Reconstruction of Personalized 3D Face Rigs from Monocular Video

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

We present a novel approach for the automatic creation of a personalized high-quality 3D face rig of an actor from just monocular video data, e.g. vintage movies. Our rig is based on three distinct layers that allow us to model the actor's facial shape as well as capture his person-specific expression characteristics at high fidelity, ranging from coarse-scale geometry to fine-scale static and transient detail on the scale of folds and wrinkles. At the heart of our approach is a parametric shape prior that encodes the plausible sub-space of facial identity and expression variations. Based on this prior, a coarse-scale reconstruction is obtained by means of a novel variational fitting approach. We represent person specific idiosyncrasies, which can not be represented in the restricted shape and expression space, by learning a set of medium-scale corrective shapes. Fine-scale skin detail, such as wrinkles, are captured from video via shading-based refinement, and a generative detail formation model is learned. Both the medium and fine-scale detail layers are coupled with the parametric prior by means of a novel sparse linear regression formulation. Once reconstructed, all layers of the face rig can be conveniently controlled by a low number of blendshape expression parameters, as widely used by animation artists. We show captured face rigs and their motions for several actors filmed in different monocular video formats, including legacy footage from YouTube, and demonstrate how they can be used for 3D animation and 2D video editing. Finally, we evaluate our approach qualitatively and quantitatively and compare to related state-of-the-art methods.

Categories and Subject Descriptors: I.3.7 [**Computer Graphics**]: Three-Dimensional Graphics and Realism—*Animation*

General Terms: Algorithms

Additional Key Words and Phrases: 3d model fitting, blendshapes, corrective shapes, shape-from-shading, facial animation, video editing

ACM Reference Format:

Garrido P., Zollhöfer M., Casas D., Valgaerts L., Varanasi K., Pérez P., and Theobalt C. 2015. Reconstruction of Personalized 3D Face Rigs from Monocular Video. ACM Trans. Graph. XX, Y, Article ZZZ (Month Year), 15 pages.
DOI = 10.1145/XXXXXXX.YYYYYYY
http://doi.acm.org/10.1145/XXXXXXX.YYYYYYY

1. INTRODUCTION

The creation of believable face animations for virtual actors in movies and in games, or for avatars in virtual reality or teleconferencing scenarios is a challenging task. Since human perception is attuned to quickly detect inaccuracies in face appearance and motion, animation artists spend tremendous effort to model and animate high quality facial animation rigs, in particular when photo-realism is the goal. A common practice for an artist is to design a face animation rig with custom-made control parameters that steer facial expression, face shape, and possibly face appearance and soft tissue deformation. The de facto standard to parametrize expression control is a blendshape model that linearly combines a set of basis expressions [Lewis et al. 2014]. Professional rigs often feature hundreds of control parameters, and it often takes many weeks of work to create such a rig for a specific actor, for instance from a laser scan of a face. The face rig is often animated from face motion capture data, a step requiring frequent manual intervention.

To simplify this complex animation pipeline, researchers developed different methods to automate some of its steps (see also Sec. 2). For instance, algorithms that use dense camera arrays and dense lighting arrays to reconstruct face geometry, facial performance and/or face appearance were developed [Beeler et al. 2010; Beeler et al. 2011; Alexander et al. 2009]. Approaches that extract components of face rigs from densely captured animation data, such as blendshape components [Neumann et al. 2013; Joshi et al. 2003], were also proposed, but despite its practical relevance, automatic rig creation received much less attention in research. Meanwhile, performance capture methods were further enhanced to work with only two or even one RGB or a depth camera, e.g. [Weise et al. 2011; Garrido et al. 2013; Cao et al. 2014; Shi et al. 2014]. However, to our knowledge, there is still no approach that fully-automatically combines both steps: reconstruct a detailed personalized modifiable face rig, *as well as* its animation, from only a single monocular RGB video of an actor filmed under general conditions.

In this paper, we propose such a method that builds a fully personalized 3D face rig, given just a single monocular input video (see Fig. 1). At the heart of our method is a new multi-layer parametric shape model that jointly encodes a plausible sub-space of facial identity, person-specific expression variation and dynamics, and fine-scale skin wrinkle formation (Sec. 4). On a coarse level, shape identity is parametrized using a principal component model, and facial expressions are parametrized with a generic blendshape model. Person-specific idiosyncrasies in expression and identity, which are not modeled in this generic space, are captured by a second layer of medium-scale corrective shapes. A generative model of wrinkle formation in the face constitutes the final most detailed layer. The medium and fine-scale layers are coupled to the coarse layer through a new sparse regression model learned from video (Sec. 6). The parameters of this model are personalized to an actor's video

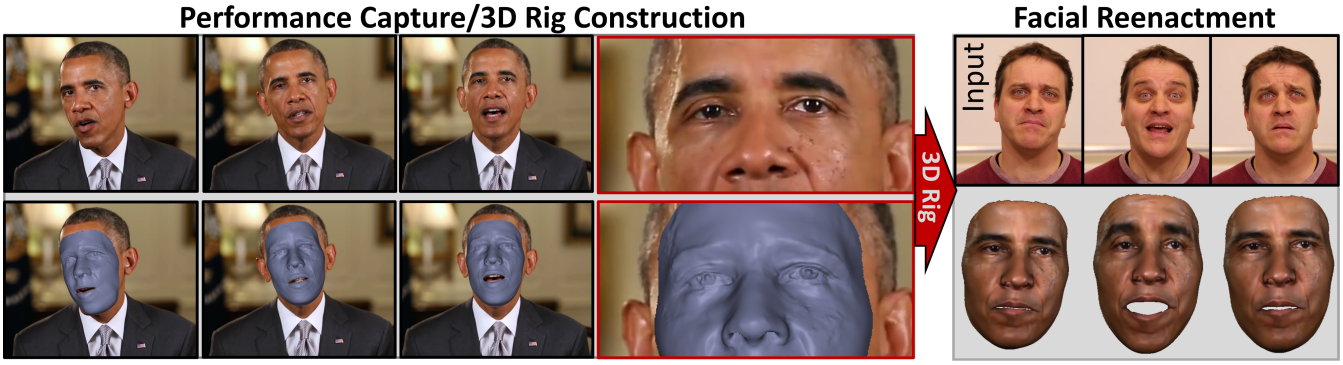

Fig. 1. Our approach reconstructs a fully personalized 3D face rig of the president of the United States of America given a single monocular video as input and learns medium, as well as fine-scale actor-specific idiosyncrasies. The facial rig can, for example, be used for reenactment.

by using a new variational fitting approach to recover the coarse and medium layers, and a shading-based refinement approach under general lighting to extract fine-scale detail (Sec. 5). The output of our algorithm is the personalized face model, blendshape expression parameters from the input video, as well as a detailed face albedo map and an incident lighting estimate. New face expressions of the rig with proper fine-scale detail can be created by simply modifying the blendshape parameters, which fits nicely into an animator's standard workflow. Our method captures detailed, personalized face rigs from arbitrary monocular video of actors, even from vintage footage, for which it would be impossible to automatically create a rig or capture the performance by any other means.

Our method improves over existing state-of-the-art approaches in several important ways. Unlike single-view or multi-view methods that only capture detailed deforming face meshes [Beeler et al. 2011; Valgaerts et al. 2012; Suwajanakorn et al. 2014], our approach additionally captures a personalized, modifiable parametric face rig. Some previous methods employed generic parametric expression and identity models for monocular facial performance capture. However, generic blendshape models and identity models alone [Cao et al. 2014; Garrido et al. 2013; Shi et al. 2014] fail to capture important person-specific expression and identity details learned by our approach. None of these approaches learns a generative wrinkle formation model from video. Generative models of face wrinkle formation were learned from high-quality expressions (out of a vast set of examples) captured with a dense sensor array [Bermano et al. 2014; Cao et al. 2015] or with depth cameras [Li et al. 2015], or also by interpolating dense high-quality scans in a video-driven way [Fyffe et al. 2014]. In contrast, our approach learns such a model from monocular RGB video alone. Some methods capture facial performances [Weise et al. 2011] and person-specific corrective shapes from RGB-D data [Bouaziz et al. 2013; Li et al. 2013], whereas our approach only requires monocular RGB video. Note also that our approach is fully-automatic and requires no manual intervention during model creation or tracking, as required in [Alexander et al. 2009; Bouaziz et al. 2013]. Our method needs no additional input other than a face video, meaning no specific sequence of face expressions [Ichim et al. 2015; Weise et al. 2011], no densely captured static face geometry [Fyffe et al. 2014; Valgaerts et al. 2012; Ichim et al. 2015], and no face detail regression model learned off-line [Cao et al. 2015].

The main contribution of this work is the *automatic extraction of a parametrized rig that models the correlation between coarse-scale blendshape weights and person-specific idiosyncrasies* on the medium and fine-scale detail layer just from monocular input data. Such a dependency has not yet been recovered by any other approach in the context of monocular video data. We show captured face rigs and their motion for several actors reconstructed from various monocular video feeds ranging from HD input to vintage video from YouTube. New face animations can be generated with these rigs and they can be used to realistically edit video footage. Additionally, our combined face modeling and capturing approach compares favorably to alternative monocular and multi-view methods in terms of reconstruction accuracy.

2. RELATED WORK

Static and Dynamic Face Capture. Several methods capture high quality static [Beeler et al. 2010] and dynamic [Beeler et al. 2011] face geometry using dense RGB camera rigs in controlled surroundings; some commercial systems, e.g. from Mova™, also fall into this category. If, in addition, the face is recorded under controlled lighting, highly detailed facial appearance or skin detail models can be captured, e.g. [Wenger et al. 2005; Graham et al. 2013; Klaudiny and Hilton 2012]. Huang *et al.* [2011] combine marker-based motion capture and high-quality 3D scanning for facial performance reconstruction, but no generative wrinkle model is learned. In contrast, our approach is designed for lightweight capture with a single RGB camera.

There is a large body of work in computer vision on face detection, face recognition, and sparse facial landmark tracking [Fasel and Luettin 2003]. A detailed survey of all these works is infeasible, and we focus on recent performance capture methods that reconstruct detailed moving geometry models. Valgaerts *et al.* [2012] took a step towards off-line lightweight capture of a deforming face mesh without an underlying face rig by using template tracking and shading-based refinement from binocular stereo. Other methods can track deforming face meshes using depth data from active triangulation scanners or RGB-D cameras [Wand et al. 2009; Popa et al. 2010], also at real-time rates [Zollhöfer et al. 2014], but require pre-designed mesh templates and do not build a detailed parametric face rig. Weise *et al.* [2011] capture facial performance in real-time by fitting a parametric blendshape model to RGB-D data. The model needs to be personalized by fitting it against a set of captured static face poses of an individual, and the approach cannot capture fine-scale detail. Recently, the first methods for facial performance capture from monocular RGB footage were proposed. Suwajanakorn *et al.* [2014] use monocular mesh deformation tracking and an identity PCA model learned from a large corpus of images captured

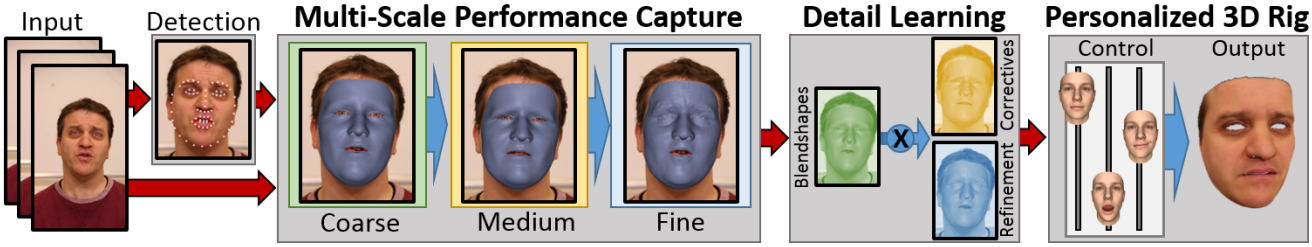

Fig. 2. Pipeline Overview: Starting from monocular video data, we first reconstruct the actor's identity and motion parameters based on a novel tracking energy, resulting in a multi-layer 3D rig. Finally, we learn the coupling between coarse-scale expression changes and medium as well as fine-scale surface detail.

under general illumination conditions to reconstruct a moving face mesh, but they do not simultaneously build a detailed parametric face animation rig. Garrido *et al.* [2013] adapt a generic template to a static 3D scan of an actor's face, then fit the blendshape model to monocular video off-line, and finally extract surface detail by shading-based shape refinement under general lighting. However, a wrinkle formation model is not learned, nor is a person-specific corrective layer built. Based on this model-based approach, Garrido *et al.* [2015] presented a method for virtual dubbing on monocular video. Shi *et al.* [2014] use a very similar tracking approach, but do not extract a high-fidelity parametrized 3D rig that contains a generative wrinkle formation model capturing the person-specific idiosyncrasies. Recently, Thies *et al.* [2015] presented an approach for real-time facial reenactment, but the method can not handle fine-scale surface detail and requires RGB-D camera input. Cao *et al.* [2014] use a learned regression model to fit, in real-time, a generic identity and expression model to RGB face video. However, no person specific correctives are learned, which reduces fitting accuracy, and no appearance and wrinkle models are built. In follow-up work [Cao et al. 2015], a regression model for face wrinkles learned on dense data from [Beeler et al. 2011] approximates but not truly reconstructs face detail, again without corrective and appearance modeling (see comparison in Sec. 7).

