Published in final edited form as:

J Neurosci Methods. 2009 January 30; 176(2): 157–165. doi:10.1016/j.jneumeth.2008.09.010.

Compensating for Camera Translation in Video Eye Movement Recordings by Tracking a Representative Landmark Selected Automatically by a Genetic Algorithm

Faisal Karmali, Ph.D. and

Department of Otolaryngology, Massachusetts Eye and Ear Infirmary, Boston, MA, Department Of Biomedical Engineering, The Johns Hopkins University School of Medicine, Baltimore, MD

Mark Shelhamer, Sc.D.

Department of Otolaryngology–Head & Neck Surgery, The Johns Hopkins University School of Medicine, Baltimore, MD, Department of Biomedical Engineering, The Johns Hopkins University School of Medicine, Baltimore, MD

Abstract

It is common in oculomotor and vestibular research to use video or still cameras to acquire data on eye movements. Unfortunately, such data are often contaminated by unwanted motion of the face relative to the camera, especially during experiments in dynamic motion environments. We develop a method for estimating the motion of a camera relative to a highly deformable surface, specifically the movement of a camera relative to the face and eyes. A small rectangular region of interest (ROI) on the face is automatically selected and tracked throughout a set of video frames as a measure of vertical camera translation. The specific goal is to present a process based on a genetic algorithm that selects a suitable ROI for tracking: one whose translation within the camera image accurately matches the actual relative motion of the camera. We find that *co-correlation*, a statistic describing the time series of a large group of ROIs, predicts the accuracy of the ROIs, and can be used to select the best ROI from a group. After the genetic algorithm finds the best ROIs from a group, it uses *recombination* to form a new generation of ROIs that inherits properties of the ROIs from the previous generation. We show that the algorithm can select an ROI that will estimate camera translation and determine the direction that the eye is looking with an average accuracy of 0.75°, even with camera translations of 2.5 mm at a viewing distance of 120 mm, which would cause an error of 11° without correction.

Keywords

eye; video; VOG; video-oculography; cross-correlation; genetic algorithm

© 2008 Elsevier B.V. All rights reserved.

Address for correspondence: Faisal Karmali, 243 Charles St, JVPL 421, Boston MA 02114, faisal_karmali@yahoo.com, Phone: +1 (410) 218-7614, Fax: +1 (410) 614-1746.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1 Introduction

1.1 Motivation

Eye-movement studies are important for understanding specific neural control systems as well as general properties of sensorimotor processing in the brain. They provide an almost direct measure of vestibular (balance) function and provide insight into many forms of neural processing, such as sensorimotor integration, adaptation, and prediction. Video eye monitoring, using high-speed cameras and appropriate image processing, has the potential to provide non-invasive measurement of eye position with high spatial accuracy and temporal resolution (Clarke et al., 2002; Abadi and Gowen, 2004; MacDougall and Moore, 2005).

We recently conducted a study of human eye movements in different gravity levels (g levels) aboard a NASA aircraft, which flies parabolic trajectories to provide alternating levels of reduced ($\sim 0 g$) and enhanced ($\sim 1.8 g$) g levels (Lackner and DiZio, 2000; Shelhamer et al., 2002; Karmali and Shelhamer, 2008). Subjects in these flight experiments had a vertical misalignment of the eyes that was dependent on g level (Karmali et al., 2006), which was confirmed with binocular eye-movement recordings using a video system (Figure 1) consisting of two infrared cameras rigidly attached to a head-mounted frame (headset) (Clarke et al., 2002). Eye position as a function of time was determined by finding the pupil position in each video image. These recordings show that the eyes diverged vertically up to 3° as a function of g level.

One possible discrepancy in this result is that the cameras might have moved relative to the head, which could have created an artifactual difference in vertical eye position. Of particular concern in this study was roll rotation of the headset relative to the head, which would result in vertical translation of the two cameras by different amounts, causing an artifact of vertical eye movements, and which would be difficult to distinguish from the true ocular vertical misalignment that was being studied. Although we are confident that the result is not an artifact because the recorded eye misalignment was accompanied by double vision, we felt it necessary to develop a quantitative approach to confirm our qualitative assessment. We note that a typical translation motion of 2.5 mm with a camera distance of 120 mm and a typical eye diameter would cause an artifact of change in eye gaze of 11° .

The solution to this problem has wide applicability. Many methods have been attempted in order to mechanically secure a camera relative to the head, such as goggles with tight elastic bands (MacDougall and Moore, 2005) and MRI thermoplastic masks (Clarke et al., 2002). However, short of drilling holes in the skull, none of these methods eliminate camera movement adequately, especially in environments with large accelerations where inertia causes larger relative displacements of the cameras and the head. A current interest in vestibulo-ocular research is the use of natural, rapid, impulsive head rotations and translations in which the most interesting component of eye movement occurs when head acceleration is highest (Walker and Zee, 2005), which is also the time when the largest movement of the cameras relative to the head occurs. Since there is no simple and feasible method to eliminate undesired camera movement, we aimed here to develop an algorithm to measure any such movement, so that it could be compensated for in the analysis of the related eye-movement recordings.

1.2 Previous approaches

Several previous techniques have been developed to address the problem of measuring camera motion relative to the eye using image processing of video eye recordings. These algorithms are applied to each video frame in the recording (the sequential set of images that make up a video recording), and many algorithms compare each video frame to a reference video frame. The general approach is to find and track a landmark in each video frame that

moves when the camera moves, but not during other events, such as eye or eyelid movement.

In one implementation of this technique, the landmark that is tracked is a “glint,” the reflection on the cornea of infrared LEDs that are mechanically coupled to the camera (Young and Sheena, 1975; Hutchinson et al., 1989; Zhu and Ji, 2005; Kolakowski and Pelz, 2006; Guestrin and Eizenman, 2006). In theory, if one assumes a spherical eyeball, these reflections translate with the camera and do not move with eyeball rotation. However, in practice the cornea is not spherical, and precise calibrations are required for this technique to be successful. One system requiring subject-specific calibration is able to estimate gaze angles with an accuracy of 0.9° (Guestrin and Eizenman, 2006), while another system requiring less calibration is able to estimate vertical gaze angles with an accuracy of 1.7° (Zhu and Ji, 2005). We preferred to develop a technique that does not rely on critical calibration procedures.

