In *Proceedings of the 2010 Symposium on Eye Tracking Research and Applications* (ETRA), ACM, New York, pp. 129-132.

Robust Optical Eye Detection During Head Movement

Jeffrey B. Mulligan*
NASA Ames Research Center

Kevin N. Gabayan†
Stanford University

Abstract

Finding the eye(s) in an image is a critical first step in a remote gaze-tracking system with a working volume large enough to encompass head movements occurring during normal user behavior. We briefly review an optical method which exploits the retroreflective properties of the eye, and present a novel method for combining difference images to reject motion artifacts. Best performance is obtained when a curvature operator is used to enhance punctate features, and search is restricted to a neighborhood about the last known location. Optimal setting of the size of this neighborhood is aided by a statistical model of naturally-occurring head movements; we present head-movement statistics mined from a corpus of around 800 hours of video, collected in a team-performance experiment.

Keywords: eye finding, pupil detection, active illumination, head movement

1 Introduction

To obtain high-accuracy, a video-based eye-tracker must acquire a high-resolution image of the eye. For any given camera, the best image will be one in which the eye occupies the entire frame. For a fixed (non-steerable) camera, however, the camera must maintain a constant position relative to the head, to prevent the eye from leaving the frame when the subject makes a head movement. Some systems solve this problem by attaching the camera(s) to the head; this provides good eye image quality, but requires the subject to wear a head-mount, and requires tracking of the head in order to refer gaze to objects in the world. Alternatively, a remote camera may be used, with the subject's head stabilized with a chin-rest or bite-bar. Neither of these approaches provides the subject with a particularly natural experience, however, and are inappropriate when the goal is to covertly observe behavior "in the wild."

To maintain high accuracy while still allowing a range of natural movement, the image acquisition system must be able to follow the eye as it moves in space. This can be accomplished by mechanical steering of the viewing axis (e.g., by using a pan-tilt-zoom camera, or steering the line-of-sight with galvonometer mirrors), or by moving a region-of-interest inset in a large, high-resolution image from a fixed camera with a wide field-of-view. Currently, commercial offerings are available using both approaches.

Regardless of how repositioning the area of analysis is accomplished, the system must determine *where* to reposition it. Additionally, when the initial position of the head is undefined, the eye(s) must be located to begin tracking, and it is desirable for this to be accomplished automatically. A number of techniques are available for following the eyes using conventional imagery [Tong and Ji 2008]; in the following section we consider the special case of optical pupil detection, and how it is affected by head motion.

*e-mail: jeffrey.b.mulligan@nasa.gov
†e-mail: kevingabayan@gmail.com

2 Optical eye detection

The optical properties of the eye itself enable a unique approach to eye detection which exploits the fact that the eye is a natural retroreflector [Ebisawa and Satoh 1993; Morimoto et al. 2000]. This is commonly observed as the "red eye" effect seen in flash photography, occurring when the light source is sufficiently close to the camera lens that light reflected by its image on the retina returns and enters the camera lens. By rapidly alternating on- and off-axis illumination, a pair of images can be obtained which are nearly identical except for the pupil region; subtracting these images results in an image in which the pupils are prominent, greatly simplifying the machine vision problem.

There are a variety of ways in which the two illumination channels can be multiplexed. One early system [Grace et al. 2001] used a pair of cameras, a beam splitter and narrow-band filters to implement *wavelength multiplexing*, allowing simultaneous acquisition of bright- and dark-pupil images. A single-camera solution using wavelength multiplexing has been demonstrated [Fouquet et al. 2004], using a custom sensor chip equipped with a checkerboard array of filters. While this approach is elegant, the sensors are not commercially available, and so it is unavailable to most would-be users. A more common approach that is amenable to a single-camera system is *temporal multiplexing*, in which the illuminator channels are energized in alternation, synchronized with the camera's frame rate. Temporal multiplexing requires additional circuitry for dynamic control of the illuminators, and more complicated software to correctly reject artifacts arising from head motion, which we will explore in this section.

