Combined Head, Lips, Eyebrows, and Eyelids Tracking Using Adaptive Appearance Models[*]

Fadi Dornaika, Javier Orozco, and Jordi Gonzàlez

Computer Vision Center
Edifici O, Campus UAB
08193 Bellaterra, Barcelona, Spain
{dornaika, orozco, gonzalez}@cvc.uab.es

Abstract. The ability to detect and track human heads and faces in video sequences is useful in a great number of applications, such as human-computer interaction and gesture recognition. Recently, we have proposed a real-time tracker that simultaneously tracks the 3D head pose and facial actions associated with the lips and the eyebrows in monocular video sequences. The developed approach relies on Online Appearance Models where the facial texture is learned during the tracking. This paper extends our previous work in two directions. First, we show that by adopting a non-occluded facial texture model more accurate and stable 3D head pose parameters can be obtained. Second, unlike previous approaches to eyelid tracking, we show that the Online Appearance Models can be used for this purpose. Neither color information nor intensity edges are used by our proposed approach. Moreover, our eyelids tracking does not rely on any eye feature extraction which may lead to erroneous results whenever the eye feature detector fails. Experiments on real videos show the feasibility and usefulness of the proposed approach.

1 Introduction

The ability to detect and track human heads and facial features in video sequences is useful in a great number of applications, such as human-computer interaction and gesture recognition. Vision-based tracker systems provide an attractive alternative since vision sensors are not invasive. Of particular interest are vision-based markerless head and/or face trackers. Since these trackers do not require any artificial markers to be placed on the face, comfortable and natural motions can be achieved. On the other hand, building robust and real-time markerless trackers for head and facial features is a difficult task due to the high variability of the face and the facial features in videos.

To overcome the problem of appearance changes recent works on faces adopted statistical facial textures. For example, the Active Appearance Models have been proposed as a powerful tool for analyzing facial images [1]. Deterministic and

[*] This work was supported by the Government of Spain under the CICYT project TIN2005-09026 and The Ramón y Cajal Program.

F.J. Perales and R.B. Fisher (Eds.): AMDO 2006, LNCS 4069, pp. 110–119, 2006.

statistical appearance-based tracking methods have been proposed and used by some researchers [2,3,4]. These methods can successfully tackle the image variability and drift problems by using deterministic or statistical models for the global appearance of a special object class: the face. A few algorithms exist which attempt to track both the head and the facial features in real time, e.g. [3] and [4]. These works have addressed the combined head and facial feature tracking using the Active Appearance Models principles. However, [3] and [4] require tedious learning stages that should be performed beforehand and should be repeated whenever the imaging conditions change. Recently, we have developed a head and facial feature tracking method based on Online Appearance Models (OAMs) [5]. Unlike the Active Appearance Models, the OAMs offer a lot of flexibility and efficiency since they do not require any facial texture model that should be computed beforehand. Instead the texture model is built online from the tracked sequence.

This paper extends a previous work [5] in two directions. First, we show that by adopting a non-occluded shape-free facial texture that excludes the eyes region more accurate and stable 3D head pose parameters can be obtained. Second, unlike feature-based eyelid trackers, we show that the Online Appearance Models can be used to track the eyelids. Thus, we can infer the eye state without detecting the eye features such as the irises and the eye corners.

Tracking the eyelids and the irises can be used in many applications such as drowsiness detection and interfaces for handicapped individuals. Detecting and tracking the eye and its features has been addressed by many researchers. A variety of methodologies have been applied to the problem of eye tracking. There are many methods for detecting eye features such as eye corners, irises, and eyelids [6,7,8,9]. However, most of the proposed approaches rely on intensity edges and are time consuming. In [8], detecting the state of the eye is based on the iris detection in the sense that the iris detection results will directly decide the state of the eye. In [6], the eyelid state is inferred from the relative distance between the eyelid apex and the iris center. For each frame in the video, the eyelid contour is detected using edge pixels and normal flow. The authors reported that when the eyes were fully or partially open, the eyelids were successfully located and tracked 90% of the time. Their proposed approach depends heavily on the extracted intensity edges. Moreover, it assumes high resolution images depicting an essentially frontal face. In our study, we do not use any edges and there is no assumption on the head pose. In our work, the eyelid motion is inferred at the same time with the 3D head pose and other facial actions, that is, the eyelid state does not rely on the detection results of other features such as the eye corners and irises. Tracking the rapid eyelid motion is not a straightforward task. In our case, we like to track the eyelid motion using the principles of OAMs. The challenges are as follows. First, the upper eyelid is a highly deformable facial feature since it has a great freedom of motion. Second, the eyelid can completely occludes the iris and sclera, that is, a facial texture model will have two different appearances at the same locations. Third, the eyelid motion is very fast compared to the motion of other facial features.

