A Robust 3D Eye Gaze Tracking System using Noise Reduction

Jixu Chen* Yan Tong Wayne Gray† Qiang Ji‡
Rensselaer Polytechnic Institute, Troy, NY 12180-3590, USA

Abstract

This paper describes a novel real-time 3D gaze estimation system. The system consists of two cameras and two IR light sources. There are three novelties in this method. First, in our system, two IR lights are mounted near the centers of the stereo cameras, respectively. Based on this specific configuration, the 3D position of the corneal center can be simply derived by the 3D reconstruction technique. Then, after extracting the 3D position of the "virtual pupil" correctly, the optical axis of the eye can be obtained directly by connecting the "virtual pupil" with the corneal center. Second, we systematically analyze the noise in our 3D gaze estimation algorithm and propose an effective constraint to reduce this noise. Third, to estimate the user-dependent parameters (i.e. the constraint parameters and the eye parameters), a simple calibration method is proposed by gazing at four positions on the screen. Experimental results show that our system can accurately estimate and track eye gaze under natural head movement.

Keywords: Gaze estimation, stereo cameras, noise reduction

1 Introduction

Gaze tracking is the procedure of determining the point-of-gaze in the space, or the visual axis of the eye. Gaze tracking systems are primarily used in the Human Computer Interaction (HCI) and in the analysis of visual scanning pattern. In HCI, the eye gaze can be served as an advanced computer input [Jacob 1991] to replace the traditional input devices such as mouse pointer [Zhai et al. 1999]. Also, the graphic display on the screen can be controlled by the eye gaze interactively [Zhu and Ji 2004]. Since the visual scanning patterns are closely related to the the person's attentional focus, the cognitive scientists use the gaze tracking system to study human's cognitive processes [Liversedge and Findlay 2000],[Mason et al. 2004].

In general, the eye gaze estimation algorithms can be classified into two groups: 2D mapping based gaze estimation methods and 3D direct gaze estimation methods. For the 2D mapping-based gaze estimation methods, the eye gaze is estimated from a mapping function by inputting a set of 2D eye features which are extracted from the eye images. For example, the widely used Pupil Center Corneal Reflection(PCCR) technique ([Huchinson et al. 1989], [Morimoto and Mimica 2005], [LC 2005], [ASL 2006], [SM 2007]) is based on the relative position between the centers of corneal reflection (glint) generated by the light source and the pupil. After the pupil and the glint are extracted from the image, the 2D pupil-glint vector is mapped to the gaze point on the screen by a mapping function. PCCR technique has been proved to be the most popular and has the advantage over the other methods in that the eye features can be easily and robustly extracted [C.H.Morimoto et al. 2000].

However, most of existing eye gaze tracking systems based on the 2D PCCR method have two common drawbacks: first, because the mapping function is person-dependent, the user has to perform complex experiment to calibrate the parameters of the mapping functions. For example, in the calibration procedure of [LC 2005], the subject need to gaze at 9 evenly distributed points on the screen or gaze at 12 points for higher accuracy. Second, if the head moves away from the original position where the user performed the gaze calibration, the accuracy of the these gaze tracking systems drops dramatically. In [Morimoto and Mimica 2005], they report detailed data showing how the calibrated gaze tracking systems decay as the head moves away from original position. So, the user has to keep his head unnaturally still.

In order to improve the tolerance of the head movement, some methods have been proposed to adapt the 2D mapping function to different head positions. In [Zhu and Ji 2004], the head position is included as an input into the mapping function implicitly, but low accuracy around 5^o is achieved. In [Zhu and Ji 2005], a complicated model is proposed to eliminate the head motion effect on the gaze mapping function so that the 2D mapping-based method can work under natural head movement with a much better accuracy, around 1^o.

Different from the 2D mapping-based gaze estimation, the 3D gaze estimation method is based on the structure of the eyeball and directly extracts the 3D direction of the gaze (visual axis). In our proposed method, the 3D eye features (the corneal center and the pupil center) can be estimated directly by the 3D reconstruction technique. Then, we propose a method to effectively reduce the error in the 3D eye features. Finally, the visual axis is estimated from the refined 3D features, and the gaze point on the screen is directly obtained by intersecting the visual axis with the screen. Since this method is not constrained by the head position, the complicated head motion elimination model can be avoided.

2 Related Work

Several techniques have been proposed to estimate the 3D direction of gaze directly, such as [Zhu and Ji 2007], [Morimoto et al. 2002], [Beymer and Flickner 2003], [Wang and Sung 2002], [Shih and Liu 2004], [Guestrin and Eizenman 2006].

A simple method for estimating eye gaze under free head movement, and without calibration is suggested by Morimoto et al [Morimoto et al. 2002]. They use one camera and two IR light sources : one light source is used to generate the bright pupil image, and the other one is used to generate the dark pupil image. Because the corneal surface can be modeled as a sphere convex mirror, by assuming the paraxial rays from the light sources, it is possible to compute the 3D corneal center. A set of user dependent parameters are used in this method, but they don't give a method to estimate these parameters. Furthermore, no working system is built from the proposed technique. Only the accuracy of about 3^o is reported using synthetic images.

In [Zhu and Ji 2007], they also assume the sphere convex corneal

*email: chenj4@rpi.edu
†email: grayw@rpi.edu
‡email: qji@ecse.rpi.edu

ETRA 2008, Savannah, Georgia, March 26–28, 2008.
© 2008 ACM 978-1-59593-982-1/08/0003 $5.00

surface and the paraxial rays. They use a set of stereo cameras and two IR light sources with known positions. First, they compute the 3D positions of two "virtual lights" inside the cornea. Then the "virtual lights" are connected with the actual lights, respectively. The corneal center position is derived from the intersection of the two lines. This method allows free head movement. However, its accuracy drops very fast when the subject leaves away from the camera and the image resolution decreases as a result.

[Shih and Liu 2004] propose a novel method to estimate the 3D gaze direction by using multiple cameras and multiple light sources. In their method, there is no need to know the user-dependent parameters of the eye. However, because the glints and the pupil center need to be extracted very accurately to reduce the noise, they use very narrow view zoom-in cameras to focus on the eye and get high resolution eye image. However, this will limit the space of head movement.