Face Modeling. Animation artists are used to manually creating face rigs of actors with custom-designed control parameters. They commonly resort to facial expression control using a set of blendshapes that span intuitive atomic face expressions and are linearly combined to obtain a new pose [Lewis et al. 2014]. Alternatively, physics-based muscle models can be used for animation control [Sifakis et al. 2005], either separately, or in conjunction with a blendshape model.

The facial anthropometry across people can also be modeled, e.g. as a parametric PCA space learned from a database of laser scans [Blanz and Vetter 1999; Blanz et al. 2003]. We employ such an identity PCA model as one component in our multi-layer face model. Automatically fitting a personalized parametric expression and identity model to an actor is a challenging problem. Dimensionality reduction techniques were applied to face animation data reconstructed with dense scanner setups to obtain parametric expression models [Tena et al. 2011]. However, such models lead to control dimensions that are often of global support and lack the semantic meaning and localized control built into blendshape models designed by artists. Generic blendshape models are used by some face tracking methods from monocular RGB video [Garrido et al. 2013] or RGB-D video [Weise et al. 2011], but need to be deformed into a static face scan or a set of scanned static expressions of an actor prior to tracking. Such generic blendshape adaptation fails to capture person-specific expression details, which is why some recent approaches estimate identity and blendshape parameters from captured face animations, and also person-specific correctives on top of this generic face model [Bouaziz et al. 2013; Li et al. 2013; Hsieh et al. 2015]. However, all these approaches require RGB-D camera input. Our model uses a corrective layer, too, but we learn it from monocular RGB video alone. Also, none of these previous methods capture a predictive fine-scale detail layer. Multi-linear models represent both identity and expression variations, and can be learned from laser scan databases [Vlasic et al. 2005]. Such a model was fitted to monocular video in [Shi et al. 2014], but is unable to capture person-specific idiosyncrasies in expression, as well as a wrinkle formation model.

Generative models of wrinkle formation were learned from a large corpus of facial performances [Bermano et al. 2014; Cao et al. 2015], or from depth camera data [Li et al. 2015]. Wrinkles can also be approximated in monocular video by video-driven interpolation of an actor-specific set of static face scans [Fyffe et al. 2014]. Small-scale transient detail was learned by a collection of local mappings using a data-driven framework [Huang et al. 2012]. Ma *et al.* [2008] infer facial detail displacement maps using a generative model, but require high-quality data captured with a professional camera and lighting setup for training. In contrast, our approach directly couples detail layer and blendshape weights by learning a generative geometric wrinkle model from monocular RGB video only.

Related to our method is the approach by Ichim *et al.* [2015] that fits a generic identity and blendshape model to a structure-from-motion-based reconstruction of the head in a static pose. They adapt the blendshape basis using monocular video of a sequence of specific expressions exercising the blendshape dimensions, making it unsuitable for legacy video footage. A parametric dynamic bump map is also learned from video to simulate some face detail. Several steps require manual intervention.

To our knowledge, our approach is the first to fully-automatically capture from general monocular video alone, without an initial 3D scan or a set of prescribed face expressions, a fully personalized face rig which is composed of a generic identity and blendshape model at the coarse level, a corrective personalized layer at the medium level, and a fine-scale generative detail layer.

3. OVERVIEW

In this section, we provide a brief overview (see Fig. 2) of our new approach to learn a high-quality personalized 3D face rig of an actor from unconstrained monocular video input, including TV programs or vintage movies. Our personalized face rig (Sec. 4) encodes the actor-specific facial geometry, appearance and motion on three layers: coarse-scale shape, medium-level correctives and

fine-scale detail on wrinkle level. To obtain this model, we first track a generic actor model from video by using a novel tracking energy (Sec. 5) that jointly optimizes for facial shape, expression and illumination parameters such that a photometric consistency measure is maximized. In this process, we also estimate camera parameters. Starting from this initial shape and motion estimate, the quality of the fit is further improved based on linear person-specific correctives. In addition, we use inverse rendering to solve for a wrinkle-level detail layer based on shading cues in the input images. A new sparse regression technique uses the recovered data as input to learn an actor-specific prediction model (Sec. 6) for the medium-scale correctives and the wrinkle-level detail based on coarse-scale expression changes. The output of our method is a personalized 3D rig, including all extracted parameters as well as a face albedo map. New realistic expressions of the rig can be conveniently created by simply modifying the blendshape weights, i.e. a small sub-set of the available control parameters, that are widely used in face animation. We evaluate the accuracy and prediction performance of our face rig qualitatively and quantitatively on several test sequences (Sec. 7). The recovered models seamlessly fit into the toolbox of animators and can be used in several applications, e.g. expression transfer, photo-realistic expression modification in video, and all fields of 3D face animation where even vintage actors can be revived.

4. MULTI-LAYER PERSONALIZED 3D MODEL

Our reconstruction process inverts the image formation and recovers the camera's extrinsic parameters, the scene lighting, and the face rig comprised by the actor's appearance, identity (shape) and expression (deformation) parameters. We parametrize facial identity and expression variation based on three different layers: a coarse-scale linear parametrization of identity and expression, medium-scale corrective shapes based on manifold harmonics and a fine-scale wrinkle-level detail layer, see Fig. 3. In the following, we explain these components in more detail.

4.1 Camera Parametrization

We assume a standard perspective pinhole camera with world space position $\mathbf{t} \in \mathbb{R}^3$ and orientation $\mathbf{R} \in \mathrm{SO}(3)$. Hence, $\mathcal{C}(\mathbf{v}) = \mathbf{R}^{-1}(\mathbf{v} - \mathbf{t})$ maps a world space point $\mathbf{v} \in \mathbb{R}^3$ to the camera's local coordinate frame. An image of the face rig in 3D world space is formed by projecting each surface point $\mathbf{v}$ of the model to the point $\mathbf{\Pi} \circ \mathcal{C}(\mathbf{v}) \in \mathbb{R}^2$ on the camera's image plane, using the camera's full perspective transformation $\mathbf{\Pi} : \mathbb{R}^3 \rightarrow \mathbb{R}^2$. To obtain $\mathbf{\Pi}$, we estimate optimal intrinsic camera parameters in a pre-processing step by jointly optimizing for the principal point, focal length and the actor specific parameters based on a sparse set of detected facial landmarks [Saragih et al. 2011] over the first 100 frames of the input video sequence.

4.2 Lighting and Appearance Model

We assume a pure *Lambertian* skin reflectance model as in [Garrido et al. 2013] and later works, e.g. [Shi et al. 2014; Suwajanakorn et al. 2014; Ichim et al. 2015]. This is a simplification of true skin reflectance that offers a good trade-off between complexity and quality of the obtained results. Since the scene is assumed to be purely Lambertian, the global illumination in the scene is represented using a spherical environment based on *Spherical Harmonics* (SH) basis functions [Müller 1966]. In spirit of Ramamoorthi and Hanrahan [2001], we use the first $B = 3$ SH bands to express the irradiance at a surface point with surface orientation $\mathbf{n}$ and skin albedo $\mathbf{c}$ in terms of the illumination coefficients $\boldsymbol{\gamma}$:

$$\mathcal{B}(\mathbf{n}, \mathbf{c} \mid \boldsymbol{\gamma}) = \mathbf{c} \cdot \sum_{b=1}^{B^2} \boldsymbol{\gamma}_b \mathrm{Y}_b(\mathbf{n}) \,. \tag{1}$$

Here, $\mathrm{Y}_b(\mathbf{n}) \in \mathbb{R}$ is the b-th SH basis function evaluated on the surface orientation $\mathbf{n}$. The irradiance is encoded using $B^2 = 9$ vector valued SH illumination coefficients $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \ \cdots, \boldsymbol{\gamma}_{B^2}^\top)^\top$, with $\boldsymbol{\gamma}_b = (\gamma_b^r, \gamma_b^g, \gamma_b^b)^\top$ a three dimensional vector that controls the irradiance separately for each color channel, leading to $3 \cdot 9 = 27$ parameters in our illumination model.

4.3 Coarse-Scale Identity and Expression Model

The head is represented as a triangle mesh $\mathcal{M} = (\mathbf{V}, \mathbf{C}, \mathbf{G})$ with the set of N vertices $\mathbf{V} = \{\mathbf{v}_n\}_{n=1}^N$, the set of per-vertex skin albedos $\mathbf{C} = \{\mathbf{c}_n\}_{n=1}^N$ and the mesh connectivity $\mathbf{G} \subset \mathbf{V} \times \mathbf{V}$. In addition, we associate with each $\mathbf{v}_n$ a normal $\mathbf{n}_n$ which is computed based on its 1-ring neighborhood. We parametrize the mesh's spatial embedding $\mathbf{V}$ and its per-vertex surface reflectance $\mathbf{C}$ using the statistical head prior of Blanz and Vetter [1999] that encodes the space of plausible human heads assuming a Gaussian distribution in the population. This linear head model is based on *Principal Component Analysis* (PCA) and has been constructed from 200 high-quality scans of Caucasian heads (100 males and 100 females). Hence, vertex positions $\mathbf{v}_n = \mathcal{P}_n^s(\boldsymbol{\alpha})$ and skin reflectances $\mathbf{c}_n = \mathcal{P}_n^r(\boldsymbol{\beta})$ can be parametrized as follows:

$$\text{Shape: } \mathcal{P}^s(\boldsymbol{\alpha}) = \mathbf{a}_s + \mathbf{E}_s \mathbf{\Sigma}_s \boldsymbol{\alpha} \,, \tag{2}$$

$$\text{Reflectance: } \mathcal{P}^r(\boldsymbol{\beta}) = \mathbf{a}_r + \mathbf{E}_r \mathbf{\Sigma}_r \boldsymbol{\beta} \,. \tag{3}$$

Here, $\mathbf{a}_s, \mathbf{a}_r \in \mathbb{R}^{3N}$ encode the per-vertex shape and reflectance of the average head. The shape and reflectance spaces are respectively spanned by the matrices $\mathbf{E}_s \in \mathbb{R}^{3N \times K_s}$ and $\mathbf{E}_r \in \mathbb{R}^{3N \times K_r}$ that contain the $K_s = K_r = 160$ first principal components of the shape and reflectance functions in their columns. Variations in shape and reflectance are controlled using the corresponding shape and reflectance parameters, $\boldsymbol{\alpha} \in \mathbb{R}^{K_s}$ and $\boldsymbol{\beta} \in \mathbb{R}^{K_r}$. The diagonal matrices $\mathbf{\Sigma}_s = \mathrm{diag}(\sigma_{\alpha_1}, \ \ldots, \ \sigma_{\alpha_{K_s}})$ and $\mathbf{\Sigma}_r = \mathrm{diag}(\sigma_{\beta_1}, \ \ldots, \ \sigma_{\beta_{K_r}})$ encode the standard deviations corresponding to the principal directions. Note, this scaling by the standard deviations guarantees a similar range of variation for the control parameters. Normally, we search for identity parameters in the range $[-3\sigma_\bullet, +3\sigma_\bullet]$, since this accounts for more than 99% of the variation and allows the model to rule out unlikely head shapes and skin reflectances.

We extend this linear shape model to also cover facial expressions by adding $K_e = 75$ delta blendshapes (i.e., displacements from the rest pose) taken from a combination of the *Emily* model [Alexander et al. 2009] and the *FaceWarehouse* database [Cao et al. 2014]:

$$\text{Expression: } \mathcal{P}^e(\boldsymbol{\alpha}, \boldsymbol{\delta}) = \mathcal{P}^s(\boldsymbol{\alpha}) + \mathbf{E}_e \mathbf{\Sigma}_e \boldsymbol{\delta} \,, \tag{4}$$

where the matrix $\mathbf{E}_e \in \mathbb{R}^{3N \times K_e}$ contains the K_e delta blendshapes in its columns, $\boldsymbol{\delta} \in [0, 1]^{K_e}$ contains the expression weights and $\mathbf{\Sigma}_e$ is a diagonal matrix of empirically determined scale factors. The delta blendshapes have been transferred to the topology of the model from [Blanz and Vetter 1999] using deformation transfer [Sumner and Popovic 2004]. Note that the blendshapes in the Emily model are redundant (i.e., the rows of $\mathbf{E}_e$ are not linear independent). We therefore use a sparsity prior on $\boldsymbol{\delta}$ (see Sec. 5).

