

A Flexible Eyetracker for Psychological Applications

D. J. DeVault

A. H. Bond

Computer Science Department
California Institute of Technology
Pasadena, CA 91125

Computer Science Department
California Institute of Technology
Pasadena, CA 91125

Abstract

We describe a practical method for measuring eye movements during psychological tests. This is an important class of applications including clinical evaluations and marketing studies. Existing methods in common use for psychological measurement, for example infrared reflection methods, are invasive involving head stabilization and special purpose lighting. In our experiments, we need to observe subjects for long periods, on the order of one hour. In addition, subjects must verbalize, which makes it difficult to stabilize their heads relative to the camera. We track the head using a lightweight spectacle framework worn by the subject. It has a set of easily visible colored balls. We segment each image into four characteristic colors, corresponding to iris, yellow ball, red ball, and background, which are obtained by sampling the images for each subject. The classification into colors is done by training a simple neural network for each characteristic color. We match a template to color-reduced image regions to find the balls and the two irises. We use a model-based object pose method, which uses a prior measurement of the relative positions of the balls, to calculate the spectacle framework pose (the head pose). A linear method is used for calibrating gaze position against head pose and iris positions. The subject's gaze position can be tracked reliably for periods of more than an hour. The locations of image features are found with an accuracy of approximately one pixel of the image. In a 640x480 image of the whole face, the eyes are each about 80 pixels across. This gives a corresponding accuracy of calculated eye gaze position on a 17 inch monitor of about 1cm horizontally and 2cm vertically. This method has shown itself in practice to be very flexible for psychological measurement, giving sufficient accuracy and being noninvasive.

Introduction

Several techniques are currently used for measuring eye gaze position. Very high accuracy is possi-

ble with a magnetic search coil technique [3]. However, it is necessary that the subject wear a contact lens containing imbedded wire coils, and the subject must remain in the center of magnetic field coils. An ideal eyetracker for psychological tests would be minimally invasive, place few constraints on the subject's head position, and not require overly expensive equipment. By these criteria, the best eyetracking method in common use for psychological tests is the pupil-center/corneal-reflection method[4].

This method works by tracking both the subject's pupil position and the reflection from an infrared light emitting diode (LED) on the cornea. The LED, which provides the corneal reflection and enhances contrast between the pupil and the iris, is placed several centimeters from the subject's eye. Images are collected with a special infrared camera. Eye gaze is calculated using the displacement between the pupil center and the corneal reflection in the image. While this method is theoretically robust against small head movements, in practice the subject's head is stabilized using a chin rest (and often a bite bar) to get sufficiently accurate results.

The use of a chin rest to stabilize the head makes talking impossible and makes long experiments uncomfortable for the subject. We describe a new eyetracking method that allows for long experiments (an hour or more) in which stabilization of the head is impractical.

General Approach

We seek to determine the subject's eye gaze position as an (x,y) coordinate pair on a computer screen. The subject's eye gaze position depends on his head position. Since we do not stabilize the subject's head, it is necessary to track the subject's head in addition to his eyes.

For this purpose, our method requires that the subject wear a light-weight framework attached to a pair of spectacles that have had the lenses removed (see

Figure 1). The framework is constructed from light-weight wood and includes five brightly colored balls (spheres) with known geometry relative to each other. The spectacle framework rotates and translates rigidly with the subject's head, so that we may consider the pose (rotation and translation) of the spectacle framework rather than the pose of the subject's head. The brightly colored balls are easily located in the camera images. The spectacle framework's pose may consequently be determined, as described below. Using this spectacle framework makes unnecessary the generally difficult task of calculating the subject's head pose given only an image of the subject's head.

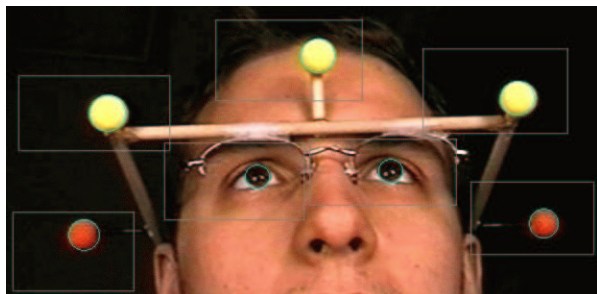


Figure 1: Example camera image. The image feature search regions are visible as rectangles enclosing the balls and irises. The circles indicate the calculated image feature locations and diameters. The annuli are not shown (even though they are used to locate the image features).

