Deep eye fixation map learning for calibration-free eye gaze tracking

Kang Wang*
Rensselaer Polytechnic Institute

Shen Wang†
University of Illinois at Chicago

Qiang Ji‡
Rensselaer Polytechnic Institute

Abstract

The existing eye trackers typically require an explicit personal calibration procedure to estimate subject-dependent eye parameters. Despite efforts in simplifying the calibration process, such a calibration process remains unnatural and bothersome, in particular for users of personal and mobile devices. To alleviate this problem, we introduce a technique that can eliminate explicit personal calibration. Based on combining a new calibration procedure with the eye fixation prediction, the proposed method performs implicit personal calibration without active participation or even knowledge of the user. Specifically, different from traditional deterministic calibration procedure that minimizes the differences between the predicted eye gazes and the actual eye gazes, we introduce a stochastic calibration procedure that minimizes the differences between the probability distribution of the predicted eye gaze and the distribution of the actual eye gaze. Furthermore, instead of using saliency map to approximate eye fixation distribution, we propose to use a regression based deep convolutional neural network (RCNN) that specifically learns image features to predict eye fixation. By combining the distribution based calibration with the deep fixation prediction procedure, personal eye parameters can be estimated without explicit user collaboration. We apply the proposed method to both 2D regression-based and 3D model-based eye gaze tracking methods. Experimental results show that the proposed method outperforms other implicit calibration methods and achieve comparable results to those that use traditional explicit calibration methods.

Keywords: calibration-free, eye gaze, saliency

Concepts: •**Computing methodologies** → **Computer vision; Machine learning;**

1 Introduction

Humans explore outside world mainly through eye. Eye gaze therefore serves an important role in this process. Gaze estimation is to predict eye gaze, mainly point-of-regard in the space or the visual axis. A typical application for eye gaze tracking is Human-Computer-Interaction (HCI). For instance, eye gaze can replace traditional input (mouse pointer) to control the computer, or serve as an additional input in games to improve user experience. Besides, since eye gaze reflects human attention, cognitive scientists use gaze tracking systems to study human's cognitive processes [Mason et al. 2004]. Other applications include marketing, advertising, medical research, etc.

*e-mail: wangk10@rpi.edu
†e-mail: swang224@uic.edu
‡e-mail: qji@ecse.rpi.edu

ETRA 16, March 14-17, 2016, Charleston, SC, USA
ISBN: 978-1-4503-4125-7/16/03 $15.00
DOI: http://dx.doi.org/10.1145/2857491.2857515

Eye gaze tracking technologies can be divided into feature-based and appearance-based approach. Feature-based methods typically rely on specific eye features like pupil and cornea reflection (glint). Feature-based approach can be further divided into 2D regression-based approach and 3D model-based approach. The 2D regression-based methods [Morimoto and Mimica 2005; Zhu and Ji 2004; Zhu et al. 2006] implicitly estimate gaze point on the display surface by mapping the vector that connects image position of the pupil with that of the cornea reflection to a gaze position on the display surface. The approach is simple but it typically does not allow much head movement. The 3D model-based approaches [Beymer and Flickner 2003; Chen et al. 2008; Guestrin and Eizenman 2008; Shih and Liu 2004; Guestrin and Eizenman 2006], on the other hand, directly estimate visual axis from the eye features in the image. It firstly builds a 3D geometric eye model and then estimates 3D eye features (pupil center, cornea center, etc.). Based on eye model and eye features, it computes the 3D gaze direction (optical axis, visual axis). Model-based method is known for its accuracy and ability to handle free head movement. But it is complex in calibration and requires a complex setup including multiple cameras or multiple IR lights. The appearance-based methods [Lu et al. 2011; Tan et al. 2002; Williams et al. 2006] perform eye gaze estimation based on the eye appearance without explicit detection of any eye features. The key idea is that similar eye appearance results in similar gaze position, based on which it constructs a regression function that maps the eye appearance to gaze position on the display surface. Compared to the feature-based method, appearance-based method does not need IR illumination. But it cannot effectively handle head pose movement and illumination variation. For a comprehensive review of existing eye gaze tracking methods, readers should refer to [Hansen and Ji 2010].

All these methods require to perform personal calibration to estimate the personal eye parameters. Calibration process is usually cumbersome, unnatural and degrades user experience. In this paper, we propose a method that combines a new calibration method with the eye fixation map learnt from deep learning model. Called distribution based calibration, the proposed method performs calibration by minimizing the distribution differences between a target gaze distribution and the estimated gaze distribution. It avoids constructing the explicit correspondences between the groundtruth gazes and the estimated gazes. To obtain a good estimate of the fixation distribution on a query image, we propose to use a regression based deep convolutional neural network (RCNN) to learn image features that are specifically used to predict eye fixations. To differ from latter fixation map/distribution predicted from eye gaze estimation method, the fixation map predicted by learned image features is denoted as deep fixation map/distribution. By combining the top-down eye fixation prediction with bottom up gaze estimation through the proposed distribution based calibration procedure, we introduce a new method that enables the existing eye tracking methods to perform eye gaze tracking without explicit personal calibration.

Compared to the existing work, the proposed work makes the following novel contributions:

- introduce a unified framework for feature learning and regression in a deep convolutional style for patch based fixation detection.

- introduce a distribution based calibration that does not require knowledge of the groundtruth gazes and their correspondences with the estimated gazes.
- introduce a method to combine the distribution based calibration with the deep fixation map to perform implicit eye personal calibration and apply the method to two main approaches for eye gaze tracking.