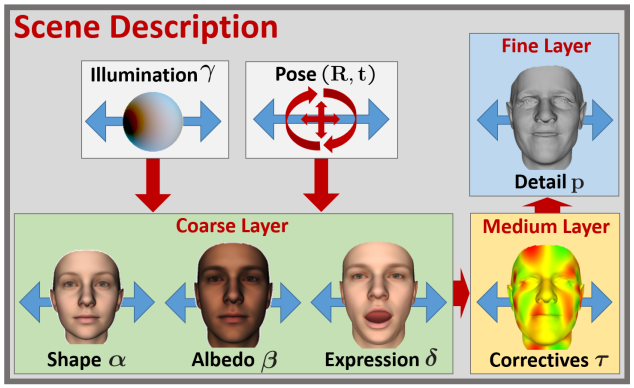

Fig. 3. Scene Description: We use a novel multi-layer person-specific rig to parametrize the identity as well as motion of an actor's face based on monocular video input. In addition, extrinsic camera parameters and the scene's illumination are extracted.

4.4 Medium-Scale Corrective Shapes

The coarse-scale model restricts the facial identity and expression to a $K_s = K_r = 160$ and $K_e = 75$ dimensional linear sub-space, respectively. Variations falling outside of this low-dimensional sub-space cannot readily be expressed with the model. Li *et al.* [2013] and Bouaziz *et al.* [2013] showed that it is beneficial to leave this limited sub-space to model characteristics in physiognomy and expression. In the spirit of [Bouaziz et al. 2013], we use *Manifold Harmonics* [Vallet and Lévy 2008; Lévy and Zhang 2010] to parametrize a medium-scale 3D deformation field:

$$\text{Correctives: } \mathcal{P}^c(\boldsymbol{\tau}) = \mathbf{E}_c \boldsymbol{\tau} \,. \tag{5}$$

Here, $\mathbf{E}_c = [H_1 \otimes I_{3\times3}, \ \cdots, H_{K_c} \otimes I_{3\times3}] \in \mathbb{R}^{3N\times3K_c}$ contains three copies of the K_c linear *Manifold Harmonics* basis functions $H_k \in \mathbb{R}^N$ as columns and the parameters $\boldsymbol{\tau} = [\boldsymbol{\tau}_1^\top, \ \cdots, \boldsymbol{\tau}_{K_c}^\top]^\top$ allow the control of the shape of the deformation field. Since we control a full 3D deformation field, each deformation coefficient $\boldsymbol{\tau}_k \in \mathbb{R}^3$ is a vector. Note that the spectral basis generalizes the *Fourier Transform* to the mesh domain. Here, H_k represent the $K_c = 80$ lowest-frequency eigenvectors of the Laplace Beltrami operator Δ_B on the average face. We use *cotan*-weights to discretize Δ_B and obtain a symmetric positive semi-definite linear operator. The eigenvectors are efficiently computed using the band-by-band *shift invert spectral transform* [Vallet and Lévy 2008; Lévy and Zhang 2010]. We apply the deformation field on vertex level, i.e. $\mathbf{v}_n + \mathcal{P}_n^c(\boldsymbol{\tau})$. Note that Bouaziz *et al.* [2013] infer correctives based on RGB-D data, while we robustly estimate them from RGB video alone (see Sec. 5). Ichim *et al.* [2015] do not learn correctives from RGB video but modify the blendshapes themselves; they mention that learning full correctives, as we do, will lead to better personalization but more involved optimization.

4.5 Fine-Scale Detail Layer

Correctives are well suited to capture medium-scale detail variations among individuals, but lack the ability to represent static and transient fine-scale surface detail such as wrinkles. To alleviate this problem, we make use of an additional per-vertex displacement field to account for such effects. These fine-scale deformations are encoded in the gradient domain based on deformation gradients [Sumner and Popovic 2004], which capture the non-translational surface deformation. Since rotation, scale and shear are inherently coupled in the per-face deformation gradients $\{\mathbf{A}_j\}_{j=1}^J$, where J is the number of triangles in the mesh, this representation does not allow for direct linear interpolation. We use polar decomposition [Higham 1986] to decompose the affine matrices $\mathbf{A}_j = \mathbf{Q}_j\mathbf{S}_j$ into their rotation $\mathbf{Q}_j$ and shear $\mathbf{S}_j$ components, and parametrize $\mathbf{Q}_j$ based on the matrix exponential (3 parameters) [Alexa 2002]. From $\mathbf{S}_j$ we extract the scaling factors (3 parameters) and the skewing factors (3 parameters), which represent the scale and parallel distortion along the coordinate axis, respectively. In total, this leads to 9 parameters per triangle, each allowing for simple direct linear interpolation. We stack these per-face representations in a feature vector $\mathbf{p} \in \mathbb{R}^{9J}$, which is used for storage and interpolation of $\mathcal{M}$'s fine-scale surface detail.

5. IDENTITY AND EXPRESSION RECONSTRUCTION

For a given video $\mathcal{F} = (f_t)_{t=1}^T$ of T image frames f_t, we seek to find the parameters of our personalized 3D model that best explain the shape (identity and expression) and skin reflectance of the actor's head, as well as the incident lighting at every frame of $\mathcal{F}$. We divide this task into two separate sub-problems:

—Recovery of the rigid head pose $(\mathbf{R}, \mathbf{t})$, the illumination $\boldsymbol{\gamma}$, and the coarse $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta})$ and medium-scale parameters $\boldsymbol{\tau}$.

—Refinement on top of the recovered medium-scale reconstruction to obtain the corresponding fine-scale detail layer $\mathbf{p}$.

We cast the first step as an energy minimization problem and recover the detail layer using shading-based refinement.

5.1 Energy Minimization

We seek the model parameters $\mathcal{X} = (\mathbf{R}, \mathbf{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\tau})$ in $\mathrm{SO}(3) \times \mathbb{R}^3 \times \mathbb{R}^{K_s} \times \mathbb{R}^{K_r} \times \mathbb{R}^{3B^2} \times \mathbb{R}^{K_e} \times \mathbb{R}^{3K_c}$ based on an *analysis-by-synthesis* approach that maximizes photo-consistency between a synthetically generated image of the head and an input RGB frame f_t. We formulate this as a constrained multi-objective optimization problem:

$$\mathcal{X}^* = \underset{\substack{\mathcal{X} \\ \text{s.t } \mathbf{0}\leq\boldsymbol{\delta}\leq\mathbf{1}}}{\operatorname{argmin}} \Big[E_{data}(\mathcal{X}) + E_{prior}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\tau}) \Big] \,. \tag{6}$$

The data objective E_{data} measures the photo-consistency of the synthetically generated image with respect to the input frame f_t. E_{prior} is a statistical prior that takes into account the likelihood of the identity and expression estimate. We impose a *box*-constraint on the expression parameters $\boldsymbol{\delta}$ to keep them in the range $[0, 1]$. To make the optimization more tractable, we relax the hard *box*-constraint on the expression parameters and model it as a soft-constraint E_{bound} directly in our reconstruction energy E_{total}. This leads to the following un-constrained highly non-linear optimization problem:

$$\mathcal{X}^* = \underset{\mathcal{X}}{\operatorname{argmin}} \underbrace{\Big[E_{data}(\mathcal{X}) + E_{prior}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\tau}) + E_{bound}(\boldsymbol{\delta}) \Big]}_{E_{total}(\mathcal{X})} \,. \tag{7}$$

Data Objective. The data term measures how well the personalized 3D model explains the input frame f_t. To this end, we consider a photo-consistency measure E_{photo} as well as the alignment to

salient facial features points $E_{feature}$:

$$E_{data}(\mathcal{X}) = w_f E_{feature}(\mathcal{X}) + w_p E_{photo}(\mathcal{X}) \, . \quad (8)$$

The weights w_f and w_p control the relative importance of these two objectives. Photo-consistency is measured on a per-vertex level. At vertex $\mathbf{v}_n = \mathcal{P}_n^e(\boldsymbol{\alpha}, \boldsymbol{\delta}) + \mathcal{P}_n^c(\boldsymbol{\tau})$, with associated reflectance $\mathbf{c}_n = \mathcal{P}_n^r(\boldsymbol{\beta})$ and normal $\mathbf{n}_n$ dependent on same parameters, it compares the surface color $\mathcal{B}(\mathbf{n}_n, \mathbf{c}_n \mid \boldsymbol{\gamma})$ synthesized according to model (1) with the actual color $f_t[\Pi \circ \mathcal{C}(\mathbf{v}_n)]$ in the input image. The corresponding energy reads:

$$E_{photo}(\mathcal{X}) = \sum_{n=1}^{N} \parallel f_t[\Pi \circ \mathcal{C}(\mathbf{v}_n)] - \mathcal{B}(\mathbf{n}_n, \mathbf{c}_n \mid \boldsymbol{\gamma}) \parallel_2^2 \, . \quad (9)$$

In addition, we take the alignment of salient facial features into account. To this end, we measure the distance between image projections $\{\Pi \circ \mathcal{C}(\mathbf{v}_{n_\ell})\}_{\ell=1}^{L}$ of a selection of $L = 66$ feature vertices on the model and corresponding L distinct detected facial landmarks $\{\mathbf{y}_\ell\}_{\ell=1}^{L}$ in the input image:

$$E_{feature}(\mathcal{X}) = \sum_{\ell=1}^{L} \parallel \Pi \circ \mathcal{C}(\mathbf{v}_{n_\ell}) - \mathbf{y}_\ell \parallel_2^2 \, . \quad (10)$$

We track the 2D facial features with an off-the-shelf algorithm [Saragih et al. 2011] and improve the landmark trajectories by using optical flow between automatically selected key-frames [Garrido et al. 2013]. To select the 3D feature points $\{\mathbf{v}_{n_\ell}\}$ on the model, we automate and extend the strategy proposed by Garrido *et al.* [2013]. In a pre-processing step, we synthesize $K_e = 75$ different facial expressions of the average person by activating one expression weight δ_k at a time and render frontal views under a fixed user-defined illumination. Afterward, we run the off-the-shelf face tracker to detect the 2D landmarks in the synthetically generated images. Landmarks are back-projected to the nearest vertices on the 3D model, discarding those that fall outside of the face region or inside the mouth cavity. Finally, the 3D positions corresponding to the same landmark are averaged and assigned to the nearest valid vertex of the model.

Prior Objective. 3D reconstruction from monocular RGB input is an ill-posed problem (depth ambiguity), since many spatial configurations of mesh vertices lead to a similar projection in the camera. We tackle this issue by incorporating suitable priors E_{prior} into our energy. This allows to disambiguate reasonable from unreasonable configurations and steer the optimization into the right direction. To this end, we use two probabilistic shape priors (E_{prob_1}, E_{prob_2}) and a sparsity prior E_{sparse} on the expression coefficients:

$$E_{prior}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\tau}) = E_{prob_1}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) + E_{prob_2}(\boldsymbol{\tau}) + E_{sparse}(\boldsymbol{\delta}) \, . \quad (11)$$

The probability of a certain scene configuration is accounted for by assuming multiple Gaussian distributions over the parameters:

$$E_{prob1}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = w_s \sum_{k=1}^{K_s} \left(\frac{\alpha_k}{\sigma_{\alpha_k}} \right)^2 + w_r \sum_{k=1}^{K_r} \left(\frac{\beta_k}{\sigma_{\beta_k}} \right)^2 + w_l \sum_{b=1}^{B^2} \left(\frac{\gamma_b}{\boldsymbol{\sigma}_{\boldsymbol{\gamma}_b}} \right)^2 , \quad (12)$$

with the division in the last term being component-wise. Here, w_s, w_r and w_l weigh the different objectives. As in [Blanz and Vetter 1999; Zollhöfer et al. 2014], we restrict the shape weights $\boldsymbol{\alpha}$ and reflectance coefficients $\boldsymbol{\beta}$ to stay statistically close to the mean using ℓ_2-regularization. Since we do not know the standard deviations of the lighting coefficients $\boldsymbol{\gamma}$, we impose *Tikhonov*-regularization constraints [Hoerl and Kennard 2000] by setting $\boldsymbol{\sigma}_{\boldsymbol{\gamma}_b} = [1, 1, 1]^\top$.

