MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

# A Qualitative Approach to Classifying Head and Eye Pose

P.A. Beardsley

TR98-10    July 1998

**Abstract**

The goal of this work is to classify the focus of attention of a subject who is switching his or her attention between a number of surrounding objects. The specific application is to classify the focus of attention of a car driver as straight-ahead, towards the rear-view mirror, towards the dashboard etc. An explicit quantitative approach to this problem requires (a) a priori information about the interior geometry of the car and the calibration of the camera, and (b) accurate computation of the subjectś location and eye direction. This paper describes a qualitative approach. The subject is observed over an extended period of time, and a p̈ose-space histogramïs used to record the frequency with which particular head poses occur. For observation of a car driver, peaks will appear in the histogram corresponding to the most frequently viewed directions–straight-ahead and toward the mirrors. Each peak is labelled, and the head pose of the driver in all subsequent images is then classified by use of the histogram. The head pose classification is refined by a qualitative measurement of the eye pose.

*IEEE Workshop on Applications of Computer Vision, Oct 1998, pp. 208-213*

# A Qualitative Approach to Classifying Head and Eye Pose

P. A. Beardsley, pab@merl.com
MERL, 201 Broadway,
Cambridge, MA 02139, USA

## Abstract

*The goal of this work is to classify the focus of attention of a subject who is switching his or her attention between a number of surrounding objects. The specific application is to classify the focus of attention of a car driver as straight-ahead, towards the rear-view mirror, towards the dashboard etc.*

*An explicit quantitative approach to this problem requires* **(a)** a priori *information about the interior geometry of the car and the calibration of the camera, and* **(b)** *accurate computation of the subject's location and eye direction. This paper describes a qualitative approach. The subject is observed over an extended period of time, and a "pose-space histogram" is used to record the frequency with which particular head poses occur. For observation of a car driver, peaks will appear in the histogram corresponding to the most frequently viewed directions - straight-ahead and toward the mirrors. Each peak is labelled, and the head pose of the driver in all subsequent images is then classified by use of the histogram. The head pose classification is refined by a qualitative measurement of the eye pose.*

## 1 Introduction

This paper addresses the classification of the focus of attention of a vehicle driver. This is an important component in the development of automatic safety mechanisms [2, 6]. One example would be a system to alert a driver who is looking to one side while other sensors are indicating a potential collision in front of the vehicle. More sophisticated systems might attempt to learn the characteristic activity of a particular driver prior to and during maneuvers, enabling anticipation of those maneuvers in the future.

An explicit quantitative approach to this problem would involve *(a)* modelling the interior geometry of the car, and obtaining the calibration of the camera, and storing this as *a priori* information, and *(b)* making an accurate computation of the driver's location and gaze direction. Generating a 3D ray for the driver's gaze direction in the car coordinate frame then determines what the driver is looking at.

There are problems with this. Firstly, although the geometry of the car's interior will usually be known from the manufacturer's design data, the intrinsic parameters of the camera and extrinsic parameters relative to the car coordinate frame need to be calibrated. That extrinsic calibration might change over time due to vibration. Furthermore, the location of the driver's head and the gaze direction must be computed in the car coordinate frame at run-time. This is difficult to do robustly, and it would prove intensive for the typical low-power processor which might be installed in a car.

Acknowledging these difficulties, we adopt an approach which shifts a significant amount of the processing to initialisation time, and avoids altogether quantitative measurements of head and eye pose. The driver is observed over an extended period. For each acquired image, the driver's head pose is computed and used to update a "pose-space histogram". Peaks in the histogram indicate those head poses which occur most frequently, and can be expected to occur for the driver looking straight-ahead, towards the rear-view and side mirrors, and out of the side window. It is straightforward to label the frequently occurring head poses from a qualitative description of the relative location of windscreen, mirrors etc. The focus of attention of the driver in all subsequent images can then be classified by measuring head pose and checking whether it is close to a labelled peak in the pose-space. [1]

Head pose alone does not of course determine gaze direction. But for our application, the head pose is often a good indicator of the driver's focus of attention. For instance, looking at the side or rear-view mirrors always seems to involve the adoption of a particular head pose. The situation is different when we wish to discriminate between a driver looking straight-ahead or looking towards the dashboard, since this may in-

[1] We use the term "qualitative" for this approach to indicate that there is no computation of absolute angles of the head pose; however, we will make accurate and repeatable measurements related to the head pose.

volve little or no head motion. To deal specifically with the latter case, we are investigating a qualitative classification of eye direction which indicates whether the eyes are directed straight-ahead or downwards.

