View-Based Appearance Model Online Learning for 3D Deformable Face Tracking

Stéphanie Lefèvre and Jean-Marc Odobez
Idiap Research Institute, Martigny, Switzerland
Ecole Polytechnique Fédérale de Lausanne, Switzerland
odobez@idiap.ch

Keywords: 3D head tracking, appearance models, structural features, view-based learning, facial expression.

Abstract: In this paper we address the issue of joint estimation of head pose and facial actions. We propose a method that can robustly track both subtle and extreme movements by combining two types of features: structural features observed at characteristic points of the face, and intensity features sampled from the facial texture. To handle the processing of extreme poses, we propose two innovations. The first one is to extend the deformable 3D face model Candide so that we can collect appearance information from the head sides as well as from the face. The second and main one is to exploit a set of view-based templates learned online to model the head appearance. This allows us to handle the appearance variation problem, inherent to intensity features and accentuated by the coarse geometry of our 3D head model. Experiments on the Boston University Face Tracking dataset show that the method can track common head movements with an accuracy of 3.2°, outperforming some state-of-the-art methods. More importantly, the ability of the system to robustly track natural/faked facial actions and challenging head movements is demonstrated on several long video sequences.

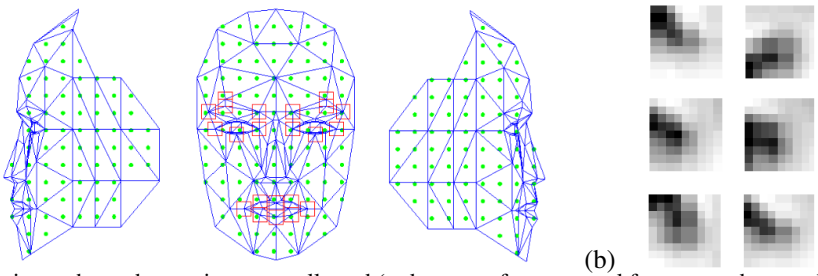
1 Introduction

The many applications of face tracking, in domains ranging from Human Computer Interaction to surveillance, urged researchers to investigate the problem for the last twenty years. Still some issues remain; the difficulties come from the variability of appearance created by 3D rigid movements (especially self occlusions due to the head pose), non-rigid movements (due to facial expressions), variability of 3D head shape and appearance, and illumination variations.

An important contribution to the problem of near-frontal face tracking was made by Cootes et al. The idea was to use Principal Component Analysis to model the 2D variations of the face shape (Active Shape Model (ASM) (Cootes et al., 1995)), or of both shape and appearance (Active Appearance Model (AAM) (Cootes et al., 1998)). Later, some works have extended the use of AAMs to more challenging poses (Gross et al., 2006), but the lack of robustness when confronted to large head pose variations is still a typical limitation of these models. Besides, extracting the 3D pose from the 2D fit is possible but not straightforward; it requires further computation (Xiao et al., 2004).

Face tracking can also be formulated as an image registration problem, and several approaches were developed to robustly track faces under large pose variations. They usually rely on a rigid 3D face/head model, which can be a cylinder (Cascia et al., 2000; Xiao et al., 2003), an ellipsoid (Morency et al., 2008), or a mesh (Vacchetti et al., 2004). The model is fit to the image by matching either local features (Vacchetti et al., 2004) or a facial texture (Cascia et al., 2000; Xiao et al., 2003; Morency et al., 2008). However, they are limited to rigid movements. In the best case the tracking is robust to facial actions; in the worst case they will cause the system to lose track; in any case they are not estimated.

To track both the head pose and the facial actions, an appropriate solution is to use a deformable 3D face/head model. Approaches using optical flow (DeCarlo and Metaxas, 2000), local structural features (Chen and Davoine, 2006; Lefèvre and Odobez, 2009), or facial texture (Dornaika and Davoine, 2006) to fit the 3D model to a face have been tried in the past. However, the tracking success is highly dependent on the recording conditions. Optical flow methods can be very accurate but are not robust to fast motions. Structural features computed at a small set of characteristic points provide useful information about both the pose and the facial actions. However, due to the set sparsity and the locality of the information, the model will not be constraining enough if too many features are hidden (e.g. when reaching a near profile view). Facial texture provides rich and precise information for tracking but is very sensitive to appearance changes. The latter is a serious problem; unless the lighting is coming uniformly from every direction, the



(a) (b)

Figure 1: (a) Set of locations where observations are collected (red squares for structural features and green dots for intensity features). (b) Samples of the training set for the structural feature located on the right corner of the right eye, before removing the patch mean.

appearance of the face will vary a lot as the head pose changes.

