DETERMINING THE GAZE OF FACES IN IMAGES

A. H. Gee and R. Cipolla

CUED/F-INFENG/TR 174

March 1994

University of Cambridge
Department of Engineering
Trumpington Street
Cambridge CB2 1PZ
England

Email: ahg/cipolla @eng.cam.ac.uk

Determining the Gaze of Faces in Images

Andrew Gee and Roberto Cipolla

University of Cambridge
Department of Engineering
Trumpington Street
Cambridge CB2 1PZ
England

March 1994

Abstract

A person's gaze is a potentially powerful input device for human-computer interaction. Current approaches to gaze tracking tend to be highly intrusive: the subject must either remain perfectly still, or wear cumbersome headgear to maintain a constant separation between the sensor and the eye. This paper describes a more flexible vision-based approach, which can estimate the direction of gaze from a single, monocular view of a face. The technique makes minimal assumptions about the structure of the face, requires very few image measurements, and produces an accurate estimate of the facial orientation, which is relatively insensitive to noise in the image and errors in the underlying assumptions. The computational requirements are insignificant, so with automatic tracking of a few facial features it is possible to produce gaze estimates at video rate.

Keywords: Gaze tracking, human-computer interaction, weak perspective, symmetry, real-time feature tracking.

1 Introduction and outline

Humans have little difficulty sensing where another person is looking, often using this information to redeploy their own visual attention. Even pre-renaissance artists were aware of this, using the gaze of characters *within* a painting to draw the viewer's eye to some significant part of the canvas. Yet this ability to determine a person's gaze, even from a single, monocular, uncalibrated view (as in paintings), is quite remarkable, especially considering the significant inter-subject variations in the facial features that provide the gaze cues.

Current approaches to gaze tracking use active sensing to measure the orientation of the subject's eyes. The eye is illuminated with infrared light, and the gaze direction inferred from the relative position of the *bright-eye* (the reflection off the retina) and the *glint* from the cornea [10]. The system's calibration is sensitive to movements of the subject's head, so the subject must either remain perfectly still, or wear cumbersome headgear to maintain a constant separation between the sensor and the eye. A passive, vision-based approach would ideally tolerate large head movements, and be able to follow a person's gaze at some distance, using little or no calibration, in much the same manner as humans do naturally. There are clearly applications for such a system in human-computer interfaces. In a general sense, a person's gaze can be used to indicate something, in much the same manner as pointing [8]. But gaze has more specific implications than general pointing,

since it additionally encodes where the person is actually looking, suggesting more specific applications, especially in the field of virtual reality.

There are two major components to gaze direction: the orientation of the subject's head, and the orientation of the subject's eyes within their sockets. Here we concentrate on the first component, presenting a simple, efficient method to extract the facial normal from a single, monocular view of a face. This is very different to the approach taken in conventional gaze tracking systems, which work by measuring only the rotation of the eyes. Such systems produce very accurate gaze estimates (errors are typically less than 1 degree [1]) for a subject looking within the narrow field of view allowed by eye movements alone, but cannot cope with the much larger gaze shifts caused by head movements (unless the subject wears cumbersome headgear). By looking *solely* at head movements, we are trading accuracy for flexibility.

For our gaze tracking system to work, it is clearly necessary to make some assumptions about the structure of the human face, though these assumptions should be as simple and generic as possible. Such issues are discussed in Section 2, leading to a compact facial model used throughout the rest of this work. In Section 3 two methods for estimating the facial normal are presented, one exploiting the planar skew-symmetry results of [12], the other using 3-D information provided by the position of the nose in the image. Both methods build on simplified imaging approximations, leading to small, systematic errors in the inferred facial normal. In Section 4 we investigate these errors, and also measure the sensitivity of the two methods to uncertainty in the model parameters and feature locations. It transpires that the planar approach is more accurate for near-profile views of the face, whereas the 3-D approach is superior for near-frontal views of the face. Fortunately, a simple switch, based on a single image measurement, can automatically choose between the two alternative methods, resulting in an accurate and robust composite scheme. Once the facial normal is known, it is relatively straightforward to extract the directions of the other principle facial axes, and estimate the gaze direction with no eyeball rotation: some illustrative results are presented in Section 5. The computational requirements of the scheme are very light, so with automatic tracking of a few facial features it is possible to produce gaze estimates at video rate, as demonstrated in Section 6. Finally, in Section 7, we present our main conclusions. The appendices contain detailed derivations and proofs relating to the material in the main text.

2 The facial model

A little thought suggests many cues to facial orientation: up-down rotation is easily inferred from visibility of the underside of the chin or the crown of the head; left-right rotation can be estimated using ear visibility, or the position of the eyes relative to the occluding contour of the face. Yet all these cues, though undeniably strong, make use of features with a high variation across different subjects: people may have double-chins, affecting the up-down rotation estimation; people may have long hair, rendering the ears invisible from any direction. In addition, these cues are rather vaguely defined, and will be difficult to detect in an image. What is required is a set of precise, geometric cues, easily extracted from an image and providing reliable estimates of facial pose across a wide variety of subjects. For this reason, the methods presented here utilise measurements taken from only the eyes, nose and mouth. It is, of course, necessary to assume some sort of underlying model for the 3-D geometry of faces, though the model should be as simple and generic as possible.

