Hierarchical Eyelid and Face Tracking

J. Orozco[1], J. Gonzàlez[2], I. Rius[1], and F.X. Roca[1]

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

Abstract. Most applications on Human Computer Interaction (HCI) require to extract the movements of user faces, while avoiding high memory and time expenses. Moreover, HCI systems usually use low-cost cameras, while current face tracking techniques strongly depend on the image resolution. In this paper, we tackle the problem of eyelid tracking by using Appearance-Based Models, thus achieving accurate estimations of the movements of the eyelids, while avoiding cues, which require high-resolution faces, such as edge detectors or colour information. Consequently, we can track the fast and spontaneous movements of the eyelids, a very hard task due to the small resolution of the eye regions. Subsequently, we combine the results of eyelid tracking with the estimations of other facial features, such as the eyebrows and the lips. As a result, a hierarchical tracking framework is obtained: we demonstrate that combining two appearance-based trackers allows to get accurate estimates for the eyelid, eyebrows, lips and also the 3D head pose by using low-cost video cameras and in real-time. Therefore, our approach is shown suitable to be used for further facial-expression analysis.

1 Introduction

Face tracking is focused on the estimation of the head pose position and predefined facial actions, usually eyebrow and lip movements. Current tracking techniques do not address eyelid tracking due to its requirements of accuracy while face tracking demands robustness. Eye states are highly involved in facial expression while determining the degree of sincerity. Once a proper description of the eyelid movement is obtained, it is possible to apply facial expressions analysis for HCI, and emotion evaluation. We need to achieve an accurate eyelid tracking in real time, and eyebrow, lip and head pose estimation. They require a robust and accurate technique for extracting facial actions, which involve a great challenge, especially for gazes and eye blinks.

In order to achieve this goal, current face tracking approaches are based on feature extraction, edge detection or image segmentation, deformable template matching [8,9,5]. However, these techniques do not allow to include eyelid tracking in real time. 3D head and face tracking can be also handled by feature-based trackers with active appearance models [1]. These provide high accuracy but require high memory/time consumptions and depend on image quality, which hinder the development of real-time applications.

Alternatively, Appearance-Based Models (ABM) have been proposed as a powerful tool for analyzing facial images [2]. Deterministic and statistical appearance-based tracking methods have been proposed and used by some researchers. They handle successfully image variability and drifting problems, since they consider an input image through reduced version in order to apply estimation models, adopting deterministic velocity and registration techniques or statistics tools [6].

In order to track eyelid movements in real time, we improve the ABM framework in three directions. Firstly, we use ABM to track eyelid motions by on-line texture learning. Thus, our proposed approach infers the state of eyes without detecting eye features, such as irises and eye corners. Secondly, we show that by adopting a no self-occluded reference facial texture, it is possible to track eyelid motions. We prove that there is a dependency between the eye region and the rest of face, in order to obtain more accurate and stable 3D head pose estimations. Thirdly, by applying the two aforementioned contributions, we build a hierarchical tracker able to register eyelid movements by reducing estimation errors and adaptation rates.

The paper is organized as follows: in section 2, the deformable 3D facial model is described. In section 3, we will explain the background of adaptive appearance models. In section 4, we explain eyelid and face tracking, the relationship between face and eye region, and the structure of the hierarchical tracking. In section 5, experimental results on hierarchical adaptive appearance-based tracker are presented, which involve tracking in real-time of the 3D head pose and some facial actions, including eyelid movements. Finally, we conclude in Section 6.

2 Tracking with Adaptive Appearance Models

We use the 3D face Candide Model [3]. The Candide provides a simple process to construct an appearance model and a single parameterization to extract facial features. The shape can be described by the matrix $\mathbf{V}$:

$$\mathbf{V}_n = V_0 + \mathbf{D}\boldsymbol{\vartheta} + \mathbf{A}\boldsymbol{\gamma}, \tag{1}$$

where V_0 is the standard shape, $\mathbf{D}\boldsymbol{\vartheta}$ determines the biometry for each person, and $\mathbf{A}\boldsymbol{\gamma}$ distorts the mesh depending on the facial actions. Thus, we encode the tracking parameters by the vector $\mathbf{q}$, which contains two variables; $\boldsymbol{\alpha}$ denotes the head angles and position, and $\boldsymbol{\gamma}$ the facial actions considered for tracking:

$$\mathbf{q} = [\boldsymbol{\alpha}_i, \boldsymbol{\gamma}_j]^t, \ for\ i = 0, ..., 5\ and\ j = 0, ..., 6. \tag{2}$$

A shape-free texture represents a facial texture which is obtained by applying an image warping technique, $\boldsymbol{\chi}(\mathbf{q}) = \Psi(\mathbf{I}, \mathbf{q})$, in order to project an input image I_t onto the reference texture, where $\boldsymbol{\chi}(\mathbf{q})$ is the ABM of the image, see Fig. 1.