A second approach is to apply markers to the skin around the eye and then assume that any translation of these landmarks is due only to movement of the cameras. There are two problems with this method. First, it requires marking of the skin before the experiment is performed, and in our specific case we have a library of data which was acquired without markers, and which cannot be repeated. Second, there remains the question of how much the skin may move due to blinks or fatigue, which is discussed below.

A third technique is to detect and track movement of the medial canthus, where the upper and lower lids converge next to the nose. Unfortunately this potential landmark is not visible in many recordings due to a small camera field of view.

Another landmark that can be tracked is the upper eyelid. We previously developed an automatic image-processing technique to measure the positions of the eyelids, which were then used to make corrections to the pupil location so that the *difference* in eye positions could be determined with less than 0.8° of error (Karmali and Shelhamer, 2005). However, eyelid position is affected by blinks, and thus is valid for our purposes only when the *difference* in vertical positions between the two eyes is required, because the eyelids can be assumed to move by the same amount, and thus any movement of the lids will cancel when the difference in eye position is computed. This method also required manual modification of intensity thresholds because of variability in skin reflectance between subjects. In another technique we developed, the *en-bloc* movement of each video frame was estimated using two-dimensional cross-correlations (Russ, 2002), but this technique had a high error, which is expected because eye and eyelid movement interfere with the cross-correlation.

Optical flow techniques (Barron et al., 1994) were briefly considered as a means to estimate camera translation. These techniques estimate the motion of objects within a series of images by estimating the optical flow, or motion velocity, of each pixel in the image, using both spatial and temporal derivatives of intensity. The equations to estimate motion are underdetermined over small regions of the image because of the “aperture problem”: there are regions of constant intensity whose derivatives change only at the edges. To solve the aperture problem, additional constraints are added to the analysis. These constraints are based on assumptions that are violated in our application, which could reduce spatial accuracy. The Lucas-Kanade algorithm (Lucas and Kanade, 1981) assumes that the amount of movement between video frames is small. When comparing each video frame to a reference, some frames would have large movements, and thus the assumption of the Lucas-Kanade algorithm would be violated. This problem could be overcome by comparing the motion in each frame to the previous frame and integrating the velocity over time, but this would cause an accumulation of error. The Horn-Schunck algorithm (Horn and Schunck,

1981) assumes global velocity smoothness over the image, which is violated in our application at the boundary of the eye and skin, and thus these errors could be propagated through the images. Although each of these existing methods is adequate under certain circumstances, none of them provide an accurate, practical, and easy-to-use solution to estimate camera translation.

2 Methods

2.1 Proposed Approach

Our goal is to develop an automatic algorithm to estimate the amount of camera translation relative to a deformable anatomical surface; this paper presents an approach and an interim solution that meets our requirements, although many improvements are possible. The surface being imaged is considered deformable because of eye movements, blinks, and fatigue which can cause warping of the skin. In this paper, we constrain the problem to vertical translations of a camera relative to the eyes and surrounding tissue. We are constrained by the small field of view of a set of existing video images (Figure 6), since our experiments were performed at great expense and it is not possible to repeat them to gather new data. Our aim is to correctly determine eye gaze angles to within 1° .

The general approach developed here is to select an arbitrary rectangular region of interest (ROI) from a reference frame and track that ROI in each subsequent frame of the video recording. (We distinguish between ROIs, which are rectangular regions in an image, and landmarks, which are anatomical features around the eye). Tracking is performed using a cross-correlation (Russ, 2002). Using a small ROI allows us to assume rigidity of the skin over a small portion of the video image. A large population of potential ROIs exists and can be generated. The question is how to select the appropriate ROI.

While there are many algorithms to automatically select "features" or landmarks based on their spatial or spatiotemporal properties, our algorithm uses only the temporal properties of an ROI to determine its quality. Since movement and deformation occur over time, our goal is to find the most stable ROI temporally, rather than the ROI with, for example, the highest spatial contrast.

In this paper we develop a method that automatically selects such an appropriate ROI from a population of putative ROI using a genetic algorithm (Figure 2). A genetic algorithm is an optimization procedure that searches a solution space using techniques inspired by evolutionary biology, such as inheritance, natural selection, and recombination (also called crossover). Genetic algorithms are most useful for solving problems when the solution space has high dimensionality or contains discontinuities which prevent conventional gradient-descent approaches. They are typically implemented by randomly creating a population of candidate solutions that evolve to become better solutions. In each generation, the fitness of each solution of the population is calculated, and a new population is created from the best solutions (Goldberg, 1989; Back et al., 1997). In addition, the best solutions undergo a process called recombination, whereby the characteristics of good solutions are combined with those of other good solutions; this process is inspired by genetic recombination, where children inherit characteristics from both of their parents. This new population of solutions form the population to be evaluated for fitness in the subsequent generation. After multiple generations, the population is expected to contain solutions that are close to optimal. Genetic algorithms have been used to filter and extract features from eye movement signals (Shelhamer, 2001; Ramat and Magenes, 2006).

Our implementation begins by creating a large population of potential ROIs. A metric, *co-correlation*, is presented which estimates the tracking accuracy of a particular ROI, and is

used as an objective function when searching the space of all possible ROIs. Optimization efficiency is improved by iteratively eliminating, early in processing, solutions that are not "worthy" – those which are unlikely to succeed because they have a low *co-correlation* – the algorithmic equivalent of neural Darwinism. After the best ROIs are found, they undergo *recombination* to produce a new population of ROIs, to which the process is applied again. This iterative procedure successfully creates and selects ROIs which enable us to accurately estimate camera translation. These automatically selected ROIs are used to correct eye gaze angles by subtracting the Cartesian location of the ROI from the Cartesian location of the pupil, before computing the eye gaze angles. We show using human eye-movement recordings that eye gaze angles are accurate to within 0.75° after the method is applied. A preliminary report of these results has been made (Karmali and Shelhamer, 2006).

