# A Probabilistic Approach to Online Eye Gaze Tracking Without Explicit Personal Calibration

Jixu Chen, *Member, IEEE*, and Qiang Ji, *Fellow, IEEE*

***Abstract*— Existing eye gaze tracking systems typically require an explicit personal calibration process in order to estimate certain person-specific eye parameters. For natural human computer interaction, such a personal calibration is often inconvenient and unnatural. In this paper, we propose a new probabilistic eye gaze tracking system without explicit personal calibration. Unlike the conventional eye gaze tracking methods, which estimate the eye parameter deterministically using known gaze points, our approach estimates the probability distributions of the eye parameter and eye gaze. Using an incremental learning framework, the subject does not need personal calibration before using the system. His/her eye parameter estimation and gaze estimation can be improved gradually when he/she is naturally interacting with the system. The experimental result shows that the proposed system can achieve <3° accuracy for different people without explicit personal calibration.**

***Index Terms*— Gaze estimation, gaze calibration, dynamic Bayesian network.**

## I. Introduction

Gaze tracking is the procedure of determining the point-of-gaze on the monitor, or the visual axis of the eye in 3D space. Gaze tracking systems are primarily used in the Human Computer Interaction (HCI) and in the analysis of visual scanning patterns. In HCI, the eye gaze can serve as an advanced computer input [1] to replace traditional input devices such as a mouse pointer [2]. Also, the graphic display on the screen can be controlled by the eye gaze interactively [3]. Since visual scanning patterns are closely related to the attentional focus, cognitive scientists use the gaze tracking system to study human's cognitive processes [4].

Various video-based gaze estimation techniques [3], [5]–[14] have been proposed (a survey of gaze estimation may be found in [15]), and the gaze estimation systems have now evolved to the point where the user is allowed free head movements while maintaining high accuracy (one degree or better). However, most of current gaze estimation systems require a personal calibration procedure for each subject in order to estimate his/her specific eye parameters. This calibration process could significantly limit the practical utility of gaze estimation. To overcome this limitation, we propose a novel gaze estimation framework without any explicit personal calibration. The main contributions of this work include:

- the introduction of a probabilistic approach (versus the existing deterministic approach) for 3D eye gaze estimation without explicit personal calibration,
- development of an incremental learning approach to incrementally refine the eye parameters when the subject is naturally viewing the screen without prompting, and
- development of a gaze estimation algorithm using Gaussian prior probability when the subject is naturally watching a video.

Our system adapts automatically online to each subject to improve eye gaze tracking accuracy without user collaboration, making natural, non-intrusive and non-collaborative eye gaze tracking closer to reality. Since the video-based eye tracking is being widely used, including many commercial eye trackers. By eliminating the explicit personal calibration for these eye trackers, the proposed research hence has significant practical impact.

## II. Related Work

Video-based eye gaze estimation can be divided into two approaches: appearance-based approach and feature-based approach. Appearance-based approach is relatively simple to implement, it, however, cannot effectively address head movements, despite much effort in this area [5], [16]. This research focuses on feature-based approach. The feature-based gaze estimation approach can be further classified into two groups: 2D mapping based gaze estimation methods and 3D model-based gaze estimation methods.

The 2D mapping approaches [3], [8] assume the mapping from 2D features (e.g., contours, eye-corners, pupil center, etc.) to gaze (3D gaze direction or 2D gaze point) a polynomial mapping function. A major issue with this approach is that the mapping function varies with head pose. It hence cannot effectively handle head movement.

On the other hand, the 3D model-based gaze estimation directly computes 3D gaze direction from eye features based on a geometric model of the eye using stereo cameras [7], [9]–[11] or a single camera with multiple

Manuscript received June 29, 2014; revised October 22, 2014; accepted December 1, 2014. Date of publication December 18, 2014; date of current version February 11, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shiguang Shan.

J. Chen is with the Computer Vision Laboratory, GE Global Research Center, Schenectady, NY 12309 USA (e-mail: chenji@ge.com).

Q. Ji is with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: jiq@rpi.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2383326

calibrated light sources [6]. Given the 3D gaze direction, 2D gaze point on the screen is estimated by intersecting the gaze direction (visual axis) with the screen, and this estimation is invariant to head movement. Despite being head-motion invariant and more accurate, 3D model-based approach still needs personal calibration to estimate the angles between the visual and optical axes since the angles vary with people.

Although the personal calibration can be performed with a single calibration target [7], it still requires active user participation, which is always thought to be intrusive for natural interaction. For some applications, any explicit personal calibration, however short they are, are not acceptable. For example, in covert monitoring of uncooperative subject like kids [7] or mentally challenged subjects [17], it is impossible to reliably perform even a one-point calibration. A "calibration-free" or implicit calibration system could also open up a new way of attentive user interface [18]. For example, if video players on mobile and tablet devices could naturally capture user's gaze data over a video clip without interrupting, it would be very useful for market research [19].

Most recently, some implicit personal calibration methods [13], [14], [20] have been proposed. Model and Eizenman [20] proposed to estimate the eye parameters by exploiting the binocular constraint. Based on the correlation between saliency map and gaze, Sugano et al. [13], [14] offered a 2D appearance-based gaze estimation without calibration. They propose to learn a probabilistic mapping (Gaussian Process Regresser) between the eye image and the gaze point. In order to collect enough data to train this complex non-linear mapping function, they ask the subject to watch a 10-minute video. For each frame of the shown video, they extract its saliency map [21], which represents the distinctive image features attracting more attention. Finally, by treating the saliency map as the probability distribution of gaze, they generate the training gaze points by sampling from the saliency map. However, watching a movie for 10 minute for training is rather burdensome for the user. Furthermore, since they employ a 2D mapping method which doesn't consider head pose, the user has to fix their head on a chin rest.

In summary, explicit personal calibration usually achieves a high accuracy for cooperative subjects, although it may be intrusive for natural interaction. Implicit personal calibration requires less or no active user participation, which makes it more suitable for applications such as covert monitoring. However, the existing "calibration free" eye tracking methods usually depend on more assumptions such as binocular constraints or saliency assumption.

In this paper, based on the extension of our prior work [22], we propose a probabilistic 3D gaze estimation method without explicit personal calibration. Our method is based on combining prior gaze distribution with 3D eye model. Although both [14] and our method use saliency map, they are fundamentally different in how the saliency map is used. Compared to Sugano's method [13], [14], our method has the following advantages:

- We propose a systematic probabilistic framework to derive the analytic solutions to eye parameter and eye gaze, while their method is primarily numerical via sampling.
- Because of the use of 3D gaze estimation, our method is more accurate and allows free head movement. The experimental result shows that our system achieves less than three degrees average accuracy for different people.
- Thanks to the incremental learning, our method does not need an explicit training procedure. It keeps improving the estimation as the user continues using the system. Different from the incremental learning method in [23], our incremental learning method is fully non-intrusive and automatic, without any cooperation from the user.
- Finally, to overcome the limitations with the saliency map, we propose to use a Gaussian prior to replace it.

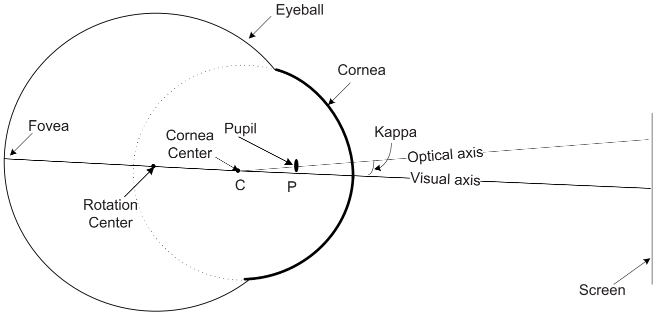


Fig. 1. Structure of the eyeball.

## III. 3D Model-Based Gaze Estimation

Before introducing our method, we briefly summarize the 3D gaze estimation techniques.

### A. 3D Eyeball Structure

As shown in Figure 1, the eyeball is made up of the segments of two spheres of different sizes [24]. The smaller anterior segment is the cornea. The cornea is transparent, and the pupil is inside the cornea. The optical axis of the eye is defined as the 3D line connecting the center of the pupil (**p**) and the center of the cornea (**c**). The visual axis is the 3D line connecting the corneal center (**c**) and the center of the fovea (i.e. the highest acuity region of the retina). Since the gaze point is defined as the intersection of the visual axis rather than the optical axis with the scene, the relationship between these two axes has to be modeled. The angle between the optical axis and visual axis is named *kappa* ($\kappa$), which is a constant value for each person. In gaze estimation methods relying on explicit calibration, $\kappa$ is estimated through a personal calibration.

