An Anatomically-Constrained Local Deformation Model for Monocular Face Capture

Chenglei Wu[2] Derek Bradley[1] Markus Gross[1,2] Thabo Beeler[1]
1) Disney Research 2) ETH Zurich

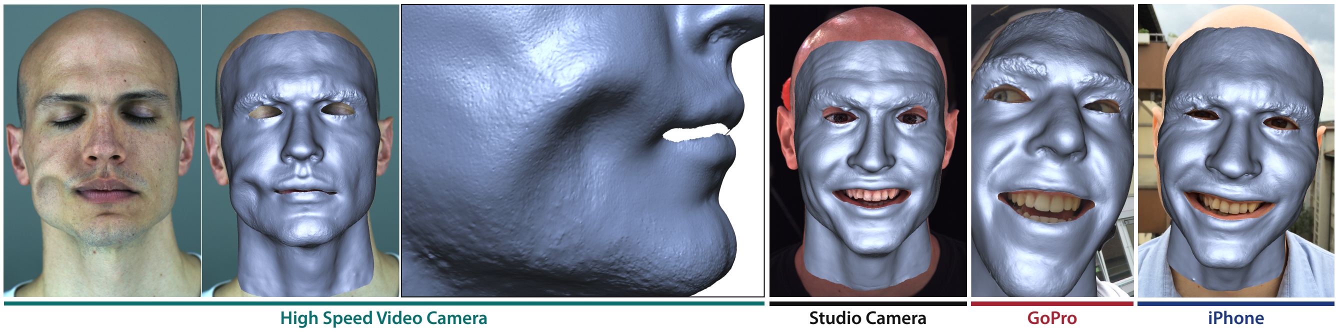

Figure 1: *Our anatomically-constrained local face model allows for high-quality and very expressive monocular performance capture from a large variety of input sources, including high-speed capture of extreme deformations caused by external forces such as wind, high-quality studio setups, or unconstrained outdoor acquisition using helmet-mounted GoPro and handheld iPhone devices.*

Abstract

We present a new anatomically-constrained local face model and fitting approach for tracking 3D faces from 2D motion data in very high quality. In contrast to traditional global face models, often built from a large set of blendshapes, we propose a local deformation model composed of many small subspaces spatially distributed over the face. Our local model offers far more flexibility and expressiveness than global blendshape models, even with a much smaller model size. This flexibility would typically come at the cost of reduced robustness, in particular during the under-constrained task of monocular reconstruction. However, a key contribution of this work is that we consider the face anatomy and introduce subspace skin thickness constraints into our model, which constrain the face to only valid expressions and helps counteract depth ambiguities in monocular tracking. Given our new model, we present a novel fitting optimization that allows 3D facial performance reconstruction from a single view at extremely high quality, far beyond previous fitting approaches. Our model is flexible, and can be applied also when only sparse motion data is available, for example with marker-based motion capture or even face posing from artistic sketches. Furthermore, by incorporating anatomical constraints we can automatically estimate the rigid motion of the skull, obtaining a rigid stabilization of the performance for free. We demonstrate our model and single-view fitting method on a number of examples, including, for the first time, extreme local skin deformation caused by external forces such as wind, captured from a single high-speed camera.

Keywords: Monocular face tracking, local face model, anatomical constraints, facial performance capture.

Concepts: •**Computing methodologies** → **Motion capture;**

SIGGRAPH '16 Technical Paper, July 24 - 28, 2016, Anaheim, CA,
ISBN: 978-1-4503-4279-7/16/07
DOI: http://dx.doi.org/10.1145/2897824.2925882

1 Introduction

Facial performance capture is key for modern visual effects in feature films and computer games. Due to its importance, this area has attracted a lot of attention from the research community and evolved rapidly over the past decades. On the one hand researchers strive to acquire the face in ever more detail to provide higher quality face shapes and dynamics. On the other hand, one can observe a clear trend towards less constrained acquisition setups, which require less hardware and give the actor more freedom to perform. The most convenient input device would certainly be just a single camera. Unfortunately, reconstruction from a single camera is ill-posed and not possible without further assumptions. Therefore, monocular performance capture methods typically rely on prior knowledge of the face, oftentimes in the form of a global blendshape rig. This approach is also considered the industry standard for facial performance capture, where an actor is first captured in a constrained setup to produce a highly accurate rig, and then the desired performance is acquired with less-constrained input devices, such as marker based helmet cameras, which drives the rig to obtain the final animation.

Global blendshape rigs have a strong tendency to over-constrain the problem, since any new face shape must lie within the space spanned by the blendshapes. Consequently, a large number of expressions must be acquired, processed and encoded into the rig in order to faithfully capture an actor's performance. A production-quality rig easily contains in the order of a hundred carefully picked expressions, requiring substantial time of both the actor and the artists creating the rig. But even a production-quality rig is unlikely to encode all shape variations of the actor's face. Take, for example, shapes caused by external forces and secondary motion, which are not typically observable in a constrained acquisition setup but are very present during under-constrained acquisition later on set. The consequence is that the performance reconstruction will not hit every expression accurately and even shift some of the error into the head pose estimation, leading to unstabilized results. To alleviate the problem, it is common practice in industry to separate head pose and expression fitting by first estimating the head pose in an often manually assisted stabilization pass before solving for the expression using the rig. Manual stabilization is a very tedious and time consuming process, and even if solved correctly, global blendshapes are typically not able to fit the expressions accurately.

Local blendshape rigs add flexibility by activating blendshapes only in pre-defined regions of the face. While these do allow to express global shapes outside of the pre-captured manifold, they are still constrained locally to linear combinations of the captured vertex positions. The increased flexibility of local blendshape models comes at the price of reduced robustness. For example, when a skin patch appears bigger on screen, this can either be due to local stretch or because the patch moved towards the camera, or a combination of both. These ambiguities have so far prevented the use of highly localized blendshapes for monocular or helmet-camera performance capture. In this paper we propose a new local subspace model that explicitly encodes local deformation rather than position to allow even greater flexibility, truly outside the positional subspace of the captured shapes. Furthermore, we additionally increase robustness over traditional localized blendshapes by using anatomical constraints. In particular we leverage the fact that the underlying bone structures move purely rigidly and that tissue thickness is directly related to local surface structure as shown by Beeler and Bradley [2014]. By globally linking local subspaces via the underlying bone, we devise a robust local face model that maintains flexibility and expressiveness. Our model is built from a minimal set of facial scans, inspired by Huang et al. [2011] who capture a minimal set of faces for performance-specific global blendshape tracking, we develop an iterative shape ranking strategy to incrementally build the local subspace of our new model. Our method can be used to automatically reconstruct both the face surface and the underlying skull from just a single view, obtaining rigidly stabilized facial performances.

The main contributions of this paper are:

1. A new local subspace model for facial performance capture that encodes local deformation rather than local shape, and is bounded by anatomical constraints, making it robust to typical ambiguities that occur with local models.
2. A method to reconstruct facial performances in very high quality from a single view, with automatic rigid stabilization. The method can incorporate both dense constraints provided, for example, from optical flow and/or sparse constraints provided, for example, by marker tracks or artist sketches.
3. An importance ranking of typically-acquired face shapes for rig creation. This informs the minimum number of shapes required for high-quality facial performance capture, and we show that by picking the right shapes our method requires significantly less expressions to be pre-acquired than traditional global blendshape tracking.