In addition to the coarse scale parameters, we also regularize the medium-scale shape correctives based on their standard deviations (squared eigenvalues of the H_k (Sec. 4.4)) and enforce temporal smoothness with respect to the corresponding result of the previous frame $\boldsymbol{\tau}^{prev}$:

$$E_{prob2}(\boldsymbol{\tau}) = w_z \sum_{k=1}^{K_c} \left(\frac{\boldsymbol{\tau}_k}{\boldsymbol{\sigma}_{\boldsymbol{\tau}_k}} \right)^2 + w_t \| \boldsymbol{\tau} - \boldsymbol{\tau}^{prev} \|_2^2 , \quad (13)$$

with component-wise divisions in the first term. Here, w_z and w_t are the weights controlling the importance of the different objectives.

Following [Bouaziz et al. 2013], we also impose ℓ_1-regularization on the expression weights $\boldsymbol{\delta}$ to enforce sparsity. This avoids potential blendshape compensation artifacts due to the inherent redundancy in the expression basis:

$$E_{sparse}(\boldsymbol{\delta}) = w_d \sum_{k=1}^{K_e} |\delta_k| \, . \quad (14)$$

Boundary Constraint. The blendshape parameters are restricted to a reasonable range ($\delta_k \in [0, 1]$) by adding a soft *box*-constraint with a weight of w_b to the energy:

$$E_{bound}(\boldsymbol{\delta}) = w_b \sum_{k=1}^{K_e} \phi(\delta_k) \, . \quad (15)$$

The function ϕ adds a penalty to the energy if and only if its parameter leaves the trusted region:

$$\phi(x) = \begin{cases} x^2 & \text{if } x < 0, \\ 0 & \text{if } 0 \leq x \leq 1, \\ (x-1)^2 & \text{if } x > 1. \end{cases} \quad (16)$$

We use a symmetric quadratic penalizer outside of the trusted region to tightly enforce the bounds of this constraint.

5.2 Optimization

Given the input video $\mathcal{F} = \{f_t\}_{t=1}^{T}$, we find the best parameters $\mathcal{X}$ by minimizing the non-linear objective $E_{total}(\mathcal{X})$ using a multi-step optimization strategy based on multiple Levenberg-Marquardt [Levenberg 1944; Marquardt 1963; Moré 1978] optimization stages. The individual steps are summarized in Alg. 1. The rigid head pose ($\mathbf{R}$ and $\mathbf{t}$) is initialized using the POSIT algorithm [David et al. 2004] on the detected facial landmarks, and ($\boldsymbol{\alpha}$, $\boldsymbol{\delta}$) are initialized by solving Eq. 10 with the parametric priors $E_{prob1}(\boldsymbol{\alpha})$, $E_{sparse}(\boldsymbol{\delta})$, and $E_{bound}(\boldsymbol{\delta})$, i.e., we optimize for ($\boldsymbol{\alpha}$, $\boldsymbol{\delta}$) using only the facial feature point subspace. The other parameters ($\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\tau}$) are initially set to zero. We start by using the first $T_{first} \approx 100$ frames of the sequence to reconstruct a coarse-scale estimate of the actor's person-specific identity ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) and of the illumination $\boldsymbol{\gamma}$ in the scene. This step does not consider the corrective parameters $\boldsymbol{\tau}$, hence the corresponding terms are removed from the energy. The resulting per-frame estimates of the actor's identity are combined using a floating average.

Before we track the complete sequence in the next stage, we generate an actor-specific skin reflectance map $\mathbf{C}_p$ that replaces the per-vertex reflectance estimates from the parametric actor model. To this end, we follow a similar strategy as in [Garrido et al. 2015], and

Algorithm 1 Multi-Step Optimization Strategy

$(\mathbf{R}, \mathbf{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\tau}) \leftarrow$ Initialize();

for (the first T_{first} frames f_t) **do** ▷ Identity Estimation
  while (not converged) **do**
    $(\mathbf{R}, \mathbf{t}) \leftarrow$ Estimate_Head_Pose();
    $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \leftarrow$ Estimate_Identity_And_Illumination();
    $(\boldsymbol{\delta}) \leftarrow$ Estimate_Expression();
  end while
end for

$(\mathbf{C}_p) \leftarrow$ Build_Person_Specific_Albedo_Map();

for every frame $f_t \in \mathcal{F}$ **do**
  while (not converged) **do** ▷ Coarse-Scale
    $(\mathbf{R}, \mathbf{t}) \leftarrow$ Estimate_Head_Pose();
    $(\boldsymbol{\delta}) \leftarrow$ Estimate_Expression();
  end while
  while (not converged) **do** ▷ Medium-Scale
    $(\mathbf{R}, \mathbf{t}) \leftarrow$ Estimate_Head_Pose();
    $(\boldsymbol{\tau}) \leftarrow$ Estimate_Correctives();
  end while
  $(\mathbf{p}) \leftarrow$ Compute_Detail_Layer(); ▷ Fine-Scale
end for

compute per pixel albedo values by dividing through the lighting term (sum on the right hand side of (1)) on a subset of 10 frames. The resulting albedo values are averaged in the final map $\mathbf{C}_p$ using the aligned model. This refined appearance step drastically improves the subsequent tracking performance, since the generated reflectance map better resembles the actor's appearance (i.e. facial hair and fine-scale skin detail are explicitly accounted for, see also [Zollhöfer et al. 2014]). Then, we keep the identity parameters $\boldsymbol{\alpha}$ fixed and the complete sequence is tracked again, starting from the first frame. For each frame f_t, we first re-estimate the head pose $(\mathbf{R}, \mathbf{t})$ and compute the best fitting blendshape coefficients $\boldsymbol{\delta}$. The coarse-scale shape estimate and the head pose are then improved by optimizing for the best corrective parameters $\boldsymbol{\tau}$, as well as $\mathbf{R}$ and $\mathbf{t}$, based on the full reconstruction energy (see Eq. 6). Note that in this step the blendshape coefficients $\boldsymbol{\delta}$ stay fixed.

The next processing step (see below) reconstructs a fine-scale detail layer $\mathbf{p}$ based on shading-based shape refinement by exploiting shading cues in the input RGB frame.

5.3 Shading-based Refinement

Given the medium-scale result $\mathcal{M}$ (at every frame) of the previous optimization, fine-scale static and transient surface details (i.e. wrinkles and folds) are recovered from shading cues in the input RGB images by adapting the shading-based refinement approach under unknown lighting and albedo proposed by Garrido *et al.* [2013]. We compute shading-based refinement on a per-vertex level, yielding a high-quality refined mesh $\hat{\mathcal{M}}$. We use the previously estimated reflectance and illumination as initialization. A refinement optimization then adapts the mesh's vertex positions via inverse rendering optimization such that the synthesized shading gradients match the gradients of the illumination in the corresponding input RGB image as best as possible. To further regularize this ill-posed problem, spatial and temporal detail smoothness is enforced as a soft constraint [Garrido et al. 2013; Valgaerts et al. 2012]. The final vertex normals are computed by averaging over a temporal window of size 5 for stability [Nehab et al. 2005]. We store the deformation field

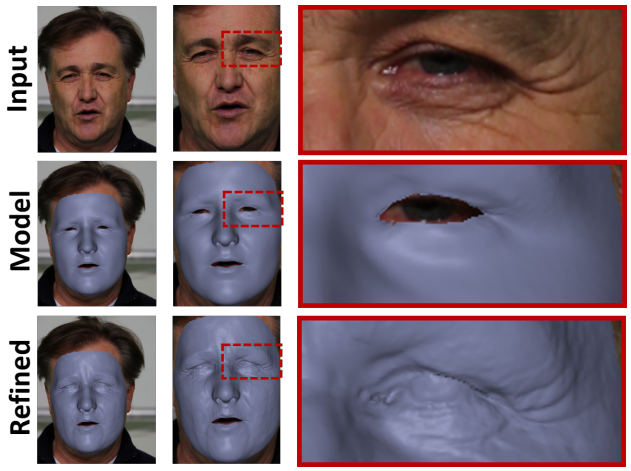

Fig. 4. Shape Refinement: We exploit shading cues in the input image (top) to augment the medium-scale model (middle) with fine-scale static and transient surface detail (bottom).

between the medium-scale result $\mathcal{M}$ and the refined high-quality geometry $\hat{\mathcal{M}}$ using our deformation gradient-based feature vector representation $\mathbf{p}$ introduced in Sec. 4. Compared to $\mathcal{M}$, the resulting high-quality reconstructions exhibit a considerable amount of fine-scale surface detail (see Fig. 4).

6. LEARNING TO PREDICT THE DETAIL LAYERS

The output of the previous processing step is a personalized 3D model $\hat{\mathcal{M}}_t$ for each of the T frames f_t that includes a coarse-scale, medium-scale and fine-scale detail layer. While the coarse-scale parametric blendshape rig allows for intuitive modification of the rig – e.g. by an artist – there is no equally convenient and semantically meaningful way to create medium and fine-scale details that match new expressions. To alleviate this problem, we learn the correlation between blendshapes and the higher detail layers, thus enabling full control of all detail levels by only using the blendshape coefficients.

6.1 Input Data

Our learning algorithm takes as input the reconstructed sequence of blendshape weights $\Delta_{\mathcal{F}} = \{\boldsymbol{\delta}^{(t)}\}_{t=1}^{T}$, the correctives $\mathcal{T}_{\mathcal{F}} = \{\boldsymbol{\tau}^{(t)}\}_{t=1}^{T}$ and the deformation gradients $P_{\mathcal{F}} = \{\mathbf{p}^{(t)}\}_{t=1}^{T}$ encoding the fine-scale detail layer. In the following, we propose a novel sparse and affine regression strategy to learn a mapping between activated blendshape weights and the detail layers that takes account of the local support of the expression basis.

6.2 Affine Parameter Regression

Given a sequence of input motion parameters $\Delta_{\mathcal{F}}$ and a corresponding sequence of details $\mathcal{H} \in \{\mathcal{T}_{\mathcal{F}}, P_{\mathcal{F}}\}$, we seek to find an affine mapping to encode their correlation. To this end, we stack the weights of the $K_e = 75$ blendshapes in a matrix $\mathbf{W}$:

$$\mathbf{W} = \left[\begin{array}{c|c|c} \boldsymbol{\delta}^{(1)} & \cdots & \boldsymbol{\delta}^{(T)} \\ \hline 1 & \cdots & 1 \end{array}\right] \in \mathbb{R}^{(K_e+1)\times T} . \quad (17)$$

Note, the last row of $\mathbf{W}$ implements a constant bias in the estimation that is especially important if certain blendshape weights are not activated in the training set. The detail layer $\mathcal{H}$ is stacked accordingly

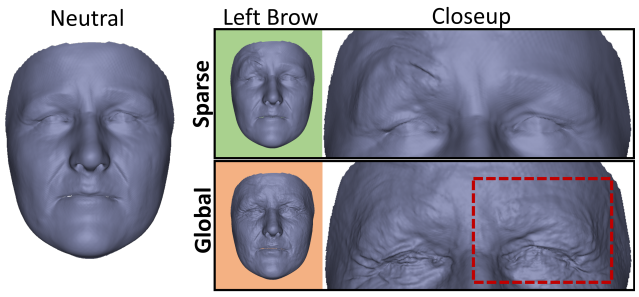

Fig. 5. Sparse vs. global fine-scale detail prediction: Our novel sparse regression formulation (top) obtains more realistic results than global regression (bottom). Note the wrong transient detail around the right eye (red) when the left eyebrow's blendshape is triggered.

in a corresponding matrix $\mathbf{H} \in \mathbb{R}^{H \times T}$. We remark that the fine-scale detail $H = 9J$ (with J the number of mesh triangles), since we regress the per-face deformation gradients. For the medium-scale detail layer, we regress the weights $\boldsymbol{\tau}$, therefore $H = 3K_c = 240$.

The task is to learn an affine mapping $\mathbf{X} \in \mathbb{R}^{H \times (K_e+1)}$ that maps the blendshape weights to the corresponding details $\mathbf{XW} = \mathbf{H}$. We solve this problem in a least-squares sense by adding a ridge regularizer on $\mathbf{X}$:

$$\mathbf{X}^* = \underset{\mathbf{X}}{\operatorname{argmin}} \parallel \mathbf{XW} - \mathbf{H} \parallel_{\mathcal{F}}^2 + \lambda \parallel \mathbf{X} \parallel_{\mathcal{F}}^2 \ , \qquad (18)$$

where $|| \ . \ ||_{\mathcal{F}}$ denotes the Frobenius norm, and $\lambda = 1.0$ is a user-defined ridge parameter. Such a linear model is known as ridge regression [Hoerl and Kennard 2000]. A closed form least-squares solution for $\mathbf{X}^*$ is given by:

$$\mathbf{X}^* = (\mathbf{W}^\top \mathbf{W} + \lambda \mathbf{I})^{-1} \mathbf{W}^\top \mathbf{H} \, , \qquad (19)$$

where $\mathbf{I}$ denotes the identity matrix.