### B. Personal Calibration

Here, we implement the 3D gaze estimation system in [6], where the cornea center **c** and optical axis **o** are directly estimated from a single camera and two infrared lights. The estimated 3D optical axis can be represented by horizonal and vertical angles ($\mathbf{o} = (\theta, \varphi)$) as shown in Figure 2. By adding $\kappa = (\alpha, \beta)$ to the optical axis, the unit vector of the visual

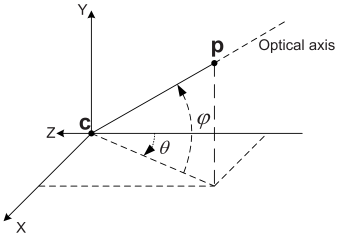


Fig. 2. Orientation of optical axis.

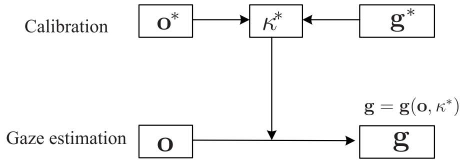


Fig. 3. Diagram of conventional 3D gaze estimation. where $\mathbf{g}^*$ is ground-truth gaze.

axis is represented as:

$$\mathbf{v}_g = \begin{pmatrix} \cos(\varphi+\beta)\sin(\theta+\alpha) \\ \sin(\varphi+\beta) \\ -\cos(\varphi+\beta)\cos(\theta+\alpha) \end{pmatrix} \tag{1}$$

Finally, the gaze point $\mathbf{g}$ on the screen is estimated by intersecting $\mathbf{v}_g$ with the screen. Here, we use a coordinate frame affixed on the screen, with the screen plane as $Z = 0$, thus $\mathbf{g}$ can be written as $\mathbf{g} = (g_x, g_y, 0)^T$. This gaze point is determined by the optical axis and $\kappa$:

$$\begin{aligned} \mathbf{g} &= \mathbf{g}(\mathbf{o}, \kappa) = \mathbf{g}(\varphi, \theta, \alpha, \beta) \\ &= \mathbf{c} + k_c \cdot \begin{pmatrix} \cos(\varphi+\beta)\sin(\theta+\alpha) \\ \sin(\varphi+\beta) \\ -\cos(\varphi+\beta)\cos(\theta+\alpha) \end{pmatrix}, \end{aligned} \tag{2}$$

where $\mathbf{c}$ is the cornea center. Because the z-component of $\mathbf{g}$ equals 0, the value of $k_c$ can be computed by solving equation: $0 = c_z - k_c \cdot \cos(\varphi+\beta)\cos(\theta+\alpha)$:

$$k_c = \frac{c_z}{\cos(\varphi+\beta)\cos(\theta+\alpha)}. \tag{3}$$

However, because $\kappa$ varies for different subjects, it needs to be estimated beforehand through calibration. In explicit personal calibration, the subject is asked to look at $N$ specific calibration points on the screen: $\mathbf{g}_i^*, i = 1, \ldots, N$. The eye parameter can then be estimated by minimizing the distance between the estimated gaze points and these ground-truth gaze points:

$$\kappa^* = \arg\min_{\kappa} \sum_i \|\mathbf{g}_i^* - \mathbf{g}(\mathbf{o}_i^*, \kappa)\| \tag{4}$$

where $\mathbf{o}_i^*$ is the estimated optical axis when subject is looking at the $i$th gaze point $\mathbf{g}_i^*$. The conventional gaze estimation method relying on explicit calibration can be represented as Figure 3. During calibration, the eye parameter $\kappa^*$ is estimated from the calibration gaze point $\mathbf{g}^*$ and the predicted optical axis $\mathbf{o}^*$. During gaze estimation, the eye parameter $\kappa^*$ is fixed, and a new optical axis $\mathbf{o}$ is estimated from the camera. The gaze point is determined by $\mathbf{o}$ and $\kappa^*$ through Eq. 2.

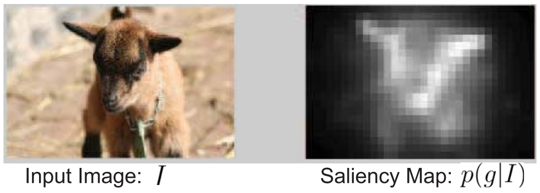


Fig. 4. An example of saliency map ($p(g|I)$).

## IV. Probabilistic Gaze Estimation

In the explicit personal calibration, in order to acquire the ground-truth gaze points to estimate $\kappa$, the subject has to look at some specific points. This procedure is often inconvenient and unnatural. Here, we propose a novel framework to estimate the probability of $\kappa$ and eye gaze without requiring the subject to look at specific calibration points.

### A. Proposed Probabilistic Framework

The basic idea is to replace the need of looking at specific gaze points on the screen with a probability distribution of the gaze when the subject is naturally viewing the screen. It is already known that where people look at in an image is affected both by bottom-up (image saliency) and top-down (subject intention/task) mechanisms. The seminal work [25] and recent evidence [26] have shown that the top-down mechanism is influenced by many factors, including the states of subject (memories, age, gender, experiences) and the tasks the subject is performing (searching, browsing, recognizing), and it is hard to model all possible factors. On the other hand, a number of bottom-up saliency estimation methods have been developed and the state-of-the-art work [27] can successfully predict the attention of an average observer in free-viewing task. By free-viewing, we mean the scenario in which a subject is viewing the screen without a specific goal. In this scenario, a correlation between bottom-up saliency and fixation location has been observed [21], [28]. Thus, for this research, we focus on free-viewing scenarios, without giving subjects any specific instructions, and model the gaze prior probability distribution based on bottom-up mechanism. Two methods are employed to model the gaze probability, namely the saliency map and Gaussian distribution.

*1) Gaze Probability Distribution From the Saliency Map:* Saliency map estimation is based on the assumption that biologically a human tends to gaze at a region containing unique and distinctive visual features compared to the surrounding regions. After Koch and Ullman's seminal work [29], various saliency map estimation methods have been proposed based on this biological motivation [21], [30] or based on a learning from ground-truth gaze data [28], [31] (see [27] for an extensive review and comparison of saliency models).

We utilized the method in [21] to estimate the saliency map of each image, which represents the distinctive features in the image. This model shown it efficacy for gaze estimation in [14]. An example of saliency map is shown in Figure 4. The experimental results in [21] show remarkable consistency between the saliency map and the gaze. Thus, given the image $I$ on the screen, the gaze distribution can be represented

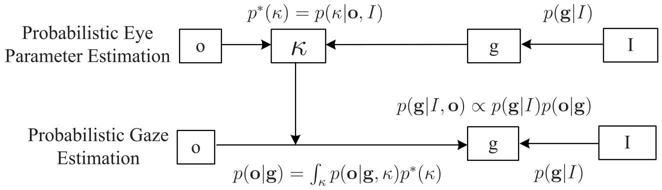

$p^*(\kappa) = p(\kappa|\mathbf{o}, I)$
$p(\mathrm{g}|I)$
Probabilistic Eye Parameter Estimation
o
$\kappa$
g
I
$p(\mathrm{g}|I, \mathbf{o}) \propto p(\mathrm{g}|I)p(\mathrm{o}|\mathrm{g})$
Probabilistic Gaze Estimation
o
g
I
$p(\mathrm{o}|\mathrm{g}) = \int_\kappa p(\mathrm{o}|\mathrm{g}, \kappa)p^*(\kappa)$
$p(\mathrm{g}|I)$

Fig. 5. Diagram of the probabilistic gaze estimation.

as the conditional probability of the gaze position $p(\mathbf{g}|I)$. Here, we make the same assumption as the one in [14] that the user is more likely gazing at the salient regions of the image.

*2) Gaussian Gaze Distribution:* In the above section, we approximate the gaze probability distribution with the saliency map. While the saliency map can approximate the gaze distribution for static image-viewing task, its predictive power reduces significantly when subject is viewing continuous movies of a dynamic environment [32]. Furthermore computing the saliency map frame-by-frame is time consuming.

In this section, we extend our method to more general scenarios where the subject is watching a video or movie. Under such scenarios, we assume the subject is naturally watching the computer screen, and that most of the fixations are concentrated on the center of the screen, with peripheral vision on the margins of the screen. This assumption simply means the probability of a gaze is located near the center is higher than away from the screen center. This assumption works well when people are watching videos or movies because the movie cameraman usually capture the videos with objects of interest in the center. Similar assumptions have been widely used in saliency map estimation [28], and used as a baseline to evaluate other saliency models [27]. In [28], it was shown that a saliency map only based on the distance of each pixel to the center of the image provides a better prediction of the gaze than many previous saliency models. However, its efficacy for gaze estimation has not been proven before. Nevertheless, these work provides a basis for the proposed Gaussian gaze prior.

Given this understanding, we can characterize the gaze probability distribution as a simple Gaussian distribution $\mathcal{N}(\mu, \Sigma)$. Its mean is located in the center of the display ($x = 640, y = 512$), and its variances can be empirically estimated based on historical data. Thus, the gaze probability can be either computed from the saliency map $p(g) = p(g|I)$ or from the assumption of Gaussian gaze distribution, i.e, $p(g) = \mathcal{N}(\mu, \Sigma)$, where we have omitted the image "I" from $p(g)$ since $p(g)$ is the same for all images.

