Projective Structure from Facial Motion

Stella Graßhof, Hanno Ackermann, Felix Kuhnke, Jörn Ostermann
Leibniz Universität Hannover
Germany

Sami S. Brandt
Univ. Copenhagen
Denmark

Abstract

Nonrigid Structure-From-Motion is a well-known approach to estimate time-varying 3D structures from 2D input image sequences. For challenging problems such as the reconstruction of human faces, state-of-the-art approaches estimate statistical shape spaces from training data. It is common practice to use orthographic or weak-perspective camera models to map 3D to 2D points. We propose to use a projective camera model combined with a multilinear tensor-based face model, enabling approximation of a dense 3D face surface by sparse 2D landmarks. Using a projective camera is beneficial, as it is able to handle perspective projections and particular camera motions which are critical for affine models. We show how the nonlinearity of the projective model can be linearized so that its parameters can be estimated by an alternating-least-squares approach. This enables simple and fast estimation of the model parameters. The effectiveness of the proposed algorithm is demonstrated using challenging real image data.

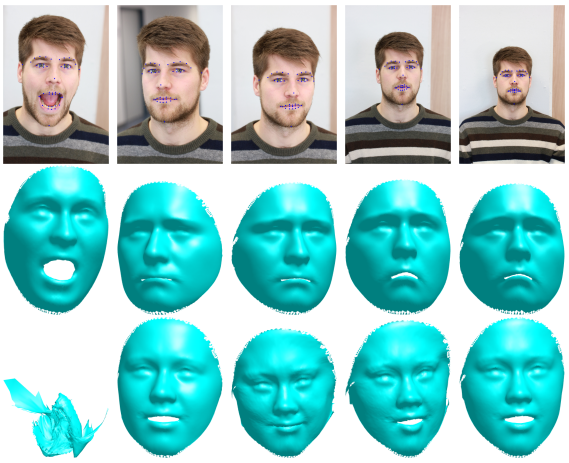


Figure 1. (first row) Input images with detected 2D landmarks, (second row) corresponding dense 3D reconstructed shapes by our model with projective camera, (third row) with weak-perspective camera. The first shape of the weak-perspective model is heavily distorted.

1 Introduction

Factorization for rigid 3D reconstruction was first introduced in [1] using a low rank constraint on the orthographic camera matrix. It was later extended to nonrigid shapes in [2]. Principal Component Analysis (PCA) was used on high-resolution 3D face scans to build a so-called *morphable model* [3]. This model captured variations in person and texture inside the dataset, enabling 3D reconstructions of persons in neutral expression. In [4], the authors presented an extended morphable model with variations for person and expression.

Tensor factorization models are an extension of matrix-based approaches. In [5], the authors applied a tensor-based version of the conventional SVD (Singular Value Decomposition) on 3D face meshes to represent variation in person, expression and viseme parameters.

Instead of using the actual 3D face scans, a multilinear tensor model was computed on the wavelet coefficients of patches of dense 3D shapes in [6]. In [7] the authors transferred expressions from one person to another in videos, by using the morphable model [3] with an additional expressive blend shape model extension.

The presented works have two major drawbacks: The target surface has to be known or estimated to penalize orthogonal deviations to it [6], [7]. In contrast to that we have proposed to use penalties directly on the model parameters, to take an advantage of the structure in model space [8]. Furthermore orthographic and weak-perspective camera models are typically used for facial 3D-reconstruction [5], [6]. We propose to use a projective camera model. In [7], a projective camera model is used but the intrinsic camera parameters are assumed to be known.

To summarize our contributions are:

- A projective camera motion model for nonrigid dense 3D face shapes.
- The non-linearity of projective camera model can be linearized and included into the model parameter estimation procedure.
- Simple and fast estimation procedure for parameters of a multilinear 3D face model by alternating least squares. The model parameters can be estimated by a linear equation system in each step.
- Estimation of sensible projective camera parameters.
- Dense nonrigid 3D face reconstruction from sparse 2D landmarks.