We have previously incorporated a stereo pair of wide-field cameras equipped with such an active illumination system into a multicamera platform for monitoring workstation use [Brolly and Mulligan 2004]. Our system employs standard analog cameras outputting interlaced video, with field-rate alternation of illuminator channels, as in Ebisawa and Satoh [1993]. This system works quite well for eye acquisition when the head is still. Because the two illuminator channels are alternated in time, however, the system suffers from motion artifacts (see figure 1). One approach that has been employed to reduce the deleterious effects of head motion in this situation is to measure the motion in the field and shift one of the images prior to computing the difference image [Ebisawa 2009]. Here we present an alternative approach that does not depend on measurement of image motion. The key to our method is the observation that the motion artifacts differ in polarity for *within-frame* and *across-frame* difference images.

Figure 1 shows a series of images obtained with this system during a lateral head movement. The top two rows show the raw images, with bright pupil illumination delivered during the odd fields, and dark pupil illumination delivered during the even fields. The third and fourth rows show the difference images: the third row shows the "within frame" difference, in which the even field of the current frame is subtracted from the odd field of the current frame, while the fourth row shows the "across frame" differences, in which the

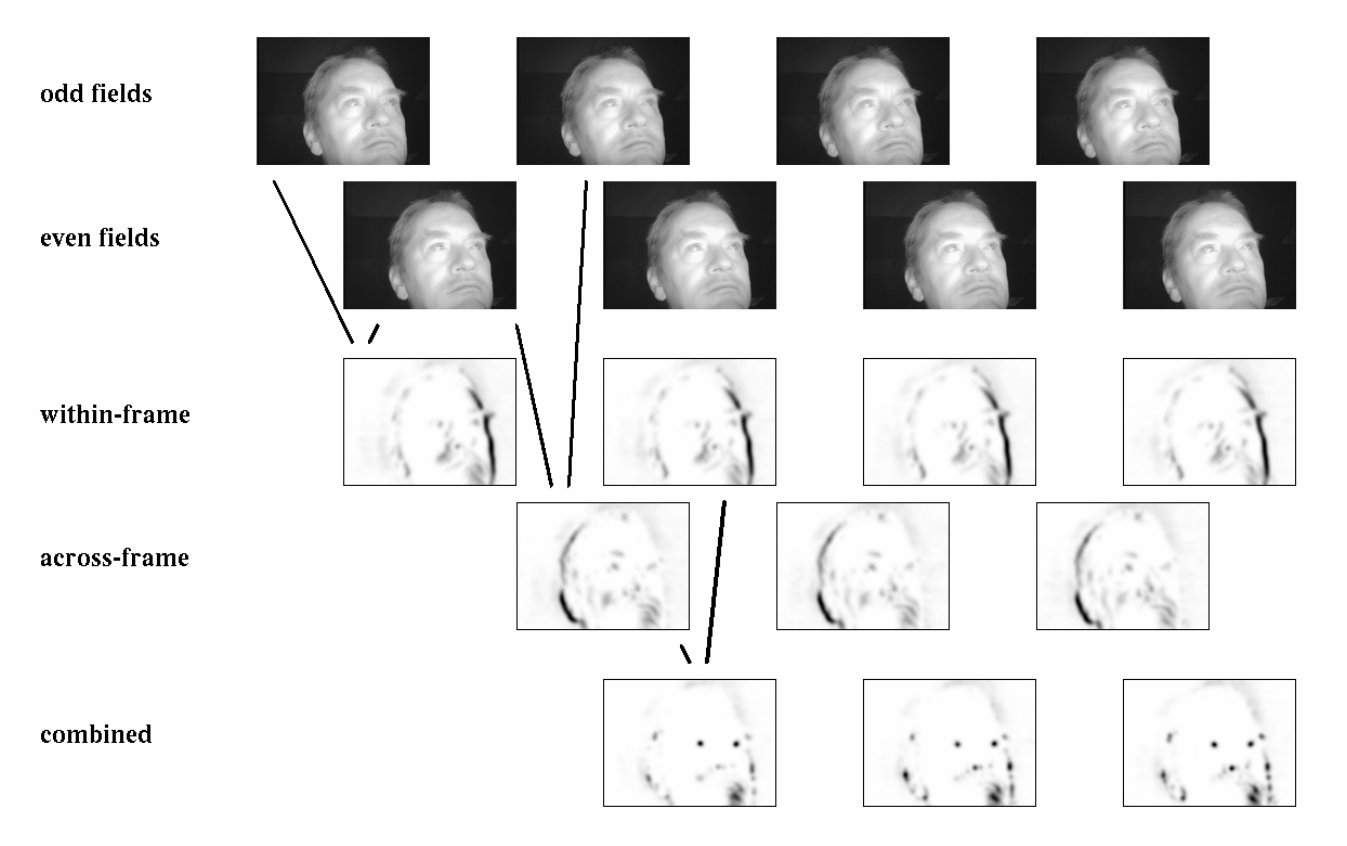

Figure 1: A series of images from a wide-field camera equipped with an active illumination system. Bright pupil illumination is provided during odd fields, while dark pupil illumination is provided during even fields. Within-frame difference images are obtained by subtracting the two fields from a single frame, while across-frame difference images are obtained by subtracting fields from different frames. These two types of difference images show different types of motion artifacts, which may be in large part suppressed by taking their product, as shown in the bottom row. To enhance visibility, all difference images have been inverted, blurred, and normalized.

second (even) field of the previous frame is subtracted from the first (odd) field of the current frame. The difference images are clipped from below at zero, in order to suppress regions where the dark pupil image is brighter than the bright pupil image. The difference images have been inverted and slightly blurred to enhance visibility in the present reproduction.

