FACIAL ACTION TRACKING USING AN AAM-BASED CONDENSATION APPROACH

Soumya Hamlaoui, Franck Davoine

Laboratoire HEUDIASYC - CNRS / UTC Université de Technologie de Compiègne

BP 20529, 60205 Compiègne Cedex - France
{shamlaou, fdavoine}@hds.utc.fr

ABSTRACT

In this paper, we address the problem of tracking a near-frontal view face and its facial features in a video sequence. To this purpose, a particle filtering scheme is proposed, where the distribution of observations is derived from an active appearance model. As in [8], the dynamics are adaptive in the sense that they are guided by a deterministic search, and the explored area of the state space is adjusted to the quality of the prediction. The number of particles is adapted accordingly, which enables a substantial gain in computing time. In order to account for occlusions, the observation model uses a robust distance measure. Experiments on real video show encouraging results.

1. INTRODUCTION

This work addresses the problem of tracking in a single video the global motion of a face as well as the local motion of its inner features. Note that in the applications targeted by this work, the person looks approximately in the direction of the camera. The face remains thus in a near frontal orientation, so that a 2D model of the face is assumed to be able to capture the expected variations. In the object tracking problem, the goal is to infer at each time step t the unobserved state of the object, denoted $\mathbf{x}_t \in \chi$, given all the observed data until time t, denoted $\mathbf{z}_{1:t} \equiv (\mathbf{z}_1,\ldots, \mathbf{z}_t)$. When tracking a face in 2D, the unobserved state includes motion or pose parameters like the position, scale and orientation of the face; when facial features are also tracked, the unobserved state should contain parameters describing the face inner motion. The observed data $\mathbf{z}_t$ consists of measurements derived from the current video frame such as greylevel patches. The tracking task then essentially consists in searching the current state $\mathbf{x}_t \in \chi$ that matches at best the measurements $\mathbf{z}_t$ in the current image.

In a non-probabilistic formulation of the tracking problem, the state $\mathbf{x}_t$ is usually seeked so as to minimize an error functional d [$\mathbf{g}_{image}(\mathbf{z}_t, \mathbf{x}_t)$; $\mathbf{g}_{model}$], e.g. an Euclidean or robust distance. The eigen-tracking method is based on such a principle [2]. Using principal component analysis, the Active Appearance Models (AAMs) encode the variations of face appearance by learning the shape and texture variations [3]. They enable thus the tracking of both global motion and inner features. In practice, tracking using the deterministic frame-by-frame AAM search appears to work well while the lighting conditions remain stable and only small occlusions are present. However, large occlusions often make the AAM search converge to incorrect positions and loose track of the face.

In probabilistic formulations, the hidden state and the observations are linked by a joint distribution; this statistical framework offers rich modeling possibilities. A Markov dynamic model describes how the state evolves through time. An observation model specifies the likelihood of each hypothesized state. Based on such a generative model, Bayesian filtering methods recursively evaluate the posterior density of the target state at each time step conditionally to the history of observations until the current time. Stochastic implementations of Bayesian filtering are generally based on sequential Monte Carlo estimation, also known as particle filtering [5]. Particle filtering approximates the posterior state density by a set of random weighted samples (particles) at each time step.

For video tracking, the CONDENSATION algorithm consists in propagating this sample set through time using a dynamic model and in weighting each sample proportionally to its likelihood function value [6]. This algorithm employs the Monte Carlo technique of factored sampling in order to recursively approximate the posterior state density. Approximation is done by means of the empirical distribution of a system of particles. The

particles explore the state space following independent realizations from a state evolution model, and are redistributed according to their consistency with the observations, the consistency being measured by a likelihood function [6].

The idea proposed in this paper consists in combining the AAM with the CONDENSATION stochastic search in order to augment its robustness to occlusions. Regarding existing works we are aware of combining AAM with temporal dynamics, they model facial behaviors in order to generate video-realistic animated faces (see e.g. [1]). In those papers, the tracking itself uses the AAM frame-by-frame search with no temporal dynamics.

In section 2, we present the proposed tracking algorithm. In section 3, experimental results are shown on real video. Finally, in section 4, we draw concluding remarks and discuss the perspectives opened by this work.

