Low Bandwidth Eye Tracker for Scanning Laser Ophthalmoscopy

Zachary G. Harvey\textsuperscript{a,b}, Alfredo Dubra\textsuperscript{c,d}, Nathan D. Cahill\textsuperscript{e} and Sonia Lopez Alarcon\textsuperscript{b}

\textsuperscript{a} Flaum Eye Institute, University of Rochester, Rochester, NY 14642
\textsuperscript{b} Department of Computer Engineering, Rochester Institute of Technology, Rochester, NY 14623
\textsuperscript{c} The Eye Institute, Medical College of Wisconsin, Milwaukie, WI 53226
\textsuperscript{d} Biophysics Department, Medical College of Wisconsin, Milwaukie, WI 53226
\textsuperscript{e} School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY 14623

ABSTRACT

The incorporation of adaptive optics to scanning ophthalmoscopes (AOSOs) has allowed for \textit{in vivo}, noninvasive imaging of the human rod and cone photoreceptor mosaics.\textsuperscript{1,2} Light safety restrictions and power limitations of the current low-coherence light sources available for imaging result in each individual raw image having a low signal to noise ratio (SNR). To date, the only approach used to increase the SNR has been to collect large number of raw images ($N > 50$), to register them to remove the distortions due to involuntary eye motion,\textsuperscript{3,4} and then to average them. The large amplitude of involuntary eye motion with respect to the AOSO field of view (FOV) dictates that an even larger number of images need to be collected at each retinal location to ensure adequate SNR over the feature of interest. Compensating for eye motion during image acquisition to keep the feature of interest within the FOV could reduce the number of raw frames required per retinal feature, therefore significantly reduce the imaging time, storage requirements, post-processing times and, more importantly, subject’s exposure to light. In this paper, we present a particular implementation of an AOSO, termed the adaptive optics scanning light ophthalmoscope (AOSLO) equipped with a simple eye tracking system capable of compensating for eye drift by estimating the eye motion from the raw frames and by using a tip-tilt mirror to compensate for it in a closed-loop. Multiple control strategies were evaluated to minimize the image distortion introduced by the tracker itself. Also, linear, quadratic and Kalman filter motion prediction algorithms were implemented and tested using both simulated motion (sinusoidal motion with varying frequencies) and human subjects. The residual displacement of the retinal features was used to compare the performance of the different correction strategies and prediction methods.

Keywords: Eye tracking, Kalman filtering, Image registration

1. INTRODUCTION

Taking pictures of the retina requires providing the subject with a fixation target, to keep the retinal region of interest in the field of view. Despite the subject’s best efforts to fixate, involuntary fixational eye motion significant impacts high magnification adaptive optics (AO) ophthalmoscopes with very small FOVs (typically between 0.5 and 3\textdegree{}). These involuntary eye motion while fixating varies greatly across the population and it is affected by factors such as age and retinal disease.\textsuperscript{5–8} Fast and small amplitude motion such as micro-saccades or tremors can potentially be compensated for with previously described image registration methods.\textsuperscript{4} The larger amplitude of eye drift however, can move features of interest partially or completely outside of the FOV of AO equipped ophthalmoscopes. This combined with the fact that AOSLO images typically have low SNR due to light safety and light sources restrictions, dictates large number of images have to be collected, registered and then averaged in order to produce a single image of a given feature with high SNR. This results in long imaging times, unnecessary exposure to light, larger storage requirements and longer image registration times. This last point is very costly, because current image registration methods for this instruments are not fully automated.

A number of eye-trackers have been previously demonstrated for applications other than retinal imaging. A notable example is the first noninvasive method for with acceptable accuracy developed by Cornsweet and

Further author information: send correspondence to Zachary Harvey: E-mail: zgh7555@rit.edu


Proc. of SPIE Vol. 8314  831450-1
Crane,\textsuperscript{9} that relies on tracking two Purkinje reflections. Shining light into the eye anterior part of the eye through the pupil results in four reflections, that correspond to the front and back surfaces of the cornea and the crystalline lens. These reflections are known as the first through the fourth Purkinje images, respectively, and their relative motion can be used to estimate eye motion. This method is particularly robust to lateral head translations as this causes the reflections to move together, and only rotations will change the separations between the images. The wobbling of the crystalline lens with saccades, fundamentally limits the performance of the method for high magnification retinal imaging.