We use an ordinary color video camera to record an image stream of the subject's head during the experiment. Ambient office lighting is used, with the addition of two standard office lamps near the computer screen for additional image brightness. All eyetracking is performed offline, after the experiment. Individual images from the video stream recorded by the camera are considered separately. In each image, we locate seven image features (the five balls on the spectacle framework and the two irises) by exhaustively searching a predetermined search region for each image feature. The locations of the five balls are used to determine the subject's head pose. The head pose plus the locations of the two irises, taken relative to one of the balls on the spectacle framework, completely characterize the subject's eye gaze position in any given frame. Each subject performs a two minute calibration procedure so that the relationship between these data and the subject's eye gaze position may be deduced.

Color Reduction

We begin the tracking process by reducing the number of colors in the image feature search regions to four: iris, front ball, back ball, and background. We refer to these four colors as "reduced colors". Three neural networks are trained to recognize pixels corresponding to iris color, front ball color, and back ball color. The network architecture is shown in Figure 2. These networks are trained on characteristic pixels sampled from the images. The experimenter selects the training pixels by dragging the mouse over rectangular image regions.

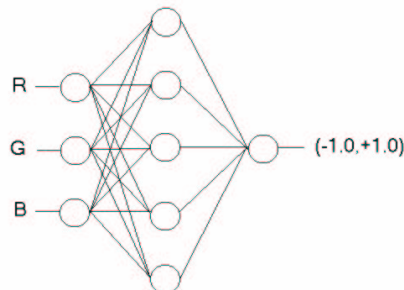


Figure 2: Neural network architecture. The three inputs are red, green, and blue color values for the pixel to be classified. There is a hidden layer with five neurons. The output is positive if the pixel matches the color this particular network is responsible for, and negative otherwise.

The network responsible for identifying iris colored pixels is trained against non-iris pixels that appear in the iris search regions, typically sclera and skin colored pixels. The iris network ideally outputs +1 for iris colored pixels and -1 for pixels of any other color. The iris network is only responsible for correctly identifying pixels in the two image regions enclosing the subject's eyes, so it is not necessary to train the iris network to reject pixels of colors that will not appear in the iris search regions. Likewise, the networks responsible for identifying front ball (in our case, yellow) and back ball (in our case, red) colored pixels are trained against colors that typically appear adjacent to the front and back balls in the images.

The three neural networks are trained using ordinary gradient descent, with an adaptive learning rate[1]. Once the networks have been satisfactorily trained, classification of an arbitrary pixel into one of the four colors is accomplished by feeding the pixel into the three neural networks. If no network gives an

output greater than zero, the pixel is classified as background color. Otherwise it is classified as the color whose network gave the highest positive output. Two examples of color-reduced regions are given in Figure 3.

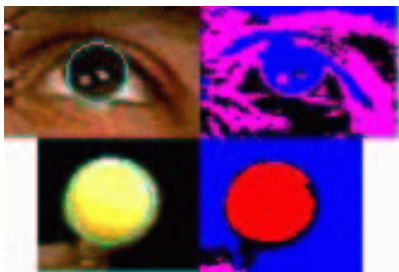


Figure 3: Color reduction examples. The original search regions for an iris and a yellow ball are shown at the left. Their corresponding reduced color images are shown at the right.

Image Feature Locations

Each of the seven image features is located using the same method. Before tracking begins, the experimenter selects a search region for each image feature. This is done by dragging rectangular image regions with the mouse. The search regions must be sufficiently large that the image features will remain within them despite any head or eye movements by the subject during the experiment. In our experience subjects remain relatively still, even though their heads are not stabilized, and the search regions do not need to be unreasonably large (see the search regions in Figure 1).

Circular Template with Annulus

Each image feature is assumed to appear as a solid circle of pixels of constant color, surrounded by an annulus of pixels of a different color (see Figure 4). Since the balls are projective projections of spheres, they appear as circles in the images regardless of their 3D positions in space. The irises appear circular when the subject looks directly into the camera, and become elliptical as the subject’s gaze direction deviates from the camera axis. In practice, a circular template provides a close enough approximation to locate the irises accurately.