2 Related Work

2.1 Eye gaze tracking

Eliminating or reducing personal calibration is an active area of research. Much work has been done in this area. For model-based eye gaze tracking, Chen *et al.* [2008] proposed a system with two cameras and two IR lights. Their system starts with the reconstruction of the optical axis. Visual axis is then computed by adding the constant angle to the optical axis. However, the system requires a 4-point calibration to estimate the constant angle. Guestrin *et al.* [2008] introduced a two cameras and four IR lights system. By using two cameras instead of one camera, their method only requires a 1-point personal calibration. Model and Eizenman [2010] proposed to automatically estimate eye parameters using both eyes by exploiting the binocular constraint. The underlying assumption is the visual axis of both eyes ideally intersect at the same gaze position on the display surface. However, the proposed method is sensitive to noise and thus requires that the PoRs span a wide range of gaze directions (e.g., looking at larger areas of the display surface). Therefore the proposed system cannot produce accurate results with ordinary displays. Maio *et al.* [2011] proposed to alleviate the noise problem by introducing additional generic person-independent constraints to the binocular-based gaze estimation framework. Chen [2011] proposed an implicit calibration method with saliency map. They build a Bayesian network to represent the probabilistic relationship among optical axis, visual axis and eye parameters. Saliency map served as gaze prior in the probabilistic model. Eye parameters estimation is formulated as an inference problem by computing the gaze posterior given the image and the optical axis. They further proposed a dynamic bayesian network to on-line incrementally update the eye parameters. However, their saliency map is from shallow bottom-up estimation. It is difficult for such saliency maps to capture semantic information that attracts more human attention. Thus the saliency map may not well represent human fixation. Later Chen [2014] extended their method which can work with Gaussian distribution. However, the method requires large amount of frames so that the gaze points distribution can be approximated as Gaussian distribution.

For appearance based eye gaze tracking, Sugano *et al.* [2007] proposed an implicit calibration method relying on mouse clicks. Their method assumes that when subjects click the mouse, they would unconsciously look at the cursor. Therefore they can implicitly collect the gaze positions when mouse click happens. Later, Sugano *et al.* [2013] introduced a visual saliency based gaze estimation method without explicit personal calibration. Their system extracts eye image and saliency map pair simultaneously while subjects watch a video clip. Saliency maps with similar eye appearance are aggregated to produce a probability map with a vivid peak around the gaze point. A Gaussian process regression is then built upon the probability maps and the eye images. Gaze estimation for a new eye image can be performed using parameters learned from the regression model. Like Chen's [2011] method, their saliency maps are from shallow bottom-up estimation which may not correlate well with human fixation. Alnajar *et al.* [2013] proposed a calibration-free gaze estimation method with the help of human gaze pattern. They assume that different subjects tend to have similar gaze patterns on the same stimuli. Therefore, model parameters for a new subject can be effectively estimated by mapping his/her initial gaze pattern to the off-line collected gaze patterns from other subjects. However, the underlying assumption remains too strong and unrealistic. Different types, content of the stimuli may results in different gaze patterns for different subjects, therefore the proposed method may be limited in real world applications. Besides, compared to model-based methods, appearance-based methods cannot handle free head movement well and thus cannot produce accurate gaze estimation.

2.2 Saliency and fixation estimation

Visual attention is a general procedure about the idea of a selection mechanism and a notation of relevance, which can be data-driven bottom-up or top-down depending on expectation or identification of objects. The foundation of most attention models is the Feature Integration Theory by [Treisman and Gelade 1980], which concentrates on the categories of visual features and the way they are combined. Based on this concept, the saliency is well investigated in the context of bottom-up computations [Itti et al. 1998; Harel et al. 2006b]. These works use low-level features such as intensity, color contrast and orientation to generate models for saliency map prediction. But saliency models based on low level features has shown its limitation to capture semantic information. Recently, some works start focusing on predicting the eye fixation. They borrow the ideas in saliency detection and apply it to eye tracking data, resulting a new research area: fixation detection. Kienzle *et al.* [2009] and Judd *et al.*[2009] trained a linear SVM directly from the human eye tracking data. The former one used the low-level features to generate the fixation map while the latter used a combination of low, middle and high level features and achieved a good performance. In addition, Judd *et al.*[2009] employed the face and human detectors as the high level features, which can bridge the semantic gap.

Recently, much work has been done on designing good features for fixation detection. Most of them are based on hand-crafted low level image features which are not specifically developed to characterize eye gazes. The ability of predicting eye fixations are hence limited. To address this issue, recent works leverage on deep learning's powerful learning and representation capability to learn image features that correlate well with eye fixations, like objects, human/faces, contexts, etc. Among them, [Lin et al. 2014; Shen and Zhao 2014; Vig et al. 2014] train a convolutional neural network (CNN) directly from eye tracking data to image features to predict eye fixation. Different convolutional layers are treated independently in their network. In [Kümmerer et al. 2014], the authors also rely on the eye tracking data, but use the pre-trained Alex Net [Krizhevsky et al. 2012] as starting point and fine-tune the network by maximizing the log-likelihood of the point process. All of these CNN based methods except for [Krizhevsky et al. 2012] formulate the fixation detection as a supervised patch classification problem and describe the fixation value with binary label. This kind of fixation description ignores the probabilistic nature of the fixation and the uncertainties associated with each fixation. Besides, all the above methods employ center prior (a strong bias for human fixations to be near the center of the image) through post-processing but cannot integrate center prior during the feature learning process. In this paper, we propose to extend traditional classification based CNN to regression based to treat the fixation value as a continuous probability value. Furthermore, we directly integrate center prior as regularization term during the feature learning process.

3 Eye Fixation Map Estimation using RCNN

In this section, we introduce the Regression Convolutional Neural Networks (RCNN) to learn center-biased fixation-discriminative

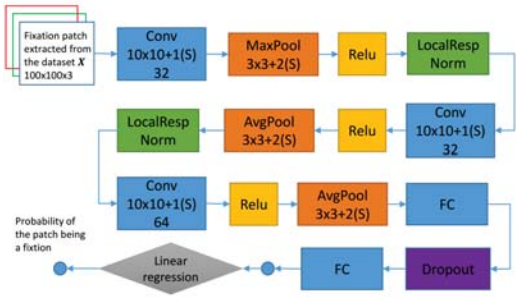


Figure 1: *The architecture of RCNN. The blue blocks represent the convolutional layer. The orange blocks represent the polling layer. The yellow blocks represent the non-linear gating layer. The green blocks represent the local response normalization layer. The purple block represents the dropout layer and the grey block represents the loss function.*

feature for eye fixation prediction. Fixation prediction is usually formulated as a classification problem with a specified feature representation of the image region $\mathbf{X}$ and a function $f(\mathbf{X};\mathbf{W})$ using learned parameters $\mathbf{W}$ to map $\mathbf{X}$ to a binary fixation value $y \in \{0,1\}$. As discussed in the last section, this kind of fixation description is not informative for gaze prediction. It ignores the uncertainty in gaze fixations. To overcome its limitation, we consider the fixation value as a probability in a soft way and use linear regression to estimate it.

