Gaze estimation using appearance models

Fred Nicolls

Department of Electrical Engineering, University of Cape Town
fnicolls@eng.uct.ac.za

Abstract

A simple gaze estimation procedure based on appearance models is described. The method uses minimum norm linear regression to directly relate eye appearance to gaze direction. No explicit feature extraction is required.

The method is fast, requiring only linear operations on the data. It is also accurate, in spite of using only a small amount of training data.

1. Introduction

Appearance models make use of large sets of training samples to characterise the appearance of an object from different viewing directions and distances [2]. The characterisation is obtained by projecting the image of the object being viewed into a dominant linear subspace, and applying statistical reasoning to the projected data. The modelling paradigm is unusual in that no attempt is made to relate the views of the object to its 3-D structure — the object is simply characterised by its appearance from different views.

This paper describes a method for determining the direction of a person's gaze from a front-on image of their face. Images are obtained from a camera mounted on the top of a computer monitor. The computational method used is appearance based, in that the estimation makes use of the variation in the appearance of an eye as the gaze direction changes.

A constrained linear regression formulation is used which entirely eliminates the need for systematic feature extraction. Instead, linear features are learned from the data, and these features are used directly in subsequent gaze direction estimation. It is shown that the constraint on the solution is equivalent to selecting the minimum norm solution to an under-constrained set of linear equations.

The conditions under which the method is developed are highly controlled — users are required to keep their head as stationary as possible during the entire process. Although appearance model formulations can in principle be extended to cases where the head rotates, tilts, or changes distance from the camera, this is beyond the scope of the work.

The structure of the paper is as follows. Section 2 describes in detail the method used for collecting a dataset of face and eye appearances for different gaze directions. The basic method for estimating the gaze is presented in Section 3, and Section 4 provides some results in the form of a performance analysis of the algorithm on the training dataset. Section 5 describes an attempt to incorporate some regularisation into the algorithm training, along with some further results.

2. Dataset generation

A small Matlab application was written which displays a square at one of 36 locations on a 17 inch computer screen. The locations are defined on a 6×6 grid, with increments of 0.2 times the height and width of the screen in the vertical and horizontal directions respectively. The user looks at the square and presses the mouse button, at which point a camera mounted on the top of the monitor captures an image of his or her face and saves it to disk. This process is performed once for each of the 36 grid locations. In an attempt to avoid systematic errors, the locations are presented to the user in a random ordering. The distance from the camera to the user's face is a typical working distance of approximately 0.5m.

An example image of the author's face, looking at the top left corner of the screen, is shown in Figure 2. The image is greyscale, with 576 pixels in the horizontal direction, and 768 pixels vertically. To avoid introducing additional degrees of freedom into the appearance model, in this work the user is required to keep his or her head as static as

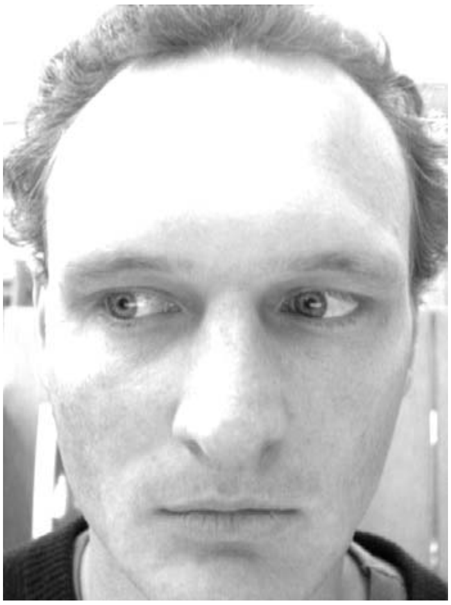


Figure 1: Example of user looking at location (0, 0) (the top left corner) on the screen.

possible during the entire capture process.

The appearance modelling paradigm could in principle operate on the full image as displayed, and the learning stage would find the variations in the images that correlate with the gaze direction. However, for this approach to be successful, a large number of training images would be required to average out spurious effects arising from changing facial expression and head position. To keep the analysis simple, the gaze estimation was therefore performed entirely from the image of the left eye.

To this end, subimages of the left eye were extracted from each of the face images. To ensure normalisation of the data with respect to position, it was decided to centre these subimages on the pupil of the eye. Unfortunately, the task of locating the pupil in an image is not trivial, primarily due to the presence of highlights and reflections on the cornea. In an operational system this problem would have to be overcome. However, since the emphasis of this work lies in the appearance modelling aspects, the locating the locating procedure was simply performed by hand.