We demonstrate the performance of the proposed algorithm on three different actors over a large variety of monocular input data, including dense optical flow from high-quality cameras, outdoor footage from smart phones and helmet-mounted GoPro cameras, as well as sparse marker tracks from MoCap data, and even artist-created input sketches. We quantitatively assess the improvement of the proposed anatomically-constrained local deformation tracking over traditional global blendshape tracking as well as over our local deformation tracking without anatomical constraints. Finally, to really highlight the flexibility of our approach we capture, for the first time, extreme skin deformations that occur from external forces (such as blowing wind) and secondary motion.

2 Related Work

Our work falls into the category of facial reconstruction from images and video, so we highlight related methods in both multi-view and monocular 3D face capture, and discuss other methods that create parametric models of faces for the application of face tracking. We end by positioning our work among others who also perform rigid stabilization of facial performances.

Multi-view Face Capture. Facial scanning and performance capture is one of the long-standing topics of research in computer graphics, driven by the high demand for realistic digital humans. In the last decade we have witnessed tremendous advances in high quality 3D static scanning [Beeler et al. 2010; Ghosh et al. 2011] and dynamic performance capture [Zhang et al. 2004; Bradley et al. 2010; Beeler et al. 2011; Huang et al. 2011; Valgaerts et al. 2012]. These methods leverage computer vision concepts like stereo reconstruction to acquire the facial geometry from multiple views.

Monocular Face Capture. A more recent trend is to reconstruct highly detailed and dense facial performances from a single view, easing the hardware burden of multi-view capture methods. Our proposed technique falls into this category. Garrido et al. [2013] use sparse feature tracking to drive a global blendshape model and account for out-of-model deformation using optical flow and photometric stereo. Shi et al. [2014] use multi-linear face models, also combined with sparse feature tracking and fine-scale shape from shading cues. Suwajanakorn et al. [2014] build a person-specific face model from a large photo collection, and then fit the model to video frames using 3D flow estimation and shape from shading. In contrast, our method does not require a large data collection, and rather than a global model we use a *local* model to allow more expressiveness with a smaller model size. We also incorporate anatomical constraints in order to resolve local model ambiguity, which also allows us to compute the rigid stabilization of the head, a feature that is not addressed by any previous monocular face capture method.

Fyffe et al. [2014] reconstruct video performances using dense optical flow, as well as image correspondences computed between a set of static reconstruction poses and the video, and then reconstruct the performance by fitting a template mesh to all frames. Fitting is performed by solving a large optimization problem over all frames and all meshes using a shape preservation constraint. In contrast, our local deformation model explicitly encodes the range of deformation of local face patches and can be used to more faithfully constrain face reconstruction in under-constrained fitting scenarios. This prior allows our method to operate even on very sparse input, such as markers or hand-drawn sketches, unlike the method of Fyffe et al., which is inherently tied to appearance and requires dense input. Such versatility has not been demonstrated with any previous approach, and is only possible due to our new local anatomical face model.

Another common trend has been real-time facial performance capture and retargeting. These methods use either a single-view depth sensor [Weise et al. 2009; Weise et al. 2011; Li et al. 2013; Bouaziz et al. 2013; Chen et al. 2013] or a web camera [Rhee et al. 2011; Cao et al. 2013; Cao et al. 2014; Cao et al. 2015] to track the face in 3D. However, to achieve real-time performance, these methods typically use a global face model as a prior, and thus cannot achieve the same level of expressiveness and fidelity as our local model.

Parametric Face Models. Parameterizing the face as a 2D or 3D face model is a common way to overcome the ambiguities associated with monocular face tracking [Li et al. 1993; Black and Yacoob 1995; DeCarlo and Metaxas 1996; Essa et al. 1996; Saragih et al. 2011]. Some common models include Active Appearance Models (AAM) [Cootes et al. 2001], blendshapes [Lewis et al. 2014], principle components analysis (PCA) on a set of training shapes [Lau et al. 2009], morphable models [Blanz and Vetter 1999], and multilinear models [Vlasic et al. 2005]. These models are used extensively throughout the monocular face capture methods mentioned above. The main drawback of these models is that they are designed to be global, meaning that the entire face is parameterized holistically, which limits local expressiveness unless the model is very large.

Local or region-based shape models have also been proposed, which offer more flexibility at the cost of being less constrained to realistic human face shapes. Joshi et al. [2003] use a region-based blendshape model for keyframe facial animation, and automatically determine the best segmentation using a physical model. Na and Jung [2011] use local blendshapes for motion capture retargeting, and they devise a method for choosing the local regions and their corresponding weighting factors automatically. Tena et al. [2011] learn a region-based PCA model based on motion capture data, allowing direct local manipulation of the face. Neumann et al. [2013] extract sparse localized deformation components from an animated mesh sequence, also with the goal of intuitive editing as well as statistical processing of the face. Brunton et al. [2014] use many localized multilinear models to reconstruct faces from noisy or occluded point cloud data. Our key contribution is a new local 3D face model that parameterizes the face into many overlapping patches and explicitly encodes the local deformation of each patch rather than local positions. In contrast to traditional region-based blendshape models that encode both local and global deformation for each region, we explicitly formulate many local subspaces that encode only the local deformation and we handle global motion through a tracking optimization step. In order to make this tractable and robust, we incorporate anatomical constraints in the form of a skull and jaw bone, and also parameterize the local skin thickness in the subspace. By decoupling rigid motion and non-rigid deformation, our model exceeds the flexibility and expressiveness of previous local blendshape models, yet still the model exhibits superior robustness thanks to the anatomical constraints, allowing monocular face reconstruction and single-view direct editing with unprecedented fidelity, even during extreme local deformations and without falling victim to depth ambiguities.

Rigid Stabilization of Faces. Our approach to monocular face capture uses the underlying bone structure to anatomically constrain the local skin thickness. As a result, we can simultaneously solve for the skin surface and the skull position for every video frame, yielding a rigidly stabilized performance. Rigid stabilization of faces is not a topic of vast research to date. Most similar to our approach is the method of Beeler and Bradley [2014], who use similar anatomical constraints learned from CT scans to stabilize a set of tracked meshes in correspondence. In this work, we use the technique of Beeler and Bradley to learn our local anatomical subspace, and then we can perform stabilization during reconstruction from just a single view.

3 Overview

We will begin by defining our new local face model, which consists of a local patch deformation subspace and an underlying anatomical bone structure (Section 4). The face model can then be used for motion reconstruction, given an initial face mesh and either sparse or dense 2D motion data, for example in the scenario of monocular video tracking. Motion reconstruction is performed in two steps, first by tracking the local patches and bones using the anatomical constraints (Section 5) and then combining the patches into a global face mesh (Section 6). To highlight the characteristics of the local face model we analyze the required patch subspace and rank the importance of typically-acquired face shapes for model building (Section 7). Finally, we show various results of our model and tracking approach for monocular face capture from different camera inputs, as well as additional application scenarios including sparse motion capture data and direct face manipulation from artist sketches (Section 8). An overview of the proposed pipeline is given in Fig. 2.