2 3D Multilinear Face Tensor Model

Given a set of 3D face scans, the data is ordered in a data tensor $\mathcal{T}_0 \in \mathbb{R}^{3N \times P \times E}$ where N is the number of corresponding 3D points, P the number of persons and E the number of expressions per person. First all shapes are globally aligned in 3D space, we then define $\mathcal{T}$ as the data tensor with subtracted mean shape. $\mathcal{T}$ can be approximated by the Higher-Order-SVD (HOSVD) as

$$\widehat{\mathcal{T}} = \mathcal{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)}, \tag{1}$$

where $\mathcal{S} \in \mathbb{R}^{L_1 \times L_2 \times L_3}$ is the cropped core tensor, and $\mathbf{U}^{(1)} \in \mathbb{R}^{3N \times L_1}$, $\mathbf{U}^{(2)} \in \mathbb{R}^{P \times L_2}$, $\mathbf{U}^{(3)} \in \mathbb{R}^{E \times L_3}$ are

orthogonal matrices containing the singular vectors of the nth mode of the nth mode unfolded data tensor, with $L_1 \leq 3N$, $L_2 \leq P$, $L_3 \leq E$.

Let $\mathbf{v} \in \mathbb{R}^{3N}$ be a mean-corrected face shape, then the approximation $\widehat{\mathbf{v}}$ using HOSVD by Eq. (1) is calculated as

$$\widehat{\mathbf{v}} = \mathcal{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{w}_2^{\mathrm{T}} \times_3 \mathbf{w}_3^{\mathrm{T}}, \tag{2}$$

where $\mathbf{w}_2 \in \mathbb{R}^{L_2}, \mathbf{w}_3 \in \mathbb{R}^{L_3}$.

To reconstruct shapes $\mathbf{v}$ of persons or expressions which are not part of the training data $\mathcal{T}$, the corresponding model parameters for person $\mathbf{w}_2$ and expression $\mathbf{w}_3$ could be estimated by minimizing $\|\widehat{\mathbf{v}} - \mathbf{v}\|_2^2$. However the model parameters defined in Eq. (2) do not contain any information of the parameter space structure. In other words $\mathbf{w}_2$, $\mathbf{w}_3$ can define arbitrary points, without taking any advantage of the substructures present in $\mathbf{U}^{(k)}$, $k = 2, 3$.

We therefore rewrite the model from Eq. (2), such that the new parameter vectors correspond to weights of the rows of $\mathbf{U}^{(k)}$, $k = 2, 3$, which each correspond to one specific person or expression:

$$\widehat{\mathbf{v}} = \mathcal{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{p}_2^{\mathrm{T}} \mathbf{U}^{(2)} \times_3 \mathbf{p}_3^{\mathrm{T}} \mathbf{U}^{(3)}, \tag{3}$$

with parameters $\mathbf{p}_2 \in \mathbb{R}^P$ and $\mathbf{p}_3 \in \mathbb{R}^E$. This representation is more stable compared to Eq. (2), because only linear combinations of the training shapes are used, thereby approximating the distribution of the training data.

Additionally, we use the standard Tikhonov regularizer to penalize large parameter values. Considering that we constructed a canonical parameter vector space, we restrict $\mathbf{p}_k$, $k = 2, 3$ to a sum of one. We thus formulate the total minimization problem as

$$\begin{aligned} \min_{\mathbf{p}_2, \mathbf{p}_3} \|\widehat{\mathbf{v}} - \mathbf{v}\|_2^2 + \lambda_1 \|\mathbf{p}_2\|_2^2 + \lambda_2 \|\mathbf{p}_2^{\mathrm{T}} \mathbf{1} - 1\|_2^2 \\ + \lambda_3 \|\mathbf{p}_3\|_2^2 + \lambda_4 \|\mathbf{p}_3^{\mathrm{T}} \mathbf{1} - 1\|_2^2. \end{aligned} \tag{4}$$

We use leave-one-out cross validation is used to determine the weights λ_k.

An alternating least squares scheme can be used to minimize Eq. (4), as the introduced function is separately linear in $\mathbf{p}_2$ or $\mathbf{p}_3$. In more detail, let $\mathbf{m} \in \mathbb{R}^{3N \times 1}$ be the mean shape, which was subtracted from the input data tensor $\mathcal{T}_0$. Reordering the elements of the tensor for all persons $\mathcal{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{p}_3^{\mathrm{T}} \mathbf{U}^{(3)} \in \mathbb{R}^{3N \times P \times 1}$ in the matrix $\mathbf{M}_2 \in \mathbb{R}^{3N \times P}$, we rewrite a model generated 3D shape $\mathbf{s}^{\mathrm{3D}} = \widehat{\mathbf{v}} + \mathbf{m}$ as $\mathbf{s}^{\mathrm{3D}} - \mathbf{m} = \widehat{\mathbf{v}} = \mathbf{M}_2 \mathbf{p}_2$ that shows the linear relationship to $\mathbf{p}_2$. The linear relationship to the the expression parameter vector $\mathbf{p}_3$ can be shown similarly.

