A New Methodology for Determining Point-of-Gaze in Head-Mounted Eye Tracking Systems

Lawrence H. Yu, *Member, IEEE,* and Moshe Eizenman*

Abstract—**The ability to determine point-of-gaze with respect to an observed scene provides significant insight into human cognitive processes, since shifts in gaze position are generally guided by shifts in attentional focus. Using a head-mounted eye tracking system, a new methodology based on four or more point correspondences in two views was developed to reconstruct the subject's point-of-gaze. For exact point correspondences, 95% of the reconstruction errors are less than 0.32° when the homography algorithm with distortion compensation is used to determine gaze position. In a typical visual scanning experiment, 95% of the reconstruction errors are less than 0.90°. Analysis of normalization techniques that reduce the sensitivity of the homography algorithm to input errors suggests that the point correspondences should be arranged in a radially symmetric distribution around the area to be scanned. The new methodology was used in a clinical study on visual selective attention and mood disorders; this study showed that depressed subjects spent significantly more time looking at images with dysphoric themes than normal control subjects.**

Index Terms—**Eye tracking systems, homography, point-of-gaze, visual selective attention.**

I. Introduction

VISUAL information plays a crucial role in our ability to interact with the world. Traditionally, we tend to think of the eyes as passive receivers that relay the information required for a particular task to the brain. The role of eye movements as potential indicators of attentional behavior is often overlooked, since our eye movements are generally inaccessible to conscious scrutiny. Under normal viewing conditions, eye movements are automatic in that individuals commonly look at stimuli that attract their attention [1]; indeed, shifts in gaze positions closely follow and are guided by shifts in attentional focus [2], [3]. It is this insight into human cognitive processes that motivates many of the practical applications for point-of-gaze tracking technologies. Point-of-gaze tracking technologies can be used, for example, to measure and analyze the visual scan patterns of pilots in the cockpit of an aircraft [4]. By comparing the scanning behavior of expert and amateur pilots in standard operational sequences, efficient scanning strategies can be identified, and inferences can be made between visual scan patterns and hazard perception abilities. Point-of-gaze tracking technologies are also commonly used in the design and evaluation of human-machine interfaces [5], [6].

The objective of a point-of-gaze estimation methodology is to calculate the intersection of the gaze vector with the observed scene, so that the elements in a visual scene that are being fixated upon by the subject can be determined. Head-mounted eye tracking systems are the preferred choice for estimating the gaze vector in applications that require accurate point-of-gaze estimates while allowing abrupt and relatively free head movements. Most head-mounted eye tracking systems include a head-mounted video camera to continuously record the scene. The use of a scene camera allows eye position estimates to be superimposed on images of the scene, enabling real-time viewing of the subject's point-of-gaze.

In order to estimate point-of-gaze with head-mounted eye tracking systems, knowledge of the following variables is required: the angular rotation of the eye relative to the head, the position of the head relative to the scene, and the locations of objects in the observed scene. The angular position of the eye relative to the head can be measured by head-mounted eye trackers [7]. The position of the head relative to the scene is typically measured by position sensors that determine the three-dimensional (3-D) head position with respect to a fixed coordinate system. Various types of transducers (magnetic, ultrasonic, mechanical, etc.) are used to sense head position, of which the most commonly used is the magnetic position transducer [8]. If the objects in the visual field are also defined with respect to the fixed coordinate system, then Euclidean transformations between the coordinate systems centered on the eye and head [9] can be used to calculate where the gaze vector intersects with the defined objects.

The above approach has several limitations: 1) high system complexity due to the use of separate eye and head tracking systems; 2) restricted subject mobility due to the limited range of the head sensor; and 3) susceptibility to distortion from electromagnetic interference and ferrous materials (for magnetic position sensors). Also, since accurate measurements of the 3-D positions of objects in the environment are required, the ability to use this system in applications that require portability and a changing visual scene is limited.

The point-of-gaze estimation methodology presented in this paper overcomes all of the above limitations. The new methodology uses features extracted from the video of the scene camera to determine the position of the head relative to the objects in the scene. Since the eye position data is provided with respect to the head, these two data sets are readily combined to deter-

Manuscript received June 3, 2002; revised January 14, 2004. The work of L. H. Yu was supported by a scholarship from the Natural Sciences and Engineering Research Council (NSERC), Canada. The work of M. Eizenman was supported by the Canadian Institute of Health Research under Grant 36505. *Asterisk indicates corresponding author.*

L. H. Yu is with the Department of Electrical and Computer Engineering, and the Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, ON M5S 3G9, Canada.

*M. Eizenman is with the Departments of Ophthalmology and Electrical and Computer Engineering, and the Institute of Biomaterials and Biomedical Engineering, University of Toronto, 4 Taddle Creek Road, Toronto, ON M5S 3G9, Canada (e-mail: eizenm@ecf.utoronto.ca).

Digital Object Identifier 10.1109/TBME.2004.831523

mine the fixation behavior on objects in the scene. Section II of this paper provides a description of the point-of-gaze estimation methodology, and the performance of the methodology is assessed in Section III. The utility of the methodology is subsequently demonstrated in a clinical study on visual selective attention in major depressive disorder.

II. Description of Methodology

A. Approach

In head-mounted eye tracking systems where eye position estimates are superimposed on images from a scene camera, the following steps can be taken to determine fixation behavior on objects in a static scene: 1) define the objects of interest on a single image from the scene camera (denoted as the *reference image*); 2) determine the geometric relationships (mappings) between objects of interest in the reference image and each subsequent image; and 3) use these mappings to transfer eye position data to the reference image, where the point-of-gaze relative to the objects defined in step 1 can be determined.

In order to determine the image-to-image mappings, the point-of-gaze estimation methodology uses two-dimensional (2-D) feature correspondences. Each frame in the video sequence is processed to locate, extract, and label 2-D features (typically points) in the scene. The detected 2-D points are associated with their corresponding points in other frames to form a set of *point correspondences*. These point correspondences are then used to solve the two-view transfer problem, so that the eye position data can be transferred to the reference image where the objects are defined.

B. Two-View Transfer

The two-view transfer problem is concerned with the transfer of features seen in one view to a second view, given the location of features available in each image coordinate frame. In this paper, bold letters are used to denote vectors (e.g. $\mathbf{x}$), while typewritten upper case letters are used to denote matrices (e.g. $\mathtt{A}$). Homogeneous coordinates are used to represent both 2-D and 3-D points; since homogeneous coordinates are defined up to a scalar, the scale factor S will be explicitly included in the expression when $S \neq 1$.