The approach in (Lefèvre and Odobez, 2009) showed the advantages of combining both types of cues: it relied on both structural features similar to (Chen and Davoine, 2006) and on intensity values computed at a sparse set of face points. The appearance model was continuously adapted to deal with appearance changes. However this approach suffered from two main problems: first, because the majority of observations are located in the face region, there is very few information when the pose reaches profile view. This issue is common to many models. To our knowledge, models for head tracking which cover the head sides are either coarse rigid models (cylinder, ellipse) or person-specific rigid models (3D model acquired with a scanner). Secondly, the system is memoryless: the appearance model of the intensity features always needs to adapt in the same way when coming back to the same pose.

In this paper, our contribution is to propose a modeling that addresses these two issues.

First, we propose to extend the Candide face model to cover head sides. Although collecting features from the head sides would allow to track challenging poses that face models cannot, the vast majority of face tracking approaches do not consider such information. Indeed, such an extension brings in additional difficulties. The appearance changes issue is even more present than before, since, most of the time, between near frontal view and profile view the intensity of points located on the head sides varies drastically. These variations are accentuated by the fact that the mesh extension is very coarse, in the sense that the approximation of the depth of the points on the head surface is usually inaccurate. In fact, it is quite difficult to built a precise person specific head model, and this is a reason why many approaches do not consider such head side extensions (AAM, Candide, etc.).

Secondly, to add memory in the appearance modeling, we propose to represent the head using a set of view-based template learned online. This is in contrast with the majority of approaches that propose to handle the appearance variation problem using either template adaptation of all sorts (e.g. doing recursive adaptation (Lefèvre and Odobez, 2009; Dornaika and Davoine, 2006), combining current observations with the initial template (Matthews et al., 2004), or using short and long term adaptation models (Jepson et al., 2003)) or incremental model learning techniques (e.g. incremental PCA (Li, 2004) or an EM algorithm (Tu et al., 2009)). None of these methods consider the fact that in most applications the appearance of the face mainly depends on the pose, since the location of the camera and of the illumination sources are usually fixed. The approach we propose that relies on templates learned online and representing appearance under different poses addresses this issue. Furthermore, it is well adapted to handle the coarse depth modeling of the additional head side mesh elements. The main difficulty of our approach lies in the building of the template set, as the risk is to learn an incorrect combination pose/template when the head motion is heading towards a region of the pose space that was not visited before. This issue is dealt with to a large extent by exploiting a fixed (i.e. not subject to adaptation) likelihood term relying on structural features. The fact that this likelihood model is learned off-line and is built on illumination-invariant cues reduces the risk of drift.

The performances of our approach are evaluated on the Boston University Face Tracking (BUFT) database (on both Uniform-light and Varying-light datasets) and on several long video sequences of people involved in natural conversation. They show that the combination of head-side and view-based modeling allows us to outperform some recent state-of-the-art techniques (Cascia et al., 2000; Morency et al., 2008) and to robustly track challenging head movements and facial actions.

2 Candide, a deformable 3D model

In this work we use an extended version of the Candide (Ahlberg, 2001) face model. The original model consists in a deformable 3D mesh defined by the 3D coordinates of 113 vertices (facial feature points) and by the edges linking them. By displacing the vertices of a standard face mesh $\overline{M}$ according to some shape and action units, one can reshape the wireframe to the most common face shapes and expressions. The transformation of a point $\overline{M}_i$ of the standard face mesh into a new point M_i can be expressed as follows: $M_i(\alpha,\sigma) = \overline{M}_i + S_i.\sigma + A_i.\alpha$, where S_i and A_i are respectively the 3×14 shape unit matrix and the 3×6 action unit matrix that contain the effect of each shape (respectively action) unit on point $\overline{M}_i$. The 14×1 shape parameters vector σ and the 6×1 action parameters vector α contain values between -1 and 1 that express the magnitude of the displacement. In our case, σ is learned once for all for a given person before tracking using a reference image (a frontal view of the person) by manually or automatically annotating several points on reference image and by finding the shape parameters σ that best fit the Candide model to the data points.

