to collect the gaze labels $\mathbf{g}_i$. In this work, we only have four constraints without an objective since we have no explicit gaze labels. Any solution satisfying the constraints can be our solution. However, it is time-consuming if we perform greedy-search to find the candidate solutions. Alternatively, we can formulate a constrained unsupervised regression problem to remove the need for explicit gaze labels. For general unsupervised learning problem, model parameters can be estimated by maximizing the marginal likelihood of the parameter θ given data $\mathbf{X}$: $\theta^* = \arg\max_\theta p(\mathbf{X}|\theta)$ subject to constraints. In our deterministic case, we can equivalently minimize the regression error by marginalizing all possible label values $\mathbf{g}_j$:

$$\theta^* = \arg\min_{\theta}\sum_{i=1}^{N}\sum_{\mathbf{g}_j}||\mathbf{g}_j - \mathbf{h}(\mathbf{c}_i, \phi_i, \gamma_i; \theta)||^2$$

subject to the constraints. In practice, marginalizing all possible labels is infeasible, we therefore approximate the regression error by using the gaze position predicted by θ^{t-1} from last iteration:

$$\theta^{\star t} = \arg\min_{\theta^t}\sum_{i=1}^{N}||\mathbf{h}(\mathbf{c}_i, \phi_i, \gamma_i; \theta^{t-1}) - \mathbf{h}(\mathbf{c}_i, \phi_i, \gamma_i; \theta^t)||^2 \tag{11}$$

$$s.t \quad ||\mathbf{h}(\mathbf{c}_i, \phi_i, \gamma_i; \theta) - \mathbf{A}\delta_i|| \le \epsilon_1^+ \ \forall i,$$

$$||\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}(\mathbf{c}_i, \phi_i, \gamma_i; \theta) - \begin{pmatrix} \frac{W}{2} \\ \frac{H}{2} \end{pmatrix}|| \le \epsilon_2^+ \ \forall i,$$

$$\theta \in S_a = \left\{ \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \in R^2 : \alpha_l \le \alpha \le \alpha_h, \beta_l \le \beta \le \beta_h \right\}.$$

Notice that if we ignore the constraints in Eq. (11), the solution collapse to the first tentative solution. But the purpose of implicit calibration is to use the natural soft constraints to gradually refine the personal eye parameters and give better gaze estimation accuracy.

To solve the constrained unsupervised regression problem as in Eq. (11), we plan to use the hard Expectation Maximization (EM) algorithm [34] Before introducing the algorithm, we discuss how to incorporate the four constraints into the hard EM framework.

First, complementary gaze constraint and center prior constraint are interpreted as regularization terms by introducing two Lagrangian multipliers. By doing so, we do not need to specify the values of ϵ_1^+ and ϵ_2^+, which may result in an empty solution space if not specified correctly. Second, display boundary constraint and angular constraint impose hard constraints on θ. From display boundary constraint, each gaze point $(x_i, \ y_i)^T$ will determine a feasible region S_i for $[\alpha, \ \beta]$. Combining angular constraint in Eq. (10) and all display boundary constraints S_i, we know θ must satisfy: $\theta \in S_f = \bigcap_{1 \le i \le N} S_i \cap S_a$, where S_f denotes the final feasible region of θ.

Given the objective function as well as the constraints, we solve the unsupervised regression problem through the hard EM framework. The hard EM consists of an E-step and an M-step. In the E-step, given estimated θ^{t-1} from the last iteration, we compute all the PoRs given the data. In the M-step, we update θ^t by solving a constrained unsupervised regression problem. Algorithm 1 summarizes the procedure.

Algorithm 1 Hard EM algorithm.

```
1: Given gaze data: $\{\mathbf{c}_i, \phi_i, \gamma_i, \delta_i\}, i = 1, \ldots, N$.
2: Initialize $\theta$ to the human average value $\theta_0$. Compute initial PoRs with $\theta_0$ via Eq. (3). These initial PoRs are used to compute initial regression parameter $\mathbf{A}_0$ via Eq. (7).
3: Compute gaze points $\mathbf{g}_i^t$ given $\theta^{t-1}$ from last iteration:$\mathbf{g}_i^{\star t} = \mathbf{h}(\mathbf{c}_i, \phi_i, \gamma_i; \theta^{t-1}) \quad \forall i$
4: Update $\theta^t$ for current iteration by solving the following optimization problem:
   $\theta^{\star t} = \arg\min_{\theta^t} \frac{1}{N}\sum_{i=1}^{N}||\mathbf{g}_i^{\star t} - \mathbf{P}_{g,i}^m||^2 + \frac{\lambda_1}{N}\sum_{i=1}^{N}||\mathbf{P}_{g,i}^m - \mathbf{P}_{g,i}^f||^2 + \lambda_2||\frac{1}{N}\sum_{i=1}^{N}\mathbf{P}_{g,i}^m - \begin{pmatrix} \frac{W}{2} \\ \frac{H}{2} \end{pmatrix}||^2$   (12)
   $s.t \quad \theta^t \in S_f$ ,
   where $\mathbf{P}_{g,i}^m = \mathbf{h}(\mathbf{c}_i, \phi_i, \gamma_i; \theta^t), \mathbf{P}_{g,i}^f = \mathbf{A}\delta_i$   (13)
5: Repeat step 3 and 4 until convergence.
```

The first term in Eq. (12) is the regression error, while the second and third terms impose the two soft constraints as regularizations, with multipliers λ_1 and λ_2 respectively. They represent a trade-off between the regression errors and the violation of the constraints. In the experiments, λ_1 and λ_2 are set to 0.4 and 0.02 for all subjects. They are chosen based on some experiments to ensure different terms in Eq. (12) are at roughly similar scale. The average performance for all subjects is not that sensitive to their values, but individual subject's performance might be affected. Actually, if subjects perform a video watching task, it might give better results if we set λ_2 larger (center prior constraint is dominant in video watching task). In practice, depending on tasks(subjects' gaze distribution), the optimal settings of λ_1 and λ_2 might be different, however, providing the two values in the paper can usually give reasonable accuracy.

