A STOCHASTIC APPROACH FOR APPEARANCE-BASED 3D FACE TRACKING

F. Dornaika, F. Davoine, and M. Dang

CNRS HEUDIASYC
Compiègne University of Technology
60205 Compiègne Cedex, FRANCE
{*dornaika, fdavoine, dang*}*@hds.utc.fr*

ABSTRACT

We propose a framework that combines active appearance models and a particle filter-like step to realize robust face and facial feature tracking. The developed framework aims at retaining the strengths of deterministic and stochastic optimization techniques with a statistical texture model. Adapting each frame consists of three steps. First, the adaptation utilizes a gradient descent algorithm based on the reconstruction error (Distance-from-feature-space). Second, a stochastic diffusion of the solution is then performed and the Maximum Likelihood solution is chosen (*Distance-in-feature-space* and *Distance-from-feature-space*). Third, the ML solution is refined using a gradient descent algorithm. Comparisons with a simple gradient descent method demonstrate the efficiency and robustness of the developed framework. The resulting robust 3D face tracker can find applications in many fields especially in the fields of human-computer interaction.

1. INTRODUCTION

Object tracking is required by many vision applications, especially in video technology and visual interface systems. Work on visual tracking can be divided into two groups: deterministic and stochastic tracking. Probabilistic video analysis has recently gained significant attention in the computer vision community through the use of stochastic sampling techniques. Visual tracking based on probabilistic analysis is formulated from a Bayesian perspective as a problem of estimating some degree of belief in the state of an object at the current time given a sequence of observations. The idea of a particle filter [1] (also known as Sequential Monte Carlo (SMC) algorithm) was independently used and proposed by several research groups. These algorithms provide flexible tracking frameworks as they are neither limited to linear systems nor require the noise to be Gaussian and proved to be more robust to distracting clutter as the randomly sampled particles allow to maintain several competing hypotheses of the hidden state. These algorithms have gained prevalence in the tracking literature due in part to the CONDENSATION algorithm [2].

Within deterministic approaches, appearance-based techniques have been widely used in object tracking. These techniques have the advantage that they are easy to implement and are generally more robust than feature-based methods. Of particular interest are Active Appearance Models (AAMs) [3] which contain two key elements: (i) a statistical appearance model of shape and texture, and (ii) an optimization algorithm. AAMs were used for face analysis/synthesis as well as for 3D face tracking [4].

In this work, we aim at designing a robust Active Appearance Model search for face and facial feature tracking by integrating a stochastic diffusion step in the search algorithm. The developed framework combines the merits of both stochastic and deterministic optimization techniques without the use of any additional visual cues such as optical flow or feature tracking. The rest of the paper is organized as follows. Section 2 describes the 3D deformable face model. Section 3 summarizes the concept of Active Appearance Model search. Section 4 describes the search algorithm utilizing a stochastic diffusion. Section 5 presents some experimental results.

2. AN ACTIVE FACE MODEL

In our study, we use the 3D face model *Candide*. The 3D shape of this deformable 3D wireframe model is directly recorded in coordinate form. The 3D face model is given by the 3D coordinates of the 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 $\mathbf{g}$ can be written as:

$$\mathbf{g} = \overline{\mathbf{g}} + \mathbf{S}\,\sigma + \mathbf{A}\,\alpha \tag{1}$$

where $\overline{\mathbf{g}}$ is the standard shape of the model, and the columns of $\mathbf{S}$ and $\mathbf{A}$ are the Shape and Animation Units, respectively. Thus, the term $\mathbf{S}\,\sigma$ accounts for shape variability

(inter-person variability) while the term $\mathbf{A}\,\alpha$ accounts for the facial animation (intra-person variability). In this study, we use 12 modes for the Shape Units matrix and six modes for the Animation Units matrix. As Animation Units, the following Action Units have been chosen: 1) Jaw drop, 2) Lip stretcher, 3) Lip corner depressor, 4) Upper lip raiser, 5) Eyebrow lowerer, 6) Outer eyebrow raiser. These Action Units are enough to cover most common facial expressions (mouth and eyebrow movements). More details about the used face model can be found in [5].

A face texture is represented as a geometrically normalized image. This geometry is represented by a triangular 2D mesh. The texture of this geometrically normalized image is obtained by texture mapping from the triangular 2D mesh in the input image using a piece-wise affine transform. Thus, the texture **x** of any geometrically normalized image is given by (using Principal Components Analysis):

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{X}\,\xi + \gamma \tag{2}$$

where $\overline{\mathbf{x}}$ is the mean texture, the columns of **X** are the texture modes (the first M eigenvectors), ξ is the vector of texture parameters, and γ accounts for the reconstruction error.

