Towards Accurate and Robust Cross-Ratio based Gaze Trackers Through Learning From Simulation

Jia-Bin Huang[1], Qin Cai[2], Zicheng Liu[2], Narendra Ahuja[1], and Zhengyou Zhang[1]

[1]University of Illinois, Urbana-Champaign
[2]Microsoft Research

Abstract

Cross-ratio (CR) based methods offer many attractive properties for remote gaze estimation using a single camera in an uncalibrated setup by exploiting invariance of a plane projectivity. Unfortunately, due to several simplification assumptions, the performance of CR-based eye gaze trackers decays significantly as the subject moves away from the calibration position. In this paper, we introduce an adaptive homography mapping for achieving gaze prediction with higher accuracy at the calibration position and more robustness under head movements. This is achieved with a learning-based method for compensating both spatially-varying gaze errors and head pose dependent errors simultaneously in a unified framework. The model of adaptive homography is trained offline using simulated data, saving a tremendous amount of time in data collection. We validate the effectiveness of the proposed approach using both simulated and real data from a physical setup. We show that our method compares favorably against other state-of-the-art CR based methods.

CR Categories: I.5.4 [Applications]: Computer vision; I.3.6 [Methodology and Techniques]: Interaction techniques;

Keywords: Cross-ratio; Eye gaze tracking; Simulation

1 Introduction

Gaze tracking research has been extensively studied for the past few decades [Duchowski 2007; Holmqvist et al. 2011; Hansen and Ji 2010]. An effective gaze tracking system is of great interest because it can enable two important applications: 1) multi-modal natural interaction [Morimoto and Mimica 2005; Zhai et al. 1999] and 2) understanding and analyzing human attention [Pantic et al. 2007].

Typical remote gaze tracking systems consist of one or more cameras for capturing subject's eye images and multiple infrared light sources for generating corneal reflections (glints). The captured images are then processed to extract informative features that are invariant to illumination and viewpoint. Commonly used features include pupil center, corneal reflections and limbus contour. Note that we focus mainly on feature-based gaze tracking methods. For other approaches such as appearance-based methods, we refer the readers to a recent survey [Hansen and Ji 2010] for a comprehensive review.

There are two types of feature-based approaches for gaze prediction: interpolation-based and model-based [Hansen and Ji 2010]. *Interpolation-based methods* [Morimoto and Mimica 2005; Cerrolaza et al. 2008; Cerrolaza et al. 2012] directly map eye features to gaze points through 2D regression functions without considering the optical properties, the eye physiology, and the geometric relationship between eye, screen and camera. Therefore, interpolation-based methods are sensitive to head movements, especially to depth variation. Yet, they are simple to implement and do not need to go through tedious hardware calibration procedures. *Model-based methods* [Guestrin and Eizenman 2006; Hennessey et al. 2006; Model and Eizenman 2010] estimate the 3D gaze vector and compute 2D point of regard by intersecting 3D rays with the 2D screen plane. Unlike interpolation-based methods, model-based methods are able to accommodate larger head movements. However, they require more complex system setup and fully calibrated hardware.

Cross-ratio (CR) based approaches offer the two advantages from both interpolation-based and model-based methods: (1) do not require hardware calibration and (2) allow free head motion. However, as shown in [Kang et al. 2008], the subject-dependent estimation bias arises from two main causes: (1) the angular deviation of the visual axis from the optic axis and (2) the virtual image of the pupil center is not coplanar with corneal reflections. Many extensions have been proposed to improve the basic CR-based approach [Yoo et al. 2002] along two directions: accuracy at the calibrated position and robustness under head movements. In Figure 1, we show a qualitative comparison with existing CR-based methods using the improvement over accuracy and robustness as main axes.

First, several efforts have been made to correct the estimation bias induced from the simplification assumptions for accurate gaze estimation at the calibrated position. The bias compensation is usually achieved by applying a 2D planar transformation on the predicted gaze by the basic CR method [Yoo et al. 2002]. These 2D planar transformation can be computed by a subject-dependent calibration, i.e., asking subjects to look at a few predefined calibration targets on the screen. Examples include scale correction in [Yoo and Chung 2005], scale and translation correction in [Coutinho and Morimoto 2006] and homography-based correction [Kang et al. 2007; Hansen et al. 2010]. Note that some of the bias correction methods [Yoo and Chung 2005; Coutinho and Morimoto 2006] require the projected position the corneal center in image, which can be obtained using an additional on-axis light source. For further accuracy, error compensation methods have also been introduced to model and compensate spatially-varying errors, e.g., via polynomial-based regression [Cerrolaza et al. 2012; Cerrolaza et al. 2008] or Gaussian process regression [Hansen et al. 2010].

Second, as the CR-based methods do not explicitly compensate for head motion, they are not robust under head movements. The performance of CR-based methods decays rapidly when the head position deviates from the calibrated position, especially along the depth axis. This problem could be partially alleviated by additional calibration positions, yet it involves significant increase in subject-dependent calibration time. In [Coutinho and Morimoto 2010], a depth compensation method was proposed using depth-adaptive displacement vector correction, extented from [Coutinho and Morimoto 2006]. Later, they proposed a hybrid (model-based and cross-ratio based) approach to account for the errors from both eye translation and rotation [Coutinho and Morimoto 2012]. They estimate the visual axis and the plane formed by the virtual im-

ETRA 2014, March 26 – 28, 2014, Safety Harbor, Florida, USA.

ages of the glints under a weak perspective camera model. After correcting the angular deviation of visual and optical axis and the co-planarity assumptions, the basic CR-based model was applied to predict the final gaze .

Despite these attempts on extending CR-based methods, two fundamental problems remain unsolved. First, the pursuits of improving the accuracy and robustness of gaze trackers are usually separately addressed. Second, the state-of-the-art head pose adaptation method in [Coutinho and Morimoto 2012] is derived under weak-perspective camera models. In other words, the camera is required with large focal length, has limited field-of-view and can only capture limited head positions.