*3) Probabilistic Gaze Estimation:* Based on this gaze probability, we propose the new gaze estimation framework shown in Figure 5. Notice the differences between our method and the conventional method in Figure 3:

1) Firstly, the conventional method needs to collect the ground-truth gaze ($\mathbf{g}^*$), when the subject is looking at specific points during calibration, while our method only needs the gaze probability $p(\mathbf{g}|I)$, when the subject is looking at the image $I$.
2) Secondly, the conventional method estimates the eye parameter $\kappa^*$ deterministically. However, without ground-truth gaze, we cannot deterministically estimate the value of $\kappa^*$. Instead, we estimate $\kappa^*$ probabilistically through the probability distribution of its measurement $\kappa$, i.e., $p^*(\kappa)$. Notice that the ground-truth value of $\kappa^*$ is a constant, but its measurement $\kappa$ is a random variable following $p^*(\kappa)$.[1]
3) Thirdly, during gaze estimation, the conventional method estimates gaze only from the optical axis and $\kappa^*$, while our method first estimates the gaze likelihood $p(\mathbf{o}|\mathbf{g})$ from the optical axis and $p^*(\kappa)$, then combines it with the gaze's prior probability $p(\mathbf{g}|I)$ (e.g. from the saliency map) to estimate gaze posterior probability.

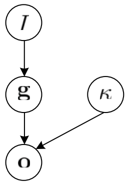


Fig. 6. Probabilistic relationships in BN.

This framework is mainly composed of two parts: *probabilistic eye parameter estimation* and *probabilistic gaze estimation*. We discuss them separately in the following two sections.

### B. Probabilistic Eye Parameter Estimation

In this section, we discuss the method to estimate the eye parameter ($\kappa$) probability from gaze probability (e.g. the saliency map). Firstly, we introduce a general graphical model to represent the relationships between the shown image ($I$), eye gaze ($\mathbf{g}$), optical axis ($\mathbf{o}$), and the eye parameters ($\kappa$).

Figure 6 is the Bayesian Network (BN) [33] that represents the probabilistic relationships. The nodes in the BN represent random variables, and the links represent the conditional probability distributions (CPDs) of nodes given their parents. Based on the gaze probability map and the eye model, we define the CPDs as follows:

1) *$p(\mathbf{g}|I)$:* $\mathbf{g}$ is a 2D vector $\mathbf{g} = (x, y)$, which represents the location of the gaze on the screen (According to the resolution of the monitor, the gaze position is discretized in the range: $0 < x < 1280, 0 < y < 1024$). The link $I \rightarrow \mathbf{g}$ is quantified by $p(\mathbf{g}|I)$ which is the gaze probability distribution estimated from image.
2) *$p(\mathbf{o}|\mathbf{g}, \kappa)$:* $\mathbf{o}$ has two parents $\mathbf{g}$ and $\kappa$. As discussed above, the camera in a gaze system cannot directly observe the visual axis and the gaze. It can only observe the optical axis ($\mathbf{o}$) as the measurement of gaze ($\mathbf{g}$). In the conventional method, $\mathbf{o}$ is a deterministic function of $\mathbf{g}$ by subtracting a constant bias $\kappa$, ignoring any uncertainties. In our proposed method, considering the noise in the gaze system, we model the conditional probability as a Gaussian distribution:

$$p(\mathbf{o}|\mathbf{g}, \kappa) = \mathcal{N}(f(\mathbf{g}, \kappa), \Sigma) \tag{5}$$

where $\mathbf{o} = (\theta, \varphi)^T$ is a 2D vector. $f(\mathbf{g}, \kappa)$ is the inverse function of Eq. 2, which estimates the

[1] The measurement $\kappa$ follows conditional probability $p^*(\kappa|\kappa^*)$. Because $\kappa^*$ is a constant, we use $p^*(\kappa)$ for notational clarity.

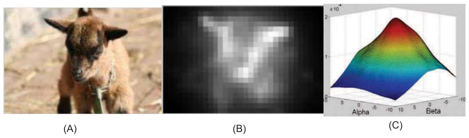


Fig. 7. An example of probabilistic eye parameter estimation. (a) is the image shown to the user; (b) is the gaze probability map $p(g|I)$; (c) is the estimated probability distribution of eye parameters $p^*(\kappa)$.

optical axis by subtracting $\kappa$ from the visual axis. Based on Eq. 2, the directional vector of visual axis can be computed as $\mathbf{d} = \mathbf{g} - \mathbf{c}$, and the horizontal and vertical angles of the visual axis are $\arctan(d_x/d_z)$ and $\arctan(d_y/\sqrt{(d_x^2 + d_z^2)})$ respectively. Finally, the optical axis is $\mathbf{o} = (\arctan(d_x/d_z) - \alpha, \arctan(d_y/\sqrt{(d_x^2 + d_z^2)}) - \beta)^T$. $\Sigma$ models the noise of the optical axis which is estimated from 3D gaze estimation system in [6]. (According to previous tests of this system, we set the standard deviation of the optical axis as one degree on both $\theta$ and $\varphi$.)

Now, based on the BN model, eye parameter estimation is solved as an inference problem in the BN, which estimates the posterior probability $p(\kappa|\mathbf{o}, I)$ given the optical axis and the shown image. Based on the conditional independencies in the BN model in Figure 6, the probability of $\kappa$ can be written as:

$$\begin{aligned} p(\kappa|\mathbf{o}, I) &= \int_{\mathbf{g}} p(\kappa, \mathbf{g}|\mathbf{o}, I) \\ &= \int_{\mathbf{g}} \frac{p(\mathbf{g}|I)p(\mathbf{o}|\mathbf{g}, \kappa) \cdot p(\kappa) \cdot p(I)}{p(\mathbf{o}, I)} \\ &\propto \int_{\mathbf{g}} p(\mathbf{g}|I)p(\mathbf{o}|\mathbf{g}, \kappa) \end{aligned} \tag{6}$$

$p(\mathbf{g}|I)$ is the gaze probability map; $p(\mathbf{o}|\mathbf{g}, \kappa)$ is the Gaussian distribution as defined in Eq. 5. Notice that, the prior probability $p(\kappa)$ is initially assumed to be a uniform distribution, thus $p(\kappa)$ is a constant. For a specific $I$ and $\mathbf{o}$, $p(I)$ and $p(\mathbf{o}, I)$ are constant, and $p(\kappa|\mathbf{o}, I)$ becomes a function of single variable $\kappa$. Therefore, Eq. 6 is a one-step belief propagation that propagates the probability from the gaze to $\kappa$ given one optical axis. The gaze position is discrete in limited range; thus the integral in the above equation can be approximated by summation.

Figure 7(C) shows an example of the estimated eye parameter probability. Here, we collected 40 optical axes when the subject was looking at the image in Figure 7(A). i.e., the training optical axes are $\mathbf{o}_{1,\ldots,40}$ and their corresponding shown images $I_{1,\ldots,40}$ are the same. Based on the probabilistic dependencies in the Bayesian network (Figure 6), we assume that optical axes $\mathbf{o}_{1,\ldots,40}$ are conditionally independent to each other given $\kappa$ and $I_{1,\ldots,40}$. We can then estimate the $\kappa$ probability as the product of each single probability:

$$p^*(\kappa) = p(\kappa|\mathbf{o}_{1,\ldots,40}, I_{1,\ldots,40}) \propto \prod_{i=1}^{40} \int_{\mathbf{g}_i} p(\mathbf{g}_i|I_i)p(\mathbf{o}_i|\mathbf{g}_i, \kappa) \tag{7}$$

Based on the biological study [34], eye parameters should be in a limited range for normal eyes. Here we restricted the eye parameter in the range $-10^o < \alpha < 10^o$ and $-10^o < \beta < 10^o$.

### C. Probabilistic Gaze Estimation

Given the estimated eye parameter probability $p^*(\kappa)$, we can estimate the gaze probability. For consistency, this derivation is based on the same BN model in Figure 6. Unlike the eye parameter estimation, the estimated $p^*(\kappa)$ is now used as the prior probability of the $\kappa$ node. Then, the probability of the gaze, given the optical axis and the shown image, can be written as:

$$p(\mathbf{g}|\mathbf{o}, I) \propto p(\mathbf{g}|I)p(\mathbf{o}|\mathbf{g}) \tag{8}$$

where $p(\mathbf{g}|I)$ is the prior probability of gaze from either the saliency map of the shown image $I$ or the Gaussian gaze distribution, and $p(\mathbf{o}|\mathbf{g})$ is the gaze likelihood, which can be derived from $p^*(\kappa)$ as:

$$p(\mathbf{o}|\mathbf{g}) = \int_{\kappa} p(\mathbf{o}|\mathbf{g}, \kappa)p^*(\kappa) \tag{9}$$

Note that all the derivations above are only valid based on the conditional independencies in the BN model.

