DETECTION OF EYE LOCATIONS IN UNCONSTRAINED VISUAL IMAGES

Ravi Kothari, and Jason L. Mitchell

Department of Electrical & Computer Engineering and Computer Science
University of Cincinnati
Cincinnati, OH 45221-0030
E-Mail: ravi.kothari@uc.edu

ABSTRACT

This paper describes a computational approach for accurately determining the location of human eyes in unconstrained monoscopic gray level images. The proposed method is based on exploiting the flow field characteristics that arise due to the presence of a dark iris surrounded by a light sclera. A novel aspect of the proposed method lies in its use of both spatial and temporal information to detect the location of the eyes. The spatial processing utilizes flow field information to select a pool of potential candidate locations for the eyes. Temporal processing uses the principle of continuity to filter out the actual location of the eyes from the pool of potential candidates. Extensions for gaze angle determination, and the tracking of human point-of-regard are indicated.

1. INTRODUCTION

The ability to estimate the location, gaze angle, and the trajectory of eyes from unconstrained visual images finds application in the design and implementation of next-generation man-machine interfaces. For example, such a system would allow the eyes to act as an alternate input modality for computer users in general and disabled users in particular [3] by replacing *point-and-click* of a mouse with *glance-and-blink* of the eyes (see also [4]). In addition to potential use as a pointing device, it is also possible to make use of point-of-regard data to enhance human computer interaction in ways which are *not* apparent to the user. For example, a user may be presented with additional information based upon current area(s) of interest in the visual display—as determined by recent gaze fixations [5]. Additionally, it would be possible to render complex images such that maximum detail is present at the user's current area of interest while fine details outside of the area of interest are ignored—until such time as the user's focus of attention shifts. Such a system would give the impression of displaying higher quality graphics than might ordinarily be capable at a given throughput. Further, if the movements of the eyes at different levels of alertness can be characterized, then it may be possible to assess operator vigilance.

Invasive eye-tracking techniques have been employed for a number of years by cognitive psychologists seeking to study eye movements with regard to perception [6]. Such methods involved mounting cumbersome devices on a subject's head and often required the head to be immobilized by means of a bite-plate or other stabilizing mechanism. These techniques have produced acceptable results in the controlled scenarios of cognitive psychology experiments, but they are undesirable if one hopes to leave the subject completely uninhibited. A truly non-invasive eye-tracking system should make no assumptions about lighting conditions, should not inhibit the subject(s), or alter their surroundings in an appreciable way. More recently, a number of non-invasive eye-location and eye-tracking systems employing digital cameras have been proposed. These systems all use monoscopic grayscale images of the subject being tracked. Vincent [7] uses a hierarchical technique which employs neural networks to locate the features. The first stage performs a coarse segmentation of the image by locating areas in a low-resolution version of the original image which may contain the desired features (i.e. the left eye, right eye and mouth). The high resolution search areas isolated by the first stage are then processed by another neural network to extract more detailed information about the precise position of the eyes. Starker [5], and Hutchinson [3] place an infrared light source in the location of the camera and use a camera which is sensitive to infrared. Due to the reflectivity of the eye, a bright specular reflection appears on its surface (see also [1]). With this technique, the gross location of the eyes are determined and the image may be processed further to extract detailed point-of-regard information. Additionally, the location of the infrared spot itself may be used since it will remain essentially static relative to the moving pupil. Once the

eyes or potential candidates for eyes are located within the image, detailed information about their exact position must be extracted. Yuille [8] uses deformable templates to describe eye-location and shape. It is assumed that a gross image segmentation has already been performed (i.e. the center of the deformable templates are close to their eventual values following minimization of an energy functional).

In the following, we propose a method which accurately locates eyes in monoscopic gray scale images under normal lighting conditions and a variety of poses without using infrared reflections to locate the irises. A novel aspect of the proposed method lies in its use of both spatial and temporal information to detect the location of the eyes. The spatial processing utilizes flow field (gradient direction field) information to select a pool of potential candidate locations for the eyes. Temporal processing uses the principle of continuity to filter out the actual location of the eyes from the pool of potential candidates.

2. DETERMINING EYE LOCATION

The most significant feature of the eyes, in a grayscale image, are the iris' ellipsoidal shape and the stark contrast between the iris and the surrounding area—the sclera. A dark circular area, like the iris, on a light background, causes an outwardly radiating flow field to appear in a gradient direction plot of the image. Figure 1 shows the local gradient direction at each location of the image, and is a typical example of the outwardly radiating flow field around the iris.