The remainder of this paper proceeds as follows. Section 2 introduces our deformable 3D facial model. Section 3 states the problem we are focusing on, and describes the online adaptive appearance model. Section 4 summarizes the adaptive appearance-based tracker that tracks in real-time the 3D head pose and some facial actions. It gives some comparisons obtained with different facial texture models. In Section 5, we present some tracking results associated with the head, lips, eyebrows and eyelids.

2 Modeling Faces

A deformable 3D model. In our study, we use the 3D face model *Candide* [10]. This 3D deformable wireframe model was first developed for the purpose of model-based image coding and computer animation. The 3D shape of this wireframe model is directly recorded in coordinate form. It is given by the coordinates of the 3D vertices $\mathbf{P}_i, i = 1, \ldots, n$ where n is the number of vertices. Thus, the shape up to a global scale can be fully described by the $3n$-vector $\mathbf{g}$; the concatenation of the 3D coordinates of all vertices $\mathbf{P}_i$. The vector $\mathbf{g}$ is written as:

$$\mathbf{g} = \mathbf{g}_s + \mathbf{A}\, \boldsymbol{\tau_a} \tag{1}$$

where $\mathbf{g}_s$ is the static shape of the model, $\boldsymbol{\tau_a}$ the animation control vector, and the columns of $\mathbf{A}$ are the Animation Units. In this study, we use seven modes for the facial Animation Units (AUs) matrix $\mathbf{A}$. We have chosen the seven following AUs: lower lip depressor, lip stretcher, lip corner depressor, upper lip raiser, eyebrow lowerer, outer eyebrow raiser and eyelid lowerer. These AUs are enough to cover most common facial animations. Moreover, they are essential for conveying emotions. Thus, the lips are controlled by four parameters, the eyebrows are controlled by two parameters, and the eyelids by one parameter.

In equation (1), the 3D shape is expressed in a local coordinate system. However, one should relate the 3D coordinates to the image coordinate system. To this end, we adopt the weak perspective projection model. We neglect the perspective effects since the depth variation of the face can be considered as small compared to its absolute depth. Thus, the state of the 3D wireframe model is given by the 3D head pose parameters (three rotations and three translations) and the internal face animation control vector $\boldsymbol{\tau_a}$. This is given by the 13-dimensional vector $\mathbf{b}$:

$$\mathbf{b} = \, [\theta_x, \, \theta_y, \, \theta_z, \, t_x, \, t_y, \, t_z, \, {\boldsymbol{\tau_a}}^T \,]^T \tag{2}$$

Shape-free facial textures. A face texture is represented as a shape-free texture (geometrically normalized image). The geometry of this image is obtained by projecting the static shape $\mathbf{g}_s$ (neutral shape) using a centered frontal 3D pose onto an image with a given resolution. The texture of this geometrically normalized image is obtained by texture mapping from the triangular 2D mesh in the input image (see figure 1) using a piece-wise affine transform, $\mathcal{W}$ (see [10] for more details). The warping process applied to an input image $\mathbf{y}$ is denoted by:

$$\mathbf{x}(\mathbf{b}) = \mathcal{W}(\mathbf{y}, \mathbf{b}) \tag{3}$$

where $\mathbf{x}$ denotes the shape-free texture and $\mathbf{b}$ denotes the geometrical parameters. Several resolution levels can be chosen for the shape-free textures. The reported results are obtained with a shape-free patch of 5392 pixels. Regarding photometric transformations, a zero-mean unit-variance normalization is used to partially compensate for contrast variations. The complete image transformation is implemented as follows: (i) transfer the texture $\mathbf{y}$ using the piece-wise affine transform associated with the vector $\mathbf{b}$, and (ii) perform the grey-level normalization of the obtained patch. Figure 1 illustrates two shape-free patches associated with an input image.