Thus, the probabilistic gaze estimation is composed of the following steps:

1) First, we estimate the gaze prior probability distribution $p(\mathbf{g}|I)$ either from the saliency map or from the Gaussian gaze distribution.
2) Then, we estimate the likelihood gaze map $p(\mathbf{o}|\mathbf{g})$, given the current optical axis and the eye parameter prior $p^*(\kappa)$, based on Eq. 9.
3) Finally, the product $p(\mathbf{g}|I)p(\mathbf{o}|\mathbf{g})$ represents the gaze posterior probability map $p(\mathbf{g}|\mathbf{o}, I)$. Given the posterior probability, the maximum posterior point is selected as the gaze point $\mathbf{g}^*$:

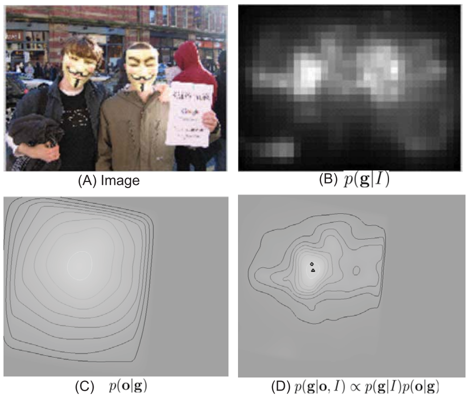
$$\mathbf{g}^* = \arg\max_{\mathbf{g}} p(\mathbf{g}|\mathbf{o}, I) \tag{10}$$

The results of the three steps are shown in Figure 8. Here we compare our method with the conventional gaze estimation method relying on explicit personal calibration. The conventional method can achieve one degree of accuracy. The peak in our posterior probability map is very close to the estimated gaze of the conventional method as shown in Figure 8, but our method does not need any explicit calibration.

## V. Incremental Learning for Gaze Estimation

In order to provide a more natural user experience, we propose an incremental learning algorithm for our probabilistic framework. This new framework does not need any prior training. It can quickly adapt to the user, and incrementally improves gaze estimation accuracy as the subject uses the system.

We first assume the initial distribution of $\kappa$ as uniform. When the subject starts to use the system, we record a sequence of his optical axes $\mathbf{o}_{t,\ldots,1}$. Given the corresponding



(A) Image (B) $p(\mathbf{g}|I)$

(C) $p(\mathbf{o}|\mathbf{g})$ (D) $p(\mathbf{g}|\mathbf{o}, I) \propto p(\mathbf{g}|I)p(\mathbf{o}|\mathbf{g})$

Fig. 8. Probabilistic gaze estimation. (A) is the shown image. (B) is the saliency map $p(\mathbf{g}|I)$ of the image. (C) is the gaze likelihood map given the optical axis. (D) is the gaze posterior probability map. The triangle shows the maximum posterior point. The circle shows the estimated gaze using the conventional method.

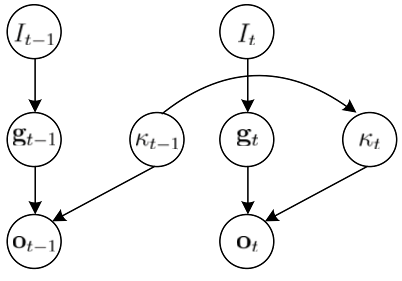


Fig. 9. DBN for incremental learning.

shown image sequence $I_{t,\ldots,1}$, the incremental learning framework continually updates the estimations of $\kappa$ and gaze given all previous information, i.e. estimating $p(\kappa_t|I_{t,\ldots,1}, \mathbf{o}_{t,\ldots,1})$ and $p(\mathbf{g}_t|I_{t,\ldots,1}, \mathbf{o}_{t,\ldots,1})$. We employ a recursive updating procedure detailed as follows.

For incremental learning, we extend the BN to a dynamic BN (DBN) model as shown in Figure 9. In general, a DBN is comprised of interconnected time slices of static BNs. One important assumption of DBN is first-order Markovian, i.e., given the state of the closest previous time slice, the current time slice is independent from other past time slices. Thus, in Figure 9, we only show the DBN of the current and previous time slices. It includes two kinds of links. *Intra-frame links* in the current time slice are the same as the BN model we set before and *inter-frame link* from $\kappa_{t-1}$ to $\kappa_t$ captures the temporal relationships. Based on the anatomy, $\kappa$ cannot vary much over time. Thus, we model it as a Gaussian distribution:

$$p(\kappa_t|\kappa_{t-1}) = \mathcal{N}(\kappa_{t-1}, \Sigma_k) \tag{11}$$

where $\Sigma_k$ is the covariance matrix which allows $\kappa_t$ to vary in a small range around the previous estimation $\kappa_{t-1}$. It depends on the uncertainty in our system. Here we empirically set the standard deviations of $\alpha_t$ and $\beta_t$ to one degree, i.e. $\Sigma_k$ is an identity matrix.

Given the above temporal relationship in the DBN, the probability of $\kappa$ can be updated recursively. Firstly, we predicted the prior probability of the current $\kappa_t$ based on its previous

**Algorithm 1** Incremental Gaze Estimation Algorithm

```
$t \leftarrow 1$
Set $p^*(\kappa_1)$ as uniform distribution.

Estimate the first gaze:
$p(\mathbf{g}_1|I_1, \mathbf{o}_1) \propto p(\mathbf{g}_1|I_1) \cdot \int_{\kappa_1} p(\mathbf{o}_1|\mathbf{g}_1, \kappa_1)p^*(\kappa_1)$

Update $\kappa$ probability :
$p'(\kappa_1) \propto \int_{\mathbf{g}_1} p(\mathbf{g}_1|I_1)p(\mathbf{o}_1|\mathbf{g}_1, \kappa_1)$
loop
  $t \leftarrow t+1$
  Temporal belief propagation $p'(\kappa_{t-1}) \rightarrow p^*(\kappa_t)$ :
  $p^*(\kappa_t) = \int_{\kappa_{t-1}} p(\kappa_t|\kappa_{t-1})p'(\kappa_{t-1})$
  Probabilistic gaze estimation:
  $p(\mathbf{g}_t|I_{t,..,1}, \mathbf{o}_{t,..,1}) \propto p(\mathbf{g}_t|I_t) \cdot \int_{\kappa_t} p(\mathbf{o}_t|\mathbf{g}_t, \kappa_t)p^*(\kappa_t)$

  Update $\kappa$ probability:
  $p'(\kappa_t) \propto \int_{\mathbf{g}_t} p(\mathbf{g}_t|I_t)p(\mathbf{o}_t|\mathbf{g}_t, \kappa_t) \cdot p^*(\kappa_t)$
end loop
```

probability, as shown in Eq. 12.

$$\begin{aligned} & p(\kappa_t|I_{t-1,\ldots,1}, \mathbf{o}_{t-1,\ldots,1}) \\ & \quad = \int_{\kappa_{t-1}} p(\kappa_t|\kappa_{t-1})p(\kappa_{t-1}|I_{t-1,\ldots,1}, \mathbf{o}_{t-1,\ldots,1}) \end{aligned} \tag{12}$$

where $p(\kappa_{t-1}|I_{t-1,\ldots,1}, \mathbf{o}_{t-1,\ldots,1})$ is the $\kappa$ probability from the previous time frame. Since the temporal CPD $p(\kappa_t|\kappa_{t-1})$ is a Gaussian distribution, this integral is implemented by a convolution of previous $\kappa$ probability map with a Gaussian kernel. (In the first time frame, when no prior information of $\kappa$ is available, we assume $p(\kappa_1)$ is uniformly distributed.)

Based on the predicted temporal prior probability $p(\kappa_t|I_{t-1,\ldots,1}, \mathbf{o}_{t-1,\ldots,1})$, the current probability of $\mathbf{g}_t$ and $\kappa_t$ can be derived as the filtering problem in the DBN:

$$\begin{aligned} & p(\mathbf{g}_t|I_{t,t-1,\ldots,1}, \mathbf{o}_{t,t-1,\ldots,1}) \\ & \quad \propto p(\mathbf{g}_t|I_t) \cdot \int_{\kappa_t} p(\mathbf{o}_t|\mathbf{g}_t, \kappa_t)p(\kappa_t|I_{t-1,\ldots,1}, \mathbf{o}_{t-1,\ldots,1}) \end{aligned} \tag{13}$$

$$\begin{aligned} & p(\kappa_t|I_{t,t-1,\ldots,1}, \mathbf{o}_{t,t-1,\ldots,1}) \\ & \quad \propto \int_{\mathbf{g}_t} p(\mathbf{g}_t|I_t)p(\mathbf{o}_t|\mathbf{g}_t, \kappa_t) \cdot p(\kappa_t|I_{t-1,\ldots,1}, \mathbf{o}_{t-1,\ldots,1}) \end{aligned} \tag{14}$$

The current estimation of $p(\kappa_t|I_{t,t-1,\ldots,1}, \mathbf{o}_{t,t-1,\ldots,1})$ is updated recursively from its previous estimation $p(\kappa_{t-1}|I_{t-1,\ldots,1}, \mathbf{o}_{t-1,\ldots,1})$, based on Eq.12 and Eq. 14.