Until now 3D representations of faces were considered, while the desired input are 2D landmarks of faces to reconstruct a 3D face. In the following Section, we show two common camera models to project 3D points to the 2D image plane.

3 Camera Models

3.1 Weak-Perspective Camera

The weak-perspective camera model [9], [10] is the standard model for 3D reconstructions and other applications. It maps a 3D point $\mathbf{x} \in \mathbb{R}^3$ to 2D image coordinates $\mathbf{u} \in \mathbb{R}^2$ as $\mathbf{u} = \begin{pmatrix} u_x & u_y \end{pmatrix}^\top = c\, \mathbf{K}_a (\mathbf{R}\mathbf{x} + \mathbf{t})$, where $\mathbf{K}_a := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \in \mathbb{R}^{2 \times 3}$, where $c \in \mathbb{R}^+$ is a scaling factor, $\mathbf{R}$ is a 3D rotation matrix and $\mathbf{t}$ is a 3D translation vector.

3.2 Projective Camera Model

In contrast to the related work using the weak-perspective camera model [11], [12], we take an advantage of the projective camera model, which maps a 3D point $\mathbf{x} \in \mathbb{R}^3$ to 2D image coordinates $\mathbf{u} \in \mathbb{R}^2$ as

$$\begin{aligned} \widetilde{\mathbf{u}} &= \begin{pmatrix} \widetilde{u}_x \\ \widetilde{u}_y \\ \widetilde{u}_z \end{pmatrix} = \mathbf{K} (\mathbf{R}\mathbf{x} + \mathbf{t}), \\ \mathbf{u} &= \begin{pmatrix} u_x \\ u_y \end{pmatrix} = \begin{pmatrix} \widetilde{u}_x / \widetilde{u}_z \\ \widetilde{u}_y / \widetilde{u}_z \end{pmatrix}, \end{aligned} \tag{5}$$

whereas $\mathbf{K} := \mathrm{diag}(f, f, 1)$, $f \in \mathbb{R}^+$ is the focal length, $\mathbf{R}$ is a 3D rotation matrix and $\mathbf{t}$ is a 3D translation vector. We assume that pixels on the image sensor are square. This property is satisfied for many recent consumer cameras. Due to the non-linearity, 3D reconstructions using this model are difficult to estimate.

4 3D Face Tensor Model with Motion

In the following, the multilinear face model of Eq. (3) is combined with a nonlinear projective camera model. Given fixed camera parameters, the 2D mapping of the 3D model coordinates is rewritten, so that it is linear in the model parameter vector for person or expression.

Let $[\cdot]_x$ be the x-component of the vector argument, with analogous notation for the y- and z-component. Given a 3D face shape $\mathbf{s}^{\mathrm{3D}} = \widehat{\mathbf{v}} + \mathbf{m}$ consisting of 3D points $\mathbf{s}_i^{\mathrm{3D}} = \widehat{\mathbf{v}}_i + \mathbf{m}_i$ and camera parameters $\mathbf{K} := \mathrm{diag}(f, f, 1)$, $f \in \mathbb{R}^+$, $\mathbf{R}$ and $\mathbf{t}$, the corresponding projected 2D points $\mathbf{s}_i^{\mathrm{2D}}$ are calculated with the equation of motion Eq. (5), then the components of $\mathbf{s}_i^{\mathrm{2D}}$ are reformulated separately, as

$$\begin{aligned} \widetilde{u}_{i,z} s_{i,x}^{\mathrm{2D}} = \widetilde{u}_{i,x} \Leftrightarrow \\ \left[\mathbf{K}\left(\mathbf{R}\mathbf{s}_i^{\mathrm{3D}} + \mathbf{t}\right)\right]_z s_{i,x}^{\mathrm{2D}} = \left[\mathbf{K}\left(\mathbf{R}\mathbf{s}_i^{\mathrm{3D}} + \mathbf{t}\right)\right]_x \Leftrightarrow \\ \left[\mathbf{K}\left(\mathbf{R}(\widehat{\mathbf{v}}_i + \mathbf{m}_i) + \mathbf{t}\right)\right]_z s_{i,x}^{\mathrm{2D}} = \left[\mathbf{K}\left(\mathbf{R}(\widehat{\mathbf{v}}_i + \mathbf{m}_i) + \mathbf{t}\right)\right]_x. \end{aligned} \tag{6}$$

