# Improving Cross-Ratio-Based Eye Tracking Techniques by Leveraging the Binocular Fixation Constraint

Zhengyou Zhang[*]  Qin Cai[†]
Microsoft Research, One Microsoft Way, Redmond, WA, USA

## Abstract

The cross-ratio approach has recently attracted increasing attention in eye-gaze tracking due to its simplicity in setting up a tracking system. Its accuracy, however, is lower than that of the model-based approach, and substantial efforts have been devoted to improving its accuracy. Binocular fixation is essential for humans to have good depth perception, and this paper presents a technique leveraging this constraint. It is used in two ways: First, in estimating jointly the homography matrices for both eyes, and second, in estimating the eye gaze itself. Experimental results with both synthetic and real data show that the proposed approach produces significantly better results than using a single eye and also better than averaging the independent results from the two eyes.

**CR Categories:** I.3.7 [Computer Vision]: Eye Tracking—Binocular;

**Keywords:** eye gaze tracking, cross-ratio, binocular, fixation, computer vision

## 1 Introduction

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]. It has been extensively studied for the past few decades, and the reader is referred to [Hansen and Ji 2010] for a recent review.

Cross-ratio (CR) based approaches [Yoo et al. 2002] 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. Special efforts have been made to correct the estimation bias induced from the simplification assumptions. The bias compensation is usually achieved through subject-dependent calibration, i.e., asking subjects to look at a few predefined calibration targets on the screen. A 2D planar transformation is then computed between the predicted gaze locations by the basic CR-based technique and the calibration targets. One noticeable work is the homography-based correction [Kang et al. 2007; Hansen et al. 2010].

In this paper, we investigate how to improve CR-based techniques by leveraging the binocular fixation constraint. Binocular fixation is the process of having both eyes directed at the same object at the same time, which is essential for good depth perception. When a user directs her gaze to a point on the screen, she in general fixes at the point, i.e., the gaze directions of the two eyes converge toward the same point on the screen. We are aware that some people may have uncoordinated motions for two eyes due to heterotropia or occlusions. The technology developed in this paper are for people who can converge their eyes at the screen with un-obstructed view. In the case when the two eyes do land at different locations on the screen, portion of the technology which jointly estimates the CR homography matrices from the lights to glints in the images is still applicable. On the other hand, although the dominant eye may play an essential role for gaze tracking, it is better to be aided by detecting the gaze direction of the non-dominant eye when the information from the dominant eye is unavailable. This constraint could be especially useful when depth perception (3D eye gaze) is needed in an application.

The remaining of the paper is organized as follows. First, we briefly describe the homography-based correction technique, on which our technique is based. Second, assuming the system has already been calibrated, we describe how eye gaze can be estimated from both eyes under the binocular fixation constraint. Third, we present our gaze calibration process using both eyes under the binocular fixation constraint. Fourth, results with simulation data and real data are described.

## 2 Overview of the Homography-Based Correction Technique

The homography-based CR method, illustrated in Figure 1, works as follows. Four or more infrared (IR) lights, $L_i$, are positioned

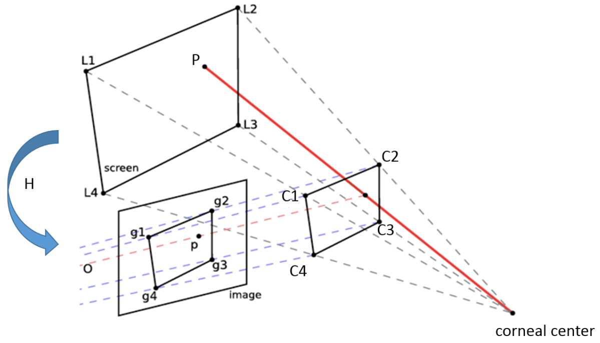

**Figure 1:** *A basic paradigm of Homography-Based CR methods, adapted from [Coutinho and Morimoto 2013].*

[*] e-mail: zhang@microsoft.com
[†] e-mail: qincai@microsoft.com

ETRA 2014, March 26 – 28, 2014, Safety Harbor, Florida, USA.
2014 Copyright held by the Owner/Author.
ACM 978-1-4503-2751-0/14/03

around a screen, where $i = 1, \dots, N$ and $N \geq 4$. They are coplanar. A camera looks at the user, and captures images of the user's eyes. The corneal reflection of each $L_i$ on the eye is called a glint, denoted by $C_i$. In the captured image, the detected glints are denoted by $\mathbf{g}_i$. We also denote by $\mathbf{p}$ the center of the detected pupil in the image.