Consider the situation where two images, $\mathcal{I}$ and $\mathcal{I}'$, are obtained at two different time instants, t_1 and t_2, by a camera moving in a static environment with an unknown trajectory (see Fig. 1). If N feature points $\{\mathbf{M}_i | i = 1, \ldots, N\}$ are visible to the camera at both time instants, then N point correspondences: $\{\mathbf{m}_i \longleftrightarrow \mathbf{m}'_i | i = 1, \ldots, N\}$ are formed in $\mathcal{I}$ and $\mathcal{I}'$ from the projection of these 3-D points to the respective image planes. Given N distinct point correspondences, we are interested in the mapping of an arbitrary point $\mathbf{m}'_{N+1}$ specified in image plane $\mathcal{I}'$ to its corresponding position $\mathbf{m}_{N+1}$ in image plane $\mathcal{I}$. To solve this problem, the image acquisition process first must be characterized to determine the appropriate mapping between a 3-D point $\mathbf{M} = [X, Y, Z]^T$ in the world coordinate system and its 2-D image coordinates $\mathbf{m} = [x, y]^T$. The standard photogrammetric camera model is an idealized geometric model that accounts for perspective projection and the transformations between the different coordinate systems used [10]. Using this

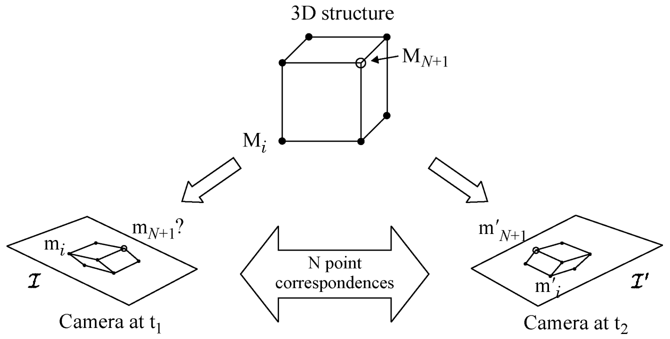


Fig. 1. Two-view transfer illustrated using two images gathered by a monocular camera in a static scene.

model, the mapping between Euclidean 3-D space coordinates (in units of length) and 2-D image coordinates (in pixel coordinates) can be expressed as a 3 × 4 projection matrix based on ten camera parameters (four intrinsic and six extrinsic)

$$S \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = \overbrace{\begin{bmatrix} f_x & 0 & x_o \\ 0 & f_y & y_o \\ 0 & 0 & 1 \end{bmatrix}}^{\mathtt{C}} [\mathtt{R} \quad \mathbf{t}] \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix} \tag{1}$$

where $\mathtt{R}$ is a 3 × 3 rotation matrix, $\mathbf{t}$ is a 3 × 1 translation vector, and $\mathtt{C}$ is the intrinsic matrix. The six extrinsic camera parameters define the position and orientation of the camera with respect to the scene, and consist of the three rotation angles that uniquely define $\mathtt{R}$ and the three components of $\mathbf{t}$. The four intrinsic camera parameters in $\mathtt{C}$ determine how the image coordinates of a 3-D point are derived, given its 3-D position with respect to the camera. The principal point (x_o, y_o) represents the pixel coordinates of the intersection of the optical axis and the digitized image. The ability to model rectangular pixels is provided by f_x and f_y, which represent the focal length in units of horizontal and vertical pixels, respectively.

Given two perspective views of a single rigid object obtained using (1), it can be shown (using epipolar geometry) that the two-view transfer problem cannot be solved uniquely regardless of the number of point correspondences used [11]. If the configuration of the object points in 3-D space is restricted to coplanar regions, however, a point-to-point mapping can be calculated between two perspective views with only point correspondence information. This is a reasonable simplification in many studies of visual scanning behavior where the visual stimuli are presented on a computer screen or projected onto a flat screen. Given N point correspondences $\{\mathbf{m}_i \longleftrightarrow \mathbf{m}'_i | i = 1, \ldots, N\}$ in two views and restricting the 3-D point configuration to be coplanar, the point-to-point mapping can be expressed as a 3 × 3 invertible matrix $\mathtt{H}$ called a homography

$$S\mathbf{m}'_i = \mathtt{H}\mathbf{m}_i = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \mathbf{m}_i. \tag{2}$$

Numerous methods are available for estimating the homography $\mathtt{H}$ [12]. In this paper, the approach is based on a linear homography algorithm [13] that provides a unique closed-form solution for $\mathtt{H}$, and avoids the high computational complexity and convergence issues of nonlinear iterative minimization algorithms. For each point correspondence, (2) yields two equations that are linear in the matrix elements of $\mathtt{H}$, so solving for $\mathtt{H}$ is equivalent to solving for the vector $\mathbf{h}$ in the homogeneous matrix equation

$$\mathtt{A}\mathbf{h} = \mathbf{0} \tag{3}$$

where $\mathbf{h} = [h_{11}, h_{12}, h_{13}, h_{21}, h_{22}, h_{23}, h_{31}, h_{32}, h_{33}]^T$ is formed from the elements of the matrix $\mathtt{H}$, and the $2N \times 9$ matrix $\mathtt{A}$ is formed by vertically stacking the matrices $\{\mathtt{A}_i^T | i = 1, \ldots, N\}$

$$\mathtt{A}_i^T = \begin{bmatrix} x_i & y_i & 1 & 0 & 0 & 0 & -x_i x_i' & -y_i x_i' & -x_i' \\ 0 & 0 & 0 & x_i & y_i & 1 & -x_i y_i' & -y_i y_i' & -y_i' \end{bmatrix}. \tag{4}$$

With four noncollinear point correspondences, it is possible to solve for $\mathtt{H}$ uniquely; taking the singular value decomposition of $\mathtt{A}$, the solution will be $\mathbf{s}_9$, the right singular vector that corresponds to the smallest singular value of $\mathtt{A}$ (0 in this case). With $N > 4$ point correspondences where no $(N-1)$ of the N point correspondences are collinear, there is no $\mathbf{h} \neq \mathbf{0}$ that solves (3) exactly, so a total least squares (TLS) approach is used to estimate $\mathbf{h}$ that minimizes $||\mathtt{A}\mathbf{h}||^2$, the squared Euclidean vector norm of $\mathtt{A}\mathbf{h}$. Using eigenanalysis to solve this minimization problem subject to $||\mathbf{h}|| = 1$, the best estimate of $\mathbf{h}$ must be the eigenvector corresponding to the smallest eigenvalue of $\mathtt{A}^T\mathtt{A}$, or the equivalent right singular vector $\mathbf{s}_9$ corresponding to the smallest singular value of $\mathtt{A}$.

C. Sensitivity to Errors in the Input Point Correspondences

An important issue regarding the homography algorithm is its sensitivity to errors in the estimation of the point correspondences. In this section, modifications to the point-of-gaze estimation methodology that improve its robustness to measurement errors in the point correspondences are explored. When more than four point correspondences are available, an appropriate input data normalization can be derived using an approach similar to that taken for estimation of the fundamental matrix [14]. If we denote $\mathbf{u}_i = [x_i, y_i, 1]^T$ and $\mathbf{v}_i = [x_i', y_i', 1]^T$ as the observed point correspondences in the reference and subsequent images, respectively, and $\bar{\mathbf{u}}_i = [\bar{x}_i, \bar{y}_i, 1]^T$ and $\bar{\mathbf{v}}_i = [\bar{x}_i', \bar{y}_i', 1]^T$ as the respective normalized point correspondences, we can define two nonsingular 3×3 normalization matrices, $\mathtt{J}$ and $\mathtt{K}$, such that $\bar{\mathbf{u}}_i = \mathtt{J}\mathbf{u}_i$ and $\bar{\mathbf{v}}_i = \mathtt{K}\mathbf{v}_i$. Note that the last rows of $\mathtt{J}$ and $\mathtt{K}$ are constrained to be [0,0,1] so that the third components of $\mathbf{u}_i$ and $\mathbf{v}_i$ remain 1. If the relationship between the normalized point correspondences $\bar{\mathbf{u}}_i$ and $\bar{\mathbf{v}}_i$ is written as $\bar{\mathbf{v}}_i = \bar{\mathtt{H}}\bar{\mathbf{u}}_i$, then the homography between the point correspondences can be expressed as $\mathtt{H} = (\mathtt{K}^{-1}\bar{\mathtt{H}}\mathtt{J})$.

