Deterministic and Stochastic Methods for Gaze Tracking in Real-Time

J. Orozco[†], F.X. Roca[†] and J. Gonzàlez[‡]

[†] *Computer Vision Center & Dept. de Ciències de la Computació, Edifici O, Campus UAB, 08193 Bellaterra, Spain*
[‡] *Institut de Robòtica i Informàtica Industrial (UPC – CSIC), C. Llorens i Artigas 4-6, 08028, Barcelona, Spain*

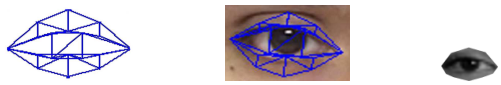
Abstract. Psychological evidence demonstrates how eye gaze analysis is requested for human computer interaction endowed with emotion recognition capabilities. The existing proposals analyse eyelid and iris motion by using colour information and edge detectors, but eye movements are quite fast and difficult for precise and robust tracking. Instead, we propose to reduce the dimensionality of the image-data by using multi-Gaussian modelling and transition estimations by applying partial differences. The tracking system can handle illumination changes, low-image resolution and occlusions while estimating eyelid and iris movements as continuous variables. Therefore, this is an accurate and robust tracking system for eyelids and irises in 3D for standard image quality.

1 Introduction

Eyelid and iris motion description is demanded from human emotion, truth and deception evaluation by combining psychological and pattern recognition techniques. Ekman and Frisen [2] already established that there are perceptible human emotions, which can be early detected by analysing eyelid and iris movements.

Applications on Human Computer Interaction (HCI) demand robustness and accuracy in real-time, which determine the performance evaluation of already proposed techniques. In the literature, there are approaches for gaze analysis dealing with contour detectors, colour segmentation, Hough transform and Optical Flow for eyelids and irises [10,5,9]. These methods are time-consuming and depend on the image quality. On the other hand, restricted detailed textures and templates have been proposed to apply template matching, skin colour detection and image energy minimization [8,6]; for example Moriyama et al [6] deal with three eyelid states namely open, closed and fluttering by constructing detailed templates of skin textures for eyelid, iris and sclera. This approach requires training and texture matching. So, these methods are difficult to be generalized for different image and environment conditions.

We propose in this paper a robust and accurate eyelid and iris tracking by combining stochastic and deterministic approaches with reduced image-data. To provide robustness, we reduce the input image by applying Appearance-Models



(a) (b) (c)

Fig. 1. The 3D mesh **(a)** is projected onto the input image **(b)** to construct the corresponding appearance **(c)**.

[3], which learn the skin texture on-line based on multi-Gaussian assumptions. To provide accuracy, we construct two Appearance-Based Trackers (ABT) to estimate the transition by partial differences with respect to appearance parameters. The first one excludes the sclera and iris information, while achieving fast and accurate eyelid adaptation for any kind of blinking and fluttering motion. The second one is able to track the iris movements, while retrieving the correct adaptation after eyelid occlusions or iris saccade movements. Both trackers agree on the best 3D mesh pose that depends on the head position. The head pose estimation enhances the system capabilities for tracking eyelids and irises in different head position, instead of a frontal face as restriction.

Compared to existing gaze tracking methods, our proposed approach has several advantages. First, our system achieves an efficient eyelid and iris tracking, whose movements are encoded as continuous values. Second, this approach is suitable for real-time applications while handling occlusions, illumination changes, faster saccade and blinking movements. Third, we deal with small images and low resolution, which extends the capabilities of the system for different type of applications.

The paper is organized as follows: section 2 describes the theoretical foundations for appearance-based trackers, the stochastic observation and deterministic transition models. Section 3 presents experimental results and discussion. Finally, section 4 concludes the paper with the main conclusion and future avenues of research.

