3D Face Pose and Animation Tracking via Eigen-Decomposition based Bayesian Approach

Ngoc-Trung Tran[1], Fakhr-Eddine Ababsa[2], Maurice Charbit[1], Jacques Feldmar[1] , Dijana Petrovska-Delacrétaz[3] and Gérard Chollet[1]

[1] ENST, 75014 Paris, France[1]
{trung-ngoc.tran,maurice.charbit,gerard.chollet}@enst.fr
jfeldmar@gmail.com
[2] IBISC, 91020 Evry, France[2]
ababsa@iup.univ-evry.fr
Telecom Sudparis, 91000 Evry, France[3]
dijana.petrovska@telecom-sudparis.eu

Abstract. This paper presents a new method to track both the face pose and the face animation with a monocular camera. The approach is based on the 3D face model CANDIDE and on the SIFT (Scale Invariant Feature Transform) descriptors, extracted around a few given landmarks (26 selected vertices of CANDIDE model) with a Bayesian approach. The training phase is performed on a synthetic database generated from the first video frame. At each current frame, the face pose and animation parameters are estimated via a Bayesian approach, with a Gaussian prior and a Gaussian likelihood function whose the mean and the covariance matrix eigenvalues are updated from the previous frame using eigen decomposition. Numerical results on pose estimation and landmark locations are reported using the Boston University Face Tracking (BUFT) database and Talking Face video. They show that our approach, compared to six other published algorithms, provides a very good compromise and presents a promising perspective due to the good results in terms of landmark localization.

1 Introduction

Tracking 3D face pose is an important issue and has received much attention in the last decades because of *multiple applications* involved such as: video surveillance, human computer interface, biometrics, *etc.* And it is much more challenging if the face animation or expression needs to be recognized in the meantime in variety of applications. Difficulties come from a number of factors such as projection, multi-source lighting biological appearance variations, facial expressions as well as occlusions with accessories, *e.g.,* glasses, hats... In this paper, we present a method using the model of landmarks to track pose efficiently as well as model facial animation. Note that the face is controlled by shape and animation which could be validated as landmark tracking problem.

Since the pioneer work of [1, 2], it is well-known that the Active Shape Model (ASM) and Active Appearance Model (AAM) provide an efficient approach for face pose estimation and tracking landmarks of frontal or near-frontal faces. Some extensions [3, 4] have been developed to improve the method in terms of accurate landmarks or profile-view fitting. Recently, Saragih *et al.* [5] via exhaustive local search around landmarks constrained by a 3D shape model, can track single face of large Pan angle in well-controlled environment. However, it needs a lot of annotated data, which is costly in unconstrained environments, to learn 3D shape and local appearance distributions. One another approach tracks faces and estimate pose uses 3D rigid models such as semi-spherical or cylinder [6, 7], ellipsoid [8] or mesh [9]. These methods can estimate three rotations well even profile-view; however, non-rigid transformation can not be applied for animation problem.

For those who using synthesized databases or online tracking technique with 3D face. An early proposal [10] concerns optical flow and does adaptable changes. Optical flow can be very accurate but not robust on fast movements. Moreover, this approach accumulates errors to drift away and is not easy to recover in long video sequences. With the help of local features, which provides invariant descriptors to non-rigid motions, Chen and Davoine [11] took advantages of local features constrained by a 3d-face paramerized model, called Candide-3, to capture both rigid and non-rigid head motions. But this methods does not work well in profile-view due to the large variation of landmarks. Ybanez *et al.* [12] found linear correlation between 3D model parameters and global appearance of stabilized face images. This method is robust for face and landmark tracking but limited just around frontal faces. Lefevre *et al.* [13] extended Candide by collecting more appearance information at profile-views and chose more random points to represent facial appearance. Their error function consists of structure and appearance features combined with dynamic modeling, is high dimension and is easy to fall into local minimum. Recently, faceAPI [14] showed impressive results in pose and face animation tracking; however, this is a commercial product that unable to be accessed to investigate and compare with other methods.