The matrix $\mathtt{A}$ in (3) can be written as $\mathtt{A} = \mathtt{A}_0 + \mathtt{D}$, where $\mathtt{A}_0$ is the actual (but unknown) $2N \times 9$ matrix associated with the case where no input measurement errors are present, and $\mathtt{D}$ is the error matrix. Since the point correspondences in the reference frame are manually selected and verified, it is reasonable to assume that there are no errors in the reference image. If we assume measurement errors in each point correspondence in subsequent images so that $(x_i', y_i') = (x_{i0}', y_{i0}') + (\epsilon_{x_i'}, \epsilon_{y_i'})$, then the matrix $\mathtt{D}$ is formed by vertically stacking the matrices $\{\mathtt{D}_i^T | i = 1, \ldots, N\}$

$$\mathtt{D}_i^T = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -\epsilon_{x_i'} x_i & -\epsilon_{x_i'} y_i & -\epsilon_{x_i'} \\ 0 & 0 & 0 & 0 & 0 & 0 & -\epsilon_{y_i'} x_i & -\epsilon_{y_i'} y_i & -\epsilon_{y_i'} \end{bmatrix}. \tag{5}$$

Using a linear first-order approximation of the error in the eigenvector $\mathbf{s}_9$, it is demonstrated in [14] that the constraint $\mathrm{E}(\mathtt{D}^T\mathtt{D}) = c\mathtt{I}$ must hold in order to ensure unbiased estimates of $\mathbf{h}$. Using a TLS technique that accounts for the error-free columns in $\mathtt{A}$ to estimate $\bar{\mathtt{H}}$ [15], the requirement can be modified to

$$\mathrm{E}(\mathtt{D}^T\mathtt{D}) = \sum_{i=1}^{N} \mathrm{E}\left(\mathtt{D}_i\mathtt{D}_i^T\right) = \sum_{i=1}^{N} \mathrm{Cov}(\mathtt{D}_i) = c\tilde{\mathtt{I}} \tag{6}$$

where $\tilde{\mathtt{I}}$ is defined as diag(0,0,0,0,0,0,1,1,1).

In order to minimize the sensitivity to errors for the normalized point correspondences, the condition expressed by (6) has to be satisfied for the covariance of $\bar{\mathtt{D}}_i$

$$\sum_{i=1}^{N} \mathrm{Cov}(\bar{\mathtt{D}}_i) = c\tilde{\mathtt{I}}. \tag{7}$$

Using (7), the normalization matrices $\mathtt{J}$ and $\mathtt{K}$ are derived in the Appendix. As shown in the Appendix, $\mathtt{K}$ can be set to the identity matrix $\mathtt{I}$, and $\mathtt{J}$ is an upper triangular matrix such that the following equation is satisfied:

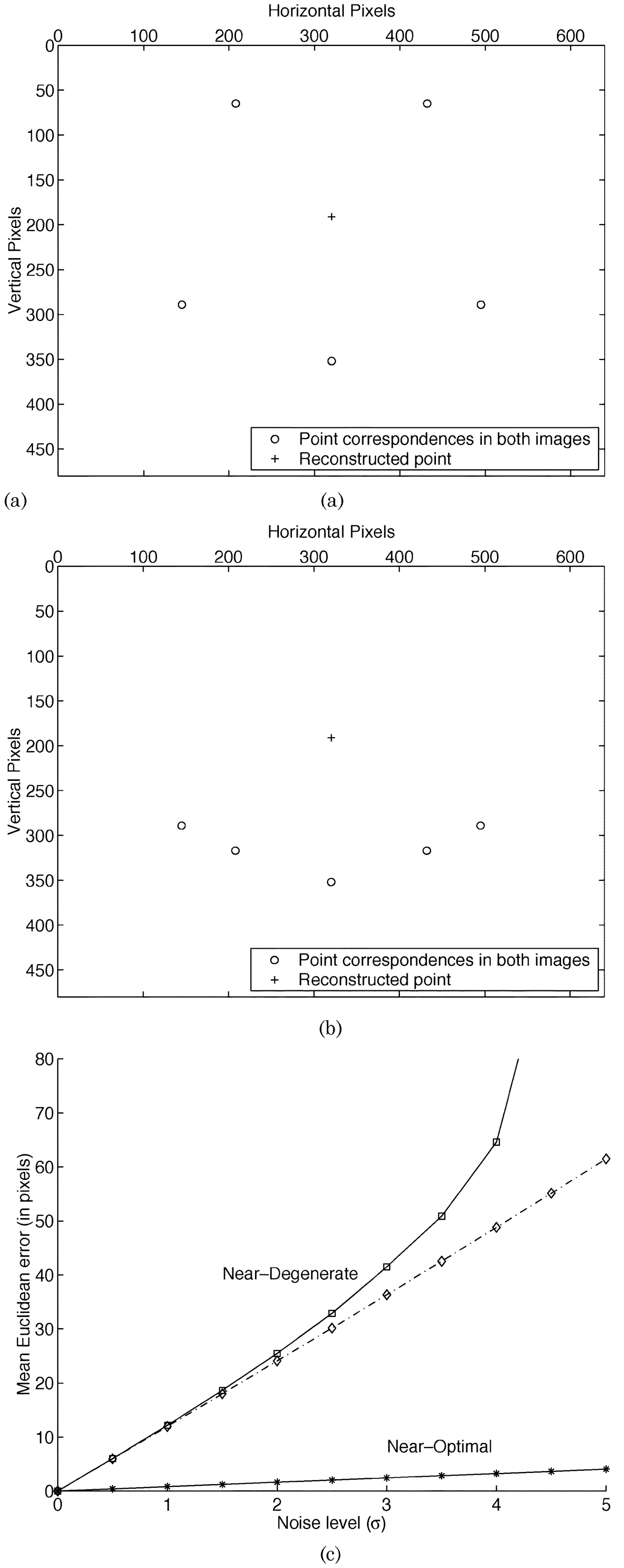
$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{u}_i\mathbf{u}_i^T = \mathtt{J}^{-1}\mathtt{J}^{-T}. \tag{8}$$

The above normalization matrix $\mathtt{J}$ is similar to that proposed in [16] for calculating the fundamental matrix. This normalization translates the centroid of the point correspondences in the reference image to the origin and sets the two principal moments of the point correspondences to unity

$$\frac{1}{N}\sum_{i=1}^{N} \bar{x}_i^2 = \frac{1}{N}\sum_{i=1}^{N} \bar{y}_i^2 = 1$$
$$\sum_{i=1}^{N} \bar{x}_i = \sum_{i=1}^{N} \bar{y}_i = \sum_{i=1}^{N} \bar{x}_i\bar{y}_i. = 0 \tag{9}$$

The effect of this anisotropic scaling (different scaling for the x and y coordinates) is to form an approximately symmetric circular cloud of points of radius one about the origin.