To allow free head movement, some 3D gaze estimation systems combine the wide view face camera with the narrow view eye camera. For example, Wang and Sung [Wang and Sung 2002] combine a face pose estimation system with a narrow FOV zoom-in camera (focal length=55mm) to compute the gaze direction. In [Beymer and Flickner 2003], Beymer and Flicker use a more complicated system, which includes two sets of stereo system. One wide angle stereo system for head detection, and one narrow FOV stereo system for high resolution eye tracking. After a set of eye image features are extracted, a complicated 3D eye model is fitted to these features via a nonlinear estimation technique. However, this numerical fitting process is very complicated in computation.

In [Guestrin and Eizenman 2006], they summarize the previous 3D estimation methods and give a general mathematical model for gaze estimation system that utilized the estimates of the centers of the pupil and one or more glints. However, in their mathematical model, there are many non-linear equations. They solve these equations with numerical method. It is not only complicated but also unstable when there is some noise.

In summary, most of the existing 3D gaze tracking techniques have the following limitations. (1) First, because the 3D gaze estimation is very sensitive to the image noise, it need to extract the eye features very accurately. Most of these systems use zoom-in cameras to capture high-resolution eye images. However, this narrow FOV camera will limit the head movement. Although another wide FOV system can be used to control the eye camera to allow a larger head movement, the system is very complicated. (2) Second, most of the 3D gaze algorithms need to solve non-linear equations. The numerical solutions for these non-linear equations are usually complicated and sensitive to noise.

In our 3D gaze tracking system, we propose to use a simple stereo camera system with 8mm lens to get the image of the whole face, so that the head can move in a large region without losing the eye. In addition, we located two IR lights near the camera centers. This system configuration not only allows easy pupil and glint detection due to the bright/black pupil effect, but also simplify the equations to estimate the 3D corneal center. Finally, to accurately estimate the 3D virtual axis, we impose a constraint to refine the extracted 3D eye features from the stereo camera system. We will show that the constraint can effectively improve the final gaze estimation result.

This new system is an extension to [Zhu and Ji 2007]. Compared with the previous work, our new system has three improvements : (1) Based on our special configuration of the lights and the cameras, the 3D corneal center position can be derived directly by linear triangulation method. (2) In previous work [Shih and Liu 2004], the noise analysis of the 3D point estimation was proposed. However, they didn't analyze the subsequent noise in gaze estimation. In our work, we analytically show that even the small noise in 3D point estimation will cause very large noise in gaze estimation. So, we propose a constraint to reduce the noise in gaze estimation. Because of this noise reduction method, our gaze estimation can work even on low-resolution eye image. (3) To estimate the user dependent parameters, a simple 4-point calibration procedure is proposed.

3 3D gaze estimation algorithm

3.1 Eyeball structure

As shown in Figure 1, the eyeball is made up of the segments of two spheres with different sizes [Oyster 1999]. The anterior smaller segment is the cornea. The cornea is transparent, and the pupil is inside the cornea. Optical axis of the eye is defined as the 3D line connecting the center of the pupil ($\mathbf{p}^*$) and the center of the cornea ($\mathbf{c}$). The visual axis is the 3D line connecting the corneal center ($\mathbf{c}$) and the center of the fovea (i.e. the highest acuity region of the retina). Since the gaze point is defined as the intersection of visual axis rather than the optical axis with the scene, the relation between these two axes has to be modeled. The angle between the optical axis and visual axis is named as *kappa*, which is a constant value for each person.

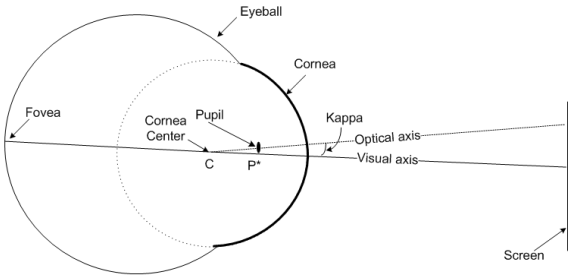

Figure 1: *The structure of the eyeball.*

3.2 Computing the 3D corneal center

Our system is composed of a set of stereo cameras and two IR lights which are mounted near the camera centers, respectively (Figure 2). The reflection ray diagram of our system is shown in Figure 3. (In this diagram, the lights are located on the camera centers.) When light passes through the eye, the sphere surface of the cornea will act like a reflective surface, and the reflection point on the corneal surface is called *glint*.

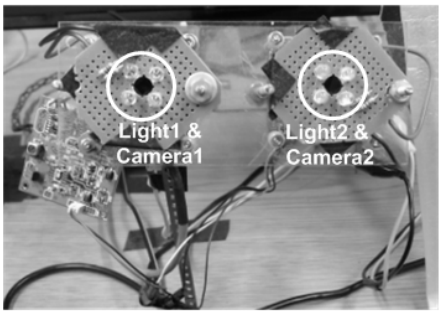

Figure 2: *Our stereo gaze tracking system.*

In Figure 3, $\mathbf{c}$ is the corneal center. $\mathbf{q}_{22}$ and $\mathbf{q}_{11}$ are the corneal reflections (glints) on the corneal surface. $\mathbf{u}_{11}$ and $\mathbf{u}_{22}$ are the glint centers in the image. According to the properties of the convex mirror, an incident ray that is directed towards the center of curvature of a mirror is reflected back along its own path (since it is normally

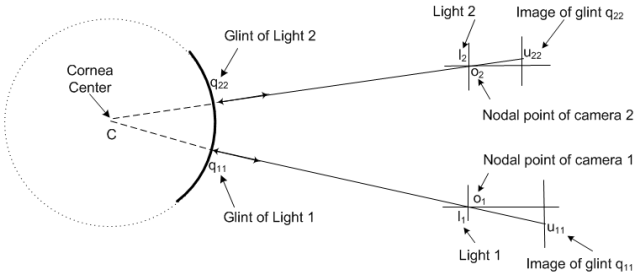

Figure 3: *Ray diagram to estimate Corneal Center. (**c**)*

incident on the mirror). Therefore, as shown in Figure 3, because the two LED light $\mathbf{l}_1$ and $\mathbf{l}_2$ are located at the origin of the camera $\mathbf{o}_1$ and $\mathbf{o}_2$ respectively, the glint image $\mathbf{u}_{11}$ ($\mathbf{u}_{22}$), the origin of the camera $\mathbf{o}_1(\mathbf{o}_2)$ and the curvature center of the cornea $\mathbf{c}$ will be co-linear. Thus, the 3D location of the corneal center $\mathbf{c}$ can be obtained by intersecting the line $\overline{\mathbf{u}_{11}\mathbf{o}_1}$ and $\overline{\mathbf{u}_{22}\mathbf{o}_2}$ as follows:

$$\begin{cases} \mathbf{c} = \mathbf{o}_1 + k_1\overline{\mathbf{u}_{11}\mathbf{o}_1} \\ \mathbf{c} = \mathbf{o}_2 + k_2\overline{\mathbf{u}_{22}\mathbf{o}_2} \end{cases} \quad (1)$$

