Machine Vision and Applications (2009) 20:353–364
DOI 10.1007/s00138-008-0130-6

ORIGINAL PAPER

Real-time gaze tracking with appearance-based models

Javier Orozco · F. Xavier Roca · Jordi Gonzàlez

Received: 27 July 2007 / Accepted: 22 December 2007 / Published online: 4 April 2008

Abstract Psychological evidence has emphasized the importance of eye gaze analysis in human computer interaction and emotion interpretation. To this end, current image analysis algorithms take into consideration eye-lid and iris motion detection using colour information and edge detectors. However, eye movement is fast and and hence difficult to use to obtain a precise and robust tracking. Instead, our method proposed to describe eyelid and iris movements as continuous variables using appearance-based tracking. This approach combines the strengths of adaptive appearance models, optimization methods and backtracking techniques. Thus, in the proposed method textures are learned on-line from near frontal images and illumination changes, occlusions and fast movements are managed. The method achieves real-time performance by combining two appearance-based trackers to a backtracking algorithm for eyelid estimation and another for iris estimation. These contributions represent a significant advance towards a reliable gaze motion description for HCI and expression analysis, where the strength of complementary methodologies are combined to avoid using high quality images, colour information, texture training, camera settings and other time-consuming processes.

Keywords Eyelid and iris tracking · Appearance models · Blinking · Iris saccade · Real-time gaze tracking

J. Orozco (✉) · F. X. Roca
Computer Vision Center & Dept. de Ciències de la Computaciò,
Edifici O, Campus UAB, 08193 Bellaterra, Spain
e-mail: jorozco@cvc.uab.es

J. Gonzàlez
Institut de Robòtica i Informàtica Industrial (UPC–CSIC),
C. Llorens i Artigas 4-6, 08028 Barcelona, Spain

1 Introduction

Gaze information is important for the psychological analysis of deceit, truth detection and emotion evaluation [4]. Ekman and Frisen [10] have already shown that there are perceptible human emotions, which can be detected early by analysing eyelid and iris motion, see Fig. 1.

Human Computer Interaction (HCI) applications require gaze analysis in real-time. Existing techniques are evaluated according to robustness and accuracy [8]. On one hand, gaze tracking is approached as an eyelid and iris detection problem by applying edge detectors, Hough transform, optical flow and thresholding techniques [17,19]. These methods are time-consuming and depend on both the image quality and acquisition, for example by IR cameras [20]. On the other hand, restricted detailed textures and templates have been proposed for template matching, skin colour detection and image energy minimization [1,15]. Moriyama et al. [15], for example, use three eyelid states: open, closed and fluttering. They have created detailed templates of skin textures for the eyelid, the iris and the sclera. Their approach needs training and texture matching.

Tan and Zhang [18] applied segmentation and colour space transformations for iris tracking using a valley-peak field approach to obtain a binary image for the iris region. However, the accuracy of the results is strongly affected by illumination and skin colour. Therefore, it is difficult to use these methods into different images and environment conditions.

By constructing a low-dimensional representation of non-rigid objects, the Appearance-Based Models (ABM) provide an accurate statistical analysis of complex shapes [9]. ABMs are commonly used for face tracking because of their robustness to handle changes in imaging conditions and different skin colours. However, they have not been used for eyelid and iris tracking yet.

Fig. 1 Psychological studies interpret eyes movements as early and spontaneous facial expressions [10]

In this paper, we propose a gaze tracking method, which combines two Appearance-Based Trackers (ABT), one for eyelid and another one for iris tracking. The first one excludes the sclera and iris information, while achieving fast and accurate eyelid adaptation for blinking and any fluttering motion. The second one is able to track iris movements and recover the correct adaptation, even in cases of eyelid occlusions and iris saccade movements. Both trackers agree with the best 3D mesh pose that depends on the head position. The head pose estimation enhances the system capabilities for eyelid and iris tracking in different head positions.

We model appearance-textures as a multivariate normal distribution. The Gaussian parameters are estimated by a recursive filtering technique. This is an on-line learning process of any facial texture. Once the expected appearance is calculated, we estimate the facial actions by applying gradient descent methods and backtracking. Thus, the algorithm converges faster to the best adaptation.

Our method has several advantages over existing methods. Firstly, we can handle occlusions, illumination changes and faster saccade and blinking movements, while it is suitable for real-time applications. Existing methods, which are predominantly created for medical image analysis, require specific image quality, training and camera set-tings. Secondly, eyelid and iris movements are represented as continuous variables according to the Facial Animation Parameters (FAP) of MPEG4, whereas previous methods deal only open, closed and fluttering state of eyelids.

The paper is organized as follows Section 2 describes the theoretical foundations of deformable models, appearance models and appearance-based trackers. Section 3 describes the tracking system. Section 4 presents the experimental results and the discussion. Finally, Sect. 5 is the conclusion and future avenues of research.

2 Problem definition

We use boldface upper-case letters for matrices (e.g. $\mathbf{D}_{k,l}$) and the corresponding dimensions as sub index. Vectors are written using boldface lower-case letters (e.g. $\mathbf{x}$, $\mathbf{r}$) and the corresponding components with lower-case letters with the sub-indices (e.g. $x_0, \ldots, x_n$). Greek letters are used for functions or constants (e.g. $\Phi(\mathbf{q})$), where vector $\mathbf{q}$ is the independent variable.

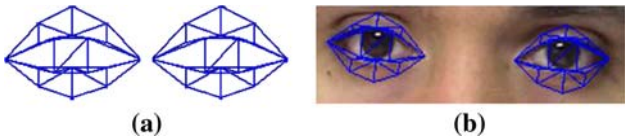


Fig. 2 The 3D mesh (**a**) is placed on the input image to model the eye region (**b**). According to FAP codes, this shape model encodes 3D mesh pose, eyelids, iris yaw and iris pitch

We model left and right eyes using a 3D deformable model composed of 36 vertices and 53 triangles, see Fig. 2a. This wire-frame covers eyeballs, upper and lower eyelids, sclera and iris, see Fig. 2b. The *nxi* matrix $\mathbf{F}$ determines the mesh deformation:

$$\mathbf{F}_{n,i} = \mathbf{D}_{n,i} + \mathbf{A}_{n,i,k} * \mathbf{g}_k \tag{1}$$

where n is the number of vertices and i indicates the Cartesian coordinates in the image. Matrix $\mathbf{D}_{n,i}$ encodes the biometry of each person according to eyes width, eyeballs height, eyes separation, eyes vertical difference, eyes outer corner, iris size, and iris asymmetry. Matrix $\mathbf{A}_{n,i,k}$ deforms the mesh depending on eyelid and iris movements. Vector $\mathbf{g} = (g_0, g_1, g_2)$ encodes three facial actions, eyelids, iris yaw, and iris pitch. Each component follows the MPEG-4 encoding for FAPs as continuous values in the range $[-1.0,1.0]$.

Facial feature detectors may give an automatic tracking initialization [7]. However, manual placement of the mesh guarantees a better tracking performance, since the biometry is kept along the sequence providing accurate estimations.

Consequently, tracking parameters are encoded in vector $\mathbf{q} = [\mathbf{r}, \mathbf{g}]$ to adjust the mesh to the eye region. Vector $\mathbf{r} = [\theta_x, \theta_y, \theta_z, x, y, s]$ contains the 3D orientation of the appearance according to three Euler's angles (the head pose), the image plane position and the scale. Tracking initialization provides $\mathbf{q}_0$ for the first frame to estimate the respective vectors along the image sequence.

An appearance model $\mathbf{x} = (x_0, ..., x_l)^T$, is created to represent the eye region [5], where l is a number of pixels in the appearance. The 3D mesh provides the shape and a warping function $\Psi(\mathbf{I}, \mathbf{q})$ allows to construct the texture. Each pixel from the input image $\mathbf{I}$ is mapped on a reference texture patch $\mathbf{x}$, according to shape deformation, $\mathbf{q}$, see Fig. 2b.

Assuming a weak perspective, the 3D mesh is projected onto the image plane according to vector $\mathbf{q}$. Similarly, a reduced mesh in frontal position is placed onto the reference texture. Regarding photometric transformations, we use a zero-mean-unit-variance normalization to partially compensate contrast variations. For the sake of clarity, from now on, all appearances are normalized.