Face tracking consists of the estimation of the 3D head pose and several facial actions encoded by the vector $\mathbf{q}_t$, Eq. (2). In order to estimate the corresponding vector $\mathbf{q}_t$ at each frame, we obtain the ABM associated with the animation

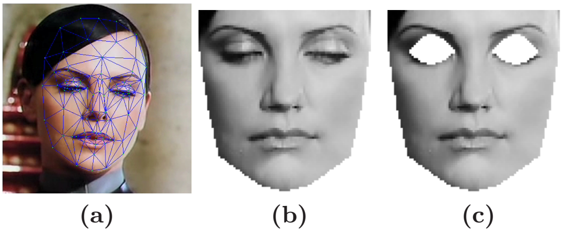


Fig. 1. **(a)**Input Image. **(b)** Texture with Eyes. **(c)** Texture without Eyes.

parameters $\boldsymbol{\gamma}_t$ through an appearance registration process. We represent as $\hat{\chi}_t$ the tracked parameters and the estimated textures. We suppose each appearance following a Gaussian distribution with mean μ_i and variance σ_i^2, which are vectors of n pixels. The probability for each observation is given by:

$$p(\mathbf{I}_t|\mathbf{q}_t) = p(\boldsymbol{\chi}_t|\mathbf{q}_t) = \prod_{i=0}^{n} N(\boldsymbol{\chi}_i;\mu_i,\sigma_i) \qquad (3)$$

When the appearance is tracked for the current image, μ_i and σ_i^2 are updated for the next frame by applying this technique, where ω is the updating factor:

$$\mu_{t+1} = \omega\mu_t + (1-\omega)\hat{\boldsymbol{\chi}}_t,\ and\ \sigma_{t+1}^2 = \omega\sigma_t^2 + (1-\omega)(\hat{\boldsymbol{\chi}}_t - \mu_t)^2, \qquad (4)$$

where μ_i and σ_i are initialized with the first patch $\boldsymbol{\chi}$.

In order to estimate the vector $\mathbf{q}_t$ for the next frame, an adaptive velocity model is adopted, where the adaptive motion velocity is predicted using a fixed function to estimate the transition state from the previous frame:

$$\mathbf{q}_t = \hat{\mathbf{q}}_{t-1} + \Delta\mathbf{q}_t \qquad (5)$$

where $\Delta\mathbf{q}_t$ is the shift of geometric parameters. The current input image $\mathbf{I}_t$ is registered with the current appearance model by minimizing the Mahalanobis distance between the warped texture and the current average of appearance. Here, the appearance parameters μ and σ are known, and the distance is minimized by an iterative first-order linear approximation. As a result, given Eq. (5), the warped texture will be the closest to the average of appearance,

$$\Psi(\mathbf{I}_t,\mathbf{q}_t) \quad \approx \quad \mu_t \qquad (6)$$

Approximating Eq. (6) via a first-order Taylor series expansion around $\hat{\boldsymbol{q}}_t$ and using the Vanilla gradient descent method [7], we have:

$$\Psi(\mathbf{I}_t,\mathbf{q}_t) \quad \approx \Psi(\mathbf{I}_t,\hat{\mathbf{q}}_{t-1}) + \frac{\partial(\boldsymbol{\chi}_t,\mathbf{q}_t)}{\boldsymbol{q_t}}(\mathbf{q}_t - \hat{\mathbf{q}}_{t-1}) \qquad (7)$$

Thus, the increment of vector $\mathbf{q}_t$ is related to the change of the previous adapted shape, and the current average appearance. The solution for this vector is found iteratively until the error measure for the distance is minimum. The gradient matrix is computed by partial differences and the Vanilla gradient descendent method is able to accommodate appearance changes while achieving precise estimation of the gradient matrix.