4 Local Face Model

Our new face model consists of two components: a local deformation subspace and an underlying anatomical bone structure (see Fig. 3). For the local deformation subspace, we represent the face geometry with overlapping patches. The local deformation of each patch is constrained by a local deformation subspace learned from a set of training shapes and the global motion of each patch is defined by a rigid transformation. Specifically, we define for each patch i the following parameters: the rigid transformation denoted as a matrix M_i, and the local deformation subspace coefficients $\boldsymbol{\alpha}_i$. The anatomical component is modeled as an actor-specific skull and jaw bone. A generic skull is fit to the actor using the method of Beeler and Bradley [2014], and the jaw is modeled by an artist. The skull motion is defined by a rigid transformation matrix M_s and the jaw motion is linked to the skull via a pivot point, $\boldsymbol{o}$, represented as a joint with two degrees of freedom for rotation and one for translation, denoted as the jaw motion parameter $\boldsymbol{\Theta} = \{\theta_X, \theta_Y, t_Z\}$. Later on it will be convenient to refer to the global rigid motion M_j of the jaw explicitly, which can be computed as $M_s \cdot M(\boldsymbol{\Theta})$, where $M(\cdot)$ computes the transformation matrix corresponding to $\boldsymbol{\Theta}$.

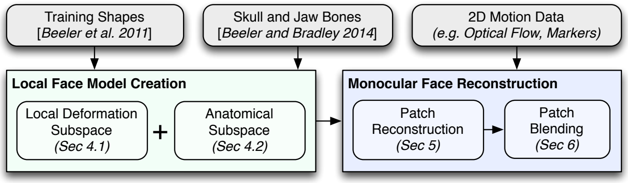

Figure 2: ***Method Overview*** *– The proposed pipeline consists of two main steps. First a local anatomical deformation model is created from K+1 facial shapes and bone structures. This model is then fit to 2D trajectories, such as optical flow or marker tracks, to reconstruct a highly accurate 3D model of the face from a single view.*

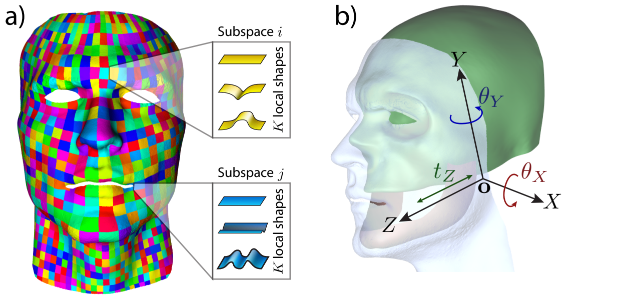

Figure 3: *Our face model consists of (a) a local deformation subspace, and (b) an underlying anatomical bone structure consisting of a skull and jaw linked by a pivot point* $\boldsymbol{o}$*. The jaw moves relative to the skull and has two rotational (θ_X, θ_Y) and one translational degree of freedom (t_Z).*

4.1 Local Deformation Subspace

Before defining our local deformation subspace, we need to first segment the face into patches. Ideally, this patch segmentation would be semantically meaningful, exploiting the physical properties and motion of the skin, e.g. following flow lines. However, achieving such a segmentation is very challenging (in fact a topic of research in itself [Joshi et al. 2003; Na and Jung 2011]), especially considering that these properties are usually person specific. In view of this and also for generality, we define our patches via a uniform segmentation in the UV space of the face. Note, however, that any other patch segmentation can be used here. As the deformation of each patch is also influenced by its neighboring patches, we require the segmented patches to overlap with their neighbors. This can be easily achieved by first performing a non-overlapping segmentation (Fig. 3.a) and then dilating each patch by a fixed amount (20% in each direction for all our datasets). To account for holes or concave boundaries in the

UV plane, we explicitly split patches if they contain disconnected regions, ensuring that each patch is a single connected region. The number of patches to create, N, (or equivalently the patch size) is a user-defined parameter. An evaluation of different patch sizes will be given in Section 7.

To obtain the local skin deformation subspace, we need to capture the actor-specific deformation for each patch in correspondence. To accomplish this, we capture a neutral scan and a set of K extreme expressions using an off-the-shelf performance capture method [Beeler et al. 2011]. From these tracked face reconstructions we can extract the local deformation subspace. Specifically, given several shapes in correspondence, we segment the neutral mesh into N patches and then build a $K+1$ subspace for each patch by first aligning the K patch shapes to the neutral patch shape using Procrustes alignment [Gower 1975] and then subtracting the neutral patch to obtain a deformation component for each expression. The resulting subspace for a patch i consists of the neutral shape U_i and K deformation components $\{D_i^1, ..., D_i^K\}$. Therefore, the deformed patch shape at time t can be computed as

$$X_i(t) = M_i(t) \left(U_i + \sum_{k=1}^{K} \alpha_i^k(t) D_i^k \right), \tag{1}$$

where $M_i(t)$ is the rigid motion of the patch and $\{\alpha_i^1(t), ..., \alpha_i^K(t)\}$ are the coefficients of the deformation components. An interesting parameter is the number and description of the K expressions required to build an expressive subspace. In Section 7 we will show that our local model requires far fewer training expressions than typical global models, and we will rank the importance of different expressions to achieve high-quality reconstructions using our model.

The proposed local deformation model differs fundamentally from traditional position-based local blendshape models (e.g. [Tena et al. 2011]), where positions are constrained to be a linear combination of the positions of the individual blendshapes. Our model explicitly decouples local rigid motion from non-rigid local deformation, which enables greater flexibility and accuracy since vertices are allowed to move outside the subspace spanned by the positions of the shapes, for example on a curve instead of a straight line. Also, contrary to local blendshape models where a particular local shape can only occur at a certain position, our model can express the non-linearity of local shape since it decouples motion from deformation. Such non-linearity is caused, for example, by some wrinkles where the skin starts folding only after a certain amount of strain has built up through local motion. Still, globally our model does respect physically plausible face shapes as it links motion and deformation, as described in Section 5.

4.2 Anatomical Subspace

Exploiting the local deformation of a skin patch is not only physically more intuitive than a holistic approach but also practically more expressive in terms of deformations that can be represented. However, similar to other local models [Joshi et al. 2003; Na and Jung 2011; Tena et al. 2011; Brunton et al. 2014], fitting our model as described so far to real world data can be more ill-posed than a global model due to the larger set of parameters to estimate, and thus could be more vulnerable to noise and outliers and suffer from depth ambiguities. With this in mind, and also considering the specific anatomical structure of a human face, we propose to employ the anatomical skull and jaw bone to constrain the patches globally, such that the deformation of the patches are jointly-constrained to be physically feasible. As a reminder, the anatomical structure we use is shown in Fig. 3, and its motion is described by the rigid motion of the underlying bones, which we write as M_b to mean either M_s when referring to the skull or M_j when referring to the jaw.