To get the most up-to-date estimate of the pupil location, we should use the most recently acquired pair of bright- and dark-pupil images. But, as illustrated in figure 1, the temporal ordering of the bright- and dark-pupil fields alternates, and so the polarity of motion-induced artifacts alternates as well. The half-wave rectification performed after the differencing suppresses roughly half of the motion-induced artifacts; in figure 1 the head is moving to the left, and so strong motion-induced artifacts are seen at the left edge of the head in the within-frame differences, and on the right edge of the head in the across-frame differences. Because a particular head-edge artifact is in general only present in one of the two types of difference image, we are able to eliminate most of them (and reduce the strength of the remainder relative to the eye signals) by *pixel-wise multiplication* the pair of difference images derived from a single bright-pupil image. We can express this more formally as follows: let $\mathbf{o}_i$ and $\mathbf{e}_i$ represent the odd and even fields (respectively) from frame i. We further let $\mathbf{w}_i$ and $\mathbf{a}_i$ represent the within- and across-frame differences, respectively, and let $\mathbf{c}_i$ represent the combined difference image. Then

$$\mathbf{w}_i = \max(\mathbf{o}_i - \mathbf{e}_i, 0), \quad (1)$$

$$\mathbf{a}_i = \max(\mathbf{o}_i - \mathbf{e}_{i-1}, 0), \quad \text{and} \quad (2)$$

$$\mathbf{c}_i = \mathbf{a}_i \mathbf{w}_i, \quad (3)$$

where the all vector operations are understood to represent the corresponding pixel-wise operation on the images.

It can be seen in figure 1 that the combined difference image has a strong signal in the eye regions, and has reduced the strength of the motion artifacts. The source of the eye signal in this case can be quite different than in the case of a stationary head; when the head and eyes are still, the *only* difference between the two images will be the brightness of the pupil region - the glints (often the brightest part of the image) will cancel. When the head and/or eye is in motion, on the other hand, the glints will not cancel each other; but when our goal is to simply find the eye (as opposed to accurately segmenting the pupil region), then this is not a problem: because the glint in the bright pupil image can be expected to be brighter than anything else with which it might be (mis)aligned, we are likely to get a strong signal in the difference image even when the pupil signal is degraded.