Letting $p^*(\kappa_t) = p(\kappa_t|I_{t-1,\ldots,1}, \mathbf{o}_{t-1,\ldots,1})$ and $p'(\kappa_t) = p(\kappa_t|I_{t,\ldots,1}, \mathbf{o}_{t,\ldots,1})$, the above incremental learning algorithm can be summarized in Algorithm 1.

An example of the incremental learning of $p'(\kappa_t)$ is shown in Figure 10. The estimated $p'(\kappa_1)$ for the first time frame has a high probability in multiple regions. By updating its probability incrementally, it gradually converges to a single peak after twenty time frames.

The estimation of $\kappa$ is shown in Figure 11. Note that our algorithm used the whole $\kappa$ probability map rather than

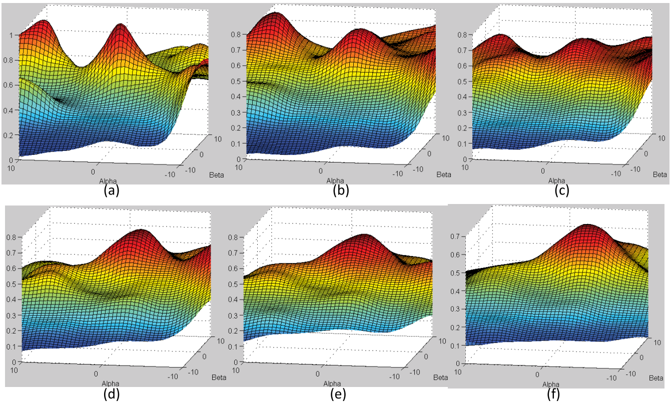
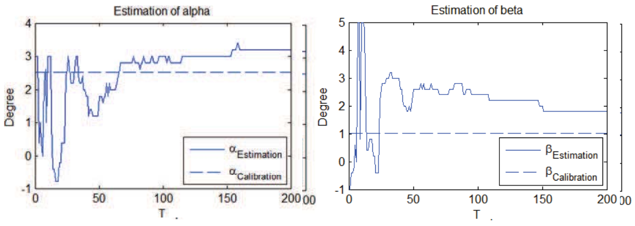


Fig. 10. The probability distribution of eye parameters after the first (a) frame, (b) 4 frames, (c) 8 frames, (d) 12 frames (e) 16 frames and (f) 20 frames.



(a) (b)

Fig. 11. Estimation of $\kappa$ in incremental learning. The estimated $\alpha$ and $\beta$ are shown as solid lines, and their ground-truth value from calibration are shown as dashed lines. In the beginning, this estimation oscillates significantly because the probability map hasn't converged, and it includes several peaks (Figure 10). Finally, the estimate $\alpha$ converges to the ground-truth values. (a) Estimation of alpha. (b) Estimation of beta.

a single point. Here, we show the maximum point in the probability map (Figure 10) as our $\kappa$ estimation. $\kappa = (\alpha, \beta)$ includes two parameters which are shown in separate figures.

The only difference between the DBN and the BN is that the DBN considers the temporal prior of $\kappa_t$, and continues updating it over time. For example, if letting $p^*(\kappa) = p(\kappa_t|I_{t-1,\ldots,1}, \mathbf{o}_{t-1,\ldots,1})$, Eq. 13 is the same as Eq. 8; if letting $p(\kappa_t|I_{t-1,\ldots,1}, \mathbf{o}_{t-1,\ldots,1})$ be uniform distribution, Eq. 14 is the same as Eq. 6.

## VI. Experimental Results

We evaluate our system when the subject is looking at a standard 19-inch monitor (37.63cm×30.11cm). Our system allows free head movement, the range of the distance between the monitor and the subjects' eyes is about 45−70cm. All the subjects were unaware of the purpose of the experiment. We perform an explicit 9-point calibration only afterward to get accurate estimation of their eye parameters. To evaluate a gaze estimation method, the subjects are often asked to look at some points on the screen. The gaze estimation error can be computed as the distance between these points and the estimated gaze points. However, in our method the user does not need to look at any specific points. To evaluate our system, we implemented the 3D gaze estimation system [6]. This system is first calibrated by asking the subject to look at nine points on the screen. The average accuracy of this system is one degree for different subjects. We evaluate our proposed method by comparing with this calibrated system.

To evaluate our method, we collected the optical axes of ten subjects while they viewed the images on the screen. Each image displayed for about four seconds on the screen. Our gaze system collects eighty optical axes for each image (our gaze system captured the video of the eye and estimated the optical axes at 20 frames per second). Because the relationship between the saliency and gaze cannot be guaranteed during saccadic eye movements, we remove the saccadic movement in two steps. First, when the subject is looking at images, when he/she switches the image, we filter out the data in the beginning (< 300ms) since we believe most of the gaze movements in the beginning are saccadic movements. Second, as the study continues, most of eye gaze movements are fixations and we filter out some very short gaze fixations (less than 100ms) and treat them as saccadic eye movements. Our study shows that less than 10% movements are saccadic eye movements after the initial stage.

In the experiment, the main computation cost is on the optical axis estimation (50ms per frame), which includes pupil and glints detection from image. Our gaze estimation from optical axis only added a small cost (5ms per frame). Notice that this doesn't include the cost of computing saliency map, because we compute the saliency of the test images beforehand.

To show the advantages of the incremental learning, we compared the incremental learning algorithm (Section V) to the batch training method in Section IV-B.

### A. Batch Training for Gaze Estimation

For batch training, we divided the eighty time frames when the subject viewed one image into training data (forty frames) and testing data (forty frames). Each subject viewed five images in this experiment. Please notice that we only use images with clear salient objects in our experiment. The saliency entropy is used as the criteria to select images from Google image search. Specifically, only images with entropy lower than 13 are selected in our experiment. These images usually have some clear salient objects as shown in Figure 12. For comparison, we also selected some poor images with high saliency entropy to study the impact of poor saliency map in section VI-D.

We used leave-one-image-out cross validation, i.e. when testing on forty frames of one image, we first learned the eye parameter probability from the training data of the other four images (Section IV-B). Saliency map is used to approximate the gaze prior distribution.

For a more effective system, we wanted to use less training data, because more training time may make the subject bored and easily distracted. We tested the dependency of our method on the amount of the training data by using 160, 80, 40, and 20 frames of training data, i.e. 40, 20, 10, and 5 frames for each training image.

The average error and the standard deviation of error (over 200 test frames) are shown in Table I. When the number of training frame increases, the average error decreases slowly

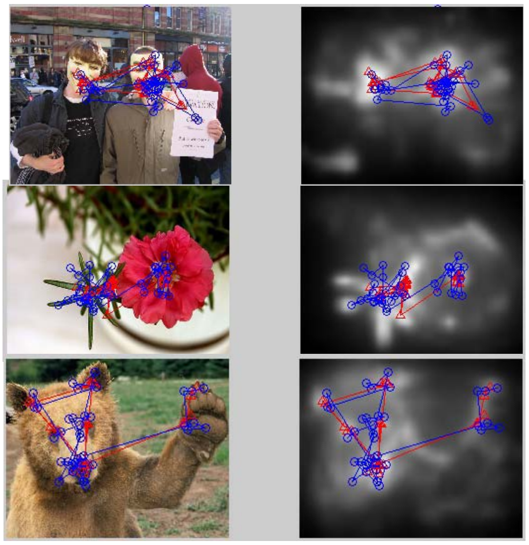


Fig. 12. An example of probabilistic gaze estimation result on three images. Red dots are the results of our proposed method. Blue dots are the results of the conventional method [6] with 9-point calibration. The left column shows the gaze fixations superimposed on the original image, while the right column shows the fixations superimposed on the saliency map.

TABLE I
GAZE ESTIMATION ERROR (IN DEGREE) OF TEN SUBJECTS WITH DIFFERENT TRAINING DATA SIZE. EYE PARAMETERS ARE TRAINED THOUGH BATCH TRAINING

| Training data size | 20 frames | 40 frames | 80 frames | 160 frames |
|---|---|---|---|---|
| Subject 1 | 2.45 | 2.43 | 2.40 | 2.18 |
| Subject 2 | 1.98 | 1.83 | 1.92 | 1.93 |
| Subject 3 | 1.70 | 1.66 | 1.64 | 1.61 |
| Subject 4 | 2.95 | 2.93 | 2.77 | 2.89 |
| Subject 5 | 3.43 | 3.39 | 3.34 | 3.40 |
| Subject 6 | 2.24 | 2.20 | 2.02 | 2.01 |
| Subject 7 | 3.41 | 3.23 | 3.21 | 3.19 |
| Subject 8 | 1.97 | 1.90 | 1.88 | 1.87 |
| Subject 9 | 2.21 | 2.19 | 2.17 | 2.16 |
| Subject 10 | 1.97 | 1.91 | 1.90 | 1.86 |
| Average | 2.43 | 2.37 | 2.33 | 2.31 |
| Std. | 0.62 | 0.61 | 0.59 | 0.62 |

and standard deviation remains high. Even so, the average gaze estimation accuracy can achieve 2.31° using 160 frames. In the following section VI-B we will show that the gaze estimation accuracy can increase much faster with incremental learning.

