Real-Time Detection of Eyes and Faces

Carlos Morimoto Dave Koons Arnon Amir Myron Flickner
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120
{*carlos,dkoons,arnon,flick*}*@almaden.ibm.com*

Abstract

Perceptual user interfaces will require the detection, tracking, and recognition of faces and other body and facial features. This paper introduces a robust, accurate, and low cost real-time solution for the eye and face detection problem. The method uses two infra-red illumination sources to generate bright and dark pupil images, which are combined to robustly detect pupils. Once the pupils are detected, the inter-ocular distance is used to determine the size and position of the bounding box around the face. The position of other facial features such as eye brows, nose, and mouth can be estimated once the face is detected. A real-time implementation of the system, which process 30 frames per second using interlaced images of resolution 640×480 pixels, is also presented.

1 Introduction

Perceptual user interfaces will require more sophisticated input and output devices than the current dominant WIMP (windows, icons, mouse, and pop-up menus) interfaces. Many problems still remain to be solved, such as the development of new input sensory devices, and multi-modal architectures that will deal with and combine the information coming from multiple input modes to deliver more natural user interfaces.

We believe that speech understanding and vision will constitute the most fundamental input modes due to their unobtrusiveness and communication power. In particular, this paper presents a new device to detect and track faces and facial features based on computer vision techniques. Once faces and facial features are detected and tracked, they can be used for recognition and identification, to help disambiguate voice commands, as a communication means through head gestures and facial expressions, etc.

The performance of all face tracking and face recognition techniques could improve from a robust face and facial feature detector. Several techniques have been proposed for automatic face detection and tracking, mostly relying on color, texture, shape, or motion cues. Face detection techniques based primarily on skin color were suggested by Fieguth and Terzopoulos [5], and Yang and Waibel [12]. Darrell *et al.*[2] detects faces from the subtraction of a known background. Birchfield [1] uses geometric constraints in addition to color for tracking heads in real-time; and De Silva *et al.*[10] detects and tracks facial features using edge-based methods and templates. Other techniques based on templates and other geometrical constraints [6, 13] as well as artificial neural networks [9] also exist.

A different strategy is to first search for eye candidates and then try to combine them into faces. Kothari and Mitchell [8] use spatial and temporal information to detect eye locations. A pool of potential candidates are selected using gradient fields. The gradient along the iris/sclera boundary always point outward the center of the iris (dark pupil), thus by accumulating along these lines, the center of the iris can be estimated by selecting the bin with highest count. Heuristic rules and a large temporal support are used to filter erroneous candidates.

The technique introduced in this paper explores some geometric and physiological properties of the eye to first detect pupils and then use the pupil locations to determine the size and position of faces. It does not require models (color, geometry, templates, examples, etc), and is based on geometrical and physiological properties of the eye. Although special lighting and synchronization schemes are required, the scene background becomes irrelevant and the pupils can be detected in a wide range of scales and illumination conditions. The next section introduces the system's principle of operation, Section 3 describes how pupils and faces are segmented, and the real-time implementation is presented in Section 4. Section 5 concludes the paper.

2 Principle of Operation

Commercial remote eye-tracking systems used for the estimation of a person's gaze such as those produced by ISCAN and Applied Science Laboratories (ASL), rely on a single light source that is positioned off-axis in the case of the ISCAN ETL-400 systems, and on-axis in the case of the ASL E504 systems. Illumination from an off-axis source (and regular ambient illumination) generates dark pupil im-

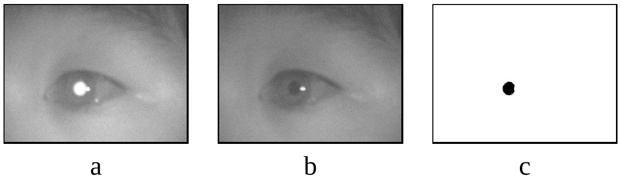

Figure 1. (a) Bright and (b) dark pupil images. (c) Shows the difference of the dark from the bright pupil after thresholding

ages (Figure 1b). When the light source is placed on-axis with the camera optical axis, the camera is able to detect the light reflected from the interior of the eye, and the image of the pupil appears bright [7, 14], as seen in Figure 1a. This effect is often seen as the red-eye in flash photographs. These systems require the initial localization of the pupil in order to begin tracking.