Given an image of a face (or a video sequence), the tracking consists of estimating the 3D pose of the face, σ and α for each image. In a tracking context, the model parameters associated with the current frame will be handed over to the next frame. For a given person, σ is constant. The geometry of the model is parameterized by the parameter vector **b** (the six degrees of freedom associated with the 3D head pose and the facial animation parameters):

$$\mathbf{b} \;=\; [\theta_x,\;\; \theta_y,\;\; \theta_z,\;\; \lambda\, t_x,\;\; \lambda\, t_y,\;\; \lambda\, t_z,\;\; \alpha^T\;]^T$$

3. THE ACTIVE APPEARANCE MODEL SEARCH

The goal is to find the optimal adaptation of the model to each input image. In [4], the parameter vector **b** is recovered by minimizing the reconstruction error which is given by the norm of the residual error between the geometrically normalized image and its best fit with the texture modes.

$$\min_{\mathbf{b}} e(\mathbf{b}) = \|\mathbf{r}(\mathbf{b})\|^2 = \|\mathbf{x} - \hat{\mathbf{x}}\|^2$$

This is carried out using a gradient descent method in which the gradient matrix is precomputed. Figure 1 illustrates the concept of the active appearance model search. It should be noticed that there are two differences between the work described in [4] and Cootes work [3]. In [4], (i) the estimated parameters deal with the 3D geometry of the face (shape and motion), (ii) the texture and geometry are separated.

When adapting a 3D face model to a video sequence, the above iterative algorithm is used for each input frame. For each frame, the initial estimate of **b** is given by its refined estimate in the previous frame.

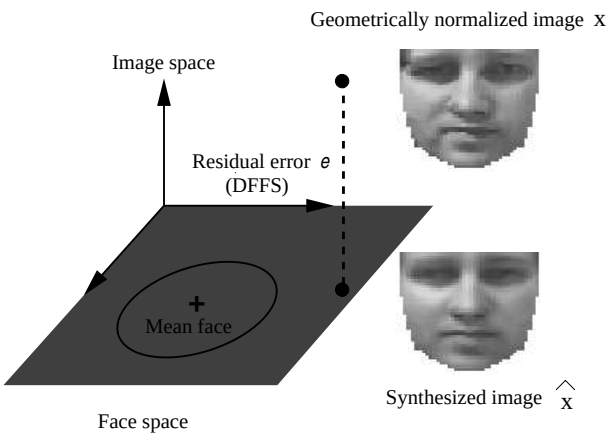


Fig. 1. The active appearance model search aims at finding the geometric parameters by minimizing the residual error between the geometrically normalized image and its best approximation in the face space. This residual error is usually called the Difference From Feature Space (DFFS).

4. ROBUST TRACKING WITH A STOCHASTIC DIFFUSION STEP

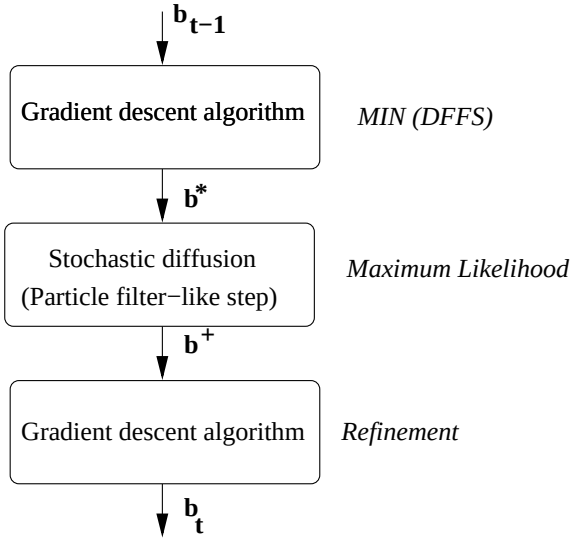


Fig. 2. The developed framework consists of three consecutive steps.

Since the active appearance model search adopts a directed continuous search and since the minimized reconstruction error is a highly multi-modal function, the gradient descent search can easily be stuck in non-desired local minima. Thus, fast head motions as well as sudden changes in the face illumination make the gradient descent algorithm inefficient in the sense that the solution obtained at conver-

gence can be completely different from the actual one. As a result, the adaptation becomes erroneous and the tracker often looses the face. When track is lost, the only chance for recovery is that the face visits a location (3D pose) which belongs to the convergence zone associated with the last correct optimal geometrical parameters.

To overcome the convergence problems encountered by the gradient descent algorithm, we add a stochastic search process which provides a mechanism for escaping non-desired sub-optimal solutions. The whole search algorithm is illustrated in Figure 2. It shows how the geometric parameters are computed for the current frame in the sequence. We proceed as follows.