2.2 ROI Tracking

Each ROI is a small rectangular section of the face on a video image. While rectangular ROIs were chosen to simplify implementation, other geometries may lead to improved tracking accuracy and could be tested in the future. ROIs are typically identified and selected on a reference video frame, and then tracked in each successive frame. The amount that the ROI moves in each frame relative to the reference frame is an estimate of camera translation. The ROI location is found in each non-reference video image using a two-dimensional cross-correlation (Russ, 2002) between corresponding regions on the reference frame and the frame being compared. The cross-correlation is a measure of similarity between the two images and is found by multiplying respective pixels from the two images and summing the products over the entire image, repeated numerous times with systematic shifts in the horizontal and vertical offsets of the ROI relative to the overall image. This produces what might be thought of as a 2-D probability distribution, in which the location of the maximum value in the distribution corresponds to the most likely horizontal and vertical location (in pixels) of the ROI in the overall image; in other words, it determines the amount of horizontal and vertical translation required to align the ROI in that video frame relative to the reference. This required translation is an estimate of how much camera translation has occurred compared to the reference frame.

A few additional processing operations are performed to improve the sensitivity of the cross-correlation. First, edge enhancement to accentuate the difference between dark and light areas is applied to improve contrast and thus the ability to find the ROI in each video image. This is applied to both the reference frame image and to each subsequent video image. For each image, the edge-enhanced version is created by adding an edge-detected overlay to the original image. A first-order difference equation is applied in both the horizontal and vertical directions to find the edges that are used to create the edge-detected overlay. The relative weighting when adding the edge-detected overlay and the original image is optimized to increase the sharpness of the cross-correlation peak (Karmali et al., 2005), and the weighting is kept constant for all images. After edge enhancement, the cross-correlation is normalized by the number of pixels that overlap between the two images, because otherwise ROIs that fall partly outside the image would be penalized. Finally, some ROIs may have a spatial similarity to several different locations in an image, producing multiple peaks in the cross-correlation. To increase the chance of selecting the correct location when the cross-correlation detects multiple candidates, we assume that the actual ROI is likely to be near the location of the ROI in the reference image. To implement this, a 2-D Gaussian distribution is centered on the location of the ROI in the reference image, and multiplied with the cross-correlation. This enhances the size of the correlation peaks near the original ROI location, relative to the other peaks, thus favoring ROIs close to the original location. Finally, even with a well-defined peak in the cross-correlation, using the location of the maximum yields a resolution that is, at best, one pixel. We achieve sub-pixel resolution by

computing the centroid of a set of points in the cross-correlation surrounding the maximum. These points are selected using a threshold (e.g. >0.99 of the maximum), and their centroid is computed to determine the improved estimate of the amount of translation between the images.

2.3 Co-correlation: Automatic Evaluation of ROI Tracking Accuracy

A measure that we term *co-correlation* was developed as a way to estimate how accurately any given ROI tracks the motion of the camera. The process of computing *co-correlation* begins by creating a large population of ROIs (typically 100). Although we attempted to use a population of rectangles of random sizes and positions, the algorithm is more successful when the starting population is a grid of overlapping rectangles that span the image. Each rectangular ROI is 94 pixels high and 52 pixels wide; these dimensions were selected during a preliminary analysis in which the accuracy of the algorithm was measured with ROIs of different sizes. ROIs close to the pupil are omitted from analysis, since they are more strongly influenced by eye movements than by camera translation. The pupil, which is the darkest region in infrared eye images, is detected using an intensity threshold; any ROIs that overlap the pupil are excluded from analysis by the genetic algorithm.

The location of each ROI n , $n \in 1 \dots N$, is found in each video frame via cross-correlation as described above, producing a time series of horizontal ($L_{X,n}(t)$) and vertical ($L_{Y,n}(t)$) translation estimates for each ROI. (It is only the vertical translations that are of interest to us in the present application.) The correlation coefficients of each time series compared to every other time series are calculated using the formula

$$C_{i,j} = \text{corr}(L_{Y,i}(t), L_{Y,j}(t)) \quad \forall i, j \in 1 \dots N.$$

This results in an $N \times N$ correlation matrix C containing estimates of the level of correlation in the movement of each ROI relative to each of the others:

$$C = \begin{bmatrix} \text{corr}(L_{Y,1}(t), L_{Y,1}(t)) & \text{corr}(L_{Y,1}(t), L_{Y,2}(t)) & \cdots & \text{corr}(L_{Y,1}(t), L_{Y,N-1}(t)) & \text{corr}(L_{Y,1}(t), L_{Y,N}(t)) \\ \text{corr}(L_{Y,2}(t), L_{Y,1}(t)) & \text{corr}(L_{Y,2}(t), L_{Y,2}(t)) & \cdots & \text{corr}(L_{Y,2}(t), L_{Y,N-1}(t)) & \text{corr}(L_{Y,2}(t), L_{Y,N}(t)) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \text{corr}(L_{Y,N-1}(t), L_{Y,1}(t)) & \text{corr}(L_{Y,N-1}(t), L_{Y,2}(t)) & \cdots & \text{corr}(L_{Y,N-1}(t), L_{Y,N-1}(t)) & \text{corr}(L_{Y,N-1}(t), L_{Y,N}(t)) \\ \text{corr}(L_{Y,N}(t), L_{Y,1}(t)) & \text{corr}(L_{Y,N}(t), L_{Y,2}(t)) & \cdots & \text{corr}(L_{Y,N}(t), L_{Y,N-1}(t)) & \text{corr}(L_{Y,N}(t), L_{Y,N}(t)) \end{bmatrix}$$

Finally the *co-correlation*, which is a measure of how much mutual movement a ROI has relative to all other ROIs, is computed as the sum of each row of the correlation matrix:

$$(\text{co-correlation})_i = \sum_{j=1}^N C_{i,j}.$$

The ROIs with the highest *co-correlation* are labeled *worthy* (typically the top 10% are selected).

2.4 Recombination to Create a New Generation of ROIs

Co-correlation is used to select *worthy* ROIs: those that can be used to best estimate camera translation. After the best ROIs have been found, the *recombination* step creates a new

generation of ROIs with similar properties to the *worthy* ROIs from the previous generation, in an attempt to find new ROIs that further reduce error.