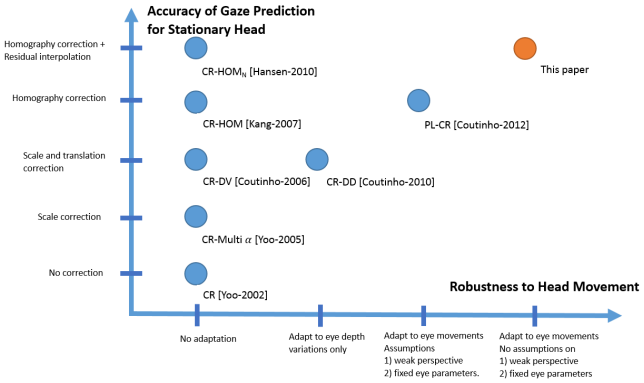


Figure 1: *Qualitative comparison of our approach with existing cross-ratio based methods along two directions: accuracy for static head and robustness under head movements.*

In this paper, we propose a learning-based approach to push the envelope of CR based methods for accurate and robust eye gaze tracking. Our base bias correction method is based on the homography-based methods [Kang et al. 2007; Hansen et al. 2010]. The homography-based methods work well at the calibration position. However, they are not robust to head movement and may not be sufficiently accurate due to the lack of spatial-varying error modeling. To compensate both types of errors in a unified framework, we introduce an adaptive homography mapping, which depends on two types of predictor variables capturing 1) the head movement relative to the calibration position and 2) the position of the gaze on the screen. We collect the groundtruth data for training the adaptive homography mapping through a series of subject-dependent calibration at various head positions using simulation. During testing, the trained model is used to adaptively correct the bias induced from both types of errors.

We make the following three contributions.

- We introduce a learning-based adaptation approach for simultaneously compensating spatially-varying errors and errors induced from head movements. We improve both the accuracy and the robustness of CR based gaze tracking systems in a unified framework.
- Our method generalizes previous works on compensating head movement using glint transformation, e.g., the distance between glints [Cerrolaza et al. 2012] or size variation of the glint pattern in [Coutinho and Morimoto 2010] by considering the geometric transformation between the glint patterns. The resultant model not only compensates depth variations, but also movements parallel to the screen plane.
- As we obtain the adaptation function through a learning process trained on simulated data, any prior knowledge about the system setup (if available) can be easily incorporated into our system.

2 An Accurate and Robust Gaze Estimation Method

2.1 Overviews

Our method is built upon the recent advances of homography-based methods for gaze estimation bias correction [Kang et al. 2007; Hansen et al. 2010]. The bias correcting homography transformation can be computed via solving the point set registration problem from the predicted gaze points by the basic CR method [Yoo et al. 2002] to the groundtruth targets on the screen during subject-dependent calibration. Homography-based methods can be considered as generalization of methods allowing limited 2D planar transformation, e.g., scaling [Yoo and Chung 2005] or scaling and translation [Coutinho and Morimoto 2006] because of their ability to correct perspective distortion.

Improving accuracy at static head position Homograph-based methods generally work well at the calibration position because they effectively model the optical and visual axis offsets. However, due to the model error from the planarity assumption on pupil center and the plane formed by glints, spatially-varying errors arise. Therefore, for accurate prediction, the bias correcting homography mapping should depend on the subject's gaze direction.

Improving robustness under head movements It is well known that the performance of homography-based methods degrade significantly when the subject moves away from the calibration position because the optimal bias correcting homography is a function of head positions. We illustrate this effect by performing a series of subject-dependent calibrations at head positions located at different distances to the screen using simulation. In Figure 2, we plot the values of the optimal bias-correcting homography computed at different head positions along the depth axis. From left to right, we plot the scaling terms for x, y and translation terms for x, y, respectively. We can see that these values change smoothly with the head positions. Supposing we can "predict" how the bias-correcting homography changes at a new head position, the performance of the gaze tracker will be as if it were calibrated even at that new head position.

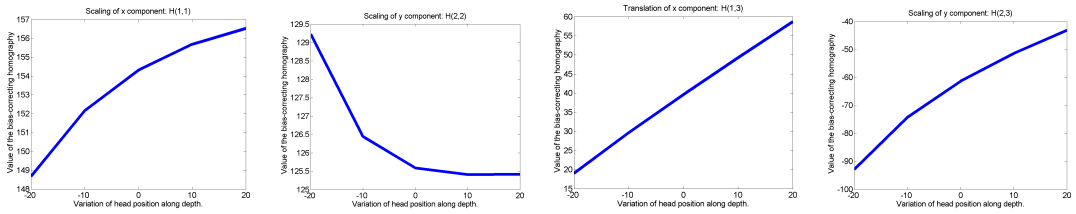
With these two insights, our goal is to design a scheme to predict the variation of the bias correcting homography computed at the calibration position based on two factors: 1) the relative changes between the current head position and the calibration position and 2) the current gaze direction.

2.2 Adaptive Homography Mapping

CR with homography-based bias correction Denote L_i as the point light sources located at the four screen corners $(1 \leq i \leq 4)$, G_i as the corresponding corneal reflection and g_i as the images of G_i. We also denote P as the pupil center in 3D and p as its projection in image. The CR method [Yoo et al. 2002] assumes each of the group (L_i, G_i, g_i) is co-planar, denoted as plane Π_L, Π_G, Π_g, respectively. We can thus describe the transformation between planes Π_L, Π_G, Π_g through homographies. Under the assumption that the pupil center P lies in Π_G, the point of regard prediction is given by

$$\mathrm{PoR}_{\mathrm{CR}} = \mathbf{H}_{\mathrm{GL}}(\mathbf{H}_{\mathrm{gG}}(p)) = \mathbf{H}_{\mathrm{CR}}(p), \qquad (1)$$

where $\mathbf{H}_{\mathrm{gG}}$ maps plane Π_g to plane Π_G, $\mathbf{H}_{\mathrm{GL}}$ maps plane Π_G to plane Π_L, and $\mathbf{H}_{\mathrm{CR}}$ is the combined transform of $\mathbf{H}_{\mathrm{GL}}$ and $\mathbf{H}_{\mathrm{gG}}$.



