A New 3D Line of Gaze Estimation Method with Simple Marked Targets and Glasses

Samil Karahan[1,2], Yakup Genc[2], Yusuf Sinan Akgul[2]
[1]TUBITAK BILGEM, Anibal Street, 41470, Gebze, Kocaeli, Turkey
[2]GIT Vision Lab, http://vision.gyte.edu.tr, Department of Computer Engineering,
Gebze Institute of Technology,41400, Gebze, Kocaeli, Turkey
samil.karahan@tubitak.gov.tr
{ygenc, akgul}@bilmuh.gyte.edu.tr

This paper presents a new Line of Gaze (LoG) method that uses a paper target with a hole for training and simple glasses for the head tracking. Both the target and the glasses are marked with fiducials for 3D localization and they are easy to construct. The system does not need any extra cameras or IR light sources. As opposed to many LoG methods in the literature, our method does not impose any restrictions on the user head movements or the LoG orientations, yet it produces low error rates comparable to the state of the art. The proposed method introduces many novel contributions such as a linear system for the cornea center and radius estimation and a new training method for preventing user errors. The experiments performed with users produced encouraging numerical results.

Keywords: Line of Gaze, Eye Tracking, Head Pose, Fiducial Marker

Introduction

Estimation of human eye gaze directions is crucial for a number of areas including human computer interaction, human cognitive and emotional state analysis, and attentive user interfaces. The gaze estimation is the process of determining the 3D Line of Gaze (LoG) of a user's eye in a known coordinate system (Hansen and Ji, 2010).

The existing gaze estimation methods can be classified with respect to their main assumptions about the problem and the equipment required. The first type of methods employ infrared (IR) type light sources near the camera to produce a light reflection on the cornea (For example, Villanueva et al., 2006). The distance between the reflection of the light(s) and the pupil center can be used for estimating the eye gaze directly. These systems usually produce accurate results but they may not work when there are other light sources in the environment. The second type of gaze estimation methods use more than one camera to allow relaxed head movement restrictions (Boening et al., 2006). The third type of these methods do not use any lights or extra cameras. However, these systems have serious head movement restrictions and their accuracies are not as good as others (For example, Valenti, et al., 2009). It should be noted that some of the methods mentioned above report Point of Regard (PoR) estimations as gaze estimation values. PoR usually involves a mapping between the pupil center positions and the screen coordinates without any detailed 3D processing. LoG estimation, on the other hand, requires extensive 3D modeling of the head and the eyes, which makes them robust against head movements.

In this paper, we present a novel LoG estimation system that employs simple glasses (Figure 2 left), which can be built very easily from cheap 3D movie glasses, and a paper target (Figure 2 right), which is required only during calibration. The simple glasses are used for the accurate head pose and 3D cornea center estimation. The glasses are modified by removing the color filters, printing a few fiducial marks on a piece of paper and gluing it on the glasses (Figure 2 left). Similarly, the target includes a few fiducial marks and a small hole near the center (Figure 2 right). For the system training, the user puts on the glasses and holds the paper target at his/her hand. The user looks at the shown positions on