In our implementation, each *worthy* ROI is recombined with every other *worthy* ROI based on the geometric properties of the rectangular ROIs. It is also recombined with itself so that copies of all *worthy* ROIs are automatically included in the next generation. For example, recombination of 10 *worthy* ROIs yields a new generation of 100 ROIs. Each pair of parents produces two new children as depicted in Figure 3. The first child is a rectangle just large enough to completely encompass both ROIs. If the rectangles of the parents intersect, the second child is the overlapping region. If they do not intersect, the top-left corner of the child is selected as the geometric point midway between the top-left corners of the parents, and the other corners are similarly computed. Once the new generation of ROIs is created, the entire process of determining *worthy* ROIs using *co-correlation* is repeated. When the *worthy* ROIs are found for the final generation, the ROI with the highest *co-correlation* is chosen as the ROI to be used to estimate the motion of the camera.

2.5 Error Introduced by ROI Instability

Since we are only interested in estimating camera translation, it is important that errors from other sources of motion are either eliminated or characterized. In this application, there are several potential causes of changes in pixel intensity in the image which might lead to erroneous movement estimates:

1. The eyeball can rotate about three axes, which, when projected into the image plane, manifests mostly as a horizontal and vertical movement of the pupil and iris, as well as small rotations and compressions of the image. The algorithm minimizes error from movement of the eyeball by detecting the pupil and excluding ROIs near it.
2. The camera can translate and *rotate* relative to the eye. Because of the way the camera is secured to the head in our particular experiments, most of this movement manifests as vertical and horizontal translation of the image. Errors introduced by camera rotation are discussed below (section 2.6).
3. The eyelids move due to blinks and fatigue, which manifests as translation and stretching of skin above and below the eyes, mostly in the vertical direction.
4. The skin around the eye can move due to external forces. Evidence that this movement is small is presented below (in this section and sections 2.7 and 3.4).
5. Movement of background light reflections can selectively illuminate one part of the image. This is not predictable, but is also not common.

To quantify the amount of error that could be introduced due to deformation of the soft tissue around the skin during physiological events such as blinks and fatigue, we assumed that the worse-case deformation occurs during blinks. We created the *intentional blinking* collection of video recordings in which the experimenter moved the headset vertically while the subject moved his eyes and made frequent blinks. A second camera tracked the motion of the headset to accurately record the actual translation of the camera (the second camera recorded a patch with an artificial pupil which was covering the other eye). The error between the estimated and actual camera translations was computed as an upper bound on the error that might be introduced by physiological events such as blinks and fatigue.

2.6 Error Introduced by Camera Rotation

Motion of the headset is possible in both translation and rotation. When the headset translates relative to the head, the cameras translate relative to the eyes. In contrast, when

the headset rotates relative to the head, it causes the cameras to rotate by the same amount, and also causes the cameras to translate relative to the eyes by different amounts for each camera. Because the physiological ocular response of interest occurs in the vertical direction, vertical translation of the headset is our main concern, which results in vertical translation of the cameras. Pure horizontal and fore-aft translation of the headset is unlikely because of the physical characteristics of the headband. While the headset is relatively unconstrained in roll rotation, the amount of pitch and yaw rotation is limited by the physical characteristics of the headset such as a nose rest, and would result in, respectively, vertical and horizontal translation of the cameras. They also cause small changes in the viewing angle of the camera, which introduces an insignificant amount of warping error. As a worst-case scenario, a rotation in either pitch or yaw, with the pivot point at the back of the head, and a 1 cm movement of the front of the headset, causes a 2° tilt of the image plane, in addition to the translation of the image described above. This causes warping compression of the image plane to $\cos(2^\circ) \times 100\% = 99.94\%$ of its original size, so that if an ROI is at the edge of the image and the pupil is at the center of the image, the ROI will move closer to the pupil by 0.08 pixels, introducing an error in the eye gaze angle of 0.02°, which we consider negligible.

Roll rotation of the headset is less constrained than pitch and yaw rotations by the geometry of the headset; rotation is about the midline of the head, and manifest mostly as a vertical translation of the two cameras, in opposite directions. However, roll rotation of the headset also results in roll rotation of the cameras, which is not measured by our algorithm; we consider here the error that may be introduced due to this. Error may be introduced because roll rotation is not accounted for when the center of the pupil and the center of the ROI being used to determine camera translation are offset from each other. As an example, if the ROI being tracked is to the left of the pupil, and if roll of the camera causes the ROI to move up in the image while the pupil does not move, then the algorithm will determine that the pupil has moved relative to the camera. The size of the error introduced is proportional to the distance between the pupil and ROI; for vertical camera translation, the error is dependent on horizontal distance only. We determined empirically that headset roll rotations of 2° are typical. We assume a worst-case horizontal distance between the ROI and pupil, converted into an eye gaze angle, of 30°. Then the worst-case scenario for the amount of vertical error is $\tan(2^\circ) \times 30^\circ = 1.0^\circ$. Although this error could be an issue, in circumstances where it could be a significant factor, only a simple extension of the algorithm would be required to compensate for headset roll. The search for ROIs could be restricted to those near the vertical line passing through the pupil, where roll-induced motion is at a minimum. Alternatively, ROIs on either side of the pupil could be found and their relative motion used to measure roll tilt, and to compensate for it.

2.7 Implementation and Evaluation

The algorithm was applied to nine video recordings obtained from human experimental trials. (Three additional recordings were excluded from analysis because of excessive vertical motion that caused the algorithm to saturate and exceeded our constraint of 50 pixels of vertical motion, or 4.1 mm. Two additional recordings were excluded from analysis because of inadequate illumination of the eye.) To evaluate the algorithm, we created a set of data for each video recording that contained our best estimate of the actual position of the camera relative to the eye at each point of time, which we call the *benchmark*. The benchmark was created only to evaluate the algorithms, and is not required during actual application of the algorithm. The benchmark was used to evaluate the ability of the algorithm to select those ROIs that result in the most accurate estimate of camera translation. All subjects gave their consent to participate; consent procedures were approved by the Johns Hopkins Medicine Institutional Review Board (JHM-IRB) and the NASA

Johnson Space Center Committee for the Protection of Human Subjects (JSC-CPHS). Recordings consisted of left and right eye images, each 256×256 pixels with 256 levels of gray, which were imported into the MATLAB software environment and analyzed using the Image Processing Toolbox. The output of the analysis for each ROI, which is a time series of estimated vertical camera translation, was compared to a benchmark, to determine its $\text{error}_{\text{RMS}}$. $\text{error}_{\text{RMS}}$ was defined as the root mean square (RMS) difference between the time series for the ROI and the time series for the benchmark, and was calculated for ROI i using the formula:

$$(\text{error}_{\text{RMS}})_i = \sqrt{\sum_t (L_{y,i}(t) - B_{y,i}(t))^2},$$

where $B_{y,i}(t)$ is the vertical location of the benchmark at time t . $\text{error}_{\text{RMS}}$ is in units of degrees, and provides the error in determining the direction the eye is looking, after ROI position is subtracted from pupil position. *Co-correlation* was also computed for each ROI. Note that although $\text{error}_{\text{RMS}}$ is a measure of linear motion, we report it in degrees as the equivalent error that the linear motion introduces to angular eye gaze-position determination during the conversion to three-dimensional angular coordinates.

