Real-Time Multiple Face Detection Using Active Illumination

Carlos H. Morimoto
Departamento de Ciência da Computação - IME/USP
Rua do Matão 1010, São Paulo, SP 05508, Brazil
hitoshi@ime.usp.br

Myron Flickner
IBM Almaden Research Center
650 Harry Rd, San Jose, CA 95120, USA
flick@almaden.ibm.com

Abstract

This paper presents a multiple face detector based on a robust pupil detection technique. The pupil detector uses active illumination that exploits the retro-reflectivity property of eyes to facilitate detection. The detection range of this method is appropriate for interactive desktop and kiosk applications. Once the location of the pupil candidates are computed, the candidates are filtered and grouped into pairs that correspond to faces using heuristic rules. To demonstrate the robustness of the face detection technique, a dual mode face tracker was developed, which is initialized with the most salient detected face. Recursive estimators are used to guarantee the stability of the process and combine the measurements from the multi-face detector and a feature correlation tracker. The estimated position of the face is used to control a pan-tilt servo mechanism in real-time, that moves the camera to keep the tracked face always centered in the image.

1. Introduction

As image sensor technology shifts from CCD to CMOS, cameras will become inexpensive commodity items enabling new applications in the areas of login authentication, video conferencing, and vision assisted user interfaces. To fully utilize low cost cameras, better techniques to detect and track faces need to be created. This paper presents a novel low-cost real-time robust solution to the problem of detecting and tracking multiple faces.

Past work on finding faces in images have mostly used still images [11, 14, 15], but future interactive technology will require real-time face detection and tracking in live video streams with low latency. Most current real-time systems for face detection and tracking are color or model based. Yang and Waibel [19] use color information to track face regions and have done rigorous studies to validate the chosen color space for human face detection [18]. Oliver *et al.*[13] present a 2D real-time single person lip and face tracker that uses color information to detect and track the face candidates. Bradski [2] uses a robust non-parametric statistical technique to track flesh tone regions in 3D, and Toyama [17] integrates color, intensity templates and dark features on the face to estimate the full 3D pose of a single head and uses this information to control the cursor in a computer interface. Birchfield [1] uses the interior color and boundary gradient of an elliptical region to control a camera to follow a single subject as it moves in a room, and La Cascia and Sclaroff [3] have developed a 3D head tracking technique that is robust to varying illumination conditions. In their technique, the head is modeled as a texture mapped cylinder, and the tracking is formulated as an image registration problem in the cylinder's texture map image.

Real-time multiple person tracking is a considerable harder problem and a good example of such a system is from Darrell *et al.*[6]. Range data is extracted using special stereo vision hardware and combined with color information to segment the close skin tone regions. Then a neural network based static face description is used to detect faces. The complete system uses several processors and its cost is a barrier to many applications.

The approach described here is a low-cost real-time multiple person detection system that uses an active lighting technique to find pupil candidates. The current functioning prototype uses a $30 B/W board camera and the complete camera/illuminator system can be build for under $50 using standard off the shelf components. The only other system requirements are a $2K PC with a $150 digitizer card.

The next section introduces the pupil detector technique, and Section 3 describes how the pupils are grouped into faces. A dual mode face tracker used to track the most salient face is presented in Section 4 and experimental results are given in Section 5. Section 6 concludes the paper and discusses future work.

2. Pupil Detection

The multiple face detection system prototype is shown in Figure 1. It uses one low cost black and white cam-

era mounted on a pan-tilt servo mechanism and two light sources. For convenience, near infra-red (IR) light with wavelength 875nm is used. The IR illumination is invisible to the human eye and it makes the system insensitive to changes in indoor ambient illumination, i.e., the room lights can be turned on/off without affecting the system operation.

The camera is a 1/3" CCD board camera with a visible light blocking filter. It is about $40 \times 40 \times 15$mm in size, and the lens is 12mm in diameter, with focal length 5.6mm.