The algorithms used are appropriate for the Artificial Retina (AR) camera [3], a low-cost (a few tens of dollars) image detector which has programmable on-chip processing. We have a version with a 32x32 detector array which is satisfactory for experiments to classify head pose. However, this is too coarse for eye pose. We did experiments to classify head and eye pose on 128x128 images captured by a camcorder, anticipating the use of these algorithms on a newer version of the AR which has 128x128 resolution.

The next section is an overview of our approach. Section 3 describes the measurement of the driver's head pose, Section 4 describes the construction of a pose-space histogram which encodes the head pose over time, Section 5 describes classification of eye pose, and Section 6 contains experimental results.

## 2 Overview

The head is modelled with an ellipsoid [1, 5]. An ellipsoid is a crude model but is sufficient in the context of the overall system since our aim is not to make accurate quantitative measurements, but to identify frequently adopted head poses, and to classify these poses based on a qualitative description of their relative orientations.

Figure 1 is a diagrammatic overview of the four components of the head pose processing - **(a)** initialise an ellipsoid coincident with the driver's head (Section 3.1), **(b)** use the ellipsoid to generate an array of synthetic views for a range of head motion (Section 3.2); this is done offline as part of the initialisation process, **(c)** search for the synthetic view which best matches a target image of the driver (Section 3.3), **(d)** accumulate head pose information over time, in order to classify subsequently observed head poses (Section 4). Once head pose has been classified, eye pose is processed as described in Section 5.