Two collections of video recordings were used to evaluate the algorithm: the *parabolic flight* collection and the *intentional blinking* collection (Table I). The first collection of video recordings, the *parabolic flight* collection, consisted of actual data from parabolic flight experiments. The benchmark datasets for these recordings were created by a user who interactively located an anatomical marker in each video frame. Although such an anatomical marker could move during blinks or fatigue, the marker used was a crease in the skin far below the lower eyelid, which has been shown to move less than 0.7° during blinks (Karmali et al., 2005; Karmali et al., 2006). We emphasize that inaccuracies in locating these anatomical markers do not affect the actual performance of the algorithm and would likely reduce its measured performance.

To confirm that the error associated with this benchmark ROI was small, a second collection of recordings, the *intentional blinking* collection and associated benchmark, was created. As described in section 2.5 above, a second camera tracked the actual motion of the headset while the experimenter moved the headset vertically and the subject moved the eyes and eyelids. We computed the error between the camera translation estimated by the algorithm for this video recording and the actual camera translation measured by the second camera, to determine an upper bound on the error that would be introduced by physiological events such as blinks and fatigue.

3 Results

3.1 ROI Tracking

Figure 4 shows the results for tracking of ROIs over 430 seconds in one video recording belonging to the *parabolic flight* collection. The gray lines show the vertical location of 10 ROIs which were automatically located in each subsequent video frame. The ROIs were generated in the reference frame with random locations and uniform size. The dashed line is the vertical position of the benchmark ROI, which was interactively located by the user. The solid black line is the vertical location of the *worthy* ROI with the highest *co-correlation* (see below). *Co-correlation* was found by computing the correlation coefficient of each ROI's time course with each other time course, and then summing the correlation coefficients, as described in Methods.

3.2 Co-correlation Predicts Tracking Accuracy

Our central result is that *co-correlation* predicts how accurately a particular ROI estimates camera translation. As described in Methods, for each ROI, the correlation coefficient is computed between the time series for that ROI, and the time series for each other ROI. The *co-correlation* is the sum of these correlation coefficients. The accuracy of each ROI is also calculated, by finding the RMS difference between its time series and the time series of the benchmark. ROIs with a low $\text{error}_{\text{RMS}}$ are the most accurate. Figure 5 shows the relationship between *co-correlation* and $\text{error}_{\text{RMS}}$ for the video recording on which Figure 4 is based (trial B1P-R). In Figure 5, each dot represents one ROI, with its *co-correlation* plotted on the ordinate, and its $\text{error}_{\text{RMS}}$ plotted on the abscissa. The results for *recombination*, shown as black dots, are described below. The gray dots show the results for the first generation of ROIs. They show a trend such that as *co-correlation* increases, the $\text{error}_{\text{RMS}}$ in tracking the ROI position decreases. This shows that *co-correlation* can be used as an objective measure when searching the solution space of all candidate ROIs. The gray open circles indicate the *worthy* ROIs: those with the highest *co-correlation*, with $\text{error}_{\text{RMS}}$ ranging from 0.24° to 1.08° . These errors for the first generation of ROIs are acceptable compared to our goal of $\text{error}_{\text{RMS}}$ less than 1° , and compared to other approaches to the problem. We emphasize that $\text{error}_{\text{RMS}}$ is calculated using the benchmarks created to evaluate the algorithm, which are not generally available when the algorithm is used. However, these results show that *co-correlation* alone can determine which ROIs are *worthy* and thus would serve as accurate estimates of camera translation. Figure 6 shows the top 10 ROIs from generation 2, corresponding to the data in Figure 4 and Figure 5 (trial B1P-R), overlaid on an image from the video recording.

In the next section, we demonstrate how average $\text{error}_{\text{RMS}}$ is reduced, using *recombination* to produce a second generation of ROIs which better estimate camera translation.

We applied the algorithm to video recordings from both the *parabolic flight* and *intentional blinking* collections. The results are summarized in Table II. $\text{error}_{\text{RMS}}$ is listed for the top ten *worthy* ROIs; the range of values is shown in addition to the average. Over all recordings, the average $\text{error}_{\text{RMS}}$ was 0.75° . This shows that the algorithm can accurately track camera motion across a range of conditions.

3.3 Recombination of ROIs Improves Accuracy

Another significant result is a reduction in $\text{error}_{\text{RMS}}$ for a new generation of ROIs created by *recombination*, compared to the $\text{error}_{\text{RMS}}$ for the previous generation based on which the new generation was created. For the individual subject discussed above, the black dots in Figure 5 are the results for the second generation of ROIs, created by *recombination* of the *worthy* ROIs in the first generation. The black open circles indicate *worthy* ROIs for the second generation. The $\text{error}_{\text{RMS}}$ values for these ROIs are between 0.26° and 0.45° . When comparing the *worthy* ROIs (top ten) from the first and second generations, the lowest $\text{error}_{\text{RMS}}$ values of the groups are similar, but the highest $\text{error}_{\text{RMS}}$ are much lower in the second generation. While in the first generation the worst-case accuracy from amongst the *worthy* ROIs was 1.08° , in the second generation it is 0.45° . In the second generation, the range of $\text{error}_{\text{RMS}}$ is small, and any of the *worthy* ROIs can be selected without a significant penalty in $\text{error}_{\text{RMS}}$.