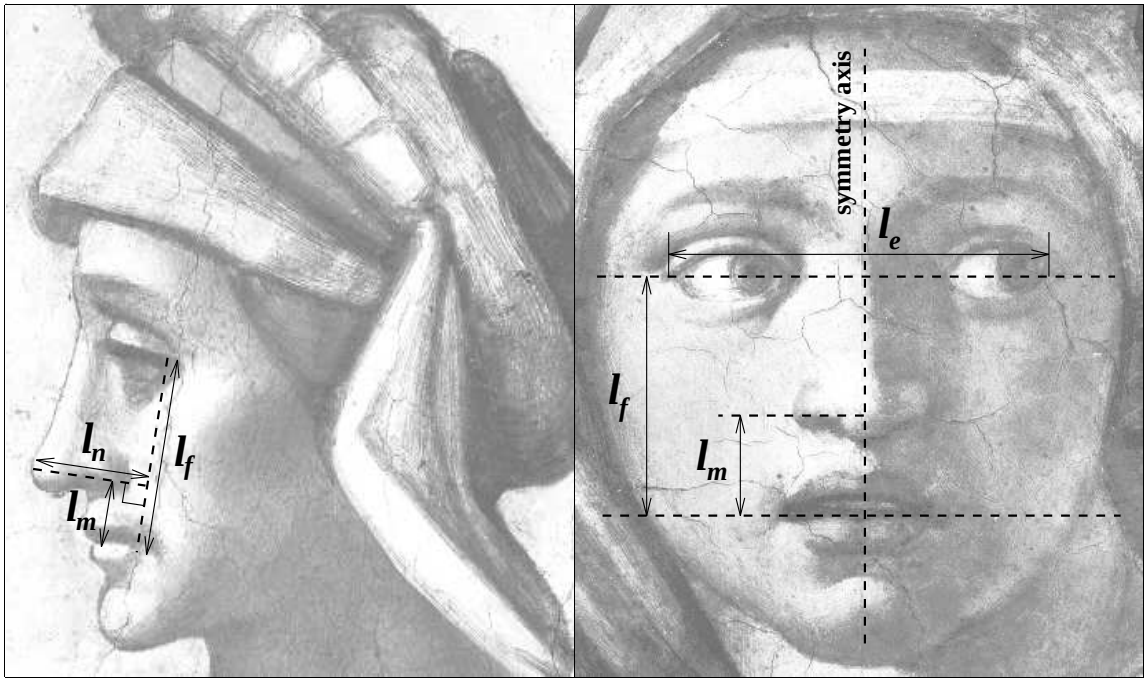


Figure 1: The facial model.

The figure shows frontal and profile views of two faces, the Delphic Sibyl and one of the Ancestors of Christ depicted by Michelangelo on the ceiling of the Sistine Chapel. The facial model used in the 3-D method comprises the ratios $L_f : L_m : L_n$, while the planar method requires only $L_e : L_f$: the corresponding image measurements are denoted using lower case letters.

The facial model is based on the ratios of four world lengths, L_f, L_n, L_e and L_m: the corresponding image quantities (denoted using lower case letters) are shown in Figure 1. The far corners of the eyes and mouth define a plane, which we term the *facial plane*. L_f and L_m are measured on the symmetry axis of this plane, while L_e is measured along the direction of symmetry correspondence. L_n, the distance between the *nose tip* and the *nose base*, is measured along the normal to the facial plane, the *facial normal*. Note that the nose base lies on the symmetry axis of the facial plane. The model is based on the distances between relatively *stable* features: we would not expect the distances to change very much for different facial expressions.

In Section 3 we present two methods for estimating the direction of the facial normal. The first method, which uses 3-D information provided by the position of the nose tip in the image, requires the model ratios $R_m \equiv L_m/L_f$ and $R_n \equiv L_n/L_f$. The second method, which exploits the planar skew-symmetry results of [12], requires only the ratio $R_e \equiv L_e/L_f$. All the model ratios are fairly constant over a range of "normal" faces. Model calibration should only be necessary when dealing with an unusual face, or when high precision is important. The ratios R_n and R_m can be measured from a single profile view of the face, as in Figure 1(left), while R_e is immediately available from a fronto-parallel view of the face, as in Figure 1(right)[1].

3 Estimating the facial normal

Throughout this work a *weak perspective* [14] imaging process is assumed, valid when depth changes on the face are small compared with the distance between the face and the camera. This is generally a good approximation, except when viewing the face from close range with a short focal length lens, which results in significant perspective distortion in the image. The renaissance artists were well aware of this, and consciously avoided short viewing distances, since the resulting images were displeasing to the eye [6, 9]. Indeed, Leonardo's own rule of thumb was to depict figures as viewed from at least ten times the depth change across the figure [6]: the same ratio is often used in more recent vision research to justify a weak perspective imaging assumption [15].

Consider a single image of a face in general pose, as in Figure 2. Assume a camera-centered coordinate system, with x and y axes aligned along the horizontal and vertical directions in the image, and z-axis along the normal to the image plane. Assume also that the far corners of the eyes and mouth, and the tip of the nose, have been located in the image[2]: these points are marked with crosses. Weak perspective preserves length ratios along parallel lines, and particularly midpoints [11]. So it is possible to locate (in the image) the symmetry axis of the facial plane, by finding the midpoints of the eye and mouth points, and joining them up. Furthermore, using the model ratio R_m, the nose base can be located along this line, since length ratios along the symmetry axis are preserved: this point is marked with a blob. Joining the nose base and the nose tip gives immediately the projection of the facial normal in the image, and hence the facial plane's *tilt* direction

[1] In these views the relevant world lengths all lie in a plane parallel to the image plane, so the ratios of the lengths in the image will equal the ratios of the world lengths.

[2] Both techniques presented here for finding the facial normal require visibility of the eye and mouth corners. Although this is a serious problem for near-profile views of the face, these points *are* visible for a wide range of poses, including cases where the face is heavily slanted (see Figures 10 and 11). When the points are obscured, their positions could be estimated using the locations of other facial features.

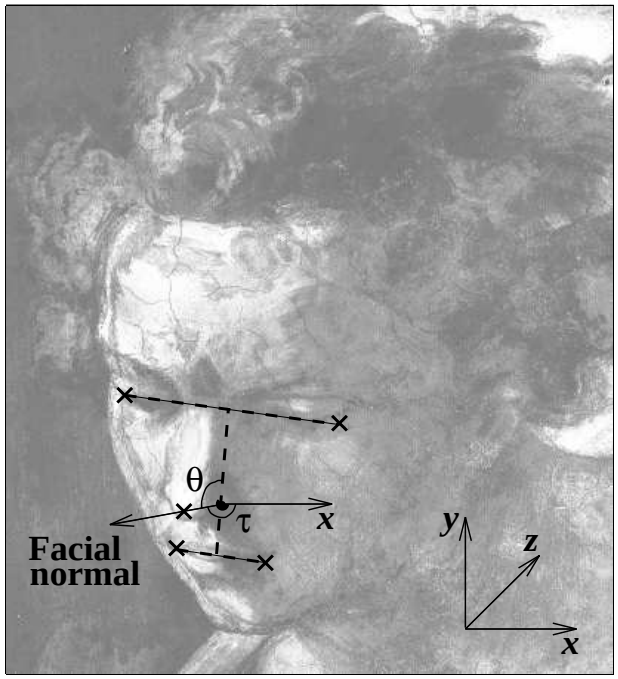


Figure 2: Estimating the facial normal using the 3-D method.

