

A Pose-adaptive constrained local model for accurate head pose tracking

Lucas Zamuner, Kevin Bailly, Erwan Bigorgne

▶ **To cite this version:**

Lucas Zamuner, Kevin Bailly, Erwan Bigorgne. A Pose-adaptive constrained local model for accurate head pose tracking. International Conference on Pattern Recognition (ICPR 2014), 2014, Stockholm, Sweden. pp.1-6, 10.1109/ICPR.2014.436 . hal-03181849

HAL Id: hal-03181849
https://hal.science/hal-03181849

Submitted on 25 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Pose-Adaptive Constrained Local Model For Accurate Head Pose Tracking

Lucas Zamuner
Eikeo
11 rue Leon Jouhaux,
F-75010, Paris, France
lucas.zamuner@eikeo.com

Kevin Bailly
Sorbonne Universités
UPMC Univ Paris 06
CNRS UMR 7222, ISIR
F-75005, Paris, France
kevin.bailly@upmc.fr

Erwan Bigorgne
Eikeo
11 rue Leon Jouhaux,
F-75010, Paris, France
erwan.bigorgne@eikeo.com

***Abstract*—Robust and precise face tracking under unconstrained imaging conditions is still a challenging task. Recently, the Constrained Local Model (CLM) framework has proven to be very powerful to track frontal and near frontal facial movements. In this paper, we introduce a Pose-Adaptive CLM which is able to accurately track large 3D head rotations. This model relies on two main parts: (1) an adaptive 3D Point Distribution Model that ensures consistency between a tracked point in the image and the corresponding point in the shape model and (2) an adaptive appearance model that deals with appearance variation of a point under different viewing angle. We present comparative experimental results highlighting the improvement in both robustness and accuracy of our method. We also introduce a new challenging dataset with accurate head pose annotation.**

***Keywords*—head pose; head tracking; constrained local model; 3D model; facial landmarks**

I. Introduction

Head pose is a crucial step in many facial processing applications such as human-machine interaction, social signal processing, biometrics or media analysis. While it has received extensive attention in the last two decades, robust and accurate head pose estimation in real life applications remains a challenging task as the tracker has to deal with occlusions, illumination changes, facial expression...

Numerous approaches have been suggested to recover the 3D head pose motions [1] including detector arrays [2][3] and nonlinear regression [4] [5]. They greatly differ in term of accuracy, processing time and robustness. Recently, a large part of head pose estimation methods are explicitly based on shape models of the face. They differ in the complexity of the underlying model that may be a simple plane [6], an ellipse [7] or extracted from actual human faces [8] and in the way they track the global appearance of the face, *e.g.*particule filtering [9], [7] or iteratively reweighted least-squared [10], [8].

Among these methods, Parameterised Appearance Models have demonstrated particularly good performances. First developed by Cootes and Taylor, Active Shape Models (ASM) [11] established the basis of this framework, in which a 2D Point Distribution Model (PDM) is fitted into the image under local constraints. Active Appearance Models [12] further improved the ASM by conveying a coherent way to jointly optimize the shape and the appearance models parameters. Additional improvements have been made to these pioneering works: 3D-AAM [13] and 3D-Morphable Models [14][15] used 3D-shapes of the face while the ASM and AMM work only on the alignment of 2D-shapes in the image frame.

More recently, the Constrained Local Model (CLM) [16] and Subspace Constrained Mean-Shifts (SCMS) [17] have shown to be able to robustly and efficiently track local facial landmarks. CLM implements a two-step strategy: Firstly a local and exhaustive search of each landmark is performed using dedicated landmark detectors and secondly, parameters of a PDM are jointly optimized over all the detector responses. The local support of the appearance model, combined with a global shape constrained regularization makes CLM very robust to identity, illumination and occlusion.

Unfortunately, this approach has two severe limitations to track large pose variation. Let consider the tracking of a particular point on the jawline using a 3D PDM. If the face is frontal, the point is visible and easily tracked. If the head performs a rotation along the vertical axis, the target point will be occluded and the appearance model will hardly cope with this change. In this case, the point tracked by the CLM tends to be located on the edge of the cheek rather than to its actual position on the jawline. In order to address the inconsistency issue between the tracked point in the image and the corresponding point of the PDM, we propose a pose-adaptive 3D PDM which is dynamically updated.