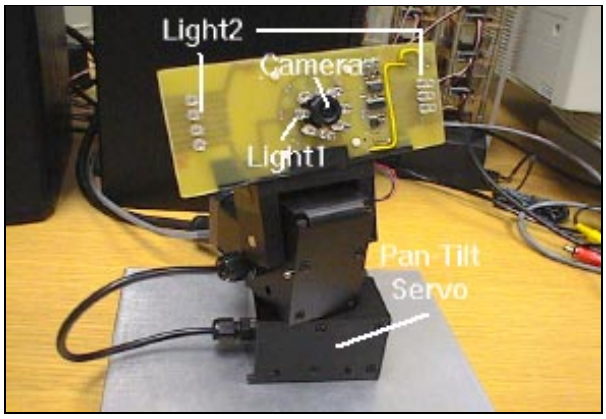


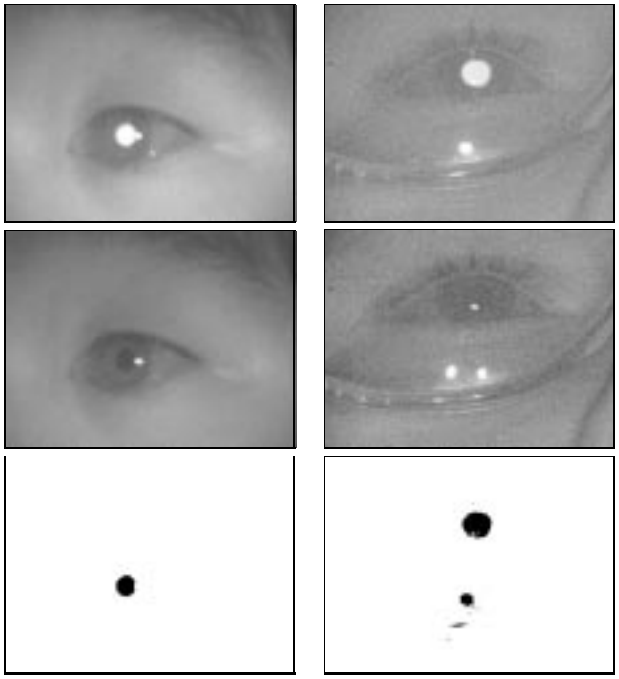
Figure 1. Camera and IR illumination setting.

The light sources LIGHT1 and LIGHT2 in Figure 1 are composed of sets of 7 IR LED's each. LIGHT2 was split into two sets, placed symmetrically in each side of the camera optical axis. Symmetry around the optical axis is desired because it reduces shadow artifacts by producing more uniform illumination, but asymmetrical configurations also perform adequately.

The LIGHT1 set of IR LED's is positioned closer to the camera optical axis and generates bright pupil images, which can be seen in the top row images of Figure 2. This effect is well known as the red-eye effect from flash photographs [21]. Eyes behave as retro-reflectors, reflecting the IR light back exactly along its incoming path, thus the need to place the light sources very close to the camera optical axis. The second set of IR LED's, LIGHT2, is placed farther from the optical axis in order to generate dark pupil images (middle row images of Figure 2), but similar brightness in the rest of the scene.

The retro-reflective property of eyes is uncommon in man-made and natural objects resulting in pupils generally being the only objects appearing with high contrast between the two illumination conditions. Pupil detection is greatly facilitated by the enhanced signal-to-noise ratio, and the simple process of thresholding the difference between bright and dark pupil images is generally sufficient, as shown in the bottom left image of Figure 2. Our experience shows that most retro-reflectors we tested, used for example in running shoes, reflect light in a reasonable wide angle, so that they appear bright in both images and do not cause pupil artifacts. Certain lamps, like table lamps and ceiling lamps, have reflectors that when pointed to the camera can cause such artifacts.

Eyeglasses and contact lens do not change the retro-reflectivity of the eye and unless the glasses/contacts are tinted with an IR blocking coating, they do not inhibit detection. Spurious false pupil candidates, shown in the bottom right image of Figure 2, can be generated by the specular reflections from eyeglasses, and these reflections can also block the dark pupil response under very particular head positions and orientations. In practice, these conditions do not persist for more than a split second, not long enough for the system to loose track of a face.