Extending the model: A limitation of the Candide model is that it only covers the face region. In our experiments we have to deal with some challenging head poses under which the face is half-hidden (e.g. self-occlusion at profile view). In that case it is useful to collect some information on the sides of the head. Indeed the texture and contrast in this region, and especially around the ears, is a strong indicator of the head movement. For this reason we extended the Candide model so that the mesh reaches the ears. Twenty vertices forming a unique planar region (for each head side) in the continuity of the original mesh were added to the standard mesh as well as a "Head width" shape unit vector. None of these new points are displaced by the action units. Note that the part of the mesh that covers the sides is very coarse; however it will bring useful information during the tracking. An illustration of the extended Candide model can be found in Fig. 1.

State space: In the Candide model, the points of the mesh are expressed in the (local) object coordinate system. They need to be transformed into the camera coordinate system and then to be projected on the image. The first step involves a scale factor s (the Candide model is defined up to a scale factor), a rotation matrix (represented by three Euler angles θ_x, θ_y and θ_z) and a translation matrix $T = (t_x\ t_y\ t_z)^T$. The camera is not calibrated and we adopt the weak perspective projection model (i.e. we neglect the perspective effect) to map a 3D point M_i to an image point m_i. Thus the vector of the head pose parameters to estimate can be expressed as $\Theta = [\theta_x\ \theta_y\ \theta_z\ \lambda t_x\ \lambda t_y\ s]$ where λ is a constant. The whole state (head pose and facial actions parameters) at time t is defined as follows:

$$X_t = [\Theta_t\ \alpha_t]\,. \tag{1}$$

3 Tracking faces

We set the problem as a Bayesian optimization problem. The objective is to maximize the posterior probability $p(X_t|Z_{1:t})$ of the state X_t at time t given observations $Z_{1:t}$ from time 1 to time t. Under standard assumptions, and assuming that the distribution of the posterior $p(X_{t-1}|Z_{1:t-1})$ is a dirac $\delta(X_{t-1} - \hat{X}_{t-1})$ (we only exploit a point estimate of the state at the previous time step), $\hat{X}_{t-1}$ being the previous estimate of the state, this probability can be approximated by:

$$p(X_t|Z_{1:t}) \propto p(Z_t|X_t) \cdot p(X_t|\hat{X}_{t-1})\,. \tag{2}$$

This expression is characterized by two terms: the likelihood $p(Z_t|X_t)$, which expresses how good are observations given a state value, and $p(X_t|\hat{X}_{t-1})$ which represents the dynamics, i.e. the state evolution. Our observations are composed of structural features and intensity features, i.e. $Z_t = (Z_t^{str}, Z_t^{int})$. Assuming that they are conditionally independent given the state, Eq. (2) can be rewritten as:

$$p(X_t|Z_{1:t}) \propto p(Z_t^{str}|X_t) \cdot p(Z_t^{int}|X_t) \cdot p(X_t|\hat{X}_{t-1})\,. \tag{3}$$

Each component is detailed below.

3.1 Likelihood model of structural features

Our goal is to learn a fixed appearance model valid under variations of head pose and illumination for patches located around characteristic points of the face. The advantage of these features is that, when they are visible, they give useful information about both the head pose and the facial actions. By learning a robust likelihood model, we aim at constraining the tracking strongly enough under any illumination condition for near-frontal to mid-profile poses.

Observations: We call S^{str} the index set of 22 structural features. Given the state X_t, observations Z^{str} will be 9×9 zero-mean patches collected around the projected points $\{m_i(X_t)\}_{i \in S^{str}}$, i.e. $Z_t^{str}(X_t) = \{Z_{i,t}^{str}(X_t)\}_{i \in S^{str}} = \{patch(m_i(X_t))\}_{i \in S^{str}}$. The locations of the observations are illustrated in