To capture semantic representation, we integrate the CNN and linear regression into a unified framework. To mimick the structure of ventral stream in visual cortex, we build the network in a deep style. The architecture of the proposed RCNN is illustrated in Figure 1. It consists of two stages: convolutional stage and full connection stage. Convolutional stage includes 3 convolutional blocks. A basic convolutional block consists of a linear convolution layer, a nonlinear gating layer, a spatial pooling layer and a local feature normalization layer. The full connection stage contains fully connected layer followed by dropout layer. The input of the network is an image patch $\mathbf{X} \in R^{100\times100\times3}$. Considering the whole network as a function and $\mathbf{W}$ as all network parameters, the output of the network $y = f(\mathbf{X};\mathbf{W}), y \in R)$ represents the probability of the corresponding patch being a fixation patch. In addition, we propose to incorporate center prior to the learning framework instead of applying center prior at post-processing stage. Besides minimizing the distance between predicted fixation probability and groundtruth fixation probability, we add a regularization term to incorporate the center prior. Therefore the loss function is defined as:

$$L = \frac{1}{2N}\sum_{i=1}^{N}[(f(\mathbf{X}_i;\mathbf{W}) - y_i^{prob})^2 + \lambda(f(\mathbf{X}_i;\mathbf{W}) - cp_i^{prob})^2]$$
$$cp_i^{prob} = \frac{1}{\sqrt{4\pi^2|\Sigma|}}\exp(-\frac{1}{2}(\mathbf{x}_i-\mu)^T\Sigma^{-1}(\mathbf{x}_i-\mu)) \quad (1)$$

where i represents the ith patch and N represents the total number of patches. $f(\mathbf{X}_i;\mathbf{W})$ and y_i^{prob} represent respectively the predicted and ground truth fixation probabilities. y_i^{prob} is obtained by applying softmax function to the groundtruth fixation map built from eye tracking data. Center prior is approximated by a Gaussian distribution $\mathcal{N}(\mu,\Sigma)$. μ is set to the coordinates of image center and Σ is empirically determined. $\mathbf{x}_i$ represents the center coordinates of patch i in the whole image. λ is the parameter to balance the first term and second term.

The training process of RCNN network consists of two stages, prediction and updating. During prediction stage, we feed forward the network and obtain the predicted fixation probability of the input patch. Given an image patch, the convolutional stage with three convolutional blocks is first performed. In a convolutional block, the input is first convolved with a filter bank and results in a feature map. The resulted feature map is then gated by a nonlinear function, which introduces the non-linearity. In this step, sigmoid, hyperbolic tangent, rectified linear unit or its variation are usually used as the nonlinear gating functions. After that the gated feature map is down-sampled into a more compact representation, which introduces translation invariance, noise robustness and improved generalization. Spatial pooling is used for down-sampling in the process. It captures local information of each pixel location by integrating the response from nearby locations. Common spatial pooling includes max, average and min operations. After convolutional stage, two full connection layers are applied to project feature map to the probability space. The dropout is implemented to prevent over-fitting between these two layers. During updating stage, the network parameters (weight and bias) are updated across layers by loss back-propagation. During the back-propagation, the network parameters are updated layer by layer by stochastic gradient decent.

To generate the fixation map of a new image, candidate image patches are extracted sequentially, then their fixation probabilities are computed through the proposed RCNN framework.

4 Distribution-based Calibration

Traditional calibration procedure typically includes a set of training data $\{g_i, \hat{g}_i(\theta)\}$, i=1,2,..,N, where g_i is a groundtruth gaze point and $\hat{g}_i$ is the corresponding estimated gaze point. Gaze calibration involves estimating the calibration parameters θ by minimizing the total prediction errors, i.e.,

$$\theta^* = \arg\min_{\theta}\sum_{i=1}^{N}(g_i - \hat{g}_i(\theta))^2 \quad (2)$$

It is clear that traditional gaze calibration requires the knowledge of groundtruth eye gaze points as well as the correspondences between the groundtruth and estimated gaze points. This means traditional calibration requires explicit collaboration from the user. To remove the explicit collaboration, we introduce a stochastic approach that estimates the eye parameters by minimizing the distributions of eye gazes. For an input image I, let $p_f(x,y|I)$ represent the deep fixation distribution estimated by RCNN and $p_g(x,y|I,\theta)$ be the estimated gaze distribution from gaze estimation on the same image. The goal of distribution based calibration is to find the parameters θ to minimize the differences between $p_f(x,y|I)$ and $p_g(x,y|I,\theta)$. Mathematically, this can be formulated as minimization of the Kullback-Leibler (KL) divergence [Kullback and Leibler 1951]:

$$\theta^* = \arg\min_{\theta} D_{KL}(p_f(x,y|I)||p_g(x,y|I,\theta)) \quad (3)$$

KL divergence between two distributions $p(x,y)$ and $q(x,y)$ is defined as:

$$\begin{aligned} &D_{KL}(p(x,y)||q(x,y)) \\ &= \sum_{x,y} p(x,y)\log(\frac{p(x,y)}{q(x,y)}) \\ &= \sum_{x,y} p(x,y)\log(p(x,y)) - \sum_{x,y} p(x,y)\log(q(x,y)) \end{aligned} \quad (4)$$

In practice, we use the symmetric divergence in order to make it a real metric. Thus we solve the following optimization problem:

$$\theta^* = \arg\min_{\theta} D_{KL}(p_f(x,y|I)||p_g(x,y|I,\theta)) + D_{KL}(p_g(x,y|I,\theta)||p_f(x,y|I)) \quad (5)$$

The optimization problem is solved through gradient-based interior-point algorithm.