The face belongs to the prophet Daniel, painted by Michelangelo on the roof of the Sistine Chapel. The eye, mouth and nose points are located in the image, and then a simple construction generates the symmetry axis of the facial plane. Using the model ratio R_m, the position of the nose base can be found on the symmetry axis: the image of the facial normal follows easily, producing an immediate estimate of the tilt τ. The slant angle is obtained by simple geometrical analysis, using image measurements θ and $l_n : l_f$, and the model ratio R_n.

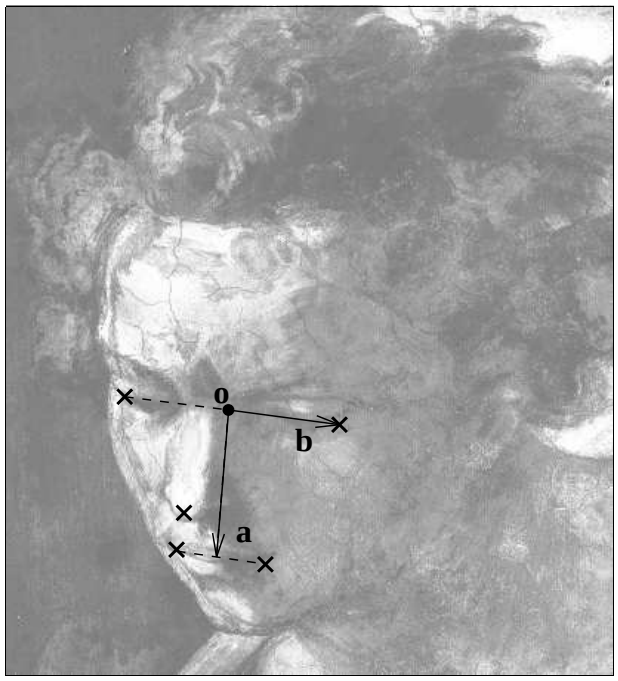


Figure 3: Estimating the facial normal using the planar method.

An alternative method for estimating the facial normal exploits planar skew-symmetry. Once again the eye and mouth points are located in the image, then a simple construction generates the vector **a** *along the symmetry axis of the facial plane, and the vector* **b** *in the direction of symmetry correspondence. The slant and tilt are derived from the* 2×2 *matrix* **U**, *which maps* **a** *and* **b** *onto the points* $(0, -1)$ *and* $(\frac{1}{2}R_e, 0)$ *in an unskewing frame. Using only eye and mouth points a* $\pm 180^o$ *ambiguity remains in the tilt, which can be resolved using a rough estimate of the nose position (or the location of some other facial feature).*

(the distance from the camera to the facial plane increases most rapidly in this direction). The tilt is quantified using the angle τ between the imaged normal and the x-axis.

To fully determine the facial normal it is also necessary to find the *slant* σ, the angle between the optical axis and the facial normal in 3-D space. Then, in camera-centered coordinates, the facial normal $\hat{\mathbf{n}}$ is given by

$$\hat{\mathbf{n}} = [\sin\sigma\cos\tau, \sin\sigma\sin\tau, -\cos\sigma] \tag{1}$$

The slant can be calculated using the model ratio R_n and two measurements in the image: the ratio $l_n : l_f$ and the angle θ. The necessary theory is presented in Appendix A. For Daniel in Figure 2, using model ratios $R_n = 0.6$ and $R_m = 0.4$[3], this gives a slant angle of 27^o, which is in good agreement with human perception.

The aforementioned technique for determining the facial normal uses 3-D information provided by the position of the nose in the image. An alternative technique exploits the skew-symmetry [12] of the facial plane, requiring only the model ratio R_e and the imaged

[3]These values were obtained by rough measurements on one of the authors' faces.

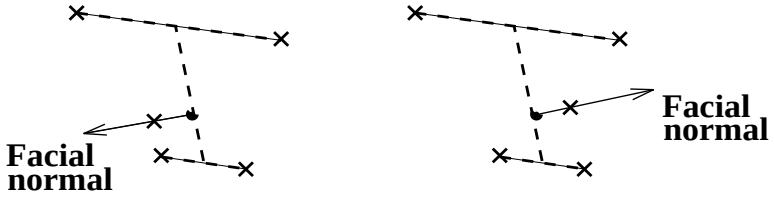


Figure 4: Eye and mouth points cannot resolve a tilt ambiguity.

Under weak perspective imaging, a face in two different poses produces identical eye and mouth points in the image. The $\pm 180^{\circ}$ *ambiguity in tilt can be resolved using a rough estimate of the nose position (or the location of some other facial feature).*

eye and mouth points — see Figure 3. The image vectors **a** and **b**, along the eye-line and symmetry axis, are extracted using the simple construction described for the 3-D method. Assuming an affine imaging process[4], **a** and **b** can be mapped onto world coordinates using a 2×2 plane projection matrix **U**:

$$\begin{aligned} \mathbf{U}\,[\,\mathbf{a}\;\;\mathbf{b}\,] &= \begin{bmatrix} 0 & \frac{1}{2}R_e \\ -1 & 0 \end{bmatrix} \\ \Leftrightarrow \mathbf{U} &= \begin{bmatrix} 0 & \frac{1}{2}R_e \\ -1 & 0 \end{bmatrix} [\,\mathbf{a}\;\;\mathbf{b}\,]^{-1} \end{aligned}$$

Given **U**, the slant and tilt of the facial plane follow (see [12] for the full theory), up to a plus/minus 180° ambiguity in the tilt — see Figure 4. The tilt ambiguity can be resolved using a rough estimate of the nose position, or the location of some other facial feature[5]. Finally, once the slant and tilt are fixed, equation (1) provides the facial normal.