Let $\mathbf{b}^\star$ be the solution obtained at the convergence of the gradient descent algorithm. Note that $\mathbf{b}^\star$ does not necessarily correspond to the desired solution, i.e. it may correspond to a local optimum. We draw J random samples according to the following model:

$$\mathbf{b}_0 = \mathbf{b}^\star \quad (3)$$

$$\mathbf{b}_j = \mathbf{b}^\star + \mathbf{n} \quad j = 1, \ldots, J-1 \quad (4)$$

In this model, $\mathbf{n}$ is a random vector having a centred normal distribution, i.e., $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$. The covariance matrix $\mathbf{C}$ is learned off-line from the state vector differences $\mathbf{b}_t - \mathbf{b}_{t-1}$ associated with previously tracked video sequences. This random drawing can be interpreted as the stochastic diffusion step in a condensation algorithm [2] where the deterministic component is given by $\mathbf{b}^\star$. In our implementation, J is set to 100.

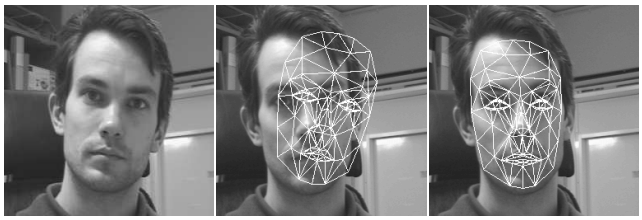
For each sample $\mathbf{b}_j$, we can easily compute the corresponding texture parameters $\xi(\mathbf{b}_j)$ and the reconstruction error $e(\mathbf{b}_j) = \|\gamma\|^2$. Thus, for each sample $\mathbf{b}_j$ the observation likelihood proposed in [6] for face detection is evaluated. This likelihood measure takes into account two distances (i) the DFFS (distance-from-feature-space), and (ii) the DIFS (distance-in-feature-space). Maximizing this likelihood is equivalent to minimizing the *Mahalanobis* distance over the original textures.

The maximum likelihood solution is:

$$\mathbf{b}^+ = \arg\max_{\mathbf{b}_j}(p(\mathbf{b}_j)) \quad (5)$$

Since the condensation algorithm weights the particles with a likelihood measure, and the weighted particles approximate the posterior distribution of the unobserved state, the solution given by Eq. (5) can also be viewed as a Monte Carlo approximation of the maximum a posteriori solution.

The final step of the method is to refine the solution found at the second step. To this end, the gradient descent algorithm is again invoked with $\mathbf{b}^+$ being the starting solution.



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

Fig. 3. **(a)** The previous frame (frame 17). **(b)** Gradient descent method: the model is (erroneously) adapted to the current frame (frame 18) due to the fast head motion. **(c)** The developed framework: the model is correctly adapted to the current frame.

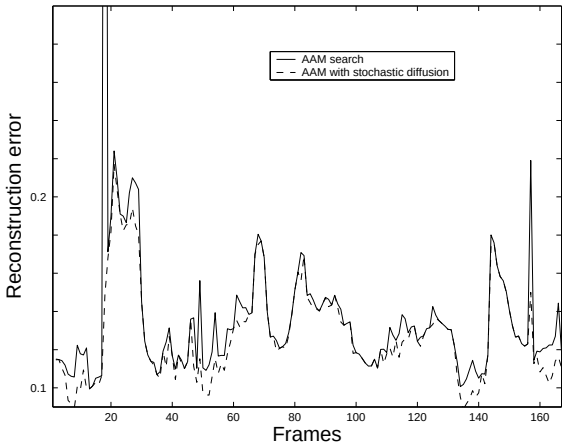


Fig. 4. The reconstruction error associated with the first test sequence (Figure 3). The solid curve corresponds to the AAM search, the dashed curve corresponds to the search algorithm which includes the stochastic diffusion step.

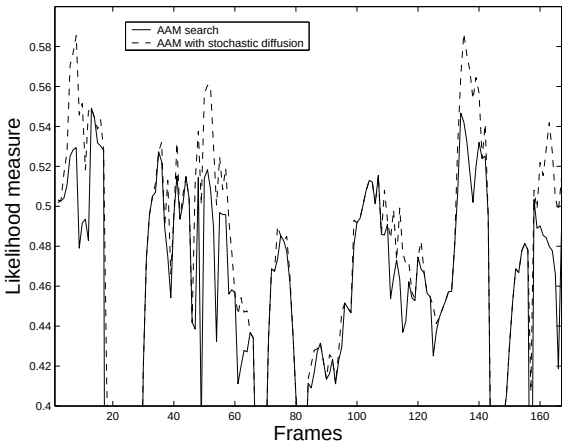
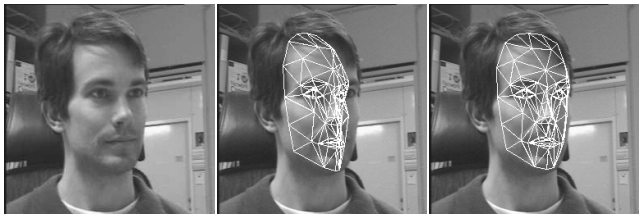