5 Eye Gaze Estimation

For this research, we propose to apply the proposed deep eye fixation map and distribution based calibration to both 2D regression-based and 3D model-based eye gaze estimation methods. Below we briefly summarize each method and identify the parameters to estimate.

5.1 2D Regression-based eye gaze estimation

5.1.1 Regression-based eye gaze estimation

Regression-based methods learn a mapping between 2D feature vectors and gaze positions on the display surface. Standard features are the pupil glint vector $(\delta x, \delta y)$ in image coordinates. During training, we collect a set of training data $\{(\delta x_i, \delta y_i), (x_i, y_i)\}$, i=1,2,..,N, where $(\delta x_i, \delta y_i)$ is a feature vector, and (x_i, y_i) is the x and y coordinate of corresponding gaze positions on the display surface. We then learn a linear regression between them:

$$x_i = a_x \delta x_i + b_x \delta y_i + c_x \quad (6)$$

$$y_i = a_y \delta x_i + b_y \delta y_i + c_y \quad (7)$$

where $\theta = (a_x, b_x, c_x, a_y, b_y, c_y)$ are the calibration parameters. After obtaining the calibration parameters, we can map any testing feature vector to its gaze position.

5.1.2 Implicit calibration for the regression method

For implicit calibration, we do not know the exact gaze positions. Instead we only have a set of pupil-glint vectors $\{\delta x_i, \delta y_i\}$, i=1,2,..,N, plus the deep eye fixation map $p_f(x,y|I)$. Our goal is to learn the calibration parameters θ through the proposed distribution based calibration method. We start with the estimation of $p_g(x,y|I,\theta)$. Mixtures of Gaussian are used to model $p_g(x,y|I,\theta)$, where each mixture is centered at each estimated gaze position:

$$p_g(x,y|I,\theta) = \sum_{i=1}^{N} w_i \mathcal{N}([x_i, y_i], \Sigma) \quad (8)$$

where gaze position x_i and y_i are obtained by Eqn. 6 and 7, Σ represents the variance of the Gaussian mixture and it is determined empirically. w_i represents the uncertainty/importance of each gaze point. Thus we set w_i of gaze point i according to its distance to image center to incorporate the center prior. Plugging deep eye fixation map $p_f(x,y|I)$ and estimated eye fixation map $p_g(x,y|I,\theta)$ into Eqn. 5, we can then solve the optimization problem to obtain the calibration parameters θ implicitly without any user collaboration.

5.2 3D model eye gaze estimation

5.2.1 Model-based eye gaze estimation

3D model-based method is based on the 3D geometric eye model as shown in Figure 2. Optical axis is defined as the line that connects the cornea center and pupil center. However, the real gaze direction is determined by the visual axis, which passes through cornea center and fovea. PoR is defined as the intersection of visual axis and the display surface. Optical axis and the corresponding visual axis differ by fixed angle kappa. We use $\theta = (\alpha, \beta)$ to express the angle difference between optical and visual axis. θ is a constant vector for each person. Personal calibration for the model-based methods amounts to estimate θ for each person.

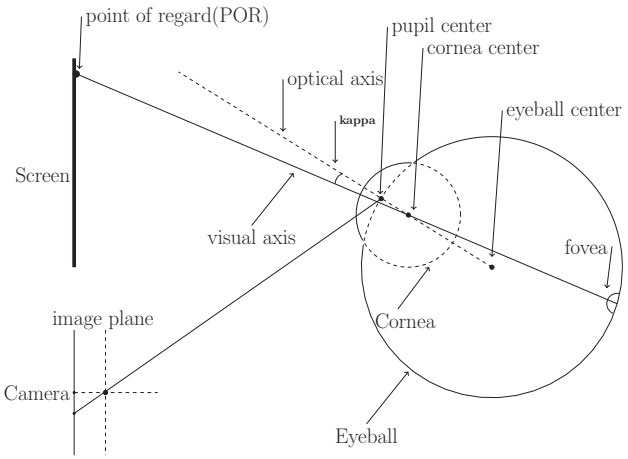


Figure 2: *3D eye model.*

Typical 3D model-based systems involve multiple cameras and IR lights as described in [Guestrin and Eizenman 2006]. The system allows to obtain the 3D coordinates of cornea center $\mathbf{c}$ and pupil center $\mathbf{p}$, from which we can estimate the 3D optical axis $\mathbf{v}_o$:

$$\mathbf{v}_o = (\mathbf{p} - \mathbf{c})/||\mathbf{p} - \mathbf{c}|| \quad (9)$$

$\mathbf{v}_o$ is a unit length vector and thus can be expressed as two angles ϕ and γ:

$$\mathbf{v}_o = \begin{pmatrix} \cos(\phi)\sin(\gamma) \\ \sin(\phi) \\ -\cos(\phi)\cos(\gamma) \end{pmatrix} \quad (10)$$

Visual axis $\mathbf{v}_g$ can be computed by adding $\theta = (\alpha, \beta)$ to optical axis:

$$\mathbf{v}_g = g(\mathbf{v}_o; \theta) = \begin{pmatrix} \cos(\phi+\alpha)\sin(\gamma+\beta) \\ \sin(\phi+\alpha) \\ -\cos(\phi+\alpha)\cos(\gamma+\beta) \end{pmatrix} \quad (11)$$

where $g(\cdot)$ is the function to compute visual axis from optical axis by adding the constant angle θ. To estimate calibration parameters $\theta = [\alpha, \beta]$, subjects are asked to look at pre-defined points (PoR) on the display surface. We can collect a set of training data $\{\mathbf{c}_i, \mathbf{p}_i, \mathbf{g}_i\}$, i=1,2,..,N, where $\mathbf{c}_i$ is the 3D coordinates of cornea, $\mathbf{p}_i$ is the 3D coordinates of pupil and $\mathbf{g}_i$ is the 3D coordinates of PoR. From which we can obtain N optical and visual axis:

$$\mathbf{v}_{oi} = (\mathbf{p}_i - \mathbf{c}_i)/||\mathbf{p}_i - \mathbf{c}_i|| \quad (12)$$

$$\mathbf{v}_{gi} = (\mathbf{g}_i - \mathbf{c}_i)/||\mathbf{g}_i - \mathbf{c}_i|| \quad (13)$$

Finally, θ can be estimated by

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} g^{-1}(\mathbf{v}_{oi}, \mathbf{v}_{gi}) \quad (14)$$

5.2.2 Implicit calibration for model-based method

For the implicit calibration method, we can only collect a set of samples $\{\mathbf{c}_i, \phi_i, \gamma_i\}, i = 1, 2, ..., N$ (ϕ_i and γ_i are the two angles representing optical axis), and the eye fixation map $p_f(x, y|I)$. Similar to 2D implicit calibration, we first estimate $p_g(x, y|I, \theta)$ by building a mixture of Gaussian based on the estimated PoRs.