4 Accuracy and sensitivity

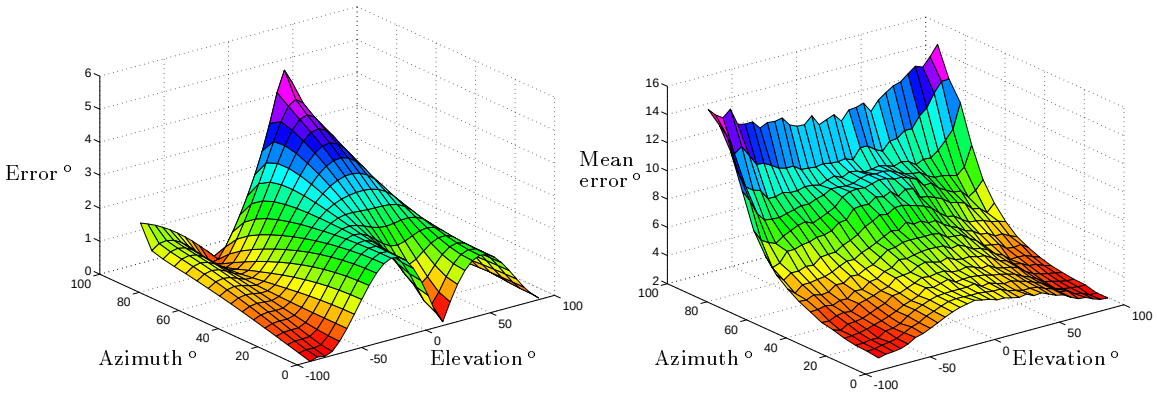
Both the methods presented in the last section build on simplified camera models, leading to small, systematic errors in the inferred facial normal. In this section we investigate these errors, and also measure the sensitivity of the two methods to uncertainty in the model parameters and feature locations. Since both methods are highly nonlinear, an analytical approach would be difficult. Instead an empirical approach is taken, using artificial images of a model face viewed from every possible direction.

The model face comprises five points, the eye, mouth and nose points, positioned so that $R_n = 0.6$, $R_m = 0.4$ and $R_e = 1.0$. The face was placed at a particular azimuth and elevation on an imaginary sphere: by varying the azimuth over the range 0° to 90° and the elevation over the range -90° to $+90^{\circ}$, the full range of possible viewing directions can be covered[6]. For each pose, an artificial image was generated using a full perspective

[4]The weak perspective camera is a special case of the more general affine camera — see [13] for a full discussion of the various imaging models.

[5]In this work, the tilt estimate from the 3-D method was used to choose between the two tilts hypothesised by the planar method.

[6]The bilateral symmetry of the face means that it is not necessary to investigate poses with azimuth -90° to 0°.



(a) Weak perspective error (b) Sensitivity

Figure 5: Performance of the 3-D method.

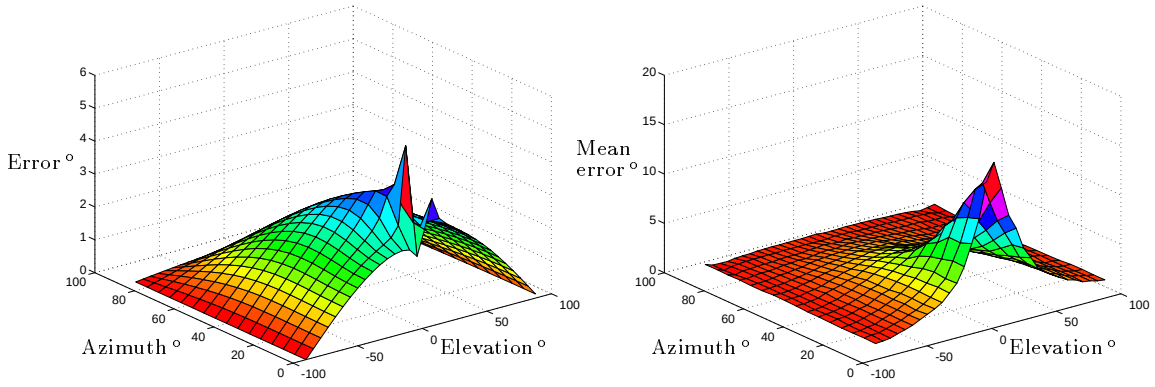
The figure shows the performance of the 3-D method for faces orientated over the full range of possible poses. Plot (a) shows the error (due to the weak perspective assumption) in the calculated facial normal when the face is viewed from a distance of ten times the eye-to-mouth length L_f: greater viewing distances will further reduce this error. Plot (b) shows the method's sensitivity to uncertainty in the model ratios and feature locations. The imaged eye, nose and mouth points were corrupted by zero-mean Gaussian noise with standard deviation 4 pixels[7]. In addition, the assumed model ratios were corrupted by zero-mean Gaussian noise with standard deviation 0.02. The plot shows the mean error observed over 1000 trials with this additive noise.

projection at a viewing distance of $10L_f$ (about 80 cm for a typical face, a fairly close camera position). The two methods were then used to estimate the facial normal, which was compared with the actual normal, allowing the accuracy of the estimate to be assessed. Following this, the imaged feature locations were corrupted by zero-mean Gaussian noise with standard deviation 4 pixels[7]. In addition, the assumed model ratios were corrupted by zero-mean Gaussian noise with standard deviation 0.02. These noise values were chosen to represent typical imaging and model uncertainties[8]. The facial normal was re-estimated from 1000 samples of noisy data, and the mean error recorded over all 1000 trials, giving a measure of the technique's sensitivity to imaging and model uncertainties.

The results for the 3-D method are shown in Figure 5. The technique is fairly accurate, with a maximum error of about 6° for a profile view of a face; moreover, this error would

[7]The imaging conditions were such that the eye-to-mouth length (l_f) measured 200 pixels in a fronto-parallel view of the face.

[8]The experiments reported in this section were repeated with different amounts of additive Gaussian noise. For moderate levels of noise, the expected error in the facial normal was found to vary approximately linearly with the standard deviation of the noise.



(a) Weak perspective error (b) Sensitivity

Figure 6: Performance of the planar method.

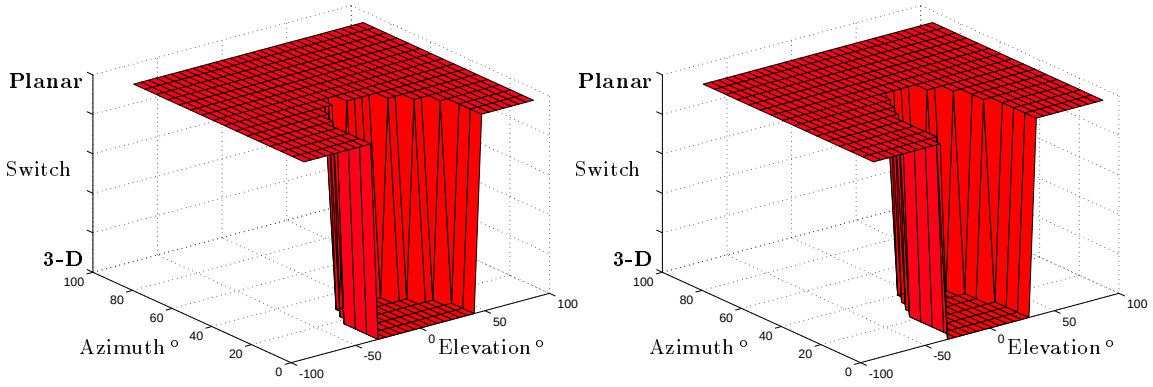
The figure shows the performance of the planar method for faces orientated over the full range of possible poses. See the caption to Figure 5 for a detailed description of the plots. The sharp peak in plot (a) near the full-frontal viewing direction is a result of poor tilt disambiguation.

be reduced for greater viewing distances. However, the sensitivity of the 3-D method is less acceptable, with an expected error of about 15° for profile views in typical noisy conditions. This is, however, to be expected, since the image measurement $l_n : l_f$, used to estimate the slant, is fairly insensitive to changes in pose for near-profile views of the face. Small changes to $l_n : l_f$, caused by small amounts of noise, can have a substantial effect on the estimated facial normal.