Highly-sophisticated eye-trackers specifically developed for retinal imaging have been successfully demonstrated in an AOSLO by Hammer \textit{et. al.} developed an eye tracking system for use with AOSLOs.\textsuperscript{10} This approach used dither scanners to move a spot of light in a circle around a user controlled feature at a frequency of 8 KHz. The amplitude of the reflection of the spot is analyzed on real-time to estimate phase shifts between consecutive circular scans, from which the amount of eye motion is estimated. The motion signal is then sent to additional galvanometric scanners that divert the imaging raster to keep the feature of interest in the FOV. Although this method for tracking achieves excellent bandwidth (1Khz), it increases overall system complexity and cost, while also reducing the optical throughput.

The low-bandwidth eye tracker presented here uses the very images recorded by the ophthalmoscope to control a tip and tilt mirror to keep the feature of interest within the field of view of an AOSLO. Because the tip-tilt mirror replaces the tip-only mirror that was used to form the imaging raster, not a single optical surface is added to the optical setup, keeping light exposure and throughput unchanged. No custom electronics had to be developed either, keeping costs and complexity to a minimum. The image motion was estimated by dividing the image in bands that span the whole image width along the fast scanning dimension, and comparing against a reference frame using the normalized cross-correlation\textsuperscript{4} using graphics processing units (GPUs) parallel processing in an Nvidia graphics card (Santa Clara, California, USA). The estimated eye motion was then fed to multiple prediction algorithms aiming an minimize both the average eye motion and the image distortion that originates from applying the motion correction itself. The eye tracker performance was evaluated using linear, quadratic and Kalman filters as predictors using sinusoidal oscillations with varying frequencies and with human subjects.

The remainder of this paper is organized in the following manner. Section 2 describes the motion prediction and eye drift compensation methods, including details of the GPU normalized cross-correlation implementation. Section 3 describes the experimental results, and finally, Section 4 presents a summary and points to future work.

2. AOSLO WITH COMPENSATION FOR EYE DRIFT

Images from AOSOs are created by recording the light reflected from scanning a focused spot of light across the retina using two optical scanners. After every oscillation of the fast (horizontal) scanner, a slower (vertical) scanner moves the spot to the next line in the raster scan. When the vertical scanner reaches the bottom of the FOV, it returns to the top to capture another image. Figure 1 shows an example a registered AOSO image of the human cone and rod photoreceptor mosaics. Typically, offline registration and registration of AOSO image sequences is used to overcome the distortion caused by involuntary eye motion and low SNR. The basis for the tracking system presented here is the image registration method developed by Dubra and Harvey,\textsuperscript{4} described next.

2.1 Motion Estimation

Given the high level of magnification provided by adaptive optics, small angles of rotation of the eye are approximated as lateral translations. Therefore the pixel mapping between two AOSLO images can be modeled by

\begin{align}
X &= x + \epsilon_x(x, y) \\
Y &= y + \epsilon_y(x, y) ,
\end{align}

where \((X, Y)\) is the location of the pixel on the reference frame, \((x, y)\) is the location of the same pixel on some subsequent, or floating, frame, and \((\epsilon_x, \epsilon_y)\) represent the approximated translation that occurred between the
reference frame and the floating frame. Horizontal scanning frequencies are sufficiently fast to assume that the eye is stationary during a single line scan. Therefore the pixel mapping can be approximated as a function of the vertical location of the pixel; i.e.,

\[
X = x + \epsilon_x(y) \\
Y = y + \epsilon_y(y).
\]  

(2)

Images are divided into strips and compared to a manually selected reference frame containing minimal distortion and significant overlap with the other images in the sequence. The normalized cross correlation (NCC) is used as a similarity metric, and is defined by:

\[
C_{R,S}(m,n) = \frac{\sum_{i,j} R(i,j)S(m + i, n + j)}{\sqrt{\sum_{i,j} R(i,j)^2 \sum_{p,q} S(p,q)^2}},
\]  