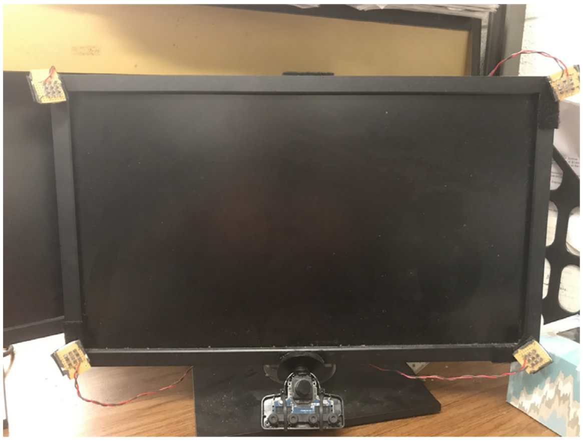


Fig. 2. Hardware setup with one web camera and 4 IR-lights on the 4 corners of the monitor.

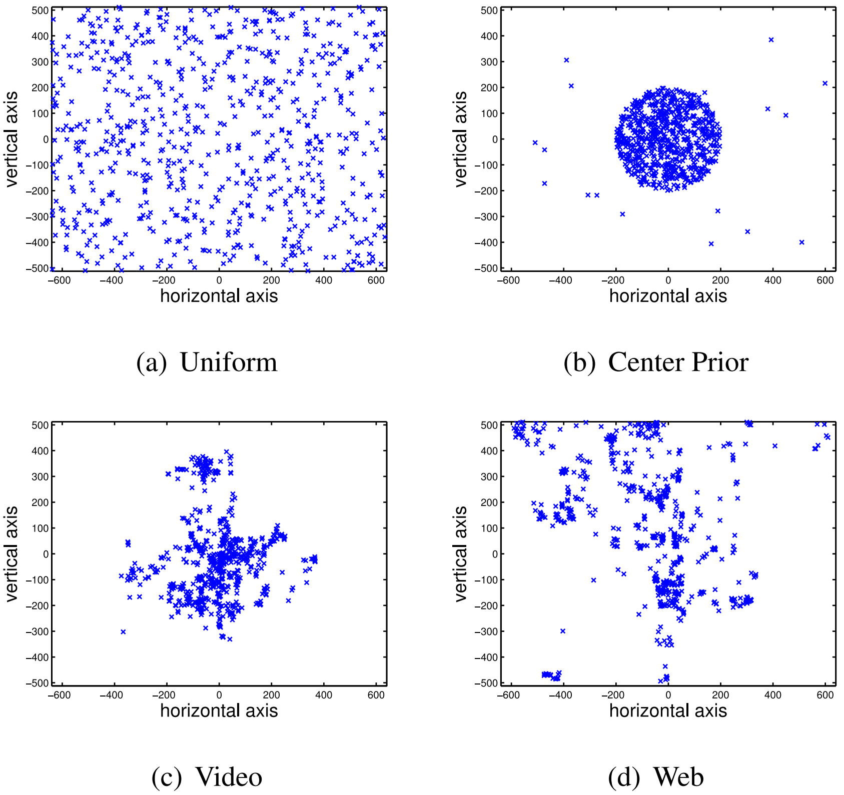
Eq. (13) represents the feasible region imposed by the two hard constraints. The nonlinear optimization problem can be efficiently solved through interior-point algorithm. To initialize the hard-EM algorithm, we first set eye parameter θ_0 to the human average values $\theta_0 = (\alpha_0, \ \beta_0)$, from which we can compute the initial PoRs $\mathbf{P}_g^{m0} = g(\mathbf{c}, \phi, \gamma; \theta_0)$ for model-based method (Eq. (3)). PoRs estimated from feature-based method (Eq. (7)) are assumed to be the same as model-based method: $\mathbf{P}_g^{f0} = \mathbf{P}_g^{m0}$, from which we can compute the initial regression parameters $\mathbf{A}_0$ using Eq. (7).

Formulated as a constrained unsupervised regression problem, the proposed implicit calibration method only uses four weak constraints on θ without any label information. Using individual constraint alone will not work. For the complementary gaze constraint alone, it is ambiguous and non-unique in the sense that different pairs of parameters θ and $\mathbf{A}$ can satisfy the constraint. This is exactly why we need other three constraints to constrain the solution to make it both physically and anatomically meaningful. The use of center prior constraint can benefit from center-biased gaze patterns, but we want to point out that our algorithm is not as sensitive to incorrect center gaze patterns as in [27]. Because center prior itself is a very weak constraint and is jointly used with the other three constraints. In fact, the proposed algorithm can be applied to non-center distributed data like the web browsing gaze data. The display boundary constraint and the angular constraint contribute to the reduction of the feasible region from $\mathbb{R}^2$ to S_f, which improves the accuracy as well as the efficiency of the algorithm. In the next section, experiments will demonstrate how these constraints work in both qualitative and quantitative manners.

5. Experiments and analysis

5.1. Experimental settings

Hardware setup: The hardware setup is illustrated in Fig. 2. We use a Sony Webcam WCX550 web camera in the experiments, its IR filter is removed and we also add a natural light filter. 4 IR-light arrays are placed on the 4 corners of the 21.5-inch monitor. We implement the one camera two IR lights system [8,9] to efficiently estimate 3D pupil $\mathbf{p}_v$ and 3D cornea $\mathbf{c}$. The purpose of using 4 IR-



(a) Uniform

(b) Center Prior

(c) Video

(d) Web

Fig. 3. Synthetic uniform and center prior distributions and real PoR distributions from one subject with video watching and web browsing scenarios. PoRs are recorded by a commercial eye tracker [35] with approximately 0.5° accuracy.

lights is to enable larger head motions and improve the robustness, the method can work with only 2 IR-lights.