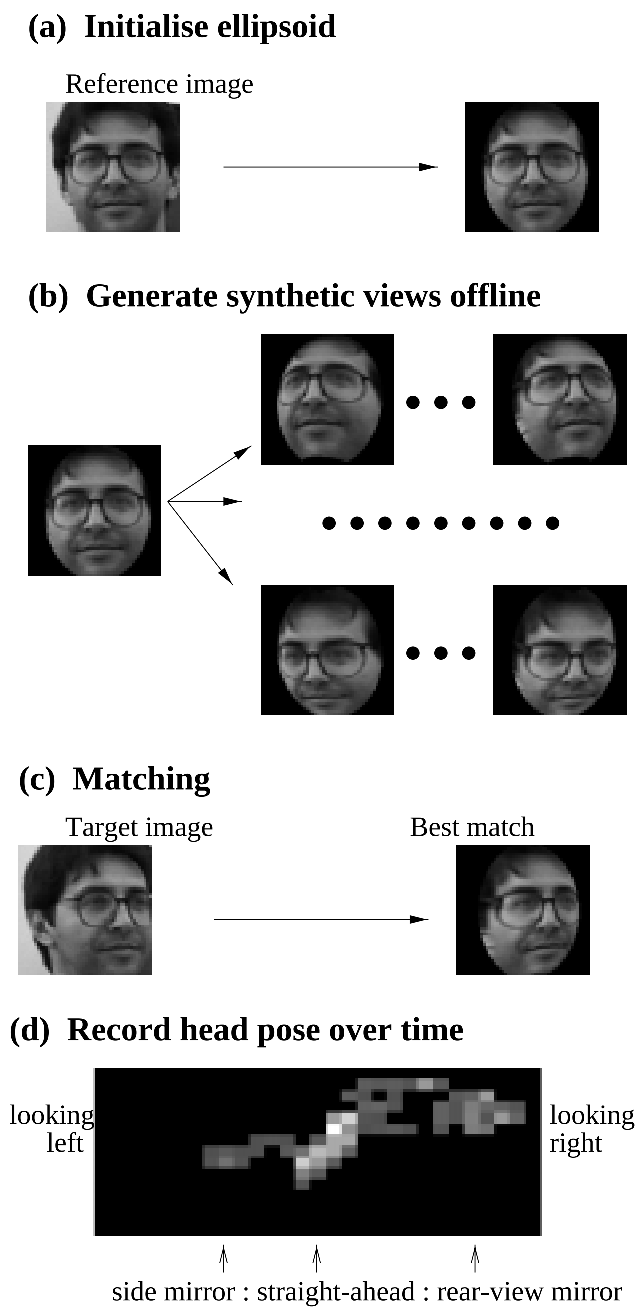


Figure 1: *(a)* a 3D ellipsoid is initialised to coincide with a reference image of the subject's head, *(b)* the reference image provides the basis for generating a series of synthetic views consistent with rotations of the head, *(c)* a target image of the subject is matched against the synthetic views, *(d)* information about head pose is accumulated over time in a pose-space histogram.

## 3 Processing Head Pose

### 3.1 Initialisation

Initialisation involves setting up a 3D coordinate frame containing a camera and ellipsoid, such that they are consistent with a fronto-parallel "reference" image of the subject. See Figure 1a. This process requires some assumptions about approximate camera intrinsic/extrinsic parameters and typical human head sizes as described below, but note that there is no requirement for exact camera calibration or other information here.

A quadric surface can be described by a 4x4 symmetric matrix $\mathtt{Q}$.

$$\mathbf{X}^\top \mathtt{Q} \mathbf{X} = 0 \tag{1}$$

where $\mathbf{X} = (X, Y, Z, 1)^\top$ are homogeneous coordinates for a 3D point. For an ellipsoid in canonical position, matrix $\mathtt{Q}$ is diagonal with diagonal elements $\left[a^{-2}, b^{-2}, c^{-2}, -1\right]$, where the axis lengths of the ellipsoid along the $x-$, $y-$ and $z-$axes are $2a$, $2b$, $2c$ respectively. Assume the $x$-axis is the horizontal axis through the ears, the $y-$axis is aligned with the vertical direction through the head, and a horizontal cross-section through the head is a circle (the ellipsoid is prolate) so that $a = c$.

The initialisation process is manual. We start with reasonable estimates of camera intrinsic and extrinsic parameters, and ellipsoid parameters, for the particular setup which is used to generate the reference image.[2] Based on these assumed parameters, the ellipsoid is projected onto the image plane, and any discrepancy between the projected location and the actual outline of the driver's head is manually corrected by adjusting the extrinsics and the ratio $b/a$.

### 3.2 Generating a Synthetic View

Once the ellipsoid model has been initialised, we can generate a synthetic view of the face consistent with a specified rotation and translation of the head.

For mathematical convenience, we develop the description by an equivalent scenario in which the head is assumed fixed, and the camera is moving relative to it. Generating a synthetic view involves the following conceptual steps - *(a)* the texture from the reference image is backprojected onto the ellipsoid, *(b)* a new location and orientation of the camera is specified, and the texture is reprojected to the image plane of the new camera to generate the required synthetic view. In practice, the image texture is mapped directly from the reference image to the synthetic image.

**(a) Backprojection of texture:**

Assume that for the reference image, the ellipsoid is in the canonical position and the camera has rotation $\mathtt{R}$ and translation $\mathbf{T}=(t_x, t_y, t_z)$ in the world coordinate frame. The perspective projection equation is

[2]For the initial estimate of the camera intrinsics, the focal length (in pixels) is obtained by taking coarse estimates of the subject's distance and head size, and using a similar triangles construction. For the initial estimate of the camera extrinsics, the optical axis is assumed to intersect the origin of the ellipsoid, while the distance of the camera to the subject was typically about 1m. For the initial estimate of the ellipsoid parameters, the typical width of a human head is assumed to be $a = 25$cm, and the ratio $b/a = 1.7$.

$$\mathbf{x} = \mathtt{P}\mathbf{X} \tag{2}$$

where $\mathbf{x} = (x, y, 1)^\top$ are homogeneous coordinates for an image point. For homogeneous quantities '=' indicates equality up to a non-zero scale factor.

In a Euclidean coordinate frame, $\mathtt{P}$ can be decomposed as

$$\mathtt{P} = \mathtt{C}[\mathtt{R}| - \mathtt{R}\mathbf{T}] \tag{3}$$

where $\mathtt{C}$ is the camera matrix [4].