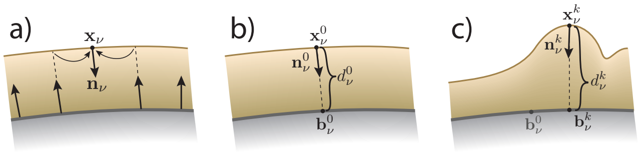

Figure 4: *Computing the anatomical subspace. A point $\boldsymbol{x}_\nu$ at the center of a patch is related to the underlying bone through a skin thickness constraint d_ν^k for each shape of the subspace. (a) The direction $\boldsymbol{n}_\nu$ in which to measure the thickness is computed by interpolating back-projected normals from bone to skin. (b) Intersecting a ray from $\boldsymbol{x}_\nu^0$ in the direction of $\boldsymbol{n}_\nu^0$ intersects the bone at $\boldsymbol{b}_\nu^0$ to give d_ν^0. (c) The same procedure is repeated for each subspace shape.*

To employ our anatomical structure to constrain the patch motion, we must establish a link between the skin surface and anatomical bones. Inspired by Beeler and Bradley [2014], we link these two parts by modeling the behavior of the soft tissue in-between them. Specifically, as skin compresses Beeler and Bradley postulate it will bulge out to preserve its volume, increasing the distance between the skin surface and the bone, and vice-versa for skin stretch. We incorporate this idea into our subspace model for patches, with the goal of predicting how a patch moves relative to the bone given its current local deformation. To accomplish this, we expand our local deformation subspace to include also the skin thickness for each subspace shape of each patch. We express the thickness of the skin tissue within the patch as a single value d_ν at a vertex ν close to the center of the patch. Selecting an actual vertex as reference point instead of the patch centroid will prove advantageous during optimization in Section 5 since the position $\boldsymbol{x}_\nu$ of the vertex is guaranteed to lie on the surface. Computing the skin thickness for the subspace is not trivial, because as a patch deforms it typically slides over the bone, and so we must account for shape-specific projections from the patch to the bone in order to compute the distance. To complicate matters, when the patch deforms, the normal at ν typically changes and is thus not a temporally stable direction to compute the distance along. Since the skull and jaw are relatively smooth and consistently rigid, a more stable approach is to use the inverse of the bone normal to compute the projection, however this introduces a chicken-and-egg problem since we do not know the bone point that corresponds to ν before projecting. Finding the corresponding bone point can be accomplished by computing Phong displacements, a well-established method in multi-resolution shape deformation [Kobbelt et al. 1999], however this approach requires several steps of Newton iterations. In our case the bone geometry is smooth enough to simply project backwards from all the vertices of the bone to the patch and interpolate the inverse bone normals at ν (Fig. 4.a). The interpolated normal $\boldsymbol{n}_\nu$ provides the direction to cast a ray and intersect with the bone (Fig. 4.b), at a point we call $\boldsymbol{b}_\nu$, yielding the skin thickness $d_\nu = ||\boldsymbol{b}_\nu - \boldsymbol{x}_\nu||$. This process is repeated for all shapes in the deformation subspace to compute skin thicknesses d_ν^k for the patch (Fig. 4.c).

To compute an estimate $\tilde{\boldsymbol{x}}_\nu$ of the vertex position later on, we additionally need to store bone points $\boldsymbol{b}_\nu^k$ and normal directions $\boldsymbol{n}_\nu^k$ for each shape k in the subspace. These quantities are represented in the coordinate frame M_b^k of the underlying bone, which removes any rigid motion and renders them compatible. Note that some patches, such as on the cheek, do not have an underlying bone and are thus not anatomically constrained.

Finally, for any time t, the position $\tilde{\boldsymbol{x}}_\nu(t)$ of vertex ν can be predicted as

$$\tilde{\boldsymbol{x}}_\nu(t) = M_b(t)\left(\tilde{\boldsymbol{b}}_\nu(t) - d_\nu(t)\tilde{\boldsymbol{n}}_\nu(t)\right), \tag{2}$$

where $\tilde{\boldsymbol{b}}_\nu(t), \tilde{\boldsymbol{n}}_\nu(t)$, and $d_\nu(t)$ are computed as

$$\tilde{\boldsymbol{b}}_\nu(t) = \boldsymbol{b}_\nu^0 + \sum_{k=1}^{K} \alpha_i^k(t)(\boldsymbol{b}_\nu^k - \boldsymbol{b}_\nu^0), \tag{3}$$

$$\tilde{\boldsymbol{n}}_\nu(t) \cong \boldsymbol{n}_\nu^0 + \sum_{k=1}^{K} \alpha_i^k(t)(\boldsymbol{n}_\nu^k - \boldsymbol{n}_\nu^0), \tag{4}$$

$$d_\nu(t) = d_\nu^0 + \sum_{k=1}^{K} \alpha_i^k(t)(d_\nu^k - d_\nu^0). \tag{5}$$

Contrary to the formulation proposed by Beeler and Bradley [2014], the estimated skull point $\tilde{\boldsymbol{b}}_\nu(t)$ and skull normal $\tilde{\boldsymbol{n}}_\nu(t)$ are only approximations in our case. We found this approximation to be reasonable since the underlying skull varies only smoothly in-between the samples such that the introduced inaccuracy is negligible, especially considering the fact that the underlying skull is an estimation in itself. The benefit of using this approximate formulation is that the problem can be cast as a system of linear equations, which can be solved uniquely and efficiently as elaborated in Section 5, where the method of Beeler and Bradley [2014] is non-linear and non-continuous due to the piecewise planar tessellation of the skull, and therefore much harder to solve and less stable due to local minima.

The combination of local deformation plus anatomical subspace completes our local face model. We next describe how our model can be fit to real data, in particular 2D motion data from a monocular view.

5 Local Patch Reconstruction

We now describe a new algorithm for 3D face tracking using our local face model, which is particularly designed for monocular facial performance capture or other applications where only 2D motion prediction is available. As with most other model-based reconstruction techniques, the goal is to estimate the model parameters that best describe the observed motion under the given constraints through optimization. In our case, the unknowns to solve for are: a) the rigid local patch motion $\{M_i\}$; b) the local patch deformation, namely the local blend coefficients $\{\boldsymbol{\alpha}_i\}$; and c) the rigid motion of the anatomical bones, including skull M_s and jaw motion $\boldsymbol{\Theta}$. We formulate the solution as an energy minimization problem for each frame t

$$\underset{\{M_i\},\{\boldsymbol{\alpha}_i\},M_s,\boldsymbol{\Theta}}{\text{minimize}} \quad E(t), \tag{6}$$

where our energy contains several terms, defined as

$$E(t) = E_M(t) + E_O(t) + E_A(t) + E_T(t). \tag{7}$$

E_M is the *2D motion* term, essentially our main data term that considers the input 2D motion vectors, e.g. from optical flow. We call E_O the *overlap* constraint, which is a spatial regularization term to enforce neighboring patches to agree with each other wherever they have shared vertices. E_A is the *anatomical* constraint, ensuring that patches remain plausibly connected with the bone structure. Finally E_T is a *temporal* regularizer. In the following we explain each term in detail. Note that we solve for all patches in a coupled way, however the result is still a set of disjoint patches that must be combined into a single global face mesh, as described in Section 6.

5.1 2D Motion Constraint

Monocular facial performance capture is an ill-posed problem due to the fact that the depth information is missing. In order to estimate a 3D face out of 2D input, some form of prior is typically needed, e.g. a blendshape subspace. Here we make use of our local deformation subspace to constrain the deformation of the patches while attempting to match the projected 2D motion as closely as possible. Given a face mesh observed from a single view, let $V(t)$ be the set of visible vertices and $\boldsymbol{p}_\nu(t)$ be the predicted 2D pixel location corresponding to vertex $\nu \in V$ at time t, and let Q be the calibrated projection matrix, then the motion energy term is defined as

$$E_M(t) = \lambda_M \sum_{\nu \in V(t)} \sum_{i \in \Omega(\nu)} \psi\left(\|Q(\boldsymbol{x}_{\nu,i}(t)) - \boldsymbol{p}_\nu(t)\|\right), \tag{8}$$

where $\boldsymbol{x}_{\nu,i}(t)$ is the unknown 3D position of vertex ν in patch i expressed in the form of Eq. 1 via the unknown blend coefficients $\boldsymbol{\alpha}_i$ and the unknown rigid transformation M_i, and $\Omega(\nu)$ is the set of patches which contain vertex ν. λ_M is a weighting factor for this term, and $\psi(\cdot)$ is a robust kernel typically used to reduce the impact of outliers [Zollhöfer et al. 2014], which takes the form

$$\psi(e) = \min_w \left(w^2 e^2 + 2\left(1 - w^2\right)^2\right). \tag{9}$$

Note that the set of motion constrained vertices V depends on the type of input and might range from very dense for flow based performance capture to very sparse in the case of marker based MoCap.