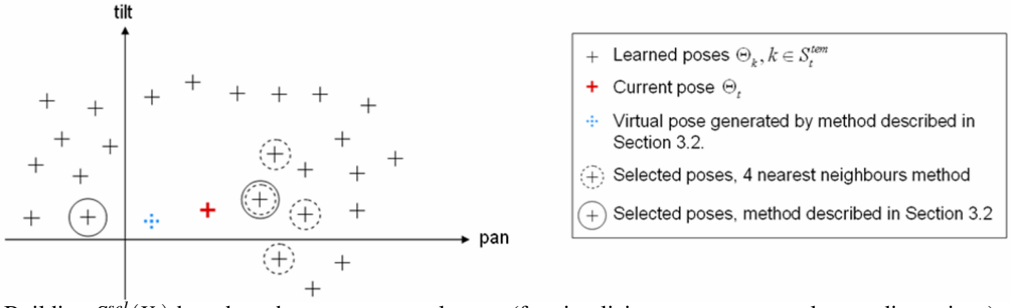


Figure 2: Building $S_t^{sel}(X_t)$ based on the poses: example case (for simplicity we represent only two dimensions). Selection with the k nearest neighbors approach, $k = 4$ v.s. selection with the approach described in Section 3.2.

Fig. 1.

Likelihood modeling: Assuming conditional independence between the features given the state[1]:

$$p(Z_t^{str}|X_t) = \prod_{i \in S^{str}} p(Z_{i,t}^{str}|X_t)\,. \tag{4}$$

This model is learned off-line using a reference image of the face. For each feature we extract a patch in the reference image, subtract the mean value to make it invariant to illumination changes, and simulate what it would look like under different head poses. This is done by applying a set of affine transformations to it, assuming the patch is planar. More precisely, for each of the three rotation parameters we sample uniformly seven values from $-45°$ to $45°$. This is illustrated in Fig. 1 (b). From this training set we compute the 1×81 mean vector μ_i and the 81×81 covariance matrix Σ_i, and define the likelihood model for a normalized 9×9 image patch $Z_{i,t}^{str}$ as:

$$p(Z_{i,t}^{str}|X_t) \propto e^{-\rho(\sqrt{(Z_{i,t}^{str}-\mu_i)^T \Sigma_i^{-1}(Z_{i,t}^{str}-\mu_i)}, \tau_{str})} \tag{5}$$

where ρ is a robust function (we used the truncated linear function) and τ_{str} is the threshold above which a measurement is assumed to be an outlier.

3.2 Likelihood model of intensity features using a set of view-based templates

The intensity features are located on both the face and the head sides, and their location distribution is much denser than the locations of the structural features. Therefore the intensity features bring precise and rich information about the appearance of the whole face. In many cases, however, although the illumination conditions are fixed the lighting is not uniform over the face (e.g. the light might be coming from the side). Thus the intensity of a face point is highly pose-dependent and can vary quite fast depending on the head movements. In order to handle this problem, we define a likelihood model that relies on a set of view-based templates.

Observations: The observations Z^{int} are defined by the intensity values at the projected points $\{m_i(X_t)\}_{i \in S^{int}}$, i.e. $Z_t^{int}(X_t) = \{Z_{i,t}^{int}(X_t)\}_{i \in S^{int}} = \{intensity(m_i(X_t))\}_{i \in S^{int}}$, where S^{int} denotes the index set of intensity features. The locations of the observations are illustrated in Fig. 1.

Likelihood modeling: The likelihood of the observations $Z_t^{int}(X_t)$ is evaluated by comparing them to a set of view-based templates. This set is built online by adding a new template each time a new region of the head pose space is reached, as described later. We call S_t^{tem} the complete set of view-based templates learned so far at time t. A template $T_k = (\mu_k, \Theta_k), k \in S_t^{tem}$ is defined by a vector of intensities μ_k and a pose Θ_k.
The observations $Z_t^{int}(X_t)$ will be compared to a set of selected templates $S_t^{sel}(X_t)$, with $S_t^{sel} \subseteq S_t^{tem}$. From this set of selected templates we create a mixed template whose appearance μ_t^{mix} is defined as: $\mu_t^{mix} = \sum_{k \in S_t^{sel}} w_{k,t} \cdot \mu_k$, where $w_{k,t}$ is the weight associated to the selected template T_k. The methodology to select S_t^{sel} and the weights is described below.
Assuming conditional independence between the features given the state, we have:

$$p(Z_t^{int}|X_t) = \prod_{i \in S^{int}} p(Z_{i,t}^{int}|X_t) \tag{6}$$

where the likelihood model for a single intensity value can be expressed as:

$$p(Z_{i,t}^{int}|X_t) \propto e^{-w_{k,t} \cdot \rho(\frac{(Z_{i,t}^{int}-\mu_{i,t}^{mix})^2}{\sigma_{int}^2}, \tau_{int})} \tag{7}$$

where ρ is a robust function (we used the truncated linear function), τ_{int} is the threshold above which a measurement is assumed to be an outlier, and σ_{int} is a constant.