In this paper, we propose an Bayesian method using a 3D face model to build the face pose and animation tracking framework. Our contribution is that in our framework, the SIFT [15] is supposed to be local descriptor to track landmarks which are constrained by the 3D shape. And eigen decomposition is proprosed to use through Singular Value Decomposition (SVD) to update the tracking model robustly and balance between what we learned in training and what we are seeing at the moment. This approach is different what previous methods of face tracking did. We also take advantages of a synthesized database [11–13] without the need of big annotated data and propose the use of robust features to rigid and non-rigid changes. During tracking, candidate of new pose and animation is estimated via the posterior probability and the appearance model are then adjusted from new observations to environmental changes. This technique can make the system robust to changes of facial expression, pose and as well as environmental factors. The results on two public datasets show that our approach, compared to six other published algorithms, provides a very good compromise in terms of pose estimation and landmark localization.

The remaining of this paper are organized as follow: Section 2 gives some background face representation. Section 3 shows the proposed framework for tracking. Experimental results and analysis are presented in Section 4. Finally, we draw conclusions in Section 5.

2 Face Representation

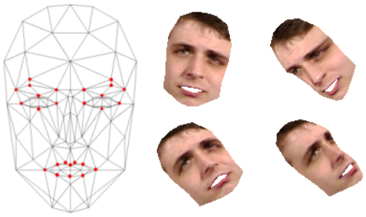

Fig. 1. Candide-3 and some sample synthesized images.

Candide-3 [16] is a very commonly used face shape model. It consists of 113 vertices and 168 surfaces. Fig. 1 represents the frontal view of the model. It is controlled both in translation, rotation, shape and animation:

$$\mathrm{g}(\sigma,\alpha) = \mathrm{R}s\left(\overline{\mathrm{g}} + \mathrm{S}\sigma + \mathrm{A}\alpha\right) + \mathrm{t} \tag{1}$$

where $\overline{\mathrm{g}}$ is 3N-dimensional mean shape (N = 113 is the number of vertices) containing the 3D coordinates of the vertices. The matrices S and A control respectively shape and animation through σ and α parameters. R is a rotation matrix, s is the scale, and t is the translation vector. The model makes an perspective projection assumption to project 3D face onto 2D image. Like [11, 13, 12], only 6 dimensions r_a of the animation parameter are used to track eyebrows, eyes and lips. Therefore, the full model parameter b of our framework has 12 dimensions: of 3 dimensions for rotation (r_x, r_y, r_z), 3 dimensions for translation (t_x, t_y, t_z) and 6 dimensions for animation r_a:

$$\mathrm{b} = [r_x, r_y, r_z, t_x, t_y, t_z, r_a] \tag{2}$$

Texture model: In the Candide model, appearance or texture parameters are not available. Usually, we warp and map the image texture onto the triangles of the 3d mesh by the image projection.

3 Proposed Method

Our framework consists of two steps: training and tracking. The framework benefits a database of synthesized faces to train tracking model and applies new way of tracking face pose and animation. In this section, we describe our method in detail.

3.1 Training

In the work of [11], the authors align manually the Candide model on the first video frame and warp and map the texture from the image to the model. In our work, landmarks are annotated manually on the first video frame, then the POSIT algorithm [17] is used to fit and estimate the pose automatically from these landmarks to get the initial model parameters b_0.

The acquisition of ground-truth is very costly and time consuming. In order to circumvent this drawback, synthetic database [11–13] using the Candide model is a good alternative. In order to collect training data, we do three following steps to obtain images using Candide and build appearance model for the next tracking step:

Data Generation After initialization, the texture is warped and mapped from the first video frame to the Candide model. Our database is built by rendering different views around the frontal image. Note that the full dimension of the parameters to track is 12, consists of pose and animation, that makes difficult to explore finely. However, the translation parameters t_x and t_y will not affect the face appearances as well as facial animation will not be significant influence because the use of local features in tracking. Hence, only rotations are gridded for building the training database. Specifically, 7 values of Pan and Tilt and Roll from -30 to +30 by step of 10 are taken to create $7^3 = 343$ pose views as some examples in Fig. 1.