6.3 Sparse Affine Regression of Fine Scale Details

For the medium-scale layer of correctives ($\mathcal{H} = \mathcal{T}_{\mathcal{F}}$), simple affine regression is sufficient to obtain high-quality results, since the spectral basis has global support. However, the same strategy leads to artifacts when used for the prediction of small-scale surface detail ($\mathcal{H} = P_{\mathcal{F}}$), e.g. detail showing up even if the triggered blendshape does not influence the corresponding surface region (see Fig. 5). To alleviate this problem, we exploit the spatial support of the blendshape basis during training and find the best affine mapping $\mathbf{X}_j^*$ for each triangle j independently:

$$\mathbf{X}_j^* = \underset{\mathbf{X}_j}{\operatorname{argmin}} \parallel \mathbf{X}_j \mathbf{D}_j \mathbf{W} - \mathbf{H}_j \parallel_{\mathcal{F}}^2 + \lambda \parallel \mathbf{X}_j \parallel_{\mathcal{F}}^2 \ , \qquad (20)$$

where $\mathbf{H}_j = [\mathbf{p}_j^{(1)}, \ \cdots, \mathbf{p}_j^{(T)}] \in \mathbb{R}^{9 \times T}$ and $\lambda = 0.1$. The spatial support of the k-th blendshape with respect to the j-th triangle is encoded in the diagonal discriminator matrix $\mathbf{D}_j = \operatorname{diag}(d_1^j, \ \cdots, \ d_{K_e}^j, 1) \in \mathbb{R}^{(K_e+1) \times (K_e+1)}$. This allows each triangle to switch on or off certain blendshapes based on their influence:

$$d_k^j = \begin{cases} 1 & \text{if } \delta_k \text{ influences the } j\text{-th triangle,} \\ 0 & \text{otherwise.} \end{cases}$$

Due to some outlier support regions in the blendshapes, we use $K_e = 75$ manually corrected support masks rather than the actual spatial support to compute $\mathbf{D}_j$. This novel affine sparse regression strategy for fine-scale details produces superior results, as illustrated in Fig. 5.

6.4 Synthesizing Medium-Scale Correctives

Given new blendshape weights (with 1 appended) $\hat{\boldsymbol{\delta}} \in [0,1]^{K_e+1}$, the medium-scale corrective layer is predicted as $\hat{\boldsymbol{\tau}} = \mathbf{X}^* \hat{\boldsymbol{\delta}}$, where $\mathbf{X}^*$ is defined as (19) with $\mathcal{H} = \mathcal{T}_{\mathcal{F}}$. Afterward, we reconstruct the deformation field $\mathcal{P}^c(\hat{\boldsymbol{\tau}})$ and apply it on a per-vertex level to the coarse-scale model, yielding $\hat{\mathbf{v}}_n = \mathcal{P}_n^e(\boldsymbol{\alpha}, \hat{\boldsymbol{\delta}}) + \mathcal{P}_n^c(\hat{\boldsymbol{\tau}})$. Since the regressed 3D displacements are not rotation invariant, this step is executed in canonical model coordinates.

6.5 Synthesizing High-Frequency Detail Variation

The high-frequency detail is synthesized on top of the medium-scale result $\hat{\mathbf{v}}_n$ leading to the final embedding $\tilde{\mathbf{v}}_n$. Given the new blendshape weights $\hat{\boldsymbol{\delta}}$, we predict the detail $\tilde{\mathbf{p}}_j = \mathbf{X}_j^* \hat{\boldsymbol{\delta}}$ for the j-th triangle, where $\mathbf{X}_j^*$ is defined as (20). From the 9-dimensional vector $\tilde{\mathbf{p}}_j$, we recover the per-face affine transformation matrix $\tilde{\mathbf{A}}_j$. Finally, we use the deformation transfer approach by Sumner and Popović [2004] to augment the medium-scale result with the fine-scale surface detail. For rotation invariance, we apply this transformation in canonical model coordinates. Note that we do not learn nor regress fine-scale detail for the surface region inside the eyes. Instead, we compute the mean deformation over the entire sequence and keep it fixed in the synthesis.

7. RESULTS

In this section, we show applications for the reconstructed 3D rig, present a qualitative and quantitative evaluation and perform a thorough comparison with respect to the state-of-the-art (see also supplementary video). First, we give a general overview of the used test sequences, parameter values and runtime requirements.

Input. We demonstrate the robustness of our approach for a wide range of scenarios, from controlled studio setups to uncontrolled legacy video footage. In total, we evaluated our approach on 9 test sequences; three indoor sequences captured in a controlled setup (SUBJECT1, SUBJECT2, SUBJECT3), two outdoor sequences (SUBJECT4, SUBJECT5) and four legacy videos (ARNOLD YOUNG, ARNOLD OLD, OBAMA, BRYAN) freely available on the Internet and downloaded from YouTube (see additional document for links to all sequences and their specs). The reconstructed facial rig consists of $N = 200$k vertices and $J = 400$k triangle faces.

Parameters. The facial performance capture stage of our pipeline relies on weights that specify the relative importance of the different objectives. During our tests, it turned out that our approach is insensitive to the specific choice of parameters. We use the following fixed weights in all our experiments: $w_f = 0.5$, $w_p = 1$, $w_s = 0.01$, $w_r = 1$, $w_l = 0.1$, $w_z = 40$, $w_t = 4$, $w_d = 100$ and $w_b = 10^9$.

Runtime. Overall, our CPU implementation takes several hours to process a sequence of 1k frames when executed on an Intel Core i7-3770 CPU (3.4 GHz). Per frame, our approach requires 30ms for facial landmark extraction, 1.5sec for landmark refinement, 40sec for identity fitting (only run for the first 100 frames), 15sec for coarse-layer tracking, 9sec for medium-layer correctives and 110sec for fine-scale shape refinement. Our sparse regression takes 10ms for the medium layer and 2sec for the fine-scale detail layer. We believe that a drastic reduction of the computation time is possible

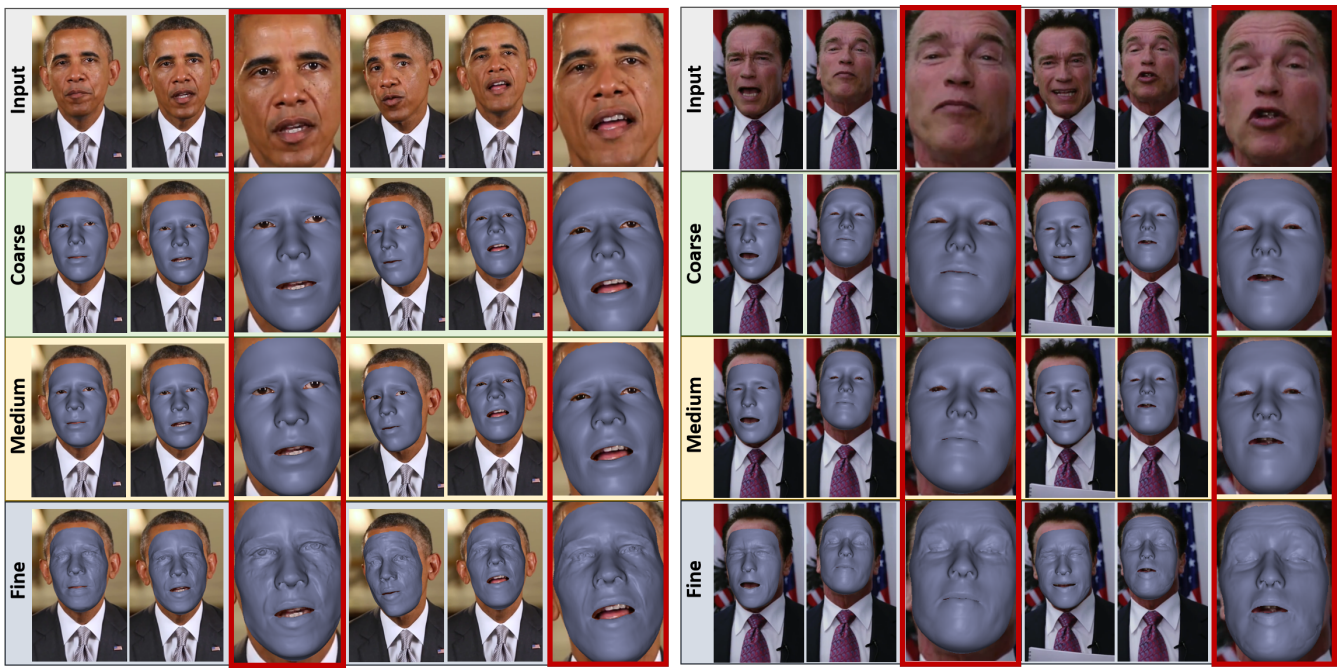

Fig. 6. Facial performance capture results on OBAMA (left) and ARNOLD OLD (right) sequence: Given a monocular video of an actor as input (first row), our approach obtains a high-quality reconstruction of his shape and motion on multiple parametrized layers: Coarse-scale shape and motion (second row), medium-scale correctives (third row) and fine-scale wrinkle-level surface detail (forth row).

by harnessing the data parallel processing power of modern GPUs, as recently demonstrated for non-linear optimization [Thies et al. 2015; Wu et al. 2014; Zollhöfer et al. 2014].

7.1 Application Scenarios

Our method automatically creates a fully parametrized facial 3D rig of an actor given just monocular video data as input. The obtained rig can be exploited for many different application scenarios, *e.g.* interactive modeling, video modification and facial reenactment.

Interactive Editing. To demonstrate the versatility of our representation, we allow the interactive modification of blendshape parameters to explore the rig's expression space, see Fig. 7 (SUBJECT2). The automatically predicted person-specific medium and fine-scale surface detail plausibly matches the new coarse-scale facial expression. Note that these novel expressions are not included in the training set that was used to learn the regressor.

Video Modification. Since we recover an estimate of the scene lighting as well as the intrinsic and extrinsic camera parameters, we can exploit our high-quality facial 3D rig to photo-realistically modify the face in the original video. To this end, we render a modified face model under the estimated lighting and then overlay the correctly lit face on top of the video. For instance, we exchange the regressed fine-scale detail layer of ARNOLD YOUNG and SUBJECT2 with that of the fine-scale layer learned on ARNOLD OLD which contains more face wrinkles. We then overlay the resynthesized face that contains Arnold Old's wrinkles on top of the original video, akin to a virtual aging edit. Keeping the medium and fine-scale detail layer of SUBJECT2, we additionally change the expression of this subject by lifting the left eyebrow and overlay the modified face rig with the video in a photo-realistic way, as shown in Fig. 9 (see also accompanying video).

Facial Reenactment. Since the facial rig is completely parametrized, we can transfer facial performances between different actors, see Fig . 8. Note that we infer the target actor's person-specific medium and fine-scale detail for every transferred expression. This leads to more natural and realistic results, since it preserves person specific idiosyncrasies. The creation of the rig and the animation is fully automatic and solely based on one single monocular video sequence, i.e. neither a high-quality face scan [Garrido et al. 2013] nor a community photo collection [Suwajanakorn et al. 2014] of the actor has been used in the process.

7.2 Qualitative and Quantitative Analysis

Our approach is based on a monocular performance capture method that estimates the actor's facial identity and tracks his facial expressions. Tracking progresses in a coarse-to-fine manner on the three layers: Coarse-scale shape, medium-scale correctives and fine-scale wrinkle-level detail. Fig. 6 shows the output tracking results on the three layers of our personalized 3D rig for OBAMA and ARNOLD OLD. Note, the finer scale layers do not only lead to more realistic results in terms of high-frequency detail, but also deliver tracking results of superior accuracy. In addition, we evaluate the geometric accuracy of the reconstruction in a neutral pose (the mean error is 1.8mm, as shown in Fig. 10). For this comparison on SUBJECT1, a sequence of high-quality ground-truth meshes has been generated using the binocular facial performance capture approach of Valgaerts *et al.* [2012].