[1] Note that such assumption would not be valid if patches would overlap.

Selection of the subset $S_t^{sel}(X_t)$: The set of templates $S_t^{sel}(X_t)$ plays an important role, as it defines the mixed appearance μ_t^{mix}. The main idea for our method to build $S_t^{sel}(X_t)$ follows the principle that, whenever possible, to synthesize a view it is usually much better to interpolate it than to extrapolate it. This is illustrated in Fig. 2. A classical approach would use the k nearest neighbors to build μ_t^{mix}. However, this is not always a good solution because the set of templates is learned online, and therefore the learned templates do not uniformly populate the pose space. Most of the views selected in this manner may be located on one side only of the current pose (see Fig. 2), leading to the extrapolation of the view from $S_t^{sel}(X_t)$ rather than its interpolation. If instead we select poses not only based on their distance to Θ_t but also based on their spread in the pose space, we might increase the accuracy of the view synthesis. Thus the proposed solution consists of defining $S_t^{sel}(X_t) = \{T_1, T_2\}$ where T_1 is the template whose pose Θ_1 is the closest to the current pose Θ_t, and T_2 is the template whose pose Θ_2 is the closest to the pose symmetrical to pose Θ_1 with respect to Θ_t. This way we make sure that the two selected poses will draw the current pose towards two opposite directions, as much as possible given the current set of templates. This is illustrated in Fig. 2. This simple approach provides a good compromise between the distance to Θ_t and repartition in the pose space. Finally, each of the two selected poses is associated a weight defined as $w_{k,t} = \frac{1}{d(\Theta_k,\Theta_t)}, k \in S_t^{sel}$, where $d(\Theta,\Theta')$ is defined as the euclidean distance between two poses Θ and Θ' in the pose space. That way the contribution of a template varies with the distance of its pose to the current pose. The weights are normalized so that their sum is equal to 1.

Addition of a template to the set of view-based templates: S_{t+1}^{tem} is built from the set of templates S_t^{tem} and the estimated pose $\hat{\Theta}_t$ by adding a new template only if it models a new region of the pose space, i.e. only if its pose is far enough from the poses of the templates already learned. That is, when the following condition is verified:

$$\forall k \in S_t^{tem}, d(\hat{\Theta}_t, \Theta_k) > \tau \qquad (8)$$

the template $T = (\hat{Z}_t^{int}, \hat{\Theta}_t)$ is added to the set S_t^{tem}. Otherwise $S_{t+1}^{tem} = S_t^{tem}$. As a value for τ we used 10°, a good compromise between appearance modeling and pose densities.

Updating the set of view-based templates: There is always a risk that a bad template is learned, for example if one part of the mesh is temporarily not well fit on the face when a new template is added to the list. For this reason, it is useful to have an adaptation mechanism that allows the appearance of a learned template to be updated when the same pose is visited again. Under some specific conditions, we update the appearance of the closest template T_k in the following way: $\mu_{k,t+1} = \beta \cdot \hat{Z}_t^{int} + (1-\beta) \cdot \mu_{k,t}$, with $\beta = 0.5 - 0.5 \cdot \frac{d(\hat{\Theta}_t,\Theta_k)}{\tau}$, i.e. β will vary between 0 and 0.5 depending on the distance d. The conditions to perform this update are 1) No template has just been created from the current pair pose/observations (see description in the above paragraph) and 2) The same template cannot be updated twice in a row. This last criterion drastically reduces the risk of drift that occurs when appearance is adapted continuously.

Dealing with global illumination changes: The appearance model as we described it so far is not robust to global illumination changes. We deal with this issue in a coarse way, so that the tracking is not perturbed by a sudden change of camera gain or by a long-term change in the lighting. Before processing any frame, all intensities are corrected by a constant value so that the average intensity of the image is the same as the one in the first frame.