Scaling on x-axis | Scaling on y-axis | Translation along x-axis | Translation along y-axis

Figure 2: *The values of the optimal bias-correcting homography computed at different head positions along the depth axis. Note that the last element of each homography is normalized to one.*

However, because these simplification assumptions are not valid in practice, large gaze estimation biases are observed. Homography-based methods [Kang et al. 2007; Hansen et al. 2010] apply another homography transformation to correct these gaze estimation bias. In [Hansen et al. 2010], the glints in images are first mapped onto a normalized space (e.g., a unitary square Π_N) and then use the bias-correcting homography to map the estimated gaze points in the normalized space to the expected gaze points in the screen space Π_L. The point of regard prediction by homography-based prediction is given by

$$\mathrm{PoR}_{\mathrm{HOM}} = \mathbf{H}_{\mathrm{NL}}(\mathbf{H}^{\mathrm{N}}_{\mathrm{CR}}(p)), \tag{2}$$

where $\mathbf{H}^{\mathrm{N}}_{\mathrm{CR}}$ maps the image space to the normalized space and $\mathbf{H}_{\mathrm{NL}}$ maps the normalized space to the screen space.

Denote v as the index for the target position on the screen, $\mathcal{V}$ as the set of the target index, and t_v as the target position in the screen space. The goal of the subject-dependent calibration is to find the optimal bias-correcting homography $\mathbf{H}^*_{\mathrm{NL}}$ that minimizes the re-projection errors:

$$\mathbf{H}^*_{\mathrm{NL}} = \underset{\mathbf{H}_{\mathrm{NL}}}{\operatorname{argmin}} \sum_{v\in\mathcal{V}} ||t_v - \mathbf{H}_{\mathrm{NL}}(\mathbf{H}^{\mathrm{N}}_{\mathrm{CR}}(p_v))||^2_2, \tag{3}$$

where p_v is the 2D pupil center position in the image when gazing at target v.

Adaptive Homography Mapping We propose to model the variation of the bias-correcting homography $\mathbf{H}_{\mathrm{NL}}$ using another homography mapping $\mathbf{H}_{\mathrm{A}}$. The point of regard by the adaptive homography is given by

$$\mathrm{PoR}_{\mathrm{AH}} = \mathbf{H}_{\mathrm{NL}}(\mathbf{H}_{\mathrm{A}}(\mathbf{H}_{\mathrm{CR}}(p))) \tag{4}$$

Note that in Eqn 4, the bias-correcting homography $\mathbf{H}_{\mathrm{NL}}$ is computed by the same minimization process in Eqn 3 at the calibration and remain unchanged for the same subject. The adaptive homography mapping $\mathbf{H}_{\mathrm{A}}$, on the other hand, needs to vary adaptive to the current head position relative to the calibration position as well as the gaze direction. We thus pose the adaptive homography as a regression problem. That is, given predictor variables describing the relative head position and gaze direction, we want to predict the values in $\mathbf{H}_{\mathrm{A}}$.

We propose two types of predictor variables $\mathbf{x} = [\mathbf{x_m}, \mathbf{x_g}]^{\mathrm{T}}$. First, we capture the head movements relative to the calibration position using the geometric transformation between the glints quadrilateral stored at the calibration and the current glints quadrilateral. In practice, we use affine or similarity transformation to encode the relative movement. For example, when the subject moves toward the screen after calibration, the scale term of the transformation will be greater than one. We obtain the first type of predictor variable $\mathbf{x_m}$ by vectorizing the motion parameters. Therefore, we have 6-dimensional and 4-dimensional vector for $\mathbf{x}_m$ when using affine transformation and similarity transformation, respectively. Second, for encoding the gaze direction for spatially-varying mapping, we use the pupil center position in the normalized space $\mathbf{x_g} = \mathbf{H}_{\mathrm{CR}}(p - p_0)$ as features, where p_0 is the pupil center position when gazed at the center of the screen.

With these predictor variables, we model the adaptive homography as polynomial regression of degree two (i.e., quadratic regression):

$$\mathbf{H}_{\mathrm{A},\mathbf{x}} = f(\mathbf{x}, \beta), \tag{5}$$

In the quadratic regression, the values of the adaptive homography are linear with the predictor variables, which contain a constant term, linear terms, quadratic terms, as well as the interaction terms.

Relationship with prior works In [Coutinho and Morimoto 2010], error compensation for depth variation is achieved by adaptively scaling the translational correction vectors using the relative size of the glint quadrilaterals at the calibration position and the current position. This could be considered as a special case within our formulation. Our formulation considers a richer set of transformation than scaling for prediction and use homography instead of translation only for correction. As observed in the simulation in Figure 2, the values of the optimal bias-correcting homographies are all dependent on the head movements. It suggests that using only translation for correction is suboptimal.

In [Cerrolaza et al. 2012], the error compensation is achieved by predicting the translation vector using data collected by a series of subject calibrations. In contrast, we use simulated training data for learning the adaptation, saving tremendous amount of subject calibration time. In addition, the use of simulation allows us to use more complex model than just translation for prediction.

2.3 Learning Homgraphy Adaptation

Denote u as the head position in 3D and $\mathcal{U}$ as the set of sampled head positions. Our objective function is defined as

$$\mathcal{L}(\beta) = \sum_{u\in\mathcal{U}} \sum_{v\in\mathcal{V}} ||t_{u,v} - \mathbf{H}^*_{\mathrm{NL}}(\mathbf{H}_{\mathrm{A},\mathbf{x}}(\mathbf{H}^{\mathrm{N}}_{\mathrm{CR}}(p_{u,v})))||^2_2, \tag{6}$$

where $\mathbf{H}_{\mathrm{A},\mathbf{x}} = f(\mathbf{x}, \beta)$ is the quadratic regression model for adaptive homography. The goal of learning adaptive homography is to find the optimal matrix of coefficients that minimize the reprojection errors by summing all the squared errors between the predicted gaze positions and the groundtruth ones on the screen when the simulated subjects are located at all sampled head positions.