Taking advantage of the linear representation in section 2 for $\widehat{\mathbf{v}}_i$ yields to

$$\begin{aligned} \left[\mathbf{K}\mathbf{R}\mathbf{M}_{2,i}\mathbf{p}_2 + \mathbf{K}\mathbf{R}\mathbf{m}_i + \mathbf{K}\mathbf{t}\right]_z s_{i,x}^{\mathrm{2D}} = \\ \left[\mathbf{K}\mathbf{R}\mathbf{M}_{2,i}\mathbf{p}_2 + \mathbf{K}\mathbf{R}\mathbf{m}_i + \mathbf{K}\mathbf{t}\right]_x. \end{aligned} \tag{7}$$

For each 3D point we then obtain the two equations

$$\begin{aligned} \begin{pmatrix} \left[\mathbf{K}\mathbf{R}\mathbf{M}_{2,i}\right]_x - s_{i,x}^{\mathrm{2D}} \left[\mathbf{K}\mathbf{R}\mathbf{M}_{2,i}\right]_z \\ \left[\mathbf{K}\mathbf{R}\mathbf{M}_{2,i}\right]_y - s_{i,y}^{\mathrm{2D}} \left[\mathbf{K}\mathbf{R}\mathbf{M}_{2,i}\right]_z \end{pmatrix} \mathbf{p}_2 = \\ \begin{pmatrix} \left[\mathbf{K}\mathbf{R}\mathbf{m} + \mathbf{K}\mathbf{t}\right]_z s_{i,x}^{\mathrm{2D}} - \left[\mathbf{K}\mathbf{R}\mathbf{m} + \mathbf{K}\mathbf{t}\right]_x \\ \left[\mathbf{K}\mathbf{R}\mathbf{m} + \mathbf{K}\mathbf{t}\right]_z s_{i,y}^{\mathrm{2D}} - \left[\mathbf{K}\mathbf{R}\mathbf{m} + \mathbf{K}\mathbf{t}\right]_y \end{pmatrix}. \end{aligned} \tag{8}$$

Stacking all N points and in both dimensions, one person parameter can be estimated for multiple shapes. An equation system for $\mathbf{p}_3$ can be derived similarly. Finally given person and expression parameters, a 3D shape can be calculated by Eq. (3), while the corresponding 2D projected shape is defined by Eq. (5).

4.1 Parameter Estimation

Given N images with corresponding 2D landmarks of the same person, the camera and model parameters are estimated in an alternating scheme, by minimizing the euclidean distance between input landmarks $\mathbf{s}_k^{2\mathrm{D}}$ and the estimated and projected 3D model shapes $\widehat{\mathbf{s}}_k^{2\mathrm{D}}$. The estimation procedure is described in Algorithm 1.

Algorithm 1 3D reconstruction from 2D landmarks

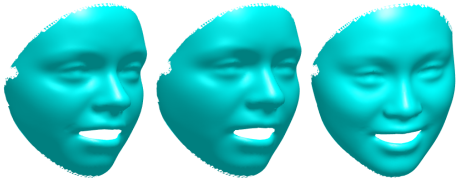
Input: 2D face landmarks $\mathbf{s}_k^{2\mathrm{D}}$, $k = 1, \ldots, N$

- **Initialization:**
  - Initialize N camera parameter vectors
  - Initialize $\widehat{\mathbf{p}}_{3,k}$ with mean expression $\forall k$
- **Repeat until convergence:**
  - **Model Parameter Estimation**
    Repeat until convergence:
    - Given $\widehat{\mathbf{p}}_{3,k}$ and camera parameters, estimate person parameter vector $\widehat{\mathbf{p}}_2$, using Eq. (8)
    - Given $\widehat{\mathbf{p}}_2$ and camera parameters, estimate expression parameter vectors $\widehat{\mathbf{p}}_{3,k}$
  - **Camera Parameter Estimation**
    - Given $\widehat{\mathbf{p}}_2$, $\widehat{\mathbf{p}}_{3,k}$, estimate the camera parameters by second order gradient descent algorithm