Actually, $\mathbf{u}_{11}$ and $\mathbf{u}_{22}$ can be seen as the images of the 3D point $\mathbf{c}$ in two cameras. So we can obtain $\mathbf{c}$ using traditional 3D reconstruction techniques. In practice, we use triangulation 3D reconstruction method known as triangulation [Trucco and Verri 1998] to obtain $\mathbf{c}$.

Here, we make an important assumption : the LED light is located at the origin of the camera. This assumption is validated in Appendix B.1.

3.3 Computing the 3D pupil center

As discussed earlier, the optical axis can be obtained by connecting the corneal center $\mathbf{c}$ and the pupil center $\mathbf{p}^*$. However, due to the refraction on the corneal surface, we can only see the virtual image of the pupil ($\mathbf{p}$), instead of the pupil itself ($\mathbf{p}^*$), as shown in Figure 4.

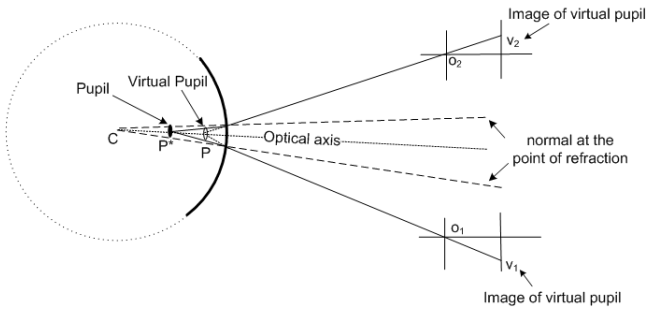

Figure 4: *Ray diagram to estimate virual pupil(**p**) and optical axis.*

$\mathbf{v}_1$ and $\mathbf{v}_2$ are images of the virtual pupil ($\mathbf{p}$) in the two cameras. From these two image, we can also use the 3D reconstruction method to estimate the 3D position of $\mathbf{p}$. Due to the symmetry of the pupil, the virtual pupil ($\mathbf{p}$) is still in the optical axis of the eye. As a result, the virtual pupil ($\mathbf{p}$) and the corneal center ($\mathbf{c}$) can be used to estimate the optical axis directly.

Here, we make another important assumption: the virtual pupil is on the optical axis. This assumption is validated in Appendix B.2.

3.4 Reducing noise in 3D reconstruction

After we obtain the virtual pupil and the corneal center positions, we can connect them to get the optical axis. However the 3D reconstruction method in Section 3.2 and 3.3 is not accurate. There is a noise (≈1mm, see Appendix A.1) in the estimated 3D virtual pupil and corneal center positions. Because the typical distance between the pupil and the corneal center is only 4.2mm ([Guestrin and Eizenman 2006]), 1mm noise will cause significant noise in the estimated optical axis and the subsequent gaze estimation. (Please refer to Appendix A for the detailed noise analysis.) In this section, we will present a method to reduce this kind of 3D reconstruction noise.

In our experiment, the subject is asked to fixate on 9 points on the screen sequentially, and 60 estimates of point-of-gaze are obtained for each fixation point. If we directly compute the virtual pupil and corneal center positions by 3D reconstruction, the result is shown as Figure 5(A). The solid circles are the intended fixation points, and the small crosses are the estimated gaze points. We can see that there is a significant noise for each fixation point, as indicated by a large spread in the estimated gaze position. (Here, eye parameters for this subject are already obtained through the calibration procedure in Section 4.2. And the results of Figure 5(A) and 5(B) are using the same parameters.)

This noise comes from the 3D reconstruction noise. In Appendix A.1, we analytically show that this 3D noise is mainly on the z direction and that it will cause the gaze estimation noise like Figure 5(A).

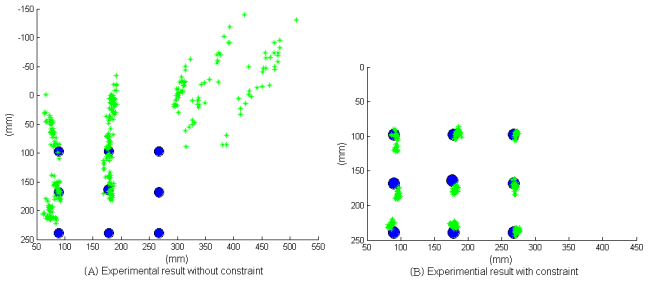

Figure 5: *Gaze estimation result before and after using K constraint. The solid circles are different fixation points and "+" s are the estimated gaze points.*

To minimize the 3D reconstruction error, we impose a constraint on the virtual pupil-cornea distance: we assume this distance is a constant (K) for each person. Then we change the virtual pupil's z-coordinate while keeping x and y coordinates fixed to satisfy this constraint. For example, given the estimated corneal center $\mathbf{c} = (x_c, y_c, z_c)^T$ and the virtual pupil center $\mathbf{p} = (x_p, y_p, z_p)^T$, we can recompute the z-coordinate of $\mathbf{p}$ as follows:

$$z'_p = z_c - \sqrt{K^2 - (x_c - x_p)^2 - (y_c - y_p)^2} \quad (2)$$

Thus, the refined pupil is $\mathbf{p}' = (x_p, y_p, z'_p)^T$.

There are two reasons for setting this constraint:

1. First, for each person, the distance between the corneal center and the pupil center is a constant. Although we actually use virtual pupil instead of pupil in our algorithm, the experiment result in Appendix B.2 shows that the refraction effect will not cause the distance between the corneal center and the virtual pupil change too much (<0.1mm)
2. Second, the noise of 3D reconstruction is mainly on the z-coordinate (see Appendix A.1). So we assume the x and y coordinates are accurate, and thus only refine the z-coordinate.