Using two appearance resolutions of $l = 170$ and $l = 580$ pixels, the complete image transformation is implemented as follows: (a) Adapt the shape model $\mathbf{F}$ to image $\mathbf{I}$, see Fig. 2a. (b) Construct texture $\mathbf{x}$ using warping function $\Psi(\mathbf{I}, \mathbf{q})$.

(c) Perform zero-mean-unit-variance normalization on the obtained patch.

3 Appearance-based tracking

Identifying head and facial movements is challenging for marker-less approaches and appearance-based trackers (ABT). 3D head motion involves six degrees of freedom. Moreover, eyelids and irises are non-rigid surfaces that move fast and are more sensitive to illumination changes and self-occlusions.

In a given image sequence showing the eye region motion with small head movements, tracking consists of estimating the eyelid and iris positions in 3D for each frame. Therefore, the goal of ABTs is estimating vector $\mathbf{q}=[\mathbf{r}, \mathbf{g}]$ at each frame t, where $\mathbf{r}$ is the 3D shape pose and $\mathbf{g}$ is the facial action vector. In the context of tracking, adaptation results associated with the current frame will be propagated to the next frame.

We represent the estimated vector as $\hat{\mathbf{q}}_t$ and the respective estimated appearance as $\hat{\mathbf{x}}_t(\hat{\mathbf{q}}_t)$ [5]. With a view to make it succinct, we assume that $\mathbf{x}_t(\mathbf{q}_t)$ and $\mathbf{x}_t$ are equivalent.

3.1 Observation process

Let us consider an appearance sequence $\mathbf{X}_{l,t}$, corresponding to input image sequence $\mathbf{I}$:

$$\mathbf{X} = \begin{bmatrix} x_{0,0} & \cdots & x_{0,t} \\ \vdots & \ddots & \vdots \\ x_{l,0} & \cdots & x_{l,t} \end{bmatrix} = [\mathbf{x}_0, \ldots, \mathbf{x}_t] \tag{2}$$

where column vectors $\mathbf{x}_{i,t} = (x_{0,t}, \ldots, x_{l,t})^T$ are appearances of l pixels and row vectors $\mathbf{x}_{l,j} = (x_{l,0}, \ldots, x_{l,t})$, contain the pixel variation at the position l over time. We assume that row vectors $\mathbf{x}_{l,j}$, which are the values at the same position in the appearance, follow single Normal Distributions. Therefore, we can assume that appearance $\mathbf{x}_{i,t}$ follows a Multivariate Normal Distribution over time, $\mathbf{x}_{i,t} \sim N_l(\boldsymbol{\mu}, \boldsymbol{\Sigma}^2)$.

Since we assume that the variation in the intensity of pixels is mutually independent among them, there is no correlation $\boldsymbol{\Sigma}$ between them. Hence, the covariance matrix $\Sigma = \text{diag}(\sigma_0, \sigma_1)$. Consequently, we collect means and standard deviations of all Gaussians in vectors $\boldsymbol{\mu} = (\mu_0, \ldots, \mu_l)^T$ and $\boldsymbol{\sigma}^2 = (\sigma_0^2, \ldots, \sigma_l^2)^T$. Therefore, $\mathbf{x}_{i,t} \sim N_l(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ and the probability for each observation is given by the conditional likelihood function:

$$P(\mathbf{x}_t|\mathbf{q}_t) = \prod_{i=0}^{l} \frac{\mathrm{e}^{-(x_i-\mu_i)^2/2\sigma_i^2}}{\sigma_i\sqrt{2\pi}} \tag{3}$$

To obtain the expected appearance, we assume all estimations under an exponential envelope. Subsequently, we calculate both Gaussian parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$, by applying a linear recursive combination based on previous adaptations, the current estimation and a learning coefficient λ:

$$\begin{aligned} \boldsymbol{\mu}_{t+1} &= (1-\lambda)\boldsymbol{\mu}_t + \lambda\hat{\mathbf{x}}_t \\ \boldsymbol{\sigma}_{t+1}^2 &= (1-\lambda)\boldsymbol{\sigma}_t^2 + \lambda(\hat{\mathbf{x}}_t - \boldsymbol{\mu}_t)^2 \end{aligned} \tag{4}$$

where $\boldsymbol{\mu}$ is initialized with the first patch $\mathbf{x}_0$ and a constant value for the standard deviation $\boldsymbol{\sigma}$. Gaussian parameters gain significance after 50 frames, in accordance to the binomial distribution approximation and the central limit theorem for big sets of data. Learning factor λ is $1/t$ until the 50th frame, otherwise, it is a constant value.

3.2 Registration process

New pixel values are registered into the cumulative Multivariate Gaussian $\mathbf{X}$, according to shape deformation $\mathbf{F}$. Vector $\mathbf{q}_{t+1}$, shape deformation parameters, is estimated for the next frame using an adaptive velocity model. This is a deterministic function that consists of the last estimated vector and the increment vector $\delta\hat{\mathbf{q}}_t$ as follows:

$$\mathbf{q}_{t+1} \approx \hat{\mathbf{q}}_t + \delta\hat{\mathbf{q}}_t \tag{5}$$

where $\delta\hat{\mathbf{q}}_t$ is the increment vector for the mesh deformation.

The registration quality depends on estimating the optimal increment vector, which minimizes the error function between *expected* $\boldsymbol{\mu}$, and *estimated* $\mathbf{x}$ appearances.

$$\xi(\mathbf{q}_t) = \frac{1}{2}\sum_{i=0}^{l}\frac{(x_i-\mu_i)^2}{\sigma_i^2} = \frac{1}{2}||\mathbf{e}(\mathbf{q}_t)||^2 \tag{6}$$

where $\mathbf{e}(\mathbf{q}_t)$ is the residual appearance vector.

Let us consider the minimization problem of $\xi(\mathbf{q}_t)$: $\mathbf{R}^l \rightarrow \mathbf{R}$, which is a convex and a twice continuously differentiable function (Eq. (6)). The condition for vector $\mathbf{q}$ to be an optimal solution is $\nabla\xi(\mathbf{q}_t) = 0$. This problem is usually solved using an iterative first-order linear approximation based on the updated Gauss–Newton Iteration (GNI) algorithm [16]:

$$\xi(\mathbf{q}_t) = \frac{1}{2}||\nabla\xi(\mathbf{q}_t) + \mathbf{e}(0)||^2 \tag{7}$$

where $\nabla\xi(\mathbf{q}_t)$ is the Jacobian matrix $\mathbf{J}_{l,i}$, containing the derivatives of the distance function. This matrix is defined as $\mathbf{J}(\mathbf{q}) = \frac{\partial e_j}{\partial q_i}$ $(1 \leq j \leq l, 1 \leq i \leq 9)$, according to the number of pixels in the appearance and to the number of facial actions in vector $\mathbf{q}$.

The Vanilla method is the simplest and the most intuitive technique to find the optimal solutions for (Eq. (7)) [2];

$$\mathbf{q}_{t+1} = \mathbf{q}_t - \delta\nabla\xi(\mathbf{q}_t) \tag{8}$$

where δ gives the size step of descent in the search direction of the negative gradient $\nabla\xi(\mathbf{q}_t)$. Gradient descent suffers convergence problems such as heavy time-consumption

finding the optimal δ according to the slope. Another issue is that curvature of the error surface may not be the same in all directions.

Given the Jacobian matrix $\mathbf{J}$, we can essentially get the Hessian ($\nabla^2\xi(\mathbf{q}_t)$) by neglecting the highest order terms in the Taylor series of the distance function. The Hessian in this case becomes

$$\nabla^2\xi(\mathbf{q}_t) = \mathbf{J}(\mathbf{q}_t)^T\mathbf{J}(\mathbf{q}_t) \tag{9}$$

To provide a faster convergence, we propose to use an updated GNI, which includes information about the curvature. With a quadratic assumption for $\xi(\mathbf{q}_t)$ around $\mathbf{q}_t$, convergence is rapid but sensitive to the starting location, more precisely, to the linearity around the starting location:

$$\mathbf{q}_{t+1} = \mathbf{q}_t - \delta[\mathbf{J}(\mathbf{q}_t)^T\mathbf{J}(\mathbf{q}_t)]^{-1}\nabla\xi(\mathbf{q}_t) \tag{10}$$

For facial actions such as eyelids and irises, linearity has low probability. Nevertheless, we avoid local minima using backtracking procedures.