Figure 2 shows the extracted image data for each captured gaze direction. Each eye subimage is of dimension 128 × 64, and can therefore be considered to be a point in 8192-dimensional space. For each image the true row and column coordinate of the gaze is known. The combined set of 36 images and their ground truth coordinates therefore constitute the training data for the algorithm.

3. Basic Formulation

Consider firstly the problem of estimating the vertical position (or row) of the gaze. It is assumed that a linear estimator can be used, so

$$r_i = \mathbf{w}_r^T \mathbf{x}_i \tag{1}$$

where r_i is an estimate of the row position of the gaze, and $\mathbf{x}_i$ is a vector description of the eye image, obtained for example by raster reordering. Specification of the estimator involves determining a suitable value for $\mathbf{w}_r$.

The vector $\mathbf{w}_r$ is 8192-dimensional, and 36 training pairs $(\mathbf{x}_i, r_i)$ are available. Forming the data matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_M \end{pmatrix} \tag{2}$$

and the vector of dependent variable observations

$$\mathbf{y}_r = \begin{pmatrix} r_1 \\ \vdots \\ r_M \end{pmatrix}, \tag{3}$$

the condition on $\mathbf{w}_r$ is therefore $\mathbf{X}^T \mathbf{w}_r = \mathbf{y}_r$. Here the elements of $\mathbf{y}_r$ are with respect to the coordinate system described in the previous section, taking values in the range [0, 1] in increments of 0.2. This represents 36 linear equations in 8192 unknowns, so the specification of $\mathbf{w}_r$ is highly underdetermined. Some form of regularisation is therefore needed to arrive at a unique solution.

The observed data vectors $\mathbf{x}_i$ for $i = 1, \ldots, 36$ span a 36-dimensional subspace of the image space. Each vector $\mathbf{x}_i$ therefore lies in the span of $\mathbf{X}$. Since no data is observed outside of this subspace, it is reasonable to require that the position estimate not depend on components of the data outside of this subspace. A suitable constraint on the vector $\mathbf{w}_r$ is therefore that it too lie in the span of $\mathbf{X}$. In this case the required vector can be written in the form $\mathbf{w}_r = \mathbf{X}\boldsymbol{\theta}_r$ for some $\boldsymbol{\theta}_r$.

Under this condition the problem becomes well-posed: the condition that must be satisfied on the training data is

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\theta}_r = \mathbf{y}_r. \tag{4}$$

Here $\boldsymbol{\theta}_r$ is 36-dimensional, so if $\mathbf{X}^T\mathbf{X}$ is full-rank then a single unique solution exists. Solving, $\boldsymbol{\theta}_r =$

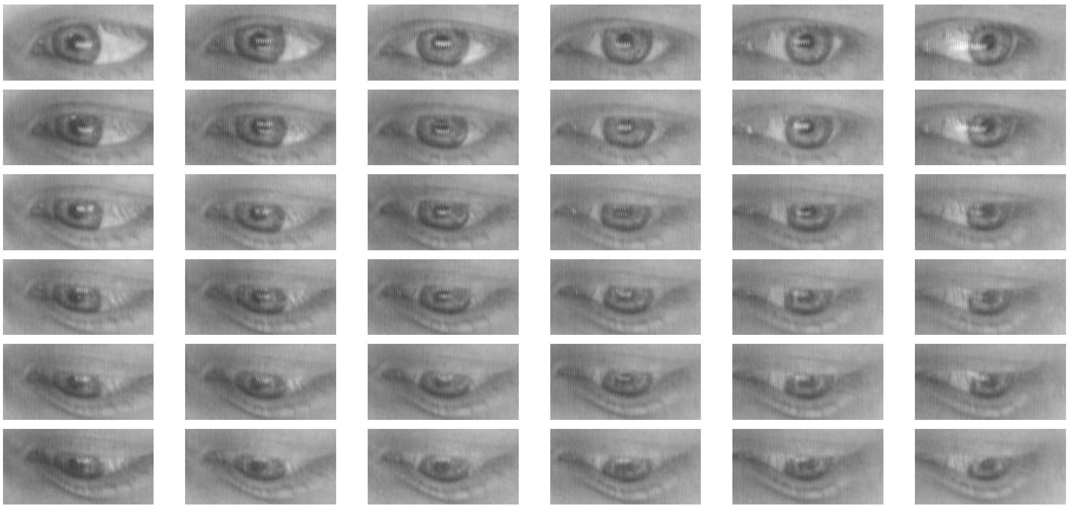


Figure 2: Dataset of images of left eye, centered on the pupil.

$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{y}_r$, so the final solution can be written as

$$\mathbf{w}_r = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{y}_r. \qquad (5)$$

It is observed that this is just the minimum norm solution of the underconstrained linear system $\mathbf{X}^T\mathbf{w}_r = \mathbf{y}_r$ [1, p. 412].