Monte Carlo simulations were performed to characterize the effect of input data normalization on the performance of the homography algorithm for two configurations of point correspondences. In the first configuration shown in Fig. 2(a), the point correspondences were distributed in an approximately symmetric circular pattern about the point to be reconstructed, i.e. a near-optimal configuration. In the second configuration shown in Fig. 2(b), the point correspondences were distributed such that four of the points were nearly collinear, i.e. a near-degenerate configuration. In the simulations, the point correspondences in the reference image were error-free, while zero-mean Gaussian noise (with standard deviations ranging



(a) (a)

(b)

(c)

Fig. 2. Comparison of reconstruction errors in reference image for different point configurations with additive zero-mean Gaussian noise σ in subsequent images. (a) Near-optimal configuration of point correspondences; (b) near-degenerate configuration of point correspondences; (c) reconstruction error versus noise level σ for 1) near-optimal configuration with normalization ($+$) and without normalization ($\times$); and 2) near-degenerate configuration with normalization ($\diamond$) and without normalization ($\Box$). Each point in the plot represents the mean of 1000 trials.

from $\sigma = 0$ to 5 pixels) was added to the point correspondences in subsequent images. For simplicity, these point correspondences were obtained using the identity homography mapping $\mathtt{H} = \mathtt{I}$. Using the resulting point correspondences, estimates of the homography mapping, denoted by $\hat{\mathtt{H}}$, were obtained. The actual positions, $\mathbf{u}_i$, and the estimated positions, $\hat{\mathbf{u}}_i = \hat{\mathtt{H}}^{-1}\mathbf{v}_i$, of the reconstructed point were then calculated, along with the mean Euclidean error $\|\mathbf{u}_i - \hat{\mathbf{u}}_i\|$. Fig. 2(c) shows the mean Euclidean error as a function of the standard deviation σ of the Gaussian noise. When a near-degenerate point configuration is used, the relative performance of the normalized algorithm improves as the standard deviation of the Gaussian noise increases. Examination of the reconstruction errors suggests that the effect of the normalization is to reduce the *bias* of the errors. As expected, when the point correspondences are arranged in a near-optimal configuration, the normalization had a minimal effect on the reconstruction errors.

In conclusion, point correspondences that are evenly distributed in a radially symmetric manner about the point to be reconstructed are more robust to input measurement errors, and do not require normalization. The effect of normalization is related to the condition number of the matrix $\mathtt{A}$ defined in (3). A large condition number for $\mathtt{A}$ (i.e. $\mathtt{A}$ is nearly rank deficient) will amplify the effect of noise on the reconstruction error [12]. Appropriate normalization enhances the numerical stability of the algorithm in such cases and reduces the bias of the reconstruction errors.

D. Compensation for Geometric Distortion

In head-mounted eye tracking systems, the scene camera often has a wide-angle lens to capture large portions of the subject's field of view. For such optical systems, perspective projection is typically insufficient to model the imaging process accurately, since nonlinear distortion is introduced to the optical paths and the resulting images. This geometric distortion is commonly decomposed into radial and tangential components [10], and incorporated into (1) to form the augmented perspective projection camera model [17], [18]

$$\begin{bmatrix} \tilde{x}_i \\ \tilde{y}_i \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & x_o \\ 0 & f_y & y_o \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{x}_i + \delta_r(\hat{x}_i) + \delta_t(\hat{x}_i) \\ \hat{y}_i + \delta_r(\hat{y}_i) + \delta_t(\hat{y}_i) \\ 1 \end{bmatrix};$$

$$S \begin{bmatrix} \hat{x}_i \\ \hat{y}_i \\ 1 \end{bmatrix} = [\mathtt{R} \quad \mathbf{t}] \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix} \tag{10}$$

where the radial $\delta_r(\cdot)$ and tangential $\delta_t(\cdot)$ distortions are defined as

$$\begin{bmatrix} \delta_r(\hat{x}_i) \\ \delta_r(\hat{y}_i) \end{bmatrix} = \begin{bmatrix} \left(a_1\hat{r}_i^2 + a_2\hat{r}_i^4 + a_3\hat{r}_i^6\right)\hat{x}_i \\ \left(a_1\hat{r}_i^2 + a_2\hat{r}_i^4 + a_3\hat{r}_i^6\right)\hat{y}_i \end{bmatrix} \tag{11}$$

$$\begin{bmatrix} \delta_t(\hat{x}_i) \\ \delta_t(\hat{y}_i) \end{bmatrix} = \begin{bmatrix} 2b_1\hat{x}_i\hat{y}_i + b_2\left(\hat{r}_i^2 + 2\hat{x}_i^2\right) \\ 2b_2\hat{x}_i\hat{y}_i + b_1\left(\hat{r}_i^2 + 2\hat{y}_i^2\right) \end{bmatrix} \tag{12}$$

and $\hat{r}_i^2 = \hat{x}_i^2 + \hat{y}_i^2$. Three distortion coefficients (a_1, a_2, a_3) are employed to represent radial distortion, while two coefficients (b_1, b_2) are used to denote tangential distortion.

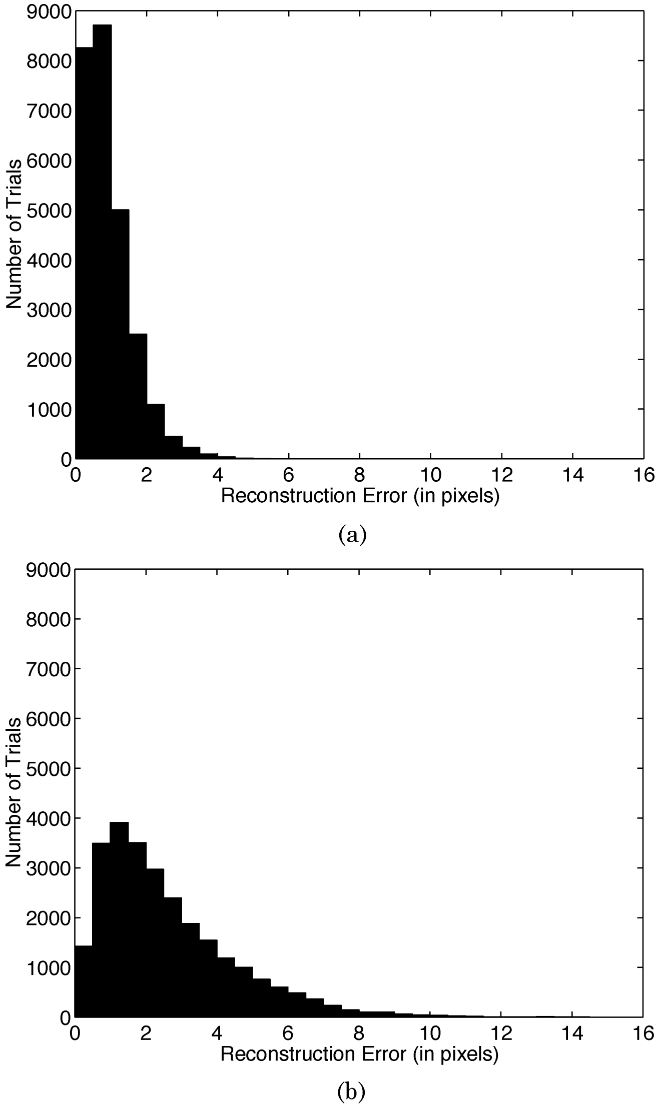
Given this camera model, it is possible to compensate for geometric distortion by converting the observable image coordinates $(\tilde{x}_i, \tilde{y}_i)$ to the ideal (but unobservable) image coordinates (x_i, y_i) used in (1). First, the camera intrinsic parameters and distortion coefficients are calculated using a flexible technique implemented in a MATLAB toolbox that relies on multiple views of a single 2-D planar calibration pattern [18]. A key advantage of this technique is that planar calibration patterns are simple and inexpensive to produce, unlike 3-D calibration objects that require perfectly orthogonal planes. Using the intrinsic parameters and the distortion coefficients, the inverse mapping of $(\tilde{x}_i, \tilde{y}_i)$ to $(\hat{x}_i, \hat{y}_i)$ can be estimated iteratively [17], and the ideal image coordinates (x_i, y_i) can be subsequently calculated by multiplying the intrinsic matrix C by $(\hat{x}_i, \hat{y}_i)$ in homogeneous coordinates.