The second main limitation relates to the local landmark detectors. The appearance of landmarks drastically changes according to the viewing angle. If detectors are only trained using near frontal samples, the CLM will accurately track these landmarks in the same condition but will fail to track them under large 3D rotations. Conversely, to train a unique detector per landmark with frontal and profile landmark samples will lead to inaccurate detections because it will be difficult to capture the heterogeneity of the landmark appearance. We overcome this limitation by using pose specific appearance model: dedicated detectors are trained for different groups of pose and during the tracking phase, detectors are selected and combined according to the current estimate of the pose.

Evaluations are carried out on two publicly available datasets: the Biwi Kinect head pose database (Biwi) [18], [19] and our new Head Pose Database (this dataset will be referred as ISIR-Eikeo Database in the rest of the article).

The remaining paper is organized as follows: Section

II briefly explains the general CLM framework. Section III presents our Pose-Adaptive CLM approach. In particular, we focus on the pose-adaptive PDM (III-A) and the pose specific appearance model (III-B). Section IV presents our new head pose dataset and the performance of our method in this challenging context.

II. Constrained Local Models

CLM fitting consists mainly in a joint optimization of the PDM parameters to minimize the misalignment error over all landmarks constituting the PDM. Recent works falling into the CLM framework [20][17][21][22] demonstrate overall good performances, all taking advantage of a two-stage implementation: local response maps are firstly computed around each landmark support. Optimization is then performed over these response maps. Saragih *et al.*[17] draw a comparison between sophisticated strategies that can be used and further propose a fitting by Regularized Landmark Mean-Shift (RLMS). Among them we choose this last method which shows particularly good results and has been extended to RGB-D images [21].

The Point Distribution Model we use describes an explicit 3D-Model coming under a weak perspective parameterization. Let $\boldsymbol{x}_i$ be the projection in the image of one landmark of the PDM given by:

$$\boldsymbol{x}_i(\boldsymbol{q}) = s\boldsymbol{R}(\bar{\boldsymbol{x}}_i + \boldsymbol{\psi}_i\boldsymbol{p}) + \boldsymbol{T}, \tag{1}$$

where $\bar{\boldsymbol{x}}_i$ denotes the mean shape of the face model and $\boldsymbol{q} = \{s, \boldsymbol{R}, \boldsymbol{p}, \boldsymbol{T}\}$ determines the current state of the model where s is a global scaling, $\boldsymbol{R}$ and $\boldsymbol{T}$ are the 3D rigid transformation parameters, and $\boldsymbol{p}$ account for the non rigid deformation along the eigenvectors $\boldsymbol{\psi}_i$ of the shape model.

In a probabilistic framework, the fitting consist in maximizing the posterior probability over the PDM parameters:

$$p(\boldsymbol{q} \mid \{l_i = 1\}_{i=1}^{n}, \mathcal{I}) \propto p(\boldsymbol{q}) p(\{l_i = 1\}_{i=1}^{n} \mid \boldsymbol{q}, \mathcal{I}) \tag{2}$$

where l_i is a discrete variable accounting for the alignment of the i^{th} landmark. Considering the conditional independence on detections for each landmark, this equation can be written as:

$$p(\boldsymbol{q} \mid \{l_i = 1\}_{i=1}^{n}, \mathcal{I}) \propto p(\boldsymbol{q}) \prod_{i=1}^{n} p(l_i = 1 \mid \boldsymbol{x}_i, \mathcal{I}) \tag{3}$$

The likelihood $p(l_i = 1 \mid \boldsymbol{x}_i, \mathcal{I})$ refers to the local appearance model and will be detailed further (section III-B). Following the RLMS Strategy[1], the likelihood maps are approximated by means of Kernel Density Estimates, offering a good trade off between a smooth estimation of the response map, which is often noisy and multi-modal, and precision about the true landmark position. An Expectation-Maximization algorithm is then used for maximizing the MAP term of (3). Please refer to [23] for a detailed description of this process.