### B. Incremental Learning for Gaze Estimation

Based on our incremental learning algorithm, the system doesn't need to estimate the eye parameter probability beforehand using training frames. This system can automatically update the eye parameter probability and estimate the gaze when the subject starts using the system. Again, saliency map is used to approximate the gaze prior distribution.

The gaze estimation error and the standard deviation for the first 20, 40, 80, 120, 160, and 200 frames are shown in Table II. Although the error is large for the first few frames (<20 frames), it decreases quickly as the subject uses the systems. Compared with the batch training, the incremental learning achieves similar performance for the first 40 frames. However, when the subject is using the system, incremental learning continues improving the performance and can achieve an average accuracy of 1.78° for the first 200 frames. This process is done automatically, naturally, and without any user knowledge. This result outperforms the batch method in Section VI-A because when a subject is looking at one image, most of his/her gaze converges to a few salient regions, and this gaze distribution is temporally consistent in a short period of time. In incremental learning, there is a good chance that the testing gaze and the most recent training gazes are collected from the same image, and they follow a same distribution. In batch training, however, the training and testing gazes are always from different images.

TABLE II
GAZE ESTIMATION RESULTS OF TEN SUBJECTS FOR THE FIRST N FRAMES (N = 10,20,40,80,120,160,200). EYE PARAMETERS ARE AUTOMATICALLY UPDATED AFTER EACH FRAME

| Training data | 20 frames | 40 frames | 80 frames | 120 frames | 160 frames | 200 frames |
|---|---|---|---|---|---|---|
| Subj. 1 | 1.73 | 2.01 | 2.07 | 2.04 | 1.89 | 1.80 |
| Subj. 2 | 2.50 | 2.06 | 1.90 | 1.84 | 1.89 | 1.80 |
| Subj. 3 | 1.60 | 1.57 | 1.54 | 1.50 | 1.59 | 1.59 |
| Subj. 4 | 4.09 | 2.95 | 2.20 | 1.91 | 2.06 | 2.07 |
| Subj. 5 | 2.95 | 2.81 | 2.09 | 1.68 | 1.60 | 1.60 |
| Subj. 6 | 3.41 | 2.91 | 2.07 | 1.76 | 1.77 | 1.64 |
| Subj. 7 | 3.52 | 3.45 | 2.77 | 2.53 | 2.33 | 2.16 |
| Subj. 8 | 2.54 | 1.93 | 1.88 | 1.87 | 1.83 | 1.75 |
| Subj. 9 | 1.99 | 1.97 | 1.91 | 1.89 | 1.79 | 1.80 |
| Subj. 10 | 1.77 | 1.75 | 1.60 | 1.57 | 1.58 | 1.58 |
| Average | 2.61 | 2.34 | 2.01 | 1.86 | 1.83 | 1.78 |
| Std. | 0.86 | 0.63 | 0.34 | 0.29 | 0.23 | 0.20 |

Some gaze estimation results (in both the original image and the saliency map) of subject 1 are shown in Figure 12. Without explicit calibration, the results of our method are close to the results of the system with 9-point calibration. The subject may look at some region with low saliency, such as the white paper in the person's hand in Figure 12(A). In this case, by incrementally improving the eye parameter estimation and by combining gaze likelihood with the saliency map, our method can still follow the true gaze positions.

Compared to the most recent implicit personal calibration method [13], [14], which asks the subject to watch a ten-minute video for training and achieves an accuracy of 3.5 degrees, our proposed method doesn't need training data beforehand and can adapt to the user very quickly (in 80 frames or less than four seconds), and continues to improve the accuracy as person uses it. The average accuracy can achieve 1.78 degrees. Furthermore, in our 3D gaze estimation framework, the subject can have natural head movement without fixing his/her head in a chin-rest.

### C. Effect of Gaze Prior Probability

Please notice that both batch and incremental algorithm depends on the gaze prior probability (saliency map) $p(\mathbf{g}|I)$ in two aspects. First, eye parameter estimation depends on the integral of this prior probability as shown in Eq. 14. Second, gaze posterior probability is estimated as the production of this prior and the gaze likelihood as shown in Eq. 13. Prior probability is necessary in eye parameter estimation step, but its effect becomes less important for gaze estimation step when the eye parameter is already accurately estimated so

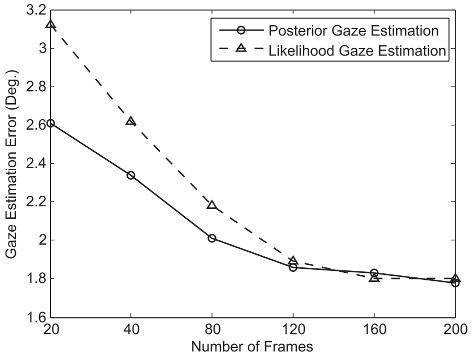


Fig. 13. Comparison of gaze estimation with posterior probability and gaze estimation with likelihood only.

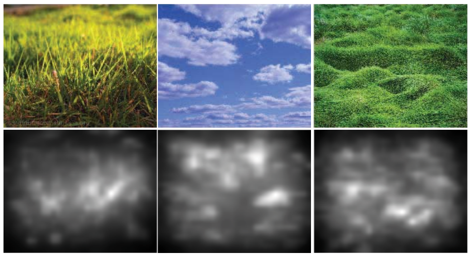


Fig. 14. Saliency maps (bottom row) of images without salient objects (top row).

that the gaze likelihood itself is sufficient. Figure 13 shows the average error of posterior gaze estimation and the error of the gaze estimation with likelihood only, e.g., changing the prior in Eq. 13 to uniform distribution. We can see that proposed gaze prior plays an important role in correcting the gaze estimation errors in the initial 150 frames but its role gradually diminishes as the frames go on. In fact, after 160 frames, the gaze likelihood achieves the same level of accuracy as the posterior estimation. This demonstrates that the posterior estimation is beneficial in the beginning when the eye parameter estimation has not converged, but the two methods are asymptotically equivalent when there is enough training data.

### *D. Gaze Estimation With Low-Quality Saliency Map*

Above saliency-based gaze estimation depends on the quality of saliency map. Here, we consider two cases of low-quality saliency map.

First, we tested our incremental learning algorithm when the subjects are looking at images without salient objects. As discussed in section VI-A, we selected 100 bad images with high entropy from Google image search. Some example images are shown in Figure 14. In this case, the high-salient regions are evenly distributed in the images. We randomly select 5 images out of 100 bad images and perform the same image-viewing experiment with the first five subjects. The gaze estimation error is summarized in Table III. Compared to Table II, the error increases significantly because the salient map cannot provide good prior information for these images.

Second, we consider the case when the image includes salient objects but the saliency estimation algorithm cannot predict an accurate saliency map. To simulate the saliency estimation error, we add noise to the saliency map. For each pixel in the saliency map, we add a uniform noise $\varepsilon = \mathcal{U}(0, \sigma)$. The saliency map with noise level $\sigma = 0.8, 1.6, 2.4$ are shown in Figure 15. The average gaze estimation error are shown in Table IV. When the noise is large, the saliency region is ambiguous in the map and the gaze error increase significantly. The above experiments show that the success of saliency-based gaze estimation highly depends on the quality of saliency map.

TABLE III

GAZE ESTIMATION ERROR (IN DEGREE) OF FIVE SUBJECTS LOOKING AT IMAGES WITHOUT SALIENT OBJECTS

| Training data | 20 frames | 40 frames | 80 frames | 120 frames | 160 frames | 200 frames |
|---|---|---|---|---|---|---|
| Subj. 1 | 3.47 | 1.97 | 1.85 | 2.06 | 2.35 | 2.52 |
| Subj. 2 | 3.34 | 3.38 | 2.88 | 2.61 | 2.29 | 2.14 |
| Subj. 3 | 3.96 | 3.82 | 3.34 | 2.96 | 2.76 | 2.76 |
| Subj. 4 | 8.15 | 6.16 | 4.59 | 3.75 | 3.35 | 3.03 |
| Subj. 5 | 6.67 | 5.85 | 4.98 | 4.58 | 3.92 | 3.52 |
| Average | 5.12 | 4.24 | 3.53 | 3.19 | 2.94 | 2.80 |

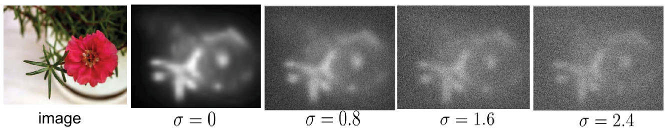


Fig. 15. Saliency maps with different levels of noise.