3.3 Dynamical model

This term defines how large we assume the difference in the state between two successive frames can be. The N_p components of the states are assumed to be independent and to follow a constant position model:

$$p(X_t|\hat{X}_{t-1}) = \prod_{i=1:N_p} \mathcal{N}(X_{i,t}; \hat{X}_{i,t-1}, \sigma_{d,i}) \qquad (9)$$

where $X_{i,t}$ denotes the i^{th} component of X_t, and $\{\sigma_{d,i}\}_i$ are the noise standard deviations.

3.4 Optimization of the error function

In practice we minimize the negative logarithm of the posterior defined in Eq. (3). Besides, we use our knowledge of the geometry of the mesh to infer if some of the feature points are occluded under a pose Θ_t. We introduce for each feature i a visibility factor $v_i(X_t)$ defined so that it is equal to 0 when the feature is hidden, and 1 when it is maximally visible: $v_i(X_t) = max(0, \vec{n_{i,t}}(X_t).\vec{z})$. where $\vec{n_{i,t}}(X_t)$ is the normal to the mesh triangle to which the point belongs, and $\vec{z}$ the direction of the camera axis. The visibility of a feature point is taken into account as a weight

Figure 3: Performances of three trackers on the same sequence - Frames 95, 210, 260, 310, 360. From top to bottom: our tracker (Tracker 1), our tracker without using the side mesh (Tracker 2), our tracker using a continuous adaptation scheme (Tracker 3). For clarity, in all cases only the face part of the mesh is drawn.

factor in the likelihood terms of the error function:

$$\begin{aligned} E(X_t) \quad = \quad & -\sum_{i \in S^{str}} v_i(X_t) \cdot log(p(Z_{i,t}^{str}|X_t)) \\ & -\sum_{i \in S^{int}} v_i(X_t) \cdot log(p(Z_{i,t}^{int}|X_t)) \\ & -\sum_{i=1}^{N_p} log(p(X_i|\hat{X}_{i,t-1}))\,. \end{aligned} \tag{10}$$

The downhill simplex method was chosen to perform the minimization. This iterative non-linear optimization method has several advantage: it does not require to derive the error function (which would be difficult to extract in our case) and it maintains multiple hypothesis (which ensures robustness) during the optimization phase. The dimension of the state space being quite large, the optimization is done in two steps: we first run the optimization algorithm to estimate the pose parameters Θ_t, then we estimate the whole state X_t.

4 Experiments and results

Our implementation of the described algorithm processes an average of 3 frames per second. However, execution time was not our priority and we believe that the algorithm could run much faster with minor revisions of the code.

The system was tested on several long video sequences in order to evaluate qualitatively its ability to track challenging head poses and facial actions in natural conditions and evaluate its stability over time, which is our primary aim. However, to provide quantitative evaluation, we also used the BUFT database (Cascia et al., 2000) to measure the precision of the head pose estimation and compare with state-of-the-art results.

4.1 Qualitative results on long video sequences

We tested our system on 8 long video sequences to evaluate its ability to track in the long term the head pose and facial actions. Sample results are given on Fig. 3 and 4, but the quality of the results is better assessed from the videos given as supplementary material. The first sequence is the publicly available Talking Face video from PRIMA - INRIA, a video of a person engaged in a conversation. The second sequence is an extract of a politician's speech in a TV broadcast. The six other sequences are videos that we recently recorded in order to test the system on more challenging head poses and facial actions.

We compared the performances of three trackers. Tracker 1 is the system described in this paper. Tracker 2 is the same as Tracker 1, but with no extension of the Candide model, i.e. no information is collected on the sides of the head. Tracker 3 is the same as Tracker 1, but using a recursive adaption method as proposed in (Lefèvre and Odobez, 2009) instead of the view-based templates.