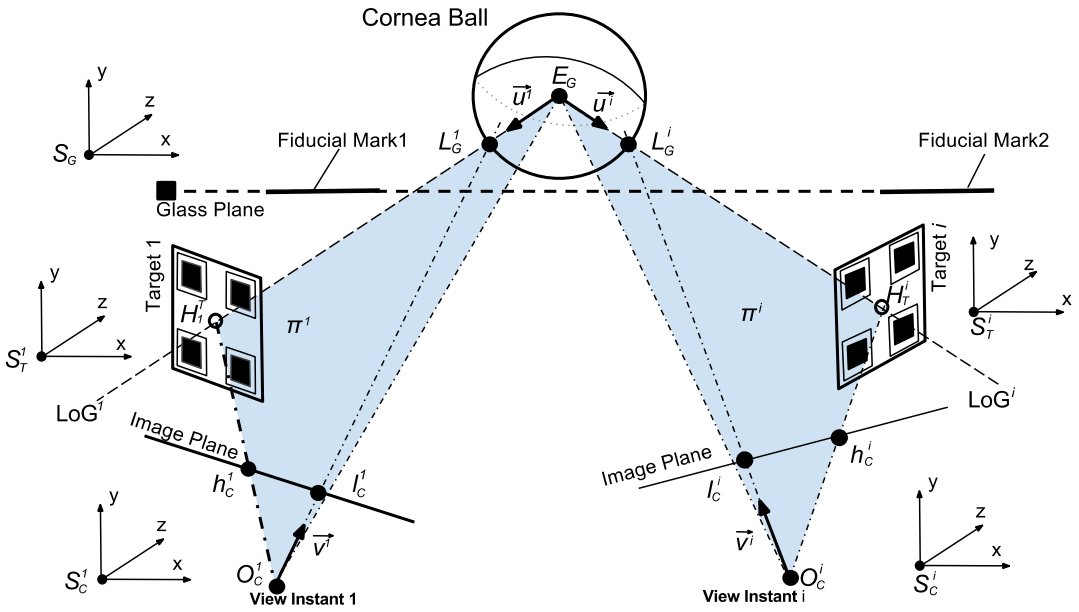

Figure 1.The 3D Geometry of Line of Gaze (LoG).

a monitor[1] through the hole on the target. The camera takes the images of both the user and the target in the same image frame. These images are used for calculating the cornea center and the radius of the eye. After the system calibration, the LoG is estimated using only the glasses and the estimated pupil positions.

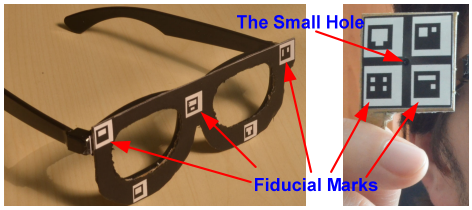

Figure 2.left: the modified3D movie glasses, right: the target with the small hole.

The proposed system has many advantages: First, the user head movements are not restricted as long as the camera can see the user's eye and the fiducials on the glasses. Second, the method of seeing the positions on the monitor[1] through the hole on the target makes the calibration process more robust because it is known that the eye gaze can be a few degrees off when looking at a certain target (Hansen and Ji, 2010). Finally, since our system continuously measures the 3D position and orientation of the users head, we can use this information for other tasks like head gesture analysis.

The 3D Geometry of Line of Gaze (LoG)

Throughout the paper, we use capital letters to represent 3D real-world points and small case letters to represent positions on the image plane, which are still 3D points. Figure 1 shows the general 3D geometry of our LoG estimation system, which includes a 3D glass coordinate system S_G. For a given view instant i, our system also includes a camera coordinate system S_C^i and a target coordinate system S_T^i. The translational and rotational transformations between these coordinate systems are estimated by the fiducial localizations as explained in the next section. Any 3D position can be explained in any coordinate system using these transformations. We use the notation A_Q^i, where A is the 3D point in homogenous coordinates, Q is the coordinate system, and i is the view instant. Without the loss of generality, our system assumes that the cornea center stays fixed at position E_G while the camera centers, O, and the target positions, H, can change for each view instant i with respect to S_G. The center of the pupil for the view instant i is represented by L_G^i, whose image is projected on the image plane as l_C^i. Similarly, the center of the target hole is represented by H_T^i, whose image is projected on the image plane as h_C^i. Note that for a given

[1]Our calibration does not necessarily need a computer monitor because the user can look at any positions in the real world as long as he/she sees the target through the hole.

view instant i, the points O_G^i, l_G^i, and h_G^i form a plane, which we call π_i in S_G. Note also that the cornea center E_G is the same for all the planes π_i, $i = 1..n$ for given n view instants. This novel observation is central to our method of estimating the cornea center and radius from n views as explained in next section. Finally, the answer for the LoG estimation problem can be found by estimating the vector u^i in S_G, which is the direction from the cornea center E_G to the pupil center L_G^i as explained in next section.