[1]An Implementation of the RLMS algorithm is freely available at https://github.com/kylemcdonald/FaceTracker (accessed Dec. 2013)

III. Pose-Adaptive Constrained Local Model

A. Pose-Adaptive Point Distribution Model

The CLM-class methods appeared to be effective to address alignment problems based either on a 2D modeling [16][17] or on a 3D Point Distribution Model [24]. All the methods that run under this paradigm must cope with the necessary consistency between the PDM and the tracked 2D facial landmarks. Three main kinds of facial landmarks can be distinguished.

The first category includes landmarks that are stable both in appearance and semantically. By semantic stability we mean that each point of the PDM corresponds to a well identified real-world feature such as the of the mouth and eyes. These landmarks are really useful but not sufficient to infer the 3D head pose for two reasons. First, they are mostly located in the frontal plane and any estimate only based on such points tends not to be accurate enough and, second, landmarks with a discriminative appearance tends to be located on the deformable parts of the face.

To overcome this limitation, it is important to include out of the frontal plane landmarks such as those located along the jaw. This second category of points are semantically stable but, unfortunately, their appearance varies greatly depending on the orientation of the head and are often self-occluded.

For this reason, we are interested in the last category that includes semantically weak landmarks with a stable and discriminant appearance such as points remaining along the *visible* edge of the cheek (see Fig. 1).

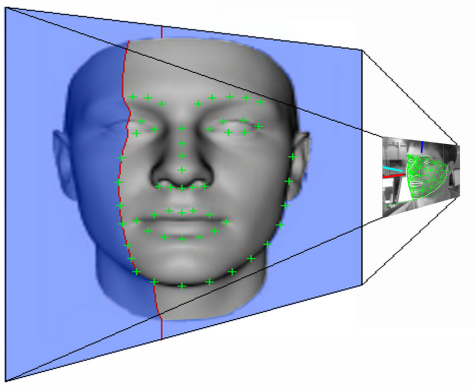


Fig. 1. The Pose Adaptive Point Distribution Model. Given a pose hypothesis, the PDM continuously describes a set of visible points among which the "limbs" remain along the visible edge of the cheek

These points are not related to existing points in the classical 3D PDM and we have to introduce a pose adaptive form for the PDM. The sparse set of points along the cheek is dynamically updated using a 3D dense model according to the current estimate of the pose:

$$\boldsymbol{x}_i(\boldsymbol{q}) = s\boldsymbol{R}(\bar{\boldsymbol{x}}_i(\widetilde{\boldsymbol{R}}) + \boldsymbol{\psi}_i(\widetilde{\boldsymbol{R}})\boldsymbol{p}) + \boldsymbol{T}, \tag{4}$$

where $\tilde{\boldsymbol{R}}$ corresponds to the estimate of the current pose. $\tilde{\boldsymbol{R}}$ is initialized with the value in the previous frame and is dynamically updated during the fitting process.

Using this new PDM, the Jacobian matrix involved in the M step of the EM optimization is modified. An analytical

expression for $\bar{\boldsymbol{x}}_i(\widetilde{\boldsymbol{R}})$ and $\boldsymbol{\psi}_i(\widetilde{\boldsymbol{R}})$ is not available and therefore the contribution of the PDM to the new Jacobian matrix is numerically estimated all along the fitting process.

B. Pose-Adaptive Appearance Model

The appearance of a landmark varies over different head poses. For this reason, we decided to train one specific landmark detector per class of pose $\mathbf{R}^k$. The probability of alignment at a particular landmark location $\mathbf{x}$ is given by:

$$p(l_i = 1 \mid \mathbf{x}, \mathbf{R}^k, \mathcal{I}) = \frac{1}{1 + e^{(aC_i(\mathbf{x};\mathbf{R}^k,\mathcal{I}))}} \tag{5}$$

where $C_i(\mathbf{x}; \mathbf{R}^k, \mathcal{I})$ is a classifier that discriminates aligned from misaligned locations and a is the regression coefficient evaluated by cross validation. In this study, we use the same regressor as [20] and [17]:

$$C_i(\mathbf{x}; \mathbf{R}^k, \mathcal{I}) = \mathbf{w}_i^T(\mathbf{R}^k)\mathcal{F}(\mathbf{x}; \mathcal{I}) + b_i(\mathbf{R}^k) \tag{6}$$

where $\mathbf{w}_i^T(\mathbf{R}^k)$ and $b_i(\mathbf{R}^k)$ are the gain and the bias of the classifier and $\mathcal{F}(\mathbf{x}; \mathcal{I})$ are the features extracted from pixels in the neighborhood of the landmark location $\mathbf{x}$. In our case, grey levels are concatenated with intensity gradients and edge directions in order to enhance the discriminative power of our patches.