Not surprisingly, the three systems perform equally well on the first two sequences. These two sequences

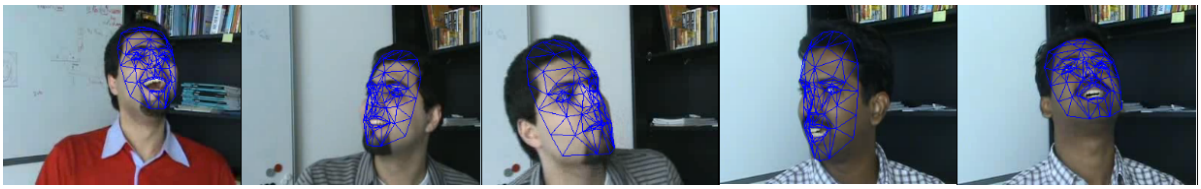


Figure 4: Sample images from various sequences obtained with our tracker

	Uniform-light dataset					Varying-light dataset				
Approach	P_s	E_{pan}	E_{tilt}	E_{roll}	E_m	P_s	E_{pan}	E_{tilt}	E_{roll}	E_m
La Cascia (Cascia et al., 2000)	75%	5.3°	5.6°	3.8°	3.9°	85%	-	-	-	-
Xiao (Xiao et al., 2003)	100%	3.8°	3.2°	1.4°	2.8°	-	-	-	-	-
Morency (Morency et al., 2008)	100%	5.0°	3.7°	2.9°	3.9°	-	-	-	-	-
Adaptation (Lefèvre and Odobez, 2009)	100%	4.4°	3.3°	2.0°	3.2°	100%	4.1°	3.5°	2.3°	3.3°
View-based	**100%**	**4.6°**	**3.2°**	**1.9°**	**3.2°**	**100%**	**6.2°**	**4.4°**	**2.7°**	**4.4°**

Table 1: Comparison on the BUFT database of robustness and accuracy between our approach (in bold) and state-of-the-art face trackers. The Three first results were extracted from the corresponding papers.

are useful to evaluate long-term and subtle lip movements tracking, but the head poses do not go very far from frontal view. The difference of performance between the different approaches shows when they are tested on the more challenging sequences. Sample results obtained by the different systems on the same sequence are illustrated in Fig. 3.

One can notice that Tracker 3 correctly estimates the movement towards profile view, but looses track when trying to come back to a more frontal pose. This phenomenon is actually observed in most of the sequences in which such a movement (frontal-profile-frontal) occurs. Indeed, the information that allows to follow the movement back to frontal view is mainly contained by the intensity features on the head side. As mentioned before, the appearance of these features varies a lot under such pose variations, and the memoryless adaptive system cannot follow.

Tracker 2 is more robust, since it never looses track in all our sequences. Despite the absence of measurements on the head sides, the memory of the learned appearances under different poses allows the tracker to find its way under all kinds of head motions. However, the loss of information compared to Tracker 1 leads to a lack of precision, and thus to a less accurate fit. An example can be seen in Fig. 3. On the second, third and fourth frames the eyes are not correctly fit, and on the fifth image the mouth and the eyebrows are not well positioned.

Out of the three systems, Tracker 1 is the one that demonstrates the best results. The use of a set of view-based templates over an adaptive template for the intensity features allows to robustly track challenging poses, and the extension of the mesh allows to gather more information and leads to an accurate tracking. The system can follow both natural and faked facial action under difficult head poses, as illustrated in Fig. 4.

4.2 Results on the BUFT database

The BUFT database contains 72 videos presenting 6 subjects performing various head movements (translations, in-plane and out-of-plane rotations). Each sequence is 6 seconds long and has a resolution of 320×240 pixels. Ground truth was collected using a "Flock of Birds" magnetic tracker. The databased is divided into two datasets. The Uniform-light dataset contains 45 sequences recorded under constant lighting conditions. The Varying-light dataset contains 27 sequences recorded under fast-changing challenging lighting conditions.

We can define the robustness of a tracker as the percentage P_s of frames successfully tracked over all the video sequences. The accuracy of a tracker is defined as the mean pan, tilt and roll angle errors over the set of all tracked frames: $E_m = \frac{1}{3}(E_{pan} + E_{tilt} + E_{roll})$. We compared the performances of five trackers; the results are shown in Table 1. The "View-based" approach corresponds to the method described in this paper.