The LoG Estimation Method

In this section, we describe the steps of our proposed method, which are shown on Figure 3 in detail. Our method includes two main phases: training and LoG estimation. The training phase is for the estimation of the radii and the center of the cornea ball. The LoG estimation is for the final result from our system.

Detecting the Fiducial Marks, Calculating the Transformation Matrices and Finding the Pupil on the Image

Our method needs the transformation matrices (rotation and translation) between the S_G and the other coordinate systems, S_T^i and S_C^i. For this purpose, we use well known calibration method of Tsai (Tsai, 1987) that uses correspondences between different coordinate spaces. The point correspondences between the camera space S_C^i and the glasses space S_G are estimated using the detected fiducial positions on the images and the known 3D fiducial positions on the glasses. The fiducials on the images are detected using open source libraries. The 3D fiducial positions on the glasses in S_G are known because we carefully place the printed fiducials on a plane with measured distances between them using image editing software. Although, the calibration method needs at least 5 different correspondences between different spaces, we use extra correspondences for robustness against detection errors and occlusions. The estimated transformation matrix between S_G and S_C^i is called P_{GC}^i, which is a 3x4 matrix that includes rotation and translation components. The transformation matrix P_{TC}^i between S_G and S_T^i is estimated similarly with the help of P_{GC}^i. Note that all the transformation matrices are found at very good precision levels because our method is based on very well established camera calibration techniques. As a result, our head pose estimations are much better than alternative markerless head pose estimation techniques. The intrinsic camera parameters are assumed fixed for the whole system.

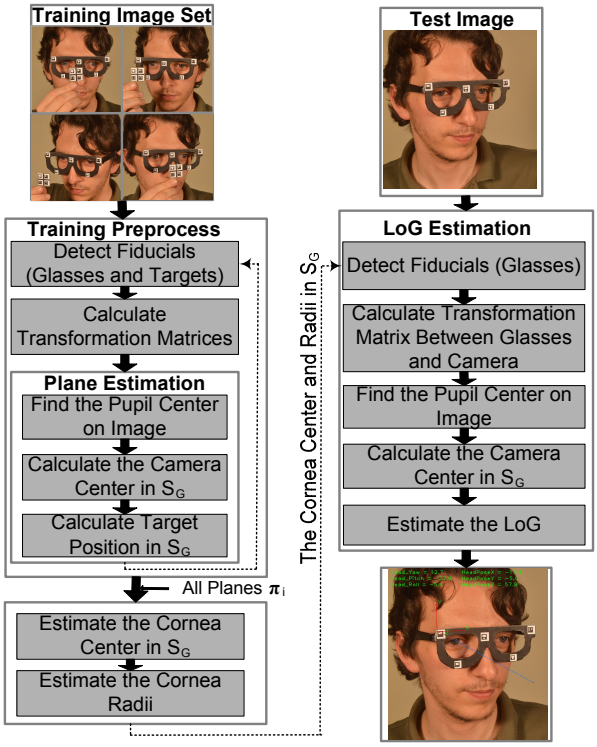

Figure 3. The flowchart of LoG estimation system includes training and testing stages

We detect the center of the pupil of the user on an image i by using the gradient field method of (Timm and Bart, 2011). We use the fiducial marks on the glasses to restrict our search space for the pupil center estimation, which makes our system more robust and efficient. The estimated pupil center gives us the 2D image position $(l_x^i,\ l_y^i)$, which is called l_C^i in S_C^i (Figure 1) and can be calculated by

$$l_C^i = \left[(l_x^i - x_0)/d_{px} \quad l_y^i - y_0/d_{py} \quad f \quad 1\right], \qquad \textbf{(1)}$$

where x_0 and y_0 are principal points, d_{px} and d_{py} are pixel sizes, and f is the focal length of the camera.