Learning Appearance Model The framework adopts local descriptors which are robust to rigid and non-rigid motion. In this paper, we also use SIFT descriptor [15] to extract local features around 26 given landmarks in Fig. 1 as observed appearance. SIFT is invariant to affine transformation and helpful to localize accurate landmarks. In order to get the appearance model, we compute mean and covariance matrices of landmark descriptors on 343 images of the synthesized database which is generated from the first image. Each pair of mean and covariance matrix (μ^i, Σ^i) plays the role of learning data for ith landmark which are 128×1 and 128×128 matrices respectively. And these matrices will be adjusted during tracking.

3.2 Tracking

Here we propose a Bayesian approach approximated from posteriori distribution:

$$p(b_t|Y_{1:t}) = \frac{p(Y_t|b_t, Y_{1:t-1})p(b_t|Y_{1:t-1})}{p(Y_t|Y_{1:t-1})} \propto p(Y_t|b_t, Y_{1:t-1})p(b_t|Y_{1:t-1}) \qquad (3)$$

Equation (3) is normally controlled by the observation model $p(Y_t|b_t, Y_{1:t-1})$, and the evolution $p(b_t|Y_{1:t-1})$ as the prior. Because Eq. 3 is still complicated to solve, we provide some assumptions to make it simpler.

Evolution Model The model $p(b_t|Y_{1:t-1})$ of state b_t is dependent on only previous observation $Y_{1:t-1}$. We know $\hat{b}_{t-1}$ was able to estimated from $Y_{1:t-1}$. So, we assume that $p(b_t|Y_{1:t-1}) \propto p(b_t|\hat{b}_{t-1})$ which means b_t is modeled independently by a Gaussian distribution around its previous estimated state $\hat{b}_{t-1}$, where $b_t = (r_x, r_y, r_z, t_x, t_y, t_z, r_a)_t$ is the 12-dimensional vector in our context expressed as:

$$p(b_t|\hat{b}_{t-1}) = \mathcal{N}(b_t; \hat{b}_{t-1}, \Psi) \tag{4}$$

where Ψ is a diagonal covariance matrix whose elements are the corresponding variances of parameters of the state vector $\sigma^i, i = 1, .., 12$. This model can be considered as the prior information during tracking.

Observation Model The tracking system starts from the frontal face where Candide is fitted onto, and then it finds the candidate of face in the next frame $t + 1$ from the state vector at time t, with $t = 0$ at the first frame. In order to obtain the observation Y_t, the 3d Candide model is projected onto the next 2D frame at t to localize 2D landmark positions. The appearance Y_t is a vector of local textures $(y_t^1, y_t^2, ..., y_t^n)$ around these landmarks as the observation. These observations can then be used to establish the observation model for tracking and the crucial point is to find an efficient observation model.

We make the assumption that the local appearances around landmarks are independent. The observation model is defined as a joint probability of Gaussian distributions, and the tracking problem can be solved as a maximum likelihood problem of a non-linear function.

$$p(Y_t|b_t, Y_{1:t-1}) = \prod\nolimits_{i=1}^{n} p(y_t^i|b_t, y_{1:t-1}^i) \tag{5}$$