3.3 Handling outliers

By combining the observation and the registration processes, appearance textures are learnt on-line. The correct adaptation of an image agrees with the best match between shape $\mathbf{F}$ and image $\mathbf{I}$ according to vector $\mathbf{q}$. Therefore, the respective appearance is the closet estimation to the expected one.

However, illumination changes, occlusions, perturbing objects, and fast movements may introduce outlier pixels to the statistical model and learnt textures. Drifting problems occur when the ABT learns outliers by introducing them into the Gaussian distribution and the gradient estimation.

In order to handle outliers, occlusions and faster movements, we constrain texture learning and gradient descent using Huber's function [14,3], which is as follows:

$$\eta(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| \le c \\ c|x| - \frac{c^2}{2} & \text{if } |x| > c \end{cases} \tag{11}$$

where x is the normalized pixel value in the appearance $\mathbf{x}_t$ and c is a constant outlier threshold equivalent to $3{*}\boldsymbol{\sigma}$. Pixel x_i is an outlier when $||x_i|| > c$.

Subsequently, we combine the $\eta(\mathbf{x})$ function with the observation process to lessen the influence of outlier pixels:

$$P(\mathbf{x}_t|\mathbf{q}_t) = \prod_{i=0}^{l} \frac{e^{-\eta(x)}}{\sigma_i\sqrt{2\pi}} \tag{12}$$

Similarly, to down-weight the influence of outlier pixels in the registration process, we combine the GNI algorithm with the diagonal matrix $\Theta(\mathbf{x})$, whose terms are:

$$\Theta(x_i) = \frac{1}{x_i}\frac{\partial\eta(\mathbf{x})}{\partial x_i} = \begin{cases} 1 & \text{if } |x| \le c \\ \frac{c}{|x|} & \text{if } |x| > c \end{cases} \tag{13}$$

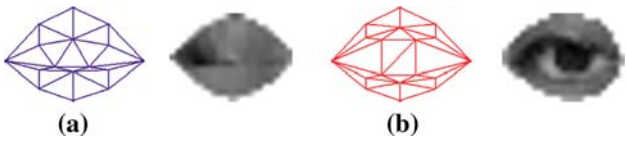


Fig. 3 Shape and appearance for the eyelid tracker (**a**). Shape and appearance for the iris tracker (**b**)

Consequently, we obtain the modified GNI:

$$\mathbf{q}_{t+1} = \mathbf{q}_t - \delta[\mathbf{J}(\mathbf{q}_t)^T\mathbf{J}(\mathbf{q}_t)]^{-1}\Theta(\mathbf{x}_t)\nabla\xi(\mathbf{q}_t) \tag{14}$$

3.4 Sequential tracking

Eyelids and irises have both smooth and spontaneous movements, which are difficult to track using statistical and deformable models. Eye region images are small and with low resolution when using monocular cameras. Eyelids and irises have a special interaction, suggesting correlation between them. On one hand, iris motion deforms the eyelid surface, which demands additional adaptation for ABTs. On the other hand, eye blinking occludes the iris region, forcing the iris tracker to recover the correct position after the occlusion. These small iris movements are called saccade and they are difficult to predict.

Tracking eyelids and irises with the same appearance models may produce drifting problems due to different intensity textures and occlusions. Therefore, we propose to construct two appearances for two independent trackers. Firstly, an appearance $\mathbf{x}(\mathbf{w})$ for eyelid tracking, excluding iris' FAP with vector $\mathbf{w} = [\mathbf{r}, g_0]$ from the shape. Those pixels in the inner eye region are warped as eyelid pixels in the appearance texture, see Fig. 3a.

Secondly, an appearance $\mathbf{x}(\mathbf{q})$ for iris tracking includes eyelid and iris pixels, see Fig. 3b. However, this tracker has special strengths to estimate irises rather than eyelids. Once the eyelid tracker gives its estimation, the iris tracker can estimate iris movements while refining the previous eyelid position.

3.4.1 Eyelid tracker

Using the current shape $\mathbf{F}_t$ based on the geometrical vector $\mathbf{w} = [\mathbf{r}, g_0]$, we construct an ABT for eyelids, $\mathbf{T}_\mathbf{w}$, taking the following steps, see Fig. 3a:

1. Construct the appearance $\mathbf{x}_t(\mathbf{w})$ for the image $\mathbf{I}_0$:
   (a) Obtain Shape, $\mathbf{F}_t = \mathbf{D} + \mathbf{A} * g_0$
   (b) Project the shape according to $\mathbf{w} = [\mathbf{r}, g_0]$
   (c) Apply the warping function, $\Psi(\mathbf{I}, \mathbf{w}) = \mathbf{x}_t(\mathbf{w})$
2. Obtain Gaussian parameters using the likelihood function:
   (a) $\boldsymbol{\mu}_{t+1} = (1-\lambda)\boldsymbol{\mu}_t + \lambda\mathbf{x}_t(\mathbf{w})$

(b) $\boldsymbol{\sigma}_{t+1}^2 = (1-\lambda)\boldsymbol{\sigma}_t^2 + \lambda(\mathbf{x}_t(\mathbf{w}) - \boldsymbol{\mu}_t)^2$

3. Calculate the Jacobian and Hessian matrices:
   (a) Calculate the residual appearance, $\xi(\mathbf{w}_t) = \frac{1}{2}||\mathbf{e}(\mathbf{w}_t)||^2$
   (b) Compute the partial derivatives, $\mathbf{J}(\mathbf{w}_t) = [\partial e_j/\partial w_i]$, $1 \le j \le l$, $1 \le i \le 7$
   (c) Obtain the Hessian, $\nabla^2\xi(\mathbf{w}_t) = \mathbf{J}(\mathbf{w}_t)^T\mathbf{J}(\mathbf{w}_t)$

   The eyelid FAP g_0, may vary in $\pm$ 2.0 in two successive frames. Therefore, the partial differences include the whole range [−1.0,1.0].
4. Estimate the new shape $\mathbf{F}_t$ applying GNI algorithm:
   (a) Construct the diagonal matrix $\Theta(\mathbf{x}_t)$
   (b) Compute the search direction, $d_k = -[\mathbf{J}(\mathbf{w}_t)^T\mathbf{J}(\mathbf{w}_t)]^{-1}\Theta(\mathbf{x}_t)\nabla\xi(\mathbf{w}_t)$
   (c) Choose the step length δ via backtracking line-search procedure:
      (i) Consider the search direction in (b) and the starting vector $\mathbf{w}_k = \mathbf{w}_t \in dom\ \xi$. Set $\delta_k = \sum_0^k (-1)^k(1/k)$
      (ii) [1]While $\xi(\mathbf{w}_k + \delta_k d_k) > \xi(\mathbf{w}_k) + \delta_k d_k$
      (iii) Set $\delta = \delta_k$.
   (d) Update variables, $\mathbf{w}_{k+1} = \mathbf{w}_k - \delta d_k$
   (e) Test convergence for stopping iterations, otherwise, consider $k = k + 1$ and go to (b).

As we mentioned above, there are two main steps to estimate eyelids correctly; step 3(b) considers the whole FAP range to estimate the gradient descent. Step 4(c) calculates the damping factor, δ, using backtracking procedures. Therefore, the eyelid tracker provides a space of solutions and a faster convergence.