To evaluate the prediction accuracy, we trained our sparse affine regressor on the first 700 frames of the test sequence (2000 frames in total) and regressed the medium and fine-scale detail layers on the

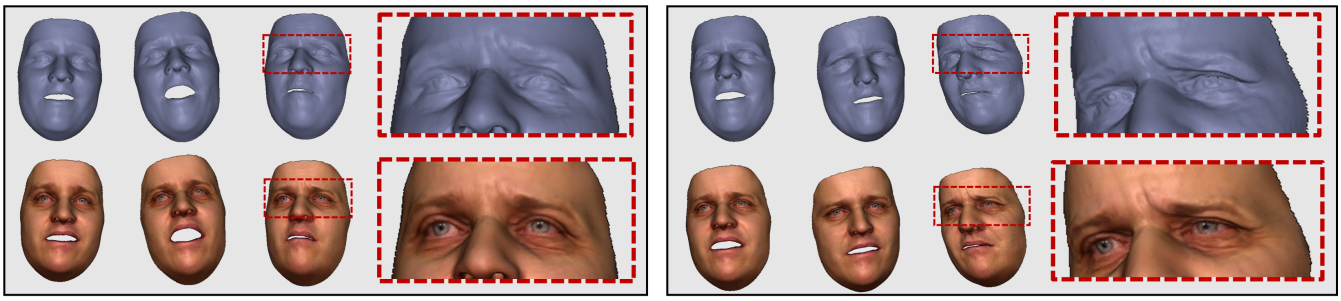

Fig. 7. Interactive Editing: Our high-quality parametrized 3D rig allows the creation of novel and expressive poses of an actor by interactively adapting the corresponding blendshape weights. Here, we show 6 poses of SUBJECT2 without (top) and with texture (bottom). Note that the medium and fine-scale details (top) have been automatically predicted using the learned sparse affine regression model.

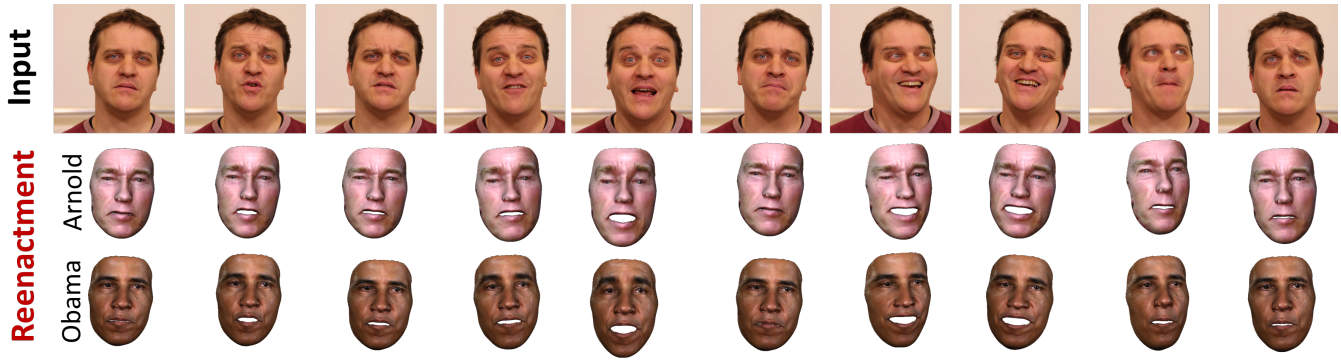

Fig. 8. Facial Reenactment: We retarget the rigid and non-rigid head motion of an input actor (top row) to the high-quality 3D rigs of ARNOLD OLD (middle) and OBAMA (bottom). Note that the target actor's characteristics are maintained, since we regress the detail layers.

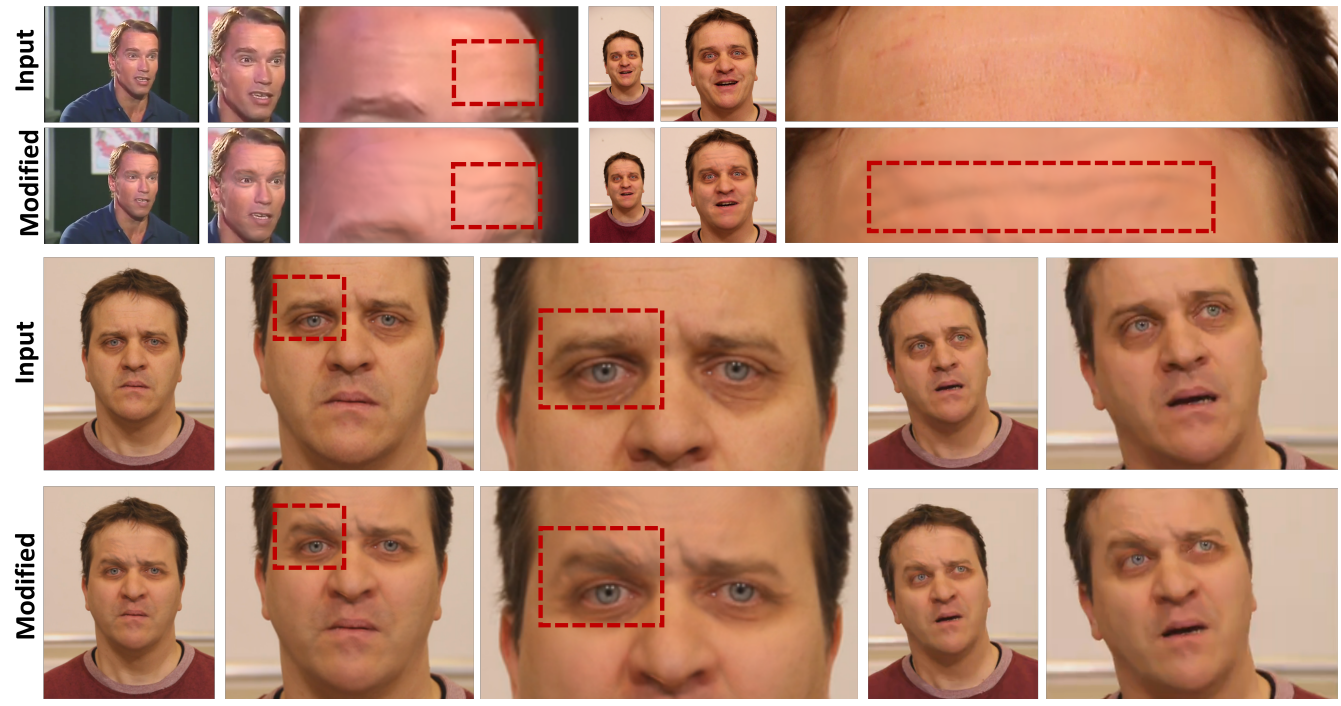

Fig. 9. Video Modification: We exchange the fine-scale detail layer of ARNOLD YOUNG and SUBJECT2 with that of the fine-scale layer estimated on ARNOLD OLD, thus adding slight wrinkles to the sequence. We also virtually lift the left eyebrow of SUBJECT2 (see the complete sequences in the accompanying video).

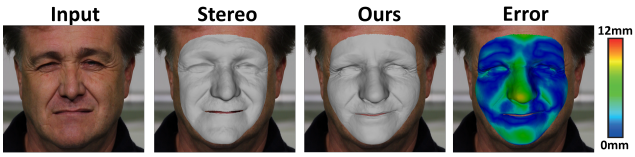

Fig. 10. Geometric Accuracy: Our monocular approach obtains similar quality (1.8mm mean error) to that of the high-quality binocular approach of Valgaerts *et al.* [2012].

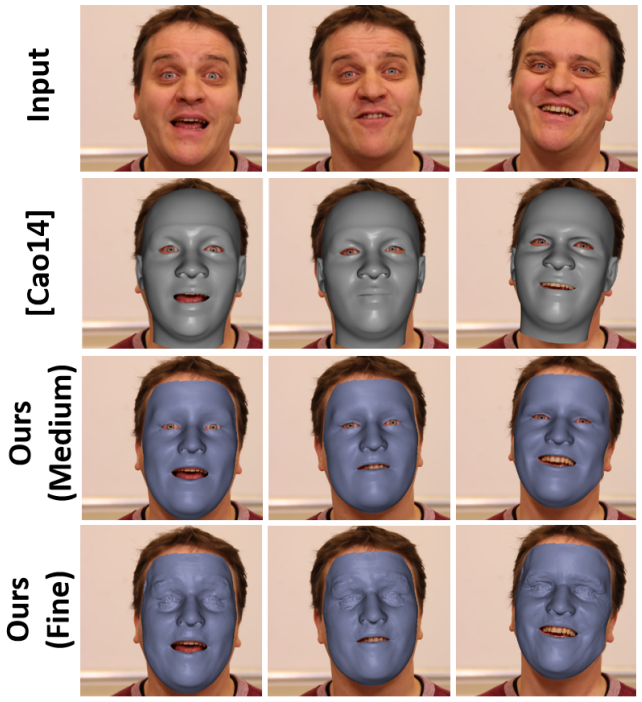

Fig. 11. State-of-the-art comparison to [Cao et al. 2014]: Monocular input (first row), result obtained by the approach of Cao *et al.* [2014] (second row), our medium-scale result (third row) and our final fine-scale reconstruction (forth row). Note that our medium-scale result matches the actor more closely and the fine-scale reconstruction adds even more realism.

second half. As ground truth for the comparison, we use the actually fitted medium and fine-scale layers by running our reconstruction pipeline on the complete dataset. Fig. 12 shows the qualitative and quantitative results. We are able to generalize well beyond the set of expressions used for training.

7.3 Comparison to Performance Capture Approaches

We compare the reconstruction part of our approach to related state-of-the-art monocular performance capture methods. Additional comparisons to monocular and multi-view methods can be found in the supplemental document. Remember that the person-specific rig building, which is an important contribution of this paper, is not performed by any approach we compare with in this section.

Comparison to [Cao et al. 2014]. The state-of-the-art monocular performance capture approach of Cao *et al.* [2014] is able to reconstruct the actor's identity and motion at a coarse-scale. While reconstructions can be obtained at video rate, they lack fine-scale surface detail and do not capture person-specific idiosyncrasies in identity and motion, see Fig. 11. In contrast, our off-line approach reconstructs person-specific medium and fine-scale surface detail and additionally learns the correlation with respect to the performed expression. Therefore, our reconstructions fit the input more closely as seen in the overlays. Our approach thus estimates a high-quality *and* intuitively controllable facial 3D rig.

Comparison to [Cao et al. 2015]. Recently, an extension to [Cao et al. 2014] that additionally regresses a wrinkle-level displacement map has been proposed [Cao et al. 2015]. This approach learns the correlation between image patches and surface detail from a database of 3D scans. While this augments the coarse-scale reconstruction with detail, the inferred geometry is not metrically correct. Thanks to the medium-scale corrective layer, our face model overlays with the input better, even if the fine-scale detail is ignored for a moment. Furthermore, our inverse rendering approach obtains detail reconstructions that match the true detail in the image closer than the regression result, which can only approximate as close as possible (see especially the shape of the eyebrows in Fig. 13). Please note that the fine-scale pores in the meshes from [Cao et al. 2015] are not reconstructed but part of the high quality template model used for learning their representation. The detail regression of Cao *et al.* [2015] is based on cues in the input image; therefore, it can not generate a detail layer for an arbitrary novel expression specified by user-defined blendshape weights. In contrast, our approach leverages the inherent semantics of the blendshape weights and allows for this scenario, which is the de facto standard for creating novel animations.

Comparison to [Garrido et al. 2013]. We are able to obtain similar or even higher quality reconstructions than those of the off-line monocular state-of-the-art facial performance capture method of Garrido *et al.* [2013], see Fig. 14. This method is able to track facial expressions including fine-scale surface detail, but it heavily relies on a static high-quality 3D scan of the actor as prior. Therefore, unlike our method, theirs is not applicable to reconstructing rigs in legacy video footage. Also, Garrido *et al.*'s approach is a pure capture method that does not learn any generative model for person-specific correctives and fine-scale details. Thus, person-specific idiosyncrasies are also better captured by our method.

Comparison to [Shi et al. 2014]. Finally, we compare to the high-quality monocular approach of Shi *et al.* [2014]. Their method employs a multi-linear face model for reconstruction and can be applied to legacy footage, see Fig. 15. We attain higher-quality reconstructions on the coarse as well as on the fine-scale due to the use of dense correspondences to jointly optimize for identity and expression. Additionally, we obtain a better model personalization and expression tracking by using medium-scale corrective shapes. On the other hand, Shi *et al.* mainly resort to sparse correspondences to estimate large-scale deformations, which are then slightly improved using normal maps estimated in their shade-from-shading framework. This leads to a less accurate head pose, as well as less accurate coarse and fine-scale surface reconstructions. Please refer to the the supplemental document for further comparisons. We remark that Shi *et al.* do not learn a correlation model for person-specific correctives and fine-scale details. Thus, their approach is unable to automatically adapt the detail layers to match person-specific idiosyncrasies, which is the foundation for realistic video editing (see Fig. 9) and reenactment tasks (see Fig. 8).