(a) (b)

Figure 2. The top and middle rows show the bright and dark pupil images respectively. The bottom row shows the difference of the dark from the bright pupil images after thresholding. Column (a) is for a person without glasses, and column (b) for a person with glasses. The reflections on the glasses can create false pupil candidates.

Pupil detection using only the dark or only the bright pupil images, as it is done by most commercial eye-trackers for gaze estimation, is harder to solve because other regions of the image might have similar shapes and appear as dark or as bright as the pupils.

Similar techniques for eye detection using differential

lighting with thresholding are presented in [8, 16], and were used for eye gaze tracking. The technique described in this paper is a much simpler and inexpensive solution for the pupil detection problem. It can be considered as an extension of the work presented in [12]. This novel system does not use any complex specialized hardware, and is able to process wide field of view images in frame rate, allowing it to be used in real-time desktop and kiosk [4] applications which require a detection range of a few meters.

2.1. Implementation Issues

Figure 3 shows a block diagram of the multiple face detection system. The video signal from a NTSC camera is composed of interlaced frames, where frames are composed of even and odd fields taken alternately. When a scene is being imaged, the even field which corresponds to the even lines of the frame is scanned first, followed by the odd field, which constitutes the odd lines of the frame.

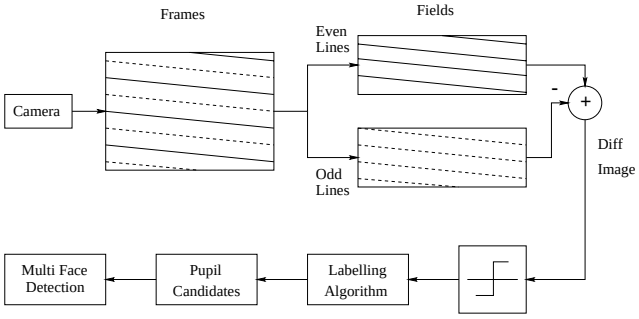


Figure 3. Multiple face detection system diagram.

A very simple circuitry was developed to synchronize LIGHT1 and LIGHT2 with the even and odd fields respectively, i.e., when the even field is being scanned LIGHT1 is on and LIGHT2 is off, and LIGHT2 will be on and LIGHT1 off during the scanning of the odd field. Thus for a frame taken at time t (F_t), the even field (E_t) will contain the bright pupil image, and the odd field (O_t) will contain the dark pupil image.

For the computation of the set of pupil candidates once the fields are decoupled, the dark pupil image is always subtracted from the bright pupil image, so that the pixel values of the difference image ($D_t = E_t - O_t$) at pupil regions are always positive. A thresholding operation is performed on D_t, and the resulting binary image is processed by a connected component labeling algorithm. Note that it is also possible to detect pupils between frames, i.e., using $D'_t = E_t - O_{t-1}$, to obtain twice the frame rate (field rate) detection speed.

Geometrical constraints based on the shape and size of each connected component are then applied to eliminate false positives and obtain the pupil candidates which are the input for the multiple face detection algorithm described in Section 3.

A frame-based pupil detection technique was also developed. In this technique, the light sources have to be synchronized with the frames, so that LIGHT1 is turned on during even frames, and LIGHT2 only during odd frames. If t is even, the difference image would be computed as $D_t = F_t - F_{t-1}$, otherwise $D_t = F_{t-1} - F_t$, for an odd t. This alternative allows the full frame resolution to be used and although could also provide frame rate speed, it was not pursued further because the synchronization was harder to keep, particularly when the system drops frames due to other system constraints. Another advantage of the faster field rate technique is that motion artifacts are considerably reduced.

3. Finding Faces

In the simplest single user scenario, the position of the user's face can be determined once two pupils are detected. The presence of noise and other faces in the image makes the face detection problem more challenging.

Past proposed solutions for detecting faces are based on templates and other geometrical constraints [10, 20] as well as artificial neural networks [14], color histograms [9, 13, 19], or fusion of several modes or cues [1, 5, 7]. The performance of most of these techniques would benefit from a robust pupil finding system by considerably reducing the search space.