2 Appearance-Based Tracking

2.1 Image-Data Reduction

The appearance model components are the deformable model and texture. In order to model the eye region, we construct a 3D model of both left and right eyes, which is composed of 36 vertices and 53 triangles, see Fig. 1.(a). This mesh covers the eyeballs, the upper and the lower eyelids, the sclera and the iris. The mesh deformation is determined by the matrix $\mathbf{M}$, which is a n x i matrix:

$$\mathbf{M}_{n,i} = \mathbf{m}_{n,i} + \mathbf{G}_{n,i,k} * \boldsymbol{\gamma}_k \ , \tag{1}$$

where n is the number of vertices and i corresponds to the Cartesian coordinates in the image plane. The matrix $\mathbf{m}_{n,i}$ is determined by the biometry of each person in neutral position. The matrix $\mathbf{G}_{n,i,k}$ deforms the mesh depending on

the eyelid and iris movements. The eyelid and iris movements are controlled by the vector $\boldsymbol{\gamma}_k$ for $k = 0, 1, 2$ for eyelid, iris yaw and iris pitch respectively. These variables are encoded according to the Facial Action Codifying System (FACS), which obey the MPEG-4 codification. Thereby, the 3D mesh $\mathbf{M}_{n,i}$ can be adapted to the eye region, see Fig. 1.(b). The 3D pose of the mesh is taken in to account $\boldsymbol{\rho} = [\theta_x, \theta_y, \theta_z, x, y, s]$, while assuming a weak perspective projection model. Therefore, each 3D point $P_i = (X_i, Y_i, Z_i) \subset \mathbf{M}$ will be projected onto the image point $p_i = (u_i, v_i)$ by using a projection matrix $\mathbf{B}$. That is, $(u_i, v_i) = \mathbf{B}(X_i, Y_i, Z_i, 1)$.

Consequently, given an input frame $\mathbf{F}$ and the parameters to modify the mesh, which are encoded as $\mathbf{q} = [\boldsymbol{\rho}, \boldsymbol{\gamma}]$, we construct an appearance model to represent the eye region [1]. The shape is provided by the 3D mesh and the texture is obtained by applying a warping function $\Psi(\mathbf{F}, \mathbf{q})$ which transfers each pixel from the input frame $\mathbf{F}$ into a reference texture according to the vector $\mathbf{q}$. In this way, the final appearance $\boldsymbol{A}(\mathbf{q}) = [a_0, ..., a_l]$ is obtained, which depends on the mesh configuration $\mathbf{q}$, see Fig. 1.(c).

2.2 Stochastic Appearance Observation

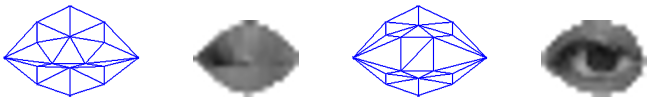
The observation model aims to provide the expected appearances over time, which are based on previous estimations. Therefore, a likelihood distribution is an appropriate method to generate the expected appearances while learning previous estimations. Given an appearance model, $\boldsymbol{A}(\mathbf{q}) = [a_0, ..., a_l]$, which depends on the mesh configuration $\mathbf{q}$, we assume each pixel of the appearance a_i following a Gaussian distribution over time. Thus, we can collect all the variables in a multi-dimensional vector, which can be assumed following a Gaussian distribution as well, $N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are vectors of l values according to the appearance variables a_i, which means that the variance must be computed component by component and not through inner product of the vector $\boldsymbol{\sigma}$. Therefore, the probability for each observation is given by the conditional likelihood function:

$$P(\mathbf{A}_t|\mathbf{q}_{t-1}) = \prod_{i=0}^{l} N(a_i; \mu_i, \sigma_i). \tag{2}$$