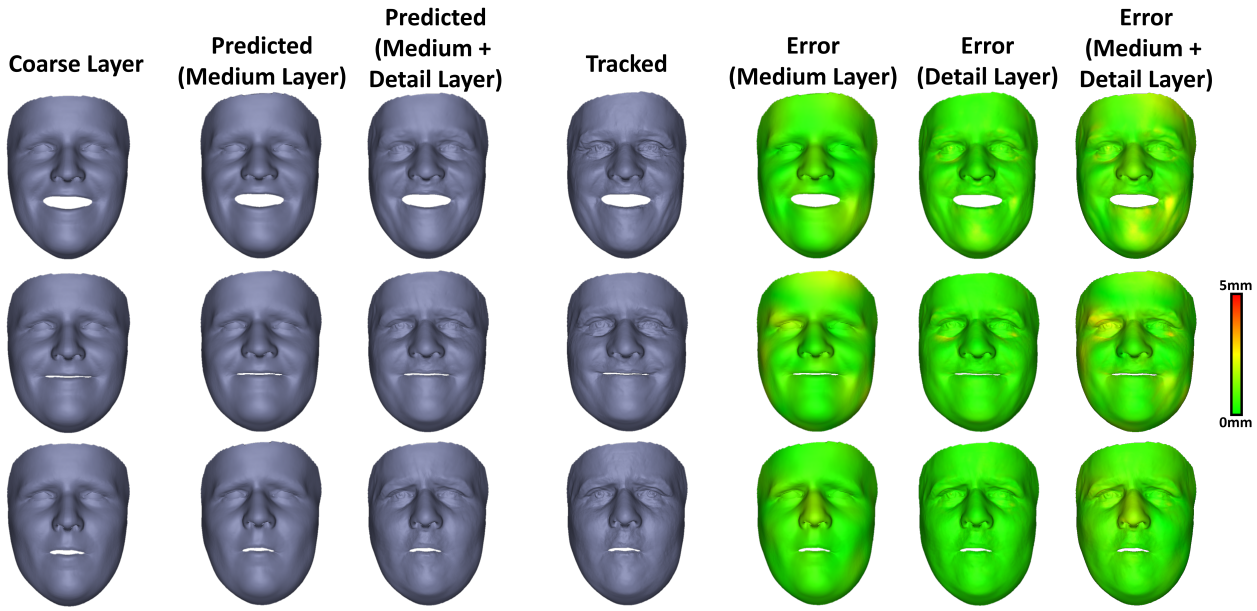

Fig. 12. Evaluation of the prediction accuracy: Our novel sparse regression strategy infers high-quality medium and fine-scale detail layers given a novel expression. Note that we compare quantitatively to the tracked ground truth reconstruction which is accurately reproduced. The prediction error of the medium and detail layer together is always smaller than 3.5mm (1mm mean and 0.16mm standard deviation). The error is mainly explained by residuals in the medium layer, while the error of the detail layer is mostly negligible ($<$ 0.4mm on average).

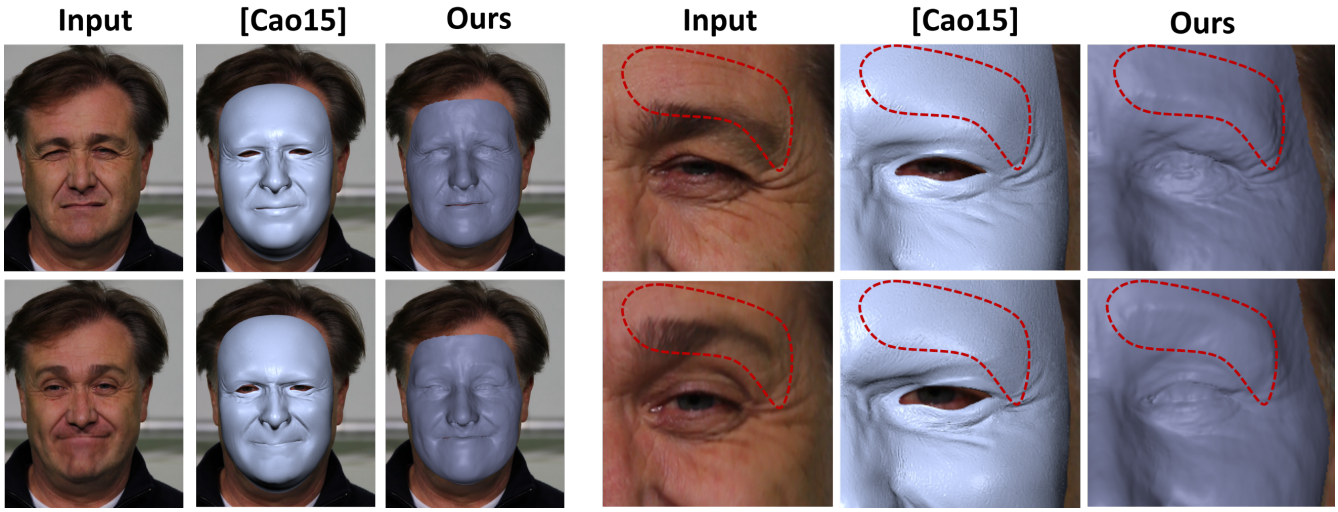

Fig. 13. State-of-the-art comparison to [Cao et al. 2015]: While the regression-based approach of Cao *et al.* [2015] infers some of the actor's fine-scale details, it produces less accurate results if poses and identities are far from the training set. In particular, note the overall less accurate reconstruction of identity (left), as well as the only approximate reconstruction of some wrinkles and the shape of the eyebrow (right). In contrast, our reconstruction-based approach delivers results closer to the real input video. Please note that fine-scale pores in the results of Cao *et al.* [2015] are merely hallucinated, as they are part of the model learned from high-quality face scans.

7.4 Comparison to Detail Prediction Methods

We compare our two-layer detail regression approach to the state-of-the-art method by Bermano *et al.* [2014] for the prediction of actor-specific idiosyncrasies and detail. Fig. 16 demonstrates that our sparse regression formulation for medium and fine-scale detail prediction achieves results of comparable quality. Note that Bermano *et al.*'s method requires a bespoke set of expressive training sequences that are captured with a multi-view camera system under controlled lighting from which the fine-scale detail and actor-specific expressiveness are extracted. In contrast, we are able to train our sparse regression technique using only a subset of frames of the monocular input footage.

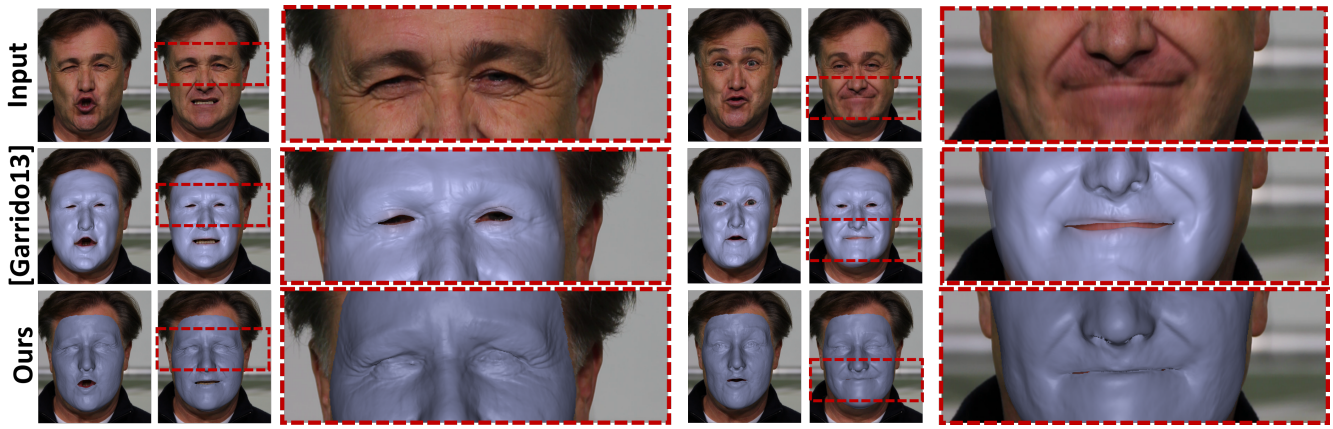

Fig. 14. State-of-the-art comparison to [Garrido et al. 2013]: Compared to the off-line approach of Garrido *et al.* [2013], our reconstructions better match the actor's static and transient small-scale surface details. Note, the method of Garrido *et al.* requires a high-quality laser scan of the actor as input, making it unsuitable for legacy video footage.

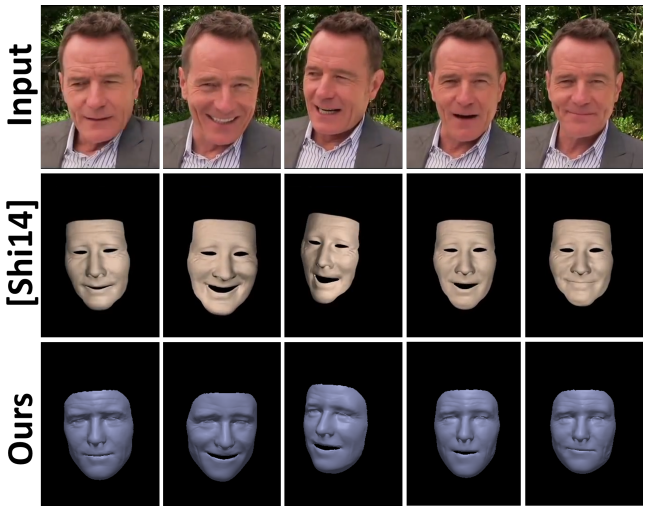

Fig. 15. State-of-the-art comparison to the approach by [Shi et al. 2014]: Our approach obtains a closer fit than Shi *et al.*'s method. Note the higher amount of fine-scale surface detail obtained by our approach.

Fig. 16. State-of-the-art comparison to the approach by [Bermano et al. 2014]: Our approach obtains predicted correctives and fine-scale detail comparable to Bermano *et al.*'s method, which requires a tailor-made set of training sequences to enhance fine-scale detail and expressiveness.

8. DISCUSSION

We presented the first approach to create a high-quality modifiable facial 3D rig of an actor from just monocular video data along with the captured facial performance. Related to our approach is the recent paper by Ichim *et al.* [2015] which aims at building a 3D face avatar from video. Our approach differs in several ways. Firstly, their approach requires a structure-from-motion 3D face reconstruction from several hundred frames of video taken around the static head to which a default model is fitted. The expression basis is then learned from a specific video sequence of facial expressions. Some of their steps also need manual intervention. In contrast, our approach only needs an RGB video of a general unscripted facial expression sequence as input and is automatic. Secondly, Ichim *et al.* do not learn medium scale correctives, but optimize the blendshapes themselves. They discuss that learning a full personalized corrective layer, as we do, would lead to better personalization.

Limitations. Our approach assumes *Lambertian* reflectance. Although this is a fairly common assumption also made in other works, it introduces artifacts in the presence of specular highlights, as shown in Fig. 17 (a). In addition, we do not model sub-surface scattering effects; the scene's light transport is parametrized using a low-dimensional SH representation which assumes smooth distant illumination and no shadows. Extreme lighting (e.g. directional spotlights) and cast shadows lead to artifacts.

Since our reconstruction approach is based on temporal frame-to-frame coherence, videos that exhibit lots of cuts are hard to handle automatically, requiring re-initialization of the parameters. Reconstructing multiple actors from a single video also requires an extra face detection and recognition component to keep the approach automatic. Mild occlusions on the face, such as hair can be handled by our approach, but may be wrongly learned as facial features, see Fig. 16. Strong occlusions, such as a dense beard, pose a problem to

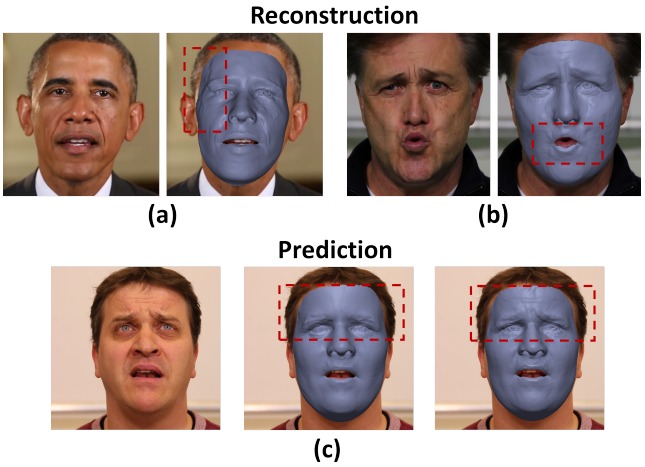

Fig. 17. Limitations: Reconstruction artifacts (top) and prediction problems (bottom). (a) Artifacts due to specular highlights on the face. (b) Artifacts due to the lack of a local-support corrective basis and constraints to handle mouth deformations. (c) Despite our reconstructions (right) accurately match the input data (left), our regressor fails to predict person-specific nuances and details (middle) when trained on a short and static face sequence.

both the 2D face tracker and the identity reconstruction (non-skin reflectance and occluding objects are not explained by our statistical prior). Our optimization relies on global-support corrective functions to correct tracking residuals and assumes all facial features contribute equally. Thus, our reconstruction approach is challenged by fast and complex local facial deformations, especially in the mouth region as shown in Fig. 17 (b). We believe that additional constraints and a semantic basis for local corrections may further improve the results.