For example, when we set this distance (K) to be 5.5mm, we see that the noise is reduced effectively as shown in Figure 5.B. The

constraint K is a subject-specific parameter. We will estimate it by a calibration procedure in section 4.2.

3.5 Visual axis estimation

After we use the constraint to refine the optical axis, we try to use the $Kappa$ angle to transfer the optical axis to visual axis. Here, we use the same method as in [Guestrin and Eizenman 2006]. The optical axis is transferred to visual axis by adding a horizontal angle (α) and a vertical angle (β). The two angles are subject dependent, and can be obtained by the calibration process in section 4.2.

4 Parameter Estimation

4.1 System (camera & screen) parameters estimation

Two steps are performed to calibrate the system. First, the parameters of the stereo camera system are obtained through camera calibration [Zhang 2000]. The second step is to obtain the 3D positions of the computer screen. Since the screen is located behind the view of the stereo camera system, it cannot be observed directly by the cameras. Therefore, similar to [Beymer and Flickner 2003], a planar mirror with a set of fiducial markers attached to the mirror surface is utilized. With the help of the planar mirror, the virtual images of the screen reflected by the mirror can be observed by the cameras. Thus, the 3D location of the screen can be calibrated after knowing the virtual image of it.

4.2 Subject-specific eye parameters estimation

The three subject-specific eye parameters (K, α, β) are obtained through a calibration procedure that is preformed once for each subject.

In the calibration procedure, the subject is asked to fixate on 4 evenly distributed reference points that are presented on the screen sequentially (Theoretically, we only need 2 points to do calibration. Because of noise, we finally use 4 points to add redundancy and improve the robustness). During calibration, the subject is not allowed to move head. For each fixation point, 8 estimates of each gaze points are obtained and their median is computed. Using the median gaze points, the three eye parameters are optimized to minimize the error between the reference points on the screen and the estimated gaze points.

But in practice, if we optimize the three parameters together, this non-linear optimization problem is very slow to converge, and it can converge to different local minima. So, to solve this optimization problem more efficiently and robustly, we take the following 3 steps to optimize K and α, β separately.

4.2.1 Step 1: K Calibration

First, we fix both the α and β as zero, and optimize K to minimize the relative distance error.

As shown in Figure 6, the four "o"s indicate the reference points which are displayed on the screen. To show the effect of K, we fix $(\alpha, \beta) = (0, 0)$ and only change the K value from 4.5 to 6.5 (Figure 6 (a)). The "*"s are the estimated gaze points on the screen when using different K values (The subject fixates on each reference point for 8 frames, we just show the "median" gaze point for each reference point.) The estimated gaze points with K= 4.5, 5.1, and 6.5 are indicated. Obviously, when the K increases, the relative distance between the 4 gaze points decreases. So we can select the best K which can keep the true relative distance, as shown in Figure 6 (b)

Let the coordinates of the 4 reference points be $\mathbf{r}_i = (x_r^i, y_r^i), i = 1...4$ and the coordinates of the four estimated gaze points be $\mathbf{p}_i = (x_p^i, y_p^i), i = 1...4$. We optimize K to minimize the relative distance error in Eq. 3. Figure 6(b) shows the selected K, and the estimated gaze points with the selected K.

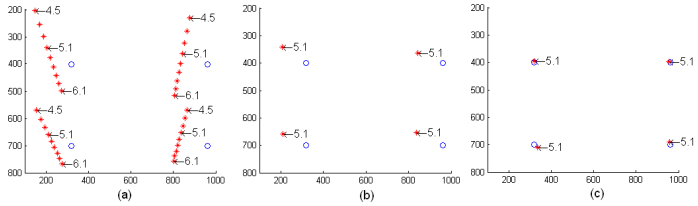

Figure 6: *(a) The estimated points with different K ($(\alpha, \beta) = (0, 0)$). The gaze points when K=4.5, 5.1 and 6.1 are pointed out. (b) The selected K=5.1 can minimize the relative distance error (Eq.3). (c) The estimated gaze points after Kappa calibration. ("o"s show the reference points which are displayed on the screen and "*"s show the estimated median gaze points.)*

$$Err = \sum_{i=2}^{4} \|(\mathbf{r}_i - \mathbf{r}_1) - (\mathbf{p}_i - \mathbf{p}_1)\| \quad (3)$$

4.2.2 Step 2: Kappa(α,β) Calibration

After we obtained K, we just need to optimize α and β to minimize the distance between the gaze points to the reference points:

$$Err = \sum_{i=1}^{4} \|\mathbf{r}_i - \mathbf{p}_i\| \quad (4)$$

In calibration procedure, the distance between the eyeball and the screen is a constant (D). So, adding the small angles α and β will cause the estimated gaze point move an approximately constant distance αD horizontally and βD vertically on the screen.

For example, based on the K value in Figure 6(b), we finally obtain the optimized $\alpha = -3.9^o, \beta = 1.0^o$. Using these angles, we can estimate the gaze points which are shown as "*"s in Figure 6(c). Compared with Figure 6(b), we can see that all the four gaze points undergo almost the same shift.

4.2.3 Step 3: Global Calibration

Finally, we use the estimated K, α and β above as our initial values, and optimize them together according to the objective function in Eq. 4. Because the parameters are already optimized separately, this non-linear optimization procedure can converge to the closest minimum very quickly using simplex search [Lagarias et al. 1998].

Finally, the optimized parameters are K= 5.06, $\alpha = -4.11^o, \beta = 1.22^o$. We see that they are very closed to their initial values. With these parameters, the average error between the estimated gaze points and the reference points is about 13 pixels ($\approx$ 3.7mm). Considering the distance from the subject to the screen (500mm), the calibration error is 0.42^o.

5 Pupil and Glint Tracking

Our gaze tracking system starts with the detection and tracking of the user's pupil, as well as glints. In previous sections, we show that the special setup of our system (two lights located on the camera centers) can simplify the gaze estimation procedure. In this section, we will show that this setup can also make it convenient to perform pupil/glint tracking.

5.1 Pupil Detection and Tracking

Based on the differential lighting scheme [Haro et al. 2000], the pupil can be detected robustly by using the difference image between the dark pupil image and the bright pupil image.