Training data collection We define the screen plane as the $x-y$ plane and the depth from screen as the z-axis in the world coordinate system. We sample a typical working space in front of the

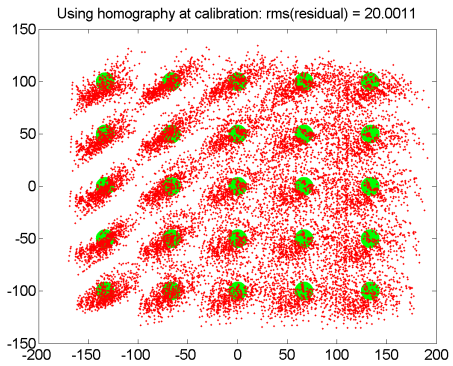


Figure 3: *Eye gaze prediction results using the bias-correcting homography computed at the calibration position. Green dots are groundtruth and red dots are predictions.*

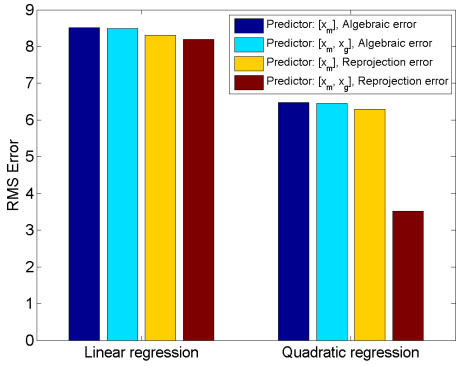


Figure 5: *Error comparison of using different training models in terms of root mean sqaure errors.*

screen using a $5 \times 5 \times 5$ grid with ranges from -200mm to 200mm, centered at position [0, 0, 600] mm. At each head position u, we perform subject-dependent calibration in Eqn 3 using a 5×5 calibration pattern on the screen. To account for subjects with different eye parameters, we also randomly sample 50 virtual subjects using Gaussian distributions with means of typical eye parameters in [Guestrin and Eizenman 2006] and standard deviations of 10% of the values of the parameter. For example, the typical size of corneal radius is 7.8 mm. We then draw random samples using a Gaussian distribution with mean 7.8 and standard deviation 0.78.

Objective function minimization To minimize the objective function defined in Eqn 6, we take a two-step approach. First, we estimate the prediction function by minimizing an algebraic error. At each head position u, we can compute the optimal bias-correcting homography $\mathbf{H}^u_{\mathrm{NL}}$ by performing a subject-dependent calibration at position u. Ideally, $\mathbf{H}^u_{\mathrm{NL}} = \mathbf{H}^*_{\mathrm{NL}}\mathbf{H}_{\mathrm{A},\mathbf{x}}$ up to a scale factor. We can thus minimize the algebric errors between the prediction $\mathbf{H}_{A,\mathbf{x}} = f(\mathbf{x}_{u,v}, \beta)$ and the difference of the bias-correcting homography $(\mathbf{H}^*_{\mathrm{NL}})^{-1}(\mathbf{H}^u_{\mathrm{NL}})$ (with the last element normalized to 1), where the $\mathbf{H}^*_{\mathrm{NL}}$ is the bias-correcting homography computed at the default calibration position. The algebraic error minimization can thus be formulated as

$$\beta^a = \underset{\beta}{\operatorname{argmin}} \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{V}} ||(\mathbf{H}^*_{\mathrm{NL}})^{-1}(\mathbf{H}^u_{\mathrm{NL}}) - f(\mathbf{x}_{u,v}, \beta)||_2^2, \quad (7)$$

where β^a is the estimated matrix of coefficients after minimizing the algebraic errors. Second, to minimize the reprojection errors in Eqn 6, we start with the initial solution using β^a and perform non-linear least square optimization using the Levenberg-Marquardt algorithm.

Training process visualization Here we use a subset of training data (10 virtual subjects) to visualize the training process and training model selection. In Figure 3, we show the gaze prediction results at various head positions using the bias-correcting homography $\mathbf{H}^*_{\mathrm{NL}}$ computed at the default calibration position. We can see that the gaze predictions are widely scattered, indicating the need for head position adaptation.

In Figure 4, we show the prediction using adaptive homography under various model selection choices. We show the results of the linear and quadratic regression model in the first and second row, respectively. The training errors in terms of root mean square errors can be found in Figure 5. With quadratic regression, the model of using both motion parameters of the glints quadrilaterals $\mathbf{x_m}$ and pupil center in the normalized space $\mathbf{x_g}$ as predictor variables has the best performance in training. This training process visualization provides two insights. First, the linear regression model may not be sufficiently expressive in capturing the homography variation among different head positions. For example, from the training errors in RMSE, we can see that the reprojection minimization is not able to reduce the errors further. Second, from the visualization of algebraic errors and reprojection errors minimization, we observe that including the pupil center position in the normalized space $\mathbf{x_g}$ as features can significantly improve the adaptation, reducing the training errors by half. Also, as can be seen from Figure 4 (second row, column 3 and 4), the adaptation is able to compensate the spatially-varying errors due to the non-coplanarity of the pupil center and the plane formed by the glints.

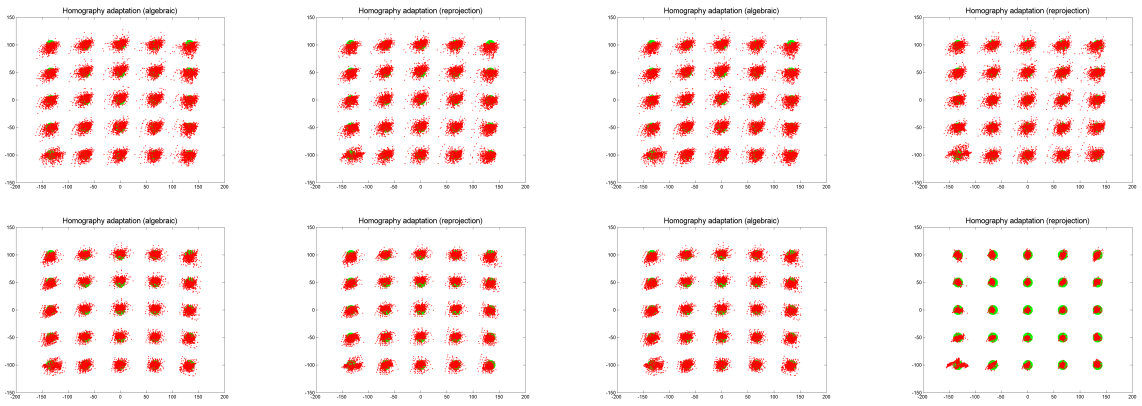
3 Evaluation using Simulated Data

In the simulated evaluation, our goal is to investigate and understand how various factors affect the performance of the gaze tracking. These factors including system setup, sensor resolution, noise level, number of calibration points, eye parameters, and head movements.