The results for the planar method are shown in Figure 6. Here we see similar levels of performance, except that the technique is least satisfactory for near-frontal views of the face, and fairly good for near-profile views of the face. Again, this is to be expected, since the image measurement $l_e : l_f$, implicitly used by the planar method, is fairly insensitive to changes in pose for near-frontal views of the face.

These results suggest that a hybrid approach, using the planar method for near-profile views and the 3-D method for near-frontal views, should deliver much improved performance. An ideal switch between the two techniques, based on the sensitivity plots in Figures 5 and 6, is shown in Figure 7(a). A similar switch can be achieved using a simple threshold on the image measurement $l_n : l_f$: if $l_n / l_f < 0.7 R_n$ use the 3-D method, else use the planar method. The action of this switch for the model face is shown in Figure 7(b).

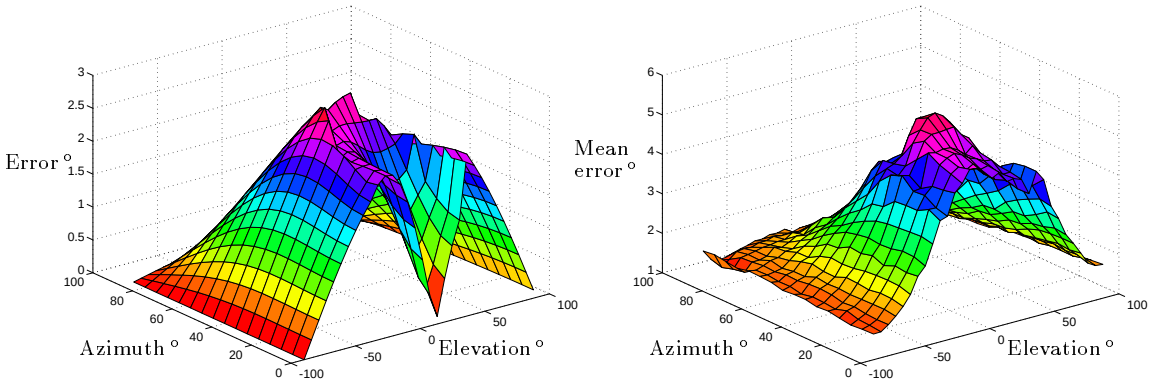
Figure 8 shows the performance of the hybrid method, which is both accurate and robust. For clean data, the technique delivers estimates within 3° of the true facial normal. For typical noisy data, the expected error is less than 6°. This performance is good enough to provide a useful, qualitative estimate of where somebody is looking.



(a) Ideal switch (b) Actual switch

Figure 7: Switching between the alternative methods.

Plot (a) shows an ideal switch between the two alternative methods, based on the sensitivity plots in Figures 5 and 6. Plot (b) shows how a good approximation is achieved using a simple threshold on the image measurement $l_n : l_f$.



(a) Weak perspective error (b) Sensitivity

Figure 8: Performance of the hybrid method.

The figure shows the performance of the hybrid method for faces orientated over the full range of possible poses. See the caption to Figure 5 for a detailed description of the plots.

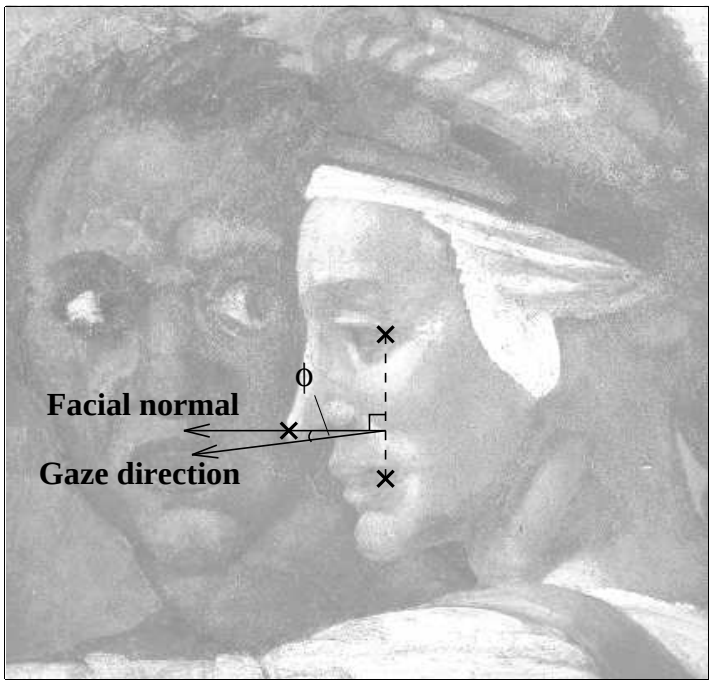


Figure 9: Gaze direction and the facial normal.

This Ancestor of Christ, depicted by Michelangelo on the ceiling of the Sistine Chapel, is looking slightly downwards at the baby she is holding (off the picture). To estimate the gaze direction with no eyeball rotation, the facial normal is rotated by a small angle $\phi \approx 10^o$ around the eye-line.

5 Facial axes and gaze direction

Once the facial normal is known, it is possible to calculate the directions of the eye-line and symmetry axis relative to the camera-centered coordinate system: the necessary geometry is presented in Appendix B. Of course, a person does not necessary look along their facial normal, but, assuming no eyeball rotation, a good estimate of the gaze direction can be found by rotating the facial normal by about 10^o around the eye-line[9] — see Figure 9. For Daniel in Figure 2, using the 3-D method with model ratios $R_n = 0.6$ and $R_m = 0.4$, this gives the following unit vectors for the facial axes and gaze direction, all of which are perfectly credible:

Facial normal	$(-0.40,\ -0.22,\ -0.89)$	**Symmetry axis**	$(0.11,\ 0.95,\ -0.29)$
Eye-line	$(0.92,\ -0.06,\ -0.40)$	**Gaze direction**	$(-0.39,\ -0.39,\ -0.84)$

The estimated eye-line and symmetry axis are not quite orthogonal, reflecting errors in the model ratios and feature locations.