In order to achieve the proposed differential lighting scheme for our stereo camera system, a circuitry has been developed to synchronize the two IR lights with the even and odd fields of the two interlaced images from the two cameras, so that one light can be either on-axis or off-axis light for different cameras.

Specifically, when the left light is on, the even fields of the images are grabbed by the cameras (Figure 7(a) and (b)). Because the left left light is the on-axis light for left camera and the off-axis light for right camera, the even field of the left camera captures a bright pupil image (Figure 7(a)), while the even field of the right camera captures a dark pupil image (Figure 7(b)). On the other hand, when the right light is on, the odd field of the left camera captures the dark pupil image, and the odd field of the right camera captures the bright pupil image.

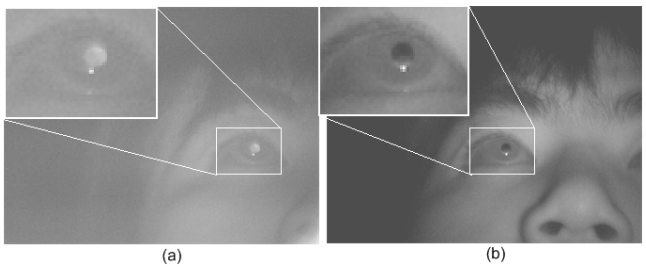

Figure 7: *Even field from stereo cameras (Left light is on): (a) a bright pupil in the even field of the left camera, and (b) a dark pupil in the even field of the right camera.*

Similar to the image subtraction method proposed in [Haro et al. 2000], the pupil blob can be located efficiently in the difference image between the even-field and the odd-field images. In practice, there are some non-pupil blobs in the difference image due to the image noise. Therefore, an appearance-based SVM classification technique [Zhu and Ji 2005] is utilized to identify the pupil blob successfully. Once the pupil blob is detected, the pupil center can be estimated accurately by ellipse fitting.

5.2 Glint Detection and Tracking

Via the proposed differential lighting scheme, the two glints can be effectively separated into different field images. Therefore, unlike most of other methods that will have more than one glints in the eye image [Morimoto et al. 2002; Beymer and Flickner 2003; Shih and Liu 2004], the difficulty of identifying these ambiguous glints can be avoided by our method.

Actually, in our gaze estimation algorithm, we only use the lights' own glints. For example, for the left camera, we only use the glint of the left light. So, we only need to detect the glints in the two bright pupil images from the left and the right cameras, respectively.

6 Gaze Estimation Accuracy

A prototype gaze system is built as shown in Figure 2. This system uses two CCD cameras (MINTRON MTV-03K9HE) with 8mm lenses. And two infrared lights (875nm) are attached on the cameras. The image resolution of the cameras is 640×480 pixels, and our system can run at approximately 25 fps on a PC with a Xeon (TM) 2.80GHz CPU. In order to test the accuracy of the gaze tracking system, we did the following experiment.

First, the 4-point calibration procedure in Section 4.2 is needed for each subject. After the calibration, a marker will display at nine fixed locations on the screen randomly, and the user is asked to gaze at the marker when it appears at each location. The nine marker locations and the gaze points are shown in Figure 8. In order to test the accuracy in different distance. The experiment contains several 1-minute sessions. At each session, the user is required to position his head at a different position. The allowed head movement region is approximately $140 \times 140 \times 220$ mm (width$\times$height$\times$depth).

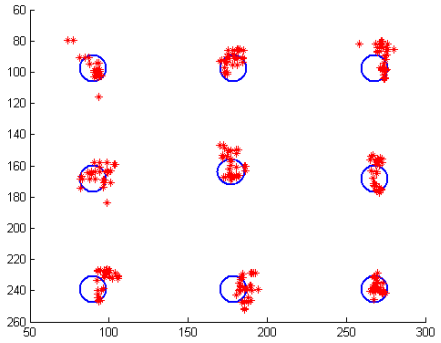

Figure 8: *Experiment result for the first subject. The nine circles (8mm radius) indicate the maker locations on the screen and the estimated gaze points are shown as "*"s. (X,Y axes are in mm)*

6.1 Gaze estimation accuracy under head movement

The gaze estimation result accuracy (RMS error) for the first subject is summarized in Table 1. We can see that the most accurate region is from 380mm to 500mm ($< 1^o$). This is because the depth-of-view of the camera is in this region. When the eye is too close ($<$ 360mm) or too far ($>$ 520mm) from the camera, the eye image will be blurred, and as a result, the glint and the pupil center cannot be extracted accurately. Also, the vertical accuracy is lower than the horizontal accuracy due to lower vertical image resolution. Compared with Zhu et al's 3D method [Zhu and Ji 2007], which is shown in Table 2, we can see that our new 3D gaze estimation algorithm is more accurate in the 380mm-500mm region. The average horizontal and vertical angular accuracies in the whole head movement range are 0.76^o and 0.95^o respectively. In addition, in this experiment, the allowed head movement in the X,Y directions is around 140mm, respectively. So, our system can accurately estimate the eye gaze under natural head movement.

Table 1: *The 3D Gaze Estimation Accuracy for the First Subject*

Distance to the Camera	Horizontal accuracy	Vertical accuracy	Total accuracy
360mm	7.4mm (0.91^o)	9.7mm (1.21^o)	12.2mm (1.5^o)
380mm	4.3mm(0.51^o)	6.1mm(0.73^o)	7.4mm (0.89^o)
410mm	4.1mm(0.46^o)	7.2mm(0.80^o)	8.2mm (0.93^o)
440mm	4.2mm(0.45^o)	6.7mm(0.71^o)	7.9mm (0.84^o)
470mm	8.8mm(0.88^o)	6.6mm(0.67^o)	11mm (1.1^o)
500mm	7.4mm(0.70^o)	7.9mm(0.75^o)	10.8mm (1.0^o)
520mm	12.5mm(1.15^o)	16.2mm(1.45^o)	20.4mm (1.9^o)
580mm	12.1mm(1.03^o)	15.1mm(1.29^o)	19.3mm (1.7^o)

Table 2: *Zhu's 3D gaze estimation result*

Distance to the Camera	Horizontal Accuracy	Vertical accuracy
280mm	5.02mm (0.72^o)	6.40mm (0.92^o)
320mm	7.20mm(0.92^o)	9.63mm(1.22^o)
370mm	9.74mm(1.24^o)	13.24mm(1.68^o)
390mm	12.47mm(1.37^o)	17.30mm(1.90^o)
440mm	19.60mm(1.97^o)	24.32mm(2.45^o)

6.2 Gaze estimation accuracy on different subjects

To estimate the accuracy of the gaze estimation algorithm, we also do the same experiment on 3 other subjects and none of them wears glasses. The average gaze estimation accuracy for each subject is shown in Table 3. In addition, the average horizontal and vertical angular accuracies for all the 4 subjects are 0.77^o and 0.95^o respectively, which is acceptable for many HCI applications, allowing natural head movements.