We compare the performance of our method to that of other three state-of-the-art CR-based gaze trackers: CR-DV [Coutinho and Morimoto 2006], CR-HOM [Hansen et al. 2010], CR-DD [Coutinho and Morimoto 2010]. [1] Among them, CR-D [Coutinho and Morimoto 2006] improves the accuracy of basic CR method [Yoo et al. 2002] via scale and translation correction. CR-HOM [Hansen et al. 2010] maps estimated gaze points from the basic CR method [Yoo et al. 2002] to the groundtruth gaze points using a homography transformation. CR-DD [Coutinho and Morimoto 2010] and PL-CR [Coutinho and Morimoto 2012] are two recent efforts on improving the robustness aspect of the CR-based gaze tracking system.

In the following evaluation, we use the gaze estimation error (in degree) to measure the accuracy. At each head position, the subject (or simulated eye model) is asked to gaze at a pre-defined group of targets on the screen. An averaged gaze error is then computed by taking the average of the gaze estimation errors (in degree) for all screen targets.

[1] We implemented the method PL-CR in [Coutinho and Morimoto 2012]. However, we found that the method perform poorly (gaze erros $\geq 3^\circ$) when the weak perspective camera assumption was violated. As noted in [Coutinho and Morimoto 2012], the camera needs to have a narrow field of view ($\approx 5^\circ$) and needs to be repositioned whenever the subjects move.



$\mathbf{x} = [\mathbf{x_m}]$ (Algebraic) $\mathbf{x} = [\mathbf{x_m}]$ (Reprojection) $\mathbf{x} = [\mathbf{x_m}, \mathbf{x_g}]$ (Algebraic) $\mathbf{x} = [\mathbf{x_m}, \mathbf{x_g}]$ (Reprojection)

Figure 4: *Training process visualization. These plots are prediction results using adaptive homography under different models. The first row: linear regression model. The second row: quadratic regression model. Compared with Figure 3, the adaptive homography significantly reduce the errors due to head movements.*

3.1 Setup

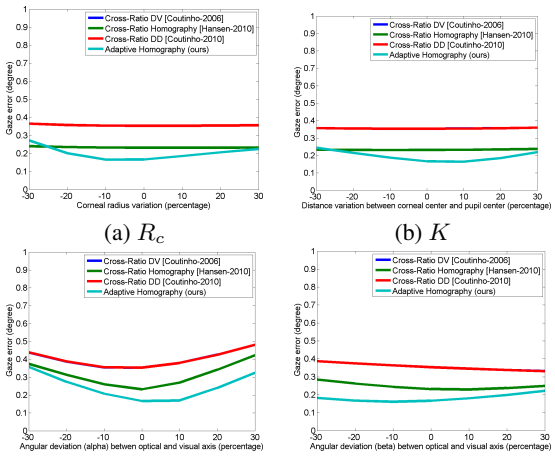
Screen and camera model The simulated screen is of size 400mm×300mm. The four IR light sources are placed at the screen plane with horizontal offset 27 mm and vertical offset 29 mm to each screen corner. The camera is located slightly below the screen. The camera has a 13mm focal length with sensor size (7.18mm×5.32mm) to allow large head movement without repositioning the camera. The horizontal field of view can be computed as $\text{FoV} = 2\arctan\frac{d}{2f} \approx 30.87°$. We assume that the captured image resolution has 1920×1080 pixels.

Calibration position and eye parameters The default calibration head position is located at $(0, 50, 600)$mm, where $(0, 0, 0)$ indicates the center of the screen. We simulate the eye model by using the typical eye parameters listed in [Guestrin and Eizenman 2006]. The cornea is modeled as a sphere with a radius of 7.8mm and the distance from pupil center to corneal center is 4.2mm. The effective index of refraction is modeled as 1.3375. We use the left eye for evaluation. The visual deviation between visual axis and optical axis is $5°$ for horizontal angle and $1.5°$ for vertical angle. In the simulation, given a fixed head position, we rotate along the eyeball rotation center (*not the corneal center*) so that the visual axis intersects with the target gaze point on the screen.

Calibration and testing process During the calibration process, the subject is asked to gaze at a regular $n \times n, n \in \{2, 3, 4, 5\}$ grid pattern that is uniformly distributed over the screen. In the testing, a uniformly distributed 5×5 grid on the screen was used.

3.2 Stationary Head

Sensitivity to eye parameters We first examine the sensitivity of the proposed method to different eye parameters. Starting with typical eye parameters (corneal radius $R_c = 7.8$mm, distance from corneal center to pupil center $K = 4.2$mm, horizontal and vertical angular deviation $\alpha = 5°$ and $\beta = 1.5°$), we vary the value of each eye parameters with [-30, 30]% of the original values. We show the gaze prediction accuracy at the calibration position in Figure 6. As CR-DD [Coutinho and Morimoto 2010] is a depth-adaptive extension of CR-DV [Coutinho and Morimoto 2006], their performances are nearly identical at the calibration position. Compared with the other three methods [Coutinho and Morimoto 2006; Hansen et al. 2010; Coutinho and Morimoto 2010], our adaptive homography is stable across different eye parameters with slight accuracy improvements.



(a) R_c (b) K

(c) α (b) β

Figure 6: *Sensitivity to eye parameters. (a) Corneal radius R_c. (b) Distance from corneal center to pupil center K. (c) Horizontal angular deviation between visual and optical axis α. (d) Vertical angular deviation between visual and optical axis β*

Number of calibration points Figure 7 shows the accuracy of the methods as a function of the calibration points on the screen. The method in [Coutinho and Morimoto 2006] does not benefit much from increasing the number of calibration points due to the use of only scale and translational correction. Our method provides extra accuracy over the homography-based method [Hansen et al. 2010] because the adaptive homography also accounts for the spatially-varying gaze errors predicted by the pupil position in the normalized space $\mathbf{x_g}$.