The CR method assumes that glints $C_i$'s are coplanar. Then, there exists a plane projectivity, expressed as a homography matrix, between $L_i$'s and $C_i$'s, and their exists another plane projectivity between $C_i$'s and $\mathbf{g}_i$'s. The composition of two homography matrices is still a homgraphy matrix, so $L_i$'s and $\mathbf{g}_i$'s are related by a homography matrix, called the CR homography $\mathbf{H}$, such that

$$\mathbf{g}_i = \mathcal{H}(\mathbf{H}L_i) \ , \tag{1}$$

where $\mathcal{H}(\mathbf{x})$ converts a homogeneous vector $\mathbf{x}$ into a 2D vector, and $L_i$ should be in homogeneous coordinates, which is actually a 3D vector with the first two elements equal to $L_i$ and the last element equal to 1. In the remaining of the paper, we use the same notation for 2D vectors and their corresponding homogeneous vectors when there is no ambiguity.

Once the mapping between $L_i$ and $\mathbf{g}_i$ is known, the pupil center $\mathbf{p}$ can be mapped to the screen coordinate space, which is given by

$$P = \mathcal{H}(\mathbf{H}^{-1}\mathbf{p}) \ . \tag{2}$$

However, it is well-known that the eye optical axis, defined by the vector from the corneal center to the pupil center on the eye, is not the same as the visual axis, which defines the actual gaze. The angle between the two axes is person-dependent. Therefore, a subject-dependent calibration is necessary to correct this bias.

To perform the bias correction, a user is asked to look at a few points on the screen, which are considered as the ground truth of the gaze points, denoted by $G_j$, with $j$ as the index to the calibration points. For each gaze point, we detect the glints in the image, denoted by $\mathbf{g}_{ij}$, and the pupil center, denoted by $\mathbf{p}_j$. From $\mathbf{g}_{ij}$'s, we can compute the CR homography, $\mathbf{H}_j$, from (1). In return, the pupil center $\mathbf{p}_j$ is mapped to the screen space as $P_j = \mathcal{H}(\mathbf{H}_j^{-1}\mathbf{p}_j)$ according to equation (2). Given a set of $P_j$'s and its corresponding ground truth gaze points $G_j$'s, the homography-based CR method models the differences by a homography mapping as

$$P_j = \mathcal{H}(\mathbf{H}_b G_j) \ , \tag{3}$$

where $\mathbf{H}_b$ is the bias-correction homography matrix.

During actual eye tracking, at each frame, we are given a set of glints $\mathbf{g}_i$'s and the pupil center $\mathbf{p}$, and we compute the CR homography matrix $\mathbf{H}$ according to (1). By combining (2) and (3), the gaze is then given by

$$G = \mathcal{H}(\mathbf{H}_b^{-1}\mathbf{H}^{-1}\mathbf{p}) \ . \tag{4}$$

## 3 Gaze Estimation from Both Eyes Under the Binocular Fixation Constraint

Now let us consider both eyes. The geometry of the gaze estimation from both eyes is illustrated in Figure 2.