(3)

where \( R \) represents the reference image, \( S \) represents the current strip, and \( m \) and \( n \) are the translations in the \( x \) and \( y \) direction. The summations are performed only over the resulting overlapping region after translating the strip \((m,n)\) pixels from the center of the reference image. The value \( C_{R,S}(m,n) \) is measure of match between the reference frame and the current strip that has been translated by a factor \((m,n)\). In general this value is in the interval \([-1,1]\), however, because images have only positive intensity values, the result of the correlation is in the interval \([0,1]\) where a value of 1 means a perfect match and 0 represents no correlation. Calculating the correlation with all possible values for \( m \) and \( n \) would be an \( O(n^2) \) operation. Fourier domain techniques are used to reduce the computational complexity to an \( O(N\log N) \) operation. Using this, the solution space for the transformation can be computed by:

\[
C_{R,S} = \frac{\text{IDFT} \left[ \text{DFT} \left[ R \right]^* \text{DFT} \left[ S \right] \right]}{\sqrt{\text{IDFT} \left[ \text{DFT} \left[ P_R \right]^* \text{DFT} \left[ S^2 \right] \right] \text{IDFT} \left[ \text{DFT} \left[ R^2 \right]^* \text{DFT} \left[ P_S \right] \right]}},
\]  

(4)
Figure 2. Zero padding of the images before performing the DFT in order to avoid aliasing artifacts due to periodic nature of DFT.

Here, DFT and IDFT denote the discrete Fourier transform and the inverse discrete Fourier transform, respectively. \( P_R \) represents the unit step function over the support of the reference image and \( P_S \) is the unit step function over the support of the current strip. The * symbol represents complex conjugate. The images and step functions are zero padded as depicted in figure 2, in order to avoid periodic aliasing caused by the discrete Fourier transform. The NCC was chosen for its low computational complexity as well as its robustness to fluctuations in brightness and contrast. This can be illustrated by treating the images \( R \) and \( S \) as \( N \times M \) dimensional vectors and rewriting 3 as the dot product of the current strip and reference frame divided by the norm of the two vectors; i.e.,

\[
C_{R,S}(m,n) = \frac{\langle R, S_{m,n} \rangle}{\| R \| \| S_{m,n} \|} = \cos(\theta),
\]  

where \( \theta \) is the angle between the two vectors \( R \) and \( S_{m,n} \). The angle between the two vectors does not depend on their norm, and hence is independent of the brightness in the image. This allows for fluctuations in mean intensity between images within a sequence due to dynamic factors such as tear film, pupil misalignment and AO correction.

If we treat \( C_{R,S} \) as a matrix with \((i,j)\) element representing the NCC between the reference image and current strip translated by a factor of \( i - \frac{M_R}{2} - \frac{M_S}{2} \) in the \( y \) direction and \( j - \frac{N_R}{2} - \frac{N_S}{2} \) in the \( x \) direction, then the translation is determined by finding the maximum across a ROI of the NCC matrix such that a minimum amount of overlap between the current strip and the reference frame is guaranteed. If the maximum is found at a location \((i_{\text{max}},j_{\text{max}})\) then the displacements, \( \epsilon_x(y) \) and \( \epsilon_y(y) \) become

\[
\epsilon_x(y) = j_{\text{max}} - N_R \]
\[
\epsilon_y(y) = i_{\text{max}} - M_R .
\]  

The tracking system implemented here uses this motion estimation method during image acquisition and feeds the results to the control algorithms that in turn modify the orientation of the tip-tilt mirror to reduce the image displacement with respect to the reference image while simultaneously trying to keep image distortion to a minimum.

2.2 GPU Implementation

In order to be able to calculate the retinal displacement using multiple strips across an image keeping up with the approximately 15 frames per second recorded by the AOSLO, end-user graphics cards with a large number of GPUs that can operate in parallel were used. The key advantage of this technology is that it that the parallel processing allows for very fast DFT calculation, which is the most demanding calculation in the eye-tracking process. In fact, our GPU implementation is almost two orders of magnitude faster than that of current central
processing units (CPUs). This was achieved by optimizing memory transfers and by performing calculations involving the reference frame and the step functions that common to all NCC calculation only once, and keeping them in memory. Our implementation also takes advantage of the fact that most fast implementations of the DFT (including the CuFFT provided by Nvidia) are faster when the matrix dimensions are powers of two, by adding additional zero-padding to the the NCC input matrices, to the nearest power of two. A diagram indicating the memory buffers used for the NCC calculation and the operations performed on them is shown in Fig. 3.

### 2.3 Motion Compensation

After the translation between the ROIs in the reference and current images has been estimated, the voltage required for compensating this shift, multiplied by a user-selected gain, is applied to the optical scanner. Applying the full compensation at once and as soon as the calculation is completed results in vertical stretching and horizontal shearing of the features in the image, as illustrated in Fig. 4.

