Reconstruction of Personalized 3D Face Rigs from Monocular Video - Supplemental Document

PABLO GARRIDO and MICHAEL ZOLLHÖFER and DAN CASAS and LEVI VALGAERTS
Max-Planck-Institute for Informatics
and
KIRAN VARANASI and PATRICK PÉREZ
Technicolor
and
CHRISTIAN THEOBALT
Max-Planck-Institute for Informatics

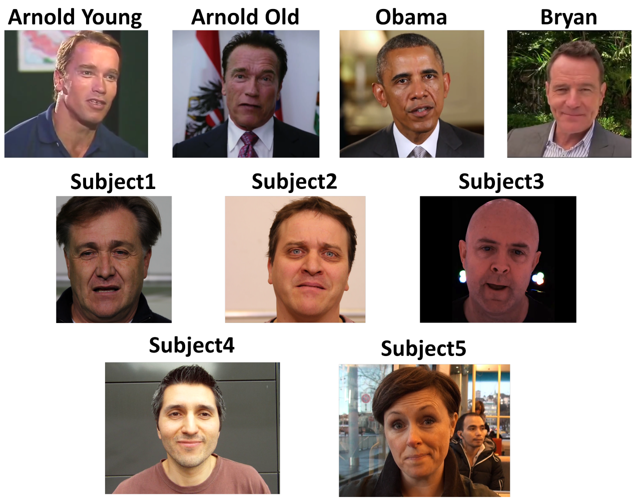

Fig. 1. Test Sequences: From left to right. ARNOLD YOUNG, ARNOLD OLD, OBAMA, BRYAN, SUBJECT1, SUBJECT2, SUBJECT3, SUBJECT4 and SUBJECT5.

1. USED TEST SEQUENCES

We evaluated our approach on 9 different sequences, shown in Figure 1. They consist of five videos (SUBJECT1[1], SUBJECT2[2], SUBJECT3[3], SUBJECT4[4], SUBJECT5[5]) captured indoors and outdoors under unknown and general lighting, and four legacy videos (ARNOLD YOUNG[6], ARNOLD OLD[7], OBAMA[8], BRYAN[9]) freely available on the Internet and downloaded from YouTube.

ARNOLD YOUNG. An interview discussing the "Predator" movie launch. We used a subset consisting of 1489 frames. The original video has a resolution of 480 × 360 pixels. We processed the video at its original full resolution.

ARNOLD OLD. Arnold Schwarzenegger's message for DECC's Energy Efficiency Mission Launch. We used a subset consisting of 1000 frames. The original video has a resolution of 1280 × 720 pixels. We processed the video at its original full resolution.

OBAMA. In this greeting address, president Obama commemorates Independence Day on the 4th of July. We used a subset consisting of 961 frames. The original video has a resolution of 1280 × 720 pixels. We processed the video at its original full resolution.

BRYAN. This video shows the actor Bryan Lee Cranston talking about the end of his journey with the TV series "Breaking Bad". We used a subset consisting of 702 frames. The original video has a resolution of 640 × 360 pixels. We processed the video at its original full resolution.

SUBJECT1. This is a studio sequence captured indoors and employed in the paper [Valgaerts et al. 2012]. A stereo reconstruction of this sequence is available. The sequence consists of 714 frames and has a resolution of 1088 × 1920 pixels. We downsampled the images to half the resolution for tracking and use the full resolution in all other steps.

SUBJECT2. This is a studio sequence captured indoors and used in the paper [Garrido et al. 2013]. There is an audio channel available. This sequence consists of 2000 frames and has a resolution of 1088 × 1920 pixels. We downsampled the images to half the resolution for tracking and use the full resolution in all other steps.

SUBJECT3. This is a studio sequence captured indoors and employed in the paper [Beeler et al. 2011]. The actual capture setup consists of 6 high-quality cameras, one recording the actor from a frontal view. This sequence consists of 347 frames and has a resolution of 864 × 1174 pixels. We downsampled the images to half the resolution for tracking and use the full resolution in all other steps.