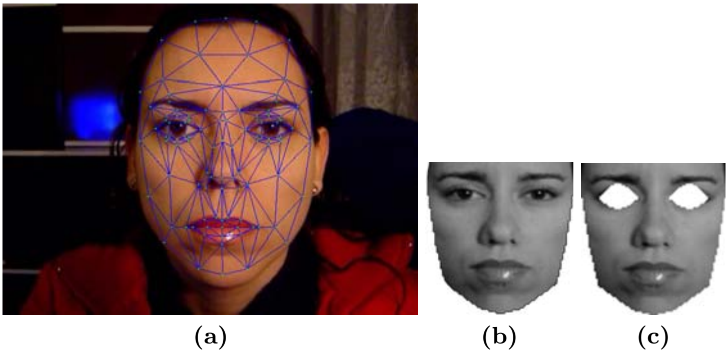


Fig. 1. (a) an input image with correct adaptation. **(b)** the corresponding shape-free facial image. **(c)** the same patch without the eyes region.

3 Problem Formulation and Adaptive Appearance Models

Given a video sequence depicting a moving head/face, we would like to recover, for each frame, the 3D head pose and the facial actions encoded by the control vector $\boldsymbol{\tau_a}$. In other words, we would like to estimate the vector $\mathbf{b}_t$ (2) at time t given all the observed data until time t, denoted $\mathbf{y}_{1:t} \equiv \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$. In a tracking context, the model parameters associated with the current frame will be handed over to the next frame.

For each input frame $\mathbf{y}_t$, the observation is simply the warped texture patch (the shape-free patch) associated with the geometric parameters $\mathbf{b}_t$. We use the HAT symbol for the tracked parameters and textures. For a given frame t, $\hat{\mathbf{b}}_t$ represents the computed geometric parameters and $\hat{\mathbf{x}}_t$ the corresponding shape-free patch, that is,

$$\hat{\mathbf{x}}_t = \mathbf{x}(\hat{\mathbf{b}}_t) = \mathcal{W}(\mathbf{y}_t, \hat{\mathbf{b}}_t) \tag{4}$$

The estimation of the current parameters $\hat{\mathbf{b}}_t$ from the previous ones $\hat{\mathbf{b}}_{t-1}$ and from the sequence of images will be presented in Section 4. In our work, the initial parameters $\hat{\mathbf{b}}_1$ corresponding to the first frame are manually provided. The automatic initialization can be obtained using the statistical technique proposed in [3].

By assuming that the pixels within the shape-free patch are independent, we can model the appearance of the shape-free facial patch using a multivariate Gaussian with a diagonal covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{\mu}$ be the Gaussian center and $\boldsymbol{\sigma}$ the vector containing the square root of the diagonal elements of the covariance matrix $\boldsymbol{\Sigma}$. $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are d-vectors (d is the size of $\mathbf{x}$) representing the appearance parameters. In summary, the observation likelihood at time t is written as

$$p(\mathbf{y}_t|\mathbf{b}_t) = p(\mathbf{x}_t|\mathbf{b}_t) = \prod_{i=1}^{d} \mathbf{N}(x_i;\mu_i,\sigma_i)_t \tag{5}$$

where $\mathbf{N}(x_i;\mu_i,\sigma_i)$ is a normal density:

$$\mathbf{N}(x_i;\mu_i,\sigma_i) = (2\pi\sigma_i^2)^{-1/2}\,\exp\left[-\rho\left(\frac{x_i-\mu_i}{\sigma_i}\right)\right],\quad \rho(x)=\frac{1}{2}\,x^2 \tag{6}$$

We assume that the appearance model summarizes the past observations under an exponential envelope, that is, the past observations are exponentially forgotten with respect to the current texture. When the appearance is tracked for the current input image, *i.e.* the texture $\hat{\mathbf{x}}_t$ is available, we can update the appearance and use it to track in the next frame. It can be shown that the appearance model parameters, *i.e.*, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ can be updated using the following equations (see [11] for more details on Online Appearance Models):

$$\mu_{i_{(t+1)}} = (1-\alpha)\,\mu_{i_{(t)}} + \alpha\,\hat{x}_{i_{(t)}} \tag{7}$$

$$\sigma^2_{i_{(t+1)}} = (1-\alpha)\,\sigma^2_{i_{(t)}} + \alpha\,(\hat{x}_{i_{(t)}} - \mu_{i_{(t)}})^2 \tag{8}$$