During testing, the head pose is roughly estimated, (*e.g.*using the pose of the previous frame) and the responses of the two regressors corresponding to the closest poses are linearly combined. Since the two detectors are linear classifiers, this is equivalent to interpolate the bias and the gain of a classifier for current pose using the selected two classifiers. Let $\mathbf{R}^{k_1}$ and $\mathbf{R}^{k_2}$ be the two closest poses of the current pose estimate $\mathbf{R}$, the interpolated classifier is defined by:

$$\mathbf{w}_i^T(\mathbf{R}) = \alpha \mathbf{w}_i^T(\mathbf{R}^{\mathbf{k_1}}) + (1 - \alpha)\mathbf{w}_i^T(\mathbf{R}^{\mathbf{k_2}}) \tag{7}$$

$$b_i(\mathbf{R}) = \alpha b_i(\mathbf{R}^{\mathbf{k_1}}) + (1 - \alpha)b_i^T(\mathbf{R}^{\mathbf{k_2}}) \tag{8}$$

The α coefficient defines the impact of each pose class on the final classifier. In this study, α is computed using only the pan angle.

IV. Experiments

To evaluate the improvement of our approach over CLM, we conducted experiments on two publicly available databases and we assessed the improvement brought by each element of the method.

A. Datasets

a) Biwi Kinect Head Pose Database: It consists of 24 video sequences collected using the Microsoft Kinect sensor. The head pose range covers about ± 75 degrees pan, ± 60 degrees tilt and ± 50 degrees roll. The head pose ground truth is obtained using a state-of-the-art template-based head tracker. The authors estimate that the mean rotation errors were around 1 degree. As pointed out by Fanelli *et al.*[18] and Baltrušaitis *et al.*[21] some consecutive frames are missing. Strong pose variations combined with missing frames make this dataset really challenging, especially for tracking based methods.

b) ISIR-Eikeo Database: In order to conduct extensive head pose tracking evaluation with a very precise ground truth, we decided to create a new database. This dataset includes 40 sequences of 20 different subjects. They were completely free of their movements and they were asked to move an object from one place to another to collect behaviors as natural as possible. For each subject, the first sequence is acquired with a ambient light of 112 lux and the second one with a lateral illumination of 215 lux. Image sequences are recorded with a HD camcorder at a framerate of 24 fps and the ground truth is obtained by a Codamotion system[2]. It results in a set of video sequences with a wide variety of poses, lightings, occlusions, identities and a precise ground truth as illustrated in figure 3. This database will be freely distributed on request.

B. Evaluation protocol and implementation details

Tests are carried out on the same 4 Biwi video sequences as in [21] and 4 video sequences of ISIR-Eikeo Database. We selected sequences with representative sources of variations. For each sequence, the model is automatically initialized using the Viola Jones frontal face detector (OpenCV implementation). At each frame, the quality of the alignment is checked by computing the response of each local landmark detector at the final position of the model. If the sum of these responses is lower than a threshold estimated by a cross validation process, the tracking is stopped and reinitialized when the face is detected again. For each frame, we measure the Mean Absolute Error (MAE) for each Euler angle.

Four systems are evaluated and compared. In this paper, the CLM refers to the freely available implementation of RMLS provided by Jason Saragih[3]. To obtain a fair comparison between our approach and CLM, we slightly modified the original CLM in two ways: (1) landmark detectors are retrained using the same features and the same training samples as our model and (2) each self-occluded landmark is removed from the model and the regression of the model is only performed based on the remaining points. This implementation will be refered as *modified CLM*. The third system, referred as *PA-PDM* is a modified CLM with a Pose-Adaptive Point Distribution Model (III-A). *PA-CLM* is our final system, which combines the two proposed improvements, the PA-PDM (III-A) and the pose adaptive appearance model (III-B).