3 Eyelid and Face Tracking Problem

As said before on the introduction, we solve the problems from previous works, by registering eyelid tracking over time for any kind of blinking, different speed, open, intermediate and closed states. In order to avoid high memory and time consumptions, we do not use low image processes like contours detectors, colour information and so on. We adopt a generic Candide model, to learn the textures on-line, and deal with eyelid facial action [1]. We improve the gradient descent method while hierarchically combining two appearance models. Illumination changes, occlusions and fast movements are considered as outliers by constraining the gradient descent for each component of $\varDelta\mathbf{q}$ with the Huber's function. The Huber's function, $\hat{\xi}$ function is defined as:

$$\eta(y) = \frac{1}{y}\frac{d\hat{\xi}(y)}{dy} = \begin{cases} 1 & \mathbf{if}|y| \leq c \\ \frac{c}{|y|} & \mathbf{if}|y| > c \end{cases} \tag{8}$$

where y is the value of a pixel in the patch $\boldsymbol{\chi}_t$ normalized by the mean μ_i and the variance σ_i^2 of the appearance at the same pixel. The constant c is set $3 * \sigma$. This restriction controls the registration of the $\boldsymbol{\chi}_t$, on the probabilistic model.

We have proven the influence of the eyes on head pose estimations, using a reference texture with size of 84 x 80 pixels. There are 5447 pixels by each shape-free texture, and the eye region is represented by only 380 pixels in the warped version of the image (7% of the patch with size 22x14 pixels) [4].

In order to achieve a robust and stable eyelid tracking we construct two different trackers. On the one hand, the first one for adapting head, eyebrows, lips and eyelids. These have thirteen parameters and a texture model with eyelid region but without sclera and iris pixels. On the other hand, the second one is a reference texture without eye region to correct head pose estimations of the first tracker. Therefore, the first tracker has one additional parameter, γ^6. The description of both trackers is detailed next, see Fig. 2.

3.1 Characteristics of Both Facial Trackers

Tracker 1: in order to estimate head, lips, eyebrows and eyelids we consider the 13 parameters of vector $\mathbf{q}$, the Eq. (2).

1. Given an input image $\mathbf{I}_t$, we obtain the warped version of the image, $\boldsymbol{\chi} = \varPsi(\mathbf{I}_t, \mathbf{q})$. This is the ABM for this tracker. We obtain μ_t and σ_t^2 of the face from previous adapted frames by applying the recursive filtering technique, Eq. (4). The ω updating factor incorporates new information depending on the velocity of facial movements.

2. The gradient matrix is computed for $[\boldsymbol{\alpha}_i, \gamma_0, ..., \gamma_5]$ by partial differences, considering an increment step δ and a small perturbation range,. On the other hand, the gradient matrix component for eyelids, γ_6 is estimated out using a biggest descent step, $k * \delta$, and a perturbation range which covers the overall FAP, since it is necessary to search the minimum error in the complete range, due to the velocity of the blinking.
3. Given the previous adaptation, $\hat{\boldsymbol{\chi}}_t$, we look for the best estimation in the face space, where the reference texture includes the eye region, comparing the average and likelihood shapes through Mahalanobis distance. We include a backward-forward searching factor for γ_6, exploitation rather than exploration. The convergence is quickly achieved for all parameters of $\mathbf{q} = [\boldsymbol{\alpha}_i, \gamma_0, ..., \gamma_5, \gamma_6]^t$, while avoiding local minimums.

Tracker 2: In order to correct the head pose estimation of the *tracker 1*, we exclude the eye regions.

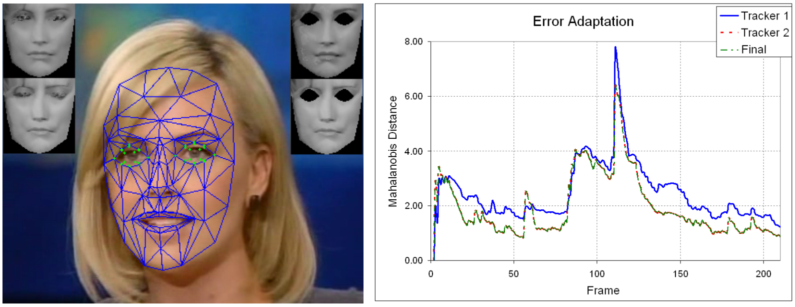
1. We obtain a warped image, $\boldsymbol{\chi}' = \mathit{\Psi}(\mathbf{I}_t, \mathbf{q})$ with a different reference texture, which excludes the eyes. We have kept more information about previous frames than *tracker 1* by using the recursive filtering with an updating factor $\beta < \omega$, learning lately the texture model, which has skill to handle outliers.
2. The gradient matrix is estimated for $[\boldsymbol{\alpha}_i, \gamma_0, ..., \gamma_5]^t$.
3. Given the previous estimation from *tracker 1*, $\hat{\boldsymbol{\chi}'}$, we measure the distance between the observation likelihood and the current average shape. The best adaptation without the pixels of the eye region is achieved with the convergence by using the same exploration range for all parameters.