Table 3: *The Gaze Estimation Accuracy for Four Subjects*

Subject	Horizontal accuracy	Vertical accuracy	Total accuracy
1	0.76^o	0.95^o	1.22^o
2	0.93^o	0.93^o	1.32^o
3	0.68^o	0.74^o	1.0^o
4	0.73^2	1.17^o	1.38^o

7 Conclusion

In this paper, a simple but robust method is proposed to estimate the 3D gaze direction of the user under natural head movement in real time. Via the properties of the convex mirror, we use a special configured stereo camera system to estimate the 3D position of the corneal center and pupil center. To reduce the noise of the estimated 3D position and the subsequent gaze estimation, we propose to use a constraint to refine the result. Compared with other 3D gaze estimation systems, our system can avoid the complicated nonlinear equations and the expensive zoom-in high resolution cameras. After a simple 4 point calibration procedure, accurate eye gaze points can be estimated under natural head movement.

8 Acknowledgement

The work was supported, in part, by grant N000140710033 to Wayne Gray from the Office of Naval Research, Dr. Ray Perez, Project Officer.

A Noise analysis in 3D gaze estimation

Due to the low resolution of the image, the detected glint and the corneal center positions are not accurate. In this section, we will discuss how these positional inaccuracies affect the estimated 3D pupil and corneal center position, and the subsequent gaze estimation.

A.1 Noise analysis in 3D reconstruction

In our algorithm, both the 3D pupil and corneal center are reconstructed by triangulation method (section 3.2 and 3.3). So, first we will discuss the 3D reconstruction noise in this method.

For simplicity, we make some assumptions on our stereo system as shown in Figure 9. In our system, the two cameras are pinhole cameras with the same focus length f. The two image planes are on the same plane and at the same height. So a 3D point projected on the same horizontal scan line in each of the two images. $\mathbf{o}_r$, $\mathbf{o}_l$ are the origins of right and left camera respectively. In our system, we use the right camera coordinate as our 3D coordinate system. b denotes the distance between optical axes of the two cameras and is usually referred to as the *baseline* of the system. $S = (x, y, z)$ is the 3D point. It is projected onto the image coordinates (u_r, v_r) and (u_l, v_l) in right and left camera respectively (Due to our assumption, the vertical coordinates are the same: $v_r = v_l$). The pixel space of the image is δ.

Thus, the 3D coordinates of S can be easily derived as:

$$\begin{cases} x = b\frac{u_r}{u_l - u_r} \\ y = b\frac{v_r}{u_l - u_r} \\ z = b\frac{f}{\delta\bullet(u_l - u_r)} \end{cases} \tag{5}$$

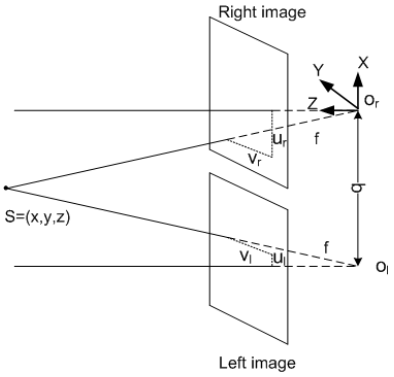

Figure 9: *Stereo camera setup.*

We can simulate the image noise by adding the same independent gaussian noise to u_r, v_r, u_l, v_l respectively. Suppose the gaussian noise has zero mean and variance σ^2. Then, we try to compute the resultant noise in 3D coordinates.

For example, the z coordinate depends only on u_l and u_r. By computing the first-order Taylor expansion, we have:

$$\hat{z} = z + \mathbf{J}_z \begin{pmatrix} du_l \\ du_r \end{pmatrix} \tag{6}$$

where $\mathbf{J}_z$ is the Jacobian matrix of z as $\mathbf{J}_z = \begin{pmatrix} \frac{\partial z}{\partial u_r} & \frac{\partial z}{\partial u_l} \end{pmatrix} = \begin{pmatrix} \frac{bf}{\delta(u_l-u_r)^2} & -\frac{bf}{\delta(u_l-u_r)^2} \end{pmatrix}$

So the noise of z can be presented as its variance (u_r, u_l have independent gaussian noise with variance σ^2):

$$\sigma_z^2 = \mathbf{J}z \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \mathbf{J}_z^T = \frac{2\sigma^2 b^2 f^2}{\delta^2 (u_l - u_r)^4} \tag{7}$$

Similarly, we can compute the noise in x and y coordinates as:

$$\sigma_x^2 = \frac{2\sigma^2 b^2 (u_l^2 + u_r^2)}{(u_l - u_r)^4} \tag{8}$$

$$\sigma_y^2 = \frac{2\sigma^2 b^2 [(u_l - u_r)^2 + 2v_r^2]}{(u_l - u_r)^4} \tag{9}$$

Since the equations in (5) are nonlinear, these expressions do not hold exactly, but we use them as satisfactory approximations.

From Figure 9, we can also present the 2D image using 3D coordinate as follows:

$$\begin{cases} u_r = \frac{f}{\delta z} x \\ u_l = \frac{f}{\delta z}(b + x) \\ v_r = \frac{f}{\delta z} y \end{cases} \tag{10}$$

Combing 7, 8, 9 and 10, we can compute the noise as:

$$\begin{cases} \sigma_z^2 = \frac{2\sigma^2 \delta^2 z^4}{f^2 b^2} \\ \frac{\sigma_x^2}{\sigma_z^2} = \frac{x^2 + (b+x)^2}{2z^2} \\ \frac{\sigma_y^2}{\sigma_z^2} = \frac{b^2 + 2y^2}{2z^2} \end{cases} \tag{11}$$

In our experiment, the distance between two camera is $b = 70mm$, the typical distance between the camera and the subject is about $z = 450mm$, and the allowed head movement in this distance is $x \sim (-105mm, 35mm), y \sim (-70mm, 70mm)$. Thus, the maximum ratio are $\max(\frac{\sigma_x}{\sigma_z}) = 0.17, \max(\frac{\sigma_y}{\sigma_z}) = 0.19$. It means the

noise on the z-axis is much larger than the noise on the x-axis and y-axis. Note that this is the maximum noise ratio. If the 3D point is in other position, the noise on x-axis and y-axis will be even smaller. For example, if the reconstructed 3D point is $S = (-35, 0, 450)$, the noise ratios are $\max(\frac{\sigma_x}{\sigma_z}) = 0.078, \max(\frac{\sigma_y}{\sigma_z}) = 0.11$.