The 3D PoR $[x, y, z]^T$ on the display surface satisfy the surface equation $f(x, y, z) = 0$, which can be estimated by a one time offline display-camera calibration. Without loss of generality, we assume the display surface is a plane and satisfy:

$$f(x, y, z) = z = 0 \tag{15}$$

3D PoR also lie on the visual axis:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \mathbf{c} + \lambda \mathbf{v}_g \tag{16}$$

Solving the line-plane intersection problem in Eqns. 15, 16 and combining Eqns. 9, 10 and 11, we can obtain the 3D PoRs from all the observations $\{\mathbf{c}_i, \phi_i, \gamma_i\}, i = 1, 2, ..., N$, given the subject-dependent eye parameters $\theta = [\alpha, \beta]$:

$$\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = \begin{pmatrix} \mathbf{c}_i[1] - \mathbf{c}_i[3] \tan(\gamma_i + \beta) \\ \mathbf{c}_i[2] - \mathbf{c}_i[3] \frac{\tan(\phi_i + \alpha)}{\cos(\gamma_i + \beta)} \\ 0 \end{pmatrix} \tag{17}$$

where $\mathbf{c}[i]$ represents the ith element of cornea center $\mathbf{c}$.

Gaze distribution $p_g(x, y|I, \theta)$ can therefore be computed as:

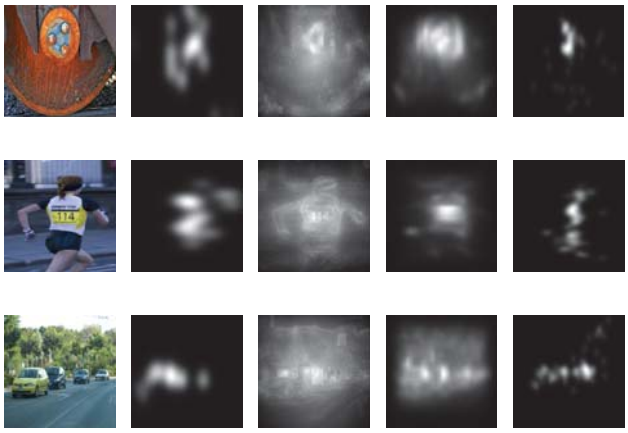
$$p_g(x, y|I, \theta) = \sum_{j=1}^{N} w_i \mathcal{N}([x_i, y_i], \Sigma) \tag{18}$$

Similarly, plugging $p_f(x, y|I)$ and $p_g(x, y|I, \theta)$ into Eqn. 5, we can solve the optimization problem to estimate the parameters.

6 Experimental Results

We briefly introduce the experimental settings. The dimension of the display is 1280×1024. 10 images from MIT dataset [Judd et al. 2009] are chosen as the displaying image during experiments. Images are from outdoors, like houses, street with cars/humans, sea with boats, etc. Subjects are asked to watch the 10 images sequentially with natural head movement. Each image is displayed for 4 seconds, and we display a black screen for 1 second between consecutive images. The first few measurements are ignored to eliminate the initial saccade eye movements. Five male and one female subjects whose age range from 22 to 30 years old participate into the experiments. To acquire eye measurements, we implement the 1 camera and 2 IR lights system as described in [Guestrin and Eizenman 2006].

We perform several experiments to evaluate the performance of the proposed method. Firstly, we visualize fixation/saliency maps from the proposed algorithm and other algorithms in section 6.1. In section 6.2, we compare the calibration parameters from implicit and explicit personal calibration methods. Since this paper proposes two novel techniques: the deep fixation map prediction algorithm and distribution-based personal calibration algorithm. Thus we evaluate these two components separately as in section 6.3 and section 6.4. Finally, we compare the proposed method with state-of-the-art methods in reducing/eliminating explicit personal calibration in section 6.5.



(a) Stimuli (b) RCNN (c) Judd (d) GBVS (e) GT

Figure 3: *Examples of predicted saliency/fixation maps from RCNN, GBVS, Judd and groundtruth.*

The gaze estimation error in section 6.3, 6.4 and 6.5 is computed as follows. Subjects are asked to look at 15 uniform-distributed points on the display surface. Data are collected and these points served as the groundtruth gaze positions. We then compute the estimated gaze positions using calibration parameters estimated from different calibration scenarios. Gaze estimation error is measured by the distance error between the estimated gaze positions and the groundtruth positions.

6.1 Fixation Map Visualization

MIT1003 dataset[Judd et al. 2009] is selected to learn the deep fixation map. The dataset includes 1003 landscape and portrait images, which contain rich objects like people, cars, faces, etc. Softmax function is applied to transform the groundtruth fixation map in the dataset into probability map. In the experiments we use patches extracted from first 500 images to train our RCNN network. We select the testing image during calibration from the rest 503 images. We visualize the saliency/fixation maps predicted with the RCNN framework and other algorithms in Figure 3. The GBVS is totally bottom up method with only low level features. The Judd's method combined the top-down and bottom-up procedure by feature integration. Qualitatively, the fixation map from RCNN is focused and centered at most salient objects. In comparison, Judd's and GBVS's saliency maps spread to the whole image and thus may degrade the personal calibration framework. To better quantitatively prove the effectiveness of the proposed RCNN framework, we compare them in terms of gaze estimation accuracy as in section 6.3.