Table II shows that tracking was improved in the second generation for almost all video recordings. The average improvement was 0.09° , an 11% reduction in error. This difference is statistically significant (one-way paired *t*-test, $p < 0.07$). This shows that *recombination* is effective in improving tracking.

3.4 Deformation of Soft Tissue does not Increase error_{RMS}

To further evaluate the effect of deformation of the soft tissue around the eye on the results, the genetic algorithm was applied to the *intentional blinking* collection of video recordings which had a benchmark created using a second camera to record motion of the headset. In this recording, the subject made frequent blinks and eye movements while the experimenter translated the headset. A second camera quantified the motion of the headset by measuring the position of an artificial pupil drawn on a patch placed over the second eye. The results from the algorithm applied to one video recording are shown in Figure 7, and show that *co-correlation* predicts error_{RMS}. The algorithm found *worthy* ROIs in the first generation with error_{RMS} ranging from 0.52° to 0.89°. In this example, error_{RMS} increased slightly in the second generation, by an amount ranging from 0.55° to 1.02°.

The average error_{RMS} for the *intentional blinking* collection is 0.96°, compared to 0.65° for the *parabolic flight* collection (Table II), although this difference does not reach a level of statistical significance ($p > 0.1$; *t*-test). Both collections have average error_{RMS} better than our goal of 1.0°. The similarity in error_{RMS} for the two collections suggests that the algorithm is robust in selecting ROIs even in video recordings with frequent physiological events such as blinks and fatigue. This verifies that the benchmarks for the *parabolic flight* collection, where creases below the eye were interactively located, are a valid *post hoc* method to evaluate the genetic algorithm, even though the lower crease may move during blinks or fatigue.

A possible concern when evaluating the algorithm is that the benchmark in the *parabolic flight* collection used in the calculation of error_{RMS} could be subject to movement due to deformation of soft tissue. We note that *co-correlation* selects the best ROIs without using the benchmark, and any movement in the benchmark caused by skin deformation will not affect this selection. In contrast, deformation will negatively affect error_{RMS}. If anything, deformation of soft tissue will cause the performance of the algorithm to be underestimated. However, the strong correlation between error_{RMS} and *co-correlation* suggests that deformation does not significantly alter the selection of ROIs.

3.5 Computational efficiency

The genetic algorithm uses image-processing techniques that are quite computationally intensive. Here we consider the number of computations required to implement the algorithm, although we emphasize that it is a prototype, and major gains are possible using relatively simple techniques¹. The total number of operations (flops) required to apply the algorithm to one video recording is dependent on the length of the recording, the number of ROIs, and the number of generations. The number of operations required for each cross-correlation is determined by the number of pixels in each ROI multiplied by the number of pixels in the video image; for example, a 94×52-pixel ROI yields approximately 5000 flops and a 256×256 video image yields 65536 flops; the product is approximately 3×10^8 flops. With three generations, two eyes, 100 ROIs, and a video sequence with 430 images, the total

¹Another technique we developed to improve efficiency was an iterative analysis that only analyzed a subset of video frames in each recording, allowing candidate ROIs which were unlikely to be successful to be pruned early. *Co-correlation* was computed over three to six iterations, starting with reduced data sets. In the first iteration, rather than finding each ROI in every video frame, a subset of video frames was used (typically every fourth frame). *Co-correlation* was computed for all ROIs on this subset of video frames, and an interim subpopulation of *worthy* ROIs was selected of those ROIs that have the highest *co-correlation* (typically the top 80% of ROIs in each iteration). In the next iteration, a larger subset of video frames was used, and *co-correlation* was recomputed for the smaller subpopulation of ROIs. In the last iteration *co-correlation* was computed using every video frame, and a small subpopulation of ROIs, to select the *worthy* ROIs (typically the top 10%, as described above). With three iterations in which 20% of candidate solutions are removed in each iteration, and only one quarter of video images are analyzed in the first iteration, the computational burden is reduced by 23%. This reduction was small compared to those that could be realized through simpler optimizations, and more aggressive pruning led to erroneous deletion of good candidates, so we did not further investigate this technique and mention it only in passing.

number of operations is 7.7×10^{13} flops. On a 3.0 GHz Pentium 4 processor, this took 1134 seconds to execute using an uncompiled Matlab script, or approximately 2.6 seconds/video image. Although Matlab was used as a development platform, more extensive processing could occur in a more efficient environment, such as C or compiled Matlab. Also, simple optimizations could reduce computation with minimal impact on the results, making real-time processing possible. One example would be to constrain the search space of the cross-correlation to the nearby region rather than searching for the ROI anywhere in the image.

3.6 Summary

We have shown that *co-correlation* accurately predicts $\text{error}_{\text{RMS}}$, suggesting that *co-correlation* can be used to evaluate ROIs and select the ROI that best estimates camera translation. The *recombination* step in the genetic algorithm is effective at creating ROIs that further reduce $\text{error}_{\text{RMS}}$. Finally, concerns that the genetic algorithm could select ROIs that are positioned on soft tissue, and are thus vulnerable to motion during blinks and fatigue, are shown to be unwarranted. We emphasize that the algorithm does not use the benchmarks or $\text{error}_{\text{RMS}}$ to compute *co-correlation* or to train the genetic algorithm, and thus *co-correlation* alone can be used to select ROIs to measure camera translation.

4 Conclusions

The genetic algorithm presented here can accurately estimate vertical camera translation in video recordings taken to measure eye movements. The *co-correlation* measure is a robust predictor of tracking accuracy for a given ROI, and can be used to select the best ROI from a population. After two generations, the genetic algorithm finds ROIs with an average accuracy of 0.75° in determining the direction that the eye is looking. This meets the requirements for the study that motivated this work (mentioned in the introduction), where the actual eye movement we are studying ranges from 1° to 3° . The accuracy of this technique is comparable to that of other methods we have previously developed (Table III), but it is more robust.

Although this accuracy is sufficient for most eye-movement studies, some studies, such as those of static eye alignment, can benefit from better accuracy. Better landmarks would be afforded by a camera lens with a larger field of view, and high-contrast markers placed on the skin. Also, incorporating more statistical temporal information about the dynamics of eye movements into *a priori* statistics would improve both eye tracking and camera translation tracking.