An image pixel $\mathbf{x}$ backprojects to a ray in the world coordinate frame described by

$$\mathbf{X} = \begin{pmatrix} \mathbf{T} \\ 1 \end{pmatrix} + \lambda \begin{pmatrix} \mathtt{D} \\ 0 \end{pmatrix} \tag{4}$$

where

$$\mathbf{D} = (d_x, d_y, d_z) = \mathtt{R}^{-1}\mathtt{C}^{-1}\mathbf{x} \tag{5}$$

Substituting into equation (1) gives a quadratic in $\lambda$,

$$a\lambda^2 + b\lambda + c = 0 \tag{6}$$

where

$$a = d_x^2 Q_{1,1} + d_y^2 Q_{2,2} + d_z^2 Q_{3,3}$$

$$b = 2(d_x t_x Q_{1,1} + d_y t_y Q_{2,2} + d_z t_z Q_{3,3})$$

$$c = t_x^2 Q_{1,1} + t_y^2 Q_{2,2} + t_z^2 Q_{3,3} + Q_{4,4}$$

and $Q_{m,n}$ is the $(m, n)$th element of $\mathtt{Q}$.

The roots of equation (6) give the intersection points of the ray with the ellipsoid (zero, one or two intersection points). For our problem, the ellipsoid is always in front of the camera. When the number of roots is two, their values are positive and the smaller value of $\lambda$ corresponds to the intersection of the ray with the front (visible) side of the ellipsoid.

**(b) Reprojection of texture:**

Assume we want to generate a synthetic image consistent with a transformation $\mathtt{R_h}, \mathbf{T_h}$ of the driver's head. The pixel location $\mathbf{x_r}$ in the reference image which corresponds to a pixel $\mathbf{x_s}$ in the synthetic image is determined by -

**1.** Transform the location of the original (reference image) camera by a rotation and translation ${\mathtt{R_h}}^{-1}, -\mathbf{T_h}$ in the world coordinate frame. Construct the matrix $\mathtt{P}'$ for the transformed camera using equation (3).

**2.** Backproject a ray from pixel $\mathbf{x_s}$ in the transformed camera, and find its intersection point $\mathbf{X_s}$ with the front side of the ellipsoid using equation (6). If there is no intersection point, discontinue. If the intersection point is not on the front (visible) side of the ellipsoid relative to the *reference* image, discontinue.

**3.** Project $\mathbf{X_s}$ onto the reference image - the projected image point is $\mathbf{x_r}$.

**4.** The pixel intensity at $\mathbf{x_s}$ in the synthetic image is then given by the intensity at $\mathbf{x_r}$ in the reference image. Note that $\mathbf{x_r}$ specifies a sub-pixel location in the reference image. To obtain a good resampling, we fit a quadric to the 3x3 patch around the required location in the reference image, and interpolate to obtain the intensity at the sub-pixel coordinates.

The synthetic view generation described above is repeated for a range of rotations around the $x-$ and $y-$ axes (the horizontal axis through the ears, and the vertical axis through the head respectively) to generate an array of synthetic views. This is done offline at initialisation time. Typically we use $\pm 35°$ and $\pm 56°$ around the horizontal axis and vertical axes respectively. We have currently omitted to include cyclorotations of the head because these are relatively uncommon motions - there is in any case some resilience in the processing to cyclorotation. Sample images from the full set are shown in Figure 2.

### 3.3 Matching Against Synthetic Views

Processing a target image of the driver now involves comparing that image with each of the synthetic views to find the best match.

Consider a target image $I$ which is being matched against a synthetic image $S$. The goodness of match $M$ between the two is found by computing

$$M = \sum 1 - \cos(I_d(i,j) - S_d(i,j)) \qquad (7)$$

where $I_d(i,j), S_d(i,j)$ are the directions of the gradient of the image intensity at pixel $(i,j)$ in the target and synthetic images respectively, and the summation is over all significant (i.e. on the ellipsoid) pixels in the synthetic view. The best-matching synthetic view is the one which minimises this score.