The point l_C^i is required for the estimation of the plane π_i. We also need to know the positions of the camera center (O_C^i) and the image of the target (h_C^i) for the estimation of the plane π_i. However, all these three

points need to be transformed to S_G, which is explained in the succeeding sections.

Representing the Camera Center O_C^i in S_G

The position of the camera center O_C^i in S_G is called O_G^i. By the definition of the camera center, if we transform the position of the camera center O_C^i to a position in camera coordinate system O_G^i with the help of P_{GC}^i, we obtain

$$P_{GC}^i O_G^i = 0, \tag{2}$$

which can be solved by (Hartley and Zisserman, 2004)

$$O_G^i = [X/T \quad Y/T \quad Z/T \quad 1], \tag{3}$$

where

$$\begin{aligned} X &= \det([p_2 \quad p_3 \quad p_4]), \; Y = -\det([p_1 \quad p_3 \quad p_4]), \\ Z &= \det([p_1 \quad p_2 \quad p_4]), \; T = -det([p_1 \quad p_2 \quad p_3]). \end{aligned} \tag{4}$$

In equations given above, p_j represents j^{th} column of transformation matrix P_{GC}^i.

Representing the Position of the Hole H_T^i in S_G

The third point required for the estimation of the π_i is the 3D position of the target H_C^i, which can be obtained by transforming the known 3D position H_T^i to the space S_C^i by

$$P_{TC}^i H_T^i = H_C^i. \tag{5}$$

Since we need to explain all three points of the π_i in S_G, we find H_G^i by

$$inv\left(\begin{bmatrix} P_{GC}^i & 0 \\ 0 & 1 \end{bmatrix}\right) \begin{bmatrix} H_{Cx}^i/H_{Ct}^i \\ H_{Cy}^i/H_{Ct}^i \\ H_{Cz}^i/H_{Ct}^i \\ 1 \end{bmatrix} = \begin{bmatrix} H_{Gx}^i/H_{Gt}^i \\ H_{Gy}^i/H_{Gt}^i \\ H_{Gz}^i/H_{Gt}^i \\ 1 \end{bmatrix} \tag{6}$$

Finding the Center of Cornea E_G

Estimating the cornea center for a given user is crucial for our system and this process needs to be repeated every time the user puts on the glasses. When the user looks through the target H_G^i, the LoG vector $\overrightarrow{E_G H_G^i}$ passes through the pupil center L_G^i (Figure 1). E_G is located on the plane π_i, which includes the points O_G^i, L_G^i and H_G^i. Since L_G^i is unknown in S_G, we cannot use it directly for the estimation of π_i. However, the image of L_G^i, which is l_C^i (formula 1), can be used for this purpose. l_C^i is transformed to l_G^i using inverse of transformation matrix P_{GC}^i.

A plane equation in 3D homogenous coordinate system may be written as (Hartley and Zisserman, 2004)

$$\pi^T X = 0, \tag{7}$$

where X is a homogenous point on π and T is for matrix transpose.

A plane can be defined by specifying at least 3 points on it. We know that O_G^i, l_G^i and H_G^i are all on π_i and hence we can write

$$\pi_i^T [O_G^i \quad l_G^i \quad H_G^i] = 0. \tag{8}$$

The plane vector can be calculated by

$$\pi_i = [D_{234} \quad -D_{134} \quad D_{124} \quad -D_{123}]^T \tag{9}$$

where the scalar D_{jkl} is the determinant formed from the jkl rows of the 4x3 matrix $[O_G^i \quad l_G^i \quad H_G^i]$ (Hartley and Zisserman, 2004).

As shown in Figure 1, when the user changes his/her LoG or when the camera/target changes their positions, a new plane is formed. The common property of these planes is that they must pass through the center of cornea, E_G. At least three planes (or three view instants) are enough to find this intersection point. However, having more than three planes can minimize the sum of distance errors between planes and the center. The linear system below is used for estimating $E_G = [E_{Gx} \quad E_{Gy} \quad E_{Gz} \quad 1]^T$ for given n view instants.

$$[\pi_1 \quad \pi_2 \quad \dots \quad \pi_n] \begin{bmatrix} E_{Gx}/E_{Gt} \\ E_{Gy}/E_{Gt} \\ E_{Gz}/E_{Gt} \\ 1 \end{bmatrix} = 0 \tag{10}$$

The linear system is solved by using Singular Value Decomposition (SVD), which minimizes the error in the least squares sense.