Data: we evaluate the proposed method with both synthetic data and real data. For synthetic data, we first generate target gaze distributions as shown in Fig. 3, and observations can be synthesized through reverse engineering. For real data, we invite 8 subjects (5 male and 3 female) whose age ranges from 22 to 30 years old to perform two natural tasks: video watching and website browsing. Subjects sit in front of the screen at a distance of $\approx$500 mm and can move horizontally and vertically as long as their eyes are visible to the camera. The real data are implicitly collected while subjects perform the tasks (approximately 120–150 s). Fig. 3(c) and (d) show an example gaze distribution from one subject.

Evaluation: To evaluate the estimated personal eye parameters, we ask subjects to look at 25 uniformly distributed points on the screen, these points served as the groundtruth positions and are used to compute the gaze estimation error. For comparison, we also implement the explicit 9-points calibration method as baseline, where Eq. (4) can be used to solve the eye parameters.

To evaluate the gaze estimation performance, we use the angular error in degree as the evaluation metric. Specifically, since the system is fully calibrated, we know the groundtruth gaze position $\mathbf{P}_g = [x_g, y_g, 0]$ and the predicted gaze position $\mathbf{P}_g^m = [x_p, y_p, 0]$ on the screen plane (see Section 3 for details). We also know the estimated eyeball center position $\mathbf{e} = [e_x, e_y, e_z]$. Given these, we can estimate the groundtruth gaze direction $\mathbf{v}_g$ and predicted gaze direction $\mathbf{v}_p$:

$$\mathbf{v}_g = \frac{\mathbf{P}_g - \mathbf{e}}{||\mathbf{P}_g - \mathbf{e}||}, \mathbf{v}_p = \frac{\mathbf{P}_g^m - \mathbf{e}}{||\mathbf{P}_g^m - \mathbf{e}||}$$

Gaze estimation error is calculated as $\mathrm{err} = \arccos(\mathbf{v}_g^T \mathbf{v}_p)$ in degree.

5.2. Evaluation on synthetic data

The proposed method depends on different viewing patterns from the subjects. Therefore we first study several common gaze patterns as shown in Fig. 3. They include the uniform-distributed gaze pattern, a center-biased gaze pattern, one real gaze pattern recorded when the subject watches a video, and one real gaze pattern from the subject browsing the website. These gaze patterns can be considered as N groundtruth gaze positions $\{\mathbf{P}_g^i\}_{i=1}^N$ on the screen. Next, we can simulate different subjects and their spatial positions. Different subjects means different kappa angles $[\alpha, \beta]$, they are drawn from a prior uniform distribution with lower bound $[\alpha_l, \beta_l] = [-8, -8]$ and higher bound $[\alpha_h, \beta_h] = [8, 8]$. Note that the true kappa distribution is not a uniform distribution but rather a Gaussian distribution ([8]). The proposed method is more general and does not assume known distributions of kappa angles. Subjects' spatial position refers to the position of the 3D eyeball center $\mathbf{e}$, which is also manually selected, so that subject appears in front of the screen and camera, with a distance around 500 millimeters. Given this information, we are able to perform reverse-engineering and estimate the 3D pupil center and 3D glints position (Fig. 1). We mainly rely on the law of reflection and refraction of light on the corneal surface (See [8] for details). Given the 3D pupil and glints positions, we can project them down to the image plane (given camera parameters) to get the measurements. We also add noise to the measured 2D pupil and glint positions. The noise level means the gaze estimation error using explicit-calibration with the noisy measurements.

With the proposed implicit calibration method, the results are shown in Fig. 4. As we can see, the center prior pattern gives the poorest result, as we cannot leverage on other constraints like the screen boundary constraint. As for the two natural scenarios web browsing and video watching, they give better results (1.5°) than

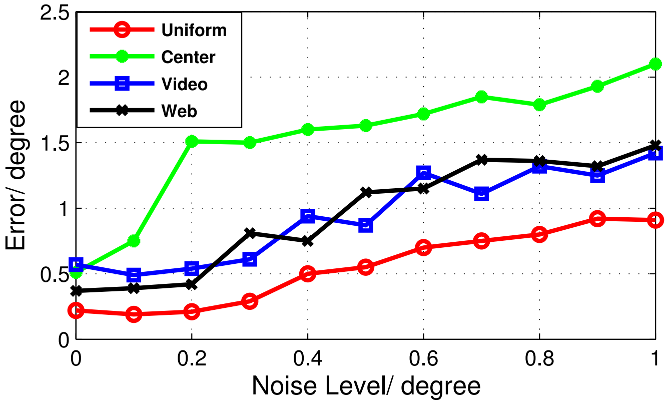


Fig. 4. Gaze estimation accuracy with different noise levels and gaze distributions.

Table 1
Estimated eye parameters [α, β] from 9-points calibration and the proposed method with Video and Web scenarios.