Pupil detection using only dark or only bright pupil images is not trivial since other regions of the image might have similar shapes and appear as dark or as bright as the pupils. But when combining both bright and dark pupil images, the pupils will correspond to regions with high contrast between the bright and dark images, as seen in Figure 1, thus simplifying the detection problem.

In a single user scenario, the position of the user's face can be determined once both pupils are detected. Our pupil detection technique uses a single wide field of view camera and two light sources. For convenience, we use near infra red (IR) light, which is almost invisible to the human eye, and a black and white camera. The wide field of view allows for the detection of multiple pupils. Geometric constraints could also be used to group the pupils, so that several heads could be detected and tracked.

Two sets of IR LED's are distributed symmetrically around the camera's optical axis in order to generate concentric reflections on the cornea. Symmetry is not a requirement for pupil detection, but it reduces shadow artifacts. Consider that the sets are composed of two concentric rings as shown in Figure 2. The inner ring is placed very close to the camera's optical axis, and when turned on, it generates a bright pupil image, such as the example shown in the Figure 1a. The outer ring has a larger radius, large enough to generate a dark pupil image, such as the one shown in the Figure 1b. Observe that the glints, or corneal reflections, from the on and off-axis light sources can be easily identified as the bright regions in the iris. Figure 1c shows the thresholded difference of the dark from the bright pupil images. Observe that only the pupil region is segmented due to the high contrast, within that region, between the dark and bright pupil images.

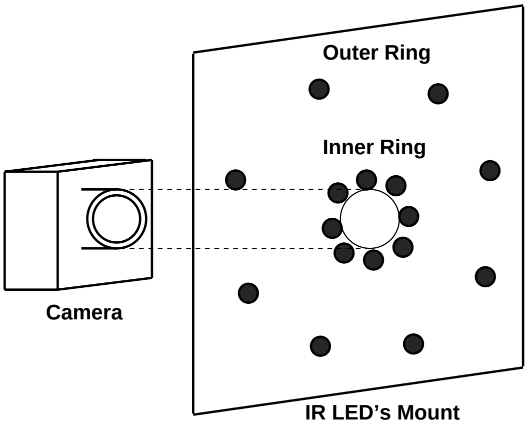

Figure 2. Camera and infra red illuminators

2.1 Related Work

Tomono *et al.*[11] and Ebisawa and Satoh [4] have developed systems very similar to the one presented in this paper. Both systems are based on the differential lighting scheme to track the pupil and estimate the point of gaze, which also requires the detection of the corneal reflections created by the light sources.

Tomono *et al.*[11] developed a real-time imaging system composed of a 3 CCD camera and 2 near infra red (IR) light sources with different wavelengths. Ebisawa and Satoh [4] also use two light sources, but both with the same wavelength to generate the bright/dark pupil images. The detection of the corneal reflection created by the light sources requires the use of a narrow field of view camera (long focal length) since the reflection is in general very small. Ebisawa [3] also presents a real-time implementation of the system using custom hardware and pupil brightness stabilization for optimum detection of the pupil and the corneal reflection.

The system presented in this paper introduces a much simpler and inexpensive solution for the pupil detection problem, and it is also based on the differencing followed by thresholding technique using bright and dark pupil images. We also extend the problem to detect multiple pupils, and group them into faces.

3 Pupils and Face Detection

Pupils are detected from thresholding the difference of the dark from the bright pupil images (Figure 1c). Since interlaced images are being used, the difference image can be computed directly from line subtraction. Figure 3 shows a magnified view of the pupil from the interlaced input frame. The pupil images in Figure 1 are subsampled horizontally for better visualization. The resolution of the input camera frames are 640×480 pixels, so that the dark and bright pupil