The local landmark detectors were trained using 1200 images from the ALFW database [25] and 100 profile images from PIE database [26]. The database was split into five classes

[2]http://www.codamotion.com/systems/indoor-3d-motion-capture.html
[3]http://jsaragih.org/



Fig. 2. Example frames from the ISIR-Eikeo Database.

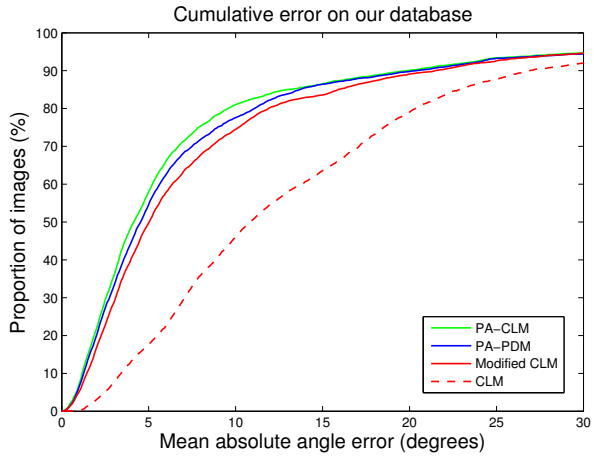


Fig. 3. Cumulative error on ISIR-Eikeo Database

of pose ranging from -80 to +80 degrees pan to train the pose-dependent local detectors. For each landmark, 20 positive and negative training patches were generated per image using random similarity transforms. Faces detected by Viola Jones detector are resized to 50×50 pixel images and the landmark localization is performed using 11×11 pixel patch experts on a 21×21 pixel search window per landmark.

C. Results and comparisons

Figure 3 provides a detailed analysis of the error distribution. Besides the clear improvement of the modified CLM over the original CLM, both pose-adaptive versions of this algorithm tend to improve the proportion of well-tracked images by 5 to 10 % for different MAE angle thresholds.

TABLE I. HEAD POSE ESTIMATION RESULTS ON THE BIWI DATASET. MEASURED IN MEAN ABSOLUTE ERROR.

Method	Mean	Pan	Tilt	Roll
Modified CLM	6.74	6.68	9.26	4.29
PA-PDM	6.02	5.59	8.81	3.65
PA-CLM	6.36	5.30	9.16	4.62

TABLE II. HEAD POSE ESTIMATION RESULTS ON THE ISIR-EIKEO DATASET. MEASURED IN MEAN ABSOLUTE ERROR.

Method	Mean	Pan	Tilt	Roll
Modified CLM	7.97	10.94	6.49	6.48
PA-PDM	6.97	7.21	7.22	6.47
PA-CLM	6.66	7.05	6.42	6.50

As all systems do not always diverge at the same time, it can be difficult to compare mean errors which have not been calculated on the same frames. For a fair comparison, we decided to compute the mean error on the frame that have been tracked by the three systems: the modified CLM, the PA-PDM and the PA-CLM. Table I and II shows the MAE per angle of the different approaches. Again, we see that both the pose adaptive shape model and the pose adaptive appearance model improve the results obtained by the modified CLM. This is especially true for the pan angle with an improvement of 1.3 and 3.9 degrees on the BIWI and ISIR-Eikeo datasets respectively.

V. CONCLUSION

In this paper, we have presented a Pose Adaptive Constrained Local Model which is well suited to track faces under large pose variations. We improve traditional CLM in two ways. First, we propose a new Pose Adaptive Point Distribution Model (PA-PDM) that insures the consistency



Fig. 4. Examples of head pose tracking on Biwi and ISIR-Eikeo datasets. Top row CLM, bottom row PA-CLM.

between the points of the 3D model and the corresponding landmarks in the image. Second, we propose a Pose-Adaptive Appearance Model, (PA-AM) that combines decision of multiple classifiers depending on the head pose estimate. Evaluations were performed on two datasets and we see that our approach improves the accuracy of the head pose tracker. Moreover, we present a new challenging head pose database with a precise ground truth which exhibits a wide variation of poses, occlusions and illumination changes

ACKNOWLEDGMENT

This work has been partially supported by the French National Agency (ANR) in the frame of its Technological Research CONTINT program (JEMImE, project number ANR-13-CORD-0004) and the Cap Digital Business cluster for digital content.

REFERENCES

[1] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, 2009.