SUBJECT4. This is an outdoor sequence employed in the paper [Garrido et al. 2013] (and also in [Shi et al. 2014]). In their capture setup, a GoPro Hero 3 camera was used to record the actor from a frontal view. This sequence consists of 651 frames and has a resolution of 1920 × 1080 pixels. We downsampled the images to

[1] http://gvv.mpi-inf.mpg.de/projects/FaceCap/

[2] http://gvv.mpi-inf.mpg.de/projects/MonFaceCap/

[3] http://graphics.ethz.ch/publications/papers/paperBee11.php

[4] http://gvv.mpi-inf.mpg.de/projects/MonFaceCap/

[5] http://www.disneyresearch.com/project/facial-performance-enhancement/

[6] https://youtu.be/BkX2CMCXhM8

[7] https://youtu.be/EgvdhvKreJI

[8] https://youtu.be/d-VaUaTF3_k

[9] http://students.cse.tamu.edu/fuhaoshi/FacefromVideo/index.htm

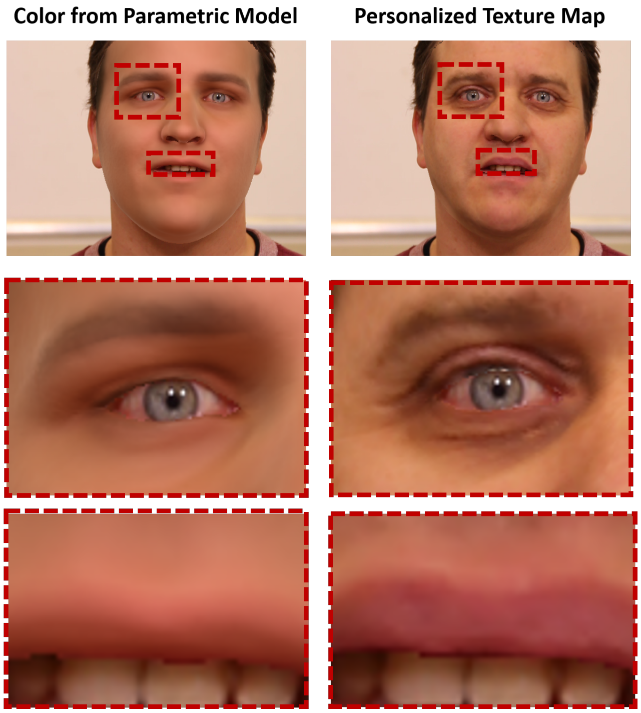

Fig. 2. Parametric model vs. personalized texture map: In contrast to the low dimensional parametric face model, the automatically computed personalized texture map captures fine-scale albedo variations.

half the resolution for tracking and use the full resolution in all other steps.

SUBJECT5. This is a cluttered scene captured outdoors and employed in the paper [Bermano et al. 2014]. This sequence consists of 806 frames and has a resolution of 1920 × 1080 pixels. We downsampled the images to half the resolution for tracking and use the full resolution in all other steps.

2. PARAMETRIC MODEL VS. PERSONALIZED TEXTURE

The automatically computed personalized texture map captures more fine-scale albedo variations than the low dimensional parametric model, see Fig. 2. Note the detail around the eyes and the nice mouth shape. Here $K_r = 160$ principal components have been used to represent the surface albedo in the parametric face model.

3. VALIDATION

To quantify the influence of the regularization in the (sparse) ridge regression of the medium- and fine-scale layer, we compared several regressors learned with different ridge regression parameters λ by measuring the geometric prediction error. To this end, we employed two test sequences (SUBJECT1 and SUBJECT2) and learned a regressor for different values of λ. As training data, we used the first half of tracked sequences. To test the accuracy, we predicted the deformation of the medium-scale layer $\hat{\boldsymbol{\tau}}$ and fine-scale layer $\tilde{\mathbf{p}}$ using the estimated blendshape weights on the second half of the tracked sequences. The prediction error has been computed as the Euclidean distance of every predicted 3D vertex position to its corresponding tracked 3D position. The average prediction error of the medium-scale and fine-scale detail layer over the two test sequences can be found in Tables I and II.

Table I. Average prediction error (medium-scale) on two sequences.

	Prediction error (in mm)			
Sequence	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$
SUBJECT1	0.98 ± 0.18	0.96 ± 0.17	0.95 ± 0.17	0.96 ± 0.17
SUBJECT2	0.87 ± 0.17	0.87 ± 0.17	0.87 ± 0.16	0.88 ± 0.16
Overall	0.93 ± 0.18	0.92 ± 0.17	0.91 ± 0.17	0.92 ± 0.17

Table II. Average prediction error (fine-scale) on two sequences.

	Prediction error (in mm)			
Sequence	$\lambda = 0.1$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 1.0$
SUBJECT1	0.30 ± 0.03	0.30 ± 0.03	0.29 ± 0.03	0.29 ± 0.03
SUBJECT2	0.53 ± 0.07	0.53 ± 0.06	0.54 ± 0.06	0.54 ± 0.05
Overall	0.42 ± 0.05	0.42 ± 0.05	0.42 ± 0.05	0.42 ± 0.04

As it can be observed, the lowest prediction error of the medium-scale layer is obtained by using $\lambda = 1.0$. On the other hand, the prediction error of the fine-scale layer stays mostly constant when increasing λ, but increasing the regularizer tends to over-smooth the results. This means that low values of λ result in more detailed, but slightly more noisy predictions due to extrapolation. Empirical experiments showed that the noise is visually negligible and $\lambda = 0.1$ achieves good results.