The glints between the left and right eyes are related by $\mathbf{H}^{LR}$, which is given by

$$\mathbf{H}^{LR} = \mathbf{H}^R{\mathbf{H}^L}^{-1} \ . \tag{5}$$

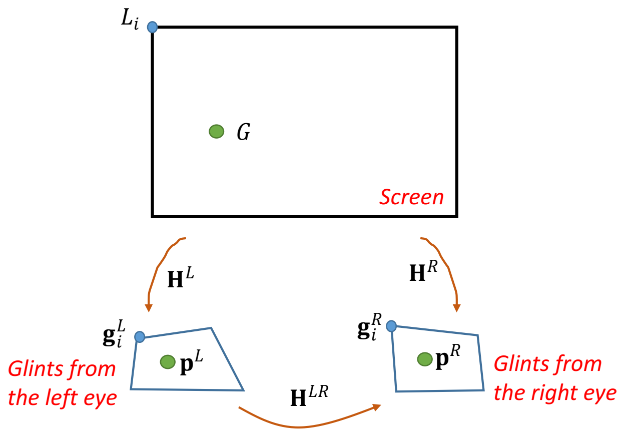

**Figure 2:** *Geometry of the IR lights and the glints in the left and right eyes*

To estimate $\mathbf{H}^L$ and $\mathbf{H}^R$, we need to solve the following problem:

$$\min_{\mathbf{H}^L,\mathbf{H}^R} \sum_i \Big( \|\mathbf{g}_i^L - \mathcal{H}(\mathbf{H}^L L_i)\|^2 + \|\mathbf{g}_i^R - \mathcal{H}(\mathbf{H}^R L_i)\|^2 + \frac{1}{2}\|\mathbf{g}_i^R - \mathcal{H}(\mathbf{H}^R{\mathbf{H}^L}^{-1}\mathbf{g}_i^L)\|^2 + \frac{1}{2}\|\mathbf{g}_i^L - \mathcal{H}(\mathbf{H}^L{\mathbf{H}^R}^{-1}\mathbf{g}_i^R)\|^2 \Big) \ . \tag{6}$$

The last two items provide constraints on the left and right homographies, $\mathbf{H}^L$ and $\mathbf{H}^R$, making them more robust to noise. The coefficient 1/2 is used to compensate for double use of pair $(\mathbf{g}_i^L, \mathbf{g}_i^R)$ in two directions.

We derive the gaze point on the screen by expanding equation (4) for both eyes

$$G^L = \mathcal{H}({\mathbf{H}_b^L}^{-1}{\mathbf{H}^L}^{-1}\mathbf{p}^L) \ , \tag{7}$$

$$G^R = \mathcal{H}({\mathbf{H}_b^R}^{-1}{\mathbf{H}^R}^{-1}\mathbf{p}^R) \ . \tag{8}$$

Note that $\mathbf{H}_b^L$ and $\mathbf{H}_b^R$ are left- and right-eye bias-correction homography matrices respectively, which adjusts the disparity from the gaze location to the mapped pupil location. In the case when the user is fixating at the same point at the screen, i.e., $G^L = G^R = G$, we can jointly estimate the gaze point by minimizing the reprojection error between the pupil centers in the image and their corresponding estimations from the gaze point using both the CR and bias-correction homography matrices, i.e.,

$$\min_G \|\mathbf{p}^L - \mathcal{H}(\mathbf{H}^L\mathbf{H}_b^L G)\|^2 + \|\mathbf{p}^R - \mathcal{H}(\mathbf{H}^R\mathbf{H}_b^R G)\|^2 \ . \tag{9}$$

Note, here we assume that the noise in the left and right pupil locations is independent, isotropic, and identically distributed.

## 4 Eye Gaze Calibration Using Both Eyes Under the Binocular Fixation Constraint

In the last section, we assumed the bias-correction homography matrices, $(\mathbf{H}_b^L, \mathbf{H}_b^R)$, for the left and right eyes are available. Now, we discuss how to estimate them using both eyes.

The user looks at $M$ gaze positions on the screen: $\{G_j | j = 1, \dots, M\}$. For each gaze position, we have a set of glints (indexed by $i$) for the left and right eyes, $\{(\mathbf{g}^L_{ij}, \mathbf{g}^R_{ij})|i = 1, \ldots, N; j = 1, \ldots, M\}$ as well as pupil locations of the left and right eyes, $\{(\mathbf{p}^L_j, \mathbf{p}^R_j)|j = 1, \ldots, M\}$.