Fig. 5. The likelihood measure associated with the first test sequence (Figure 3). The solid curve corresponds to the AAM search, the dashed curve corresponds to the search algorithm including the stochastic diffusion step.



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

Fig. 6. **(a)** The previous frame (frame 274). **(b)** Gradient descent method: the model is badly adapted to the current frame (frame 275). Although there is no fast head motion, the optimization process was stuck in a non desired local minimal. **(c)** The developed framework: the model is correctly adapted to the current frame.

5. TRACKING RESULTS

Figure 3 (**(b)** and **(c)**) displays the adaptation results associated with one frame (frame 18) of a 167-frame-long sequence, with a fast head motion from the previous frame (Figure 3**(a)**). Figure 3**(b)** displays the adaptation results obtained with the gradient descent method. One can notice that the model is still stuck to its adaptation to the previous frame. Figure 3**(c)** displays the adaptation results when the developed framework has been used.

Figure 4 displays the reconstruction error associated with the sequence. The solid curve corresponds to the AAM search (the gradient descent method described in Section 3), the dashed curve corresponds to the search algorithm including the stochastic diffusion step (Section 4). Note the large difference in the reconstruction errors obtained at frame 18 for which the adaptations are depicted in Figures 3**(b)** and 3**(c)** for the gradient-based algorithm and for the proposed framework, respectively. Figure 5 displays the likelihood measure associated with the two algorithms. As can be seen, by including a stochastic diffusion step, the likelihood measure is further maximized yielding both a better estimation and a robust tracker.

Figure 6 (**(b)** and **(c)**) displays the adaptation results at frame 275 of another test sequence (a 340-frame-long sequence). Here, the head has moved only slightly from the previous frame (Figure 6**(a)**), the simple gradient descent method fails to provide an accurate adaptation due to its lack of robustness when the face is in a non-frontal view (Figure 6**(b)**). The stochastic diffusion, however, successfully corrects the solution provided by the gradient descent method (Figure 6**(c)**). For this sequence, the proposed algorithm yields also a higher likelihood during the whole length of the sequence (Figure 7).

6. CONCLUSION

We have proposed a stochastic 3D head and facial feature tracking method using an appearance-based fit measure. The approach proves to robustly track the 3D head pose despite rapid moves and significant out-of-plane rotations. Future work includes improving the robustness of the tracking to occlusions and the use of appearance-adaptive models that does not require to learn textures from a training set.

Acknowledgment The authors are grateful to Jörgen Ahlberg for providing the active appearance model data.

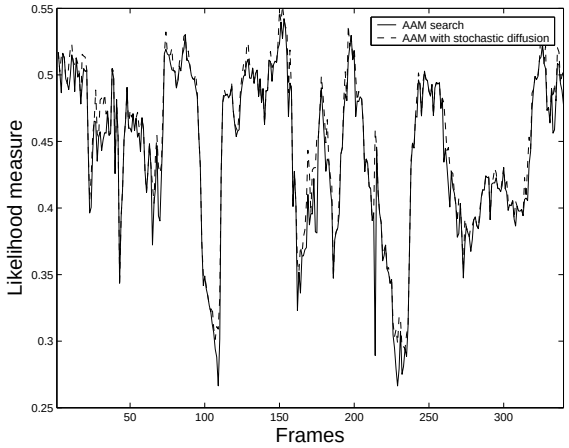


Fig. 7. The likelihood measure associated with the second test sequence (Figure 6). The solid curve corresponds to the AAM search, the dashed curve corresponds to the search algorithm including the stochastic diffusion step.

7. REFERENCES

[1] S. Arulampalam, S.R. Maskell, N.J. Gordon, and T. Clapp, "A tutorial on particle filters for on-line nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[2] M. Isard and A. Black, "Contour tracking by stochastic propogation of conditional density," in *Proc. European Conference on Computer Vision*, 1996.

[3] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–684, 2001.

[4] J. Ahlberg, "An active model for facial feature tracking," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 6, pp. 566–571, June 2002.

[5] F. Dornaika and J. Ahlberg, "Face and facial feature tracking using deformable models," *International Journal of Image and Graphics*, July 2004.

[6] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696–710, 1997.