One can notice that the results obtained by the Adaptation approach and the View-based approach are very similar. The performances on the Uniform-light dataset are in accordance with our expectations; on such short sequences and only a few profile views we did not expect to observe improvement. On the other hand, we did not expect our system to perform as well on the challenging Varying-light dataset, since it does not incorporate a way to handle fast illumination variations (appearance model updates are much less frequent than in the recursive case), but in the end our coarse estimation of the global illumination changes and the update of the set of templates was enough to successfully track all the sequences with a small loss of accuracy compared to (Lefèvre and Odobez, 2009). Remember however that using this

recursive approach in our modeling often failed on longer sequences, which showed that it was not really stable. When comparing our approach to three other trackers in the literature, we notice that it perform noticeably better than (Cascia et al., 2000) on both datasets. The performances on the Uniform-light dataset are comparable to those demonstrated in (Morency et al., 2008; Xiao et al., 2003). However, we handle the much more challenging Varying-light dataset while none of (Morency et al., 2008; Xiao et al., 2003) demonstrated successfully on this dataset.

5 Conclusion

In this paper we introduced a face tracking method that uses information collected on the head sides to robustly track challenging head movements. We extended an existing 3D face model so that the mesh reaches the ears. In order to handle appearance variation (mainly due to head pose changes in practice), our approach builds online a set of view-based templates. These two distinctive features were proved to be particularly useful when the tracker has to deal with extreme head poses like profile views. Moreover we showed the ability of our approach to follow both natural and exaggerated facial actions. However we are aware that one limitation of our system is that there is no mechanism to recover from a potential failure. One solution would be to add a set of detectors for specific points that could help to set the system back on track.

ACKNOWLEDGEMENTS

This work was supported by the EU 6th FWP IST Integrated Project AMIDA (Augmented Multiparty Interaction with Distant Access) and the NCCR Interactive Multimodal Information Management project (IM2). We thank Vincent Lepetit for useful discussions on the model.

REFERENCES

Ahlberg, J. (2001). Candide 3 - an updated parameterised face. Technical Report LiTH-ISY-R-2326, Linköping University, Sweden.

Cascia, M. L., Sclaroff, S., and Athitsos, V. (2000). Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. In *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, volume 22.

Chen, Y. and Davoine, F. (2006). Simultaneous tracking of rigid head motion and non-rigid facial animation by analyzing local features statistically. In *British Machine Vision Conf. (BMVC)*, volume 2.

Cootes, T., Edwards, G., and Taylor, C. (1998). Active appearance models. In *European Conf. Computer Vision (ECCV)*, volume 2.

Cootes, T., Taylor, C., Cooper, D., and Graham, J. (1995). Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59.

DeCarlo, D. and Metaxas, D. (2000). Optical flow constraints on deformable models with applications to face tracking. *Int. Journal of Computer Vision*, 38(2):99–127.

Dornaika, F. and Davoine, F. (2006). On appearance based face and facial action tracking. In *IEEE Trans. On Circuits And Systems For Video Technology*, volume 16.

Gross, R., Matthews, I., and Baker, S. (2006). Active appearance models with occlusion. *Image and Vision Computing Journal*, 24(6):593–604.

Jepson, A., Fleet, D., and El-Maraghi, T. (2003). Robust online appearance models for visual tracking. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 25(10):1296– 1311.

Lefèvre, S. and Odobez, J. (2009). Structure and appearance features for robust 3d facial actions tracking. In *Int. Conf. on Multimedia & Expo*.

Li, Y. (2004). On incremental and robust subspace learning. *Pattern Recognition*, 37(7):1509–1518.

Matthews, I., Ishikawa, T., and Baker, S. (2004). The template update problem. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 26(6):810 – 815.

Morency, L.-P., Whitehill, J., and Movellan, J. (2008). Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*.

Tu, J., Tao, H., and Huang, T. (2009). Online updating appearance generative mixture model for mean-shift tracking. *Machine Vision and Applications*, 20(3):163–173.

Vacchetti, L., Lepetit, V., and Fua, P. (2004). Stable real-time 3d tracking using online and offline information. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 26(10):1385–1391.

Xiao, J., Baker, S., and Matthews, I. (2004). Real-time combined 2d+3d active appearance models. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*.

Xiao, J., Moriyama, T., Kanade, T., and Cohn, J. (2003). Robust full-motion recovery of head by dynamic templates and re-registration techniques. *Int. Journal of Imaging Systems and Technology*, 13(1):85–94.