2. PROPOSED SCHEME: AAM-BASED CONDENSATION

A face AAM is a statistical model which describes shape and texture variations of the human face class [3]. The appearance variability is linearly modeled by a Principal Component Analysis (PCA) of shape **s** and texture **g**:

$$\mathbf{s} = \mathbf{s}_m + \phi_s \mathbf{b}_s \qquad \mathbf{g} = \mathbf{g}_m + \phi_g \mathbf{b}_g \tag{1}$$

where $\mathbf{s}_m$, $\mathbf{g}_m$ are respectively the mean shape and texture, $\boldsymbol{\phi}_s$, $\boldsymbol{\phi}_g$ are the eigenvectors of shape and texture covariance matrices. A third PCA is then performed on a concatenated shape and texture parameters **b**, to obtain a combined model vector **c**:

$$\mathbf{b} = \phi_c \mathbf{c} \tag{2}$$

From the combined appearance model vector **c**, a new instance of shape and texture can be generated:

$$\mathbf{s}_{model}(\mathbf{c}) = \mathbf{s}_m + \mathbf{Q}_s \mathbf{c} \qquad \mathbf{g}_{model}(\mathbf{c}) = \mathbf{g}_m + \mathbf{Q}_g \mathbf{c}. \tag{3}$$

We propose to adapt the CONDENSATION algorithm in combination with AAM to our tracking task in three aspects, each being detailed below. For a good introduction, the reader is referred to the seminal paper of Isard and Blake [6].

2.1. State space spans the global and inner motion of the face

The state vector $\mathbf{x}_t$ contains the parameters to infer about the object:

- the face global 2D pose $\mathbf{p}_t = (t_x, t_y, \alpha, \theta)^T$, representing position, scale and orientation of the face.
- the facial actions, contained in the AAM shape and texture, which are themselves captured in a compact way by the combined appearance parameter vector denoted $\mathbf{c}_t$. Our experiments suggest that retaining only the first four modes of the appearance parameter $\mathbf{c}_t$ allows spanning the facial changes of interest and provides satisfying tracking results.

The state vector $\mathbf{x}_t = (\mathbf{p}_t, \mathbf{c}_t)^T$ is thus of dimension 8.

2.2. AAM-based observation model

The observation model is based on sampled pixel grey level patches and a previously trained AAM subspace; it consists of the likelihood $p(\mathbf{z}_t \mid \mathbf{x}_t)$, according to which the particles are weighted in the CONDENSATION algorithm. This likelihood indicates the probability that a hypothesized state $\mathbf{x}_t = (\mathbf{p}_t, \mathbf{c}_t)^T$ gives rise to the observed data. This probability should be high whenever there is a good match between:

- the image patch sampled at the hypothesized pose and shape, $\mathbf{g}_{im}(\mathbf{p}_t, \mathbf{c}_t)$.
- the hypothesized appearance of the face, given by the model texture $\mathbf{g}_{model}(\mathbf{c}_t)$.

The adopted likelihood function has thus the following form:

$$p(\mathbf{z}_t|\mathbf{x}_t) = p(\mathbf{z}_t|\mathbf{p}_t, \mathbf{c}_t) = C \exp -d\left[\mathbf{g}_{model}(\mathbf{c}_t); \mathbf{g}_{im}(\mathbf{p}_t, \mathbf{c}_t)\right] \tag{4}$$

where C is the normalizing constant of this distribution, and the texture distance d [;] is an error measure, summed over all L pixels of both textures:

$$d\left[\mathbf{g}; \mathbf{g}'\right] = \sum_{\ell=1}^{L} \rho\left(\frac{g_\ell - g'_\ell}{\sigma_\ell}\right). \tag{5}$$

This error is weighted by the standard deviation $\boldsymbol{\sigma}_l$ of each pixel, computed from training data. The error function $\boldsymbol{\rho}$() can be chosen in different ways:

- a simple square error function $\rho(x) = \frac{1}{2}x^2$ yields a weighted Euclidean distance d[;] and a Gaussian density $p(\mathbf{z}_t \mid \mathbf{x}_t)$;
- a robust error function can be used instead (see e.g. [5]); in our experiments, we tested the following function:

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \le h \\ \frac{1}{2}h|x| - \frac{1}{2}h^2 & \text{if } |x| > h \end{cases} \tag{6}$$

where h is a fixed threshold above which the difference |x| is considered to be an outlier. Such a robust measure reduces the influence of occluded pixels, which would otherwise dominate the total error measure (5) and rule out a potentially good state candidate.