6.2 Comparison of Calibration Parameters

In this section, we compare calibration parameters from implicit and explicit calibration methods. For 2D regression-based method, we implement the traditional explicit method with linear regression. The average error of this system is about 0.6 degree. For 3D model-based method, we implement the explicit 9-points calibration based gaze estimation system. The error of this system is averagely 1 degree for different subjects. As we can observe from Table 1, for 2D regression-based method, most of the estimated parameters are

Table 1: *Comparison of estimated calibration parameters from traditional and proposed methods for 2D regression-based and 3D model-based methods.*

Subjects	Method	2D: $\theta = (a_x, b_x, c_x, a_y, b_y, c_y)$						3D: $\theta = (\alpha, \beta)$	
1	explicit	−11.50	−0.09	−17.40	−2.40	16.70	188.00	0.29	1.81
	proposed	−10.10	−1.50	−23.00	−2.40	11.60	169.00	1.12	0.41
2	explicit	0.04	−0.02	110.00	0.90	5.70	155.00	0.99	0.31
	proposed	0.06	−0.02	102.00	1.20	8.60	171.00	1.93	−0.25
3	explicit	−2.10	−7.00	43.00	−6.70	6.70	85.00	−1.31	1.35
	proposed	−1.30	−8.50	30.00	−8.30	8.90	72.00	−2.04	0.23
4	explicit	4.30	−2.10	79.00	−0.70	1.20	185.00	−0.74	3.76
	proposed	3.20	−4.50	53.00	−0.30	2.70	174.00	−1.24	3.03
5	explicit	−2.70	0.30	91.00	6.50	−8.70	114.00	2.70	−0.80
	proposed	−4.80	0.10	62.00	9.30	−11.90	111.00	1.94	1.08
6	explicit	−2.10	−1.40	80.50	1.80	−7.00	72.00	−3.56	−0.81
	proposed	−3.70	−2.50	49.20	2.30	−8.90	71.00	−2.32	−2.85

close to that from explicit methods for different subjects. There are a few parameters differing away from explicit method. However, the overall parameter configuration can still achieve good gaze estimation performance as illustrated in following sections. For 3D model-based method, the estimated parameters are both close to that from traditional explicit method, thus we can achieve good performance in terms of gaze estimation accuracy.

6.3 Evaluation of Deep Fixation Maps

In this section, we compare the deep fixation map with other saliency maps in terms of gaze estimation accuracy. We consider 4 different saliency/fixation maps, namely: pure center prior, saliency map generated by GBVS algorithm [Harel et al. 2006a], fixation map from Judd's model [Judd et al. 2009] and the proposed deep fixation map.

For 2D regression-based method, we firstly implement the explicit 9-points calibration method as baseline. Then we implement the distribution-based calibration method with different saliency/fixation maps. In the experiments, subjects are required to fix their head position. The average error for six subjects with the five calibration scenarios are showed in Figure 4 (a). We can see that the 9-points method achieve the best results, the average error is about 0.67 degree. For the implicit-calibration method, GBVS, Judd and the proposed fixation map give much better results than pure center prior. The average error for the proposed method is 0.99 degree, which is slighter better than GBVS (1.03 degree) and Judd (1.04 degree). Overall, the implicit calibration framework is suitable for 2D regression-based gaze estimation methods, and we can achieve averagely 1.0 degree error. Even it is larger than traditional 9-points method, the proposed method can be totally implicit without user's explicit collaboration.

Similar to 2D regression-base method, for 3D model-based method, we also implement the explicit 9-points calibration method and the proposed method with different saliency/fixation maps. Results are showed in Figure 4 (b). In this experiments, subjects can move their head naturally within the system's working range. The 9-points calibration gives the smallest average error of 1.08 degree. The average error for center prior, GBVS, Judd and the proposed fixation map are 1.66 degree, 1.54 degree, 1.56 degree and 1.40 degree respectively. Implicit calibration framework can still give us comparable results. Because of free head movement in the experiments, the variance of 3D model-based method is larger than 2D regression-based method. In practice, the large variance issue can be alleviated by proper temporal smoothing operations. With the distribution-based calibration and the learned deep fixation map, we can achieve 1.4 degree error in gaze estimation.

Table 2: *Comparison of distribution based calibration with Chen's [2011] full probabilistic approach.*

Method	Frames/Error	Frames/Error	Frames/Error
Full probabilistic	100 / 1.71°	200 / 1.49°	300 / 1.43°
Distribution-based	100 / 1.52°	200 / 1.41°	300 / 1.42°

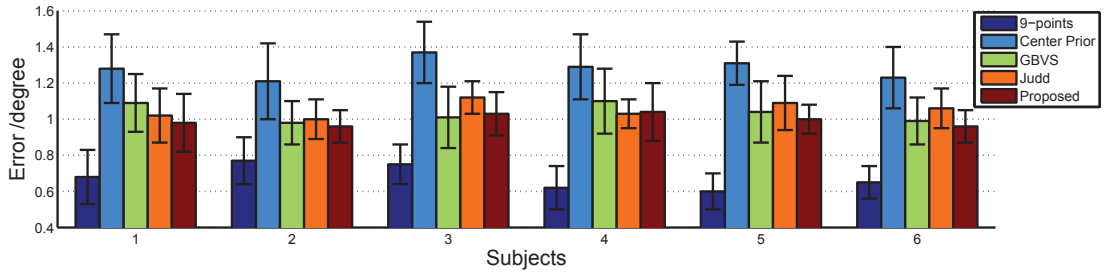
The results in Figure 4 proves that with the same distribution-based personal calibration method, the proposed deep fixation prediction algorithms outperforms other saliency/fixation prediction algorithms.