The use of minimum norm solutions is common when dealing with underconstrained systems of equations. However, justification for this particular choice from a multitude of other candidate solutions is seldom explicitly provided. The formulation presented completely justifies the minimum norm solution to the gaze direction problem based on linear subspace principles.

Figure 3(a) shows a representation of the vector $\mathbf{w}_r$ for predicting the vertical component of the gaze from an eye subimage. In order to estimate the gaze from a new image, a subimage of the left eye must be extracted, again centred on the pupil, and the inner product formed between the pixels of the subimage and the corresponding elements of this feature image. The result of this operation is the required estimate for the vertical component of the gaze. Similarly, the feature $\mathbf{w}_c$ for estimating the horizontal component of the gaze is shown in Figure 3(b).

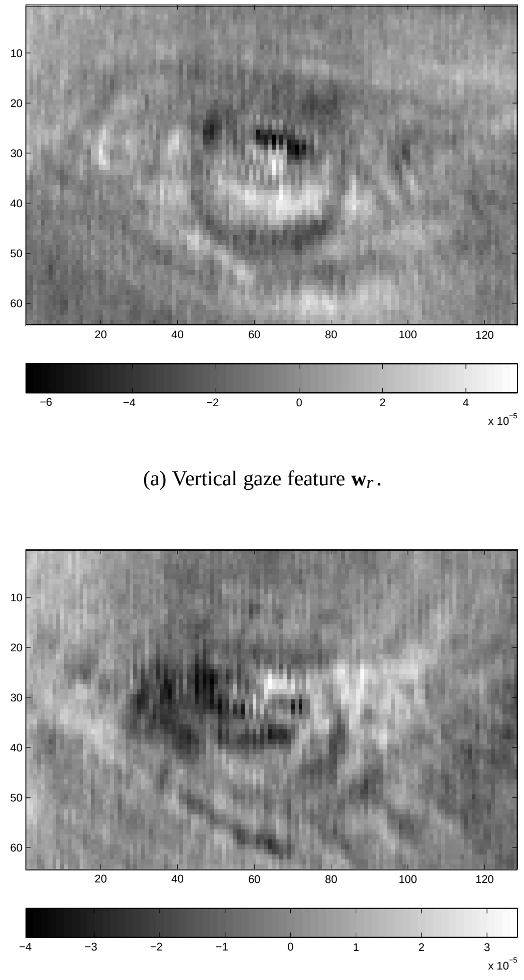
Apart from the process for locating the centre of the eye, the total computational requirement for the gaze estimation procedure is therefore 2×8192 multiply operations, and an equal number of additions.

4. Results

The dataset used in developing the gaze estimation procedure was quite limited, consisting in total of 36 images corresponding to 36 different gaze directions. In order to evaluate the performance of the method, a hold-one-out cross validation procedure was therefore used on the training data. This involves removing one of the images from the dataset, training the gaze classifier on the remaining samples, and testing the resulting system on the held-out sample. Figure 4 shows the vector differences between the actual gaze position (marked by circles) and the estimated gaze position for each of the hold-out cases.

An estimate of the overall error can be obtained by averaging over the errors obtained for each hold-out case. The RMS value of the error magnitude across all estimates for the dataset was 0.0927. However, this error estimate is highly conservative due to the sparseness of the training set — since only a single sample exists for each gaze position, the process of holding out a sample means that information regarding that specific gaze direction is entirely missing. It therefore has to be inferred from neighbouring gaze directions. One would therefore expect quite considerably better performance with the full training set.

It is also apparent from the results that errors tend to be larger near the edges of the image. This is explained by the fact that results in these regions have to be obtained by extrapolation from available data samples. In contrast, within the viewing area an interpolation procedure is implied, which



(a) Vertical gaze feature $\mathbf{w}_r$.

(b) Horizontal gaze feature $\mathbf{w}_c$.

Figure 3: Appearance-based features for linear regression.

is better-posed and more accurate.

It must be noted that the results here have been based on the assumption that the head is entirely stationary during the whole capture process. Thus no attempt has been made to use the observed position of the eye in the image to correct for movement - once the pupil has been located, the extracted subimage of the eye constitutes the only input to the algorithm. Since eye appearance almost certainly relates to viewing direction rather than position, gaze estimates should include a correction component based on the observed eye position. No attempt has been made to include such a correction.