The calibration consists of three steps.

The first step is to estimate $(\mathbf{H}^L_j, \mathbf{H}^R_j)$ for each gaze position $j$ using equation (6).

The second step is to compute the mapped pupil locations in the screen space: $(P^L_j, P^R_j)$. That is:

$$P^L_j = \mathcal{H}(\mathbf{H}^{L^{-1}}_j \mathbf{p}^L_j), \quad \text{and} \quad P^R_j = \mathcal{H}(\mathbf{H}^{R^{-1}}_j \mathbf{p}^R_j). \qquad (10)$$

The final step is to estimate the bias-correction homography matrices $(\mathbf{H}^L_b, \mathbf{H}^R_b)$ separately by minimizing the distances between the gaze positions and the corresponding mapped pupil locations in the screen, i.e.,

$$\min_{\mathbf{H}^L_b} \sum_j \|P^L_j - \mathcal{H}(\mathbf{H}^L_b G_j)\|^2, \qquad (11)$$

$$\min_{\mathbf{H}^R_b} \sum_j \|P^R_j - \mathcal{H}(\mathbf{H}^R_b G_j)\|^2. \qquad (12)$$

Note, here we assume no noise in $G_j$'s. The noise in the mapped pupil locations is independent, isotropic, and identically distributed.

## 5 Evaluation using Simulated Data

**Screen and camera model.** The simulated screen is of size 400 mm × 300 mm. We define $z = 0$ at the screen plane, and assume the origin of the world coordinate is at the center of the screen. In both simulation and real experiments, we adopted a right handed coordinate system with the $x$ and $y$ axes corresponding to the horizontal and vertical direction in the display plane, with $x$ axis points at right, $y$ axis up, and $z$ axis towards the user.

The eight 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 slighted below the screen. The camera has a 13mm focal length with sensor size (7.18mm×5.32mm) to allow large head movement without reposition the camera. The horizontal field of view is approximately $30.87°$. The captured image resolution is assumed to have $1280 \times 1024$ pixels.

**Calibration position and eye parameters.** The head position at calibration is located at $(0, 0, 600)$ mm, where $(0, 0, 0)$ is the center of the screen. We simulate the eye model by using the typical eye parameters listed in [Guestrin and Eizenman 2006]. The corneal is modeled as a sphere with a radius of 7.8 mm and the distance from pupil center on the eye to the corneal center is 4.2mm. The effective index of refraction is set to 1.3375. Both eyes are used for evaluation. The visual deviation between visual axis and optical axis is $5°$ and $-5°$ for left and right eyes in the horizontal direction and $1.5°$ in the vertical direction. The distance between the two eye ball rotation centers is 64 mm. In the simulation, given a fixed head position, we rotate along the eyeball center (not the corneal center) so that the visual axis intersects the target gaze point on the screen.

**Results with different head poses.** We show the results with bias correction calibrated at (0, 0, 600) but applied to different head positions in Table 1. Each test reports the average result of 30 trials with random noise added to the glints and to the pupil center. The noise for the glints is zero mean and 0.5 pixel standard deviation, the pupil center is zero mean and 0.25 pixel standard deviation, smaller than the glint noise because we expect that ellipse fitting gives the pupil center more accurate result. Without surprise, we get the best evaluation results at the calibration location of (0, 0, 600) mm shown from the results. As the head moves away from the calibration point, the error of gaze tracking increases. The further the distance between the head and screen, the bigger the error is. However, all the results provide a statistically convincing conclusion that the use of both eyes ("Binocular") reduces the gaze errors by almost 40% compared with using one eye ("Left Eye" or "Right Eye"). We also provide the results by averaging the independent estimations from the two eyes, as shown in Column "Averaging". The joint "Binocular" method shows some marginal improvement over "Averaging". The improvement is more noticeable with real data, to be described in the next section.