Although we have developed these techniques for a specific study of video-eye recordings, they have more general applicability. Current roadblocks to video techniques as the standard for eye movement studies include temporal speed, spatial resolution, and headset movement. The first two issues can be overcome by multi-megapixel cameras with frame rates in the kilohertz range which are rapidly becoming affordable for both research and clinical settings. Methods of mechanically fixing the cameras to the head in dynamic motion environments have many shortcomings, and thus the development of image-processing algorithms to overcome this problem is important to enable the wider use of video systems. The algorithm described here does not require the skin to be marked, avoids using features that are often outside the camera field of view, and does not require precise calibration.

While this technique was developed specifically to compensate for unwanted camera movement relative to the head, it could also be expanded to track intentional movement of the head when the camera is fixed in space. This would allow tracking of eye and head position. This is important for applications where free head movement without the

impediment of a heavy head-mounted video system is desired, such as studies of eye-head interaction and studies of eye movements in pilots.

Acknowledgments

Supported by: NIH DC006090, National Space Biomedical Research Institute (NASA/NSBRI), National Science and Engineering Research Council of Canada (NSERC). We acknowledge the technical assistance and advice of Tiffany L. Chen, Dr. Jerry Prince, and Dr. Stefano Ramat.

References

- Abadi RV, Gowen E. Characteristics of saccadic intrusions. *Vision Res.* 2004; 44:2675–2690. [PubMed: 15358063]
- Back T, Hammel U, Schwefel HP. Evolutionary computation: Comments on the history and current state. *IEEE Transactions on Evolutionary Computation.* 1997; 1:3–17.
- Barron JL, Fleet DJ, Beauchemin SS. Performance of optical flow techniques. *Int J Comput Vision.* 1994; 12:43–77.
- Clarke AH, Ditterich J, Druen K, Schonfeld U, Steineke C. Using high frame rate CMOS sensors for three-dimensional eye tracking. *Behav Res Methods Instrum Comput.* 2002; 34:549–560. [PubMed: 12564559]
- Goldberg, DE. Genetic algorithms in search, optimization and machine learning. Addison-Wesley Longman Publishing Co., Inc; Boston, MA: 1989. p. 1-9.
- Guestrin ED, Eizenman M. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Trans Biomed Eng.* 2006; 53:1124–1133. [PubMed: 16761839]
- Horn BKP, Schunck BG. Determining optical flow. *Artif Intell.* 1981; 17:185–203.
- Hutchinson TE, White J, Martin WN, Reichert KC, Frey LA. Human-computer interaction using eye-gaze input. *IEEE Trans Syst Man Cybern.* 1989; 19:1527–1534.
- Karmali, F.; Chen, TL.; Shelhamer, M. Compensating for camera translation in video eye movement recordings. Abstracts of the Biomedical Engineering Society Meeting; Baltimore. 2005.
- Karmali F, Ramat S, Shelhamer M. Vertical skew due to changes in gravito-inertial force: A possible consequence of otolith asymmetry. *J Vestib Res.* 2006; 16:117–125. [PubMed: 17312339]
- Karmali F, Shelhamer M. Automatic detection of camera translation in eye video recordings using multiple methods. *Ann N Y Acad Sci.* 2005; 1039:470–476. [PubMed: 15827002]
- Karmali, F.; Shelhamer, M. Compensating for camera translation in video eye movement recordings by tracking a landmark selected automatically by a genetic algorithm. Proceedings of the IEEE EMBS 28th Annual International Conference; New York. 2006.
- Karmali F, Shelhamer M. The dynamics of parabolic flight: flight characteristics and passenger percepts. *Acta Astronautica.* 2008; 63:594–602. [PubMed: 19727328]
- Kolakowski, SM.; Pelz, JB. Compensating for eye tracker camera movement. Proceedings of the 2006 Symposium on Eye Tracking Research and Applications; San Diego. 2006.
- Lackner JR, DiZio P. Human orientation and movement control in weightless and artificial gravity environments. *Exp Brain Res.* 2000; 130:2–26. [PubMed: 10638437]
- Lucas, BD.; Kanade, T. An Iterative Image Registration Technique with an Application to Stereo Vision. Proceedings of the 7th International Joint Conference on Artificial Intelligence; Vancouver. 1981.
- MacDougall HG, Moore ST. Functional assessment of head-eye coordination during vehicle operation. *Optom Vis Sci.* 2005; 82:706–715. [PubMed: 16127336]
- Ramat S, Magenes G. Latency detection in motor responses: a model-based approach with genetic algorithm optimization. *IEEE Trans Biomed Eng.* 2006; 53:2015–2023. [PubMed: 17019866]
- Russ, JC. The Image Processing Handbook. CRC Press; Boca Raton, FL: 2002. p. 527–529.
- Shelhamer M. Use of a genetic algorithm for the analysis of eye movements from the linear vestibulo-ocular reflex. *Ann Biomed Eng.* 2001; 29:510–522. [PubMed: 11459345]
- Shelhamer M, Clendaniel RA, Roberts DC. Context-specific adaptation of saccade gain in parabolic flight. *J Vestib Res.* 2002; 12:211–221. [PubMed: 14501099]

- Walker MF, Zee DS. Cerebellar disease alters the axis of the high-acceleration vestibuloocular reflex. *J Neurophysiol.* 2005; 94:3417–3429. [PubMed: 16033941]
- Young LR, Sheena D. Eye-movement measurement techniques. *Am Psychol.* 1975; 30:315–330. [PubMed: 1137223]
- Zhu, Z.; Ji, Q. Eye gaze tracking under natural head movements. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; San Diego.* 2005.

Figure 1. Binocular video eye tracker (Chronos Vision GmbH). Two side-mounted infrared cameras record eye movements via infrared-reflective mirrors mounted in front of the eyes. A bitebar is used to reduce movement of the headset relative to the head, although movement still occurs.

Figure 2. Flowchart depicting the genetic algorithm used to determine the best (worthy) ROIs using co-correlation, and to recombine the worthy ROIs to create a new generation.

Figure 3.

Recombination of ROIs. Each pair of parents produces two new children. The first child is a rectangle just large enough to completely encompass both ROIs. If the rectangles of the parents intersect, the second child is the overlapping region. If they do not intersect, the top-left corner of the child is the geometric point midway between the top-left corners of the parents, and the other corners are similarly computed.