We have performed a preliminary evaluation of the technique using a video sequence captured while a subject executed voluntary, side-to-side head movements, ranging from slow to as fast as possible.

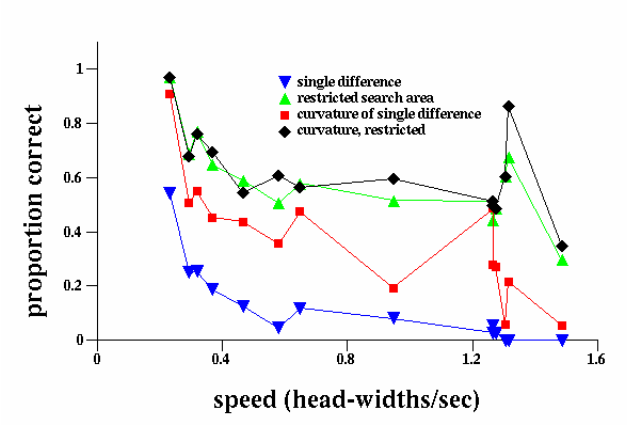

Figure 2: Proportion of frames in which both eyes were correctly localized as a function of average head speed, for four algorithms, based on: single inter-frame differences (inverted triangles); difference images with a restricted search region (upright triangles); Gaussian curvature of difference image (squares); and Gaussian curvature with a restricted search region (diamonds).

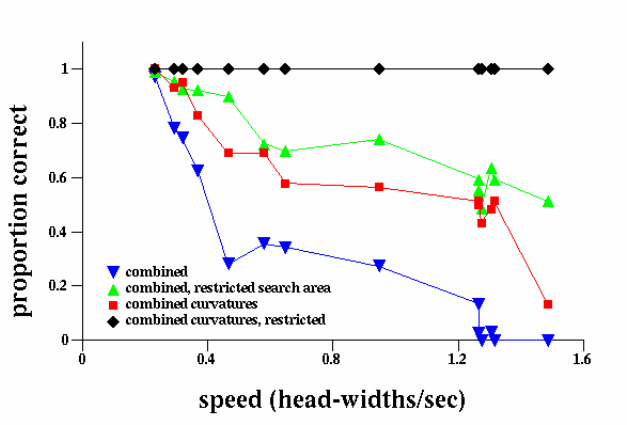

Figure 3: Similar to figure 2, for four algorithms in which within- and across-frame difference images were combined as illustrated in figure 1: combined difference images (inverted triangles); combined difference images with a restricted search area (upright triangles); combined curvature images (squares); and combined curvature images with a restricted search area (diamonds).

The imaging configuration was similar to that shown in figure 1, with the head taking up roughly half the width of the frame (about 300 pixels), and an effective frame rate of 60 Hz after deinterlacing. Difference images were computed, and blurred slightly. Our baseline algorithm was to search for the first two local maxima in the blurred difference image. While this simple procedure performs satisfactorily when the head is still, it breaks down as significant head motion is encountered (see figure 2). A couple of simple enhancements can improve performance somewhat: first we can restrict the search area for each eye to a small neighborhood centered on the previous position (in our example we used a radius of 7 pixels, larger than the highest head speed). Secondly, we can apply a transformation to the difference image which accentuates punctate features while suppressing linear features; this is motivated by the observation that many of the most severe motion artifacts occur around the edges of the face and are elongated. The transformation we chose was Gaussian curvature [Zetzsche and Barth 1990; Barth et al. 1998], a combination of second derivatives which is zero for linear features but responds strongly to spots and corners. We can see in figure 2 that, while application of the curvature operator does produce a performance benefit, the greatest benefit comes from assuming the eyes are not far from their previous locations. This too breaks down, however, at moderate head speeds.

Figure 3 shows similar results for various combinations of within- and across-frame differences. Simple combination of difference images (as illustrated in figure 1) produces a significant improvement over single difference images, but still fails frequently. The within- and across-frame combination can be applied to curvature images as well, and some improvement is seen. But as was the case for the simple differences, restriction of the search area produces the largest single improvement; for our test sequence, we obtain near-perfect performance for combination of curvatures of within- and across-frame difference images, with a search area restricted to a neighborhood about the previous location.