TABLE IV

AVERAGE GAZE ESTIMATION ERROR (IN DEGREE) OF FIVE SUBJECTS WITH NOISY SALIENCY MAP

| Noise Level $\sigma$ | 0 | 0.8 | 1.6 | 2.4 |
|---|---|---|---|---|
| Gaze Error | 1.77 | 1.79 | 2.06 | 2.75 |

TABLE V

GAZE ESTIMATION ERROR (IN DEGREE) OF TEN SUBJECTS FOR THE FIRST N FRAMES. THE GAZE PRIOR IS ASSUMED AS GAUSSIAN DISTRIBUTION

| Training data | 100 frames | 1000 frames | 2000 frames | 3000 frames | 4000 frames | 5000 frames | 6000 frames |
|---|---|---|---|---|---|---|---|
| Subj. 1 | 5.46 | 4.60 | 3.54 | 2.61 | 2.22 | 1.84 | 1.62 |
| Subj. 2 | 2.94 | 1.59 | 1.03 | 1.42 | 1.45 | 1.51 | 1.51 |
| Subj. 3 | 3.83 | 3.60 | 2.24 | 1.81 | 1.54 | 1.30 | 1.16 |
| Subj. 4 | 3.95 | 1.75 | 1.88 | 1.51 | 1.33 | 1.38 | 1.33 |
| Subj. 5 | 3.56 | 2.11 | 2.27 | 1.64 | 1.28 | 1.10 | 1.02 |
| Subj. 6 | 4.34 | 3.86 | 2.57 | 1.95 | 1.53 | 1.44 | 1.32 |
| Subj. 7 | 3.59 | 2.83 | 2.31 | 1.71 | 1.49 | 1.50 | 1.29 |
| Subj. 8 | 3.85 | 1.92 | 1.87 | 1.88 | 1.31 | 1.11 | 1.10 |
| Subj. 9 | 3.03 | 2.24 | 1.79 | 1.29 | 1.35 | 1.40 | 1.39 |
| Subj. 10 | 4.16 | 3.33 | 2.25 | 1.82 | 1.43 | 1.21 | 1.19 |
| Average | 3.87 | 2.78 | 2.18 | 1.76 | 1.49 | 1.38 | 1.29 |
| std | 0.71 | 1.02 | 0.64 | 0.36 | 0.27 | 0.22 | 0.18 |

TABLE VI

COMPARISON OF THE ESTIMATED EYE PARAMETERS $\kappa^G$, $\kappa^B$, $\kappa^S$ AGAINST THE EYE PARAMETERS $\kappa^*$ FROM CALIBRATION

| Subj. | $\kappa^*$ | | $\kappa^G$ | | $\kappa^B$ | | $\kappa^S$ | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha$ | $\beta$ | $\alpha$ | $\beta$ | $\alpha$ | $\beta$ | $\alpha$ | $\beta$ |
| 1 | -3.11 | 0.02 | -3.0 | 0.2 | -3.0 | 0.2 | -3.0 | 0.0 |
| 2 | 3.07 | -2.43 | 3.0 | -3.0 | 3.2 | -2.4 | 3.0 | -2.4 |
| 3 | 3.51 | -1.33 | 3.8 | -1.0 | 3.4 | -1.2 | 3.4 | -1.0 |
| 4 | 2.51 | 1.01 | 3.0 | 1.0 | 3.2 | 1.8 | 3.2 | 1.8 |
| 5 | -0.51 | 1.02 | -1.0 | 1.6 | -0.6 | -0.2 | -0.4 | -0.2 |
| 6 | -1.02 | 3.10 | -0.8 | 2.6 | 0.4 | 2.8 | -0.2 | 2.8 |
| 7 | 1.04 | -3.15 | 1.6 | -2.2 | 2.0 | -2.4 | 1.8 | -2.6 |
| 8 | -1.31 | 1.35 | -1.2 | 1.2 | -1.0 | 0.8 | -1.0 | 1.0 |
| 9 | 2.79 | -0.80 | 2.8 | -1.0 | 2.4 | -0.6 | 2.4 | -0.8 |
| 10 | 1.10 | -4.21 | 1.0 | -4.0 | 0.8 | -4.0 | 0.8 | -4.2 |

TABLE VII

COMPARISON OF GAZE ESTIMATION ERROR USING THE ESTIMATED EYE PARAMETERS ($\kappa^G$, $\kappa^B$, $\kappa^S$) AND THE ERROR USING THE CALIBRATED EYE PARAMETERS $\kappa^*$

| Subject | $\kappa^*$ | | $\kappa^G$ | | $\kappa^B$ | | $\kappa^S$ | | Optical Axis | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $error$ | $bias$ | $error$ | $bias$ | $error$ | $bias$ | $error$ | $bias$ | $error$ | $bias$ |
| 1 | 1.0 | (-0.3,0.1) | 1.0 | (-0.2, 0.3) | 1.0 | (-0.2,0.3) | 1.0 | (-0.2,0.2) | 2.7 | (2.6,-0.5) |
| 2 | 1.0 | (-0.4,-0.3) | 1.3 | (-0.5,-0.9) | 1.0 | (-0.3,0.3) | 1.0 | (-0.4,0.2) | 4.3 | (-3.2, 2.7) |
| 3 | 0.9 | (0.7,0.1) | 1.2 | (1.0,0.5) | 1.0 | (0.7,0.3) | 1.0 | (0.7,0.5) | 3.4 | (-2.7,1.9) |
| 4 | 1.1 | (0.1,-0.4) | 1.2 | (0.6,-0.4) | 1.3 | (0.8,0.5) | 1.3 | (0.8,0.5) | 3.1 | (-2.5, -1.5) |
| 5 | 1.3 | (0.3, 0.6) | 1.9 | (-0.1,1.3) | 1.4 | (0.2, -0.7) | 1.5 | (0.2, -0.8) | 1.4 | (0.7, -0.6) |
| 6 | 1.0 | (-0.3, -0.2) | 1.1 | (-0.2,-0.8) | 1.3 | (1.0, -0.7) | 1.0 | (0.4, -0.6) | 3.9 | (0.3, -3.8) |
| 7 | 1.2 | (0.1, -0.4) | 1.4 | (0.7,0.7) | 1.5 | (1.1, 0.5) | 1.3 | (0.9, 0.2) | 3.8 | (-0.8, 3.5) |
| 8 | 0.9 | (0.7, 0.1) | 0.9 | (0.8,-0.1) | 1.2 | (1.0, -0.6) | 1.1 | (1.0, -0.4) | 2.5 | (1.9, -1.6) |
| 9 | 1.0 | (-0.4,-0.3) | 1.1 | (-0.4,-0.5) | 1.2 | (-0.7, 0.0) | 1.2 | (-0.8, -0.3) | 3.3 | (-3.0, 0.8) |
| 10 | 1.3 | (0.2, -0.1) | 1.3 | (0.2,0.2) | 1.3 | (0.0, 0.2) | 1.3 | (-0.1, -0.1) | 5.3 | (-0.4, 5.1) |

### *E. Probabilistic Gaze Estimation With Gaussian Gaze Distribution*

In this section, we study the performance of gaze estimation without using saliency map but instead assuming gaze fixations are normally distributed with the mean in the center of the screen as discussed in section IV-A2. Specifically, each user was unconscious of the gaze tracking system. He/she was naturally watching a movie in full screen mode. The experiment lasted about five minutes. The gaze estimation algorithm is similar to Algorithm 1. The difference is that we assume gaze follows a Gaussian distribution for each gaze point $p(\mathbf{g}_t)$. Thus, the probability distribution $p(\mathbf{g}_t|I_t)$ derived from the saliency map is replaced by Gaussian probability distribution $p(\mathbf{g}_t)$.

Our system captured totally 6000 optical axes in this five-minute experiment. The gaze estimation error for ten subjects are summarized in Table V. Please note that the results in Table V and the results of saliency map in Table II are based on different testing data sets, so that they are not directly comparable. However, we can at least draw the following conclusion: The gaze estimation with Gaussian prior needs longer time to converge (from a few seconds to five minutes) and if given enough training time (after 3000 frames), this method can achieve the same level of accuracy as the saliency map based method.

### *F. Eye Parameter Estimation*

In the above experiment, we evaluate different methods under different scenarios, i.e., 'image-viewing' and 'video-viewing'. It is hard to compare their gaze estimation accuracies across scenarios. To make a fair comparison of these methods, we evaluate their performance for eye parameter estimation, since there is only one ground-truth eye parameter for each person.

In our incremental learning, we continue updating the $\kappa$ probability map in the experiment. After the experiment, we can extract the maximum point in the probability map as our estimation of eye parameter $\kappa$. This is our best estimation of $\kappa$ after improving it using all the training frames. We extract $\kappa^S$ using 200 frames in saliency-based incremental method (Sec. VI-B), $\kappa^B$ using saliency-based batch method (Sec. VI-A), and $\kappa^G$ using 6000 frames in Gaussian-based method (Sec. VI-E). We compare our eye parameter estimation with $\kappa^*$ which is obtained from nine-point calibration in Table VI. We can see that our eye parameter estimate is close to the eye parameters from calibration.