Figure 4.

Results for tracking of ROIs over 430 seconds in one video recording made in parabolic flight (trial B1P-R). The gray lines show the vertical location of 10 ROIs which were automatically located in each subsequent video frame after the reference. The dashed line is the vertical position of the benchmark landmark, which was interactively located by the user in each video frame. The solid black line is the vertical location of the worthy ROI with the highest *co-correlation*, which was calculated as described in Methods.

Figure 5.

Co-correlation predicts tracking accuracy. The results shown are from the application of the genetic algorithm to a video recording from the *parabolic flight* collection used in this study, recorded in parabolic flight (trial B1P-R). $\text{error}_{\text{RMS}}$ is computed using a benchmark created interactively by the user. During the first generation (gray dots), ten *worthy* ROIs are selected (gray open circles), with $\text{error}_{\text{RMS}}$ ranging from 0.24° to 1.08° . These ten *worthy* ROIs are recombined and the analysis is repeated on the new generation of 100 ROIs (black dots), to find ten *worthy* ROIs (black open circles). For the *worthy* ROIs from the second generation, $\text{error}_{\text{RMS}}$ is between 0.26° and 0.45° . Even if the worst ROI is chosen from this group, $\text{error}_{\text{RMS}}$ is still low at 0.45° . The benchmark data are never used by the image processing algorithms, demonstrating the robustness of the algorithm. The benchmark ROI used was a natural blemish below the eye.

Figure 6.

A video image of the right eye, showing the typical field of view for videooculography. The overlaid *worthy* ROIs are from generation 2, and correspond to those shown in gray in Figure 4 (trial B1P-R).

Figure 7.

Co-correlation predicts tracking accuracy, even with frequent blinks, eye movements, and large headset movements. The results shown are for a video recording from the *intentional blinking* collection used in this study (trial E1G-R). $\text{error}_{\text{RMS}}$ was computed using a benchmark measured by a second camera that recorded actual camera translation. The genetic algorithm found *worthy* ROIs in the first generation with $\text{error}_{\text{RMS}}$ ranging from 0.52° to 0.89° . In this example, $\text{error}_{\text{RMS}}$ increased slightly in the second generation, ranging from 0.55° to 1.02° , but these results were still close to our goal of 1.0° accuracy. These $\text{error}_{\text{RMS}}$ are similar to that for recordings from the *parabolic flight* collection, such as in Figure 5, and thus the algorithm is robust in selecting ROIs even in video recordings with frequent physiological events such as blinks and fatigue.

Table I

The genetic algorithm was applied to video recordings from two different collections.

Name of the video collection	1. <i>Parabolic flight collection</i>	2. <i>Intentional blinking collection</i>
Purpose of the video collection	Confirm that the genetic algorithm selected appropriate regions of interest (ROIs) in video recordings made during vestibular experiments, for which movement of the camera relative to the face would cause artifacts in the results.	Confirm that the genetic algorithm selected appropriate regions of interest (ROIs) even with large eye and eyelid movement, which result in large skin deformations.
Experimental methods used to produce video recordings	Both eyes were recorded during changing g levels in parabolic flight.	The experimenter moved the headset vertically while the subject moved the eyes and eyelids.
Benchmark dataset used to evaluate the results of the genetic algorithm for video recordings in the collection	A user interactively located an anatomical marker in each video frame (a crease below the eye).	A second camera tracked the actual motion of the headset.

Table II

The algorithm is able to select ROIs that accurately track the motion of the camera. Results from the analysis of nine videos from both the *parabolic flight* and *intentional blinking video* collections are shown. errorRMS reported are the mean of the top 10 *worthy* ROIs determined for each trial. By the second generation, the average errorRMS is 0.75°. The average improvement between the first and second generation is 0.09° or 11% ($p < 0.07$; paired one-tail t -test).

Video collection	Subject Trial	Eye	errorRMS for top ten <i>worthy</i> ROIs average in bold (range in brackets)		Improvement
			Generation 1	Generation 2	
Parabolic flight	B1P-L	Left	0.58° (0.32–1.15)	0.51° (0.50–0.52)	0.07°
Parabolic flight	B1P-R	Right	0.66° (0.24–1.08)	0.32° (0.26–0.45)	0.34°
Parabolic flight	D1P-L	Left	1.21° (0.30–1.57)	1.02° (0.79–1.41)	0.19°
Parabolic flight	D1P-R	Right	0.68° (0.32–0.87)	0.41° (0.28–0.56)	0.27°
Parabolic flight	C1P-L	Left	0.76° (0.51–1.13)	0.88° (0.60–1.21)	–0.12°
Parabolic flight	C1P-R	Right	0.81° (0.59–1.04)	0.77° (0.58–1.76)	0.04°
Intentional blinking	E1G-R	Right	0.75° (0.52–0.89)	0.84° (0.55–1.02)	–0.09°
Intentional blinking	B2G-R	Right	1.05° (0.54–1.99)	1.00° (0.92–1.05)	0.05°
Intentional blinking	B3G-R	Right	1.07° (1.00–1.11)	1.04° (1.01–1.07)	0.03°
Average across trials			0.84°	0.75°	0.09°

Table III

Comparing the accuracy of different camera translation estimation techniques described in a previous paper (Karmali and Shelhamer, 2005) and in this paper. In image cross-correlation a reference video image is cross-correlated with each video image. In eyelid tracking, an eyelid trace is found in each video image, and the relative location of the traces is found. In the eyelid template method, a template of the eyelid shape is produced and cross-correlated with each video image. Finally the algorithm described in this paper, with the result from Figure 5, is listed. Although the eyelid template algorithm accurately estimates the motion of the eyelid, it does not distinguish between motion of the eyelid in the image due to camera translation, and movement due to blinks or fatigue. This artifact is less important when calculating the difference between eye positions, since any blink will affect both eyelids similarly.

Method	error _{RMS}
Image cross-correlation	6.49°
Eyelid tracking	4.35°
Eyelid template	0.76°
Eyelid template (with enhancements)	0.47°
Genetic algorithm to automatically select a region of interest (ROI)	0.75°