Output: $\widehat{\mathbf{p}}_2$, $\widehat{\mathbf{p}}_{3,k}$, camera parameters, $\widehat{\mathbf{s}}_k^{2\mathrm{D}}$, $\widehat{\mathbf{s}}_k^{3\mathrm{D}}$

(a) (b) (c)

Figure 2. Color-coded point-wise approximation error for the two camera models. (a) Ground truth shape, (b) reconstruction by projective camera model, and (c) weak-perspective camera model. Using the projective model facilitates a better shape reconstruction as indicated by the mouth which is wider open in (b) than in (c).



(a) (b) (c)

Figure 3. Comparison of the two camera models. (a) Ground truth shape; reconstructed 3D shape with rotation by (b) projective model, (c) weak-perspective camera model. It can be seen that for the weak-perspective model, the rotation cannot be as accurately estimated as for the projective model.

5 Experiments

We used the BU-3DFE Database [13] consisting of 100 persons with 6 emotions (Anger, Disgust, Fear, Happiness, Sadness and Surprise) in 4 different expression levels and a neutral expression. For each 3D face scan, 83 manually labelled landmark points were provided, which we extended to 85.Based on these we computed the HOSVD of Eq. (1). We then computed the full correspondence among the 3D face scans by the Extended Coherent Point Drift Algorithm (ECPD) [14] and estimated the second model with the resulting 7308 points. As the landmark points were a subset of the dense face scans, sparse correspondences among the models were provided. The number of camera and model parameters was not altered between the models.

5.1 Evaluation on Synthetic Shapes

We evaluated the performance of our model with the weak-perspective camera and projective camera model by synthetic data with known ground truth. We assumed that the person parameter and focal lengths were the same for all input frames, while expression parameters and all other camera parameters were estimated for each frame individually. For a weak-perspective camera, it was not possible to estimate both the focal length and the z-components of the translation vectors [15], so we let the z-components be undetermined.

The results of the evaluation are shown in Figure 2–4. Figure 2 compares, the ground truth shape to their reconstructed counterparts by illustrating the point-wise approximation error. It can be seen that withthe projective camera model (Fig. 2b) the open mouth expression can be better recovered than using the weak-perspective camera (Fig. 2c). In Figure 3, the ground truth shape is compared to the reconstructed shapes by weak-perspective and projective camera model. It can be seen that the projective camera model is able to estimate the 3D shape and orientation better than the weak-perspective camera model. Figure 4 compares the ground truth camera positions with the estimated ones. It can be concluded that the projective estimates partly overlap with the ground truth camera positions while the camera positions estimated by the weak-perspective model have high distances to the real camera positions.

In summary, the projective camera model improved the 3D reconstruction of the shapes. Even for sequences with difficult motion along the z-axis, shapes of high quality could be estimated.

5.2 Reconstruction from Real Images

A DSLR Canon EOS 5D Mark III camera with a Tamron 90mm F/2.8 Macro 1:1 lens was used to capture several images of a human face. During the recordings the subject changed his expression, while the camera position was changed relative to the object. For each frame 68 face landmarks were detected by the dlib C++ implementation of [16]. Of the landmarks, 46 were selected corresponding to the landmark model vertices and used to estimate the model and camera parameters as described in Section 4.1. The results are displayed in Figure 1. It can be seen that the projective

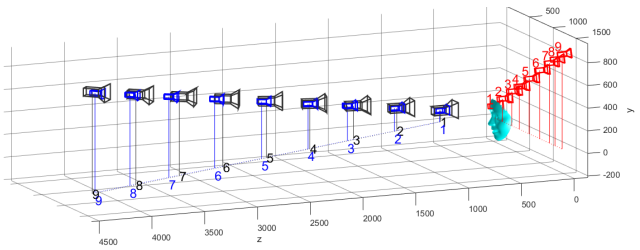


Figure 4. Comparison of ground truth camera positions (black) with the estimated ones by our model with weak-perspective (red) and projective (blue) projection of our synthetic data. The figure is best viewed in color.

camera model leads higher quality 3D reconstructions, which are more similar to the person and expression in the images, than the weak-perspective version.