**Table 1:** *Gaze errors in degrees with simulated data. Calibration is done at (0, 0, 600) mm.*

| Head Pose | Left Eye | Right Eye | Averaging | Binocular |
|---|---|---|---|---|
| (-100, 0, 500) | 1.1801 | 1.5924 | 1.0770 | 1.0427 |
| (0, 0, 500) | 1.3405 | 1.2947 | 0.9396 | 0.9396 |
| (100, 0, 500) | 1.6382 | 1.2489 | 1.0541 | 1.0198 |
| (-100, 0, 600) | 1.1649 | 1.1744 | 0.8594 | 0.8498 |
| (0, 0, 600) | 1.0694 | 1.0694 | 0.7830 | 0.7734 |
| (100, 0, 600) | 1.1935 | 1.2126 | 0.8785 | 0.8594 |
| (-100, 0, 700) | 1.4566 | 1.3258 | 0.9248 | 0.9330 |
| (0, 0, 700) | 1.3258 | 1.3258 | 0.9003 | 0.8921 |
| (100, 0, 700) | 1.3503 | 1.4812 | 0.9576 | 0.9494 |
| *average* | 1.3021 | 1.3028 | 0.9305 | 0.9177 |

The homography-based CR technology has certain robustness against head poses within the neighborhood of the calibration position, as shown in Table 1. Works in [Coutinho and Morimoto 2013] described a few close-form solutions regarding adjustment on calibrations with known head pose change. We also developed our own technology in detecting head pose change based on glint patterns and then adapting calibration parameters accordingly. Discussion on these technologies is beyond the scope of this paper, and they are complementary to the use of binocular fixation constraint described in this paper and should be used together in practice.

## 6 Experiments with Real Data

For real data evaluation, we use 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 offset of the IR lights at the corner is the same as in simulation, i.e., 27 mm horizontally and 29 mm vertically. The purpose of using 8 IR LED lights instead of 4 is to increase the robustness of the system by having at least 4 or more glints. We use a CMOS point grey camera with 1/1.8" sensor and resolution 1280×1024 pixels and a Fujinon Lens with focal length set to be 13mm. At a distance of 600mm, the glints occupy an area of only 16×12 pixels. The camera is located around 50 mm below the middle point of the bottom edge of the screen and 150 mm away from the screen plane. The subject's left eye is roughly located along the axis perpendicular to the screen center. A chin rest was used to support the subject's head during recording eye images. However, there are still inevitable head movements during calibration and testing. Figure 3 shows our actual experimental system. As one notices, we have in total 24 IR LEDs to potentially test out different locations of IR lighting in the future. For this experiment we only use 8 of them. Also, we have two cameras in our setup, but only one is used for this paper. In the future, we will explore other possibilities.

To collect the real data from people, we asked the subjects to fixate at a uniformly distributed $5 \times 5$ grid on the screen and record

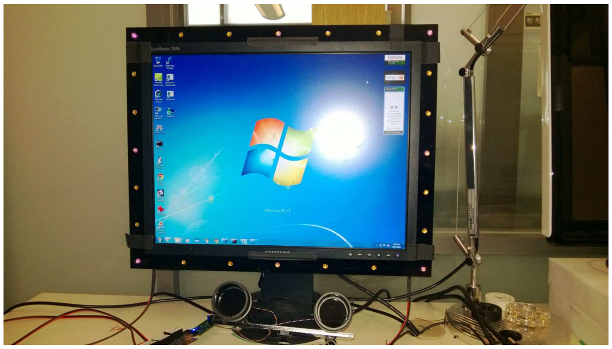

**Figure 3:** *Experimental setup*

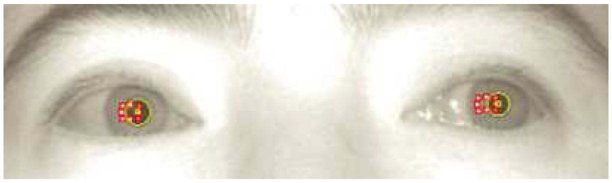

**Figure 4:** *An example of detected glints and pupil center.*

the captured eye images. We gathered maximum 60 samples (2-3 seconds) for each gaze target position. To avoid capturing the eye images during the transition of the eye movement, we only started to record the image "after" the subject clicked the mouse on the dot at the targeted location. After capturing these eye images, we performed pre-processing stage of glint and pupil detection via thresholding and ellipse fitting. Figure 4 shows an example of detected glints and pupil center.