III. Performance Analysis

In Sections III-A and III-B, the performance of the point-of-gaze estimation methodology is evaluated under ideal and typical experimental conditions, respectively. The criterion for evaluating the performance is *point-of-gaze reconstruction error*, which is defined as the Euclidean distance between the actual and the reconstructed point-of-gaze in the reference image. In order to eliminate the inherent variability of actual eye position data, the reconstruction errors were calculated for a set of points that were tracked in each image and transferred by the homography algorithm to the reference image.

A. Performance Using Exact Point Correspondences

To evaluate the performance of the methodology under ideal conditions where exact point correspondences are available, a set of 25 images consisting of multiple views of a planar black and white checkerboard grid (8 × 6 squares) was obtained by translating and rotating the grid in front of the stationary scene camera component of the eye tracking system. For the experiments described in this paper, a miniature charge-coupled device camera (ELMO ME411, Nagoya, Japan) with a viewing angle of 92.1° H × 69.1° V was used to obtain images with dimensions of 640 × 480 pixels. The entire checkerboard pattern was visible in each of the 25 images which were selected to simulate head movements significantly larger than those expected during visual scanning experiments. From an initial position where the checkerboard was centered in the image at a distance of 28 cm from the camera, the distances of the checkerboard from the camera ranged from 18 to 35 cm, while the translational distances in the x and y directions ranged from -20 to 30 cm. The orientation of the camera relative to the checkerboard ranged from $-25°$ to $115°$ in roll, and $-55°$ to $55°$ in pitch and yaw. Point correspondences were subsequently extracted with sub-pixel accuracy from the vertices of the checkerboard [18], [19]. The four outermost vertices of the checkerboard pattern were used to calculate the homography mapping between views, while the remaining 44 vertices of the checkerboard pattern were designated as the points to be reconstructed. This is consistent with typical experimental setups, where the objects of interest (and hence the points-of-gaze) are usually situated within a region formed by four or more point correspondences.



(a)

(b)

Fig. 3. Histograms of reconstruction errors for the homography algorithm with exact point correspondences. (a) With distortion compensation; (b) without distortion compensation.

To determine the performance of the methodology with and without distortion compensation, reconstruction errors were calculated for each of the 600 possible image pairs (selected from the set of 25 images) for a total of 26 400 observations. Histograms of the reconstruction errors obtained using the homography algorithm with and without distortion compensation are depicted in Fig. 3; note that reconstruction errors are expressed in pixels, but may also be written in angular form using the conversion factor of $0.1439°/\mathrm{pixel}$.

Using distortion compensation, the mean and median reconstruction errors were 0.92 and 0.75 pixels, respectively, with a maximum error of 6.00 pixels. As demonstrated in Fig. 3(a), 95% of the reconstruction errors were smaller than 2.2 pixels, while 92.7% of the errors were smaller than 2 pixels, and 64.3% were smaller than 1 pixel. When distortion compensation was not used, the reconstruction errors were significantly higher, with mean and median errors of 2.63 and 2.13 pixels, respectively, and a maximum error of 16.1 pixels [see Fig. 3(b)]. The largest differences in the performance of the methodology with and without distortion compensation were observed to occur

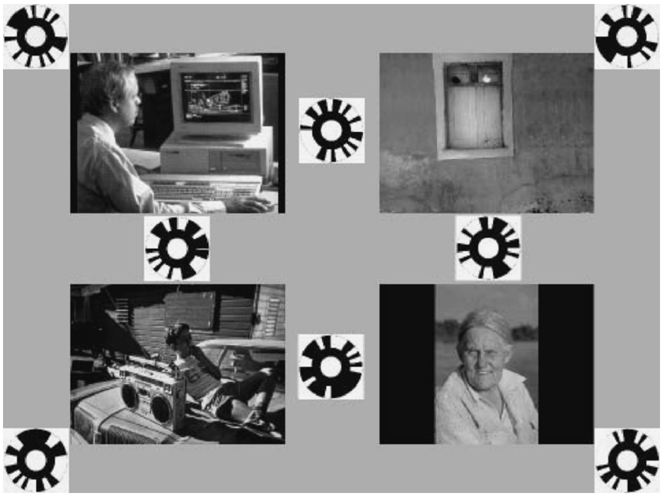


Fig. 4. Distinctive reference targets are used to determine point correspondences in the visual scene. The outer set of targets is used for the estimation of the homography mapping, while the inner set is used for the performance analysis.

when the point correspondences between image pairs suffer from varying degrees of distortions. For example, if one of the point correspondences moves from a region of low distortion in one image (central region) to a region of high distortion in the second image (peripheral region), while other point correspondences remain in a region of relatively low distortion, the algorithm without the distortion compensation can have relatively large errors.

The above results provide insight into the performance of the point-of-gaze estimation methodology under relatively ideal conditions. In the next section, the performance of the methodology is evaluated under conditions that are similar to those encountered in visual scanning experiments, i.e. where matched point correspondences cannot be consistently obtained with sub-pixel accuracy.

B. Performance in Typical Visual Scanning Experiments

In typical visual scanning experiments, subjects are presented with a set of visual stimuli, while the subjects' field of view is continuously recorded by the eye tracker's scene camera. Distinctive reference targets embedded in the periphery of the visual scene are typically used to accurately determine and match point correspondences (see Fig. 4). First, objects in the image that are consistent with the shape and intensity distribution of a reference target are detected. Each potential target is then identified by cross-correlating the pseudo-random binary sequence encoded in the outer portion of the target with a template that is unique for each target. This identification process is verified by comparing the relative locations of the potential targets to the known geometric distribution of the reference targets. Finally, the boundary points of the white circular region in the middle of each reference target are used to estimate the target center. Reference targets are, thus, detected and estimated in real-time (at a rate of 30 Hz) using a Pentium-II PC (233 MHz) equipped with a frame grabber (Meteor II, Matrox Inc., Montreal, Canada).