Detail is learned based on the correlation to the corresponding expression in the input video. Thus, we require a sufficient amount of expression variation and detail revelation in the training. If only a short sequence is provided or the actors remain mostly static, we can not sufficiently explore their expression space. Fig. 17 (c) illustrates such a limitation. Learning fine-scale detail also requires robust tracking, otherwise geometric tracking drift may be learned, leading to over-smoothed regions or undesirable artifacts in the synthesis.

We share the limitation of related work that no detailed mouth interior or eye/eyelid model can be reconstructed from video alone. We therefore do not model eye geometry or blendshapes for blinking; we decided to render the rigs with a static eye albedo map, as this looks more natural than leaving eye holes. Specially scanned eye models [Bérard et al. 2014] or synthetic eyeball templates [Ichim et al. 2015] could be used for rendering. Although our models can be used for high quality animation, our rigs may still fall short of the very high detail and control level required for some professional VFX applications in movies. Even in such cases, our reconstructions are a starting point for artists to customize/refine rigs.

9. CONCLUSION

We have presented an approach for the automatic creation of a fully parametrized high-quality facial 3D rig from just monocular video data. A novel variational fitting formulation is used to capture the actor's facial identity and expression idiosyncrasies in the rig. Our rig is composed of three distinct layers that encode the actor's geometry on all scales: Starting from coarse-scale shape detail up to a layer that accounts for static and transient fine-scale detail. We explicitly learn the correlation between expression variation and the detail layers, yielding a detail prediction model. This enables an intuitive control of the rig based on a small set of control parameters familiar to artists. We demonstrated the high fidelity of our reconstructed rigs for several actors from different sources of video, including YouTube footage, and show their use in animation, expression transfer and video editing. We see our approach as a step towards automatic rig creation from monocular video, e.g. legacy footage from feature films, and hope that it will inspire further research.

ACKNOWLEDGMENTS

We thank all the reviewers for their valuable comments and TrueVisionSolutions Pty Ltd for kindly providing the 2D face tracker. We also credit NBC 5/KXAS-TV/Dallas-Fort Worth, Texas, USA for sharing the ARNOLD YOUNG sequence and Hollywood Foreign Press Association for sharing the BRYAN sequence. This work was supported by the ERC Starting Grant CapReal and by Technicolor.

REFERENCES

ALEXA, M. 2002. Linear combination of transformations. *ACM TOG 21,* 3, 380–387.

ALEXANDER, O., ROGERS, M., LAMBETH, W., CHIANG, M., AND DEBEVEC, P. 2009. The digital emily project: Photoreal facial modeling and animation. In *ACM SIGGRAPH 2009 Courses*. 12:1–12:15.

BEELER, T., BICKEL, B., BEARDSLEY, P., SUMNER, B., AND GROSS, M. 2010. High-quality single-shot capture of facial geometry. *ACM TOG 29,* 4, 40:1–40:9.

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.

BÉRARD, P., BRADLEY, D., NITTI, M., BEELER, T., AND GROSS, M. 2014. High-quality capture of eyes. *ACM TOG 33,* 6, 223:1–223:12.

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

BLANZ, V., BASSO, C., POGGIO, T., AND VETTER, T. 2003. Reanimating faces in images and video. *CGF 22,* 3, 641–650.

BLANZ, V. AND VETTER, T. 1999. A morphable model for the synthesis of 3d faces. In *Proc. SIGGRAPH '99*. 187–194.

BOUAZIZ, S., WANG, Y., AND PAULY, M. 2013. Online modeling for realtime facial animation. *ACM TOG 32,* 4, 40:1–40:10.

CAO, C., BRADLEY, D., ZHOU, K., AND BEELER, T. 2015. Real-time high-fidelity facial performance capture. *ACM TOG 34,* 4, 46:1–46:9.

CAO, C., HOU, Q., AND ZHOU, K. 2014. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM TOG 33,* 4, 43:1–43:10.

CAO, C., WENG, Y., ZHOU, S., TONG, Y., AND ZHOU, K. 2014. Facewarehouse: A 3d facial expression database for visual computing. *IEEE TVCG 20,* 3, 413–425.

DAVID, P., DEMENTHON, D., DURAISWAMI, R., AND SAMET, H. 2004. Softposit: Simultaneous pose and correspondence determination. *IJCV 59,* 3, 259–284.

FASEL, B. AND LUETTIN, J. 2003. Automatic facial expression analysis: a survey. *Pattern Recognition 36,* 1, 259–275.

FYFFE, G., JONES, A., ALEXANDER, O., ICHIKARI, R., AND DEBEVEC, P. 2014. Driving high-resolution facial scans with video performance capture. *ACM TOG 34,* 1, 8:1–8:14.

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.

GARRIDO, P., VALGAERTS, L., SARMADI, H., STEINER, I., VARANASI, K., PEREZ, P., AND THEOBALT, C. 2015. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. *CGF 34,* 2, 193–204.

GRAHAM, P., TUNWATTANAPONG, B., BUSCH, J., YU, X., JONES, A., DEBEVEC, P. E., AND GHOSH, A. 2013. Measurement-based synthesis of facial microgeometry. *CGF 32,* 2, 335–344.

HIGHAM, N. J. 1986. Computing the polar decomposition with applications. *SIAM J. Sci. Stat. Comput. 7,* 4, 1160–1174.

HOERL, A. E. AND KENNARD, R. W. 2000. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 42,* 1, 80–86.

HSIEH, P.-L., MA, C., YU, J., AND LI, H. 2015. Unconstrained realtime facial performance capture. In *Proc. CVPR*.

HUANG, H., CHAI, J., TONG, X., AND WU, H.-T. 2011. Leveraging motion capture and 3d scanning for high-fidelity facial performance acquisition. *ACM TOG 30,* 4, 74:1–74:10.

HUANG, H., YIN, K., ZHAO, L., QI, Y., YU, Y., AND TONG, X. 2012. Detail-preserving controllable deformation from sparse examples. *IEEE TVCG 18,* 8, 1215–1227.

ICHIM, A. E., BOUAZIZ, S., AND PAULY, M. 2015. Dynamic 3d avatar creation from hand-held video input. *ACM TOG 34,* 4, 45:1–45:14.

JOSHI, P., TIEN, W. C., DESBRUN, M., AND PIGHIN, F. 2003. Learning controls for blend shape based realistic facial animation. In *Proc. SCA '03*. 187–192.

KLAUDINY, M. AND HILTON, A. 2012. High-detail 3d capture and non-sequential alignment of facial performance. In *Proc. 3DIMPVT*. 17–24.

LEVENBERG, K. 1944. A method for the solution of certain non-linear problems in least squares. *Quarter. of Applied Math.* 2, 164–168.

LÉVY, B. AND ZHANG, H. R. 2010. Spectral mesh processing. In *ACM SIGGRAPH 2010 Courses*. 8:1–8:312.

LEWIS, J. P., ANJYO, K., RHEE, T., ZHANG, M., PIGHIN, F., AND DENG, Z. 2014. Practice and Theory of Blendshape Facial Models. In *Eurographics STARs*. 199–218.

LI, H., YU, J., YE, Y., AND BREGLER, C. 2013. Realtime facial animation with on-the-fly correctives. *ACM TOG 32,* 4, 42:1–42:10.

LI, J., XU, W., CHENG, Z., XU, K., AND KLEIN, R. 2015. Lightweight wrinkle synthesis for 3d facial modeling and animation. *Computer-Aided Design 58*, 117–122.

MA, W.-C., JONES, A., CHIANG, J.-Y., HAWKINS, T., FREDERIKSEN, S., PEERS, P., VUKOVIC, M., OUHYOUNG, M., AND DEBEVEC, P. 2008. Facial performance synthesis using deformation-driven polynomial displacement maps. *ACM TOG 27,* 5, 121:1–121:10.

MARQUARDT, D. W. 1963. An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. on Applied Math. 11,* 2, 431–441.

MORÉ, J. 1978. The Levenberg-Marquardt algorithm: Implementation and theory. In *Numerical Analysis*. Lecture Notes in Math., vol. 630. 105–116.

MÜLLER, C. 1966. *Spherical harmonics*. Lecture Notes in Math., vol. 17.

NEHAB, D., RUSINKIEWICZ, S., DAVIS, J., AND RAMAMOORTHI, R. 2005. Efficiently combining positions and normals for precise 3d geometry. *ACM TOG 24,* 3, 536–543.

NEUMANN, T., VARANASI, K., WENGER, S., WACKER, M., MAGNOR, M., AND THEOBALT, C. 2013. Sparse localized deformation components. *ACM TOG 32,* 6, 179:1–179:10.

POPA, T., SOUTH-DICKINSON, I., BRADLEY, D., SHEFFER, A., AND HEIDRICH, W. 2010. Globally consistent space-time reconstruction. *CGF 29,* 5, 1633–1642.

RAMAMOORTHI, R. AND HANRAHAN, P. 2001. A signal-processing framework for inverse rendering. In *Proc. SIGGRAPH '01*. 117–128.

SARAGIH, J. M., LUCEY, S., AND COHN, J. F. 2011. Real-time avatar animation from a single image. In *Proc. FG*. 213–220.

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.

SIFAKIS, E., NEVEROV, I., AND FEDKIW, R. 2005. Automatic determination of facial muscle activations from sparse motion capture marker data. *ACM TOG 24,* 3, 417–425.

SUMNER, R. W. AND POPOVIC, J. 2004. Deformation transfer for triangle meshes. *ACM TOG 23,* 3, 399–405.

SUWAJANAKORN, S., KEMELMACHER-SHLIZERMAN, I., AND SEITZ, S. M. 2014. Total moving face reconstruction. In *Proc. ECCV*. 796–812.

TENA, J. R., DE LA TORRE, F., AND MATTHEWS, I. 2011. Interactive region-based linear 3d face models. *ACM TOG 30,* 4, 76:1–76:10.

THIES, J., ZOLLHÖFER, M., NIESSNER, M., VALGAERTS, L., STAMMINGER, M., AND THEOBALT, C. 2015. Real-time expression transfer for facial reenactment. *ACM TOG 34,* 6, 183:1–183:14.

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.

VALLET, B. AND LÉVY, B. 2008. Spectral geometry processing with manifold harmonics. *CGF 27,* 2, 251–260.

VLASIC, D., BRAND, M., PFISTER, H., AND POPOVIĆ, J. 2005. Face transfer with multilinear models. *ACM TOG 24,* 3, 426–433.

WAND, M., ADAMS, B., OVSJANIKOV, M., BERNER, A., BOKELOH, M., JENKE, P., GUIBAS, L., SEIDEL, H.-P., AND SCHILLING, A. 2009. Efficient reconstruction of nonrigid shape and motion from real-time 3d scanner data. *ACM TOG 28,* 2, 15:1–15:15.

WEISE, T., BOUAZIZ, S., LI, H., AND PAULY, M. 2011. Realtime performance-based facial animation. *ACM TOG 30,* 4, 77:1–77:10.

WENGER, A., GARDNER, A., TCHOU, C., UNGER, J., HAWKINS, T., AND DEBEVEC, P. 2005. Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM TOG 24,* 3, 756–764.

WU, C., ZOLLHÖFER, M., NIESSNER, M., STAMMINGER, M., IZADI, S., AND THEOBALT, C. 2014. Real-time shading-based refinement for consumer depth cameras. *ACM TOG 33,* 6, 200:1–200:10.

ZOLLHÖFER, M., NIESSNER, M., IZADI, S., REHMANN, C., ZACH, C., FISHER, M., WU, C., FITZGIBBON, A., LOOP, C., THEOBALT, C., AND STAMMINGER, M. 2014. Real-time non-rigid reconstruction using an rgb-d camera. *ACM TOG 33,* 4, 156:1–156:12.

ZOLLHÖFER, M., THIES, J., COLAIANNI, M., STAMMINGER, M., AND GREINER, G. 2014. Interactive model-based reconstruction of the human head using an RGB-D sensor. *Journal of Vis. and Comput. Anim. 25,* 3-4, 213–222.

Received September 2015; accepted November 2015