It means that the observation Y_t is dependent on the state variable b_t as well as previous observations Y_{t-1}. Since the database of synthesized faces is generated in the range limit of $(-30; 30)$ of three rotations that make the system limited in profile tracking. We can generate more data, however, it makes the framework less robust because of the variation for patches as well as occlusion problem at profile-view. Additionally, there are many factors such as illumination, poses and facial expression that may affect to tracking. So, the learning model needs to be adaptive to changes of environment that brings us the idea of maximum likelihood problem (5) can be rewritten as follows:

$$p(Y_t|b_t, Y_{1:t-1}) = \prod\nolimits_{i=1}^{n} \mathcal{N}(y_t^i|\mu_t^i, \Sigma_t^i) \tag{6}$$

where n is the number of landmarks, $\mathcal{N}(y_t^i|\mu_t^i, \Sigma_t^i)$ denotes multivariate Gaussian distribution of function value at observation around the ith landmark y_t^i, and μ_t^i and Σ_t^i are mean and covariance matrices updated at time t during tracking. Note that μ_0^i and Σ_0^i are pre-learned mean and covariance in training step at first frame. The likelihood in Eq. 6 is controlled by two terms: μ_t^i and Σ_t^i which model how confidence the new landmark observation is. Since trained at first frame, these terms should be adjusted to fit changes of factors, but still "remember" what it learned before. The proposed way how to update can be described as follows for mean vectors:

$$\mu_t^i = (1-\alpha)\mu_{t-1}^i + \alpha y_{t-1}^i \tag{7}$$

where forgetting factor $\alpha \in (0,1)$ is a constant. This equation is a way to correct the error between the observation and the mean vector of appearance model. In order to update covariance matrices, Singular Value Decomposition (SVD) [18] is used to factorize the previous covariance matrix at time $t-1$ into unitary matrices and singular matrix of eigen values: $svd(\Sigma_{t-1}^i) = [U_{t-1}^i, S_{t-1}^i, (U_{t-1}^i)^T]$. Note that covariance matrix is positive definite, so unitary matrices are the same. Then, updating the singular matrix before composing all of them back to obtain a new covariance matrix at time t.

$$S_t^i = (1-\alpha)S_{t-1}^i + \alpha \left\| y_{t-1}^i - \mu_{t-1}^i \right\|_2^2 I \quad \texttt{and} \quad \Sigma_t^i = U_{t-1}^i S_t^i (U_{t-1}^i)^T \tag{8}$$

where I is identity matrix, $\|.\|_2$ is norm-2. The equations denote how to do adaptive observation model, while keeping principal components of what is seen before. In order to do this, we use Eq. 7 for the eccentricity and the direction is changed when the new covariance matrix is decomposed to update in next step as Eq. 8. The updated mean and covariance matrices are used to model the observation as Eq. 6. To sum up, replacing the observation and evolution models respectively of equations (4) and (6) into (3) and taking the log of likelihood, we finally attempt to minimize the error function approximated as follows:

$$\hat{b}_t = \arg\min_{b_t} \sum_{i=1}^{n} \left\| y_t^i - \mu_t^i \right\|_{(\Sigma_t^i)^{-1}}^2 + \left\| b_t - \hat{b}_{t-1} \right\|_{\Psi^{-1}}^2 \tag{9}$$

where $\hat{b}_{t-1}$ is the model parameter estimated from previous frame. In our optimization context, the error function in (9) is a multi-dimensional function of the model parameter b_t that we wish to minimize. It is not easy to solve analytically, so a derivative-free optimizer such as down-hill simplex [19] is preferred. Like [11], thirteen initial points are chosen randomly around the current state (12-dimensional space) to form the simplex and the solution that subjects to local minimum can be found by deformations and contracts during optimization.

4 Experimental Results

We adopted the Boston University Face Tracking (BUFT) database [6] and Talking Face video[3] to evaluate the performances of face pose estimation and its animation by landmark tracking respectively.