Subjects	9 points		Video		Web	
1	1.0	0.3	0.5	0.6	0.3	0.7
2	−0.7	3.8	−1.0	2.9	−1.3	3.5
3	1.3	2.8	1.6	2.3	1.7	2.0
4	1.6	2.0	1.5	1.3	2.1	2.1
5	2.7	−0.8	3.2	−0.5	2.5	−0.7
6	−3.6	−0.8	−2.8	−1.0	−3.2	−1.1
7	0.8	−2.3	1.2	−1.0	0.1	−1.3
8	1.2	−3.0	0.4	−2.1	1.8	−2.8
9	1.9	0.3	2.2	0.5	1.7	0.4
10	−1.3	1.9	−0.5	1.8	−1.7	2.1
11	1.8	−1.5	1.7	−1.2	2.1	−1.7
12	−3.9	−3.3	−3.2	−2.7	−3.5	−3.7
13	1.7	3.9	1.6	4.0	1.5	4.1
14	1.6	2.3	2.1	2.1	1.8	2.5

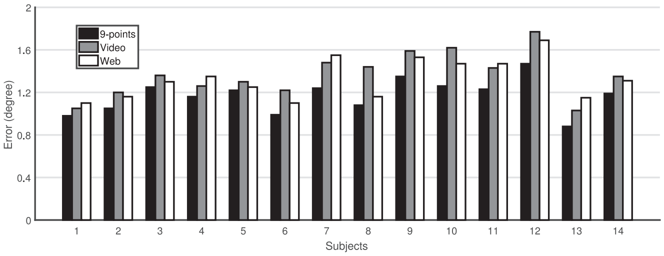


Fig. 5. Gaze estimation error with different calibration settings.

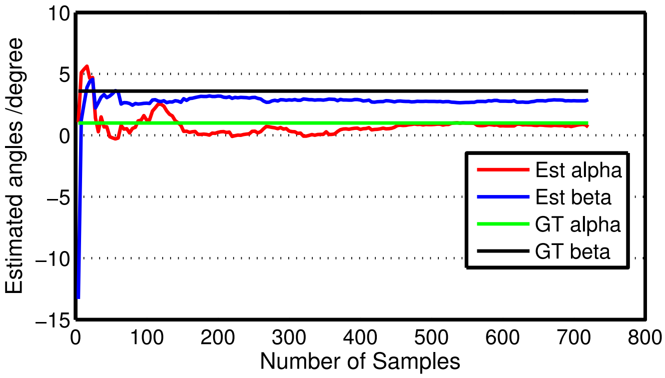


Fig. 6. Calibrated parameters with a different number of calibration samples.

center prior, but not as good as the explicit-calibration approach (1.0°). The best result is from the uniform distribution which is comparable to explicit-calibration, the underlying reason is that it narrows down the feasible solution space with the screen boundary constraints, but such pattern is not realistic in practice. Overall, the results with web browsing and video watching scenarios convince us the proposed method can implicitly calibrate eye parameters, and give reasonably good results.

5.3. Evaluation on individual subjects

Personal eye parameters: We first take a look at the estimated eye parameters for different subjects. Table 1 lists the eye parameters estimated from explicit 9-points method, and the proposed implicit method with Video and Web scenarios. We can find that some eye parameters are close to each other, while others differ a lot. For example, β from subject 7 differ a lot to the explicit method. α and β from subject 8 with Video task also different a lot to the explicit method. Since we do not know the groundtruth eye parameters for each subject, the parameter comparison is only for reference.

Gaze estimation evaluation: The gaze estimation error for different subjects with Video and Web scenarios are shown in Fig. 5. Visually speaking, 9-points method is better than Video and Web scenarios, but the accuracy is also comparable. These results prove that the proposed method can not only work with strong center-biased PoR distribution (Fig. 3 (c)) but also work with distributions where PoRs are close to the boundary of the display (Fig. 3 (d)). In [27], center prior is directly utilized to align the subjects' gaze data to learn the eye parameters; therefore strong center-biased PoR distributions are required. However, in the proposed method, center prior is only interpreted as a soft constraint and is jointly utilized with other three constraints. If PoR distributions are indeed center-biased, the center prior constraint will help find better eye parameters. If PoR distributions are similar to Fig. 3 (d) where PoRs are close to the boundary, such distribution forces the feasible region S_f to be smaller and compact, therefore we eliminate more wrong parameters and are able to find better eye parameters.

Quantitatively speaking, the average error for 9-points calibration, Video scenario and Web scenario are 1.18° ± 0.43°, 1.36° ± 0.48° and 1.32° ± 0.47° respectively. The variance is resulted from poor feature detections, head motions, and screen reflections, etc. Overall, we can see that the proposed implicit calibration method can achieve comparable results as the explicit method. The results also illustrate that the proposed framework can work efficiently with natural interactions like browsing websites and watching videos.

5.4. Evaluation on number of calibration samples

The performance of the algorithm also depends on the number of data samples. Fig. 6 shows the calibrated parameters from one subject and Fig. 7 shows the average gaze estimation error for all subjects with a different number of calibration samples. The overall error is decreasing when the number of samples increases, as the calibrated parameters approach the groundtruth parameters (9-points calibration results). Because of the randomness of gaze positions on the screen, the curve is not monotonically decreasing but with perturbations. However, after enough number of data samples, the estimated eye parameters will finally converge, which enable accurate eye gaze tracking. Depending on interaction scenarios where gaze patterns are different, the number of samples might be different. But with the proposed algorithm, we can obtain very good results in around 1000 samples as opposed to 6000 samples in [27].

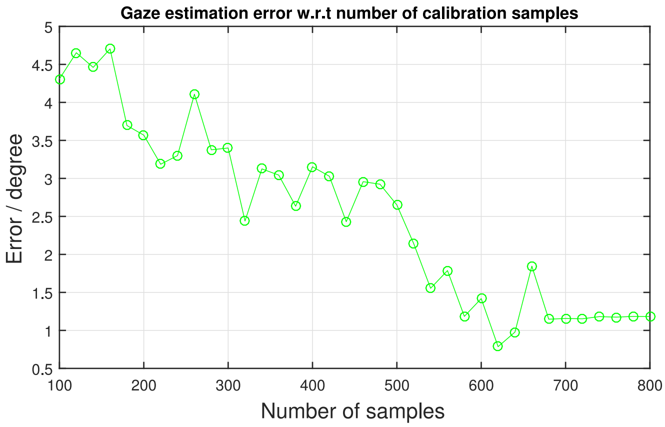


Fig. 7. Gaze estimation error with a different number of calibration samples.