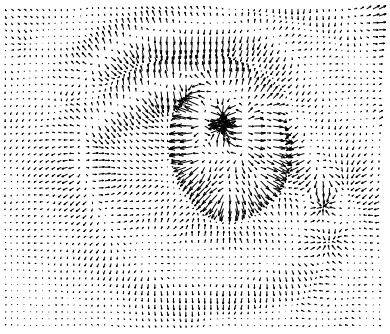

Figure 1: Gradient directions in a typical image.

Since the flow field around an iris radiates outwards it must intersect at a point if extrapolated in a direction opposite to the gradient (the gradient points in the direction of increase of a function, hence points outwards — from the darker iris to the lighter sclera). We thus use a 2-dimensional array of *bins*, initialized to 0, which serve as accumulators. A line drawn at each edge point along the direction opposite the gradient passes through several of these bins. The bins are incremented each time one such line passes through it. This process is shown for two arbitrary local gradients in Figure 2. Note that the bin which lies at the intersection of the two local gradients has the largest accumulation (shown by a lighter shade). This illustrates the fact that a bin which lies at the intersection of lines in the direction of local gradients will be incremented a large number of times compared to bins elsewhere in the image. This is the basis for our algorithm. It is clear that a radiating flow field such as that found at an iris will cause a similar additive effect in the bin(s) located approximately in the center of the iris.

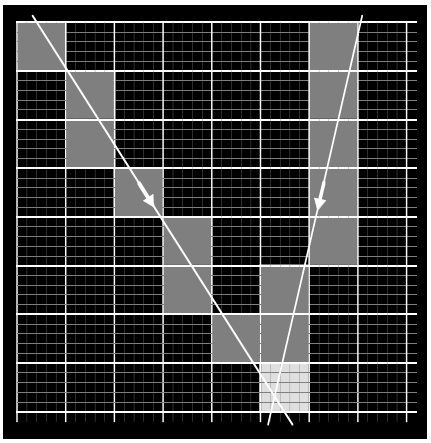

Figure 2: Illustration of the bin-incrementing scheme for two arbitrary local gradients. The broken lines separate the pixels in the image, while the solid lines show the bins. Lighter bins indicate greater accumulation.

There are two variables in the above process which have an impact both in terms of efficiency and effectiveness. The size of each bin (in terms of pixels) is one of these. Larger bins require less computational accuracy at the expense of localization accuracy. Through experimentation we have found that a bin size of 5 pixels provides a reasonable trade-off between localization accuracy and computational efficacy for a wide range of camera-subject distances ($\approx$ 3—12 feet). The second factor which adds to the computational efficiency is to only consider gradients of significant magnitude (e.g. above the rms magnitude). This serves to reduce computation and is unlikely to filter out the local gradients along the edges of the iris due to their high magnitude.

In an ideal situation one might anticipate that two bins (corresponding to the two eyes) will show the largest

accumulations. Due to the background noise, patterns on the subject's clothes etc. it is likely that more than two bins show comparably large accumulations. We have dealt with this in two ways. The first of these methods uses further processing of the bin accumulations (spatial processing), and the second uses temporal information to ascertain the correct location of the eyes. We discuss the two methods below.

Method 1: To obtain the correct eye locations from a pool of potential candidates, we use two heuristics: (i) the bin accumulations for both the eyes should be reasonably close, (ii) the eyes should be spatially close in terms of the y coordinate. Both the conditions are reasonable but only for an upright subject. Denoting the bin accumulations at location (i, j) by b_{ij}, the above translates into finding the two bins which minimize:

$$J = (b_{ij} - b_{i'j'})^2 + \lambda(j - j')^2 \qquad (1)$$

where, (i, j) and (i', j') denote the locations of the two bins, and λ controls the relative weighting that each of the above two criteria have. To prevent a combinatorial growth of the bin pairs considered, one might limit the search to r bins having the highest accumulations.