Grouping of pupil candidates into faces can be done using simple heuristic rules. These rules can be based on spatial and temporal cues. Temporal cues depict the temporal behavior of eyes. Eyes from the same face are likely to blink at the same time and with the same frequency. They also move rigidly with the head.

Spatial cues are formed by static properties of the eye such as position, size, and aspect ratio, but can also include color of the iris, skin tone surrounding the eyes, etc. Eyes from the same face are likely to appear in horizontal lines and have approximately the same size. Other constraints are imposed based on the expected size of a face, which is estimated from the properties of the camera (size of the sensor and lens) and the illuminators.

Only some of the spatial cues were implemented in the current system prototype for grouping pupils. Since the list of pupil candidates is in general very small, all possible pairs are considered. The first step is to compute the distance between pairs, and eliminate the pairs violating the expected size of a face. The remaining pairs are considered face candidates and sorted according to their inter-ocular

distance. Starting from the smallest to the largest candidate, a new face is accepted if it passes the other spatial constraints (face orientation, and similar pupil size). Once a new face is accepted, the candidates containing these pupils are also eliminated. A simple consistency check can be implemented due to the fact that the imaginary line segment connecting the eyes of a new face cannot cross the inter-eye lines from other detect faces.

4. Face Tracking

To verify the reliability of the multiple face detector, we have developed a correlation-based single face tracker that selects the most salient face (the closest to the camera or the largest one) for tracking. This face is tracked until it is lost. During tracking, the system keeps detecting other faces, but it does not react to them even if they become more salient than the tracked face. This is a desirable feature in desktop and kiosk environments for example, where the focus of attention of the system must remain with one single user during interaction. The tracker also drives a pan-tilt servo mechanism that keeps the tracked face centered in the image.

Once the face tracker is initialized, it relies on two operation modes to continue tracking. One mode uses the information from the multi-face detector, and the second is a feature correlation tracker that uses the sum of absolute differences (SAD) as the object function to be minimized. To ensure stability of this process, two zero order recursive estimators are used to combine the information from both modes, similar to [5]. The state of the tracked face is represented by its size and position, which are treated independently by the two recursive estimators (one for the position and another for the dimensions of the face box). A state of each recursive estimator is defined by a two parameter vector (the position (x,y) of the center of the face, and the width and height (w,h) of the face box). Each vector is accompanied by a covariance.

Movements of the subject's face is unpredictable, but assuming the frame rate is much faster than the rigid head motion, the predicted state vector can be considered to be the last updated estimate

$$\breve{\mathbf{X}} = \hat{\mathbf{X}} \tag{1}$$

and the predicted covariance matrix is

$$\breve{\mathbf{C}} = \hat{\mathbf{C}} + (\Delta t)^2 \mathbf{W} \tag{2}$$

where the uncertainty in position and size grows quadratically with the time interval Δt between the observation and the last estimation, and $\mathbf{W}$ captures the precision loss of each component, and depends on the properties of the underlying process.

New face observations $(\mathbf{Y}, \mathbf{C})$ are used to update the state of the estimators as follows:

$$\hat{\mathbf{C}} = [\,\breve{\mathbf{C}}^{-1} + \mathbf{C}^{-1}\,]^{-1} \tag{3}$$

$$\hat{\mathbf{X}} = \hat{\mathbf{C}}\,[\,\breve{\mathbf{C}}^{-1}\breve{\mathbf{X}} + \mathbf{C}^{-1}\mathbf{Y}\,] \tag{4}$$

The covariance matrix $\hat{\mathbf{C}}$ is an estimation of the error of the estimated state vector $\hat{\mathbf{X}}$. The face detected closest to the estimated position, and within certain error boundaries, is used to update the state.

It is not possible to get observations from the multiple face detector for every frame because of blinking and failures in the grouping process. When no face is detected, or no face closer than a certain threshold in size and position to the predicted face is detected, the SAD correlation tracker is called to determine the 2D translation of the face last used as measurement. The translated face is then used as the new measurement to updated the position estimator.