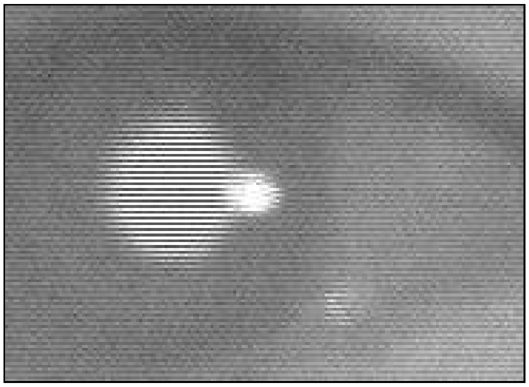

Figure 3. Magnified region around the pupil from an interlaced frame

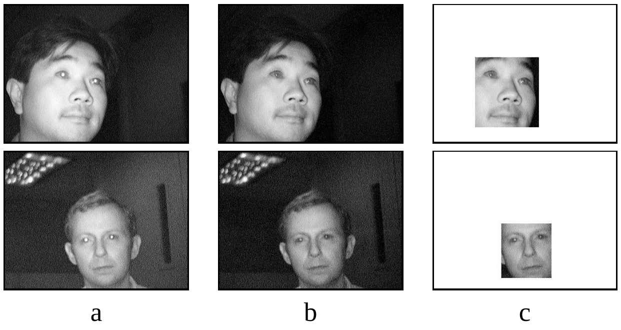

Figure 4. Example of faces detected by the system. Column (a) shows the bright pupil images, (b) the dark pupil images, and (c) the detected faces

images are 640×240 pixels, and the images of Figure 1 are 320×240 pixels.

In the case of large pupil displacements due to fast motion of the head and/or eyes, the pupils can be lost. They are detected again as soon as there is some pupil overlap between fields. If the motion is small, so that some overlap exists, pupils can still be detected, but motion artifacts might make this task more difficult.

Most motion artifacts can be filtered by considering a larger temporal support, i.e., more than two frames are considered for pupil detection. If the eyes do not move much during the sampling period of F frames, the pupils can be detected as high contrast regions from the differences between all consecutive pairs, while most motion artifacts are detected only between some of the pairs (motion from some particular textured surfaces could also be present after temporal filtering). This method introduces a delay of $F-1$ frames every time the pupil is lost, since the pupil must remain approximately still for at least F frames until it is detected again.

Pupil candidates are determined from the thresholded image using a simple region labeling algorithm. The candidates are further constrained based on the aspect ratio and size of the labelled components. The two best pupil candidates are then used to determine the size and position of the face using the following heuristic rule. It is considered that the aspect ratio of a face is 4×5, when normalized by half the inter-ocular distance given by the two pupil candidates. The aspect ratio can be changed arbitrarily, but the normalization by a factor of the inter-ocular distance makes it possible to automatically scale the box around the face to an appropriate size. The middle point between the pupils lies on the coordinate (2,2). Figure 4 shows examples of faces detected using the above scheme.

Other rules and constraints could also be used to detect multiple faces from multiple pupils. Simple methods such as nearest neighbor and temporal consensus could be used to group pairs of pupils, and then for each pair apply the same heuristic used for single faces.

4 System Implementation

We have developed a real-time (30 frames per second) implementation of the pupil and face detection system on a single processor Pentium II-333 MHz machine. It uses a PCI frame grabber with no on-board processing capabilities for image acquisition, and the interlaced frames are synchronized with the bright and dark pupil illuminators by special hardware.

The system process live NTSC video generated by a B&W camera. The video signal is interlaced, i.e., each image frame is composed by an even and an odd field, each field with half the vertical resolution of the original frame. Extra hardware was developed to synchronize the even and odd fields with the inner and outer rings respectively, i.e., when the inner ring is on, an even field is being scanned, and alternately, when the outer ring is on, an odd field is scanned. When the frame grabber digitizes an image, both fields are passed to the computer simultaneously. Hence, even if the computer drops a frame, synchronization is never lost. Different circuitry is required to synchronize the illuminators with image frames instead of fields. The advantage is that the frames are of higher resolution. On the other hand, synchronization becomes harder in the event of dropped frames, and motion artifacts become more noticeable due to the lower sampling rate, which is 30Hz for frames and 60Hz for fields.

Figure 5 shows the pupil detection system operating with typical noisy data. Simple thresholding after subtraction is used. To eliminate artifacts, high contrast regions only one pixel wide were eroded. Observe that the pupils from all subjects were detected, and the false eyes (glass marbles) in the center of the images were rejected since they do not have the correct optic and geometric properties that produce