5.2 Overlapping Constraint

The motion constraint above is applied independently for each patch, which means that overlapping patch boundaries may not agree. In the end we wish to construct a single global mesh surface, which will be easier if neighboring patches deform consistently. To accomplish this we introduce an overlap constraint, which operates like a spatial regularizer. As we explicitly generate a group of shared vertices in the patch segmentation step, we can define the overlapping constraint for these vertices, as follows

$$E_O(t) = \lambda_O \sum_{\nu \in S} \sum_{(i,j) \in \Omega(\nu), i > j} \|\boldsymbol{x}_{\nu,i}(t) - \boldsymbol{x}_{\nu,j}(t)\|^2, \tag{10}$$

where S is the set of vertices shared by patches, and λ_O is a weighting factor.

5.3 Anatomical Constraint

With motion and overlapping constraints defined above, facial surface tracking would already be possible, however, as mentioned earlier, this more expressive local model comes at the cost of lower robustness. Therefore, here we introduce our new anatomical constraint into the energy. The anatomical constraint contains two terms, one for constraining the patches given the bone structure, and one for predicting the rigid bone motion given the 2D motion data. The first term constrains patches using the sparse predicted point constraints $\tilde{\boldsymbol{x}}_\nu(t)$ computed from the anatomical subspace in Eq. 2, and is written as

$$E_{A1}(t) = \lambda_{A1} \sum_{\nu \in A} \sum_{i \in \Omega(\nu)} \omega_\nu \|\boldsymbol{x}_{\nu,i}(t) - \tilde{\boldsymbol{x}}_\nu(t)\|^2, \tag{11}$$

where A is the set of vertices that contain anatomical constraints, and ω_ν is a weighting factor, as computed by Beeler and Bradley [2014]. With this term alone, the rigid motion of the anatomical bones could be obtained, as the predicted surface point is also indirectly constrained by the flow constraint. In practice, however, we found that more stable bone tracking can be achieved by imposing the flow constraint directly on the predicted surface point in a second term, written as

$$E_{A2}(t) = \lambda_{A2} \sum_{\nu \in A} \omega_\nu \psi \left(\| Q(\tilde{\boldsymbol{x}}_\nu(t)) - \boldsymbol{p}_\nu(t) \| \right), \tag{12}$$

where $\psi(\cdot)$ is again the robust kernel from Eq. 9. The final energy for the anatomical constraint is then

$$E_A(t) = E_{A1}(t) + E_{A2}(t). \tag{13}$$

Adding this anatomical constraint significantly improves the depth reconstruction of the face, which we will show in Section 8. As a by-product, the anatomical bone tracking result can also be used to automatically estimate a rigid stabilization of the face sequence, as proposed by Beeler and Bradley [2014].

5.4 Temporal Regularizer

Due to noise in the input data, e.g. from optical flow computations, small errors in reconstruction can cause temporal flickering. We overcome this by adding a temporal regularization term, consisting of two parts. First, the head pose should change smoothly, and second, the local face deformation should change smoothly. Thanks to our new face model, these constraints can be easily formulated on a subset of our variables, namely the anatomical bone motion and the local blend coefficients. Specifically, for the skull we impose a constraint on the movement of the pivot point $\boldsymbol{o}$, and for the jaw motion and local deformation we directly minimize the change of their parameters over time. The temporal regularization term is thus written as

$$\begin{aligned} E_T(t) = & \lambda_{T1} \| \boldsymbol{o}(t) - \boldsymbol{o}(t-1) \|^2 + \\ & \lambda_{T2} \| \boldsymbol{\Theta}(t) - \boldsymbol{\Theta}(t-1) \|^2 + \\ & \lambda_{T3} \sum_{i=1}^{N} \| \boldsymbol{\alpha}_i(t) - \boldsymbol{\alpha}_i(t-1) \|^2. \end{aligned} \tag{14}$$

Note that in case of the jaw the magnitudes of the angular components expressed in radians and the translational component given in mm are compatible and therefore the terms can be used without reweighting.

5.5 Optimization

Our energy function is defined as a least squares problem, which can be solved by the Gauss-Newton method. Due to the rotational components in $\{M_i\}, M_s$ and $\boldsymbol{\Theta}$ our energy is non-linear. We therefore first linearize the energy using a Taylor expansion and explicitly compute the analytical gradient for each term. Then we compute the Jacobian matrix for the normal equations in the Gauss-Newton solver. We chose to represent the rigid transformations as exponential maps, which have proven to work well for rigid tracking [Bregler et al. 2004]. As each patch is related only to its neighbors, the Jacobian matrix is very sparse. We use the Intel MKL library to solve the sparse matrix to obtain a vector to update our current solution, and this is iterated for N_{iter} iterations.

For all of our datasets we use the following parameters: $\lambda_M = 1$, $\lambda_O = 1$, $\lambda_{A1} = 100$, $\lambda_{A2} = 10000$, $\lambda_{T1} = 40000$, $\lambda_{T2} = 40000$, $\lambda_{T3} = 400$, and $N_{iter} = 12$. The only exception is for the high-speed sequence shown in Fig. 13, where $\lambda_{T3} = 0$ because the local deformation of the real skin is actually very fast. An analysis of the number of patches N and the subspace size K will come in Section 7, but for completeness we use $N = 1000$ and $K = 9$.

The result of local patch reconstruction is a set of distinct local skin patches and the anatomical bone positions for each frame. The remaining step is to combine the patches into a single global face mesh.

6 Global Patch Blending

Our single view local tracking method provides an estimate of the reconstructed face, provided as a set of patches with local deformation and global positions. As we only impose a soft constraint on the overlapping patch boundaries, the shared vertices could have different position estimates from different patches. An example is shown for one frame in Fig. 5.a, where the patches are not connected. We desire a single global face mesh. A naïve approach to combine the patches is to directly average the positions of vertices that belong to more than one patch, as shown in Fig. 5.b, however this results in visible seams between the patches. In order to obtain a seamless reconstruction, we must actually blend all vertices (not just the ones that were in overlap regions). We propose a weighted averaging method, which gives higher influence to patches for which the vertex is geodesically closer to the center of the patch. We approximate geodesic distance by computing the shortest path along mesh edges. Specifically, for each vertex ν, we compute the approximate geodesic distance $\delta_{\nu,i}$ to the center of each patch i, and then compute a weight for the patch as

$$w_{\nu,i} = exp\left(\frac{-\delta_{\nu,i}^2}{\sigma_s^2} \right), \tag{15}$$

where σ_s is the standard deviation of a Gaussian kernel, empirically set to 1.6 times the average width of a patch, as a Gaussian kernel created with this size covers roughly half the area of a 3×3 patch neighborhood within one σ_s. After the weights from all the patches are computed they are normalized to 1. With the normalized weights $\hat{w}_{\nu,i}$, the new position for vertex ν is computed as

$$\hat{\boldsymbol{x}}_\nu(t) = \sum_{i=1}^{N} \hat{w}_{\nu,i} \boldsymbol{x}_{\nu,i}(t), \tag{16}$$

where $\boldsymbol{x}_{\nu,i}(t)$ is the estimated position from patch i. The resulting global patch blend is shown in Fig. 5.c. Note that computing the weights can be time-consuming since many geodesic paths must be traversed, however on the one hand we do not have to compute a weight for every patch, as the influence of patches becomes negligible after approximately $2\sigma_s$, and on the other hand the weights depend only on the mesh topology and the Gaussian kernel which remain fixed for a given actor and are thus computed only once.