Figure 10 shows approximate gaze directions for a group of figures in Botticelli's *Primavera*. For each figure, the 3-D method was used to find the facial normal, which was

[9]This angle was estimated from typical profile views of faces: see, for example, Figure 9 and the left hand picture in Figure 1.

Figure 10: Estimated gaze directions for several faces.

The figure shows a detail from Botticelli's Primavera*, with eye, nose and mouth positions located by hand. For each figure, the direction of the facial normal was calculated using the 3-D method, and then rotated by* 10^o *around the eye-line to obtain an estimate of the gaze direction with no eyeball rotation. The dark lines show the gaze directions in the image, with lengths proportional to the sine of the slant.*

then rotated by 10^o around the eye-line to estimate the gaze direction. The faint crosses show the eye, nose and mouth positions located (by hand) in the image, while the dark lines show the estimated gaze directions, their lengths proportional to the sine of the slant. The same model ratios used for Daniel in Figure 2, $R_n = 0.6$ and $R_m = 0.4$, were used for all three figures. The results are again credible: the left figure is staring almost directly out of the canvas, while the other figures' faces are slanted at about 50^o and looking towards each other.

6 Real-time gaze tracking

To demonstrate the feasibility of the gaze estimation scheme, a simple real-time eye and mouth tracker was developed and implemented. Following hand initialization[10], eyes were tracked by simply looking for the darkest pixel near the previous eye position: this locks on to the pupil, which is not as stable a feature as the eye corner (since the eyeball is free to rotate relative to the rest of the face), but suffices in the context of this feasibility study. The division between the lips appears as a dark line in the image, whose end points locate

[10]Although hand initialization was used here, techniques do exist for the automatic location of facial features, using either parameterized models [5, 7, 17] or grey-level templates [2, 3, 4].

Figure 11: Real-time gaze tracking.

The figure shows four frames from a typical image sequence. The eye and mouth points were automatically tracked from frame to frame, and the planar method used to obtain an estimate of the facial normal, which is displayed as a drawing pin in the top left hand corner of each frame. The combined tracking and gaze estimation process runs at 100 Hz (four times frame rate) on a Sun SparcStation 10.

the mouth corners. The line and its end points are easily tracked from frame to frame. Given the eye and mouth points, the planar method produces an estimate of the facial normal up to a $\pm 180^{\circ}$ ambiguity in the tilt: this was resolved by hand in the first frame, and subsequently by continuity of the facial normal. The combined tracking and gaze estimation process runs at 100 Hz (four times frame rate) on a Sun SparcStation 10. Several frames taken from a typical image sequence, along with a drawing pin representation of the estimated facial orientation, are shown in Figure 11. Correlation-based algorithms (which can run in real-time on dedicated hardware [16]) could track the nose tip, allowing the 3-D method to be used for near-frontal views of the face.

7 Conclusions

The facial normal, and an estimate of the gaze direction, can be extracted from a single, monocular view of a face, making minimal assumptions about the underlying facial structure. The method for doing this is accurate and fairly robust to uncertainties in the

feature locations and model parameters. It seems unlikely, with current technology, that sufficient resolution is available to measure the eyeball rotation in a typical image. However, a gaze estimate based on the facial orientation alone is useful, and, compared with conventional eye-tracking systems, very large shifts of gaze can be accommodated. Since the computational requirements of the scheme are very light, it is possible to produce gaze estimates at video rate on standard hardware, with many applications in the realm of human-computer interaction.

Acknowledgements

The authors would like to thank Jonathan Lawn and Mark Wright for many helpful discussions relating to this work, and Nick Hollinghurst for his invaluable assistance with the coding of the real-time demonstration. Andrew Gee gratefully acknowledges the financial support of Queens' College, Cambridge, where he is a Research Fellow.

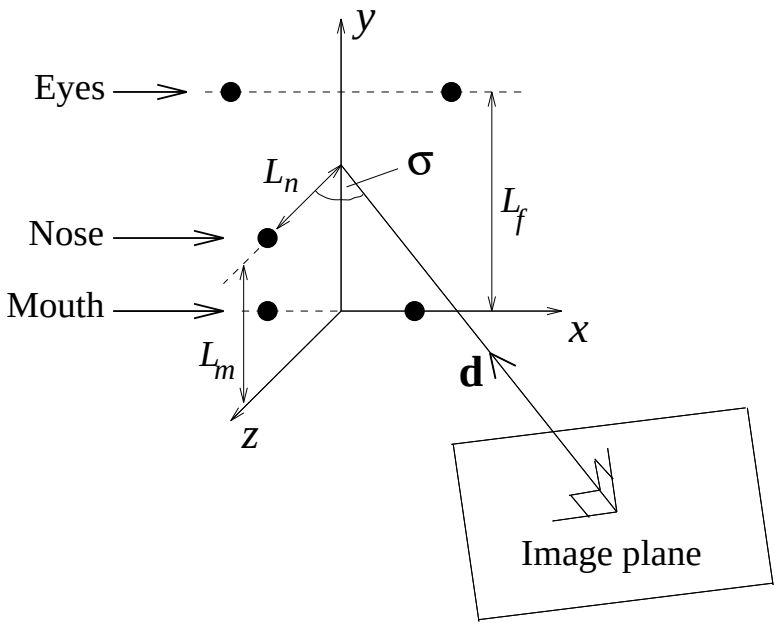


Figure 12: Face-centered coordinate system for slant calculation.