6 Conclusions

In this paper we have studied the application of the projective camera model together with tensor-based face model. We used our 3D multilinear face model, based on the HOSVD, which utilizes the inherent data structure by a canonical parametrization and constraints. The incorporated constraints do not require any explicit knowledge about the target 3D surfaces and enable an ALS scheme to estimate reasonable model parameters, and thereby a 3D face reconstruction of sparse 2D face landmarks. Furthermore, we showed that the estimation of the perspective camera parameters is possible, while retaining the linear estimation scheme of the multilinear model. In the experiments, the projective camera model improved the estimated 3D reconstruction, camera and model parameters in comparison to the weak-perspective camera model used in previous work. This is a promising result and opens the way for more accurate estimation of non-rigid facial structures.

Acknowledgements

This work was partly supported by German Research Foundation grant DFG AC 264-2/1.

References

[1] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization method", *Int. Journ. Comp. Vis. (IJCV)*, vol. 9, no. 2, pp. 137–154, 1992.

[2] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3d shape from image streams", in *Comp. Vis. and Pattern Rec. (CVPR)*, 2000, pp. 2690–2696.

[3] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces", in *Proc. 26th Conf. Computer Graphics and Interactive Techniques (SIGGRAPH)*, 1999, pp. 187–194.

[4] B. Amberg, R. Knothe, and T. Vetter, "Expression invariant 3d face recognition with a morphable model", in *8th IEEE Int. Conf. Autom. Face and Gesture Recog. (FG)*, 2008, pp. 1–6.

[5] D. Vlasic, M. Brand, H. Pfister, and J. Popović, "Face transfer with multilinear models", *ACM Trans. Graph. (SIGGRAPH)*, vol. 24, no. 3, pp. 426–433, 2005.

[6] A. Brunton, T. Bolkart, and S. Wuhrer, "Multilinear Wavelets: A Statistical Shape Space for Human Faces", in *European Conf. Comp. Vis. (ECCV)*, 2014, pp. 297–312.

[7] J. Thies, M. Zollhöfer, M. Niessner, L. Valgaerts, M. Stamminger, and C. Theobalt, "Real-time expression transfer for facial reenactment", *ACM Trans. Graph. (SIGGRAPH)*, vol. 34, no. 6, Oct. 2015.

[8] S. Grasshof, H. Ackermann, S. S. Brandt, and J. Östermann, "Apathy is the root of all expressions", in *Proc. 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, To appear., 2017.

[9] K. Kanatani, Y. Sugaya, and H. Ackermann, "Uncalibrated Factorization Using a Variable Symmetric Affine Camera", in *European Conf. Comp. Vis. (ECCV)*, 2006, pp. 147–158.

[10] S. S. Brandt, "Conditional solutions for the affine reconstruction of n-views", *Image and Vision Computing*, vol. 23, no. 7, pp. 619–630, 2005.

[11] H. Ackermann and K. Kanatani, "Iterative low complexity factorization for projective reconstruction", in *Proceedings of the 2nd Intern. Conf. on Robot Vision*, ser. RobVis'08, Springer-Verlag, 2008, pp. 153–164.

[12] H. Ackermann and B. Rosenhahn, "Projective Reconstruction from Incomplete Trajectories by Global and Local Constraints", in *CONF. for Visual Media Production*, 2011, pp. 77–86.

[13] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, "A 3d facial expression database for facial behavior research", in *7th Intern. Conf. on Automatic Face and Gesture Recog., (FG)*, 2006, pp. 211–216.

[14] V. Golyanik, B. Taetz, G. Reis, and D. Stricker, "Extended coherent point drift algorithm with correspondence priors and optimal subsampling", in *IEEE Winter Conf. on Appl. of Comp. Vis. (WACV)*, 2016, pp. 1–9.

[15] K. Kanatani and Y. Sugaya, "Factorization without factorization: Complete recipe", *Mem. Fac. Eng. Okayama Univ*, vol. 38, no. 1&2, pp. 61–72, 2004.

[16] J. S. Vahid Kazemi, "One millisecond face alignment with an ensemble of regression trees", in *Comp. Vision and Pattern Recog. (CVPR)*, 2014.