2.3. Adaptive dynamics

The state transition model $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ is used in the CONDENSATION algorithm in order to draw the particles approximating the predicted distribution. Following the ideas developed in Zhou *et al* [8], the

dynamics used here are adaptive by having the following model for state evolution:

$$\mathbf{x}_t = \hat{\mathbf{x}}_{t-1} + \mathbf{v}_t + \mathbf{S}_t \mathbf{u} \quad (7)$$

- $\hat{\boldsymbol{x}}_{t-1}$ is the estimate of the state vector at the previous time step,
- the velocity $\mathbf{v}_t = (\partial\mathbf{p}, \partial\mathbf{c})^T$ indicates the predicted shift in pose/appearance,
- the random component $\mathbf{u}$ is a vector of 8 independent normal random variants having zero mean and unit variance,
- the diagonal matrix $\mathbf{S}_t$ = diag $(\boldsymbol{\sigma}_t^{(t_x)},\ldots,\boldsymbol{\sigma}_t^{(c_4)})$ specifies the standard deviation of the random draw for each pose/appearance parameter.

The predicted shift $\mathbf{v}_t$ is obtained by an automatic AAM search in the current frame (iterative gradient-like adaptation) as in [3]. This search aims to compute the predicted pose and appearance parameters $(\tilde{\boldsymbol{p}}_t, \tilde{\boldsymbol{c}}_t)$ that best approximate the target face in the current image. This optimization is initialized at the previous state estimate $\hat{\boldsymbol{x}}_{t-1}$. The criterion to minimize is thus the norm of the texture residue vector:

$$\mathbf{r}(\mathbf{p},\mathbf{c}) = \mathbf{g}_{model}(\hat{\mathbf{c}}_{t-1}) - \mathbf{g}_{im}(\hat{\boldsymbol{p}}_{t-1}, \hat{\boldsymbol{c}}_{t-1}) \quad (8)$$

where $\mathbf{g}_{im}(\hat{\boldsymbol{p}}_{t-1}, \hat{\boldsymbol{c}}_{t-1})$ denotes the current image texture sampled at the previous state estimate $\hat{\boldsymbol{x}}_{t-1}$. The optimal successive corrections to apply $(\partial\mathbf{p}, \partial\mathbf{c})$, are linear functions of the error vector:

$$\partial\mathbf{p} = \boldsymbol{R}_p\,\mathbf{r}(\mathbf{p},\mathbf{c}) \qquad \partial\mathbf{c} = \boldsymbol{R}_c\,\mathbf{r}(\mathbf{p},\mathbf{c}) \quad (9)$$

The matrices $\boldsymbol{R}_p$ and $\boldsymbol{R}_c$ can be precomputed from training data, as in [3].

This deterministic search aims to focus the particle drawing in a region that is most likely to contain good candidates, and thus reduce the volume of the state space to explore:

$$\tilde{\boldsymbol{x}}_t = (\tilde{\boldsymbol{p}}_t, \tilde{\boldsymbol{c}}_t) = \hat{\boldsymbol{x}}_{t-1} + (\partial\mathbf{p}, \partial\mathbf{c})^T \quad (10)$$