4. ADDITIONAL COMPARISONS

4.1 Comparison to Performance Capture Approaches

In this section we compare the reconstruction quality of our monocular approach to existing multi-view and monocular facial performance capture systems.

Comparison to [Beeler et al. 2011]. Fig. 3 shows a comparison to the high-quality off-line performance capture method of Beeler *et al.* [2011]. This method requires a controlled setup with 6 high-quality cameras and controlled in-studio lighting to perform a variant of multi-view stereo in combination with a mesoscopic detail augmentation step. Furthermore, the approach does not construct a face rig from the tracked data. In contrast, our approach is based on a single monocular video under general lighting as input and is capable of achieving a reconstruction quality that comes close to their approach. Besides, our approach reconstructs a fully-modifiable face rig (see additional supplementary video).

Comparison to [Garrido et al. 2013] and [Shi et al. 2014]. Our approach attains reconstructions of higher-quality than those of Garrido *et al.* [2013] and Shi *et al.* [2014], both on the coarse geometry and on the fine-scale level, see Fig. 4. Note that our method can also handle strong out-of-plane head rotations, as in [Shi et al. 2014], while preserving the face details. We remark that none of these state-of-the-art approaches can reconstruct a highly-detailed 3D face rig as we do.

REFERENCES

BEELER, T., HAHN, F., BRADLEY, D., BICKEL, B., BEARDSLEY, P., GOTSMAN, C., SUMNER, R. W., AND GROSS, M. 2011. High-quality passive facial performance capture using anchor frames. *ACM TOG 30,* 4, 75:1–75:10.

BERMANO, A. H., BRADLEY, D., BEELER, T., ZUND, F., NOWROUZEZAHRAI, D., BARAN, I., SORKINE-HORNUNG, O., PFISTER, H., SUMNER, R. W., BICKEL, B., AND GROSS, M. 2014.

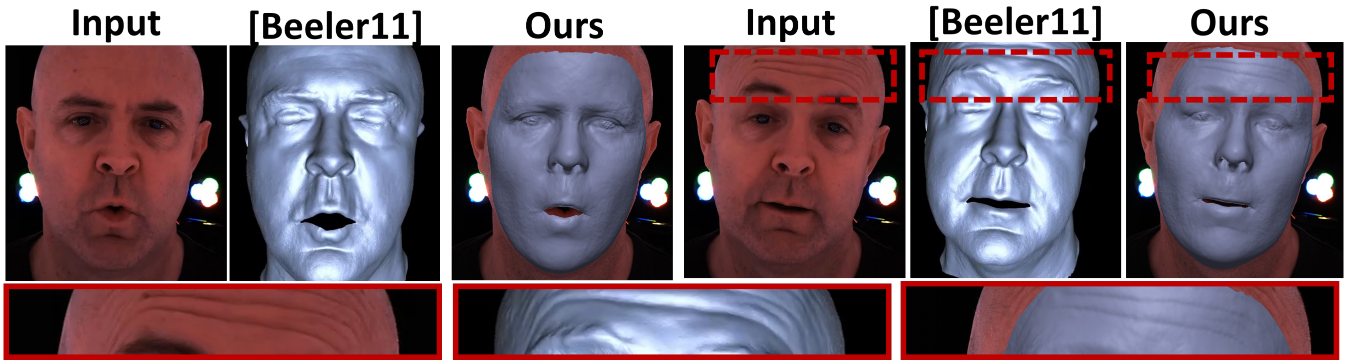

Fig. 3. State-of-the-art comparison to the multi-view in-studio approach by [Beeler et al. 2011]: Our monocular approach, which reconstructs detailed geometry from a single video under general lighting, comes close in reconstruction quality to that of Beeler *et al.*'s method which requires a professional setup with 6 high-quality cameras.