Finding the Radii of the Cornea

We model the cornea with an ellipsoid with three radii, which have to be estimated for each user separately because it changes from person to person. In Figure 1, LoG unit vector u^i (i.e., $\overrightarrow{E_G H_T^i}$) and the unit vector v^i (i.e., $\overrightarrow{O_C^i l_C^i}$), which starts from the camera center and pass through the pupil position on the image plane, intersect at the real world pupil center L_G^i, which can be expressed as

$$E_G + r.\mathrm{u}^i = O_G^i + a_i \mathrm{v}^i \quad \textbf{(11)}$$

where the vector $r = [r_x \quad r_y \quad r_z]^T$ represents the 3 radii of ellipsoidal cornea model, the scalar a_i is the distance between L_G^i and O_G^i, and the operator $\cdot$ represents element-wise vector multiplication. Given n training images from n view instants, we can write the following linear system to estimate the cornea ellipsoid radii r.

$$\begin{bmatrix} u_{\mathrm{x}}^1 & 0 & 0 & v_{\mathrm{x}}^1 & \dots & 0 & E_{G_x} - O_{G_x}^1 \\ 0 & u_{\mathrm{y}}^1 & 0 & v_{\mathrm{y}}^1 & \dots & 0 & E_{G_y} - O_{G_y}^1 \\ 0 & 0 & u_{\mathrm{z}}^1 & v_{\mathrm{z}}^1 & 0 & 0 & E_{G_z} - O_{G_z}^1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & u_{\mathrm{z}}^{\mathrm{n}} & 0 & \dots & v_{\mathrm{z}}^{\mathrm{n}} & E_{G_Z} - O_{G_z}^n \end{bmatrix} \begin{bmatrix} r_x \\ r_y \\ r_z \\ a_1 \\ \vdots \\ a_n \\ 1 \end{bmatrix} = 0 \quad \textbf{(12)}$$

The system above, which includes n*3 rows, can be solved by SVD to minimize the error in least squares sense.

Estimation of LoG

The training phase produces the estimated cornea center E_G and the cornea radii r. For the testing phase, for a given view instant i, we need to calculate the vector u^i for LoG value without using the target information such as H_G^i and h_G^i. Equation 11 can be used to estimate the LoG, but now the unit vector u^i is unknown due to the absence of the target paper in testing phase. However, we now know the values of r, so we can write the equation system

$$\left(O_G^i - E_G + a_i \mathrm{v}^i\right)/r = \mathrm{u}^i, \quad \textbf{(13)}$$

where operator / is for element-wise vector division. The system above expands into 3 equations with 4 unknowns (3 for u^i and 1 for a). Since we know that $\left|u^i\right|$ is 1, we can produce a unique solution. Note that for some cases, the solution for u^i is a complex number, which indicates that the coordinate systems of the user's head and the glasses are not the same. In this case, the calibration is not valid anymore and E_G has to be re-estimated.

Results

The fiducial marks were prepared in a picture editor program with respect to the size of the glasses. The fiducial positions are chosen at precisely known locations. The intrinsic parameters of the camera are calculated with OpenCV. The user was asked to sit in front of a screen at 50cm distance to look at shown positions through the target hole. The user was asked to move his/her head freely to obtain many head and LoG orientations. Approximately 30 images were taken from each user for training and testing. We performed the experiments in a leave-one-out methodology: An image was selected in test data and the training process was run with the others. This process was repeated for all images, thus, each image was tested at least once.