The approximate diameter of the solid inner circle for each image feature is initialized before tracking begins by dragging the mouse over each image feature. The best size for the outer annulus is determined in



Figure 4: Circular Template with Annulus. Pixels in the inner circle are of one reduced color, while pixels in the outer annulus are of any other reduced color.

practice. We currently have the outer annulus diameter always equal to $1.3 * \text{inner circle diameter}$.

Locating the Image Features

Since the precise diameter of an image feature varies with the subject’s head pose, it is determined in each image. The diameter is allowed to vary from the approximate value set before tracking begins by up to M pixels in either direction. We have used $M = 2$. For each possible template size, we consider all the integral pixel coordinates for the center of the template that keep the entire template within the search region. At each position, we consider each pixel in the template. The measure of how well the template matches the image is the fraction of pixels in the template that are the appropriate reduced color in the image given their position in the template.

Each image feature has a characteristic color C , which is one of the four reduced colors. Pixels within the solid inner circle in the template are of the appropriate color if the corresponding image pixel is of color C . Pixels within the annulus are of the appropriate color if the corresponding image pixel is not of color C (is some other color).

The template diameter D and origin coordinates (x_c, y_c) that give the highest percentage of pixels matching the template determine the image feature’s location to the nearest pixel. We then estimate the image feature’s location with subpixel accuracy. The set $\{x_i, y_i, P_i\}$ is formed from the 3×3 grid of pixels with (x_c, y_c) at the center, where P_i is the percent of pixels matching the template with diameter D and center at (x_i, y_i) . A second-order surface with height P_i at position (x_i, y_i) is fit. Its maximum may be found analytically at $(x_{subpixel}, y_{subpixel})$. This maximum is the estimated maximum response to the template, and $(x_{subpixel}, y_{subpixel})$ is the estimated image feature location with subpixel accuracy.

Restricting the search regions for the image features by using their locations in previous frames could eliminate the need to define large search regions, and would speed up the image feature location process. In our experiments, we have not been concerned with the overall speed of the tracking process. We opted

for this reliable but slow method of exhaustive searching. The image feature location process could be further sped up using a Kalman filter to locate the image features[5].

Benefits of Using a Circular Template with Annulus

Using a circle with an outer annulus as the template instead of using a simple circle provides several benefits. (1) The color reduction process sometimes results in large regions of false positives that match every pixel in a circular template. For example, this can happen when, on a subject with dark hair and eyes, the iris network classifies eyebrow pixels as iris color. The problem arises because false positive eyebrow regions can have area considerably larger than the area of the iris. Checking for an annulus of non-iris colored pixels prevents a strong match from occurring in these regions, because the annulus in the false positive region will not be mostly non-iris color, as it will for the correct iris location. (2) Since we use a variable diameter for the template, incorrectly classified subregions of an image feature could cause an incorrect location to be calculated with a simple circular template. For example, this could happen if part of the iris near the limbus is misclassified as background. A simple circular template could find a higher match with a smaller diameter that forces the false negative pixels out of the circle. (3) False negatives within the iris, combined with a region of false positives just outside the iris, can cause a circular template to move off the correct location. This can happen, for example, if a bright specular reflection within the iris is classified as non-iris color. Due to inadequate lighting, a dark region can sometimes exist outside the iris, near the corner of the eye. If the subject’s irises are dark, this region can be classified as iris-color. In this situation, a simple circular template could move off the center of the iris, as the benefit of moving off the correct location could outweigh the cost. Using an annulus will prevent this, because the annulus pixels provide a stiff penalty for moving away from the true iris location.

Head Pose

Once the locations of all the balls are known, the subject’s head pose can be calculated using an algorithm due to Dementhon and Davis[2].

The algorithm places two restrictions on the spectacle framework. At least four non-coplanar features (balls) must be located in the images, and the 3D geometry of the balls must be known. We designed the framework geometry so that five balls are always visible in the images, even when the subject rotates his

head by up to twenty degrees. The bright yellow and red colors of the balls do not occur on subjects’ faces, and make the balls very easy to locate. The geometry of the framework is measured mechanically. (The geometry of the framework is constant across subjects and needs only to be determined once.)