3.2 Combination of Eyelid and Face Tracking

Both trackers were tested without any a priori information, thus obtaining inaccurate results because the first tracker needs the uncertainty of the eye region, which the second tracker avoids. Consequently, the first tracker generates rough estimations while the second tracker further refines them. In order to keep the best estimations and to avoid the additional errors, we combine the trackers as follows.

We first begin estimating $[\boldsymbol{\alpha}_i, \gamma_0, ..., \gamma_5, \gamma_6]^t$ with of the *tracker 1* by obtaining a rough estimation for the current image, $p(\boldsymbol{\chi}_t|\mathbf{q}_t)$. Subsequently, the *tracker 2* estimates the vector $[\boldsymbol{\alpha}_i, \gamma_0, ..., \gamma_5]^t$ for the current image, while correcting the previous adaptation from the *tracker 1*, $p(\chi'_t|\mathbf{q}_t)$. Therefore, *Tracker 1* is able to estimate the eyelid tracking and the second one is able to correct the head pose estimations. We constraint the second tracker to modify the vector $\mathbf{q}$ only if the previous estimations are improved. Consequently, *tracker 2* estimation is made according to $p(\chi'|\mathbf{q}_t, M)$, where M is the minimum Mahalanobis distance of the *tracker 1*. The geometrical vector $\mathbf{q}$ is modified only if the Mahalanobis distance is lower for the *tracker 2*.

The result from this hierarchical tracking is a robust and accurate estimation for vector $\mathbf{q}_t$. The minimum distance for both of trackers reduces the adaptation



(a) **(b)**

Fig. 2. Input image of a sequence composed of 210 frames tested with both ground textures (84x80 and 42x40 pixels, corresponding to 21x5 and 11x4 pixels on eye region). **(a)** Left and right top on the image are the two trackers, the estimated and current textures. **(b)** The error adaptation for two Trackers.

error for all parameters and improves the eyelid adaptation. We do not only detect opened and closed eye states, also continuous eyelid movements with a correct adaptation.

4 Experimental Results

The tracking experiments were run in a 3.2 GHz Pentium PC, without optimized C code. It had a performance of 500 MB as RAM memory consumption and 21 average fps for each tracker, using a big mask (84x80). The real time is achieved using a small mask (42x40) with 200 MB and 67 fps for memory and time consumption.

The following experimental results were obtained by using the FGnet[1] database for face and expression recognition, as well as low-resolution video clips. There are videos modified by image compressors or published in web format, so that the hard task of the trackers is to deal with low image resolution, lack of continuity, and illumination changes. In these movie videos, the face and eyelid movements are natural and spontaneous, with large angle variations. In fact, scale variations are found related to in-depth movements or camera zooming.

In the first sequence the size of the input image is 320x240 pixels, the face size is 100x130 pixels Fig 3. It begins from closed eye position describing a simple face expression with two eye blinks at 22nd and 151st frames. We show how our method is robust with standard databases, movies, video clips, etc. See Fig. 3. We show an image sequence where the camera is static and in frontal position, Fig. 4. It has 585 images; the face size is 200x220 pixels. There is in-depth motion of head and several zoom effects.

As a result, a robust and accurate learning is achieved, see Fig. 5. On the one hand, Tracker 1 learns faster and therefore includes quickly early eye blinks into