7 Subspace Analysis

As we will demonstrate in Section 8, the proposed local deformation model is much more expressive than a global blendshape model, and will require many fewer training shapes. In addition to the number of shapes employed to create the subspace (K in Section 4), our local deformation model is also largely influenced by the size of the patches (related to N in Section 4). The patch size essentially

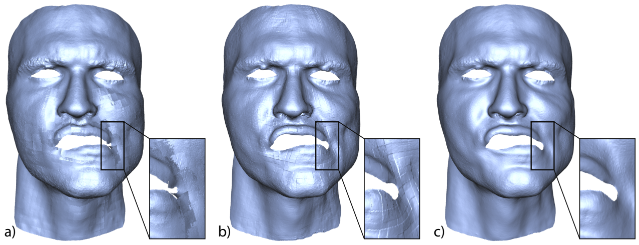

Figure 5: *Global patch blending. (a) The individual patches after optimization, (b) naïvely blending them results in visible seams, (c) our weighted blending result.*

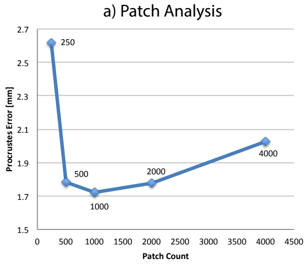

Figure 6: ***Analysis*** *– We analyze the impact of the patch size (a) as well as how many and which expressions to include (b). The nine most significant expressions plus neutral are shown in Fig. 7.*

determines the locality of our model and trades flexibility versus robustness. In this section, we analyze the impact of these two quantities on the model. We will start by determining a good patch size and then look into identifying which expressions to include. To ensure the analysis is not influenced by errors in the input data (e.g. from optical flow), we use ground truth 2D motion vectors in this section and will investigate the degradation of our method under imperfect input data in Section 8. To obtain the ground truth, we perform experiments on a sequence reconstructed by Beeler et al. [2011] and project the known mesh motion onto the image plane of one of the cameras. This approach also gives us ground truth geometry to analyze the single-view reconstruction error.

Patch Size Analysis. As mentioned above, the size of the patches directly influences the locality of our model. The smaller the patches, the better the model will fit to the monocular input data in the image plane, but at the same time the depth will be less well constrained. To identify the optimal patch size we tested varying patch sizes by fitting our model to 160 frames which contain substantial skin deformation. As error measurement we use the Procrustes distance, which corresponds to the average Euclidean distance between our fit and the provided ground truth shape. As can be seen in Fig. 6.a, a partition into around 1000 patches gives the best result, and it also shows that the exact number of patches is not critical since the quality degrades gracefully around the optimum. We therefore chose to use 1000 patches for all results presented in this paper. The subspace was constructed from all available shapes for this experiment and we will describe next how this set can be reduced.

Expression Analysis. When building an actor-specific rig, people typically acquire a well defined set of shapes by scanning the actor. One standardized set of shapes was introduced by Eckman and Friesen [1977] in the late 1970s, which is still very much used in industry with slight variations. This face set contains over 100 shapes. For practical reasons we focus our analysis on a common subset of 26 expressions, which we capture for all three of our actors. To answer the question regarding which of these expressions to include in our subspace, we take an iterative approach, starting with the neutral expression and iteratively adding the most significant missing shape to the deformation subspace. We define the most significant shape as the one that is least well explained by the current subspace, and consequently will expand the subspace the most. To rank the shapes we therefore fit all local patches to all candidate shapes using the current subspace and again compute the Procrustes distance. To reduce the danger of overfitting to one person we compute the distance on all of our three actors simultaneously. If the next most significant candidate is an asymmetric expression, we also include its counterpart into the subspace to prevent biasing the model to one side. To assess the quality of the subspace, we then test it on a validation sequence of ∼480 frames, for which we also have ground truth. Fig. 6.b summarizes the results and shows that the error is reduced exponentially by incrementally adding the most significant shape to the model. The nine most significant shapes plus neutral are shown in Fig. 7. Already with the two most significant shapes, namely the open and closed faces where the actor stretches and compresses the complete face as much as possible, our model can cut the error in half. From this analysis we found that nine expressions plus neutral provide a good tradeoff in terms of fitting accuracy over model complexity and we thus use a nine-dimensional deformation subspace for all results in this paper. In contrast to typical global blendshape models that require over 100 shapes in industry practices, our local model allows for an order of magnitude reduction in the number of shapes and hence amount of pre-processing work and actor time required to build the face model.

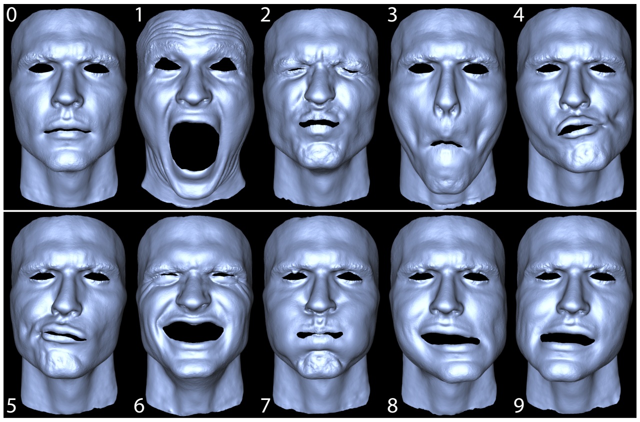

Figure 7: ***Shape Ranking*** *– We analyze which shapes are the most important ones to be included into our deformation subspace and rank them according to their significance. The two most important ones are the open and closed faces, which intuitively capture the gross deformation of a face.*

8 Results

In this section we first analyze our method quantitatively and qualitatively and compare to other models. Then we demonstrate the versatility of our approach in several different application scenarios, including dense optical flow based performance capture, sparse marker based motion capture and very sparse direct manipulation via a sketch interface.

