Real-time 3D Face Super-resolution From Monocular In-the-wild Videos

Patrik Huber[1], William Christmas[1], Matthias Rätsch[2], Adrian Hilton[1], Josef Kittler[1]

[1]Centre for Vision, Speech and Signal Processing, University of Surrey, UK
[2]Image Understanding and Interactive Robotics, Reutlingen University, Germany

Objective

Given a 2D video stream with a person's face, reconstruct a 3D face and high-quality texture of that person.

Challenges

Reconstructing a 3D face from 2D information is a highly ill-posed problem. In our scenario, the video conditions are highly unconstrained («in-the-wild»): Webcam live-streams, low-resolution, blur, pose.

Motivation

Reconstruct a person's face in 3D in real-time for applications on mobile phones, graphics, games, face analysis, HCI, and many more applications.

Contributions

- Real-time 3D face and high-quality texture reconstruction
- Combining 3D Morphable Face Model fitting with texture super-resolution on video
- Suitable for in-the-wild scenarios with low resolution, blur, and varying pose
- Incremental approach working on live-streams
- No prior training or subject-adaption required

3D Shape Fitting

A 3D Morphable Face Model with expression blendshapes is fitted to each video frame to obtain dense correspondence between the video frames.

Camera model: Affine camera model with closed-form solution from Hartley & Zisserman («Gold Standard Algorithm», H&Z 2004). A 3 x 4 camera matrix is estimated using 3D shape points to 2D landmark correspondences.

Shape identity fitting: 3D shape is fitted to landmarks by finding the most likely shape coefficients $\boldsymbol{\alpha}$ of the PCA model by minimising

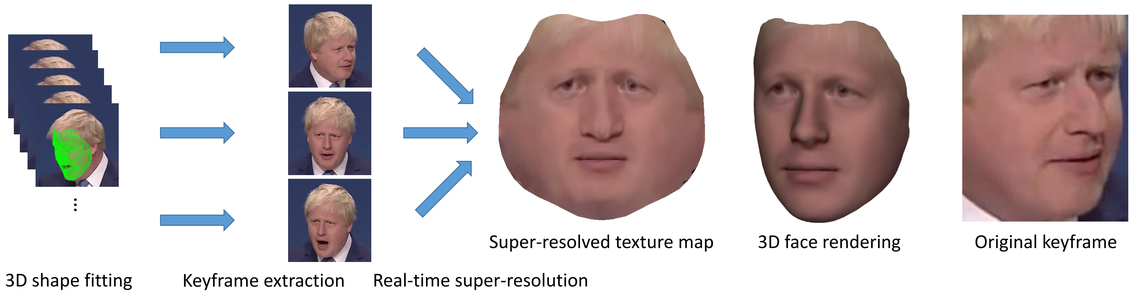
$$\mathrm{E} = \sum_{i=1}^{3N} \frac{\left(y_{m2D,i} - y_i\right)^2}{2\vartheta_{2D,i}^2} + \|\boldsymbol{\alpha}\|_2^2, \qquad (1)$$

with N the number of 2D landmarks, y_i the 2D landmark points, $y_{m2D,i}$ the projection to 2D of the corresponding 3D model points, and $\vartheta_{2D,i}^2$ variances for the landmark points. This cost function is brought into a standard linear least squares formulation.

Expression fitting: 6 expression blendshapes are fitted using the same formulation, on the basis of the current PCA shape estimate. The linear system is solved with a NNLS solver, and the process alternated with the PCA shape fitting.

Contour refinement: 2D-3D correspondences at the face contour are sought and updated to improve the shape fit at the face contour.

Algorithm Overview



Keyframe Selection

- The face region in each frame is rated with regards to its image quality and suitability to use for texture reconstruction. We rate each face patch using the variance of Laplacian measure (Pech et al., 2000), which is a measure for focus or sharpness of an image.
- Frames with fewer expressions are preferred, penalising frames with a large norm of the blendshape coefficients, $\|\boldsymbol{\psi}\|$.
- Each frame is then associated with a score

$$s = L * \frac{\tau}{\|\boldsymbol{\psi}\|} \qquad (2)$$

where L is the variance of Laplacian measure and τ a parameter to influence how much expressions are penalised.

- The $\pm 90°$ yaw pose space is divided into bins of 20° intervals. Each bin contains a maximum of p keyframes.