In order to further evaluate our eye parameter estimation, we ask the person to look at nine fixed points that are uniformly distributed on the screen. We estimate fixations using $\kappa^*$, $\kappa^B$, $\kappa^S$ and $\kappa^G$ respectively and take the fixed points as ground-truth to compute the gaze estimation errors. The gaze estimation errors and the horizontal and vertical angle bias of the ten subjects are summarized in Table VII. Here, the estimation bias is accessed using the mean signed difference (MSD): $\text{MSD}(x) = \sum_{i=1}^{n} \frac{(\hat{x}_i - x_i)}{n}$, where $\hat{x}_i$ is the estimate and $x_i$ is the ground truth. We also compute the error and bias when we directly use the optical axis to estimate gaze without any calibration, i.e., set $\kappa = (0, 0)$. As expected, the estimation error with optical axis is very large. The error of our method with saliency map or Gaussian prior is a little higher than the calibration-based method, but our method doesn't need the personal calibration procedure, and it can keep improving the gaze estimation when the user continues using the computer naturally. Notice that, compared to the average error of the first $N$ frames in Table II and V, this is the error using the eye parameter after N frames training procedure. Thus this gaze estimation error is smaller.

## VII. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new probabilistic gaze estimation framework by combining the gaze prior with the 3D eye gaze model. Compared to the conventional gaze estimation method, our proposed approach eliminates the explicit personal calibration procedure. It changes conventional deterministic eye gaze tracking to probabilistic eye gaze tracking, allowing to combine eye gaze prior with optical axis estimates to simultaneously estimate 3D gaze point and the personal eye parameters in an incremental manner without any cooperation from the user. Compared with the most recent implicit personal calibration method [13], [14], our system allows natural head movement, and achieves high gaze estimation accuracy of less than three degree (comparing to the accuracy of 3.5 degrees in [14]). By using a novel incremental learning framework, our system doesn't need any training data from the subject beforehand. It can adapt to the user quickly and improves its performance as the subject naturally uses the system. Finally, we further extend our system without computing the saliency map by assuming the prior gaze distribution follows a Gaussian distribution, with a mean located in the center of the screen. This not only improves the speed of our method (without the need of computing saliency map), but also extend its application scope.

Our approach, however, is limited to free-viewing scenarios when subjects are naturally viewing images or videos. Studies in visual attention [25], [26] have already shown that if the user is performing a specific visual task, his/her gaze distribution is driven by the task. The proposed gaze prior, either saliency map or Gaussian gaze prior, may not be applicable in this case. We will study the task-dependent gaze prior in our future work.

## References

[1] R. J. K. Jacob, "The use of eye movements in human-computer interaction techniques: What you look at is what you get," *ACM Trans. Inf. Syst.*, vol. 9, no. 2, pp. 152–169, Apr. 1991.

[2] S. Zhai, C. Morimoto, and S. Ihde, "Manual and gaze input cascaded (MAGIC) pointing," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 1999, pp. 246–253.

[3] Z. Zhu and Q. Ji, "Eye and gaze tracking for interactive graphic display," *Mach. Vis. Appl.*, vol. 15, no. 3, pp. 139–148, Jul. 2004.

[4] S. P. Liversedge and J. M. Findlay, "Saccadic eye movements and cognition," *Trends Cognit. Sci.*, vol. 4, no. 1, pp. 6–14, 2000.

[5] K.-H. Tan, D. Kriegman, and H. Ahuja, "Appearance-based eye gaze estimation," in *Proc. 16th IEEE Workshop Appl. Comput. Vis.*, Dec. 2002, pp. 191–195.

[6] E. D. Guestrin and M. Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 6, pp. 1124–1133, Jun. 2006.

[7] E. D. Guestrin and M. Eizenman, "Remote point-of-gaze estimation requiring a single-point calibration for applications with infants," in *Proc. Symp. Eye Tracking Res. Appl.*, 2008, pp. 267–274.

[8] C. H. Morimoto and M. R. M. Mimica, "Eye gaze tracking techniques for interactive applications," *Comput. Vis. Image Understand.*, vol. 98, no. 1, pp. 4–24, Apr. 2005.

[9] S.-W. Shih and J. Liu, "A novel approach to 3-D gaze tracking using stereo cameras," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 1, pp. 234–245, Feb. 2004.

[10] D. Beymer and M. Flickner, "Eye gaze tracking using an active stereo head," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2003, pp. II-451–II-458.

[11] J. Chen, Y. Tong, W. Gray, and Q. Ji, "A robust 3D eye gaze tracking system using noise reduction," in *Proc. Symp. Eye Tracking Res. Appl.*, 2008, pp. 189–196.

[12] J. Chen and Q. Ji, "3D gaze estimation with a single camera without IR illumination," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[13] Y. Sugano, Y. Matsushita, and Y. Sato, "Calibration-free gaze sensing using saliency maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* Jun. 2010, pp. 2667–2674.

[14] Y. Sugano, Y. Matsushita, and Y. Sato, "Appearance-based gaze estimation using visual saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 329–341, Feb. 2013.

[15] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 478–500, Mar. 2010.

[16] F. Lu, T. Okabe, Y. Sugano, and Y. Sato, "A head pose-free approach for appearance-based gaze estimation," in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 126.1–126.11.

[17] M. Eizenman *et al.*, "A naturalistic visual scanning approach to assess selective attention in major depressive disorder," *Psychiatry Res.*, vol. 118, no. 2, pp. 117–128, May 2003.

[18] E. Horvitz, C. M. Kadie, T. Paek, and D. Hovel, "Models of attention in computing and communication: From principles to applications," *Commun. ACM*, vol. 46, no. 3, pp. 52–59, 2003.

[19] SensoMotoric Instrum. *SMI Eye Tracking Glasses*. [Online]. Available: http://www.smivision.com, accessed, Dec. 15, 2010.

[20] D. Model and M. Eizenman, "An automatic personal calibration procedure for advanced gaze estimation systems," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 5, pp. 1031–1039, May 2010.

[21] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2006, pp. 545–552.

[22] J. Chen and Q. Ji, "Probabilistic gaze estimation without active personal calibration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* Jun. 2011, pp. 609–616.

[23] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike, "An incremental learning method for unconstrained gaze estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2007, pp. 656–667.

[24] C. W. Oyster, *The Human Eye: Structure and Function*. Sunderland, MA, USA: Sinauer Associate Inc., 1999.

[25] A. L. Yarbus, B. Haigh, and L. A. Rigss, *Eye Movements and Vision*, vol. 2. New York, NY, USA: Plenum, 1967.

[26] A. Borji and L. Itti, "Defending Yarbus: Eye movements reveal observers' task," *J. Vis.*, vol. 14, no. 3, p. 29, Mar. 2014.

[27] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. MIT-CSAIL-TR-2012-001, 2012.

[28] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sep./Oct. 2009, pp. 2106–2113.

[29] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.

[30] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[31] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. Vis.*, vol. 11, no. 3, pp. 1–15, Mar. 2011.

[32] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard, "Eye guidance in natural vision: Reinterpreting salience," *J. Vis.*, vol. 11, no. 5, pp. 1–23, May 2011.

[33] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.

[34] A. Slater and J. M. Findlay, "The measurement of fixation position in the newborn baby," *J. Experim. Child Psychol.*, vol. 14, no. 3, pp. 349–364, Dec. 1972.

**Jixu Chen** (M'11) received the Ph.D. degree in electrical engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 2011. He is currently a Researcher with the Computer Vision Laboratory, GE Global Research, Niskayuna, NY, USA. His research interests include computer vision, machine learning, and human–computer interaction. He is a member of the IEEE Computer Society.

**Qiang Ji** (F'15) received the Ph.D. degree in electrical engineering from the University of Washington, Seattle, WA, USA. He is currently a Professor with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA. He served as the Program Director with the National Science Foundation (NSF), where he managed NSF's computer vision and machine learning programs. He held teaching and research positions with the Beckman Institute, University of Illinois at Urbana—Champaign, Champaign, IL, USA, the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, the Department of Computer Science, University of Nevada at Reno, Reno, NV, USA, and the U.S. Air Force Research Laboratory, Dayton, OH, USA. His research interests are in computer vision, probabilistic graphical models, and their applications in various fields. He has authored over 200 papers in peer-reviewed journals and conferences. His research has been supported by major governmental agencies, including NSF, NIH, DARPA, ONR, ARO, and AFOSR and major companies. He is an Editor of several related IEEE and international journals, and has served as the General Chair, the Program Chair, the Technical Area Chair, and a Program Committee Member on numerous international conferences/workshops. He is a fellow of the International Association for Pattern Recognition.