Given the locations of the five balls in the image, the Dementhon and Davis method provides the rotation and translation from the coordinate system of the framework to the coordinate system of the camera.

Calibration Procedure

The eyetracker is calibrated to each subject by finding the best linear correspondence between head pose and iris positions and the corresponding eye gaze positions on the screen. The subject fixates each point in an $N \times N$ grid of points on the screen for each of 9 different head positions. We have used $N = 3$. The head positions are: looking straight at the screen, with head rotated to the left, rotated to the right, rotated up, rotated down, and with head translated to the left, translated to the right, translated forward, and translated backward. The extent of these rotations and translations is sufficient to cover the range of head motion that is likely to occur during the experiment. In our experience, the rotations need to be about twenty degrees, and the translations need to be about five centimeters. The calibration process takes approximately two minutes.

We refer to the points in the $N \times N$ grid at which the subject looks as “fixation points”, and the camera images corresponding to the fixation points as “calibration frames”. The head pose and eye position data from each calibration frame are placed into a 1×17 row vector as follows:

$$B_{1 \times 17} = [R_1 \ R_2 \ \dots \ R_9 \ T_1 \ T_2 \ T_3 \ x_l \ y_l \ x_r \ y_r \ 1]$$

where

$R_1 - R_9$: the 3×3 rotation matrix for the head pose

$T_1 \ T_2 \ T_3$: the 3×1 translation vector for the head pose

$x_l \ y_l$: the position of the left iris relative to the center front ball

$x_r \ y_r$: the position of the right iris relative to the center front ball

1 : makes possible a constant term in the linear map

The eye gaze position, which for the calibration frames is a known location in the $N \times N$ grid, is placed into a 1×2 row vector:

$$C_{1 \times 2} = [x\text{-position } y\text{-position}]$$

Calibration provides $S = 9 \times N \times N$ examples of $B_{1 \times 17}$ and $C_{1 \times 2}$. If these examples are arranged in matrices with S rows, the assumption of a linear relationship

between the subject's head pose and iris positions relative to the front center ball (B) and the subject's eye gaze position (C) is written:

$$C_{S \times 2} = B_{S \times 17} \cdot A_{17 \times 2}$$

Calibration for a particular subject amounts to solving for A . The least squares solution is:

$$A = (B^T * B)^{-1} * B^T * C$$

if B is rank 17. Once calibration is accomplished, the eye gaze position for an arbitrary image frame is estimated as:

$$C' = B' \cdot A$$

where B' contains the subject's head pose and iris positions in the image frame, A is the calibration matrix, and C' is the estimated gaze position.

Accuracy

One measure for the potential accuracy of this method is given by the error in the calculated eye gaze positions in the original calibration frames. An ideal calibration would give calculated gaze positions identical to the original fixation points. Smaller deviations from this ideal generally indicate a higher quality in the calibration.

We have found typical values for the average error in the calculated gaze positions in the original calibration frames to be 3% of the screen in the horizontal direction and 10% of the screen in the vertical direction. On a 17" computer screen (32cm x 24cm), these values correspond to 1.0cm in the horizontal direction and 2.4cm in the vertical direction.

There are three possible sources for error in the estimated gaze positions.

- (1) The image feature locations may be inaccurately determined.
- (2) The relationship between B and C , above, may not be adequately represented by a linear equation.
- (3) The subject may not have actually been looking directly at the fixation points in the calibration frames.

(1) Our calculated location for image features is accurate to within at least one pixel. Generally, sub-pixel accuracy is achieved. The head pose algorithm requires four non-coplanar points, but uses any additional points to minimize error in the calculated rotation and translation. Since we use five points, inaccuracies in locating the individual ball positions will tend to cancel each other out. Error in locating the iris positions is more detrimental to the calculated eye gaze positions. In our experiments a change in the horizontal location of either iris by 0.5 pixels changes the calculated horizontal gaze position by approximately 2% of the screen, or 0.6cm on a 32cm x 24cm screen. Our estimation of the subject's vertical gaze position

is less accurate than the estimation of horizontal gaze position because occlusion by the eyelids makes determination of the vertical position of the iris more error prone.