While translational motion of the driver's head could also be handled by searching over translations of the ellipsoid in the 3D world coordinate frame, we take a different approach. The target image is matched against a synthetic view for a range of offsets around the default position. Typically the range of offsets is $\pm$ 4 pixels in steps of 2 pixels. This is almost equivalent to searching through the space of translations in 3D space, but offers a more explicit understanding of the coverage of the search space.

Figure 2: An array of synthetic views is generated from the reference image, corresponding to head rotations around the axis through the ears and around the vertical axis. This figure shows a sample of the images - the full series is typically an 11x17 array.

### 3.4 Using Multiple Reference Images

The basic scheme above is extended to make use of three reference images of the subject in the following way. The fronto-parallel reference image is used to generate an array of synthetic views. The subject looks to the left, a left-facing reference image is taken, and the best-match synthetic view is computed. All the synthetic views in the array which correspond to more extreme left-turn rotations than the best-match are now regenerated, using the left-facing reference image. This is repeated on the right side. This provides better quality synthetic views for the more extreme rotations of the head.

The next section describes how the information obtained from matching in Section 3.3 is used to classify what the driver is looking at.

## 4 The Pose-Space Histogram

The algorithm in the previous section will not deliver accurate measurements of head orientation, because we are using an approximate head model. But it does allow identification of those head poses which are being repeatedly adopted, together with the relative orientation of those poses, and that is what we seek to capitalise on.

Corresponding to the 2D array of synthetic views

(Figure 2), a 2D histogram of the same dimensions is set up. All elements in the array are initialised to zero. For each new target image of the driver, once the best-matching synthetic view is found, the corresponding element in the histogram is incremented. Over an extended period, peaks will appear in the histogram for those head poses which are being most frequently adopted.

Ideally, we would expect to find a peak corresponding to the driver looking straight-ahead, a peak to the left of this for viewing the left-side mirror, and a peak to the right for viewing the rear-view mirror (see Figure 1d). Observed peaks can be labelled automatically in accordance with this. Thereafter, for any acquired image of the driver, we find its best-matching synthetic view, use that to index the corresponding location in the histogram, and then classify the target image according to its proximity to a significant area in the histogram. In this way, classification of the driver's focus of attention is achieved without any quantitative information about the 3D layout of the car.

## 5 Processing Eye Pose

The eye pose processing, as with the head pose processing, is not intended to provide an accurate quantitative measurement. In fact, our initial work on processing eye pose has been targeted at one specific task - discriminating whether a car driver is looking straight-forward or at the dashboard, since head pose alone is insufficient for this discrimination in many subjects.

We start with an observation about computing eye pose. Active systems which use reflected infra-red are able to identify the location of the pupil very reliably. This is more difficult in a passive system, particularly when the gaze direction is directed downwards. Note in Figure 3a how directing the gaze downwards results in dropping of the eyelid and the eyelashes, which obscures a clear view of the iris and pupil. For this reason, our measurements will be based on how much the eyelid has dropped, rather than on the position of the iris and pupil. This provides no quantitative measurement of eye direction, but it is still sufficient for a classification of whether the driver is looking straight-forward or at the dashboard.

Figure 3 is an overview of the approach: **(a)** At initialisation time, a series of images of the fronto-parallel face is taken as the gaze direction moves downwards. **(b)** The interior of the eye is segmented out. **(c)** At run-time, the segmented images are used as templates which are matched against the target image.

The matching in **(c)** uses the metric described in Section 3.3, applied only to those template pixels which surround the segmented eye interior. The interior part is not suitable for template matching because the iris can occupy a variety of positions, and also because the iris very often exhibits specularities. We are investigating algorithms which process the interior part to check if that area is consistent in appearance with an iris surrounded by white.

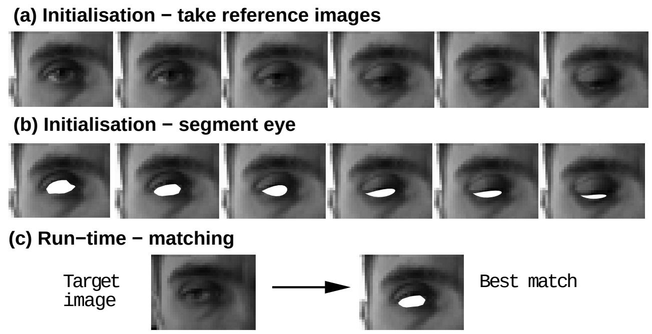


Figure 3: *(a)* a series of reference images of the eye is taken at initialisation time, *(b)* semi-automatic processing is performed to segment out the interior of the eye, resulting in a set of templates which are used for matching at run-time, *(c)* a target image of the eye is processed to find the best-matching template; the iris in the target image has moved, which would degrade the template matching if it were included in the processing.