The tracking goal is the estimation of the vector $\mathbf{q}$ at each frame t. We represent as $\hat{\mathbf{q}}_t$ and $\hat{\mathbf{A}}_t(\hat{\mathbf{q}}_t)$ the tracked parameters and the estimated appearances [1]. For the sake of clarity, hence we assume that $\mathbf{A}_t(\mathbf{q}_t)$ and $\mathbf{A}_t$ are equivalent and used depending on the specification level. The estimated average appearance is obtained by applying a recursive filtering technique. Thus, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ are updated for the next frame with respect to previous adaptations and a learning coefficient λ:

$$\boldsymbol{\mu}_{t+1} = \lambda \boldsymbol{\mu}_t + (1-\lambda)\hat{\mathbf{A}}_t \quad \textit{and} \quad \boldsymbol{\sigma}^2_{t+1} = \lambda \boldsymbol{\sigma}^2_t + (1-\lambda)(\hat{\mathbf{A}}_t - \boldsymbol{\mu}_t)^2, \tag{3}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are initialized with the first appearance $\mathbf{A}_0$.



(a) **(b)** **(c)** **(d)**

Fig. 2. The 3D mesh **(a)** and the appearance **(b)** for the eyelid tracker. The 3D mesh **(c)** and the appearance **(d)** for the iris tracker.

2.3 Deterministic Appearance Transition

In order to estimate the vector $\mathbf{q}_t$ for the next frame, we adopt an adaptive velocity model, which is predicted by using a deterministic function to obtain the transition state based on the previous prediction, $\hat{\mathbf{q}}_{t-1}$:

$$\hat{\mathbf{q}}_t = \hat{\mathbf{q}}_{t-1} + \Delta\hat{\mathbf{q}}_t \tag{4}$$

where $\Delta\hat{\mathbf{q}}_t$ is the shift of the mesh configuration.

Consequently, for each $\hat{\mathbf{q}}_t$ we construct the corresponding appearance, which is compared with the likelihood average appearance by the Mahalanobis distance. Therefore, given Eq. (4), the appearance becomes $\mathbf{A}_t \approx \boldsymbol{\mu}_t$, which can be approximated via a first-order Taylor series expansion around $\hat{\mathbf{q}}_t$ using the Vanilla gradient descent method [7]:

$$\mathbf{A}_t(\mathbf{q}_t) \approx \Psi(\mathbf{F}_t, \hat{\mathbf{q}}_{t-1}) + \frac{\partial(\hat{\mathbf{A}}_t, \hat{\mathbf{q}}_t)}{\partial\hat{\mathbf{q}}_t}(\hat{\mathbf{q}}_t - \hat{\mathbf{q}}_{t-1}). \tag{5}$$

As a result, the estimation of the vector $\hat{\mathbf{q}}_t$ depends on both the previous adapted and the current average appearance, as well as the minimization distance. The gradient is computed by partial differences with specific descent steps, due to the saccade movements and spontaneous blinking. Thus, tracking is enhanced by quickly retrieving the best adaptation while avoiding drifting problems. Illumination changes, occlusions and faster movements are considered as outliers by constraining with the Huber's function [4] the gradient descent step for each component of the shift vector $\Delta\mathbf{q}$. The Huber's function, $\hat{\xi}$ function is defined as:

$$\eta(y) = \frac{1}{y}\frac{d\hat{\xi}(y)}{dy} = \begin{cases} 1 & \textbf{if} |y| \leq c \\ \frac{c}{|y|} & \textbf{if} |y| > c \end{cases} \tag{6}$$

where y is the value of a pixel in the appearance $\mathbf{A}_t$ normalized by the appearance statistics $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\sigma}}$ according to the Gaussian assumption for the Appearance. The constant c is set $3 * \bar{\boldsymbol{\sigma}}$. Thus, we constrain the appearance registration on the probabilistic model.

2.4 Eyelid and Iris Tracking

Eyelids and irises perform fast movements, which are difficult to predict due to the spontaneous motion and the low resolution on images from monocular

cameras. Therefore, a probabilistic model for them would be quite uncertain. Instead, we propose two models for each tracker, which avoid those pixels that add uncertainty for the appearance predictions, see Fig. 2.