In the above equations, the subscript i denotes a pixel in the patch $\hat{\mathbf{x}}$. This technique, also called recursive filtering, is simple, time-efficient and therefore, suitable for real-time applications. The appearance parameters reflect the most recent observations within a roughly $L = 1/\alpha$ window with exponential decay.

Note that $\boldsymbol{\mu}$ is initialized with the first patch $\hat{\mathbf{x}}_1$ corresponding to the geometrical parameters $\hat{\mathbf{b}}_1$. However, equation (8) is not used until the number of frames reaches a given value (*e.g.*, the first 40 frames). For these frames, the classical variance is used, that is, equation (8) is used with α being set to $\frac{1}{t}$.

4 Tracking Using Adaptive Appearance Registration

We consider the state vector $\mathbf{b} = [\theta_x, \theta_y, \theta_z, t_x, t_y, t_z, \boldsymbol{\tau_a}^T]^T$ encapsulating the 3D head pose and the facial actions. In this section, we will show how this state can be recovered for time t using the previous known state $\hat{\mathbf{b}}_{t-1}$, the current input image $\mathbf{y}_t$, and the current appearance parameters. The vector $\boldsymbol{\tau_a}$ may have 6 facial actions (lips and eyebrows) or 7 facial actions (lips, eyebrows, and eyelids).

The sought geometrical parameters $\mathbf{b}_t$ at time t are related to the previous parameters by the following equation ($\hat{\mathbf{b}}_{t-1}$ is known):

$$\mathbf{b}_t = \hat{\mathbf{b}}_{t-1} + \Delta\mathbf{b}_t \tag{9}$$

where $\Delta\mathbf{b}_t$ is the unknown shift in the geometric parameters. This shift is estimated using a region-based registration technique that does not need any image feature extraction. In other words, $\Delta\mathbf{b}_t$ is estimated such that the warped texture will be as close as possible to the facial appearance given by the Gaussian parameters. For this purpose, we minimize the *Mahalanobis* distance between the warped texture and the current appearance mean,

$$\min_{\mathbf{b}_t} e(\mathbf{b}_t) = \min_{\mathbf{b}_t} D(\mathbf{x}(\mathbf{b}_t), \boldsymbol{\mu}_t) = \sum_{i=1}^{d} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \tag{10}$$

The above criterion can be minimized using iterative first-order linear approximation which is equivalent to a Gauss-Newton method. It is worthwhile noting that minimizing the above criterion is equivalent to maximizing the likelihood measure given by (5). Moreover, the above optimization is made robust by using robust statistics [5]. In the above optimization, the gradient matrix $\frac{\partial \mathcal{W}(\mathbf{y}_t, \mathbf{b}_t)}{\partial \mathbf{b}} = \frac{\partial \mathbf{x}_t}{\partial \mathbf{b}}$ is approximated by numerical differences. More details about this optimization technique can be found in [5].

On a 3.2 GHz PC, a non-optimized C code of the approach computes the 3D head pose and the seven facial actions in 70 ms.

5 Tracking Comparisons

In this Section, we compare the 3D head pose estimates obtained with different shape-free patches using the same robust optimization technique described above. To this end, we use the two shape-free patches depicted in Figure 1.**(b)** and 1.**(c)**. Note that the second patch is obtained from the first one by removing the eyes region. We assume that the state vector **b** is given by the six head pose parameters and the six facial actions associated with the lips and eyebrows.