The SAD correlation tracker determines the translation (i,j) of the feature point $F_{t-1}(x, y)$ to its corresponding tracked point $F_t(x + i, y + j)$ in the current frame F_t by minimizing the SAD(i,j) function within a search neighborhood defined by the region of support around the feature being tracked.

The SAD correlation tracker uses a small region of support and search window around the left pupil, so that it can loose track very easily. To avoid tracking of non-face objects, if the SAD correlation tracker gets called consecutively for more than a certain number of frames, without a face being detected within the predicted region by the multiple face detector, the process is re-initialized with the next most salient face. A surprisingly robust tracker is obtained from the combination of the SAD correlation tracker with the multiple face detector using the recursive estimators, given that neither mode could robustly operate by itself, and one mode has to rely on the other to compensate each others weaknesses.

5. Experimental Results

Figure 4 shows an example result of the multi-face detection algorithm. Observe that the system works reliably even for people with glasses. Detection works very well for distances of up to 3 meters from the camera for most subjects tested (further experiments will have to be conducted to determine the factors which contribute to this variance).

We have noticed a high correlation between age and low intensity bright pupils, i.e., the system sometimes performs poorly with the elderly, while it easily detects younger people. One theory for this is that since pupil size declines with age less light is available for detection from older people. This theory is consistent with the observation that the system works better with large pupils since there are more

(a) (b) (c)

Figure 4. (a) Bright and (b) dark pupil images. (c) Difference and detected faces.

bright pixels. Another theory is that since cataract cloudiness increases with age we see less IR light returned. Neither of these theories explain our observation that race has as much of an effect as age. We have seen darker bright pupil responses from Indian, African and Chinese people.

Figure 5 shows an example from the operation of the multiple people detection with tracking system. Observe that the person closest to the camera has two boxes, one around the face and a second around the left eye, which represents the search window used by the SAD correlation tracker. The tracked face is not centered because the servo was turned off for the subjects convenience, allowing them to observe themselves in the picture. Figure 5a shows the bright pupil image, Figure 5b shows the dark pupil image with the detected faces superimposed, and Figure 5c shows the difference image with the detected face boxes superimposed.

The current prototype was implemented in a dual Pentium II 400MHz machine running Windows NT4, using a commercial frame grabber compatible with Video for Windows. The multiple face detector alone runs at 30 frames per second, grabbing interlaced frames of resolution $640\times480\times8$ bits, and the multiple face detector with the face tracker and the pan-tilt control runs at about 24 frames per second, using the same image resolution. We also noticed that the machine load in this case was about 50%, i.e., one of the processors is dedicated to the face detector and tracker. The performance of the system drops to about 14 frames per second when ported to a single Pentium II 266MHz computer, which might still be fast enough for most applications.