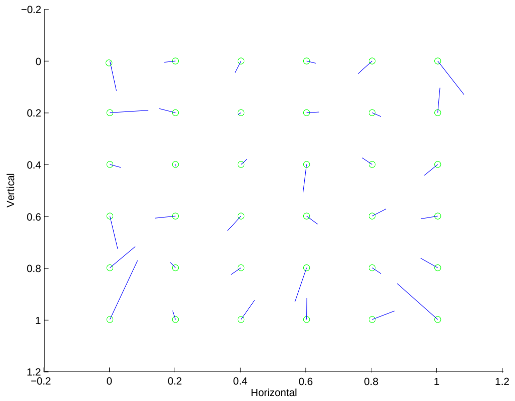


Figure 4: Gaze position estimates using conservative hold-one-out procedure.

5. Attempted regularisation

The proposed solution used a reasonable subspace constraint to reduce a system of 36 equations in 8192 unknowns to one of 36 equations in 36 unknowns. This is sufficient to make the estimate unique, but trying to estimate 36 parameters from 36 unknowns may leave the problem sensitive to noise and outliers. Some means of regularising the solutions space further is therefore desirable.

In light of the earlier discussion, a reasonable option may be to require that the vectors $\mathbf{w}_r$ and $\mathbf{w}_c$ be further constrained to lie in a subspace with dimensionality smaller than 36. The emphasis then lies in deciding on a criterion for the subspace selection procedure.

A very simple selection procedure involves performing principal component analysis on the set of training images, and restricting the solution to the regression to lie in the dominant subspace. In this case a regularisation parameter is introduced, namely the dimension d of the reduced subspace. If the value of d approaches the dimension of the training set, in this case 36, the performance will tend to that of the full-rank case. However, as the value is reduced it is conceivable that some regularity will be imposed, since a smaller number of variables are in effect being estimated. Of course, overly small values of d will lead to poor performance, since the majority of the appearance data is then ignored in the estimation.

Figure 5 shows a plot of the RMS hold-one-out error estimate for varying d. The error remains approximately constant as d is decreased from 36 down to about 10. It does not seem, however,

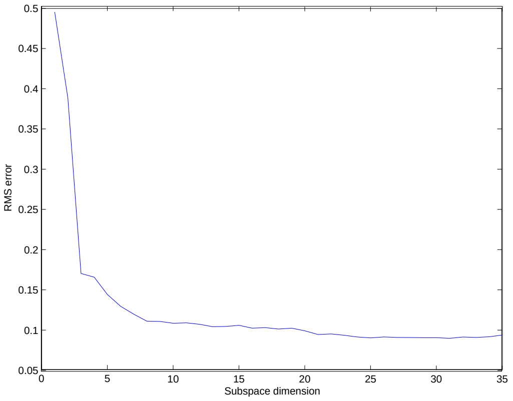


Figure 5: Error estimate versus subspace dimension for PCA-selected basis.

that better generalisation is being obtained from the smaller values. One quite likely reason for the lack of improvement in performance as d is decreased may be because the principal component subspace was already estimated from the data, so the selection process does not yield any regularity. Further analysis is required to confirm or deny this postulate. In any event it seems that the prescribed regularisation method has little merit. Detailed gaze position estimates for the case of $d = 9$ are shown in Figure 6.

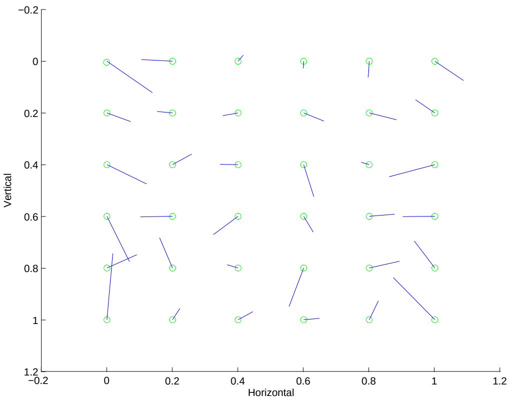


Figure 6: Gaze position estimates for PCA-selected subspace with $d = 9$.

6. Conclusion

A method has been presented for estimating the direction of a person's gaze from the appearance of their eyes. The formulation is admittedly extremely simplistic, with the requirement that the subject keep their head stationary and square to the camera. Nonetheless, surprisingly good results are obtained.

The use of appearance modelling type methods in the formulation seems valuable, particularly when contrasted with more classical methods based on feature extraction and pattern classification. Instead, the method provides a completely automatic means of learning a simple linear transformation on the data which provides the required estimates. It is additionally fast and accurate.

The method can be extended to deal with additional degrees of freedom, such as head rotation, tilt, and distance from the camera. It is quite likely however that considerably more training samples will be required in this case, to ensure that the full range of appearances are accommodated.

7. References

[1] Simon Haykin. *Adaptive Filter Theory*. Prentice Hall Information and System Sciences Series. Prentice-Hall, second edition, 1991.

[2] Hiroshi Murase and Shree K. Nayar. Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.