In order to estimate the reconstruction errors associated with such experiments, ten subjects were shown a single slide containing four reference targets in the periphery of the slide and four additional reference targets distributed through the central region (see Fig. 4). The four outer point correspondences were used to determine the homography mapping, and the reconstruction errors for the centers of the four inner targets were calculated. Using the homography algorithm with distortion compensation, the mean and median reconstruction errors across all subjects (based on 6600 video frames with four reconstructed points per frame) were 1.51 and 0.93 pixels, respectively, with a maximum error of 15.50 pixels. 95% of the reconstruction errors were smaller than 6.2 pixels, with 89.5% smaller than 2 pixels, and 64.3% smaller than 1 pixel.

In a set of similar experiments to test the performance under larger dynamic head movements, a subject was instructed to perform head movements that, in turn, spanned each of the six degrees of freedom: horizontal, vertical and torsional rotations; and horizontal, vertical, and zoom translations. The range of head movements performed by the subject exceeded the range of head movements encountered in a typical study; head rotations were in the range of ±15° in each direction, while head translations were in the range of ±10 cm in each direction. The reconstruction errors were similar to those observed in the previous experiment, with no apparent dependence on rotation angle or translation.

When the reconstruction errors in typical experiments are compared with the results in Section III-A, it is clear that the inability to consistently obtain exact estimates of the point correspondences increases the probability of relatively large errors (>2 pixels or 0.3°). With exact point correspondences, 95% of the reconstruction errors are less than 0.32°, compared to 0.90° in typical experiments. The above results can provide guidance with regard to the minimum required separation between objects in the field of view for visual scanning experiments. In order to ensure that the probability of the reconstructed point-of-gaze falling on the wrong object is less than 5%, the minimum separation between objects should be greater than 1°.

If the points of interest are not in the plane defined by the point correspondences, the reconstruction errors are dependent on the changes in the 3-D orientation of the plane defined by the point correspondences relative to the plane defined by the reference image. Changes in plane orientation are measured relative to the image plane of the scene camera. The technique described in this paper can be extended to work more generally in this situation, if the 3-D positions of the points of interest are known relative to the reference points. Without this additional information, errors due to the deviations from planarity can still be mitigated by 1) gathering intermediate reference images (with less substantial inter-frame camera motion) and 2) defining multiple sets of reference points so that each point of interest belongs to a plane defined by one of these sets.

IV. Implementation in a Clinical Study

The novel point-of-gaze estimation methodology was used in a study on visual selective attention in mood disorders (described in detail in [20]). This study *directly* examined visual selective attention by monitoring the point-of-gaze of subjects presented with multiple competing complex visual stimuli. The primary hypothesis of this study was that relative to normal controls, individuals with major depressive disorder would selectively attend to visual images with dysphoric themes. The depressed group for the study consisted of eight individuals who

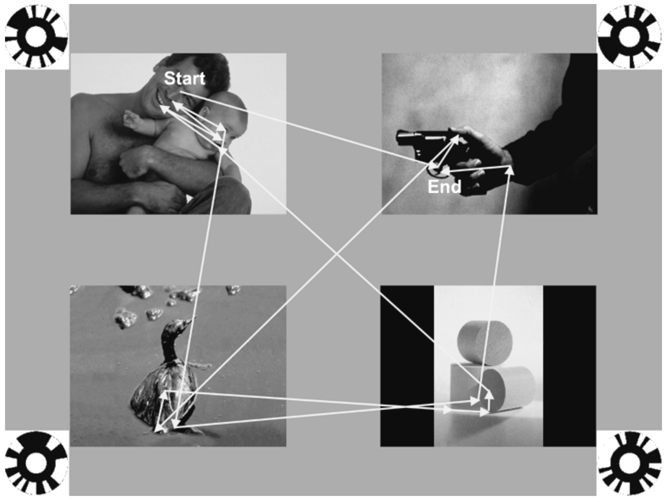


Fig. 5. Example of a study slide (top left: social theme; top right: threatening theme; bottom left: dysphoric theme; bottom right: neutral theme) with arrows representing saccadic eye movements.

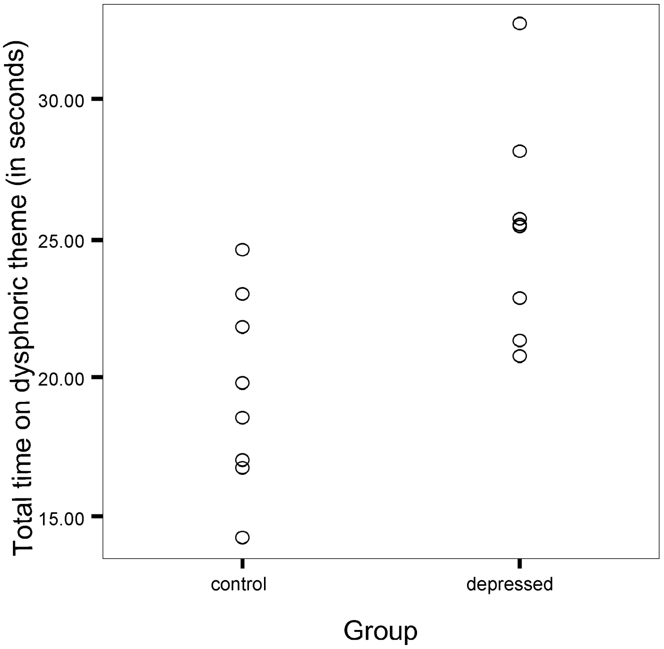


Fig. 6. Total viewing time of images with dysphoric theme by subject group. Each marker represents a single subject.

met the standard Clinical Interview for DSM-IV-TR (SCID) criteria for current major depression, and scored 16 or greater on the Beck Depression Inventory (BDI). The control group consisted of nine individuals who reported no psychological history and scored 5 or less on the BDI.

Subjects were fitted with a head-mounted eye tracking system [4], and seated in front of a portable projection screen where visual stimuli were back-projected. The visual stimuli consisted of a series of slides, with each slide containing four images. Four reference targets were placed in the corners of each slide. The images on each slide fell into four main categories: neutral stimuli, stimuli related to themes of loss and sadness (dysphoric), stimuli related to themes of threat and anxiety, and stimuli relating to themes of interpersonal attachment and social contact. The images were chosen based on the valence ratings provided by the International Affective Picture System (IAPS) [21], as well as the thematic content. Images relating to threatening and dysphoric themes had valences ranging from 2 to 4, while images relating to social themes had valences ranging from 6 to 8. A total of 15 slides, eight *study* slides and seven *neutral* slides, were shown to each subject for 10.5 s each. Each of the four images on a study slide was selected from each of the four themes previously listed (see Fig. 5). For each slide, a reference image was captured, and the boundaries of the four composite images were defined relative to the four reference targets on this image. The four reference targets were then tracked for the duration of the slide presentation, and the eye position data was mapped onto the reference image using the homography algorithm with distortion compensation. By calculating the amount of time the point-of-gaze fell within the boundaries of each image, the viewing times of each of the four images on each slide were obtained. The number of times that each subject directs and redirects attention to a particular image (or the viewing frequency) was also calculated. For each subject, viewing times and viewing frequencies on images with the same theme were summed up to generate the total viewing time and total viewing frequency for each theme. As demonstrated in Fig. 5, the visual scan path of each subject can be recorded and superimposed on each slide.