[1] ©FGnet - IST-2000-26434. http://www-prima.inrialpes.fr/FGnet/

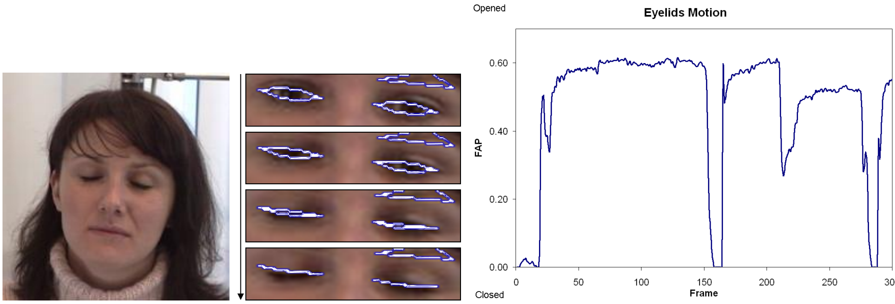


Fig. 3. Eyelid Trackin performance with images of FGnet database

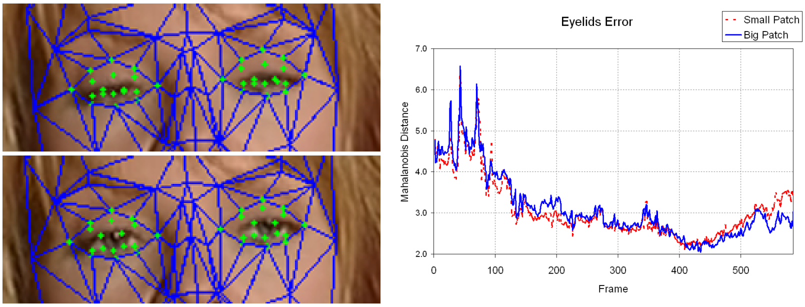


Fig. 4. Eyelid motion comparison between small and big textures

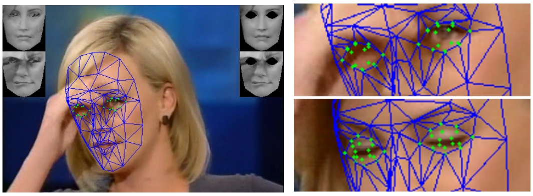


Fig. 5. 111 frame being occludes onto eyes region

appearance model. On the other hand, Tracker 2 learns slowly, in order to keep the previous adaptation whenever occlusions are occurring. In this paper, every sequence has been tested with both of the reference textures and the results are quite similar.

5 Conclusions

Our framework extends the application of ABMs to obtain continuous eyelid tracking, thereby proving the capability for estimating accurately and in real-time the 3D head pose, eyebrows, lips and eyelids. The assumption of this approach is to model an appearance like a Gaussian for a subsequent gradient descent analysis in multidimensional spaces by applying Gauss-Newton method. With respect to previous works, our framework has important contributions. Firstly, we blend the 3D head pose, eyebrows, lips and eyelid tracking, using deformable models. We propose to join two appearance-based trackers with different reference textures, face spaces and learning process. Secondly, we avoid any colour information of the image with respect to colour spaces or edge detectors. We recover the mesh parameters using statistics and estimation tools on finite series. Thirdly, we propose a hierarchical tracking with its respective face space to refine the estimations about the input image, to allow the use of small and big reference textures, in order to reduce memory and time consumption.

We are now working on automatic initialization of tracking system, by implementing a face and feature detector for a future inference of initial geometrical parameters. We are also going to track the iris motion for gaze analysis, that is the first step toward real facial expression and sincerity analysis on human faces.

Acknowledgement. This work has been supported by the EC grant IST-027110 for the Hermes Project and by the Spanish MEC under projects TIN2006-14606 and DPI-2004-5414. J. Gonzàlez acknowledges the support of a Juan de la Cierva Postdoctoral fellowship from the Spanish MEC.

References

1. Ahlberg, J.: An active model for facial feature tracking. EURASIP Journal on Applied Signal Processing 2002(6), 566–571 (2002)
2. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models - their training and application. Computer Vision and Image Understanding 61(1), 39–59 (1995)
3. Dornaika, F., Ahlberg, J.: Face model adaptation for tracking and active appearance model training. British Machine Vision Conference (2003)
4. Dornaika, F., Orozco, J., González, J.: Combined Head, Lips, Eyebrows,and Eyelids Tracking using Adaptive Appearance Models. In: Perales, F.J., Fisher, R.B. (eds.) AMDO 2006, LNCS, vol. 4069, pp. 110–119, Springer, Heidelberg (2006)
5. Zhang, Y., Tan, H.: Detecting eye blink states by tracking iris and eyelids. Pattern Recognition Letters, 2005 (2005)
6. Sclaroff, S., Cascia, M.L., Athitsos, V.: Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(4), 322–336 (2000)
7. Nocedal, J., Wright, S.: Numerical optimization. Springer, Heidelberg (1999)
8. Cohn, J.F., Tian, Y., Kanade, T.: Dual-state parametric eye tracking. In: International Conference on Automatic Face and Gesture Recognition (2000)
9. Liu, H., Wu, Y., Zha, H.: A new method of detecting human eyelids based on deformable templates. Systems, Man and Cybernetics, 2004 IEEE International Conference, vol. 1, pp. 604–609 (2004)