Eyelid tracker uses an appearance model, which excludes the sclera and the iris regions by warping these pixels like eyelid skin. Subsequently, the iris tracker uses an appearance that includes the sclera and iris region. Thus, we construct two Appearance-Based Trackers (ABT), which are combined sequentially, as described next.

Eyelid Tracker $\mathbf{T_W}$: estimates the eyelid position independently from irises, see Fig. 2.(a). The tracking vector is $\mathbf{w} = [\boldsymbol{\rho}, \gamma_0]$, and the appearance model $\mathbf{A}(\mathbf{w})$, which excludes sclera and iris, see Fig. 2.(b):

1. To obtain $\mathbf{A}_t(\mathbf{w}_{t-1})$ by applying the warping function, $\Psi(\mathbf{F}_t, \mathbf{w}_{t-1})$.
2. The Gaussian parameters[1] are estimated, $\boldsymbol{\mu}_t(\mathbf{w})$ and $\boldsymbol{\sigma}_t^2(\mathbf{w})$ by using Eq. (3). We try with five different learning coefficients $\lambda_{\mathbf{W}}$, the high values allow to learn fast movements while the low values keep more information to handle occlusions.
3. To compute the eyelid gradient based on the previous adaptation $\mathbf{w}_{t-1}$ by using Eq. (5), while testing the whole FACS range [-1,1].
4. The best estimation is obtained using Eq. (4), by comparing the average and likelihood appearances through a Mahalanobis distance in an iterative Gauss-Newton process. The search involves exploitation more than exploration to avoid local minima and to estimate the spontaneous movements.

Iris Tracker $\mathbf{T_q}$: estimates yaw and pitch orientations for irises, see Fig. 2.(c). The tracking vector is $\mathbf{q} = [\boldsymbol{\rho}, \gamma_0, \gamma_1, \gamma_2]$ to obtain the appearance $\mathbf{A}(\mathbf{q})$:

1. To obtain $\mathbf{A}_t(\mathbf{q}_{t-1})$, by applying the warping function $\Psi(\mathbf{F}_t, \mathbf{q}_{t-1})$.
2. The Gaussian parameters $\boldsymbol{\mu}_t(\mathbf{q})$ and $\boldsymbol{\sigma}_t^2(\mathbf{q})$ by applying Eq. (3). The learning coefficient $\lambda_{\mathbf{q}}$ is lower than $\lambda_{\mathbf{W}}$ to keep more information from previous frames, since the iris motion is slower than the eyelid blinking.
3. To compute the iris gradient $\mathbf{A}_t(\mathbf{q}_{t-1})$, the iris gradient is estimated for $\mathbf{q} = [\boldsymbol{\rho}, \gamma_0, \gamma_1, \gamma_2]$ in the whole FACS range [-1,1], using Eq. (5).
4. Finally, the best estimation is computed using Eq. (4), by minimizing the Mahalanobis distance. The convergence is achieved by using a shorter exploration than the previous tracker.

Both trackers are connected through the error estimation, which is standardized according to the number of pixels of each appearance. Subsequently, the eyelid tracker provides the vector $\mathbf{w} = [\boldsymbol{\rho}, \gamma_0]$ while the iris tracker estimates the iris while correcting the previous mesh orientation, $\mathbf{q} = [\boldsymbol{\rho}, \gamma_0, \gamma_1, \gamma_2]$, see Fig. 3.

[1] Let be $\boldsymbol{\mu}(\mathbf{w})$ and $\boldsymbol{\sigma}^2(\mathbf{w})$ the Gaussian parameters corresponding to the eyelid tracker $\mathbf{T_W}$. Similarly, $\boldsymbol{\mu}(\mathbf{q})$ and $\boldsymbol{\sigma}^2(\mathbf{q})$ are the Gaussian parameters for the iris tracker $\mathbf{T_q}$.