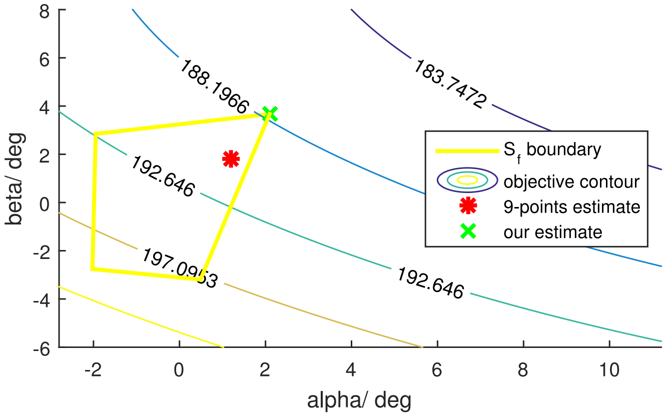


Fig. 8. Visualization of solution space.

5.5. A deeper look into four natural constraints

The goal of this paper is to solve a set of parameters given four constraints. Without an objective function, any sets of parameters within the feasible space can be our solution. One brutal-force approach might be to evaluate all possible parameters and construct the feasible space, the final estimation can be the mean of parameters in the feasible space. However, such brutal-force approaches are time-consuming and require carefully chosen hyper-parameters. For example, If ϵ_1^+ in Eq. (8) and ϵ_2^+ in Eq. (9) are set too small, we might end up with an empty feasible space, or the constraints might not work if they are set too large. Nevertheless, with the proposed hard-EM framework and iterative algorithm, we can still take advantage of the constraints and obtain good eye parameters for eye gaze tracking.

5.5.1. Visualization of solution space

We first take a look at the solving process from one iteration in Algorithm 1. As shown in Fig. 8, the objective function contour, feasible space, and solutions are shown on a 2D plot. The objective function contour is obtained by computing the objective function value at the 2D grid of the parameters, which encodes how the complementary constraint and center prior constraint contribute to our estimation. The yellow polyhedron denotes the feasible space forced by the display boundary constraint. The angular constraint does not contribute in this case since its feasible space is a superset of the yellow polyhedron. We can see that the objective function is monotonically decreasing towards the top right direction. Therefore our estimation (green cross) is the top right corner of the yellow polyhedron. For reference, the parameters estimated from 9-points method are marked as a red star. Compared to $\mathbb{R}^2$, the feasible region defined by the yellow polyhedron is much smaller. After this iteration, the algorithm continues based on the current estimate. Notice the solution space in Fig. 8 is only one particular trial from one subject, the solution space will be totally different for other trials. We cannot predict where is the optimal solution as the solution space is a function of the number of samples, as well as the distributions of the samples. However, following the algorithm in Algorithm 1, we can get a good estimation of the personal eye parameters that give good gaze estimation performance.

Table 2
Distances (pixel) of predicted gaze positions in Figs. 9 and 10. $\mathbf{p}^f$, $\mathbf{p}^m$ and $\mathbf{g}$ represent gaze positions predicted from the feature-based method, model-based method and groundtruth respectively.

		$\|\|p^f - p^m\|\|$	$\|\|p^f - g\|\|$	$\|\|p^m - g\|\|$
Fig. 9	Before calibration	64.9	196.9	208.7
	After calibration	8.6	64.3	64.8
Fig. 10	Before calibration	9.5	216.9	216.8
	After calibration	9.2	65.1	66.9

5.5.2. Complementary gaze constraint

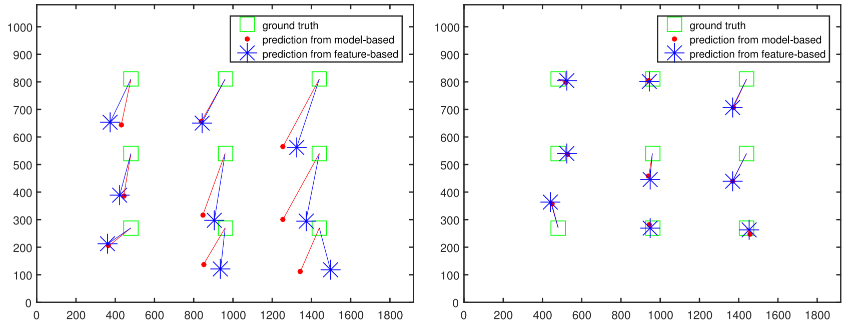
In this section, we take a detailed look at how complementary gaze constraint affects the predicted gaze positions. For simplicity, we denote $\mathbf{p}^f$, $\mathbf{p}^m$ and $\mathbf{g}$ as gaze positions predicted from feature-based method, model-based method and groundtruth respectively. As shown in Fig. 9 and Table 2, the error for feature-based ($||\mathbf{p}^f - \mathbf{g}||$) and model-based ($||\mathbf{p}^m - \mathbf{g}||$) methods are large. And the distance between feature-based and model-based ($||\mathbf{p}^f - \mathbf{p}^m||$) is also large. By solving the proposed constrained unsupervised regression problem, we are able to reduce the errors as well as the distance between the two methods. This demonstrates the effectiveness of the complementary gaze constraint, by forcing gaze positions from the two methods to be as close as possible, we are able to find more accurate personal eye parameters.

Fig. 9 actually shows a case where complementary gaze constraint is dominant, we also take a look at a different scenario in Fig. 10. Though the distance $||\mathbf{p}^f - \mathbf{p}^m||$ is already pretty small before calibration, the proposed method can take advantage of other natural constraints to find accurate parameters that give better estimation accuracy.

5.5.3. Importance of individual constraint

To evaluate the importance of the four constraints, we drop one of the 4 constraints one at a time and compare the accuracy with using all constraints. Similarly, we evaluated on both Video watching and Web browsing tasks.