3.4.2 Iris tracker

For the same image and using the current shape $\mathbf{F}_t$, based on the geometrical vector $\mathbf{q} = [\mathbf{r}, g_0, g_1, g_2]$, we construct an ABT for irises, $\mathbf{T}_\mathbf{q}$, taking the following steps, see Fig. 3b:

1. Construct the appearance $\mathbf{x}_t(\mathbf{w})$ for the image $\mathbf{I}_0$:
   (a) Obtain Shape, $\mathbf{F}_t = \mathbf{D} + \mathbf{A} * \mathbf{g}$
   (b) Project the shape according to $\mathbf{q} = [\mathbf{r}, g_0, g_1, g_2]$
   (c) Apply the warping function, $\Psi(\mathbf{I}, \mathbf{q}) = \mathbf{x}_t(\mathbf{q})$
2. Obtain Gaussian parameters:
   (a) $\boldsymbol{\mu}_{t+1} = (1-\lambda)\boldsymbol{\mu}_t + \lambda\mathbf{x}_t(\mathbf{q})$
   (b) $\boldsymbol{\sigma}_{t+1}^2 = (1-\lambda)\boldsymbol{\sigma}_t^2 + \lambda(\mathbf{x}_t(\mathbf{q}) - \boldsymbol{\mu}_t)^2$
3. Calculate the Jacobian and Hessian matrices:
   (a) Calculate the residual appearance, $\xi(\mathbf{q}_t) = \frac{1}{2}||\mathbf{e}(\mathbf{q}_t)||^2$
   (b) Compute the partial derivatives, $\mathbf{J}(\mathbf{q}_t) = [\partial e_j/\partial q_i]$, $1 \le j \le l$, $1 \le i \le 7$
   (c) Obtain the Hessian, $\nabla^2\xi = \mathbf{J}(\mathbf{q}_t)^T\mathbf{J}(\mathbf{q}_t)$

   Iris movements are more subtle than eyelids; the iris' FAP may change approximately $\pm$ 0.5 in two successive frames. Hence, gradients are estimated in the range $[\mathbf{g} - 0.5, \mathbf{g} + 0.5]$.
4. Estimate the new shape $\mathbf{F}_t$ applying GNI algorithm:
   (a) Construct the diagonal matrix $\Theta(\mathbf{x}_t)$.
   (b) Compute the search direction, $d_k = -[\mathbf{J}(\mathbf{q}_t)^T\mathbf{J}(\mathbf{q}_t)]^{-1}\Theta(\mathbf{x}_t)\nabla\xi(\mathbf{q}_t)$
   (c) Choose the step length δ via backtracking line-search procedure:
      (i) Consider the search direction in (b) and the starting vector $\mathbf{q}_k = \mathbf{q}_t \in$ dom ξ. Set $\delta_k = \frac{\delta_{k-1}}{\upsilon}$, for $\upsilon > 1$.
      (ii) While $\xi(\mathbf{w}_k + \delta_k d_k) > \xi(\mathbf{w}_k) + \delta_k d_k$
      (iii) Set $\delta = \delta_k$.
   (d) Update variables, $\mathbf{q}_{k+1} = \mathbf{q}_k - \delta d_k$
   (e) Test convergence for stopping iterations, otherwise, consider $k = k + 1$ and go to (b).

Even when the iris movements are not smooth facial actions, the backtracking procedure is more deterministic by decreasing the damping factor δ, which is a constant rate.

3.4.3 Combination of ABTs

The residual image $\mathbf{e}(\mathbf{q})$, or the Mahalanobis distance $\xi(\mathbf{q})$ can be used as error measures to estimate the effectiveness of both trackers. The number of iterations k is estimated experimentally according to the average iterations needed to converge when tracking long image sequences.

Appearance modelling, observation process and gradient matrices are based on results of previous adapted frames. Therefore, both trackers can run simultaneously and independently until the GNI algorithm starts, then, they run sequentially. Once the eyelid tracker converges, $\min[\xi(\mathbf{w}*)] = \xi(\mathbf{w}*)$, the iris tracker starts the iterative process, see Fig. 4.

On one hand, the eyelid tracker $\mathbf{T}_\mathbf{w}$ sets the starting point in 4(c)(i) by modifying vector $\mathbf{q}_t$, $\mathbf{q}_k = [\mathbf{r}, g_0^*, g_1, g_2]$, where g_0^* is the eyelid tracking solution for the current frame, $(t+1)$. On the other hand, the step-size estimation is constrained to improve the convergence error of eyelid tracker, $\xi(\mathbf{q}*) \le \xi(\mathbf{w}*)$.

Consequently, when applying twice the iterative minimization, both trackers are efficiently connected. The eyelid tracker is independent from the iris estimation while the second tracker is led to the correct eyelid position and forced to improve the eyelid convergence error.

The sequential tracking combines the strengths of both ABTs. Specific shape models contribute to avoid the high contrast between eyelids, sclera and irises. The space of solutions is extended using different FAP range and step-size

[1] Armijo Condition [16].

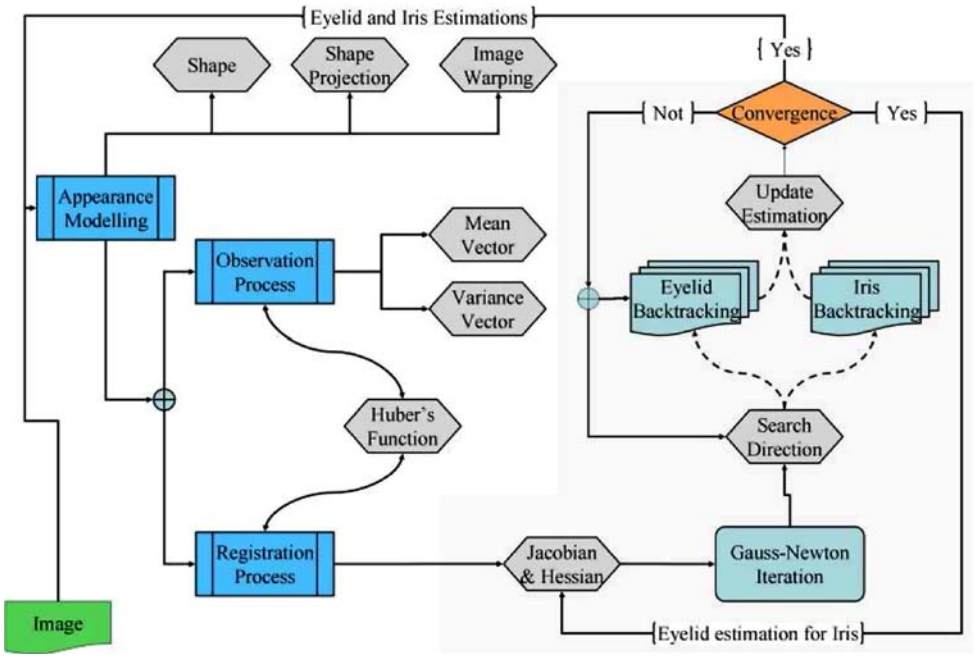


Fig. 4 Structure of the sequential gaze tracking system

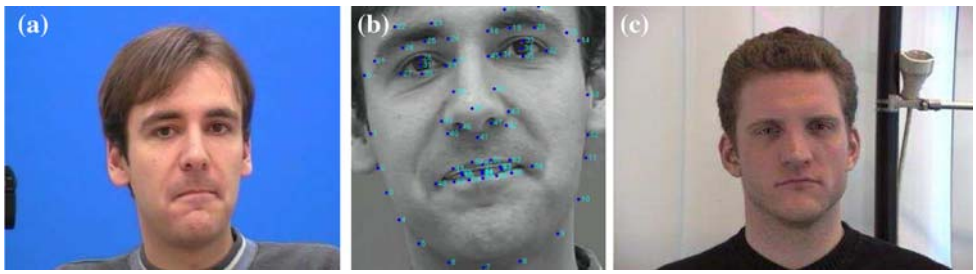


Fig. 5 Public databases and ground truth. We use the "FGnet Talking" database for face tracking (**a**) with the corresponding ground truth (**b**). Additionally, we use the FGnet Facial Expressions and Emotion Database (**c**)

for the estimation of the gradient. The use of two particular backtracking procedures improve the convergence for each tracker.

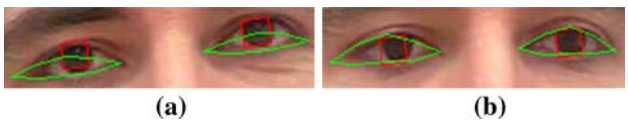


Fig. 6 The image **I** and shape **F** edges do not match in a wrong result (**a**) while matching in a correct one (**b**)

4 Experimental results

We have tested our method with two public databases; one is the FGnet Talking Face Video [12] for face tracking. It provides five image sequences, each one is composed by one thousand images of head and shoulders with the ground truth for upper and lower eyelids, and iris centre, see Fig. 5. The other database is the FGnet for facial expressions analysis [13]. Additional recorded videos are tested. They present other challenges such as gaze movements, gazes in 3D, eyeglasses, illumination changes and occlusions.