We have used a 1000-frame long sequence featuring a talking subject[1] as a test video. Note that talking is a spontaneous activity. Figure 2 illustrates the estimates of the 3D head pose parameters associated with a 150-frame long segment using the two different shape-free facial patches (this segment starts at frame 500). This video segment contains three blinks at frames 10, 104, and 145. As can be seen, the most significant deviations in the 3D head pose parameters occur at those frames (e.g., see the scale plot). Whenever eye blinking occurs the patch without the eyes region has provided more accurate and stable parameters than the patch with the eyes region. This is explained by the fact that despite the use of robust statistics the estimation of the 3D head pose with a texture model

[1] http://www-prima.inrialpes.fr/FGnet/html/benchmarks.html

containing the eyes region (sclera and iris) is affected by the eyelids motion. One can notice that the rotational deviations/errors seem to be small. However, the vertical and in-depth translation errors can be large. For example, at frame 145 the obtained scale deviation/error is about 0.025 which corresponds to an in-depth error of about 3 centimeters[2].

6 Head, Lips, Eyebrows, and Eyelids Tracking

In the previous Section, we have shown that the accuracy of the 3D head pose can be affected by the eyelids motion/blinking if the sclera and iris region is included in the texture model. This is not surprising since eye blinking corresponds to a sudden occlusion of a small part of the face. Thus, if the eyelids motion is tracked one can expect that the 3D head pose parameters can be more stable. Also, we have shown that the estimated 12 degrees of freedom associated with the head, lips ad eyebrows together with the used deformable 3D model are enough to track the eye boundaries in a video sequence. However, one needs to do more to track the eyelids motion. As we have mentioned earlier, tracking the eyelids motion is a very challenging task, and most of the proposed approaches for locating and tracking the eyelids rely on the extracted intensity edges.

To tackle the difficulties associated with the eyelids motion, we use the following. First, we adopt a shape-free facial texture model whose eyes region corresponds to closed eyes configuration (see Figure 3), which implicitly excludes the iris and sclera regions. Second, we use the same registration technique described in Section 4 where the facial action vector $\boldsymbol{\tau_a}$ is now given by 7 facial actions (lips, eyebrows, and eyelids). Note that when the eyes are open in the input image, the shape-free texture corresponding to the eyelids region (associated with a correct eyelid facial action) will be a distorted version of a very small area in the input image. However, the global appearance of the eyelid is still preserved since the eyelids have the skin appearance.

We have tracked the head, lips, eyebrows, and eyelid using the 1000-frame long sequence. Figure 4 displays the tracking results (13 degrees of freedom) associated with frames 280, 284, and 975. The middle displays zoomed views of those frames. Notice how the eyelids are correctly tracked. The upper left corner of each image shows the current appearance ($\boldsymbol{\mu_t}$) and the current shape-free texture ($\hat{\mathbf{x}}_t$). The bottom of this figure displays the estimated eyelid facial action as a function of time where the zero value corresponds to a closed eyelid and the one value to a wide open eyelid. Eye blinking is a discrete and important facial action [12,13]. In our case, it can be directly detected and segmented by thresholding the continuous eyelid facial action. As can be seen, the dual state of the eye can easily be inferred from the continuous curve. For the tracked sequence, all blinks are correctly detected and segmented.

Figure 5 displays the tracking results obtained with another two videos.

[2] The exact value depends on the camera intrinsic parameters and the absolute depth.