The figure shows the face-centered coordinate system used for the slant calculation. The x-axis is aligned with the eye-line, the y-axis runs along the symmetry axis and the z-axis corresponds to the facial normal. In these coordinates, the normal to the image plane is denoted **d**. *The facial slant, σ, is the angle between* **d** *and the z-axis.*

A Slant calculation in the 3-D method

A face-centered coordinate system used for the slant calculation is shown in Figure 12. In these coordinates, let the unit normal to the image plane be $\hat{\mathbf{d}} = [\hat{d}_x, \hat{d}_y, \hat{d}_z]$. The slant angle of the facial plane is σ, and it is simple to see that $\cos\sigma = |\hat{d}_z|$ (it is necessary to take the absolute value of $\hat{d}_z$, since $\hat{d}_z$ is negative for a face orientated towards the camera). Under orthographic projection, the imaged nose length l_n is given by

$$l_n = L_n \sin\sigma = L_n\sqrt{1 - \hat{d}_z{}^2} \tag{2}$$

Similarly, the imaged facial length is

$$l_f = L_f\sqrt{1 - \hat{d}_y{}^2} \tag{3}$$

So measuring m_1, the squared ratio of the nose to facial lengths in the image, gives one constraint on $\hat{d}_y$ and $\hat{d}_z$, valid for scaled orthographic projection (ie. weak perspective):

$$m_1 \equiv \left(\frac{l_n}{l_f}\right)^2 = R_n{}^2\left(\frac{1 - \hat{d}_z{}^2}{1 - \hat{d}_y{}^2}\right) \tag{4}$$

Another constraint on $\hat{d}_y$ and $\hat{d}_z$ comes from the image angle θ (see Figure 2). The scaled orthographic projection matrix which projects vectors onto the image plane is given

by $c(\mathbf{I}-\hat{\mathbf{d}}\hat{\mathbf{d}}^T)$, where c is an arbitrary constant. Under the action of this projection matrix, the unit vector along the facial normal, $[0,0,1]$, projects to $c[-\hat{d}_x\hat{d}_z, -\hat{d}_y\hat{d}_z, 1-\hat{d}_z{}^2]$. Likewise, the unit vector along the symmetry axis, $[0,1,0]$, projects to $c[-\hat{d}_x\hat{d}_y, 1-\hat{d}_y{}^2, -\hat{d}_y\hat{d}_z]$. Taking the inner product of these two image plane vectors, and using $\hat{d}_x{}^2+\hat{d}_y{}^2+\hat{d}_z{}^2=1$, gives

$$c^2\sqrt{1-\hat{d}_z{}^2}\sqrt{1-\hat{d}_y{}^2}\cos\theta = -c^2\hat{d}_y\hat{d}_z \tag{5}$$

This equation is made more tractable by measuring not θ, but $m_2 \equiv \cos^2\theta$, giving

$$m_2 = \frac{\hat{d}_y{}^2\hat{d}_z{}^2}{\left(1-\hat{d}_y{}^2\right)\left(1-\hat{d}_z{}^2\right)} \tag{6}$$

Eliminating $\hat{d}_y$ between (4) and (6) gives a quadratic equation in $\hat{d}_z{}^2$:

$$R_n{}^2(1-m_2)\hat{d}_z{}^4 + (m_1 - R_n{}^2 + 2m_2R_n{}^2)\hat{d}_z{}^2 - m_2R_n{}^2 = 0 \tag{7}$$

This equation allows recovery of $\hat{d}_z{}^2$ (and hence the slant $\sigma = \cos^{-1}|\hat{d}_z|$) from image measurements m_1 and m_2, and model ratio R_n. It remains only to prove that (7) gives a single root for $\hat{d}_z{}^2$ in the range 0 to 1. It is necessary to consider the cases $m_2 = 1$ and $0 \le m_2 < 1$ separately. If $m_2 = 1$ then (7) becomes

$$\begin{aligned}(m_1+R_n{}^2)\hat{d}_z{}^2 - R_n{}^2 &= 0\\ \Leftrightarrow \hat{d}_z{}^2 &= \frac{R_n{}^2}{m_1+R_n{}^2}\end{aligned}$$

So in this case $0 \le \hat{d}_z{}^2 \le 1$, since $m_1 \ge 0$. We consider now the case $0 \le m_2 < 1$. The discriminant of (7) is $(m_1 - R_n{}^2)^2 + 4m_1m_2R_n{}^2$, which is not negative, since $m_1 \ge 0$ and $0 \le m_2 < 1$: so there are two real solutions for $\hat{d}_z{}^2$. The product of the roots is $m_2/(m_2-1)$, which is zero or negative. Thus either both roots are zero, which gives the desired unique solution for the slant, or one of them is positive and the other negative. For the latter case, it remains only to prove that the positive root is less than 1. This root is given by

$$\hat{d}_z{}^2 = \frac{R_n{}^2 - m_1 - 2m_2R_n{}^2 + \sqrt{\left(m_1 - R_n{}^2\right)^2 + 4m_1m_2R_n{}^2}}{2(1-m_2)R_n{}^2}$$

Since $m_2 < 1$, replacing m_2 with 1 in the square-rooted expression will make the numerator larger (or leave it unchanged if $m_1 = 0$). Doing this, we obtain

$$\begin{aligned}\hat{d}_z{}^2 &\le \frac{R_n{}^2 - m_1 - 2m_2R_n{}^2 + \sqrt{\left(m_1+R_n{}^2\right)^2}}{2(1-m_2)R_n{}^2}\\ \Leftrightarrow \hat{d}_z{}^2 &\le 1\end{aligned}$$

Thus (7) always gives a single root for $\hat{d}_z{}^2$ in the range 0 to 1. The facial slant follows immediately using $\sigma = \cos^{-1}|\hat{d}_z|$.