We use the ground truth of FGnet to evaluate the accuracy of the results. Convergence error is used to evaluate those experiments with image sequences without ground truth such as FGnet for expression analysis and other recorded videos in our laboratory. In this case, the output image helps to validate the results, since the drawn shape depends on the estimation and the image is the original input, see Fig. 6.

The Experiments were run on a 3.2 GHz Pentium PC, in ANSI *C* code. Image sequences were recorded with monocular cameras and standard resolutions. We tested the method with two sizes of appearances, 170 and 580 pixels.

4.1 Ground truth comparison

To compare the sequential tracking results with the ground truth, we used the FGnet database [12]. The images have a size of 720×576 pixels and 800 frames. The actor performs slow and short head movements, close to the frontal position while moving eyelids and irises.

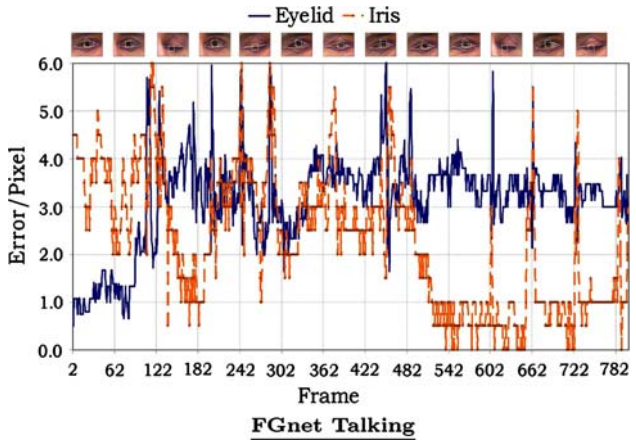


Fig. 7 Estimated positions are compared with the ground truth of FGnet DB. 3.2 and 2.2 averages for both trackers

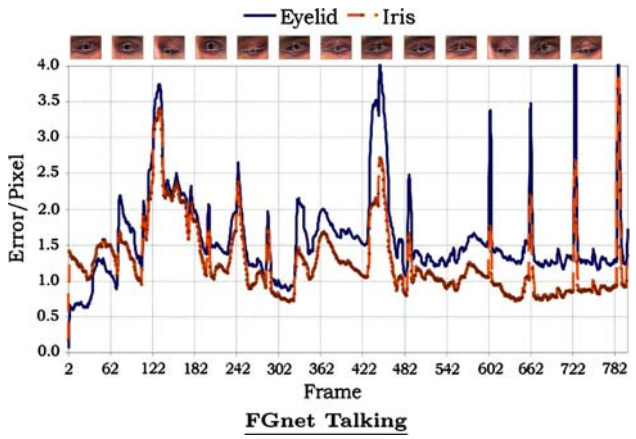


Fig. 8 The error between expected and estimated appearances as confidence value of the effectiveness

For eyelid comparison, we calculate the average difference for the vertical positions of upper and lower eyelids. It is worth to mention that those positions may vary depending on the horizontal position for both ground truth and tracking result. However, the average error for eyelids is 3.2 pixels per frame. For irises, we calculate the average difference for the iris centre. Likewise, the average error for irises is low, 2.2 pixels per frame.

In Fig. 7, it is possible to see that higher errors coincide with those frames where eye blinking or fast iris movements are occurring. Nonetheless, errors decrease when both trackers improve the convergence in the subsequent frames. Altogether, low errors compared to the ground truth agree with low errors when comparing expected and estimated appearances, see Fig. 8. Wrong adaptations are visible when the drawn contour does not match the image. Moreover, higher errors differ about 6.0 pixels with the ground truth and 4.5 pixels with the expected appearance.

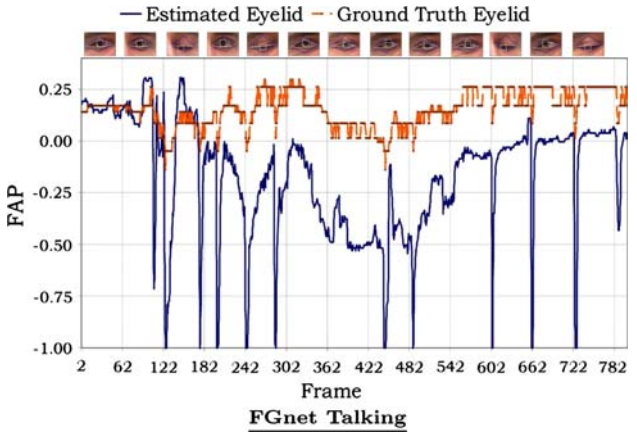


Fig. 9 The ground truth is not accurate by marking blinks

In relation to this ground truth data, it is important to mention that it may have wrong annotations, which are possible to verify with the eyelid tracker $\mathbf{T_w}$. This tracker estimates vector $\mathbf{w} = [\mathbf{r}, g_0]$, where g_0 $[-1.0, 1.0]$, -1.0 when the eyes are closed and 1.0 when they are in an open position. Frames such as 125, 126, 242, 243, 603, 604 give us a value of $g_0 = -1.0$ and the distance between upper and lower eyelids is zero pixels. However, the ground truth differences are never zero pixels (-1.0 corresponding to the FAP), see Fig. 9.

4.2 Eyelid tracking

As we mentioned before, the eyelid tracker $\mathbf{T_w}$ estimates vector $\mathbf{w}$ and eyelid facial action g_0 as continuous variables in the range $[-1.0,1.0]$ for closed and open eyes, respectively.

Given that for each iteration of the GNI algorithm, the 3D mesh varies according to rotation, translation and eyelids, with respect to the previous estimated vector, we obtain an appearance space with k possible solutions.

We have used an image sequence of 700 frames, recorded in our laboratory (Eyelid Sequence) with monocular cameras and standard illumination. Each frame includes the head and shoulders performing extreme eye facial actions, which deform the eyelid surface in 3D, see Fig. 10. The eye-cropped region shows the experimental results. The tracker does not depend on the image size because it is warped into the same appearance texture.

Without using edge detectors, the eyelid tracker is able to handle low and smooth movements like eye slitting, eye closing and eye squinting. Forced and spontaneous movements like eyelid raising, eyelid tightening, winking and blinking, are handled correctly. These estimations are also independent from the iris position because the inner eye pixels are not visible when the eyelid in the image matches the shape model.

Psychological studies have addressed the importance of analysing these movements regarding emotion analysis,

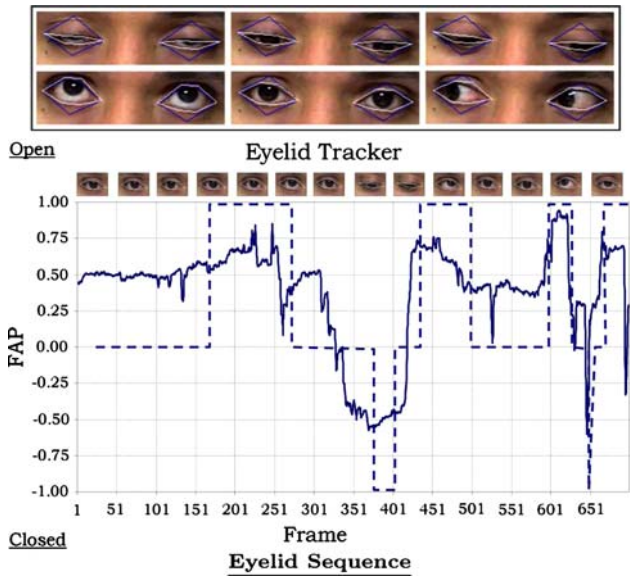


Fig. 10 The eyelid estimation $\mathbf{T_w}$ is a continuous curve, instead of the dash line for discrete states

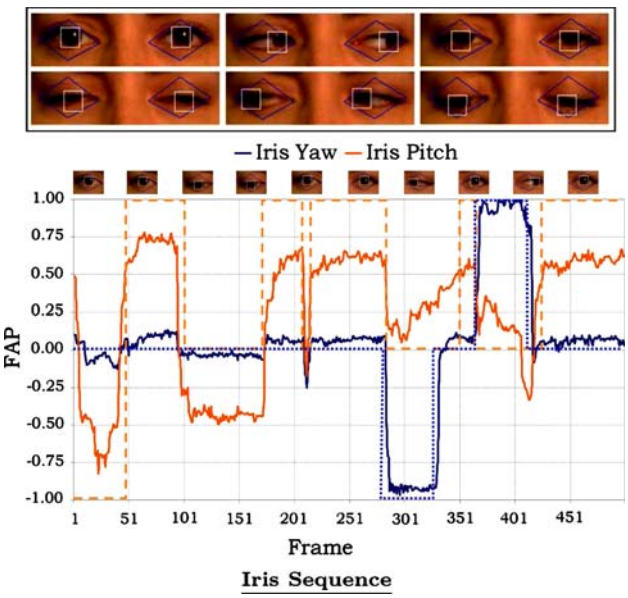


Fig. 11 The estimations $\mathbf{T_q}$ for iris yaw and pitch are continuous curves instead of estimating discrete states

image encoding and HCI. These movements are characterised by the Facial Actions Coding System (FACS) [11].