BUFT: The pose ground-truth is captured by magnetic sensors "*Flock and Birds*" with an accuracy of less than 1^o. The uniform-light set which is used to

[3] http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html

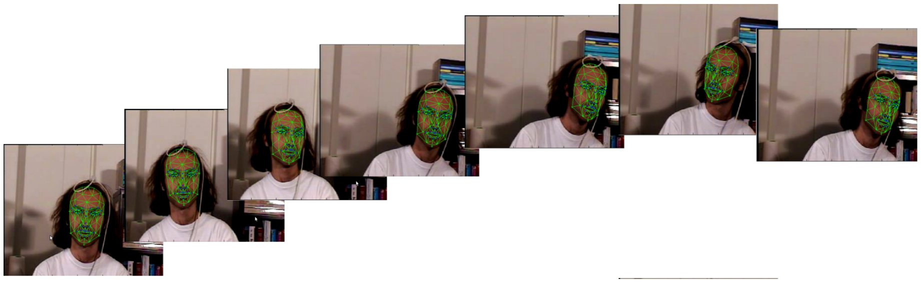

Fig. 2. An sample result of our method on one BUFT video.

evaluate, has a total of 45 video sequences (320×240 resolution) for 5 subjects (9 videos per subject) with available ground-truth which is formatted as (*X_pos, Y_pos, depth, roll, yaw (or pan), pitch (or tilt)*).

For each frame of one video sequence, we use the estimation of the rotation error $e_i = [\theta_i - \hat{\theta}_i]^T[\theta_i - \hat{\theta}_i]$ like [13] to evaluate the accuracy and robustness, where θ_i and $\hat{\theta}_i$ are $(pan, tilt, roll)$ of the ground-truth and estimated pose at frame i respectively. A frame is lost when e_i exceeds the threshold. The robustness is the number N_s of frames tracked successfully and P_s is the percentage of frames tracked over all videos. The precision measures include Pan, Tilt, Roll and average rotation errors which are computed by Mean Absolute Error (MAE) as the measure of tracker accuracy over tracked frames: $E_{pan}, E_{tilt}, E_{roll}$ and $E_m = \frac{1}{3}\left(E_{pan} + E_{tilt} + E_{roll}\right)$ where $E_{pan} = \frac{1}{N_s}\sum_{i \in S_s} |\theta^i_{pan} - \hat{\theta}^i_{pan}|$ (similarly for the tilt and roll) and S_s is set of tracked frames.

The Talking Face Video: is a freely 5000-frames video sequence of a talking person with face animations. The ground-truth is available with 68 facial points annotated manually on the whole video. Basing on movements of landmarks, we can estimate the face animation. On that account, we instead evaluate the precision of landmark tracking as the accurate animation. The Root-Mean-Squared (RMS) error is normally used to evaluate the landmark tracking performance on this database. Despite that the number of landmarks of our system and other methods is different, the same evaluation scheme could be still applied on same number of landmarks with our work as well as other comparative methods.

The performance of pose estimation in Table 1 shows the comparable results between our work and state-of-the-art methods in 3d pose tracking. Our performance is 100% robustness and the accuracy E_m is 3.9, which outperforms [11] and [6] both in terms of robustness and accuracy. And it gets the same result of mean error E_m as [5, 8], but the variance of error of [5] is higher than our work especially in Tilt. However, we are worse than [7, 13] at the accuracy. In spite of the fact that our result is quite encouraging, the Pan precision is still low compared to others. The reason why Pan rotation is bad-estimated, could probably comes from occlusion problem. When Pan is bigger than, for instance, 30^o, some landmarks are occluded that make local descriptors is inefficient that make the likelihood discontinued. For [5], the authors trained their landmarks

Table 1. The comparison of robustness (P_s) and accuracy (E_{pan}, E_{tilt}, E_{roll} and E_{avg}) between our method and state-of-the-art on uniform-light set of BUFT dataset.