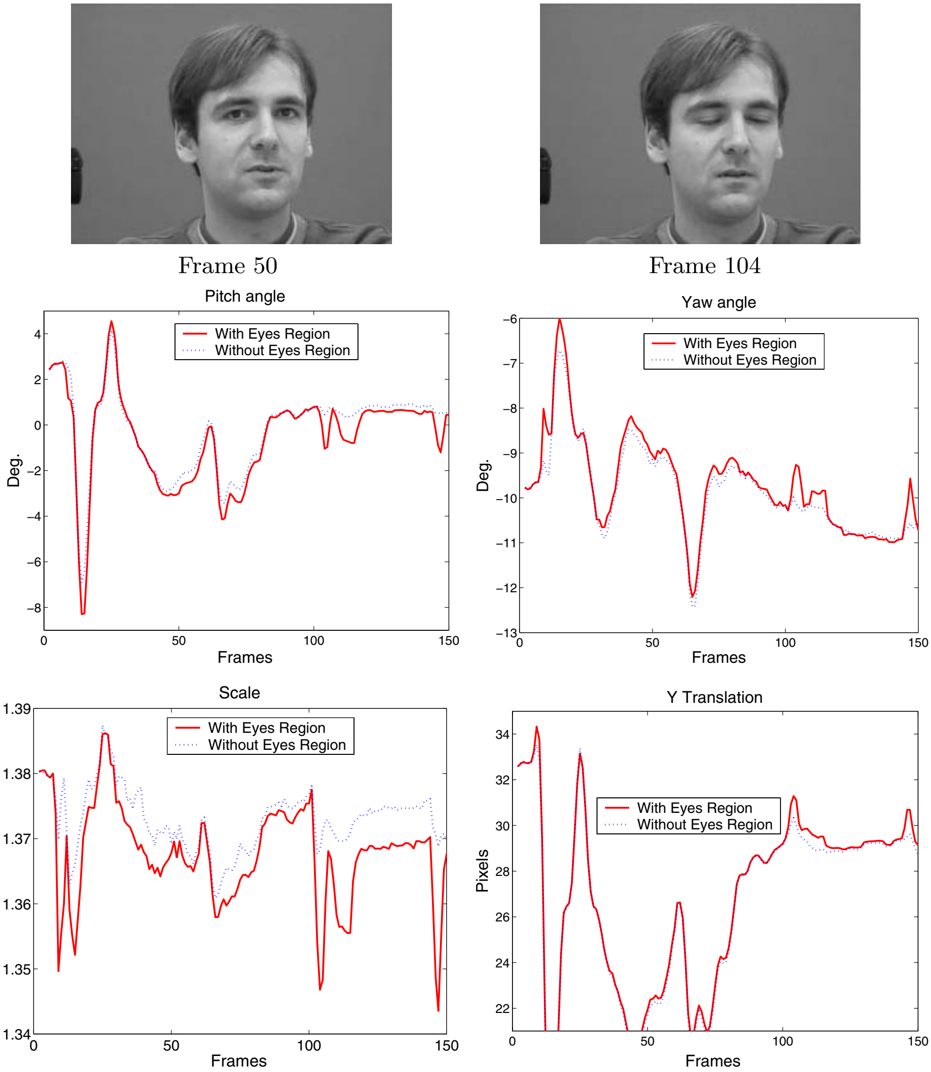


Fig. 2. 3D head pose parameters obtained with two different facial patches that differs by the eyes region

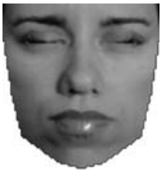
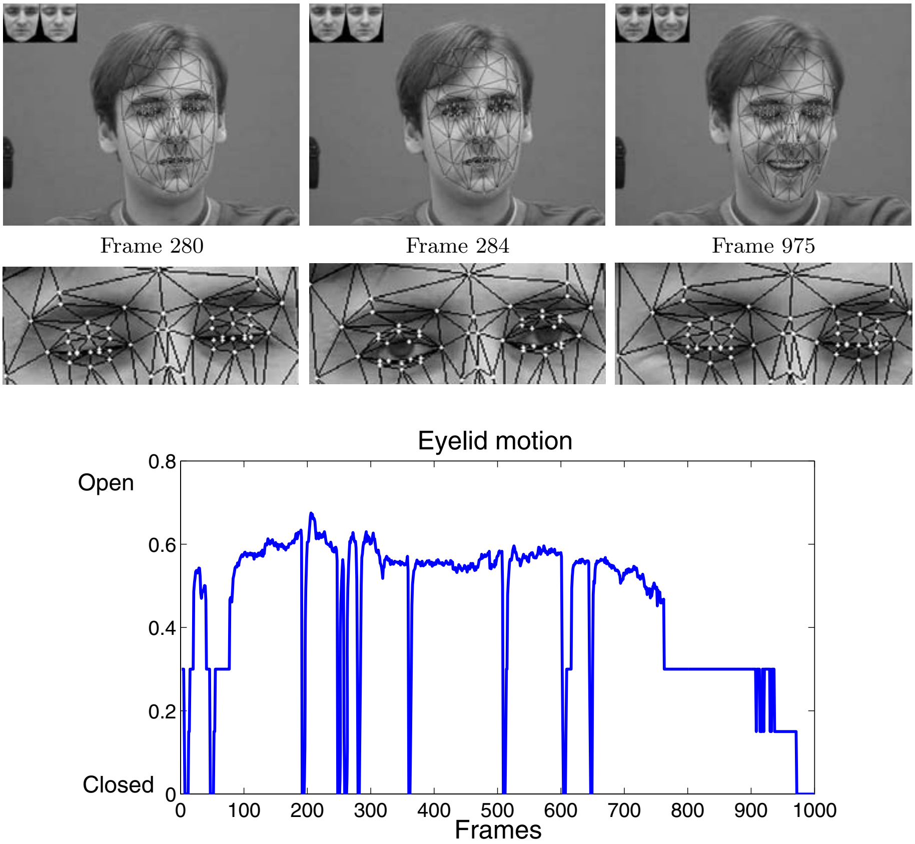