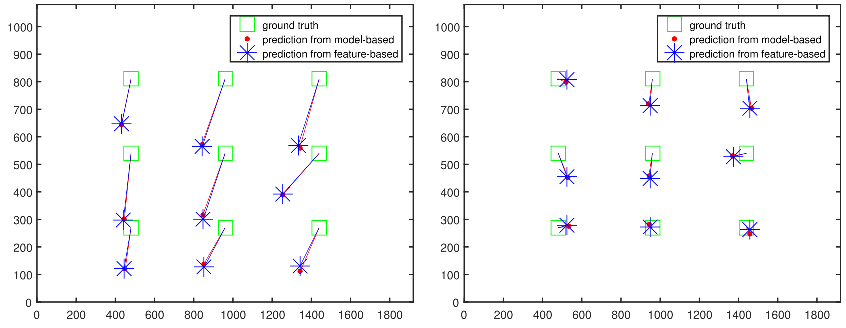
In Fig. 11 a), when dropping complementary gaze constraint, the error significantly increases for both Web browsing and Video watching tasks. This is because the complementary gaze constraint is more generic and can apply to different scenarios, and thus dropping this constraint degrades the performance. In Fig. 11 b), it is clear that when gaze distribution is center-biased as in Video watching task, the center-prior constraint is more effective and dropping the constraint causes a large error. However, the error also increases even in Web browsing tasks, demonstrating the importance of center prior constraint. In Fig. 11 c), we find that display-boundary constraint is of great importance as it can significantly reduce the feasible solution space (yellow polyhedron in Fig. 8). It also applies to both tasks and causes a large error when dropping the constraint. Finally in Fig. 11 d), it seems angular constraint if of no use as dropping the constraint does not make any difference. The reason is that with sufficiently large samples ($N = 800$), the solution space determined by display-boundary constraint is already a subset of the feasible space determined by



(a) before calibration

(b) after calibration

Fig. 9. Visualization of predicted gaze positions before and after calibration. Complementary gaze constraint is dominant.



(a) before calibration

(b) after calibration

Fig. 10. Visualization of predicted gaze positions before and after calibration. Complementary gaze constraint is not dominant.

Table 3
Comparison with state-of-the-art methods.

Method	Error/degree
Proposed	**1.3**
Wang et al. [28]	1.5
Chen and Ji [26]	1.7
Chen and Ji [27]	2.5
Guestrin and Eizenman [7]	1.3
Sugano et al. [30]	4.8
Sugano et al. [31]	3.5
Alnajar et al. [29]	4.3

angular constraint. It can also be seen in Fig. 8, where the yellow polyhedron is a subset of $S_a = \{-8 \le \alpha \le= 8 \ and - 8 \le \beta \le= 8\}$.

However, when there are not enough samples, angular constraints may also contribute to finding the personal eye parameters. As shown in Fig. 12, when we use only 200 samples, using all constraints cannot give a good performance, but we observe that dropping angular constraint indeed increases the error. This demonstrates that the angular constraint can benefit especially when we do not have enough calibration samples.