The set of eye templates in Figure 3b was generated from a fronto-parallel view of the face. To generate a set of eye templates for a different head pose, we use the method in Section 3 to find the reprojection for that head pose, and then apply that reprojection to the eye templates. This is repeated for every head pose in the 2D array of synthetic views described in Section 3.

In overview, the full approach for processing an image sequence of a subject is as follows. The head-pose computation in Section 3 is used to identify the best-matching synthetic view for an image of the subject, and the pose-space histogram is updated. The corresponding set of eye templates is matched to the target image, across a small search area surrounding the default location of the templates, and the best-matching eye template is found. The match is used to update an "eye-state histogram" for that set of eye templates (i.e. for that head pose), which records frequency of match.

As already described, we obtain peaks in the pose-space histogram corresponding to frequently adopted head poses (straight-ahead or towards the mirrors). For the straight-ahead head pose, we expect to be able to divide the the eye-state histogram into sections corresponding to eyes straight-ahead and eyes towards the dashboard. Once this division has been determined, the eye direction can be classified for all subsequent images where the driver has a straight-ahead head pose.

Finally, note that our approach - measuring the amount of drop of the eyelid - can confuse downward gaze direction with simple blinking. To deal with this, we are considering how to incorporate the *duration* of the eye state into the processing, since blinking is transitory, but attentive viewing by the driver has a longer timespan. See Section 7.

# 6 Results

Figure 4: Segments of image sequences (not consecutive images) for different subjects, showing resilience to strong illumination gradients (top), specularities on spectacles (centre), and changing facial expression (bottom).

The system runs on an SGI workstation. For the majority of the experiments, image acquisition was by a Sony Hi-8 video camera with the images subsampled to 32x32 pixels or 128x128 pixels. Other experiments were carried out directly on 32x32 images captured by the Artificial Retina. The processing speed is about 10Hz for 32x32 images on the SGI.

Since the main idea of this system is to avoid explicit measurement of the rotation angles of the head, we will not provide quantitative measurements about head pose, but will illustrate various aspects of the performance of the system.

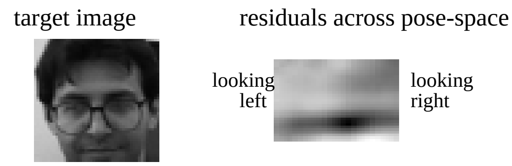


Figure 5: A typical target image together with the error surface generated by matching the image against each image in the array of synthetic views. The darker areas indicate lower residuals (better matching). The error surface is well-behaved, with a clear minimum at the expected location.

As previously discussed, Figures 1 and 2 depict a typical reference image and a selection of synthetically generated views. Synthetic views of the face were generated in the range $\pm 35°$ and $\pm 56°$ around the horizontal and vertical axes respectively, quantized at $7°$ intervals, for a total of 187 synthetic views. This initialisation process takes some tens of seconds.

Figure 4 shows tracking for a number of different subjects. For each image, the best-matching synthetic view has been found, and a 3D head model is illustrated with pose given by the pose angles which were used to generate that synthetic view. This use of the absolute angles is for illustration only, and is not part of the processing.

Figure 5 shows a typical target image together with the error surface generated by matching the target against each image in the array of synthetic views. The error surface is often well-behaved, as shown here. The horizontal elongation of the minimum probably occurs because the dominant features in the matching process are the upper hairline, the eyes, and the mouth - all horizontally aligned features so that horizontal offsets have smaller effect on the matching score in equation (7) than vertical offsets.