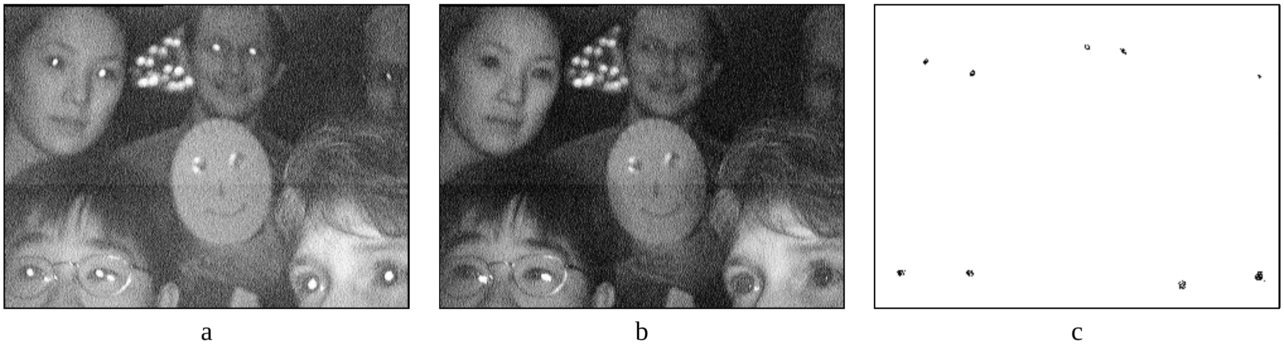

Figure 5. (a) Bright and (b) dark pupil images. (c) Multiple pupils detected from subtraction and thresholding, after erosion

the bright/dark eye effect. Note also that the system is quite robust even for people wearing glasses.

5 Conclusion

We have described a real-time system for eye and face detection that can be used to improve the performance of any interactive eye/face tracking and face recognition system. The even and odd fields of a video camera are synchronized with two IR light sources, generating dark and bright pupil images. Pupil detection is based on thresholding the difference between these images. Once the pupils are detected, they are used to determine the size and position of the face. The robustness of the technique is demonstrated by a real-time implementation that process 30 frames per second using interlaced frames of resolution 640×480×8 pixels.

The system has been successfully tested for a very large number of people, and it is inexpensive and very compact. Future extensions of this work include a system for gaze estimation, and multiple face and facial features detection and tracking.

References

[1] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 232–237, Santa Barbara, CA, June 1998.

[2] T. Darrell, B. Moghaddam, and A. Pentland. Active face tracking and pose estimation in an interactive room. Technical Report 356, M.I.T. Media Laboratory Perceptual Computing Section, Cambridge, MA, 1996.

[3] Y. Ebisawa. Unconstrained pupil detection technique using two light sources and the image difference method. *Visualization and Intelligent Design in engineering and architecture*, pages 79–89, 1995.

[4] Y. Ebisawa and S. Satoh. Effectiveness of pupil area detection technique using two light sources and image difference method. In A. Szeto and R. Rangayan, editors, *Proceedings of the 15th Annual Int. Conf. of the IEEE Eng. in Medicine and Biology Society*, pages 1268–1269, San Diego, CA, 1993.

[5] P. Fieguth and D. Terzopoulos. Color based tracking of heads and other mobile objects at video frame rates. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–27, Puerto Rico,PR, June 1997.

[6] V. Govindaraju, S. Srihari, and D. Sher. A computational model for face location. In *ICCV*, pages 718–721, 1990.

[7] T. Hutchinson, K. W. Jr., K. Reichert, and L. Frey. Human-computer interaction using eye-gaze input. *IEEE Transactions on Systems, Man, and Cybernetics*, 19:1527–1533, Nov/Dec 1989.

[8] R. Kothari and J. Mitchell. Detection of eye locations in unconstrained visual images. In *Proc. International Conference on Image Processing*, volume I, pages 519–522, Lausanne, Switzerland, September 1996.

[9] H. Rowley, S. Baluja, and T. Kanade. Rotation invariant neural network-based face detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 38–44, Santa Barbara, CA, June 1998.

[10] L. D. Silva, K. Aizawa, and M. Hatori. Detection and tracking of facial features. In *Proceedings of the SPIE Com. and Image Proc. 95*, volume 2501, pages 1161–1172, Taipei, Taiwan, May 1995.

[11] A. Tomono, M. Iida, and Y. Kobayashi. A tv camera system which extracts feature points for non-contact eye movement detection. In *Proceedings of the SPIE Optics, Illumination, and Image Sensing for Machine Vision IV*, volume 1194, pages 2–12, 1989.

[12] J. Yang and A. Waibel. A real-time face tracker. In *Proceedings of the Third IEEE Workshop on Applications of Computer Vision*, pages 142–147, Sarasota, FL, 1996.

[13] A. Yiulle, P. Hallinan, and D. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8(2):99–111, April 1992.

[14] L. Young and D. Sheena. Methods & designs: Survey of eye movement recording methods. *Behavioral Research Methods & Instrumentation*, 7(5):397–429, 1975.