Fig. 6 depicts total fixation times on dysphoric images for each subject in both the control and depressed groups. The primary finding was that the mean of the total amount of time that the depressed group spent looking at images with dysphoric themes ($\text{mean } (M) = 25.32 \text{ s}, \text{standard deviation } (\text{SD}) = 3.89$) was significantly greater than that of the normal control group ($M = 19.35$ s, $\text{SD} = 3.30$), $t(15) = 3.42, p = .004$. The mean viewing frequency of the control group on images with dysphoric themes ($M = 17.11$, $\text{SD} = 4.51$) was not significantly different from that of the depressed group ($M = 16.38$, $\text{SD} = 5.24$). Similar statistical tests were performed on subsets of eight random images from the neutral slides. In all cases, the total viewing time of the subsets of neutral images did not differ significantly between the depressed and control groups, $t(15) < 1.0$, $p > .3$. In addition, no significant differences were found in the viewing patterns of the control and depressed groups due to the relative positions of the images on the screen, i.e. the mean of the total viewing times on the left right, top and bottom portions of the slides did not differ significantly, $0.40 > t(15) > 0.14$, $.694 < p < .892$. The above results suggest that depressed subjects selectively attend to stimuli with themes of loss or sadness, and that depression appears to influence the elaborative stages of processing when dysphoric images are viewed.

V. Conclusion

A new methodology to determine the point-of-gaze with a head-mounted eye tracking system has been presented. It combines the well-known homography algorithm with distortion compensation, to determine the point-of-gaze from point correspondences in images obtained by the eye tracker's scene camera. This methodology does not require either a separate head tracking system or accurate 3-D measurements of objects in the subject's field of view to determine the visual scanning behavior (i.e. viewing time and viewing frequency of each object). The point-of-gaze estimation methodology can be used to assess visual scanning patterns accurately (to less than 0.90°). As such, it can provide insights into selective attention processes that can aid in the diagnosis and evaluation of subjects with mood disorders. The reduced complexity of the methodology allows it to be used in applications that require portability, flexibility, and a changing visual scene.

Appendix
Derivation of Normalization Matrices

In order to satisfy (7), we need to determine the covariance of $\bar{\mathtt{D}}_i$. First, we denote the errors in the point correspondences as $\mathbf{e}_i = (\epsilon_{x'_i}, \epsilon_{y'_i}, 0)^T$ and the errors in the normalized point correspondences as $\bar{\mathbf{e}}_i = (\epsilon_{\bar{x}'_i}, \epsilon_{\bar{y}'_i}, 0)^T = \mathtt{K}\mathbf{e}_i$. Substituting x_i, y_i, $\epsilon_{x'_i}$, and $\epsilon_{y'_i}$ in (5) with their normalized counterparts $\bar{x}_i$, $\bar{y}_i$, $\epsilon_{\bar{x}'_i}$, and $\epsilon_{\bar{y}'_i}$ yields

$$\mathrm{Cov}(\bar{\mathtt{D}}_i) = \mathrm{E}\left(\epsilon^2_{\bar{x}'_i} + \epsilon^2_{\bar{y}'_i}\right) \begin{bmatrix} \ddots & \vdots & & & \vdots \\ \cdots & 0 & \cdots & \cdots & 0 \\ & \vdots & \bar{x}_i^2 & \bar{x}_i\bar{y}_i & \bar{x}_i \\ & \vdots & \bar{x}_i\bar{y}_i & \bar{y}_i^2 & \bar{y}_i \\ \cdots & 0 & \bar{x}_i & \bar{y}_i & 1 \end{bmatrix}. \tag{13}$$

The expression $\mathrm{E}(\epsilon^2_{\bar{x}'_i} + \epsilon^2_{\bar{y}'_i})$ can be evaluated by calculating the covariance matrix of the errors in the *normalized* point correspondences. From the definition of $\bar{\mathbf{e}}_i$, we have

$$\mathrm{Cov}(\bar{\mathbf{e}}_i) = \mathrm{E}\left(\bar{\mathbf{e}}_i\bar{\mathbf{e}}_i^T\right) = \mathrm{E}\left(\begin{bmatrix} \epsilon^2_{\bar{x}'_i} & \epsilon_{\bar{x}'_i}\epsilon_{\bar{y}'_i} & 0 \\ \epsilon_{\bar{x}'_i}\epsilon_{\bar{y}'_i} & \epsilon^2_{\bar{y}'_i} & 0 \\ 0 & 0 & 0 \end{bmatrix}\right). \tag{14}$$

Assuming that the errors in the second image are zero-mean independent identically distributed, i.e. for all i, $\mathrm{E}(\epsilon_{x'_i}) = \mathrm{E}(\epsilon_{y'_i}) = \mathrm{E}(\epsilon_{x'_i}\epsilon_{y'_i}) = 0$ and $\mathrm{E}(\epsilon^2_{x'_i}) = \mathrm{E}(\epsilon^2_{y'_i}) = \sigma_e^2$, we can also write $\mathrm{Cov}(\bar{\mathbf{e}}_i)$ in terms of the elements of $\mathtt{K}$

$$\begin{aligned} &\mathrm{Cov}(\bar{\mathbf{e}}_i) \\ &\quad = \mathtt{K}\mathrm{E}\left(\mathbf{e}_i\mathbf{e}_i^T\right)\mathtt{K}^T \\ &\quad = \mathtt{K}\begin{bmatrix} \sigma_e^2 & 0 & 0 \\ 0 & \sigma_e^2 & 0 \\ 0 & 0 & 0 \end{bmatrix}\mathtt{K}^T \\ &\quad = \sigma_e^2 \begin{bmatrix} k_{11}^2 + k_{12}^2 & k_{11}k_{21} + k_{12}k_{23} & 0 \\ k_{11}k_{21} + k_{12}k_{23} & k_{21}^2 + k_{22}^2 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned} \tag{15}$$

Equating (14) and (15), $\mathrm{E}(\epsilon^2_{\bar{x}'_i}) = \sigma_e^2(k_{11}^2 + k_{12}^2)$ and $\mathrm{E}(\epsilon^2_{\bar{y}'_i}) = \sigma_e^2(k_{21}^2 + k_{22}^2)$. Combining (13) with (7), the following condition has to be satisfied in order to minimize the sensitivity of the matrix $\mathtt{H}$ to errors in the input point correspondences:

$$\sigma_e^2\left(k_{11}^2 + k_{12}^2 + k_{21}^2 + k_{22}^2\right)\sum_{i=1}^{N}\begin{bmatrix} \bar{x}_i^2 & \bar{x}_i\bar{y}_i & \bar{x}_i \\ \bar{x}_i\bar{y}_i & \bar{y}_i^2 & \bar{y}_i \\ \bar{x}_i & \bar{y}_i & 1 \end{bmatrix} = c\mathtt{I}_{3\times 3}. \tag{16}$$