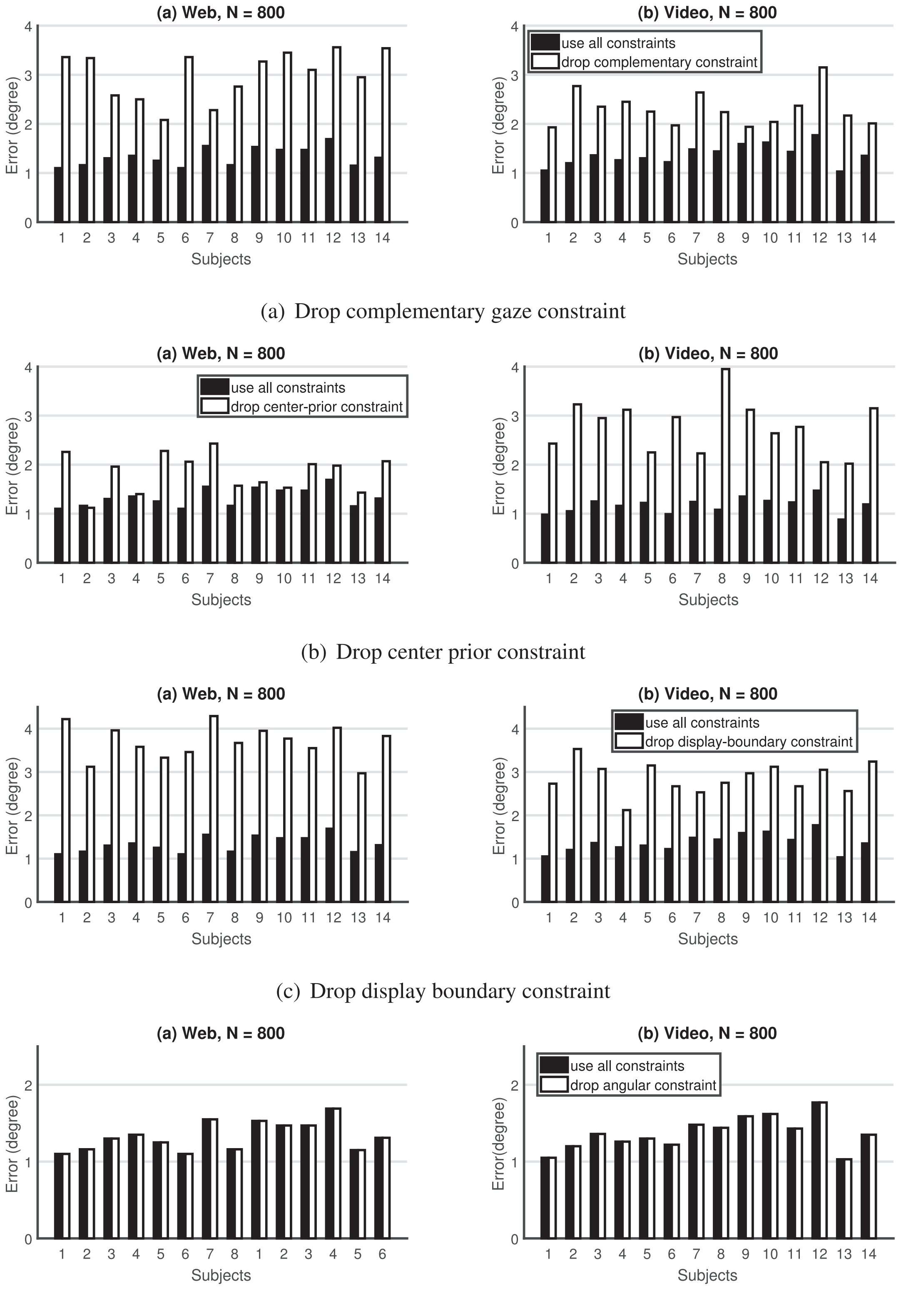
5.6. Comparison with state-of-the-art

Finally, we compare the proposed method with other state-of-the-art methods in reducing/eliminating explicit calibration, as shown in Table 3. In particular, we implement three model-based approaches [26–28] which have similar experimental settings with ours. The algorithms in [26,28] utilize saliency maps, while [27] use a simple Gaussian distribution. Since the Web browsing scenario might not give good saliency map and gaze patterns are not center-biased, we therefore only evaluate on Video watching scenarios in order to be comparable. The numbers for one model-based approach [7] and three appearance-based approaches in [29–31] are extracted from their original papers for reference. Though we did not implement and compare with these methods directly, they are limited in practical usages. For example, the limitation of complex system setup in [7] and the requirement of saliency content [29,31].

Compared to [26,28], our method is fully implicit with minimum user cooperation, while their method requires users to look at salient images/videos and requires saliency computation. The method in [27] only requires a Gaussian distribution, however, their method is limited to actual center-biased gaze patterns, and cannot apply to an noncenter-biased interactions. Besides, even for center-biased gaze patterns like Video watching, their method requires much longer time to converge (6000 frames). Compared to state-of-the-art model-based approaches, the proposed method provides a generic calibration interface for many interaction scenarios, requires minimum user cooperation, and enables efficient and accurate eye gaze tracking.

6. Conclusion

In this paper, we propose a novel implicit calibration framework for 3D model-based method with the help of four natural constraints during eye gaze tracking. By exploring the complementary nature of two gaze estimation methods (complementary gaze constraint), the center prior principle (center prior constraint), the viewing habit for screen-based scenarios (display boundary constraint), and human eye anatomy knowledge (angular constraint), we propose to formulate the implicit calibration problem as constrained unsupervised regression problem which integrates all these natural constraints. The constrained unsupervised regression problem can be solved effectively using the iterative hard EM algorithm. The proposed framework does not require any explicit user participation. Experiments for different subjects with video watching and web browsing scenarios prove the effectiveness of



(a) Drop complementary gaze constraint

(b) Drop center prior constraint

(c) Drop display boundary constraint

(d) Drop angular constraint

Fig. 11. Gaze performance without one of the 4 constraints, using $N = 800$ samples.

the proposed implicit calibration framework. Compared with traditional explicit calibration method, the proposed method achieves comparable results but is less intrusive and more friendly to users. Compared with other methods in implicit personal calibration, the proposed method is more efficient and achieves better results in less constrained experimental settings.

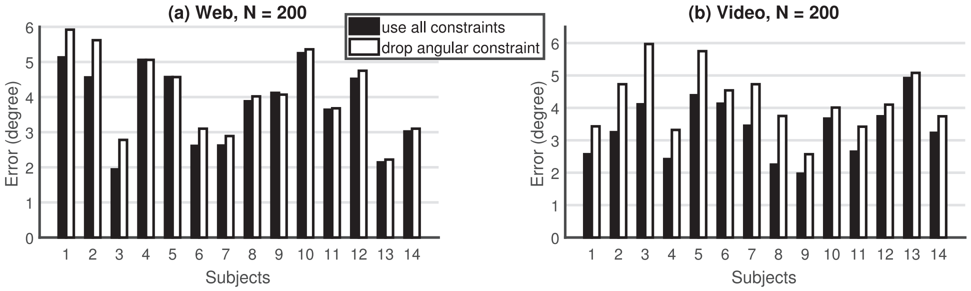


Fig. 12. Gaze performance without angular constraint, using $N = 200$ samples.

Acknowledgment

This work was supported in part by a National Science Foundation grant (IIS 1539012).