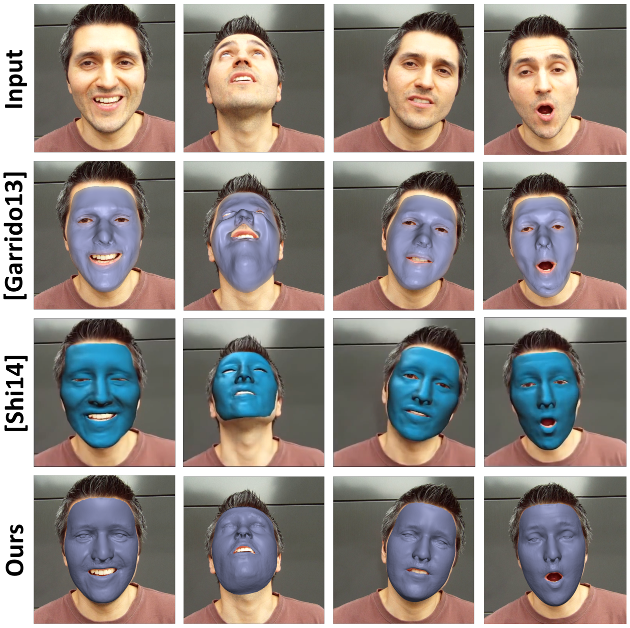

Fig. 4. State-of-the-art comparison to the approach by [Shi et al. 2014] and [Garrido et al. 2013]: Our monocular approach obtains better reconstruction quality than that of Shi *et al.*'s and Garrido *et al.*'s method. Note the better tracking on the coarse geometry as well as on the fine-scale detail layer.

Facial performance enhancement using dynamic shape space analysis. *ACM TOG 33,* 2, 13:1–13:12.

Garrido, P., Valgaert, L., Wu, C., and Theobalt, C. 2013. Reconstructing detailed dynamic face geometry from monocular video. *ACM TOG 32,* 6, 158:1–158:10.

Shi, F., Wu, H.-T., Tong, X., and Chai, J. 2014. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM TOG 33,* 6, 222:1–222:13.

Valgaerts, L., Wu, C., Bruhn, A., Seidel, H.-P., and Theobalt, C. 2012. Lightweight binocular facial performance capture under uncontrolled lighting. *ACM TOG 31,* 6, 187:1–187:11.

APPENDIX

A. LIST OF MATHEMATICAL SYMBOLS

Symbol	Description
$\mathcal{F} = \{f_t\}_{t=1}^{T}$	input video with T frames f_t
$\mathcal{M}$	triangle mesh
N, J	# of model vertices, triangle faces
$\mathbf{V}, \mathbf{N}, \mathbf{C}$	vertex, normal, reflectance set
$\mathbf{G}$	mesh topology
$\mathbf{v}_n, \mathbf{n}_n, \mathbf{c}_n$	vertex position, normal, albedo
$\mathcal{P}^s, \mathcal{P}^r, \mathcal{P}^e, \mathcal{P}^c$	shape, refl., expr., corr. model
$\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\tau}$	shape, refl., expr., corr. coeffs.
$\mathbf{E}_s, \mathbf{E}_r, \mathbf{E}_e$	linear shape, refl., expr. basis
$\mathbf{a}_s, \mathbf{a}_r$	shape, refl. average
$\mathbf{\Sigma}_s, \mathbf{\Sigma}_r, \mathbf{\Sigma}_e$	matrix of standard deviations
$\sigma_{\alpha_k}, \sigma_{\beta_k}, \boldsymbol{\sigma}_{\boldsymbol{\tau}_k}$	shape, refl., corr. std. dev.
$\mathcal{X} = (\mathbf{R}, \mathbf{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\tau})$	set of all model parameters
K_s, K_r, K_e, K_c	# of shape, refl., expr., corr. coeffs.
$\mathbf{E}_c$	manifold harmonics basis
$\mathbf{\Pi}$	perspective camera projection
B	# of spherical harmonics bands
Y_k	k-th SH basis function
$\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \cdots, \boldsymbol{\gamma}_{B^2}^\top)^\top$	spherical harmonics coeffs.
$\boldsymbol{\gamma}_b = (\gamma_b^r, \gamma_b^g, \gamma_b^b)^\top$	b-th coeff. vector
$\mathcal{B}$	illumination model
$\mathbf{C}_p$	personalized texture
$\mathbf{R}, \mathbf{t}$	camera orientation, position
$\mathcal{C}$	mapping world-to-camera
$\{\mathbf{A}_j\}_{j=1}^{J}$	per-face deformation gradients

Symbol	Description
$\mathbf{Q}$	polar decomposition (rotation)
$\mathbf{S}$	polar decomposition (shear)
$\phi(x)$	*box*-constraint on x
E_{total}	complete energy
$w_x,\ x \in \{s, r, \cdots\}$	weights in the energy function
$\mathbf{W}$	blendshape weight matrix
$\mathbf{X}$	affine regressor
$\mathbf{H}$	target attributes
λ	ridge parameter
$\mathbf{I}$	identity matrix
$\mathbf{p} = (\mathbf{p}_1^\top,\ \cdots, \mathbf{p}_J^\top)^\top$	deformation feature vectors

Received September 2015; accepted November 2015