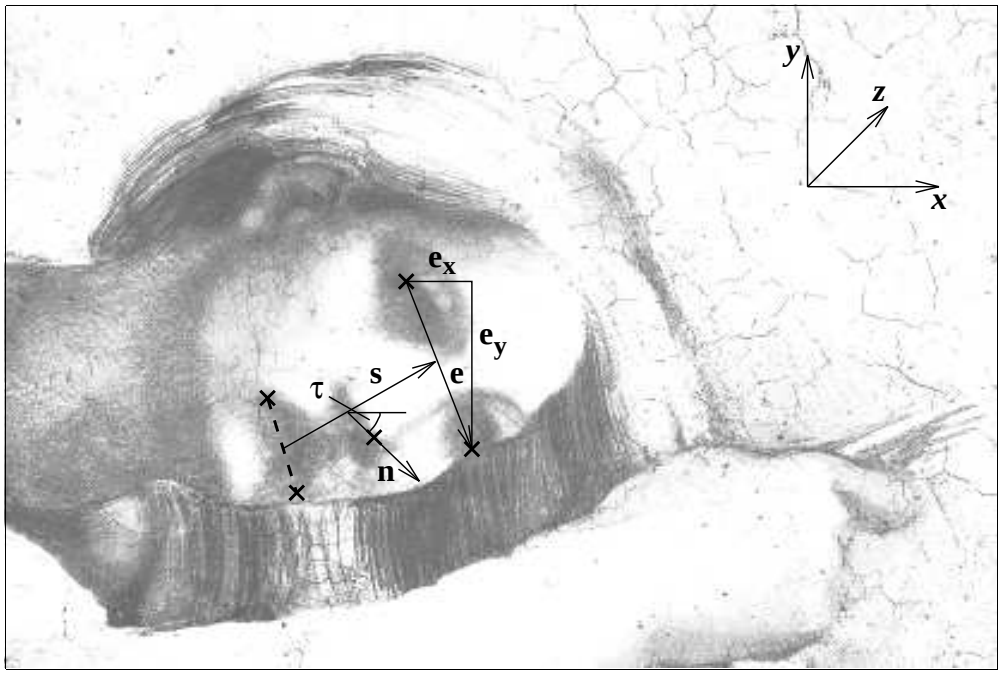


Figure 13: Recovering the facial axes.

The face belongs to one of the Ancestors of Christ, painted by Michelangelo on the ceiling of the Sistine Chapel. Assume a camera-centered coordinate system, with x-axis aligned with the image horizontal, y-axis aligned with the image vertical, and z-axis along the normal to the image plane. If $\hat{\mathbf{n}}$ *is known, it is possible to recover the directions of the two remaining facial axes,* $\hat{\mathbf{e}}$ *and* $\hat{\mathbf{s}}$*, relative to this coordinate system. The tilt angle* τ *is the angle between the imaged facial normal and the x-axis.*

B Recovering the facial axes

Knowledge of the facial normal $\hat{\mathbf{n}}$ allows recovery of the two remaining facial axes, the eye-line $\hat{\mathbf{e}}$ and the symmetry axis $\hat{\mathbf{s}}$, relative to the camera-centered coordinate system. We consider first the direction of the eye-line. The ratio $\hat{e}_y : \hat{e}_x$ is measurable directly in the image — see Figure 13:

$$\frac{\hat{e}_y}{\hat{e}_x} = \frac{e_y}{e_x} = k, \text{ say} \tag{8}$$

This constrains $\hat{\mathbf{e}}$ to be of the form

$$\hat{\mathbf{e}} = \left[\hat{e}_x, k\hat{e}_x, \pm\sqrt{1 - {\hat{e}_x}^2(1+k^2)}\right] \tag{9}$$

A further constraint on $\hat{\mathbf{e}}$ is available, since the eye-line is perpendicular to the facial normal, so $\hat{\mathbf{e}}.\hat{\mathbf{n}} = 0$. After some algebraic manipulation, this gives

$$\hat{e}_x = \frac{\pm\hat{n}_z}{\sqrt{{\hat{n}_x}^2 + k^2{\hat{n}_y}^2 + 2k\hat{n}_x\hat{n}_y + {\hat{n}_z}^2(1+k^2)}} \tag{10}$$

A plus/minus ambiguity in $\hat{\mathbf{e}}$ is acceptable, so we choose the positive solution for $\hat{e}_x$. The other elements of $\hat{\mathbf{e}}$ follow from (9), with the sign of the z-component chosen so that $\hat{\mathbf{e}}.\hat{\mathbf{n}} = 0$. The symmetry axis $\hat{\mathbf{s}}$ can be determined in much the same manner. A check on the mutual orthogonality of $\hat{\mathbf{s}}$ and $\hat{\mathbf{e}}$ reflects the accuracy of the feature locations and model ratios.

References

[1] S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. Technical Report CMU-CS-94-102, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, January 1994. A shortened version appears in *Advances in Neural Information Processing Systems 6*, J. Cowan, G. Tesauro and J. Alspector (editors), Morgan Kaufmann, San Francisco, 1994.

[2] D. J. Beymer. Face recognition under varying pose. Technical Report 1461 (CBLC number 89), Massachusetts Institute of Technology, Artificial Intelligence Laboratory, December 1993.

[3] R. Brunelli and T. Poggio. Face recognition: features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, 1993.

[4] P. J. Burt. Multiresolution techniques for image representation, analysis and 'smart' transmission. In *SPIE Volume 1199, Visual Communications and Image processing IV*, pages 2–15, 1989.

[5] T. F. Cootes, C. J. Taylor, A. Lanitis, D. H. Cooper, and J. Graham. Building and using flexible models incorporating grey-level information. In *Proceedings of the International Conference on Computer Vision*, pages 242–246, Berlin, May 1993.

[6] A. Costall. Beyond linear perspective: A cubist manifesto for visual science. *Image and Vision Computing*, 11(6):334–341, 1993.

[7] I. Craw and P. Cameron. Finding face features. In *Proceedings of the European Conference on Computer Vision*, pages 92–96, 1992.

[8] M. Fukumoto, K. Mase, and Y. Suenaga. Realtime detection of pointing actions for a glove-free interface. In *Proceedings of the IAPR Workshop on Machine Vision Applications*, pages 473–476, 1992.

[9] M. A. Hagen. How to make a visually realistic 3D display. *Computer Graphics*, 25(2):76–81, April 1991.

[10] T. E. Hutchinson, K. P. White, W. N. Martin, K. C. Reichert, and L. Frey. Human-computer interaction using eye-gaze input. *IEEE Transactions on System, Man and Cybernetics*, 19(6):1527–1533, November/December 1989.

[11] J.J. Koenderink and A.J. van Doorn. Affine structure from motion. *Journal of the Optical Society of America*, pages 377–385, 1991.

[12] D. P. Mukherjee, A. Zisserman, and M. Brady. Shape from symmetry — detecting and exploiting symmetry in affine images. Technical Report OUEL 1988/93, Oxford University Department of Engineering Science, June 1993.

[13] J. L. Mundy and A. Zisserman, editors. *Geometric Invariance in Computer Vision*. MIT Press, Cambridge MA, 1992.

[14] L.G. Roberts. Machine perception of three-dimensional solids. In J.T. Tippet, editor, *Optical and Electro-Optical Information Processing*. MIT Press, 1965.

[15] D. W. Thompson and J. L. Mundy. Three-dimensional model matching from an unconstrained viewpoint. In *Proceedings of IEEE Conference on Robotics and Automation*, pages 208–220, April 1987.

[16] W. Yang and J. Gilbert. A real-time face recognition system using custom VLSI hardware. In *IEEE Computer Architectures for Machine Vision Workshop*, December 1993.

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