Approach	P_s	E_{pan}	E_{tilt}	E_{roll}	E_{avg}
(La Casicia *et al.*, 2000) [6]	75%	5.3	5.6	3.8	3.9
(Xiao *et al.*, 2003) [7]	100%	3.8	3.2	1.4	2.8
(Lefevre *et al.*, 2009) [13]	100%	4.4	3.3	2.0	3.2
(Morency *et al.*, 2008) [8]	100%	5.0	3.7	2.9	3.9
(Saragih *et al.*, 2011) [5]	100%	4.3 ± 2.2	4.8 ± 3.3	2.6 ± 1.4	3.9
(Chen *et al.*, 2006) [11]	91%	5.5 ± 1.7	4.2 ± 1.5	2.1 ± 1.0	3.9
Our method	**100%**	**5.4 ± 2.2**	**3.9 ± 1.7**	**2.4 ± 1.4**	3.9

classifiers only with variation of Pan angles that make their estimation of Tilt and Roll inefficient. Fig. 2 is an example of our method on one video of BUFT dataset.

In order to evaluate the landmark precision, we compare our method and FaceTracker[4] proposed by [5]. Because the landmarks of our method, [5] and ground-truth are not the same, 12 landmarks around eyes, nose and mouth as in Fig. 3 are chosen to evaluate RMS error. The Fig. 4 shows the (Root Mean Square) RMS error which is computed using our method (red curve) and FaceTracker (blue curve) on the Talking Face video. The vertical axis is RMS error (in pixel) and the horizontal axis is the frame number. The model of [5] sometimes drift away the ground-truth, but recovers quickly to good location by benefiting face and landmark detectors. The Fig. 4 shows that even though our method just learned from the synthesized database, what we obtain is the same the state-of-the-art method as well and is even more robust.

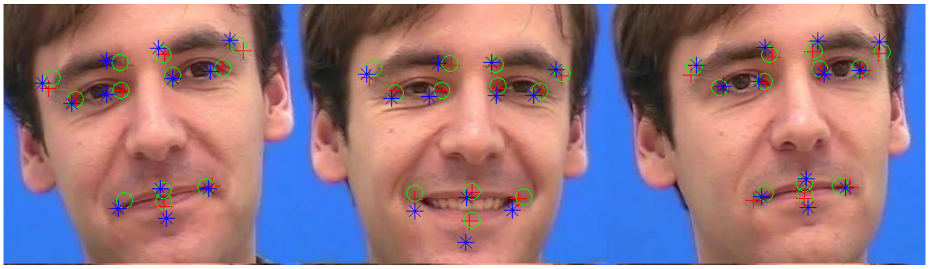

Fig. 3. The 12 landmarks is used to compute RMS error where red (+), blue (∗) and green (o) markers are ground-truth, of Saragih *et al.* [5] and our method respectively on frames 110, 2500 and 4657 of Talking Face video.

The performance of our method for pose estimation could be improved if the Pan was estimated more accurately. One possible solution is assigning weights

[4] http://web.mac.com/jsaragih/FaceTracker/FaceTracker.html

to landmarks corresponds to the Pan value. Or projecting landmarks on tangent plane at each landmark that compute mean and covariance matrices as a function of face pose to deal with occlusion. In general, how to deal with occluded landmarks is one of critical points to improve our performance. Although real-time computation is unreachable (about 5s/frame on Laptop Core 2 Duo 2.00GHz, 2G RAM) due to using down-hill simplex algorithm to optimize the energy function, it can be improved by using Gradient Descent in future work.

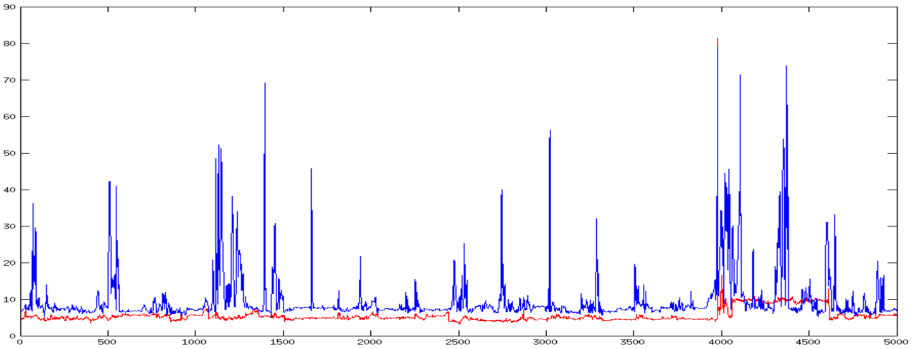

Fig. 4. The RMS error of 12 selected points for tracking in our framework (below red curve) and Saragih *et al.* [5] (above blue curve). The vertical axis is RMS error (in pixel) and the horizontal axis is the frame number.