3.3 Head Movements

Robustness to head movement is one of the most challenging objectives of gaze trackers. In the following, we demonstrate the influences of head movements away from the calibration position. We also examine several coupling factors in the case of the depth variation to better characterize the properties of the gaze trackers.

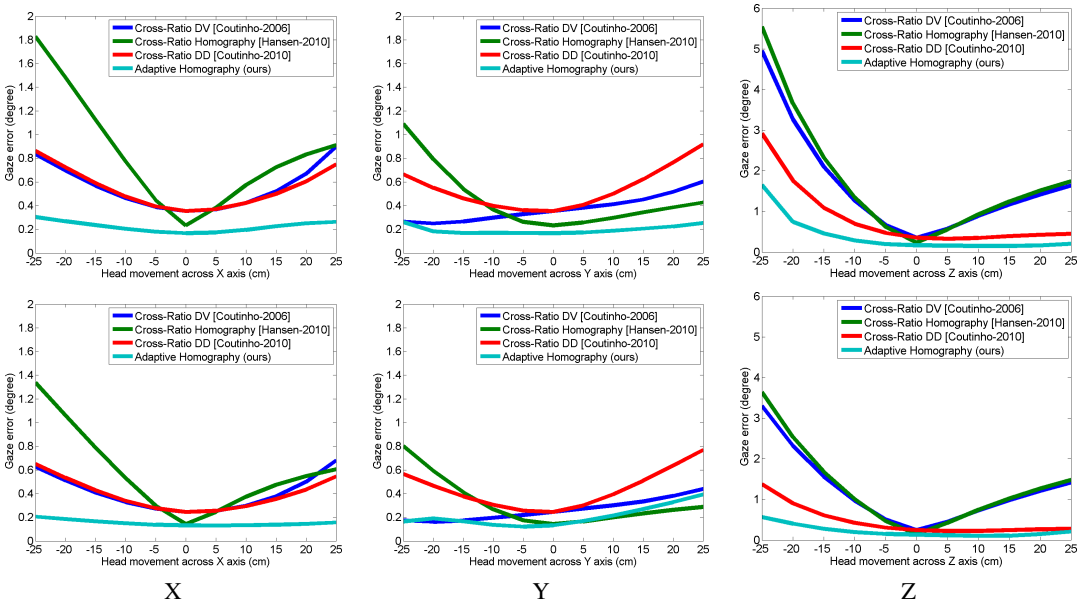


Figure 8: *Eye gaze prediction accuracy under head movements in X, Y, Z directions. First row calibrated at* $(0, 50, 600)$*mm. Second row: calibrated* $(50, 100, 700)$*mm.*

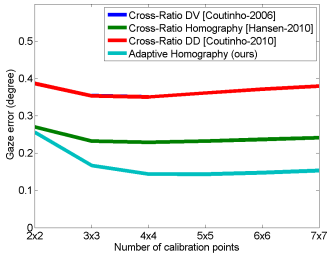


Figure 7: *Accuracy as a function of the number of calibration points on the screen.*

Head movements in three different directions We first test the eye gaze prediction accuracy in the noiseless environment. The virtual subject is calibrated at the default head positions $(0, 50, 600)$mm and $(50, 100, 700)$mm. After calibration, we then move the virtual subject along three directions, X, Y, and Z from -250mm to 250mm with a step size 50mm. Figure 8 shows the effect of head movement in three directions for both initial calibration positions. The results show that the proposed adaptive homography method is robust to head movements in all three head movement directions. The performance of the methods [Coutinho and Morimoto 2006; Hansen et al. 2010] that do not account for head movements decay rapidly when the subject moves away from the calibrated position. This issue is particularly severe along the depth variation. For example, for a subject moving toward the screen by 250mm, the gaze prediction errors by [Coutinho and Morimoto 2006; Hansen et al. 2010] increase rapidly to more than 5°. In contrast, our method remains stable with gaze error less than 2° at this extreme case. For head movements parallel to the screen, our method achieves less than 0.3° errors for all head positions. The depth adaptive method in [Coutinho and Morimoto 2010], however, is not robust to such movements. We also note that our method is insensitive to the initial calibration position. We observe similar results for both calibration positions $(0, 50, 600)$mm in the first row and $(50, 100, 700)$mm in the second row.

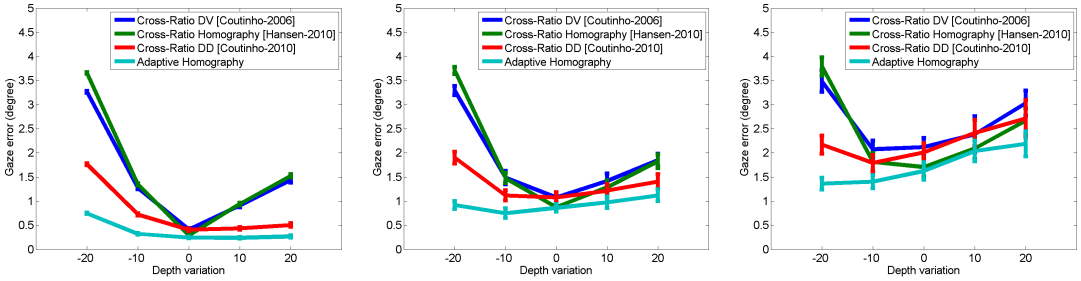
Influence of Noises We investigate the stability to noise level of our gaze tracking system. We set the sensor resolution as 1980×1080 and focal length 13mm for sufficient field-of-view. We use independently added Gaussian distributions as noises in localizing the eye features, e.g., glints and pupil center, with three different levels of standard deviations: $\sigma = \{0.1, 0.5, 1\}$. Figure 9 shows the gaze prediction accuracy under three different levels of noise. The benefits of our adaptation become less obvious for higher noise levels because the noises also negatively affect the adaptation prediction.