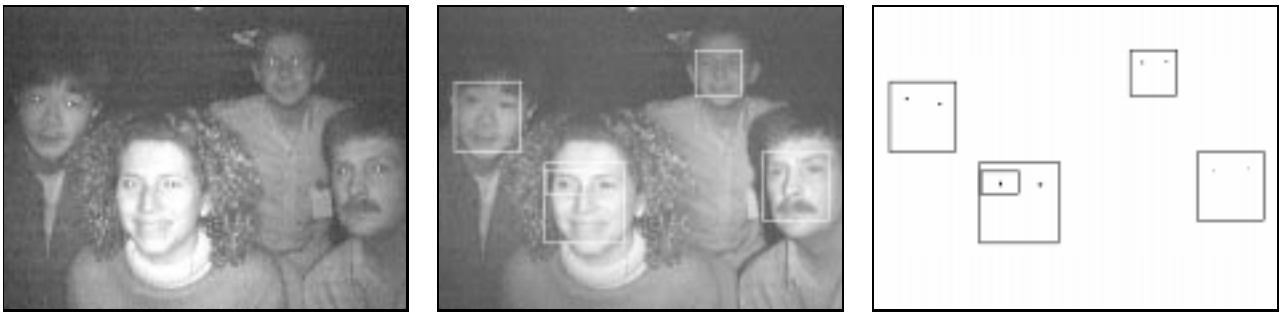
6. Conclusion

We have presented an inexpensive real-time multiple face detection system which was combined with a single face tracking system using recursive estimators. Face detection is accomplished using an active illumination scheme that exploits geometrical and physiological properties of eyes, which is considerably more computationally efficient than model or template based search. The technique is mostly suited to desktop and kiosk applications that require an operation range of a few meters. Once the location of the eye candidates are computed, they are filtered using geometrical constraints and grouped into pairs that correspond to faces using heuristic rules. When multiple faces are detected, the most salient one is used to initialize a dual-mode face tracker. In one of the modes, the tracker uses a face from the multiple face detector, and if that fails the tracker relies upon a SAD correlation technique to estimate the translation of the face. Recursive estimators are used to combine the measurements from both tracking modes, and to control a pan-tilt servo mechanism that moves the camera to keep the face being tracked centered at all times. The prototype system was developed in a dual PII 400MHz platform, and can process 24 frames per second of resolution $640\times480\times8$ bits.

We are currently developing a new face tracker that estimates the full 3D pose of the head, and also extending the pupil and multi-face detector to use pupil and face templates obtained from principal component analysis (PCA) techniques. We expect that much better filtering can be done using the templates, without excessively increasing the computational costs. Further experiments are also being conducted to determine the factors that influence the performance of the detection technique.

Acknowledgments

We would like to thank Arnon Amir, Chris Dryer, Dave Koons, Dragutin Petkovic, Steve Ihde, Wayne Niblack, Wendy Ark, Xiaoming Zhu, Shumin Zhai, and the other people involved in the BlueEyes project for their valuable discussions and contributions during the development of this project.



(a) (b) (c)

Figure 5. (a) Bright and (b) dark pupil images with detected faces. (c) Difference with detected faces. The face with two boxes is the most salient one. The small box surrounds the feature to be tracked.

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] G. Bradski. Computer vision face tracking for use in a perceptual user interface. Technical Report Q2, Intel Corporation, Microcomputer Research Lab, Santa Clara, CA, 1998.

[3] M. L. Cascia and S. Sclaroff. Fast, reliable head tracking under varying illumination. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Fort Collins, CO, June 1999.

[4] A. Christian and B. Avery. Digital smart kiosk project. In *Proc. ACM SIGCHI - Human Factors in Computing Systems Conference*, pages 155–162, Los Angeles, CA, April 1998.

[5] J. Crowley and F. Berard. Multi-modal tracking of faces for video communications. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 640–645, Puerto Rico,PR, June 1997.

[6] T. Darrell, G. Gordon, J. Woodfil, and M. Harville. A virtual mirror interface using real-time robust face tracking. In *Proc. of the 3rd Int. Conf. on Automatic Face and Gesture Recognition*, pages 616–621, Nara, Japan, April 1998.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] C. Morimoto, D. Koons, A. Amir, and M. Flickner. "framerate pupil detector and gaze tracker". In *Proc. of the IEEE ICCV'99 Frame-Rate Workshop*, Kerkyra, Greece, September 1999.

[13] N. Oliver, A. Pentland, and F. Berard. Lafter: Lips and face real time tracker. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 123–129, Puerto Rico,PR, June 1997.

[14] 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.

[15] S. Sirohey. Human face segmentation and identification. Technical Report CAR-TR-695, CfAR - Center for Automation Research, Center for Automation Research, University of Maryland, College Park, MD 20742, November 1993.

[16] 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.

[17] K. Toyama. "look, ma - no hands!" hands-free cursor control with real-time 3d face tracking. In *Proceedings of 1998 Workshop on Perceptual User Interfaces*, pages 49–54, San Francisco, CA, November 1998.

[18] J. Yang, R. Stiefenlhagen, U. Meier, and A. Waibel. Visual tracking for multimodal human computer interaction. In *Proc. ACM SIGCHI - Human Factors in Computing Systems Conference*, pages 140–147, Los Angeles, CA, 1998.

[19] 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.

[20] 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.

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