Top-scoring keyframe for each pose bin

- A frame is used as keyframe if its particular pose bin is not full, or, if any of the existing keyframes of that pose bin have a lower score than the current frame. In the latter case, the currently present keyframe is replaced with the newer one with the higher score.

Median-based Super-resolution

Texture maps of keyframes are fused with a median-based super-resolution inspired by Maier et al. (2015):

- The texture of each keyframe is remapped to a common reference texture map of higher resolution using the dense correspondence obtained from the model fitting.
- For each pixel of the final texture map, a weighted median over all keyframes is computed. The final colour value $\hat{c}$ of a pixel is given by:

$$\hat{c} = \arg\min_{c} \sum_{(c_i,\omega_i)\in O} \omega_i \|c - c_i\| \qquad (3)$$

where O is the set of all colour values and weights of all keyframes for a particular pixel.

- The weight ω for a pixel consists of the previously calculated frame score and a per-vertex weighting based on the view-angle of the vertex w.r.t. the camera's viewing direction:

$$\omega = \frac{\langle \boldsymbol{d}, \boldsymbol{n} \rangle * L * \tau}{\|\boldsymbol{\psi}\|} \qquad (4)$$

where $\boldsymbol{d}$ is the camera viewing direction and $\boldsymbol{n}$ the normal of the vertex.

- This texture map is recomputed whenever a new keyframe is added.

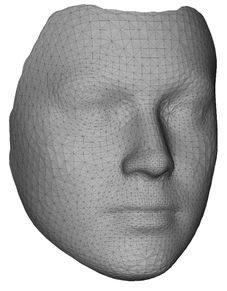
The Surrey 3D Morphable Face Model

We use a 3D Morphable Face Model to reconstruct a face in 3D. A 3D Morphable Model consists of a PCA model of 3D shape and albedo (colour), of which we use the shape model in this work.
The model is built from high-resolution example 3D scans that are brought in dense correspondence. PCA is then applied to these scans to obtain a mean shape $\bar{\mathbf{v}}$, a matrix of eigenvectors $\mathbf{V}$ and their corresponding variances $\boldsymbol{\sigma}^2$. On top of the PCA shape model, we use a set of linear expression blendshapes, modelling the 6 universal emotions by Ekman: Anger, disgust, fear, happiness, sadness and surprise.

Novel faces can be approximated as a linear combinations of the shape basis vectors $\mathbf{V}$ and the expression blendshapes $\mathbf{B}$:

$$\mathbf{v} = \bar{\mathbf{v}} + \sum_{i=1}^{l} \alpha_i \mathbf{V}_{:,i} + \sum_{j=1}^{m} \psi_j \mathbf{B}_{:,j} \qquad (5)$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_l]^{\mathrm{T}}$ is a vector of shape coefficients, l the number of shape basis vectors, $\boldsymbol{\psi} = [\psi_1, \ldots, \psi_m]^{\mathrm{T}}$ a vector of blendshape coefficients, and m the number of blendshapes.



Low-resolution 3D Morphable Face Shape Model

Model-fitting library and low-resolution shape-only model available online:
github.com/patrikhuber/eos

References

- Aldrian, O. and Smith, W. A. P., *Inverse rendering of faces with a 3D Morphable Model*, PAMI 2013
- Blanz, V., and Vetter, T., *A Morphable Model for the synthesis of 3D faces*, SIGGRAPH 1999
- Cao, C., Bradley, D., Zhou, K., and Beeler, T., *Real-time high-fidelity facial performance capture*, SIGGRAPH 2015
- Huber, P., Hu, G., Tena, R., Mortazavian, P., Koppen, W. P., Christmas, W., Rätsch, M., and Kittler, J., *A multiresolution 3D Morphable Face Model and fitting framework*, VISAPP 2016
- Ichim, A. E., Bouaziz, S., and Pauly, M., *Dynamic 3D avatar creation from hand-held video input*, SIGGRAPH 2015
- Maier, R., Stückler, J., and Cremers, D., *Superresolution keyframe fusion for 3D modeling with high-quality textures*, 3DV 2015
- Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G., and Pantic, M., *The first facial landmark tracking in-the-wild challenge: Benchmark and results*, ICCVW 2015

Future Work

- Multi-frame shape fitting
- Temporally coherent pose and expression fitting
- Shape refinement using more points and/or image information

Acknowledgements Support from the EPSRC Programme Grant FACER2VM (EP/N007743/1) is gratefully acknowledged.