8.1 Evaluation and Comparison

Just as for Section 7 we continue to use ground truth motion for our evaluation and then switch over and analyze the performance under imperfect input data. We start by assessing the importance of our two main contributions, the local deformation model and the anatomical constraints. For this we compare our model to the traditional global

		G26	G9	GA9	L9	LA9
GT Motion	μ	3.01	7.57	5.40	5.44	**1.39**
	σ	1.73	3.14	1.82	3.64	**0.61**
Dense O-Flow	μ	7.90	10.86	9.36	8.95	**5.01**
	σ	2.44	4.41	3.65	3.95	**0.76**
Sparse Markers	μ	5.22	9.26	6.11	7.83	**1.88**
	σ	2.98	2.97	2.89	4.73	**0.77**

Table 1: ***Model Comparison*** *– We list mean Procrustes error (μ) and standard deviation (σ) in mm, computed over $\sim$480 frames. For conciseness, 'G/L' specifies the model (global blendshapes vs. local deformation), 'A' indicates that anatomical constraints were used, and the number stands for the amount of shapes employed to build the subspace in addition to the neutral. The proposed model (rightmost column) performs best in all cases.*

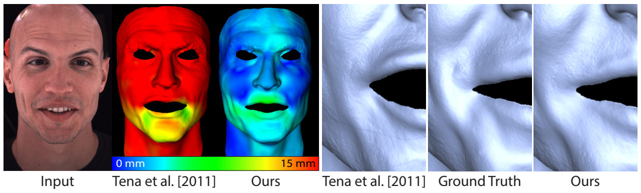

Figure 8: ***Comparison to Tena et al.*** *– Our local deformation model can reconstruct the local shape more accurately than the region-based linear model of Tena et al. [2011] since our patch layout exhibits much higher granularity and is thus more flexible (right). Even so, our method does not suffer from depth ambiguity common to local models thanks to the anatomical constraints and thus also exhibits higher accuracy in absolute position (left).*

blendshape model as well as a naïve local deformation model that does not use anatomy. For the global blendshape model we used all 26+1 blendshapes available and for the local deformation models we employ a subset of 9+1 shapes as described in Section 7. The '+1' refers to the neutral shape. For completeness we also added anatomical constraints to the global blendshape model for comparison. Table 1 lists mean Procrustes errors and standard deviations for all five models. For conciseness, 'G/L' denotes global blendshapes vs. local deformation model, 'A' indicates that anatomical constraints where employed and the number specifies the amount of shapes used to build the subspace. The first row reports errors under perfect motion input, the second row shows how the models degrade under real world conditions using optical flow ([Brox et al. 2004]) as input data, and the third row shows the impact of reducing the density of the input data. As can be seen, the proposed approach clearly outperforms the other methods for all scenarios.

Perfect Input Data. Under perfect input data, the proposed method with 9+1 shapes outperforms the full global blendshape model with 26+1 shapes by more than a factor of two and the naïve local deformation model almost by a factor of four. Fig. 9 gives an indication on how the errors are distributed and shows that both global and local models distribute the error over the full face due to the least squares norm used. Our method tends to concentrate the error predominantly at the neck, where no anatomical constraints can provide robustness. The graph on top of Fig. 9 shows that our method performs consistently well over all $\sim$480 frames, where the others show great temporal variation. For the global blendshape model, this stems from the fact that it performs very well when the expression is part of the subspace but cannot extrapolate at all. Therefore, global models require to include many more shapes than local ones. Following Fig. 6.b our local deformation model can achieve similar performance to the global blendshape model with 26+1 shapes already with only 4+1 shapes, which means it requires about 6-7 times less expressions of the actor to be captured, processed and stored. Local models have the power to extrapolate, however, without anatomical constraints the local deformation model suffers from depth ambiguities, leading to the worst overall performance. Adding anatomical constraints not only improves the quality of the local deformation model substantially but also helps to constrain the global blendshape model better, as it adds stabilization to the global model which effectively reduces the amount of error shifted from fitting the non-rigid expression to the rigid head pose.

We also compare our new model to a region-based linear face model [Tena et al. 2011] on the perfect input data. Both models have been built from the same $9 + 1$ training shapes (Fig. 7). The method of Tena et al. achieves a mean Procrustes error of $\mu = 5.84$mm with standard deviation $\sigma = 2.49$mm. Compared to other approaches (Table 1), this region-based linear model improves over pure global blendshapes, but still falls short when compared to our method ($\mu = 1.39$mm, $\sigma = 0.61$mm). These measurements are further supported by Fig. 8. The proposed model not only reconstructs the local shape more faithfully since it exhibits much higher granularity (1000 patches vs. 13 regions), it also suffers much less from depth ambiguity and can thus more faithfully reconstruct the face in depth.

Imperfect Input Data. Under imperfect input data, referring to Table 1 the performance of all methods drop and on first glance the quality improvement of our method over the others seems less significant. Considering the low standard deviation, however, indicates that the gross error of our method is due to head pose depth estimation, while the relative expression to the skull is still very reasonable. This hypothesis is confirmed by computing the Procrustes error after stabilization, which reduces the error from 5.01mm down to 1.68mm. This means that even though the absolute head pose in depth is not estimated perfectly, the relative expression motion can still be recovered very well, which is good news as this is the most important information for many applications, such as retargeting.

Sparse Input Data. The last row of Table 1 shows the performance under sparse input data provided by a set of $\sim$110 markers distributed as depicted in Fig. 14. Since marker positions are typically tracked very accurately and robustly, we again use the ground truth 2D motion at these sparse locations. Our method degrades only minimally for this sparse input data leading to visually very similar expressions as can be seen in Fig. 14. This indicates that our method is very well suited to augment traditional marker based motion capture pipelines.

Stabilization. As a very beneficial side-effect, our method produces shapes which are implicitly stabilized. Stabilization refers to the process of removing the rigid motion of the head to recover the true expression changes [Beeler and Bradley 2014] and is essential, for example, for performance transfer. Stabilization is a very tedious, time consuming and oftentimes manually assisted process. To assess the quality of the estimated skull motion we compare in Fig. 10 to the method of Beeler and Bradley [2014] and found that we achieve very similar results. As before, most of the error is concentrated on the neck, not affecting the face itself.

8.2 Applications

Now we demonstrate the robustness and versatility of the proposed method on a variety of input sources. The temporal aspects of our results are best viewed in the accompanying video.

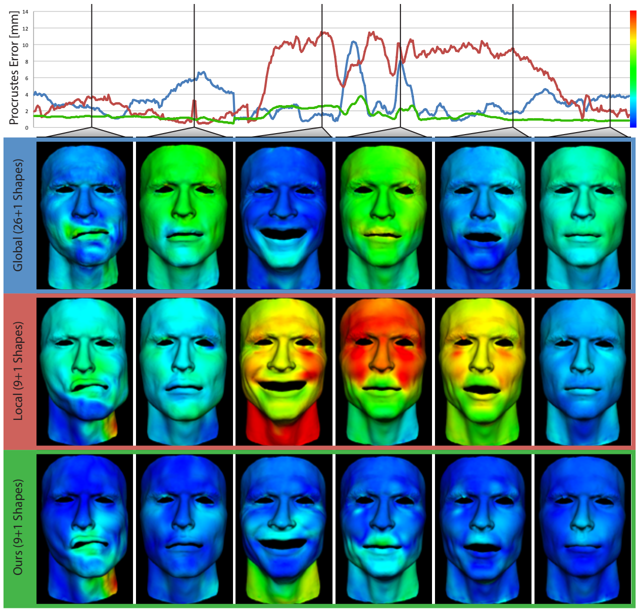

Figure 9: ***Model Comparison*** *– Comparing the global blendshape model based on 26+1 blendshapes (blue) to the local deformation models without (red) and with anatomical constraints (green) based on a subset of 9+1 shapes for a sequence of ∼480 frames (left to right).*

Dense Input. We start with three examples that use dense optical flow as the data term. Fig. 11 shows results achieved in a studio setup, where we added a synchronized side camera for validation (Dalsa Falcon 4M60). The figure qualitatively compares to results provided by Beeler et al. [2011]. Our method achieves visually very similar results from a single camera where Beeler and colleagues use seven. The main difference is visible in the less-pronounced wrinkles where they are not part of the local shape subspace.