6.4 Evaluation of Distribution-based Calibration

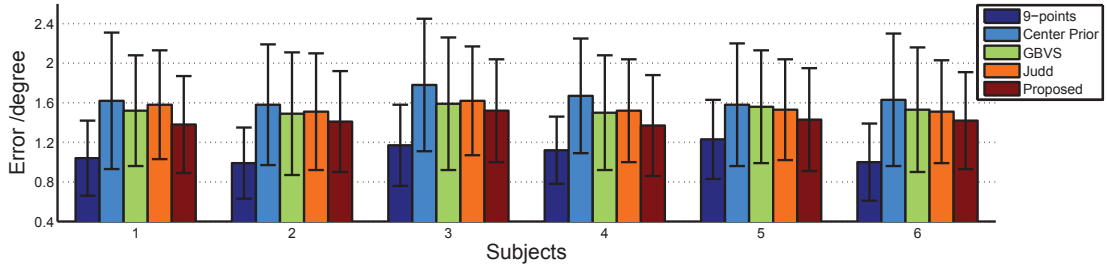
To evaluate the proposed distribution-based personal calibration, we implement Chen *et al*'s full-probabilistic method [2011] with the same deep fixation maps from RCNN framework. Since Chen *et al* proposed an incremental calibration framework to keep updating the eye parameters. Thus we evaluate the two calibration methods with different number of input frames. The gaze estimation error is shown in Table 2. Our method hence achieves better results than Chen's method. Besides, the proposed method works with fewer data since the proposed method can align partial saliency map built from gaze data to the predicted saliency map from RCNN. Therefore the proposed method is more suitable in scenarios where fast calibration is required. Furthermore, Chen's full probabilistic approach treat gaze points one by one from a local point of view, while the proposed distribution-based method integrates all the gaze points into a global gaze distribution map. Since saliency/fixation map reflects the global attention/interest to the image stimuli, global gaze distribution map is more likely to match the deep eye fixation map and therefore help implicitly estimating the eye parameters. Experiments also demonstrate the proposed distribution-based calibration method is more accurate and data efficient than Chen's full probabilistic approach.

6.5 Comparison with State of the Art

We also compare the proposed implicit calibration framework with other state-of-the-art methods in reducing/eliminating explicit personal calibration. The results are shown in Table 3, the numbers are either directly taken from the corresponding papers or the average of their reported errors. Compared with Model *et al* [2010], their system can achieve 1.3 degree, but their system requires much



(a) 2D regression-based



(b) 3D model-based

Figure 4: *Gaze estimation error with different saliency/fixation maps. 9-points represent the results from explicit calibration method.*

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

Method	Error /degree
Proposed 2D	1.0
Proposed 3D	1.4
Model *et al* [2010]	1.3
Chen *et al* [2011]	1.7
Sugano *et al* [2007]	4.8
Sugano *et al* [2013]	3.5
Alnajar *et al* [2013]	4.3

complex setup and more IR lights. The proposed method also outperforms the method proposed by Chen *et al* [2011]. With deep eye fixation map and the distribution-based calibration method, we can improve the gaze estimation accuracy by 0.3 degree. Even with the same deep eye fixation map, our method appears to be more accurate and data efficient (Sec. 6.4). The last three rows show the gaze estimation error for three appearance-based methods in reducing/eliminating explicit personal calibration. The results are for reference only, we do not intend to compare the proposed model-based method with appearance-based methods. Overall, the proposed method can achieve comparable and better results to the existing state-of-the-art work.

7 Conclusion

In this paper, we propose a novel framework to eliminate explicit personal calibration and improve gaze estimation accuracy for two major gaze estimation methods. The eye fixation learning algorithm takes advantage of top-down eye tracking data and powerful representation and feature learning ability of deep model, so that it is able to capture semantic information that attracts more human attention and represent real human eye fixation. We introduced a regression-based CNN (RCNN) that learns deep features to predict eye fixations. Compared to the traditional hand-crafted image features, the learned image features from deep model can better capture and predict eye fixations. The proposed distribution based calibration method considers the calibration problem from a global perspective in a probabilistic manner without the need of establishing gaze correspondences. Instead of treating training samples separately, we treat all the training samples together as a probabilistic distribution. By combining deep eye fixation map and the distribution based calibration method, our method gives good gaze estimation accuracy, and enables implicit calibration which gives rise to broader eye gaze tracking applications.

References

ALNAJAR, F., GEVERS, T., VALENTI, R., AND GHEBREAB, S. 2013. Calibration-free gaze estimation using human gaze patterns.

BEYMER, D., AND FLICKNER, M. 2003. Eye gaze tracking using an active stereo head. *IEEE Conference in Computer Vision and Pattern Recognition*.

CERF, M., HAREL, J., EINHÄUSER, W., AND KOCH, C. 2008. Predicting human gaze using low-level saliency combined with face detection. In *Advances in neural information processing systems*.

CERF, M., FRADY, E. P., AND KOCH, C. 2009. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision*.

CHEN, J., AND JI, Q. 2011. Probabilistic gaze estimation without active personal calibration. *IEEE Conference on Computer Vision and Pattern Recognition*.

CHEN, J., AND JI, Q. 2014. A probabilistic approach to online eye gaze tracking without personal calibration. *IEEE Transactions on Image Processing*.

CHEN, J., TONG, Y., GARY, W., AND JI, Q. 2008. A robust 3d eye gaze tracking system using noise reduction. *Proceedings of the 2008 symposium on Eye tracking research and applications*.

EINHÄUSER, W., SPAIN, M., AND PERONA, P. 2008. Objects predict fixations better than early saliency. *Journal of Vision*.

FUKUSHIMA, K. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*.

FUNES MORA, K., AND ODOBEZ, J. 2014. Geometric generative gaze estimation (g3e) for remote rgb-d cameras. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

GUESTRIN, E. D., AND EIZENMAN, M. 2006. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transcations on Biomedical Engineering*.

GUESTRIN, E. D., AND EIZENMAN, M. 2008. Remote point-of-gaze estimation requiring a single-point calibration for applications with infants. *Proceedings of the 2008 symposium on Eye tracking research and applications*.

HANSEN, D. W., AND JI, Q. 2010. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

HAREL, J., KOCH, C., AND PERONA, P. 2006. Graph-based visual saliency. *NIPS*.

HAREL, J., KOCH, C., AND PERONA, P. 2006. Graph-based visual saliency. In *Advances in neural information processing systems*, 545–552.

HOU, X., AND ZHANG, L. 2007. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*.

ITTI, L., AND KOCH, C. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*.

ITTI, L., KOCH, C., AND NIEBUR, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence 20*, 11, 1254–1259.