We assumed the ground truth LoG to be the vector from the cornea center E_G to the target hole H_G^i. The estimated LoG, u^i, was obtained using equation 13. The angular difference between the ground truth vector and the estimated vector is calculated in degrees as our error measure. Table 1 lists the error rates for four sub-experiments: results for the complete set, results for normal head poses (-20° to +20°), results for extreme head poses (<-20° or > +20°), and results for the complete set with hand detected pupil centers. Figure 4 shows some examples from our systems for the low and high error examples.

Experiments	Errors (degrees)			
	Min	Max	Mean	Median
Complete Set	0.34°	6.81°	3.52°	3.35°
Normal Head Pose	0.73°	6.60°	3.81°	3.65°
Extreme Head Pose	0.34°	6.81°	3.17°	3.34°
Complete set by Hand Detected Pupil Centers	0.54°	6.32°	3.25°	3.22°

Table 1. The error of the proposed method

Discussion

Our system works at real time rates for images of size 640x480 on common hardware. The limiting factor for the image resolution is the fiducial sizes (minimum 20x20 pixels) and the area of the eye visible from the glasses (minimum 120x80 pixels). Our error rates compare favorably with the existing methods. For example, the method of (Heyman 2011), which is a passive method, restricts the head and LoG orientations and reports average errors rates of 5.64 degrees for some cases. Similarly, (Valenti et al. 2012) reports mean error rates of 2-5 degrees. The active methods report smaller error rates (Villanueva et al., 2006, 1.0 deg) but they

require extra equipment and they impose stricter head movement restrictions.

Figure 4. Sample results with position, orientation and error numbers from low and high error examples. The blue line is the estimated LoG, the orange line is the ground truth.

Conclusions

We presented a new 3D LoG estimation method that relies on simple marked glasses for head tracking and simple paper target with a hole for training. Our method uses novel 3D techniques to estimate the cornea center and the radius robustly under no head or LoG orientation restrictions. In fact, different head and LoG orientations during the training produce more robust results. Our novel training apparatus with a hole makes the training process even more robust because it eliminates user errors while tracking targets.

In contrast to many PoR methods, our system does not depend on any fixed coordinate systems other than the glasses coordinate system. As a result, it can be freely used in the real word applications such as Google glasses and first person vision (Kanade and Hebert, 2012).

For the future work, we will test our system with more users on real world eye tracking tasks. We also plan to eliminate the need for retraining in case the glasses move with respect to the users head. Finally, we plan to enhance our system to be used with regular prescription glasses instead of the modified movie glasses.

Acknowledgements

This work is supported by TUBITAK Project 112E127.

References

Hansen, D. W., and Ji, Q. In the eye of the beholder: A survey of models for eyes and gaze. Pat. Anly. and Mach. Intel., IEEE Transac. on 32, 3, 478-500, 2010.

Villanueva, A., Cabeza, R., and Porta, S. Eye tracking: Pupil orientation geometrical modeling. Image and Vision Computing, 24(7):663–679, July 2006.

Valenti, R., Sebe, N., and Gevers, T. Combining head pose and eye location information for gaze estimation. IEEE Transactions on 21.2: 802-815, 2012.

Valenti, R., Staiano, J., Sebe, N., and Gevers, T. Webcam-based visual gaze estimation. ICIAP 2009, pp. 662-671, 2009.

Tsai, Roger. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. IEEE Trans. on Robot. and Automat., vol. 3, pp. 323-344, Aug. 1987.

Hartley, R. I., and Zisserman, A. Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521540518, 2004.

Timm, F., and Bart,E. Accurate eye centre localisation by means of gradients. In VISAPP, 2011.

Boening, G. et al. Mobile eye tracking as a basis for real-time control of a gaze driven head-mounted video camera. Eye track. research & appli.. ACM, 2006.

Kanade, T., and Hebert, M. First-Person Vision. Proceedings of the IEEE 100.8: 2442-2453, 2012.

Heyman, T. et al. 3d face tracking and gaze estimation using a monocular camera. In Proceed. of the 2nd Interna. Conf. on Pos. and Context-Awareness, 2011.