In (16), $c/(\sigma_e^2(k_{11}^2 + k_{12}^2 + k_{21}^2 + k_{22}^2))$ should be set to N to keep the third component of $\mathbf{u}_i$ equal to 1. Since c is an arbitrary constant, there are no constraints placed on the entries of the matrix $\mathtt{K}$. For convenience, $\mathtt{K}$ can be set to $\mathtt{I}$. The matrix $\mathtt{J}$ can be derved by rewriting (16) as

$$\sum_{i=1}^{N}\bar{\mathbf{u}}_i\bar{\mathbf{u}}_i^T = \sum_{i=1}^{N}\mathtt{J}\mathbf{u}_i\mathbf{u}_i^T\mathtt{J}^T = N\mathtt{I}_{3\times 3}. \tag{17}$$

Using Cholesky factorization, we can then solve for the upper triangular matrix $\mathtt{J}$, so that $(1/N)\sum_{i=1}^{N}\mathbf{u}_i\mathbf{u}_i^T = \mathtt{J}^{-1}\mathtt{J}^{-T}$.

References

[1] J. P. Jonides, "Voluntary versus automatic control over the mind's eye's movements," in *Attention and Performance IX*, J. Long and A. Baddeley, Eds. Hillsdale, NJ: Erlbaum, 1981, pp. 187–203.

[2] E. Kowler, "Eye movements," in *Visual Cognition*, S. M. Kosslyn and D. N. Osheron, Eds. Cambridge, MA: MIT Press, 1995, pp. 215–265.

[3] N. Moray, "Designing for attention," in *Attention: Selection, Awareness, and Control*, A. Baddeley and L. Weiskrantz, Eds. New York: Clarendon, 1993, pp. 111–134.

[4] P. A. Wetzel, G. Krueger-Anderson, C. Poprik, and P. Bascom, "An Eye Tracking System for Analysis of Pilots' Scan Paths," United States Air Force Armstrong Laboratory, Tech. Rep. AL/HR-TR-1996-0145, Apr. 1997.

[5] S. Merchant and T. Schnell, "Applying eye tracking as an alternative approach for activation of controls and functions in aircraft," in *Proc. 19th Dig. Avion. Syst. Conf.*, vol. 2, 2000, pp. 5A5/1–5A5/9.

[6] R. Sharma, V. I. Pavlovic, and T. S. Huang, "Toward multimodal human-computer interface," *Proc. IEEE*, vol. 86, pp. 853–869, May 1998.

[7] L. R. Young and D. Sheena, "Methods and designs—survey of eye movement recording methods," *Behav. Res. Meth. Instrum.*, vol. 7, no. 5, pp. 397–429, 1975.

[8] F. Raab, E. Blood, T. Steiner, and H. Jones, "Magnetic position and orientation tracking system," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-15, pp. 709–718, Nov. 1979.

[9] R. Allison, M. Eizenman, and B. Cheung, "Combined head and eye tracking system for dynamic testing of the vestibular system," *IEEE Trans. Biomed. Eng.*, vol. 43, pp. 1073–1082, Nov. 1996.

[10] *Manual of Photogrammetry*, Fourth ed., C. C. Slama, Ed., Amer. Soc. Photogrammetry and Remote Sensing, Falls Church, VA, 1980.

[11] Z. Zhang, "Determining the epipolar geometry and its uncertainty: a review," *Int. J. Comp. Vis.*, vol. 27, no. 2, pp. 161–195, 1998.

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. New York: Cambridge Univ. Press, 2000.

[13] A. Criminisi, I. Reid, and A. Zisserman, "A plane measuring device," *Image Vis. Comput.*, vol. 17, no. 8, pp. 625–634, 1999.

[14] M. Mühlich and R. Mester, "The role of total least squares in motion analysis," in *Proc. Eur. Conf. Comp. Vis.*, H. Burkhardt, Ed., New York, 1998, pp. 305–321.

[15] S. van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*. Philadelphia, PA: Soc. Ind. Appl. Math. (SIAM), 1991.

[16] R. I. Hartley, "In defence of the 8-point algorithm," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 580–592, June 1997.

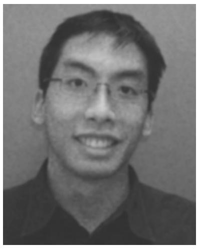
[17] J. Heikkilä and O. Silven, "A four-step camera calibration procedure with implicit image correction," in *Proc. IEEE Conf. CVPR*, 1997, pp. 1106–1112.

[18] J.-Y. Bouguet. (2001, Mar.) Camera Calibration Toolbox for MATLAB. California Institute of Technology, Department of Electrical Engineering, Pasadena, CA. [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc

[19] C. Harris and M. J. Stephens, "A combined corner and edge detector," in *Proc. 4th Alvey Vision Conf.*, 1988, pp. 147–152.

[20] M. Eizenman, L. H. Yu, L. Grupp, E. Eizenman, M. Ellenbogen, M. Gemar, and R. D. Levitan, "A naturalistic visual scanning approach to assess selective attention in major depressive disorder," *J. Psychiat. Res.*, vol. 118, no. 2, pp. 117–128, 2003.

[21] Center for the Study of Emotion and Attention [CSEA-NIMH], "*International Affective Picture System: Digitized Photographs*," The Center for Research in Psychophysiology, University of Florida, Gainesville, FL, 1999.



Lawrence H. Yu (S'01–M'03) was born in Mississauga, ON, Canada, in 1977. He received the B.A.Sc. degree in engineering science (biomedical option) at the University of Toronto,Toronto, ON, Canada in 1999. In 2002, he received the M.A.Sc. degree in biomedical engineering from the University of Toronto for his work on point-of-gaze reconstruction in head-mounted eye tracking systems.

In 2002, he was a System Design Engineer at Micro Optics Design Corp. Currently, he is with AECL (Chalk River Laboratories) as a Research Engineer in the Information and Control Systems Development Division. His research interests are computational vision and image processing.

Moshe Eizenman was born in Tel-Aviv, Israel, in 1952. He received the B.A.Sc., M.A.Sc., and Ph.D. degrees in electrical engineering from the University of Toronto, Toronto, ON, Canada, in 1978, 1980, and 1984, respectively.

He joined the faculty of the University of Toronto in 1984, and he is currently an Associate Professor in the departments of Electrical Engineering, Ophthalmology, and at the Institute of Biomaterials and Biomedical Engineering. He is also a Research Associate at the Eye Research Institute of Canada and at the Hospital for Sick Children, Toronto. In cooperation with EL-MAR Inc. he has developed advanced technologies for eye tracking and gaze estimation systems. These systems are used around the world by universities and research institutes for medical and human-factors research and for pilot training. His research interests include detection and estimation of biological phenomenon, eye-tracking and gaze estimation systems, visual evoked potentials and the development of vision.