4.3 Iris tracking

Iris tracker $\mathbf{T_q}$ estimates the whole vector $\mathbf{q} = [\mathbf{r}, \mathbf{g}]$. The iris yaw and pitch parameters are evaluated as continuous variables in the same FAP range.

In this experiment, we use an image sequence of 500 frames of size 640×480 pixels (Iris Sequence). The video was recorded in our laboratory with a photographic camera in VGA mode and standard illumination. The actor shows iris movements in all directions and looking askance.

The iris tracker can deal with the four different FACS [11]; eyes turned up, down, left, right and extreme movements like with an askance look, where the iris could be partially occluded or distorted by the 3D perspective. However, involuntary movements such as iris saccades are commonly detected after eyelid occlusions or gaze accommodation.

Figure 11 shows how the iris tracker correctly estimates yaw and pitch movements. The dash and light curves represent gaze-tracking results using the discrete scale of related approaches. A good result corresponds to the correct matching between the image and the drawn rectangle around to the iris, which comes from the estimated vector $\mathbf{q}$.

Frames between 101 and 151 show g_2 near -0.5, when the subject is looking down. In frame 300 the iris pitch is $g_1 = -1.0$ because the subject is looking askance to the left, while frame 400 has $g_1 = 1.0$ for askance to the right. Figure 11 shows the results for both eyes when the upper eyelids are drooping and occluding the iris.

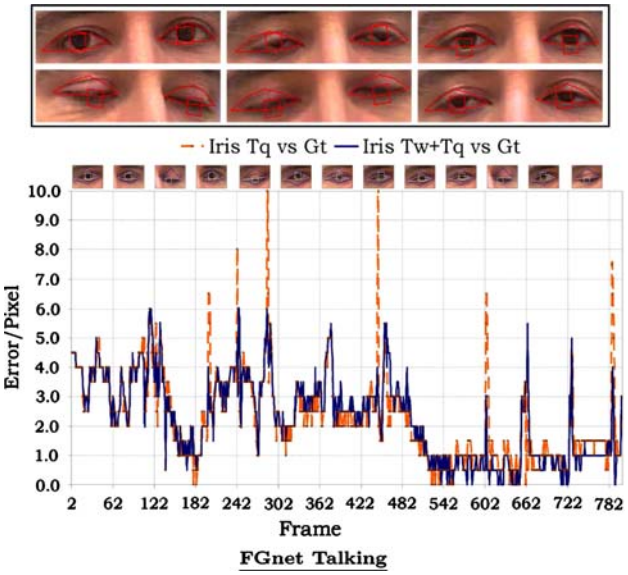


Fig. 12 The iris tracker $\mathbf{T_q}$ is compared to the sequential tracking $\mathbf{T_w} + \mathbf{T_q}$, by measuring the iris estimation in relation to the ground truth

4.4 Sequential tracking

First, we show the results when the iris tracker follows the eyelids without any previous eyelid tracker estimation, see Fig. 12. The search direction and Jacobian matrix for eyelids are calculated in the range $[-1.0, 1.0]$. However, the backtracking procedure is not the same for the eyelid tracker, because of a damping factor which influences the descent of all appearance parameters.

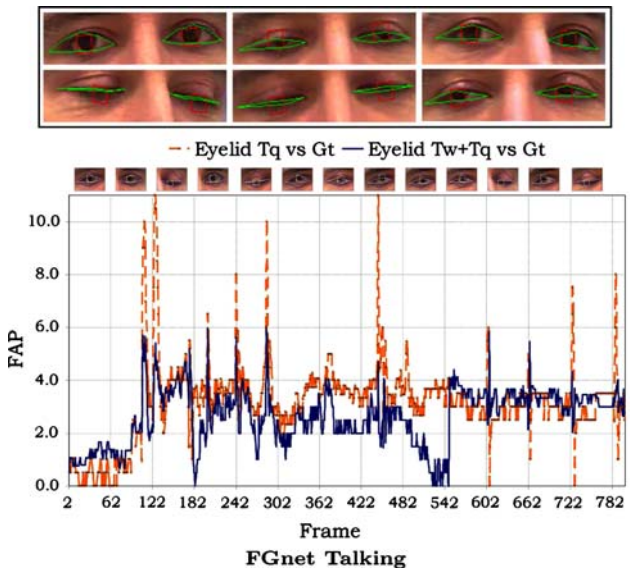


Fig. 13 Sequential tracking $\mathbf{T_w} + \mathbf{T_q}$ improves the eyelid estimations since the iris tracker $\mathbf{T_q}$ (*dash curve*) is not able to track eyelids

In order to compare both trackers, we use the previously introduced dataset of FGnet [12]. When eyes are blinking, the iris tracker is not able to adapt to the shape model. Hence, the eyelid pixels over the inner eye region are rejected and are not included in the appearance model and the shape remains in the previous correct adaptation, see again Fig. 12.

However, in the same sequence, the sequential tracking obtains first the eyelid position from eyelid tracker $\mathbf{T_w}$. Next, iris tracker $\mathbf{T_q}$, starts estimating the iris movements while refining the current eyelid estimation. We can see the same frames after applying the sequential tracking, see Figure 13. The eyelid estimation is more accurate due to the second iterative process. This is demonstrated in the descent of the error estimation in relation to the likelihood and the ground truth, see Fig. 13.

All facial actions $\mathbf{g}$ and the 3D shape pose are independently estimated since the Jacobian $\mathbf{J}$ is calculated by partial differences. Although the iris tracker includes eyelid facial action, it cannot find big changes such as closed eyes or blinks. On one hand, the contrast between the inner and the outer eye pixels is greater than the outlier threshold in Huber's function. Therefore, inner pixels are outliers and are not learnt fast enough to expect quick changes. On the other hand, the gaze vector is estimated in a small range with a ratio of $\pm$ 0.5.

Another test for the sequential gaze tracking is done using an image sequence from the FGnet DB for facial expression. It contains 100 frames of 320 $\times$ 240 pixels, where the actor performs an expression of anger while squinting, blinking and moving the iris.

In Fig. 14, it can be seen how the iris position is retrieved after eyelid occlusion. Moreover, the eyelid tracker adjusts

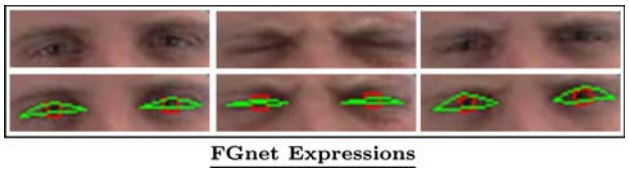


Fig. 14 The sequential tracking estimates eyelids and irises while expressing emotions, squinting and blinking

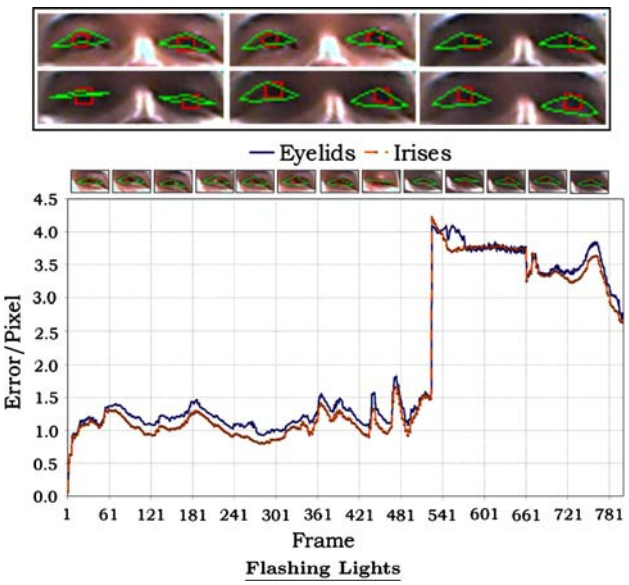


Fig. 15 The ABT is stable to illumination changes by learning the new environment conditions and decreasing the error

correctly to the eyelid position during the iris motion. Only eyelid estimation and its convergence are shared by both trackers. Therefore, iris estimations do not influence the eyelid tracker at the next frame.