(2) For large viewing angles, the relationship between the subject's head pose and iris positions and the subject's gaze position will not be linear. Our method could possibly be made more accurate by using higher order calibration equations. We have found the accuracy given by a linear mapping to be sufficient in practice.

(3) A fundamental problem facing all eyetrackers is that a subject need not fixate any closer than one degree of an object he would like to focus his attention on, since he is able to see clearly anything within the one degree foveal region. A subject can also move his attention around within the one degree foveal region (without eye movements), so that no eye tracker can determine what the subject is paying attention to with greater accuracy than one degree, no matter how accurately the eye tracker measures the subject's gaze position. If the subject is 46cm (18") from the screen, 1 degree is 0.8 cm. Thus there is a theoretical limit on the accuracy of the subject's calculated gaze position.

We have used one calibration frame for each fixation point. Though the subject is often not looking precisely at the fixation point in the calibration frames, we might compensate by averaging the subject's fixation positions over a small time interval around the current calibration frame. We expect the subject's eyes to constantly make small movements near the fixation point, so including several image frames before and after the current calibration frame could provide robustness against small inaccuracies in our assumptions about the subject's gaze position.

This method has given very convincing calculated gaze positions as a subject looked at each point in a 5x5 grid (after calibrating with 3x3 grids). The accuracy of collected data can be ensured by having the subject look at the 9xNxN fixation points both at the beginning and at the end of a lengthy experiment. Only the calibration session at the beginning is used to determine the best linear map. This map may be used to calculate the error in the subject's calculated gaze position as he looks at the fixation points at the end of the experiment. We have found the average error in the calculated gaze position to remain at approximately 3% of the screen in the horizontal direction, and 10% of the screen in the vertical direction, even after experiments up to an hour in length.

Since the spectacle framework is not uncomfortable for subjects, long experiments are possible. As long as

the headframe does not move with respect to the head, the calibration will remain accurate.

Conclusion

We have found this eyetracking method to be very flexible and reliable in practice. The 1-2 cm accuracy on a 17 inch computer screen is sufficient for many psychological applications. The subject's ability to talk provides greater flexibility in experimental design than methods requiring head fixation, and makes interaction with the subject during experiments more natural.

This method makes unnecessary the placement of an LED or other object near or in the subject's eye, and does not require head fixation with a chin rest or bite bar. While it fails to be completely noninvasive, since subjects must wear the spectacle framework, our subjects have found the spectacle framework comfortable enough to wear for extended periods.

The equipment required to use this eyetracking method is easily acquired. The necessary lighting may be provided by standard office lamps. A common color video camera suffices to capture the video stream. The only special item required is the spectacle framework, which can be constructed from readily available, inexpensive materials. Since no unusual or expensive equipment is needed, this eyetracking method is quite accessible.

Acknowledgements

This work was supported in part by NSF research grant No. IIS-9812714 and by the Caltech Neuromorphic Engineering Research Center.

References

- [1] Bishop, C.M., *Neural Networks for Pattern Recognition*, Clarendon, Oxford, 1995.
- [2] Dementhon, D.F. and Davis, L.S., "Model-Based Object Pose in 25 Lines of Code," *International Journal of Computer Vision*, Vol 15, pp. 123-141, 1995.
- [3] DiScenna, A.O., Das, V., et al., "Evaluation of a video tracking device for measurement of horizontal and vertical eye rotations during locomotion," *Journal of Neuroscience Methods*, Vol. 58, pp. 89-94, 1995.
- [4] Merchant, J., Morrissette, R., and Porterfield, J.L., "Remote Measurement of Eye Direction Allowing Subject Motion Over One Cubic Foot of Space," *IEEE Trans. Biomed. Electron.*, Vol. BME-21, No. 4, pp. 309-317, 1974.
- [5] Xie, X., Sudhakar, R., and Zhuang, H., "Real-Time Eye Feature Tracking from a Video Image Sequence Using Kalman Filter," *IEEE Trans. Systems, Man, and Cybernetics*, Vol. 25, No. 12, pp. 1568-1577, 1995.