According to the state transition model (7), pose/appearance parameters are drawn around the predicted state $\tilde{\boldsymbol{x}}_t = (\tilde{\boldsymbol{p}}_t, \tilde{\boldsymbol{c}}_t)$ with dispersions (standard deviations) given by $\mathbf{S}_t$. We consider, as in [8], adaptive dispersion given by:

$$(\boldsymbol{\sigma}_t^{(t_x)},\ldots,\boldsymbol{\sigma}_t^{(c_4)}) = R_t\,(\boldsymbol{\sigma}_0^{(t_x)},\ldots,\boldsymbol{\sigma}_0^{(c_4)}) \quad (11)$$

where $(\boldsymbol{\sigma}_0^{(t_x)},\ldots,\boldsymbol{\sigma}_0^{(c_4)})$ are fixed reference standard deviations, and the scaling factor R_t is proportional to the square root of ε_t, with bounding values $[R_{min}, R_{max}]$:

$$R_t = \max(\min(\sqrt{\varepsilon_t}, R_{max}), R_{min}) \quad (12)$$

where ε_t is a measure of variance corresponding to a texture error averaged over the L pixels of the textures:

$$\varepsilon_t = \frac{2}{L}\sum_{\ell=1}^{L} \rho\left(\frac{g_\ell^{model}(\tilde{\mathbf{c}}_t) - g_\ell^{im}(\tilde{\mathbf{p}}_t, \tilde{\mathbf{c}}_t)}{\sigma_\ell}\right) \quad (13)$$

When R_t is large, the predicted distribution has a high variance and requires therefore a large number of particles to approximate it. In other words, the larger is the area of the state space subregion covered by the predicted distribution, the more particles are needed to explore it. This suggests having an adaptive number N_t of particles, using the formula:

$$N_t = N_0 R_t \quad (14)$$

where N_0 is a fixed number of particles.

3. EXPERIMENTAL RESULTS

The proposed method was implemented in non-optimized C++ and tested on a PC running WinXP at 2.4 GHz with 512 Mb of RAM.

Results are first shown for a video sequence where a face in near-frontal view undergoes large variations in pose and expressions (see Figure 1). The tracking of both global pose and facial features appears satisfying. Setting N_0 = 500, the number of particles N_t evolves between about 20 and 80, and increases each time the change in pose and/or appearance is rapid; using such adaptive dynamics allows to process on average 2 frames per second. This represents a drastic improvement over a method using a zero-velocity state evolution model, which required 1000 particles to successfully track this sequence (according to experiments not shown here).

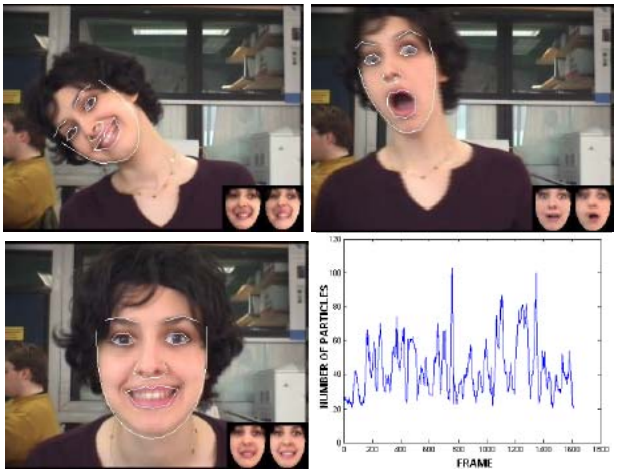

Figure 1: AAM-based adaptive CONDENSATION tracking, for frames 015, 363 and 618. On each image, the drawn shape shows the estimated state; the model and image texture $\mathbf{g}_{model}(\mathbf{c}_t)$ and $\mathbf{g}_{im}(\mathbf{p}_t, \mathbf{c}_t)$ are displayed in the lower right corner. The graph displays the variable number of particles N_t over 1600 frames.

The performance of our approach was also tested in presence of occlusions. We compared it with a purely deterministic AAM tracking. As is highlighted in the top row of Figure 2, when the occlusion occurs, the deterministic search appears to be trapped in an incorrect local optimum, and the tracking diverges thus from that moment. This problem is overcome by the stochastic tracking: the occlusion induces a high texture error ε_t for the predicted state $\tilde{\boldsymbol{x}}_t$, and consequently the variance of drawn particles and their number N_t are increased (see the peaks in Figure 3). The particles cover thus a greater area of the state space which allows to correct the deterministic search (bottom row of Figure 2).