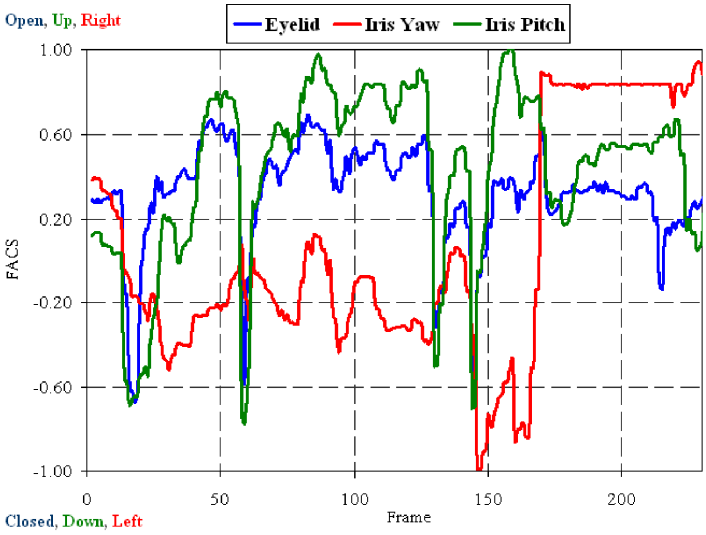


Fig. 3. The eyelid tracking handle movements and blinks as continuous values. Iris tracking for spontaneous movements and saccades. The eyelid position is corrected for the pitch variation effects.

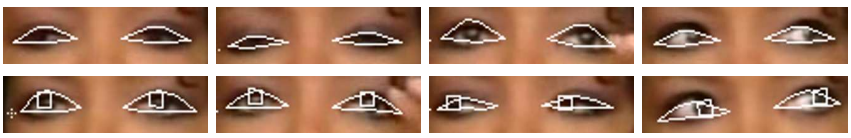


Fig. 4. Eyelid and Iris trackers are applied sequentially while learning each appearance texture on-line, which enhances the robustness and accuracy.

3 Experimental Results

Experiments were run in a 3.2 GHz Pentium PC, in ANSI C code. Three image sequences of 250 frames were used for testing both trackers, which correspond originally to facial image sequences with the cropped eye region. They were recorded with monocular and photographic cameras without illumination conditions. We do not use high image resolution, because the reference texture size is 14x18 pixels.

We computed the eyelid estimation by using $\mathbf{T_W}$ while dealing with any kind of blinking, e.g. open, closed and fluttering. In addition, the iris estimation is done by using $\mathbf{T_q}$, which estimates yaw and pitch motion, saccade movements and eyelid occlusions.

On one hand, the eyelid tracker adapts the eyelid position for slow movements and blinks while estimating the mesh orientation, see Fig. 4. This tracker does

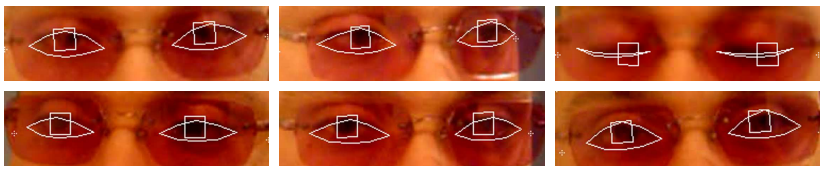


Fig. 5. Due to tracking capability to handle illumination changes and occlusions, the tracker has a good performance with eyes wearing glasses.

Fig. 6. The eyelid and iris tracking deals with images of small and low-resolution. These images are 64x20 pixels where each input eye is 18x12 pixels.

not depend on iris estimations even when the pitch movement affects the eyelid position. However, this tracker warps the sclera and iris pixels as eyelid skin, otherwise, those pixels are considered outliers when the eyelid is occluding the inner eye region. On the other hand, the iris tracker adapts well the iris position while dealing with slow yaw and pitch motion while retrieving the correct adaptation of the mesh after either saccade motion or iris movements while the eyelid occludes the iris. Both trackers use the FACS codification, which is expressed as continuous values between -1.0 and 1.0, see Figs. 3. The eyelid and iris pitch have similar plots because they commonly perform a synchronized and spontaneous movement. This is a correlation that is independently well estimated.