3 Modeling natural head movements

In the previous section, we saw that the eye regions can be tracked much more robustly when they are assumed to be relatively near to their previous locations. In order to intelligently determine the size of the search window, it is therefore useful to know something about the expected movements of the subjects. Such data are also useful for determining parameters of a remote gaze tracking system, such as field-of-view and steering speed. To this end, we analyzed a large corpus of video, provided to us by a research group studying team decision-making [Fischer et al. 2007]. In this study, 5-member teams participated in a (simulated) tele-robotic search mission, by manning console stations in each of 5 separate experimental chambers. Video footage of each subject's work session was recorded by a video camera mounted above their workstation, with a horizontal field-of-view of about 40 degrees. Although the subjects in the videos assume a variety of seated poses while at work, their detected faces typically occupied between roughly one-quarter to one-third of a frame width (80 - 106 pixels). The corpus made available to us consisted of 258 separate recordings, comprising a total of 793 hours (42.8 million images). The recordings had already been subjected to some preprocessing; as received by us, each movie had a resolution of 320 x 240, at a temporal sampling rate of 15 frames per second.

To detect the location of the face in each frame, we used a boosted cascade of classifiers [Viola and Jones 2001]. We used the implementation provided in the OpenCV library [Lienhart and Maydt 2002]. Exactly 1 face was found in approximately 50.9% of the frames analyzed. The recordings included periods that were not part of the study; typically recordings began with an empty booth, depicted a brief flurry of activity as the subject and experimenter entered, and then a relatively long period with the subject seated at the console. So, some fraction of the frames in which no faces were detected are indeed correct rejections. However, the face detector as implemented is trained to detect frontal faces, and fails when the head is turned, presenting an oblique or profile view. Similarly, it fails when the subject tilts his or her head down (for example, to read a clipboard held in the lap). We feel that these misses do not greatly diminish the value of the dataset, because in these cases a gaze tracking system would similarly be hard-put to make sense of the image.

The outputs of the face detector are the location and size of a square window containing the face. From the raw data, we computed the inter-frame displacement of the center of the face-box, for all pairs of temporally contiguous frames in which exactly one face was de-

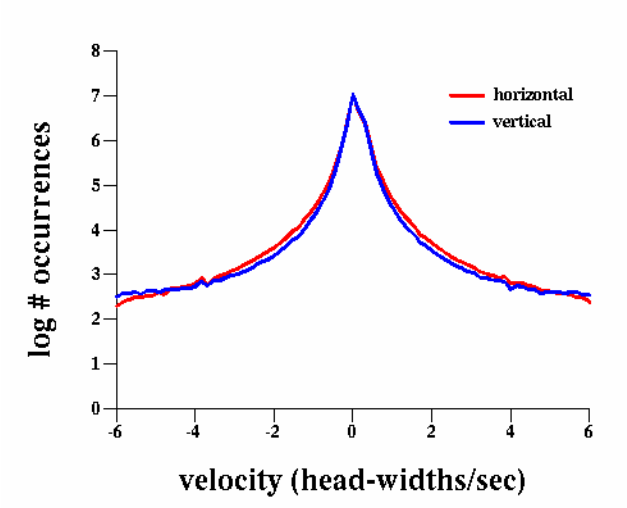

Figure 4: Plot showing histogram for inter-frame face displacement, computed from a large corpus of video data using the Viola-Jones face detector, as implemented in OpenCV.

tected in each frame of the pair. Because the detector missed the face on some frames when the subject was present, there are gaps in the records, and the number of transitions available for analysis is somewhat less than the number of frames. The number of transitions analyzed was approximately 47.1% of all transitions, corresponding to 92.4% of the frames in which one face was detected. The results are shown as a histogram of displacements in figure 4. Face displacements in pixels have been converted to face-widths by dividing each displacement by the average of the box sizes in the two frames making up the pair. Note that the vertical axis is logarithmic; the bulk of the displacements are at or near zero, reflecting the fact that the subjects sat still more than they moved.