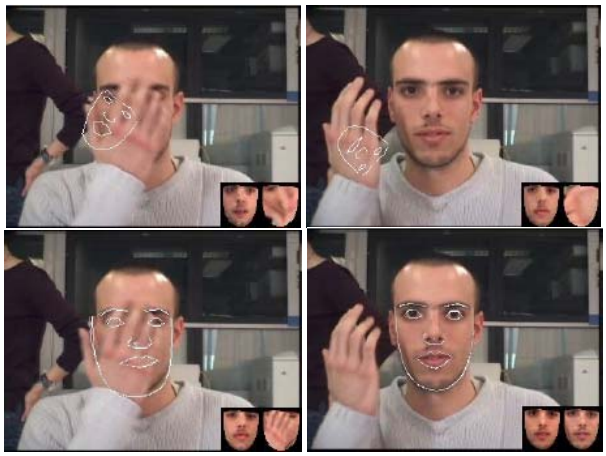

Figure 2: Tracking on a video sequence with occlusions, frames 921 and 940. Top row: deterministic AAM tracking. Bottom row: CONDENSATION based tracking.

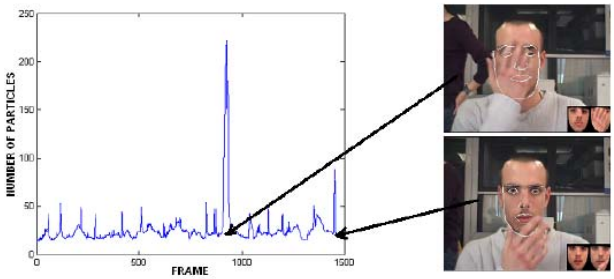

Figure 3: Evolving number of particles N_t on the video sequence with occlusions. The nearly full occlusion of frame 921 induces a high peak, while a partial occlusion induces a lower peak.

4. CONCLUSION

For the purpose of tracking the 2D global pose of a face and its inner facial actions, this paper proposes to combine an adaptive particle filtering scheme with an active appearance model. The state vector is composed of four pose parameters and four combined appearance parameters. The likelihood measures the fit between the hypothesized model texture and the image texture sampled at the hypothesized location and shape; a robust distance accounts for occluded pixels. Following the ideas of [8], the dynamics in state space are guided by a deterministic AAM search; this allows reducing significantly the number of particles, which is only increased when the AAM search fails to converge to a satisfying solution. The experiments show that the proposed algorithm can successfully track a face and its facial actions undergoing quick motion and nearly full occlusions.

On the basis of this work, several directions can now be investigated. On the one hand, training the tracking system could be made easier, by learning the texture model on the fly as in [8]. On the other hand, now that a robust tracking system is available, we can study the recognition of facial actions: the input being given by the combined appearance parameters at each time step, different recognition approaches can be tested, from a simple linear discriminate analysis on still frames, to dynamic graphical models. In this regard, the particle filter paradigm provides a natural inference framework for richer models, for instance, the facial action to be recognized could be included as a discrete component of the state vector [7].

5. REFERENCES

[1] F. Bettinger, T.F. Cootes and C.J. Taylor, "Modelling facial behaviours", *In Proc BMVC 2002*, pp. 797-806, 2002.

[2] M. Black and A. Jepson, "Eigen-tracking: Robust matching and tracking of articulated objects using a view-based representation", *Int. Journal of Computer Vision*, pp. 101–130, 1998.

[3] T.F. Cootes, G.J. Edwards and C.J. Taylor, "Active Appearance Models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 681-685, June 2001.

[4] A. Doucet, J.F.G De Freitas, N. Gordon, *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, 2001.

[5] P.J. Huber, *Robust Statistics*, Wiley, 1981.

[6] M. Isard, A. Blake, "Condensation – Conditional Density Propagation for Visual Tracking", *Int. Journal of Computer Vision*, pp. 5-28, 1998.

[7] M. Isard and A. Blake, "A mixed-state condensation tracker with automatic modelswitching", In *Proc. 6th Int. Conf. Computer Vision*, pp. 107–112, 1998.

[8] S. Zhou, R. Chellappa, and B. Moghaddam, "Visual tracking and recognition using appearance-adaptive models in particle filters", *IEEE Trans. on Image Processing*, November 2004.