We evaluated the algorithm on the data collected from the system configuration described above. Since both eyes need to be present at various head positions, we are only able to capture up to 5 locations due to the field of view of the camera. The result shown in Table 2 demonstrates the improvement with the proposed algorithm by about 18.4% over using only the left eye, by about 59.4% over using the poorly posed right eye, and by about 7.7% over the commonly used "Averaging" method.

Table 3 shows another example with calibration done at a different location. Again, it shows a relative error reduction by about 22.2% over using only the single left eye, by about 52.7% over using the right eye, and by about 7.0% over the "Averaging" method.

## 7 Conclusion

In this paper, we have presented an approach to improve the cross-ratio-based eye tracking techniques by leveraging the binocular fixation constraint. Two techniques are shown: One is the simple averaging, and the other is based on a maximum likelihood formulation which estimates the CR homography matrices for both eyes jointly. The joint "Binocular" method achieves significant (about 40%) error reduction in simulation, and over 20% with real data, compared with the best single eye. It shows a marginal improvement over the typical "Averaging" method in simulation, but a meaningful (over 7%) improvement for real data. Both experiments with simulation and real data clearly show that the binocular fixation constraint can significantly improve the gaze tracking accuracy if the constraint is valid. Future work includes developing a technique to detect whether the user has heterotropia in order to decide whether the proposed technique should be applied.

**Table 2:** *Gaze errors in degrees with real data. Calibrated at (0, 50 mm, 600 mm), but evaluated at various head poses*

| Head Pose | Left Eye | Right Eye | Averaging | Binocular |
|---|---|---|---|---|
| (0 50 600) | 0.33 | 0.41 | 0.32 | 0.30 |
| (0 50 500) | 0.78 | 0.84 | 0.65 | 0.60 |
| (0 50 700) | 0.71 | 0.72 | 0.53 | 0.51 |
| (-80 50 600) | 0.79 | 1.84 | 1.02 | 0.87 |
| (50 50 600) | 0.54 | 0.87 | 0.54 | 0.51 |
| *average* | 0.63 | 0.94 | 0.61 | 0.56 |

**Table 3:** *Gaze errors in degrees with real data. Calibrated at (0, 50 mm, 500 mm), but evaluated at various head poses*

| Head Pose | Left Eye | Right Eye | Averaging | Binocular |
|---|---|---|---|---|
| (0 50 500) | 0.36 | 0.83 | 0.49 | 0.41 |
| (0 50 600) | 0.71 | 0.53 | 0.56 | 0.53 |
| (0 50 700) | 0.79 | 0.75 | 0.57 | 0.56 |
| (-80 50 600) | 0.84 | 1.73 | 0.99 | 0.92 |
| (50 50 600) | 0.72 | 0.88 | 0.64 | 0.63 |
| *average* | 0.68 | 0.94 | 0.65 | 0.61 |

## References

COUTINHO, F. L., AND MORIMOTO, C. H. 2013. Improving head movement tolerance of cross-ratio based eye trackers. In *International Journal on Computer Vision*, vol. 101, 259–481.

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. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 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 2010 Symposium on Eye-Tracking Research & Applications*, ACM, 13–20.

KANG, J. J., GUESTRIN, E., MACLEAN, W., AND EIZENMAN, M. 2007. Simplifying the cross-ratios method of point-of-gaze estimation. In *30th Canadian medical and biological engineering conference (CMBEC30)*.

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.

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., 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*, IEEE, 94–99.

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