Method 2: The second method of choosing the correct eye locations from a pool of prospective candidates relies on the principle of continuity in the location of the eyes from one frame to the next. Succinctly,

$$B_{ij}^{(k+1)} = \sum f(B_{mn}^{(k)})e^{-\tau} + b_{ij}^{(k+1)} \qquad (2)$$

where, k has been added to index the frames (time), and τ controls the rate at which the cumulative accumulations decay, mn denotes the bin within a spatial neighborhood bin ij, and f is a function that rapidly decreases monotonically with the distance from bin ij. Thus, the total bin accumulation of each frame is a proportion of the total accumulation of itself and its neighbors, and its current accumulation. There are two advantages of this method: (i) it does not allow inconsistent results of a single frame to eliminate the correct eye locations, and (ii) it provides a more sophisticated method to mark the areas of the next frame to which processing should be restricted (those areas with the highest cumulative accumulations). Results of the this second method are particularly robust if a one time 'calibration' step is performed which identifies the eye locations. Indeed, a combination of methods 1 and 2 can be used if the assumptions of method 1 can be justified in a particular application.

We now show the results of the above algorithm on a set of sample images (Figure 3) chosen to maximize typical variations in facial characteristics (e.g. glasses, beard, long hair etc.).

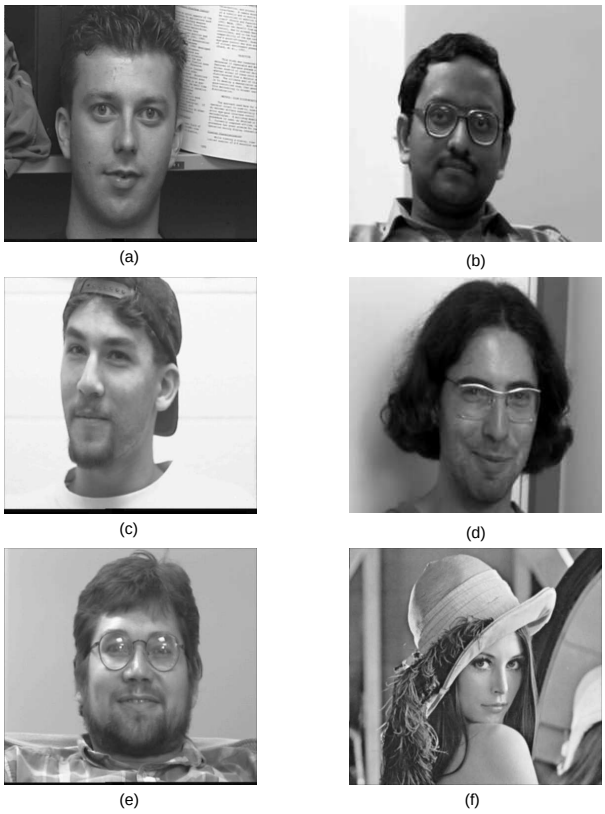

Figure 3: Test images

The results of the bin incrementing process are shown in Figure 4. Low bin counts are represented by dark pixels while high bin counts are shown as bright pixels.

Figure 5 shows the result of further spatial processing based on method 1. Shown in Figure 5 are the two bins which minimized the cost function of equation (1), overlaid on the original images. Note the correct location of the eyes in each instance.

3. DISCUSSION AND CONCLUSION

In this paper we presented a principled approach towards detecting the locations of human eyes in unconstrained images. We have obtained reliable results over a wide range of subjects using the approach presented above. Extensions to allow for gaze angle determination and tracking the human point-of-regard are underway. Our approach to gaze angle determination is based on the change in the eccentricity of the iris as the gaze angle increasingly deviates from 0 deg (straight ahead), adjusted based on deviation of the vertical axis of symmetry of the face with the 0 deg line. Due to the