4.5 Illumination changes

Appearance trackers commonly suffer drifting problems due to sensitiveness to illumination changes. This happens mainly because of their dependency on training of textures and shapes.

To prove the capability of our method to handle illumination changes, we use an image sequence of 800 frames (Flashing Light Sequence), each of a size of 352 $\times$ 288 pixels, recorded with a web camera in an indoor scenario. The subject is using his hands, creating shadows and occluding his face while lights are flashing, see Fig. 15.

The pictures in Fig. 15 show how the ABT can adapt to the 3D shape while there are changes in illumination. This is a controlled learning ability based on combining likelihood and Huber's function. Besides, both trackers handle different learning rates, $\lambda_{\mathbf{w}}$ and $\lambda_{\mathbf{q}}$, because eyelids change faster than irises.

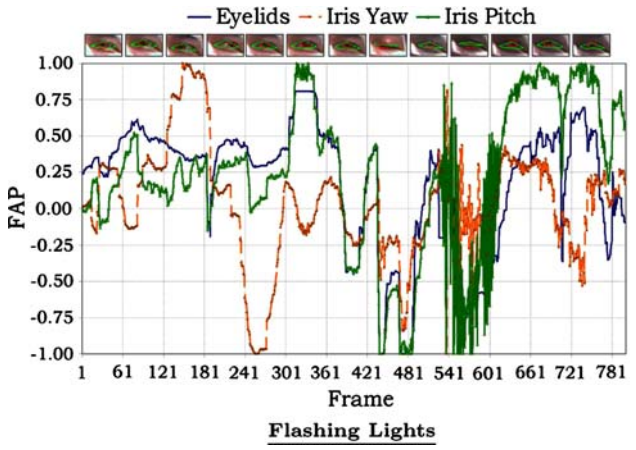


Fig. 16 The ABT recovers stability if the illumination changes are not extreme, as with flashing lights

In frame 525, a fluorescent lamp is turned off to vary illumination. The FAP plot shows how both eyelid and iris trackers need extra accommodation due to environment changes, see Fig. 16.

Appearance space in the iterative process is more diverse, as for each iteration k, the gradient descent may change the step size and the direction. However, we keep the same number of iterations, if we assume that illumination changes are not extreme, as with flashing lights.

Sequential tracking improves estimations as long as the new illumination conditions remain stable. Therefore, expected appearances have less information from the previous illumination and the GNI algorithm recovers stability, see the plot in Fig. 15.

4.6 Translucent textures

It is an interesting challenge to analyse images when subjects wear eyeglasses or sunglasses. Sequential gaze tracking is able to handle these cases where the eye region is partially occluded by a translucent surface.

In order to test tracking stability for bright and translucent surfaces such as sunglasses, we use our own-recorded sequence of 240 frames (Wearing Sunglasses Sequence). This was recorded indoors with low illumination and semi-transparent sunglasses. The camera is photographic and the size of the image is of 640 × 480 pixels, see Fig. 17.

The subject wears sunglasses from the beginning allowing the ABT to learn the translucent texture. However, when the head is in profile, one eye could be partially occluded by the nose, darkening the occluded eye and also causing specular reflectance on the sunglasses.

We consider facial symmetry up to 45° of tilt rotation, because both eyes are visible. If rotation is greater than 45°, the face is considered asymmetric and the appearance

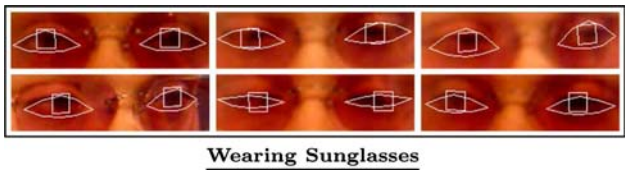


Fig. 17 Gaze tracking results under translucent textures. ABT learns on-line this textures, assumes facial symmetry after 45° to deal with profile

matches mainly the non-occluded eye. Both eyes can be tracked independently by extending the sequential tracking at the cost of more computational effort.

The ABT handles the darker side and the the specular effects as variations of illumination. In the observation process, high intensity changes are considered outliers and excluded from learning. Therefore, the 3D shape remains in the previous correct positions, see again Fig. 17.

4.7 Occlusions and real-time

There are real situations where we can see how occlusions affect the estimation errors for both trackers. In the experiment, we present an image sequence of 600 frames, recorded in an indoor scenario with a monocular camera (Wearing Glasses Sequence). The subject performs head movements and exaggerated facial actions while illumination is subtly changed. In one frame, the subject puts on a pair of eyeglasses, which produces occlusions and intensity variations.

Spontaneous eyelid blinks occlude the iris region in two or three frames. After the blinking, the iris tracker has to recover the correct adaptation because during the blinking the position is the same. This search can use one or two more frames while increasing the estimation error, for example, at frames 17, 164, 232. However, iris saccade movements deform the eyelid surface, changing the descent direction for the eyelid tracker. In Fig. 18, we can see how the iris movements influence the estimation error of the eyelid tracker in frames such as 96, 330, 483, 551.

In order to achieve the real-time requirements, we have tested the same image sequence using an ABT with a small appearance resolution of 170 pixels. Accuracy and robustness are tested by comparing the output images and the time spent to find the correct adaptation (i.e. the time needed to complete the algorithm in Fig. 19).

While tracking with a small appearance resolution, we obtained an average of 85% of correct adaptations and 32 frames per second (fps). Instead, big appearance resolution provides an average of 96% of correct adaptations and 1.1 fps. It is worth to mention that for big resolution appearances, the iris had fewer pixels in the 2 × 3 appearance than in the one of 5 × 6 pixels.

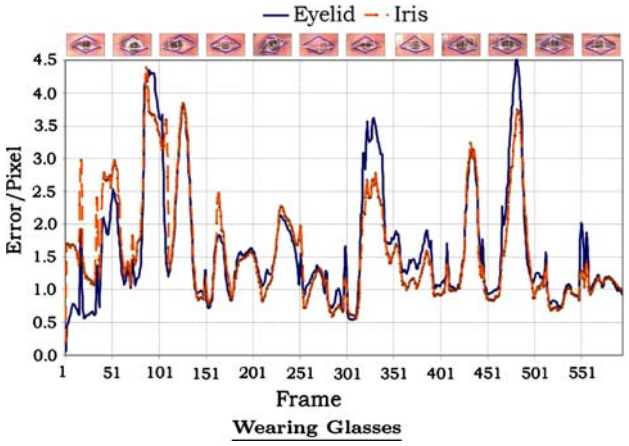
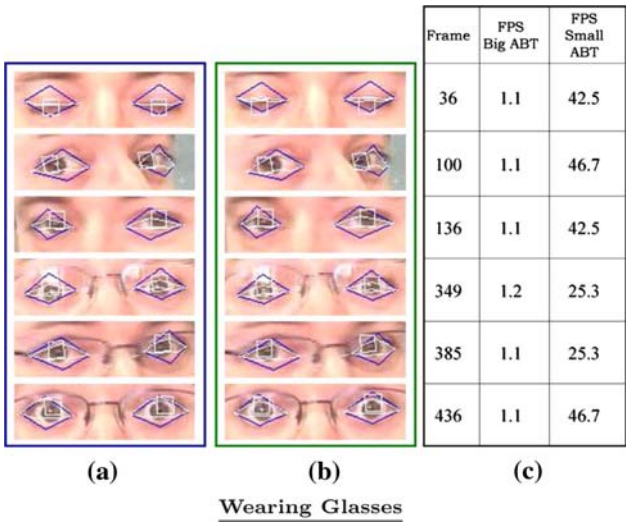


Fig. 18 This sequence exhibits both blinks and saccades. For example, we can see eyelid blinks at frames 17, 164, 232. We can also see iris saccades at frames 96, 330, 483 and 551



Frame	FPS Big ABT	FPS Small ABT
36	1.1	42.5
100	1.1	46.7
136	1.1	42.5
349	1.2	25.3
385	1.1	25.3
436	1.1	46.7

Fig. 19 Performance comparison between two ABT of 580 pixels (**a**) and 170 pixels (**b**). It is possible to obtain robust results (**c**) with similar accurate results (**a**) and (**b**)

5 Conclusions

We have shown that appearance-based trackers can achieve automatic gaze analysis by combining deterministic and stochastic methods. It is possible to extract gaze motion information without using edge detectors or colour information.