It can be seen that the central portion of the distribution in figure 4 is within ± 1 head-width per second. (Because the voluntary head movements shown in figures 2 and 3 do not exceed 2 head-widths per second, we assume that the tails of the distribution shown in figure 4, which extend beyond the range plotted, represent large jumps of the face detector to false targets.) Assuming that an eye-width is about one fifth of a head-width, this corresponds to 5 eye-widths per second, or 0.083 eye-widths per field of 60 Hz video. For a high-magnification image in which the eye fills a frame having an angular extent of 4 x 3 degrees (corresponding to a camera located at arm's length from the subject), we see that this corresponds to an angular rate at the camera of 0.33 degrees per field, or 20 degrees per second. This provides a performance goal for robotic cameras used for remote gaze tracking.

4 Conclusion

In this paper we have demonstrated an novel approach to active illumination eye detection, that combines pairs of difference images to reduce the effects of head movements. We have also examined the movement statistics of users working in a naturalistic setting at a computer console, and found that the majority of movements do not exceed a speed of 1 head-width per second. This is therefore a sensible target goal for a head-tracking system designed to keep the eye region within the area of analysis in a remote gaze-tracking system. By considering the camera field-of-view, and average distance from the subject, we can appropriately size the search region when repositioning the region-of-interest in a high-resolution image.

References

BARTH, E., ZETZSCHE, C., AND KRIEGER, G. 1998. Curvature measures in visual information processing. *Open Systems and Information Dynamics 5*, 25–39.

BROLLY, X. L. C., AND MULLIGAN, J. B. 2004. Implicit calibration of a remote gaze tracker. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, vol. 8, 134.

EBISAWA, Y., AND SATOH, S. 1993. Effectiveness of pupil area detection technique using two light sources and image difference method. In *Proceedings of the 15th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA*, A. Y. J. Szeto and R. M. Rangayan, Eds., 1268–1269.

EBISAWA, Y. 2009. Robust pupil detection by image difference with positional compensation. In *Proceedings of the 2009 IEEE International Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems*, 143–148.

FISCHER, U., MCDONNEL, L., AND ORASANU, J. 2007. Linguistic correlates of team performance: toward a tool for monitoring team functioning during space missions. *Aviation, Space, and Environmental Medicine 78*, 5 Suppl., B86–B95.

FOUQUET, J. E., HAVEN, R. E., VENKATESH, S., AND WENSTRAND, J. S. 2004. A new optical imaging approach for detecting drowsy drivers. In *IEEE Intelligent Transportation Systems Conference*, IEEE, WeA6.

GRACE, R., BYRNE, V. E., BIERMAN, D. M., LEGRAND, J., GRICOURT, D., DAVIS, B. K., STASZEWSKI, J. J., AND CARNAHAN, B. 2001. A drowsy driver detection system for heavy vehicles. In *Proceedings of the 17th Digital Avionics Systems Conference*, vol. 2, I36/1–I36/8.

LIENHART, R., AND MAYDT, J. 2002. An extended set of haar-like features for rapid object detection. In *IEEE ICIP 2002*, 900–903.

MORIMOTO, C. H., KOONS, D., AMIR, A., AND FLICKNER, M. 2000. Pupil detection and tracking using multiple light sources. *Image and Vision Computing 18*, 4, 331–335.

TONG, Y., AND JI, Q. 2008. Automatic eye position detection and tracking under natural facial movement. In *Passive Eye Monitoring*, Springer, R. I. Hammoud, Ed., 83–107.

VIOLA, P., AND JONES, M. 2001. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 511–518.

ZETZSCHE, C., AND BARTH, E. 1990. Fundamental limits of linear filters in the visual processing of two-dimensional signals. *Vision Res. 30*, 1111–1117.