Also, the camera parameter in our system is $f/\delta = 2200$. So, if we add gaussian noise $\sigma = 0.2$, the resultant noise on z-axis is about $\sigma_z = 0.37mm$. (Actually $\sigma = 0.2$ is a very small image noise. In practical, even using sub-pixel pupil and glint detection, we cannot avoid such a small image noise.)

A.2 Noise analysis in gaze estimation

As shown in A.1, only a small $\sigma = 0.2$ gaussian noise in the image will cause the noise in the reconstructed 3D point. In this section, we will show how this 3D reconstruct noise affect the gaze estimation result. To compute the resultant gaze noise exactly, we have to consider the screen position, eyeball position, optical axis, etc. Here, we only give a simple example to show this noise. For simplicity, suppose the subject is looking straight forward, as shown in Figure 10, and the optical axis is perpendicular to the screen plane (We only shown the Y-Z plane). In our system the camera is located under the screen and looking up at the subject's eye. In our system, the angle between the screen plane and the camera's Z-axis is $\varphi = 24^o$. So the angle between the camera's Z-axis and the eye's optical axis is φ. Suppose the true position of the 3D pupil center and the corneal center are $\mathbf{p} = (x_c, y_c, z_c)$ and $\mathbf{c} = (x_p, y_p, z_p)$, respectively. Based on the analysis in section A.1, there is a noise on the reconstructed 3D points. So, we can define the reconstructed coordinates of the 3D points are $\hat{\mathbf{c}} = (x_c + \Delta x_c, y_c + \Delta y_c, z_c + \Delta z_c)$ and $\hat{\mathbf{p}} = (x_p + \Delta x_p, y_p + \Delta y_p, z_p + \Delta z_p)$, respectively.

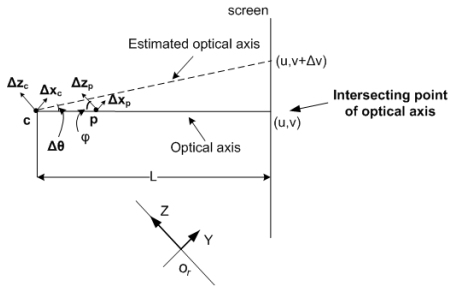

Figure 10: *Gaze estimation noise analysis.*

Thus, we can easily derive the vertical angle between the estimated optical axis and the true optical axis as:

$$\Delta\theta = \frac{(\Delta z_p - \Delta z_c) \cdot \sin\varphi + (\Delta x_p - \Delta x_c) \cdot \cos\varphi}{\|\mathbf{p} - \mathbf{c}\|} \tag{12}$$

Because the difference between optical axis and the visual axis is just a constant angle, and $\Delta\theta$ is a small angle. So the resultant vertical error between the true gaze point and the estimated gaze point is

$$\Delta v = L \cdot \Delta\theta \tag{13}$$

Based on the noise analysis in A.1, we assume the independent gaussian distribution with zero mean and standard deviation of $\sigma_x = 0.037$ and $\sigma_z = 0.37$ for $\Delta x_c, \Delta x_p$ and $\Delta z_c, \Delta z_p$, respectively. The typical distance from the screen is $L = 450mm$, and the typical distance for $\|\mathbf{p} - \mathbf{c}\|$ is $4.8mm$ (section B.2). From Eq. 12 and Eq. 14, we can easily derive the noise on $\Delta\theta$ and Δv as:

$$\begin{aligned} \sigma_\theta &= 0.077 \\ \sigma_v &= 35 \end{aligned} \tag{14}$$

We see that the resultant vertical noise on the gaze estimation is very big ($\sigma_v = 35mm$). Remember that all the noise comes from a small gaussian noise ($\sigma = 0.2$) in the image. Even such a small noise will finally cause a very big noise on the gaze estimation result. (Actually the gaze estimation noise changes, when the subject moves their head and fixates at different directions. Here we just want to show that the small 2D image noise can cause big gaze estimation noise.) In section 3.4, we give the method to reduce this noise.

B Assumptions validation

In our algorithm, there are two important assumptions about the light and the virtual pupil. In this section, we will show that the bias that is introduced by these assumptions is too small to affect the gaze estimation result.

B.1 Validation of light assumption

In section 3.2, we assume that the LED light is located at the camera's origin point. So the light ray to the corneal surface will be reflected back along its own path. Then, we can use our algorithm to estimate the corneal center. However in practice, we can put the light close to the camera center, but cannot put the light exactly on the camera's origin point. In our system, the light is located at about 20mm in front of the camera. In this section, we will use synthetic data to test the effect of this 20mm bias of the light location.

Different from Figure 3 in section 3.2, the ray diagram is shown as Figure 11 to demonstrate the reflections, when lights are not on the camera centers.

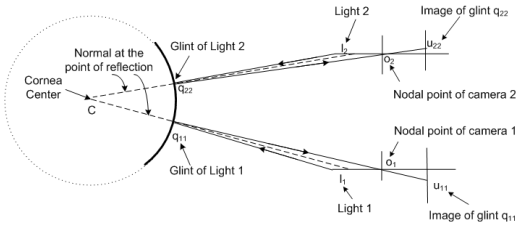

Figure 11: *Ray diagram to show the reflection when the lights are not on the camera centers.*

In this experiment, we generate the synthetic corneal center position $\mathbf{c} = (-35, 0, 450)$ and use the typical corneal radius $R = 7.8mm$ ([Guestrin and Eizenman 2006]). The two lights are located at 20mm in front of the left and right camera, respectively. Then based on the law of reflection, we can compute the image of the glint $\mathbf{u}_{11}$ and $\mathbf{u}_{22}$. Then from $\mathbf{u}_{11}$ and $\mathbf{u}_{22}$, we still use the method in section 3.2 to reconstruct the corneal center $\hat{\mathbf{c}}$ and estimate the gaze point. The estimated corneal center is $\mathbf{c} = (-35.0003, 0, 450.1875)$ and the estimated gaze points are shown in Figure 12. We see that the 20mm bias of the light position will not cause big bias to the gaze estimation result, the average error between the estimates and the ground truth is only 0.089mm. It can be ignored in our algorithm.