Our proposed technique deals with different eyelid and iris movements according to FACS codes; upper eyelid raising, tightening drooping, squint, blinking, slit, eyes closed, iris yaw, iris pitch and saccade. This algorithm is extensible to asymmetric facial actions to track winks, but it is time-consuming.

We have introduced new strengths for ABTs in general. The two backtracking procedures and the modified Gauss-Newton Iterative algorithm are essential to increase the convergence speed. Therefore, the algorithm can find the optimal combination of search direction and damping factor, producing a better appearance space and possible solutions.

Combining the strengths of two different and independent trackers, we improve gaze tracking. We have proven the need of combining two trackers to deal with different facial actions. Each tracker has its own appearance model, backtracking procedure and learning coefficient.

By applying the Huber's function, we have demonstrated robustness to handle illumination changes, occlusions and fast movements. Appearance registration and gradient descent are controlled for those pixels whose value is higher than the outlier threshold.

We have shown the system's capabilities with positive experimental results for the most challenging issues as markless trackers and gaze tracking algorithms. The experiments are focused on eyelid and iris tracking, illumination changes, occlusions, spontaneous movements and real-time.

The proposed method is a significant contribution to-wards gaze motion tracking using an appearance-based model. On one hand, the shape representation, the statistical modelling, and optimization algorithms give an alternative to already proposed methods. On the other hand, the robustness and accuracy make this method suitable for HCI applications and psychological analysis, since it can work in standard video surveillance environments without previous training. Our future work involves the inclusion of psychological models of behaviours and expressions, since gazes determine brain activity (for example, saccades are initiated by the frontal lobe of the brain called Brodmann area 8), as well as supports deceit and truth detection. We also want to include head, eyebrows and lips for a detailed facial motion description for expression analysis.

Acknowledgments This work is supported by EC grants IST-027110 for the HERMES project and IST-045547 for the VIDI-video project, by the Spanish MEC under projects TIN2006-14606 and CONSOLIDER INGENIO 2010 (CSD2007-00018). Jordi Gonzàlez also acknowledges the support of a Juan de la Cierva Postdoctoral fellowship from the Spanish MEC.

References

1. Bernogger, S., Yin, L., Basu, A., Pinz, A.: Eye tracking and animation for mpeg-4 coding. In: ICPR'98: Proceedings of the 14th International Conference on Pattern Recognition, Washington, USA, vol. 2, pp. 1281–1284 (1998)
2. Cauchy, A.: Méthodes générales pour la résolution des systémes déquations simultanées. C.R. Acad. Sci. Par. **25**, 536–538 (1847)
3. Clark, D.: Comparing huber's m-estimator function with the mean square error in backpropagation networks when the training data is noisy. The Information Science Discussion Paper Series, **2000**(19) (2000)
4. Cohen, I., Sebe, N., Chen, L., Garg, A., Huang, T.S.: Facial expression recognition from video sequences: temporal and static modeling. Comput. Vision Image Understand **91**(1–2), 160–187 (2003)

5. Cootes, T.F., Taylor, C.J.: Statistical Models of Appearance for Computer Vision. University of Manchester (2004)
6. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models - their training and application. Comput. Vis. Image Understand. **61**(1), 39–59 (1995)
7. Cristinacce, D., Cootes, T., Scott, I.: A multi-stage approach to facial feature detection. In: Proceedings of the British Machine Vision Conference, Kingston, September 2004
8. Doubek P., Svoboda T., Nummiaro K., Koller-Meier E., Van Gool L. Face Tracking in a Multi-Camera Environment Computer Vision Lab, 266, November 2003
9. Edwards, G.J., Cootes, T.F., Taylor, C.J.: Face recognition using active appearance models. In: Proceedings of the Fifth European Conference on Computer Vision, vol. 2, pp. 581–695 (1998)
10. Ekman, P.: Emotions Revealed. Times Books Henry Holt and Company, New York (2003)
11. Ekman, P., Friesen, W.V.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)
12. Face and Gesture Recognition Working Group. www-prima.inrialpes.fr/FGnet
13. FGnet Facial Expressions and Emotion Database. www.mmk.ei.tum.de/waf/fgnet/feedtum.html
14. Huber, P.J.: Robust estimation of a location parameter. Ann. Math. Statist. **35**, 73–101 (1964)
15. Moriyama, T., Xiao, J., Cohn, J., Kanade, T.: Meticulously detailed eye model and its application to analysis of facial image. IEEE Trans. Pattern Anal. Mach. Intell. **28**(5), 738–752 (2006)
16. Nocedal J., Wright S. Numerical optimization. 1999
17. Sirhey, S., Rosenfeld, A., Duric, Z.: A method of detecting and tracking irises and eyelids in video. In: International Conference on Pattern Recognition, vol. 35, pp. 1389–1401 (2002)
18. Tan, H., Zhang, Y.: Detecting eye blink states by tracking iris and eyelids. Pattern Recognition Lett. (2005)
19. Wu Y., Liu, H., Zha, HL.: A new method of detecting human eyelids based on deformable templates. In: Systems, Man and Cybernetics, 2004 IEEE International Conference, vol. 1, pp. 604–609 (2004)
20. Zhu, Z., Ji, Q.: Eye and gaze tracking for interactive graphic display. Mach. Vis. Appl. **15**, 139–148 (2004)

Author biographies

Javier Orozco has a degree in Mathematics and Physics from the University of Antioquia, Colombia (1995–2000). In 2007, he obtained a MSc in Computer Science, at the Computer Vision Center in Autonomous University of Barcelona, where he is currently researching on Human Emotion Analysis and Pattern Recognition. He is involved in national projects as Consolider Ingenio 2010 and Cicyt Sisyphus, and European projects as Hermes and the European Network for the Advancement of Artificial Cognitive Systems euCognition.

F. Xavier Roca received the degree in Computer Sciences in 1990 from the Universitat Autònoma de Barcelona and the Ph.D. degree in Computer Sciences in 1998 from the Universitat Autònoma de Barcelona. Since 1993 he is an associate professor at the Computer Sciences Department of the Universitat Autònoma de Barcelona and a staff researcher of the CVC. He has participated in a several projects in computer vision, with private and public founds. In 2002, with other members of the CVC create a spin-off from the results of a research projects. He has published 30 papers in national and international conferences and journals. From 1998–2003 he was marketing director of the CVC and from 2003-04 vice-director. From the beginning of 2006 is Director of the Computer Science Department of Universitat Autonoma de Barcelona.

Jordi Gonzàlez obtained his Ph.D. degree in 2004, from Universitat Autònoma de Barcelona. At present he is a Juan de la Cierva postdoctoral researcher at the Institut de Robòtica i Informàtica Industrial (UPC-CSIC). The topic of his research is the cognitive evaluation of human behaviours in image sequences. He has more than 70 publications about active camera control, segmentation, tracking, human motion understanding (interpretation and reasoning), natural language text generation and automatic behavioural animation. He has participated as WP-Leader in the European projects HERMES and VIDI-Video, and he is member of the European Network euCognition. He has co-founded the Image Sequence Evaluation research group at the CVC in Barcelona.