For an appearance model size of 14x18 pixels, we obtained a performance of 21 fps, and a 96% of correct adaptations. The error is principally due to the saccade motion or eyelid occlusions. Each sequence was also tested using an appearance size of 5x11 pixels, thus obtaining 52 fps for the iris tracker and 67 fps for the eyelid tracker, and average accuracy adaptation of 85%. Iris tracking is a challenge for small reference texture, where the iris size is 2x3 instead of 5x6 for the big resolution.

Learning the texture on-line and handling illumination changes are important capabilities to obtain good adaptations by analysing row-image resolution, and occluded eye regions. The Fig. 5 shows a 400 image sequence, where the subject is wearing sunglasses, the input eye region is 42x82 pixels. The tracking got a 91% of correct adaptations with a reference texture of 14x18 pixels. We obtain 82% of correct adaptation analysing small images with low resolution, where the input eye region is 10x18, see Fig. 6. The whole input image is 112x160 that is appropriate for video conference software.

4 Conclusions

Three main contributions were proven in this work. First, the information and dimensionality reduction for the input image by constructing appearance models, which are appropriate for statistical modelling. Second, the stochastic observation model provides an accurate likelihood function, which is conditioned by previous estimations and accumulative appearance information. Third, the deterministic transition model allows generating an appearance space by applying first-order Taylor approximations around to the previous adaptation.

The experimental results have proven that combining sequentially two ABT, the system is able to estimate accurately the eyelid and iris position in 3D images. The system does not require high quality images or specific illumination conditions, since we do not use colour information, edge detectors, or motion extraction algorithms. We have demonstrated in this framework that eyelid and iris motion is reliable for HCI applications and psychological systems, which demand real-time performance with accurate results. Our system provides a robust and accurate gaze description, able to handle occlusions, illumination changes, small and low-resolution images in 3D.

Acknowledgement: This work is supported by EC grants IST-027110 for the HERMES project, IST-045547 for the VIDI-Video project, by the Spanish MEC under projects TIN2006-14606 and DPI-2004-5414. Jordi Gonzàlez also acknowledges the support of a Juan de la Cierva Postdoctoral fellowship from the Spanish MEC.

References

1. T.F. Cootes and C.J.Taylor. *Statistical Models of Appearance for Computer Vision.* Imaging Science and Biomedical Engineering, University of Manchester, 2004.
2. P. Ekman and V. Friesen. Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press, Palo Alto*, 1978.
3. T. F. Cootes G. J. Edwards and C. J. Taylor. Face recognition using active appearance models. *In Proceedings of the Fifth European Conference on Computer Vision (ECCV)*, 2:581–695, 1998.
4. P.J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35:73 – 101, 1964.
5. H. Liu, Y. Wu, and H. Zha. Eye states detection from color facial image sequence. *Proc. 2nd Int. Conf. Image and Graphics (ICIG02)*, 4875:693–698, 2002.
6. T. Moriyama, J. Xiao, J. Cohn, and T. Kanade. Meticulously detailed eye model and its application to analysis of facial image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):738 – 752, May 2006.
7. J. Nocedal and S. Wright. Numerical optimization. In *Springer, New York*, 1999.
8. A. Basu S. Bernogger, L. Yin and A. Pinz. Eye tracking and animation for mpeg-4 coding. *IEEE Pattern Recognition, 1998. Proceedings, Fourteenth International Conference*, 2:1281–1284, 1998.
9. H. Tan and Y. Zhang. Detecting eye blink states by tracking iris and eyelids. *Pattern Recognition Letters*, 2005, 2005.
10. J.F. Cohn Y. Tian, T. Kanade. Dual-state parametric eye tracking. *In: International Conference on Automatic Face and Gesture Recognition*, 2000.