B.2 Validation of virtual pupil assumption

In section 3.3, we make the assumption that the virtual pupil is also on the optical axis. In section 3.4, we make the assumption that the distance from the virtual pupil and the corneal center is a constant. Actually, only the pupil position satisfies these two assumptions. In this section, we will show that they are also suitable for the virtual pupil position.

Our experiment is based on the ray diagram of Figure 4 in section 3.3. We generate the synthetic corneal center $\mathbf{c} = (-35, 0, 450)$

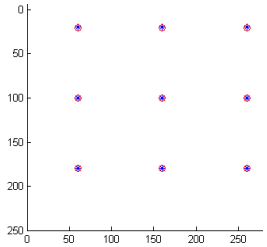

Figure 12: *The gaze estimation bias caused by the light position assumption. The nine ground truth gaze points are shown as "o"s, and their estimations with bias are shown as "*".*

and nine synthetic pupil positions $\mathbf{p}^*$. The distance between $\mathbf{p}^*$ and $\mathbf{c}$ is a constant value: $K = 4.2mm$ ([Guestrin and Eizenman 2006]). Then, by connecting $\mathbf{p}^*$ with $\mathbf{c}$, the optical axis is obtained and the ground truth gaze points can be estimated. The nine ground truth gaze points are shown as "o" in Figure 13.

Then, given the corneal radius $R = 7.8mm$ and the index of refraction of the cornea $n = 1.3375$, we can compute the refracted ray and the images of the virtual pupil $\mathbf{v}_{11}$ and $\mathbf{v}_{22}$. Given the virtual pupil images, we still use the method in section 3.3 to reconstruct the virtual pupil position and then estimate the gaze points. The result is shown in Figure 13.

We see that the distance between the virtual pupil and the corneal center ($\hat{K} = \|\mathbf{p} - \mathbf{c}\|$) changes when the subject fixates on different position. But the change is limited (<0.1mm). We also notice that ($\hat{K}$) is larger than the pupil-cornea distance $K = 4.2$. It means that the pupil is behind the virtual pupil. By connecting $\mathbf{c}$ and $\mathbf{p}$ to estimate the optical axis, the resultant gaze estimates are shown as "*"s. The average error for these nine gaze points is 0.623mm. This error is very small. So in our algorithm, we just ignore this error and assume the virtual pupil is also on the optical axis.

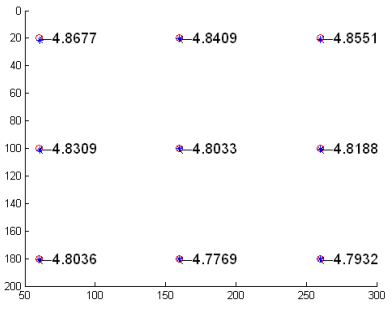

Figure 13: *The gaze estimation bias caused by the virtual pupil assumption. The nine ground truth gaze points are shown as "o"s, and their estimations with bias are shown as "*". The distance between virtual pupil and the corneal center is also shown near each gaze point.*

References

2006. *http://www.a-s-l.com*.

BEYMER, D., AND FLICKNER, M. 2003. Eye gaze tracking using an active stereo head. *IEEE Conference on CVPR03*.

C.H.MORIMOTO, KOONS, D., A.AMIR, AND FLICKNER, M. 2000. Pupil detection and trcking using multiple light sources. *Image and Vision Computing 18*, 331–336.

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*, 1124–1133.

HARO, A., FLICKNER, M., AND ESSA, I. 2000. Detecting and tracking eyes by using their physiological properties, dynamics, and appearance. *IEEE Conference on CVPR00*.

HUCHINSON, T., JR., K. P. W., AND REICHERT, K. 1989. Human computer interaction using eye-gaze input. *IEEE Transactions on Systems, Man, and Cybernetics 19*, 1527–1533.

JACOB, R. J. 1991. The use of eye movements in human computer interaction technqiues: What you look at is what you get. *ACM Tracnsations on Information Systems 9*, 152–169.

LAGARIAS, J. C., REEDS, J. A., WRIGHT, M. H., AND WRIGHT, P. E. 1998. Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM Journal of Optimization 9*, 1.

2005. Lc technologies, inc. *http://www.eyegaze.com*.

LIVERSEDGE, S., AND FINDLAY, J. 2000. Saccadic eye movements and cognition. *Trends in Cognitive Science 4*, 6–14.

MASON, M., B.HOOD, AND MACRAE, C. 2004. Look into my eyes : Gaze direction and person memory. *Memory 12*, 637–643.

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 98*, 4–24.

MORIMOTO, C. H., AMIR, A., AND FLICKNER, M. 2002. Detecting eye position and gaze from a single camera and 2 light sources. *Proceedings of the International Conference on Partern Recognition*.

OYSTER, C. W. 1999. *The Human Eye: Structure and Function*. Sinauer Associate, Inc.

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 34*, 1, 234–245.

2007. *http://www.smi.de*.

TRUCCO, E., AND VERRI, A. 1998. *Introductory Techniques for 3D Computer Vision*.

WANG, J.-G., AND SUNG, E. 2002. Study on eye gaze estimation. *IEEE Transactions on Systems, Man and Cybernetics, PartB 32*, 3, 332–350.

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

ZHANG, Z. 2000. A flexible new technique for camera calibration. *IEEE Transactions on PAMI 22*.

ZHU, Z., AND JI, Q. 2004. Eye and gaze tracking for interactive graphic display. *Machine Vision and Application 15*, 3, 139 – 148.

ZHU, Z., AND JI, Q. 2005. Eye gaze tracking under natural head movements. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05)*.

ZHU, Z., AND JI, Q. 2007. Novel eye gaze tracking techniques under natural head movement. *to appear in IEEE Transactions on Biomedical Engineering*.