These distortions lead to great information loss that make large portions of the images unusable, which is not acceptable. Thus two different motion estimation strategies were explored: full frame and sub-frame. In the full-frame approach, the motion was estimated by comparing the full reference frame against the full reference
frame, and the estimated motion was applied while the vertical (slow) scanner is returning to the top of the raster and no portion of the image is being collected, thus avoiding additional image distortion. This approach however, limits the correction rate to a single update per frame. The second method, estimates the motion and updates multiple times per frame, increasing the effective bandwidth of the eye tracker, but at the potential cost of distorting the images while the motion correction is applied. Given the higher correction rate, the estimated eye motion between iterations should be smaller, and thus reduce the impact of image distortion.

The estimated retinal displacements were used to evaluate four predictive algorithms: linear extrapolation, quadratic extrapolation, discrete cosine transform (DCT) extrapolation (i.e. low-pass filtering), and Kalman filtering. The linear and quadratic prediction methods use a user-defined number of image displacements from previous iterations to extrapolate the motion using first and second order polynomials, respectively. The DCT method uses the lower order terms of the DCT II\(^{11}\) to extrapolate the retinal displacements. Finally, the Kalman filter is an optimal recursive state estimator in which the prediction is based on a physical model.\(^{12}\)

For this work, a simple kinematics model is assumed where the velocity of the retina is constant between any two updates.\(^{13}\) The filter algorithm is described in table 1, using the following definitions,

\[
x_k = \left( \begin{array}{c} x \\ \dot{x} \end{array} \right), \quad F = \left( \begin{array}{cc} 1 & \Delta t \\ 0 & 1 \end{array} \right), \quad G = \left( \begin{array}{c} \Delta t^2/2 \\ \Delta t \end{array} \right), \quad Q = GG^T \sigma_a^2,
\]

where \(\sigma_a^2\) is the expected variance in the acceleration.

3. EXPERIMENTAL RESULTS

3.1 Sinusoidal Motion Test

Sinusoidal retinal motion was simulated by placing a non-resonant galvanometric optical scanner (VM2500+, GSI Group Corp, Billerica, Massachusetts, USA) in front of a model eye formed by a lens and a piece of paper acting as a retina. The optical scanner was driven with a sinusoidal voltage with a fixed amplitude of 1.54 degrees (approximately 2 times the typical FOV), and varying frequencies to test the performance of the eye tracker. For
Model

\[ x_k = F x_{k-1} + G u_{k-1} + w_{k-1} \]
\[ y_k = H x_k + v_k \]

Prediction

\[ \hat{x}_k = F \hat{x}_{k-1} + G u_{k-1} \]
\[ P_k^- = F P_{k-1}^+ F^T + Q_{k-1} \]

Update

\[ K_k = P_k^- H^T (H_k P_k^- H^T + R_k)^{-1} \]
\[ \hat{x}_k^+ = \hat{x}_k^- + K_k (y_k - H \hat{x}_k^-) \]
\[ P_k^+ = (I - K_k H) P_k^- \]

Table 1. Summary of the Kalman filter algorithm. \( x_k \) is a column vector containing the current state, \( F \) is a matrix represents the system dynamics, \( u_{k-1} \) is a column vector containing the control input, \( G \) is a matrix that relates the control input to the system state, \( w_{k-1} \) represents noise in the system dynamics with covariance matrix \( Q_k \), \( y_k \) is a column vector containing the measurements, \( H \) is a matrix relates the system state to the measurements, and \( v_k \) is a column vector representing the noise in the measurements with covariance matrix \( R_k \).

Each frequency, the tracking and motion signals were recorded with a digital oscilloscope (DPO 3014, Tektronix, Beaverton, Oregon, USA). The peak velocity was calculated for each frequency and normalized by the velocity of the vertical component of the imaging raster. The results are summarized in Fig. 3.1.

The relative error were calculated as the norm of the residual motion estimated off-line divided by the norm of the motion signal. The error is plotted as a function of normalized peak velocity, which is proportional to the frequency of the sinusoidal motion. This parameter was chosen as a measure of speed relative to the eye-tracking sampling frequency. If the motion is equal to the speed of the vertical scanner (normalized peak velocity = 1.0), then each line of the image would be exactly the same (if the motion was opposing the raster scan) or each line in the image would be spaced by the FOV size (if the motion is with the direction of the raster scan) resulting in a very sparse sampling.