References

[1] Y. Sawahata, R. Khosla, K. Komine, N. Hiruma, T. Itou, S. Watanabe, Y. Suzuki, Y. Hara, N. Issiki, Determining comprehension and quality of tv programs using eye-gaze tracking, Pattern Recognit. 41 (5) (2008) 1610–1626.
[2] J.D. DeJong, G.M. Jones, Akinesia, hypokinesia, and bradykinesia in the oculomotor system of patients with Parkinson's disease, Exp. Neurol. 32 (1) (1971) 58–68.
[3] C.F. Norbury, J. Brock, L. Cragg, S. Einav, H. Griffiths, K. Nation, Eye-movement patterns are associated with communicative competence in autistic spectrum disorders, J. Child Psychol. Psych. 50 (7) (2009) 834–842.
[4] M. Mason, B. Hood, C.N. Macrae, Look into my eyes: Gaze direction and person memory, Memory 12 (5) (2004) 637–643.
[5] K.R. Park, J.J. Lee, J. Kim, Gaze position detection by computing the three dimensional facial positions and motions, Pattern Recognit. 35 (11) (2002) 2559–2569.
[6] D. Beymer, M. Flickner, Eye gaze tracking using an active stereo head, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2, IEEE, 2003, pp. II–451.
[7] E.D. Guestrin, M. Eizenman, Remote point-of-gaze estimation requiring a single-point calibration for applications with infants, in: Proceedings of the Symposium on Eye Tracking Research & Applications, ACM, 2008, pp. 267–274.
[8] E.D. Guestrin, M. Eizenman, General theory of remote gaze estimation using the pupil center and corneal reflections, IEEE Trans. Biomed. Eng. 53 (6) (2006) 1124–1133.
[9] J. Chen, Y. Tong, W. Gray, Q. Ji, A robust 3d eye gaze tracking system using noise reduction, in: Proceedings of the Symposium on Eye Tracking Research & Applications, ACM, 2008, pp. 189–196.
[10] K. Wang, Q. Ji, Real time eye gaze tracking with kinect, in: Proceedings of the 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 2752–2757.
[11] K. Wang, Q. Ji, Real time eye gaze tracking with 3d deformable eye-face model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1003–1011.
[12] C. Gou, Y. Wu, K. Wang, K. Wang, F.-Y. Wang, Q. Ji, A joint cascaded framework for simultaneous eye detection and eye state estimation, Pattern Recognit. 67 (2017) 23–31.
[13] F. Song, X. Tan, S. Chen, Z.-H. Zhou, A literature survey on robust and efficient eye localization in real-life scenarios, Pattern Recognit. 46 (12) (2013) 3157–3173.
[14] N. Markuš, M. Frljak, I.S. Pandžić, J. Ahlberg, R. Forchheimer, Eye pupil localization with an ensemble of randomized trees, Pattern Recognit. 47 (2) (2014) 578–587.
[15] J. Song, Z. Chi, J. Liu, A robust eye detection method using combined binary edge and intensity information, Pattern Recognit. 39 (6) (2006) 1110–1125.
[16] C.H. Morimoto, M.R. Mimica, Eye gaze tracking techniques for interactive applications, Comput. Vis. Image Underst. 98 (1) (2005) 4–24.
[17] Z. Zhu, Q. Ji, Eye and gaze tracking for interactive graphic display, Mach. Vis. Appl. 15 (3) (2004) 139–148.
[18] Z. Zhu, Q. Ji, K.P. Bennett, Nonlinear eye gaze mapping function estimation via support vector regression, in: Proceedings of the 18th International Conference on Pattern Recognition (ICPR), 1, IEEE, 2006, pp. 1132–1135.
[19] H. Cheng, Y. Liu, W. Fu, Y. Ji, L. Yang, Y. Zhao, J. Yang, Gazing point dependent eye gaze estimation, Pattern Recognit. 71 (2017) 36–44.
[20] F. Lu, X. Chen, Y. Sato, Appearance-based gaze estimation via uncalibrated gaze pattern recovery, IEEE Trans. Image Process. 26 (4) (2017) 1543–1553.
[21] K.-H. Tan, D.J. Kriegman, N. Ahuja, Appearance-based eye gaze estimation, in: Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision (WACV), IEEE, 2002, pp. 191–195.
[22] W. Maio, J. Chen, Q. Ji, Constraint-based gaze estimation without active calibration, in: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG), IEEE, 2011, pp. 627–631.
[23] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, Appearance-based gaze estimation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4511–4520.
[24] D.W. Hansen, Q. Ji, In the eye of the beholder: a survey of models for eyes and gaze, IEEE Trans. Pattern Anal. Mach. Intell. 32 (3) (2010) 478–500.
[25] D. Model, M. Eizenman, An automatic personal calibration procedure for advanced gaze estimation systems, IEEE Trans. Biomed. Eng. 57 (5) (2010) 1031–1039.
[26] J. Chen, Q. Ji, Probabilistic gaze estimation without active personal calibration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 609–616.
[27] J. Chen, Q. Ji, A probabilistic approach to online eye gaze tracking without explicit personal calibration, IEEE Trans. Image Process. 24 (3) (2015) 1076–1086.
[28] K. Wang, S. Wang, Q. Ji, Deep eye fixation map learning for calibration-free eye gaze tracking, in: Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, ACM, 2016, pp. 47–55.
[29] F. Alnajar, T. Gevers, R. Valenti, S. Ghebreab, Calibration-free gaze estimation using human gaze patterns, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 137–144.
[30] Y. Sugano, Y. Matsushita, Y. Sato, H. Koike, An incremental learning method for unconstrained gaze estimation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2008, pp. 656–667.
[31] Y. Sugano, Y. Matsushita, Y. Sato, Appearance-based gaze estimation using visual saliency, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2) (2013) 329–341.
[32] K. Pfeuffer, M. Vidal, J. Turner, A. Bulling, H. Gellersen, Pursuit calibration: Making gaze calibration less tedious and more flexible, in: Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, ACM, 2013, pp. 261–270.
[33] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: Proceedings of the IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 2106–2113.
[34] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Stat. Soc. Ser. B Methodol. 39 (1) (1977) 1–38.
[35] Tobii Technology, Tobii eye tracker 4c, 2017 (http://www.tobii.com/en/eye-experience/).

Kang Wang received his B.S. degree from Department of Electronic Engineering and Information Science, University of Science and Technology of China in 2013. He is currently pursuing the Ph.D. degree at Rensselaer Polytechnic Institute, Troy, New York. His main research interests include computer vision and machine learning. Specifically, he is interested in human attention modeling and its applications for real-time eye gaze tracking.

Qiang Ji received his Ph.D. degree in electrical engineering from the University of Washington. He is currently a professor with the Department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute (RPI). He recently served as director of the Intelligent Systems Laboratory (ISL) at RPI. Prof. Ji's research interests are in computer vision, probabilistic graphical models, information fusion, and their applications in various fields. Prof. Ji is an editor on several related IEEE and international journals and he has served as a general chair, program chair, technical area chair, and program committee member in numerous international conferences/workshops. Prof. Ji is a fellow of IEEE and IAPR.