.

5 Conclusion

In this paper, we propose a Bayesian method to deal with the problem of face tracking using one adaptive model through eigen decomposition. The synthesized database within local features are around landmarks to learn appearance model as mean and covariance matrices. For tracking, an energy function which is approximated from posterior probability is minimized as difference between the observations and the appearance model. In order to adjust the model to changes of environments, the eigen decompostion is deployed. The results showed that the use of our model is comparable to some state-of-the-art methods of pose estimation and much more robust than state-of-the-art at landmark tracking or animation tracking. It demonstrated what we proposed is useful to both tasks of pose estimation and landmark tracking. Moreover, it is easy to build the learning database of synthesized images to learn without the need of real annotated data. With our current encouraging results, some other evolutions could be done to improve the performance. For examples, taking into account the weights of contribution to energy function which is dependent on the confidence of landmark observations at each time, computing appearance model as function of the pose to make the objective function continuous. In general, the way how to improve Pan precision by dealing with occluded landmarks is a crucial point to think as future work. Finally, the speed can be improved to real-time application by using Gradient Descent like methods instead of down-hill simplex algorithm.

References

1. Cootes, T.F., Taylor, C.J.: Cj.taylor, "active shape models - "smart snakes. In: BMVC. (1992)
2. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. TPAMI (1998) 484–498
3. Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real-time combined 2d+3d active appearance models. In: CVPR. (2004)
4. Gross, R., Matthews, I., Baker, S.: Active appearance models with occlusion. IVC **24** (2006) 593–604
5. Saragih, J.M., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. IJCV **91** (2011) 200–215
6. Cascia, M.L., Sclaroff, S., Athitsos, V.: Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. IEEE Trans. PAMI **22** (2000) 322–336
7. Xiao, J., Moriyama, T., Kanade, T., Cohn, J.: Robust full-motion recovery of head by dynamic templates and re-registration techniques. International Journal of Imaging Systems and Technology **13** (2003) 85 – 94
8. Morency, L.P., Whitehill, J., Movellan, J.R.: Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In: FG. (2008)
9. Vacchetti, L., Lepetit, V., Fua, P.: Stable real-time 3d tracking using online and offline information. IEEE Trans. PAMI **26** (2004) 1385–1391
10. DeCarlo, D., Metaxas, D.N.: Optical flow constraints on deformable models with applications to face tracking. IJCV **38** (2000) 99–127
11. Chen, Y., Davoine, F.: Simultaneous tracking of rigid head motion and non-rigid facial animation by analyzing local features statistically. In: BMVC. (2006)
12. Ybáñez-Zepeda, J.A., Davoine, F., Charbit, M.: Local or global 3d face and facial feature tracker. In: ICIP. Volume 1. (2007) 505–508
13. Lefevre, S., Odobez, J.M.: Structure and appearance features for robust 3d facial actions tracking. In: ICME. (2009)
14. FaceAPI: (http://www.seeingmachines.com)
15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
16. Ahlberg, J.: Candide-3 - an updated parameterised face. Technical report, Dept. of Electrical Engineering, Linkping University, Sweden (2001)
17. Dementhon, D.F., Davis, L.S.: Model-based object pose in 25 lines of code. IJCV **15** (1995) 123–141
18. Golub, G., Kahan, W.: Calculating the singular values and pseudo-inverse of a matrix. Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis **2** (1965) 205–224
19. Nelder, J.A., Mead, R.: A simplex algorithm for function minimization. Computer Journal (1965) 308–313