Figure 6 shows the result of an experiment in which the subject repeatedly views three different locations over an extended period, with a short pause (about 1s) at each location. In Figure 6a, the three locations correspond to the rear-view mirror, the side-mirror, and straight-ahead for a car driver. The pose-space histogram shows distinctive peaks for each location. In Figure 6b, the three locations correspond to the rear-view mirror, straight-ahead, and the dashboard. The pose-space histogram has a separate peak for the rear-view mirror direction, but the other two directions are

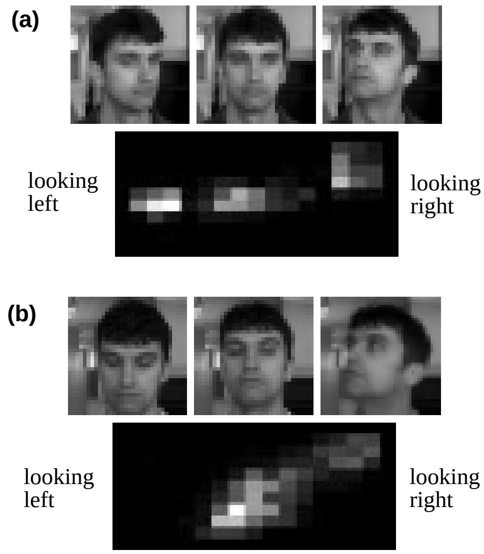


Figure 6: **(a)** The images show the three head poses adopted repeatedly over an extended sequence, together with the pose-space histogram which shows three distinct peaks for those head poses. **(b)** A similar experiment but two of the head poses (forward and forward-down) are not sufficiently far apart to register as distinct peaks in the pose-space histogram.

not differentiable. This is as expected since the array of the synthetic views has a resolution of 7 degrees between images, which is similar to the head motion for these two directions.

Figure 7 shows the pose-space histogram for a short video sequence of a driver in a car. There is a peak for the straight-ahead viewing direction, and lobes to the left and right correspond to the driver looking at the side and rear-view mirrors.

Regarding eye pose, Figure 3 showed an example of matching the eye templates for a fronto-parallel view of the face. Fully automating the segmentation of the eye templates, and matching at other head orientations, is work in progress.

# 7 Future Work

We are considering using a generic head model in place of the ellipsoid to increase sensitivity to changes of head pose. The use of a superior model for the head enables a denser 2D array of synthetic views to be employed, since smaller rotations of the head can be discriminated.

More sophisticated temporal analysis of the images will be an important part of future work - to discriminate roving eye motions from actual fixations (which are characterised by the time the gaze is maintained), and to discriminate blinking from downward fixation.


Figure 7: At top, three sample images from a driving sequence. At bottom, the pose-space histogram for this sequence, showing a peak for the driver looking straight-forward, and side lobes corresponding to viewing the side and rear-view mirrors.

# 8 Conclusion

This paper describes work in progress on a system for classifying the focus of attention of a car driver. Our approach is to address the specific problems which arise in this application (e.g. only incorporating computation of eye pose when head pose alone is insufficient for classification) and to make use of the specific constraints which are available (repeated adoption by the driver of certain head and eye poses), rather than aiming at general purpose algorithms.

We introduced a representation - the pose-space histogram and associated eye-state histograms - for classifying the driver's focus of attention without computing Euclidean information. By avoiding explicit Euclidean measurements, we avoid the need to know *a priori* information about the car's interior geometry, the camera calibration, or the driver's exact location and gaze direction, and thereby hope to achieve a more robust system.

# References

[1] S. Basu, I. Essa, and A. Pentland. Motion regularization for model-based head tracking. In *13th Int'l Conference on Pattern Recognition*, Vienna, 1996.

[2] W. Bounds. Sounds and scents to jolt noisy drivers. page B1, May 3, Wall Street Journal, 1993.

[3] K. Kyuma, E. Lange, J. Ohta, A. Hermanns, B. Banish, and M. Oita. *Nature*, 372:197, 1994.

[4] J. Mundy and A. Zisserman. *Geometric invariance in computer vision.* MIT Press, 1992.

[5] A. Shashua and S. Toelg. The quadric reference surface: theory and applications. *International Journal of Computer Vision*, 23(2):185–198, 1997.

[6] D. Tock and I. Craw. Tracking and measuring drivers' eyes. *Image and Vision Computing*, 14(8):541–547, 1996.