Sensor resolution and head movements It is of interest how the sensor resolution affects the performance. This evaluation provides a reference for choosing a suitable sensor resolution to meet the required accuracy. We fix the image noise to have standard deviation of 0.5 pixels and examine the performance under five image resolutions: 640×480, 1280×720, 1920×1080, 2880×1620 and 3840×2160. With 13mm focal length, the diagonal distances of the glint quadrilateral are around 8, 17, 25, 40, and 50 pixels for the five image resolutions, respectively. The image noise is added independently to each glint position and pupil center position. All the results are computed by averaging over 30 trials. We show two sets of results on sensor resolutions: 1) at the calibration position and 2) away from the calibrated position. First, we show in Figure 10 the gaze accuracy at the calibration position under various sensor resolutions. Second, Figure 11 shows the gaze accuracy at a new head position. Figure 10 suggests that for cameras with 13mm focal length, sensor resolution of size 1920×1080 is required for accuracy under 1°. With larger focal length (e.g., 35mm), a lower resolution of 640×480 may be sufficient.

4 Evaluation using a Physical Setup

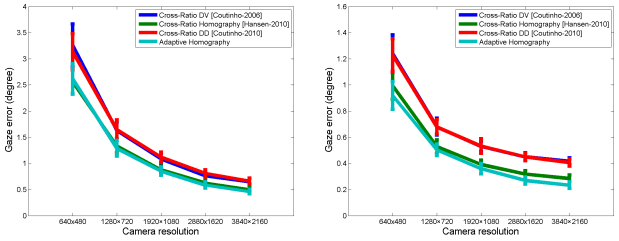
4.1 Setup

The physical setup consists of a 21 inch screen, with a total of 8 IR LED lights located on the four screen corners and the middle points of each edge of the screen border. The purpose of using 8 IR LED lights instead of 4 is to increase the robustness in glint detection. We use a CMOS point grey camera with 1/1.8" sensor of 1280×1080 pixels and Fujinon 3 MP Lens and with focal length set to be 13mm.



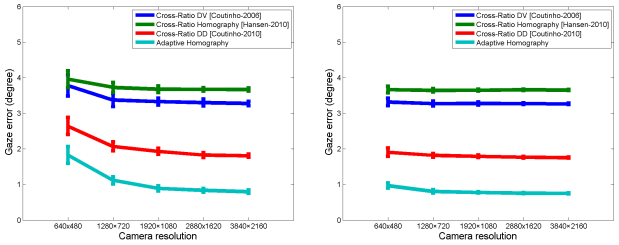
(a) $\sigma = 0.1$ (b) $\sigma = 0.5$ (c) $\sigma = 1$

Figure 9: *Accuracy under head movements for different levels of noise. (a) $\sigma = 0.1$ (b) $\sigma = 0.5$ (c) $\sigma = 1$*



$f = 13$mm $f = 35$mm

Figure 10: *The influence of sensor resolution on gaze accuracy at the calibration position.*



$f = 13$mm $f = 35$mm

Figure 11: *The influence of sensor resolution on gaze accuracy at new head position.*

The camera is located around 5cm below the middle point of the bottom edge of the screen and 15cm away from the screen plane. The subject is roughly located along the axis perpendicular to the screen center. At each position, a chin rest is used to support the subject's head during capturing eye images. However, there are still inevitable head movements during calibration and testing.

For data collection, we ask the subjects to gaze at a uniformly distributed 5×5 grid on the screen and record the captured eye images. We gather a maximum of 60 samples (2-3 seconds) for each gaze target position. To avoid capturing the eye images during the transition of the eye gaze, we show the next target first and record eye images for the next target "after" the subject clicks the mouse. After capturing these eye images, we then extract the glints using image thresholding and detect pupil center with ellipse fitting.

4.2 Experiments

Based on the physical setup, we compare our method with the homography-based method CR-HOM [Hansen et al. 2010]. All other cross-ratio methods including CR-Multiα [Yoo and Chung 2005], CR-DD [Coutinho and Morimoto 2006], CR-DV [Coutinho and Morimoto 2010], and PL-CR [Coutinho and Morimoto 2012] require additional on-axis ring right for generating projection of the corneal center.

We present two sets of experiments using data from a physical setup. First, we show the accuracy improvement at the calibration position. Figure 12 shows the performance comparison with CR-HOM [Hansen et al. 2010] using a series of subject-dependent calibrations at different depth positions. The computed bias-correcting homography $\mathbf{H}^{*}_{\mathrm{NL}}$ is the same for both CR-HOM and our method. However, we can see from Figure 12 that the accuracy at the calibration position of our adaptive homography consistently outperforms that of the CR-HOM [Hansen et al. 2010] at all depth positions. The gaze prediction accuracy values reported here are computed by averaging the gaze errors in degree using all raw samples, i.e., no temporal smoothing or post-processing are used.

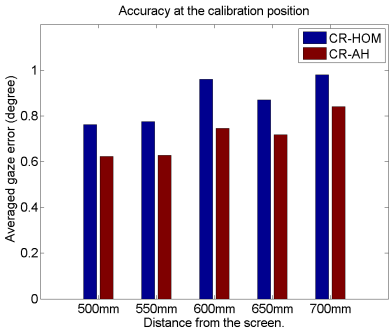
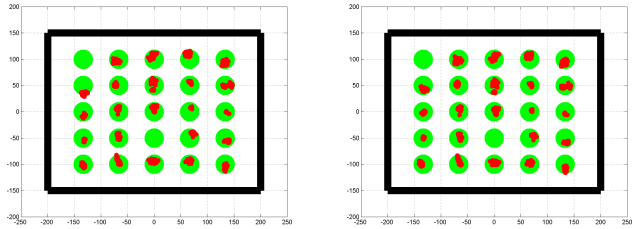


Figure 12: *Accuracy at the calibration position. Both CR-HOM and our adaptive homography use the same bias-correcting homography.*