Generally speaking, full frame tracking performs the worst, as the features being tracked fall out of the field of view before the compensation has been applied. As expected the tracking error decreases with increasing number of updates per frame. When using the finer strips, the tracker was able to follow motion up to 0.65 times the speed of the vertical scanning, while full frame tracking methods failed for relative speeds higher than 0.20. Kalman filtering performed best among the predictive methods for sub-frame tracking while LPF performed best for full frame tracking.

![Figure 5. Relative errors in the tracking signal for sinusoidal motion with varying frequencies. The normalized peak velocity is the ratio of the peak velocity in the motion signal to the velocity of the raster scan. The color of the lines indicate the prediction method used. The line style corresponds to the number of updates made per frame, with solid lines indicating once per frame (full frame), dotted lines indicating 4 updates per frame, and dashed lines indicating 8 updates per frame. The frame rate of the system was 16.9 Hz.](image)
3.2 Eye Motion in Human Subjects

The eye tracker was evaluated in two different human subjects, one of them with no known eye disease and one with amblyopia, with good and average fixation, respectively. Sequences of retinal images were recorded using each of the tracking methods discussed above, followed by posterior estimation of the residual displacement by using the NCC. The residual displacements for all image sequences are summarized in the box plots in Fig. 6, showing that all tracking strategies performed better than no tracking.

When eye motion is so large that most of the features within the reference frame no longer appear within the FOV, the tracking fails. This can occur for example after blinks, and it can explain why all the recorded image sequences show comparable minima and maxima displacements.

In agreement with the results from the tests using sinusoidal motion in a model eye, the tracking with finer strips performs best. Within these, the Kalman and LPF performed worse, against the findings using sinusoidal motion. No significant performance differences were observed between the horizontal and vertical directions, that correspond to the fast and slow optical scanning, respectively. This is interesting given the asymmetry in scanning speeds, also suggesting that the NCC appears to be equally robust to image shear and compression.

Figure 6. Residual displacements in image sequences collected using different tracking methods for two different subjects: No Tracking (NT), No Prediction (NP), Linear Extrapolation Full Frame (LFF), Quadratic Extrapolation Full Frame (QFF) Kalman Filter Full Frame (KFF), Low Pass Filter Full Frame (LPFF), No Prediction Sub-Frame (NPSF), Linear Extrapolation Sub-Frame (LSF), Quadratic Extrapolation Sub-Frame (QSF), Kalman Filter Sub-Frame (KFSF), Low Pass Filter Sub-Frame (LPSF). The box represents the inner quartile range while bars represent the maximum and minimum of the displacements.

4. CONCLUSION

AOSOs allow for noninvasive high resolution imaging of the human retina, and have the potential to become an import tool in early diagnosis of eye disease, help accelerate the development of new therapies and increase our understanding of the mechanisms underlying eye disease. Engineering challenges remain to translate this technology to the clinic, with involuntary eye motion being one of the major problems that needs to be addressed. Compensating for eye motion during image acquisition has the potential to dramatically reduce the number of images required, which would translate in reducing imaging time, light exposure, data storage requirement, operating and processing costs. Eye tracking has also the potential to enable unprecedented longitudinal studies, in which individual retinal cells can be tracked over multiple visits.
In this paper, we have demonstrated an AOSO with an elementary hardware eye tracker that does not increase the optical complexity of the instrument or reduce its optical throughput. Two different strategies using full- and sub-frame motion estimation and correction were tested, with the latter showing the better performance. Multiple motion prediction methods were evaluated in both a controlled experiment and with human subjects. Even though these tests were performed only in two human subjects, two main conclusions emerged. First, the use of fine strips is better than using coarse strips or even full frames for motion estimation and correction, in terms of reducing the residual eye displacement. Evaluation of the quality of the resulting images remains to be performed. Second, and against expectation, the control method using no prediction seems to perform as well as or better than the ones using prediction. This might suggest that the parameters on these models might require further refinement, but this is promising in that a very simple non-predictive control can achieve comparable performance to that of more complex (and therefore more difficult to operate) predictive algorithms.

ACKNOWLEDGMENTS

Alfredo Dubra-Suarez, Ph.D., holds a Career Award at the Scientific Interface from the Burroughs Welcome Fund. This research was supported financially by Research to Prevent Blindness.

REFERENCES