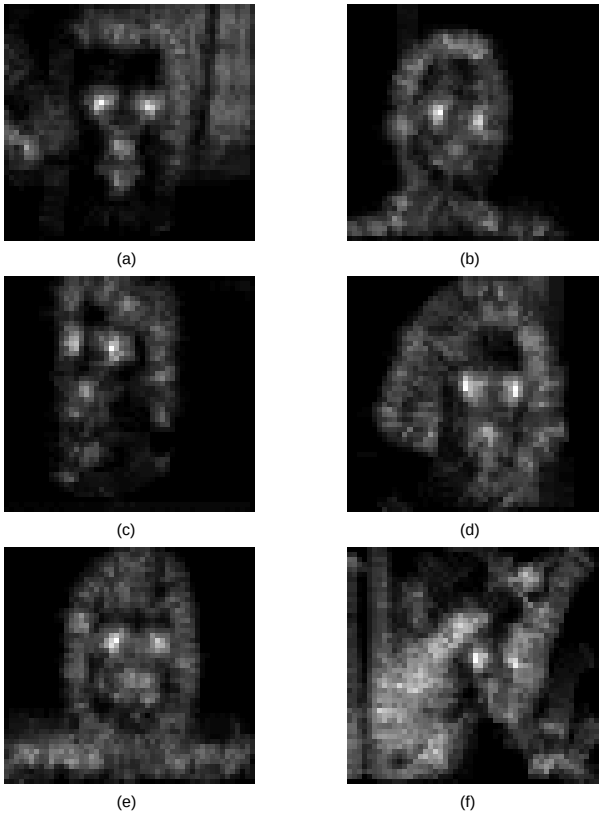

Figure 4: Bins for the images of Figure 3

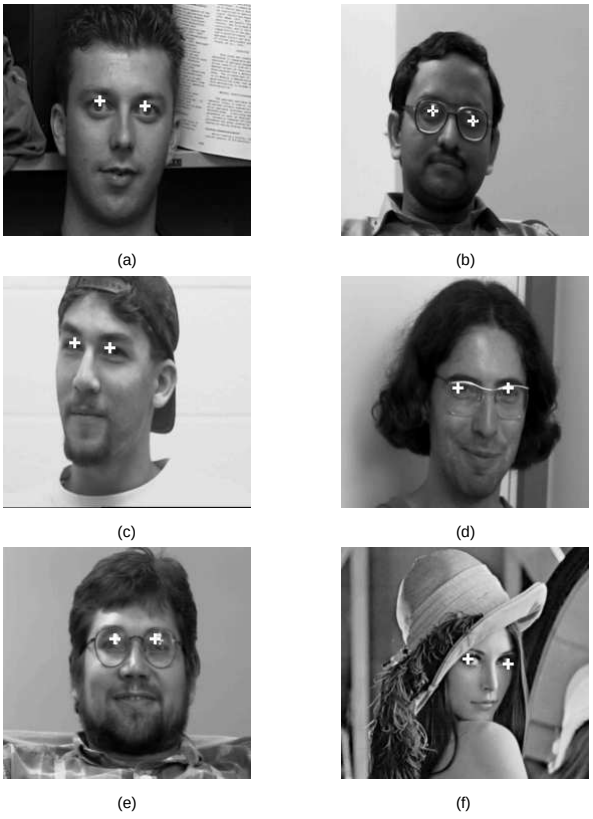

Figure 5: Location of the eyes as determined shown overlaid on the respective images

presence of saccadic and smooth movements in the eye, our approach to tracking human point-of-regard relies on the use of modular neural networks composed of two experts, each individually responsible for saccadic and smooth movements of the eyes [9].

4. REFERENCES

[1] S. Baluja and D. Pomerleau, "Non-Intrusive Gaze Tracking Using Artificial Neural Networks," *CMU Technical Report CMU-CS-94-102*, 1994.

[2] R. Bolt, "Gaze-Orchestrated Dynamic Windows," *Comp. Graph.*, vol. 25, no. 3, pp. 109-119, 1981.

[3] T. E. Hutchinson, et. al, "Human-Computer Interaction Using Eye-Gaze Input," *IEEE Trans. on Sys., Man and Cyber.*, vol. 19, no. 6, pp. 1527-1534, 1989.

[4] R. J. K. Jacob, "The Use of Eye Movements in Human-Computer Interaction Techniques: What You Look At Is What You Get," *ACM Trans. Info. Sys.*, vol. 9, no. 3, pp. 152-169, 1991.

[5] I. Starker, and R. A. Bolt, "A Gaze-Responsive Self-Disclosing Display," *Proc. ACM CHI '90 Human Factors in Computing Conference*, pp. 3-9, 1990.

[6] L. R. Young and D. Sheena, "Survey of eye movement recording methods," *Behav. Methods and Instr.*, vol. 7, no. 5, pp. 397-429, 1975.

[7] J. M. Vincent, J. B. Waite and D. J. Myers, "Precise Location of Facial Features by a Hierarchical Assembly of Neural Nets," *Proc. IEE Conf. on Artificial Neural Networks*, pp. 69-73, 1991.

[8] A. L. Yuille, P. W. Hallinan and D. S. Cohen, "Feature Extraction from Faces Using Deformable Templates," *Int. J. Computer Vision*, vol. 8, no. 2, pp. 99-111, 1992.

[9] J. L. Mitchell, and R. Kothari, "Human Point-of-Regard Tracking Using State Space and Neural Network Models," *Proc. IEEE Neural Networks for Signal Processing*, 1996 (to appear).