[2] J. Huang, X. Shao, and H. Wechsler, "Face pose discrimination using support vector machines (svm)," in *International Conference on Pattern Recognition, ICPR 98.*, vol. 1. IEEE, 1998, pp. 154–156.

[3] H. A. Rowley, S. Baluja, and T. Kanade, "Rotation invariant neural network-based face detection," in *Computer Vision and Pattern Recognition, CVPR 1998.* IEEE, 1998, pp. 38–44.

[4] M. Osadchy, Y. L. Cun, and M. L. Miller, "Synergistic face detection and pose estimation with energy-based models," *J. Mach. Learn. Res.*, vol. 8, pp. 1197–1215, May 2007. [Online]. Available: http://dl.acm.org/citation.cfm?id=1248659.1248700

[5] K. Bailly and M. Milgram, "Head pan angle estimation by a nonlinear regression on selected features," in *International Conference on Image Processing, ICIP 2009.* IEEE, 2009, pp. 3589–3592.

[6] M. J. Black and Y. Yacoob, "Tracking and recognizing rigid facial motions using local parametric models of image motion," *ICCV*, pp. 374–381, 1995.

[7] S. Choi and D. Kim, "Robust head tracking using 3d ellipsoidal head model in particle filter," *Pattern Recognition*, pp. 241(9):2901–2915, 2008.

[8] P. Phothisane, E. Bigorgne, L. Collot, and L. Prevost, "A robust composite metric for head pose tracking using an accurate face model," in *Proceedings of the 9th IEEE conference on Automatic Face and Gesture Recognition*, ser. FG 2011, 2011, pp. 694–699.

[9] S. Ba and J.-M. Odobez, "A probabilistic framework for joint head tracking and pose estimation," in *Proceedings of the 17th IEEE Int'l Conf. on Pattern Recognition*, 2004, pp. 264–267.

[10] J. Xiao, T. Kanade, and J. Cohn, "Robust full-motion recovery of head by dynamic templates and re-regristeration techniques," *FG*, pp. 156–162, 2002.

[11] T. F. Cootes and C. J. Taylor, "Cj.taylor, "active shape models - "smart snakes," in *in Proceedings of the British Machine Vision Conference*, 1992.

[12] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *Computer Vision-ECCV*, 1998.

[13] S. Baker, I. Mathews, J. Xiao, R. Gross, T. Ishikawa, and T. Kanade, "Real-time non-rigid driver head tracking for driver mental state estimation," *11th World Congress Intelligent Transportation Systems*, 2004.

[14] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," *PAMI*, 2003.

[15] F. Dornaika and J. Ahlberg, "Fitting 3d face models for tracking and active appearance model training," *Image and Vision Computing 24*, pp. 1010–1024, 2006.

[16] D. Cristinacce and T. Cootes, "Feature detection and tracking with constrained local models," in *Proceedings of the British Machine Vision Conference*, 2006, pp. 929–938.

[17] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *In International Conference of Computer Vision (ICCV'09)*, 2009, pp. 1034–1041.

[18] G. Fanelli, M. Dantone, A. Fossati, J. Gall, and L. V. Gool, "Random forests for real time 3d face analysis," *International Journal of Computer Vision (IJCV)*, 2012.

[19] G. Fanelli, T. Weise, J. Gall, and L. V. Gool, "Real time head pose estimation from consumer depth cameras," in *33rd Annual Symposium of the German Association for Pattern Recognition, DAGM 2011)*, September 2011.

[20] Y. Wang, S. Lucey, and J. Cohn, "Enforcing convexity for improved alignment with constrained local models," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.

[21] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "3d constrained local model for rigid and non-rigid facial tracking," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012*, 2012.

[22] V. Rapp, K. Bailly, T. Senechal, and L. Prevost, "Multi-kernel appearance model," *Image and Vision Computing*, vol. 31, no. 8, pp. 542–554, 2013.

[23] J. Saragih, S. Lucey, and J. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.

[24] J. M. Saragih, S. Lucey, and J. F. Cohn, "Real-time avatar animation from a single image," in *Proceedings of the 9th IEEE conference on Automatic Face and Gesture Recognition*, ser. FG 2011, 2011, pp. 213–220.

[25] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[26] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression database," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615–1618, December 2003.