Fig. 3. The shape-free texture used to track 13 degrees of freedom including the eyelid motion


Frame 280

Frame 284

Frame 975

Fig. 4. Tracking the 3D head pose, the lips, the eyebrows and the eyelids associated with a 1000-frame long sequence. Only frames 280, 284, 975 are shown. The plot depicts the estimated eyelid facial action as a function of time.

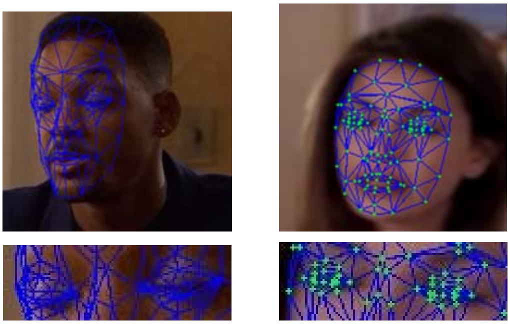


Fig. 5. Two test sequences

7 Conclusion

In this paper, we have extended our appearance-based 3D head and facial action tracker to deal with eyelid motions. The 3D head pose and the facial actions associated with the lips, eyebrows, and eyelids are simultaneously estimated in real-time using Online Appearance Models. Compared to other eyelid tracking techniques our proposed approach has several advantages. First, computing and segmenting intensity edges has been avoided. Second, the eyelid is tracked with other facial actions at the same time, and hence it does not depend on the detection of other eye features. Third, the eyelid motion is tracked using a continuous facial action. Experiments on real video sequences indicate that the eye state can be detected using the eyelid tracking results.

References

1. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(6) (2001) 681–684
2. Cascia, M., Sclaroff, S., 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) (2000) 322–336
3. Ahlberg, J.: An active model for facial feature tracking. EURASIP Journal on Applied Signal Processing **2002**(6) (2002) 566–571
4. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision **60**(2) (2004) 135–164
5. Dornaika, F., Davoine, F.: On appearance based face and facial action tracking. IEEE Transactions on Circuits and Systems for Video Technology (In press)
6. Sirohey, S., Rosenfeld, A., Duric, Z.: A method of detecting and tracking irises and eyelids in video. Pattern Recognition **35**(6) (2002) 1389–1401
7. Liu, H., Wu, Y., Zha., H.: Eye states detection from color facial image sequence. In: SPIE International Conference on Image and Graphics, vol. 4875. (2002) 693–698
8. Tian, Y., Kanade, T., Cohn, J.F.: Dual-state parametric eye tracking. In: International Conference on Automatic Face and Gesture Recognition. (2000)
9. Zhu, J., Yang, J.: Subpixel eye gaze tracking. In: International Conference on Automatic Face and Gesture Recognition. (2002)
10. Ahlberg, J.: Model-based coding: Extraction, coding, and evaluation of face model parameters. PhD thesis, No. 761, Linköping University, Sweden (2002)
11. Jepson, A., Fleet, D., El-Maraghi, T.: Robust online appearance models for visual tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(10) (2003) 1296–1311
12. Grauman, K., Betke, M., Gips, J., Bradski, G.R.: Communication via eye blinks - Detection and duration analysis in real time. In: International Conference on Computer Vision and Pattern Recognition. (2001)
13. Moriyama, T., Kanade, T., Cohn, J., Xiao, J., Ambadar, Z., Gao, J., Imamura, H.: Automatic recognition of eye blinking in spontaneously occuring behavior. In: International Conference on Pattern Recognition. (2002)