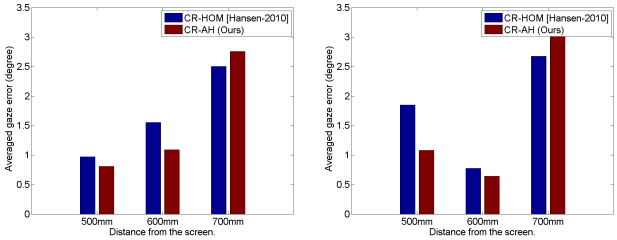
Figure 13 shows a qualitative comparison to demonstrate the effectiveness of compensating spatially-varying gaze errors. The black border indicates the screen border, the green dots are the targeted gaze positions and the red dots are the gaze prediction. For two target positions, there are no samples recorded due to eye blink or the occlusion of eye lid. Comparing Figure 13 (a) and (b), we observe that the predicted gaze points seem to have a curved pattern in (a). The curved pattern is known as the spatially-varying errors due to the non-coplanarity of pupil center and the glint plane and cannot be compensated using a single homography transformation. Our method "rectifies" the gaze prediction through the use of adaptive homography predicted by the pupil center position.
The goal of the second experiment is to validate the ability to compensate errors induced from head movements. Because both CR-HOM [Hansen et al. 2010] and CR-AH are robust with regards to head pose changes parallel to the screen plane, we show the comparison on accuracy as a function of depth variations. Figure 14



(a) CR-HOM [Hansen et al. 2010] (b) CR-AH (Ours)

Figure 13: *Qualitative comparison on calibration accuracy.*

shows the gaze prediction accuracy under head movements. The subject is calibrated at 500mm and 600mm away from the screen, respectively. The gaze prediction accuracy are then evaluated for data at three depth positions: 500mm, 600mm, and 700mm. Both methods achieve the lowest errors at the calibration positions. The performance degrades as the subject moves away from the calibration. However, we can see that the adaptive homography is more robust to the head pose changes than CR-HOM. For example, when the subject (calibrated at 600mm) moves to 500mm, our adaptive homography reduces 40% of the errors from CR-HOM. One interesting observation from Figure 14 is that both methods work poorly and ours is a little bit worse when the subjects are located at 700 mm. We attribute this to the insufficient camera resolution for detecting pupil center and glints position reliably. This result bears some similarity as in the noise analysis experiments using simulation (Figure 9 and 10). For example, in Figure 9(c), all the gaze tracking algorithms perform poorly in the high noise level when the head positions are too far from the camera. The unreliable input to our adaptive homography model in the very low resolution case apparently does not produce improvement.



(a) calibrated at 500 mm (b) calibrated at 600 mm

Figure 14: *Averaged gaze estimation errors under head movements.*

5 Conclusion

We have introduced a learning-based adaptation for simultaneously improving the accuracy and robustness of CR-based gaze tracking methods. The adaptive homography simultaneously compensates both spatially-varying gaze errors and head pose dependent errors in a unified framework. Through learning from simulation, we can effectively model how the bias-correcting homography should change depending on the head pose and gaze direction without going through an additional subject-dependent calibration procedure. We have validated the effectiveness of the adaptation using both simulated data and real data from a physical setup.

Acknowledgment

The work presented in this paper was conducted mostly at Microsoft Research during the first author's internship. Additional support of the Office of Naval Research under grant N00014-12-1-0259 is gratefully acknowledged.

References

CERROLAZA, J. J., VILLANUEVA, A., AND CABEZA, R. 2008. Taxonomic study of polynomial regressions applied to the calibration of video-oculographic systems. In *Proceedings of the symposium on Eye tracking research & applications*, 259–266.

CERROLAZA, J. J., VILLANUEVA, A., VILLANUEVA, M., AND CABEZA, R. 2012. Error characterization and compensation in eye tracking systems. In *Proceedings of the Symposium on Eye Tracking Research & Applications*, 205–208.

COUTINHO, F. L., AND MORIMOTO, C. H. 2006. Free head motion eye gaze tracking using a single camera and multiple light sources. In *Computer Graphics and Image Processing*, 171–178.

COUTINHO, F. L., AND MORIMOTO, C. H. 2010. A depth compensation method for cross-ratio based eye tracking. In *Proceedings of the Symposium on Eye-Tracking Research & Applications*, 137–140.

COUTINHO, F. L., AND MORIMOTO, C. H. 2012. Augmenting the robustness of cross-ratio gaze tracking methods to head movement. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, 59–66.

DUCHOWSKI, A. T. 2007. *Eye tracking methodology: Theory and practice*, vol. 373. Springer.

GUESTRIN, E. D., AND EIZENMAN, M. 2006. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering 53*, 6, 1124–1133.

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 32*, 3, 478–500.

HANSEN, D. W., AGUSTIN, J. S., AND VILLANUEVA, A. 2010. Homography normalization for robust gaze estimation in uncalibrated setups. In *Proceedings of the Symposium on Eye-Tracking Research & Applications*, 13–20.

HENNESSEY, C., NOUREDDIN, B., AND LAWRENCE, P. 2006. A single camera eye-gaze tracking system with free head motion. In *Proceedings of the symposium on Eye tracking research & applications*, 87–94.

HOLMQVIST, K., NYSTRÖM, M., ANDERSSON, R., DEWHURST, R., JARODZKA, H., AND VAN DE WEIJER, J. 2011. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

KANG, J. J., GUESTRIN, E., MACLEAN, W., AND EIZENMAN, M. 2007. Simplifying the cross-ratios method of point-of-gaze estimation. In *Canadian medical and biological engineering conference*.

KANG, J. J., EIZENMAN, M., GUESTRIN, E. D., AND EIZENMAN, E. 2008. Investigation of the cross-ratios method for point-of-gaze estimation. *IEEE Transactions on Biomedical Engineering 55*, 9, 2293–2302.

MODEL, D., AND EIZENMAN, M. 2010. An automatic personal calibration procedure for advanced gaze estimation systems. *IEEE Transactions on Biomedical Engineering 57*, 5, 1031–1039.

MORIMOTO, C. H., AND MIMICA, M. R. 2005. Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding 98*, 1, 4–24.

PANTIC, M., PENTLAND, A., NIJHOLT, A., AND HUANG, T. S. 2007. Human computing and machine understanding of human behavior: a survey. 47–71.

YOO, D. H., AND CHUNG, M. J. 2005. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *Computer Vision and Image Understanding 98*, 1, 25–51.

YOO, D. H., KIM, J. H., LEE, B. R., AND CHUNG, M. J. 2002. Non-contact eye gaze tracking system by mapping of corneal reflections. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 94–99.

ZHAI, S., MORIMOTO, C., AND IHDE, S. 1999. Manual and gaze input cascaded (magic) pointing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 246–253.