JUDD, T., EHINGER, K., DURAND, F., AND TORRALBA, A. 2009. Learning to predict where humans look. *IEEE 12th International Conference on Computer Vision*.

JUDD, T., DURAND, F., AND TORRALBA, A. 2012. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*.

KIENZLE, W., FRANZ, M. O., SCHÖLKOPF, B., AND WICHMANN, F. A. 2009. Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision 9*, 5, 7.

KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

KULLBACK, S., AND LEIBLER, R. A. 1951. On information and sufficiency. *Ann. Math. Statist. 22*, 1 (03), 79–86.

KÜMMERER, M., THEIS, L., AND BETHGE, M. 2014. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*.

LE, Q. V. 2013. Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*.

LECUN, Y. 2012. Learning invariant feature hierarchies. In *Computer vision–ECCV 2012. Workshops and demonstrations*, Springer, 496–505.

LEE, H., GROSSE, R., RANGANATH, R., AND NG, A. Y. 2009. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*.

LIN, Y., KONG, S., WANG, D., AND ZHUANG, Y. 2014. Saliency detection within a deep convolutional architecture. In *Cognitive Computing for Augmented Human Intelligence Workshop, in conduction with AAAI*.

LU, F., SUGANO, Y., OKABE, T., AND SATO, Y. 2011. Inferring human gaze from appearance via adaptive linear regression. *In Proc. International Conference on Computer Vision*.

LU, F., SUGANO, Y., OKABE, T., AND SATO, Y. 2012. Head pose-free appearance-based gaze sensing via eye image synthesis. *International Conference on Pattern Recognition*.

MAIO, W., CHEN, J., AND JI, Q. 2011. Constraint-based gaze estimationwithout active calibration. In *Automatic Face and Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*.

MANCAS, M., AND PIRRI, F. ADN PIZZOLI, M. 2011. From saliency to eye gaze: embodied visual selection for a pan-tilt-based robotic head. *LNCS Series from Proceedings of the 7th International Symposium on Visual Computing (ISVC)*.

MASON, M., HOOD, B., AND MACRAE, C. 2004. Look into my eyes: Gaze direction and person memory. *Memory*.

MODEL, D., AND EIZENMAN, M. 2010. An automatic personal calibration procedure for advanced gaze estimation systems. *IEEE Transactions on Biomedical Engineering*.

MORIMOTO, C. H., AND MIMICA, M. R. 2005. Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding, Special Issue on Eye Detection and Tracking*.

NUTHMANN, A., AND HENDERSON, J. M. 2010. Object-based attentional selection in scene viewing. *Journal of vision*.

OLSHAUSEN, B. A., AND FIELD, D. J. 1997. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*.

OLSHAUSEN, B. A., ET AL. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*.

RICHE, N., DUVINAGE, M., MANCAS, M., GOSSELIN, B., AND DUTOIT, T. 2013. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *Computer Vision (ICCV), 2013 IEEE International Conference on*.

RIESENHUBER, M., AND POGGIO, T. 1999. Hierarchical models of object recognition in cortex. *Nature neuroscience*.

ROSENHOLTZ, R. 1999. A simple saliency model predicts a number of motion popout phenomena. *Vision research*.

SERRE, T., WOLF, L., BILESCHI, S., RIESENHUBER, M., AND POGGIO, T. 2007. Robust object recognition with cortex-like mechanisms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*.

SHEN, C., AND ZHAO, Q. 2014. Learning to predict eye fixations for semantic contents using multi-layer sparse network. *Neurocomputing*.

SHIH, S. W., AND LIU, J. 2004. A novel approach to 3-d gaze tracking using stereo cameras. *IEEE Transactions on Systems, Man and Cybernetics, PartB*.

SUGANO, Y., MATSUSHITA, Y., SATO, Y., AND KOIKE, H. 2007. An incremental learning method for unconstrained gaze estimation. In *in European Conference on Computer Vision*.

SUGANO, Y., MATSUSHITA, Y., AND SATO, Y. 2010. Calibration-free gaze sensing using saliency maps. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

SUGANO, Y., MATSUSHITA, Y., AND SATO, Y. 2013. Appearance-based gaze estimation using visual saliency. *IEEE Trans. Pattern Analysis and Machine Intelligence*.

SUGANO, Y., MATSUSHITA, Y., AND SATO, Y. 2014. Learning-by-synthesis for appearance-based 3d gaze estimation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

TAN, K. H., KRIEGMAN, D., AND AHUJA, N. 2002. Appearance-based eye gaze estimation. *In Proc. 6th IEEE Workshop on Applications of Computer Vision*.

Tobii technology. http://www.tobii.com/en/eye-experience/.

TREISMAN, A. M., AND GELADE, G. 1980. A feature-integration theory of attention. *Cognitive psychology 12*, 1, 97–136.

VALENTI, R., SEBE, N., AND GEVERS, T. 2011. What are you looking at? improving visual gaze estimation by saliency. *Int J of Computer Vision*.

VIG, E., DORR, M., AND COX, D. 2014. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Computer Vision and Pattern Recognition, 2014. CVPR'14. IEEE Conference on*.

WEN, S., HAN, J., ZHANG, D., AND GUO, L. 2014. Saliency detection based on feature learning using deep boltzmann machines. In *ICME*, 1–6.

WILLIAMS, O., BLAKE, A., AND CIPOLLA, R. 2006. Sparse and semi-supervised visual mapping with the s3gp. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

XU, J., JIANG, M., WANG, S., KANKANHALLI, M. S., AND ZHAO, Q. 2014. Predicting human gaze beyond pixels. *Journal of vision*.

YAMINS, D. L., HONG, H., CADIEU, C. F., SOLOMON, E. A., SEIBERT, D., AND DICARLO, J. J. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 201403112.

ZHAO, Q., AND KOCH, C. 2011. Learning a saliency map using fixated locations in natural scenes. *Journal of vision*.

ZHU, Z., AND JI, Q. 2004. Eye and gaze tracking for interactive graphic display. *Machine Vision and Applications*.

ZHU, Z., JI, Q., AND BENNETT, K. P. 2006. Nonlinear eye gaze mapping function estimation via support vector regression. *International Conference on Pattern Recognition*.