Fig. 12 shows results on footage acquired in the wild, from a helmet-mounted GoPro 4 (left) and a handheld iPhone 5 (right), both captured outside under overcast sky. Such setups are very compelling as they require only very affordable hardware and impose only minimal constraints onto the actor. The GoPro sequence contains a lot of secondary motion as the actor moves, and also substantial camera-shake, which is best seen in the accompanying video. Nevertheless, our method manages to robustly reconstruct the facial expressions over time since it implicitly stabilizes the face via the anatomical constraints.

Finally, Fig. 13 demonstrates a very challenging use case, where we capture an actor with a high-speed camera (Sony F55) at 240fps while blowing compressed air at his face. The stream of air forms ripples on the actors cheek propagating upwards. Capturing such a scenario with a global blendshape model would be nearly impossible since the external forces create face shapes which are far outside the pre-acquired expression space. Our model is capable of reproducing these ripples, even though the optical flow estimation is quite inaccurate at times.

Sparse Input. Our model does not require dense input but also performs well when constraining only a very small subset of the vertices. Fig. 14 mimics a classical marker based motion capture scenario, where the actor's face is covered with a sparse set of markers. These markers are then tracked over time and used as sparse constraints to our method. The result is visually almost identical to using dense constraints and also very similar to the high-quality shapes provided by Beeler et al. [2011]. The last row in Table 1 supports these findings quantitatively, showing the error increases only slightly when using sparse constraints as input.

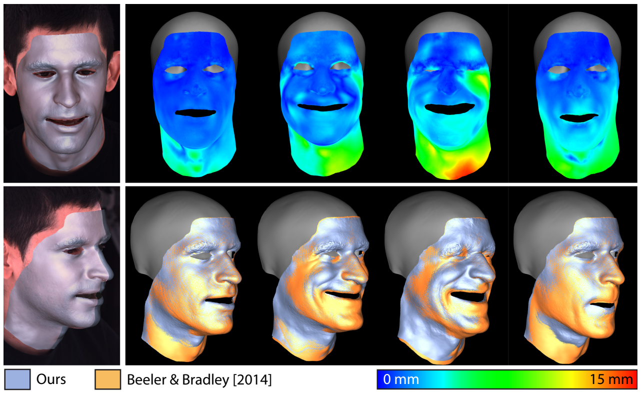

Figure 10: ***Stabilization*** *– Our method implicitly stabilizes the shapes since it estimates the rigid skull transformation. This allows to compute the skin deformation relative to the skull, which is less susceptible to depth ambiguities than the absolute position as can be seen in the lower left. Overall, our stabilized results are very similar to the ones of Beeler and Bradley [2014].*

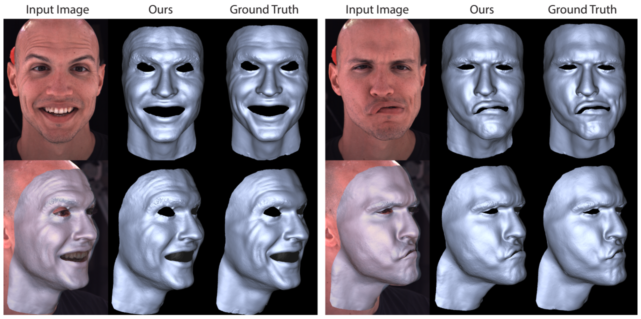

Figure 11: ***Qualitative Evaluation*** *– We show results of our method on two different expressions using dense ground truth motion. The reconstructed shapes are visually very similar to ground truth, also when seen from a side view. The biggest visible difference is in the less-pronounced wrinkles where they are not part of the local shape subspace.*

Reducing the input even further, Fig. 15 shows an application of direct, user guided manipulation. In a sketch based interface, the user can draw a set of source (green) and target strokes (blue) onto the face to control its deformation. The yellow stroke in the first column indicates that source and target strokes coincide. Drawing on an image in 2D instead of manipulating a shape in 3D can be very attractive for less technical users. For example, by fixing a stroke on the left eyebrow and moving a second stroke on the right eyebrow upwards, the system plausibly synthesizes a shape including forehead wrinkles on the right side of the face. In the third column, the chin area is translated with a single stroke causing also the underlying jaw to move. The artistic control can lead to shapes which are physically not achievable by the actor, such as shown in column four. Here the jaw has been constrained into an extremely wide open position by the artist, yet still the resulting shape does look plausible. This could allow for cartoony and stylized facial animation, where expressions are oftentimes exaggerated.

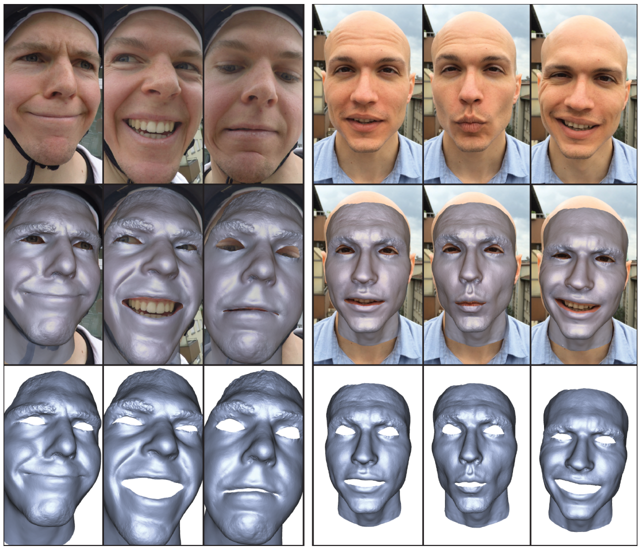

Figure 12: ***Capture In The Wild*** *– Results on footage acquired with consumer grade cameras in unconstrained outdoor setups. On the left we show results on a helmet-mounted GoPro 4 and on the right on video acquired with a handheld iPhone 5.*

Extension to multiple views. Lastly, we would like to point out that while designed for monocular input, extending the method to include multiple views is straightforward. Adding additional views, be it overlapping or not, simply adds additional equations to the motion energy E_M introduced in Eq. 8. By adding just one more camera, the slight error in absolute depth can be removed as shown in Fig. 16. We would like to point out that even though our method has minor errors in the monocular case when estimating absolute depth, the relative deformations caused by the expression can be faithfully recovered thanks to the anatomical constraints, as can be seen in Fig. 10.

8.3 Limitations and Future Work

The proposed model presents a new way to think of model-based facial tracking. In this paper we answer some of the questions but quite a few remain, presenting some interesting future work.

We currently do only consider square patches and uniform segmentation of the face. It would be very interesting to investigate the benefits of using an adaptive approach with potentially irregular patch shapes. Such an adaptive model could then better account for the spatially varying complexity of skin deformation on the face. Similarly, we currently regularize the data terms spatially uniformly, but in particular for closeup cameras with large field of view, it could prove beneficial to use spatially varying weights, for example scaled by the area occupied in the images. Also, the dimensionality of the subspaces is currently the same for all patches in our implementation. Obviously, not every patch exhibits the same range of deformation and thus an adaptive scheme could be employed to compress the subspaces where they are less expressive.

Furthermore, from the three actors used in this work it is apparent that not every person performs the same expressions in the same way. An example is shown in Fig. 17, where one actor forms a kiss mouth during the compressed expression while the others compress their lips inwards and outwards, respectively. This has implications when building the subspace and demands for very careful acquisition and actor guidance.

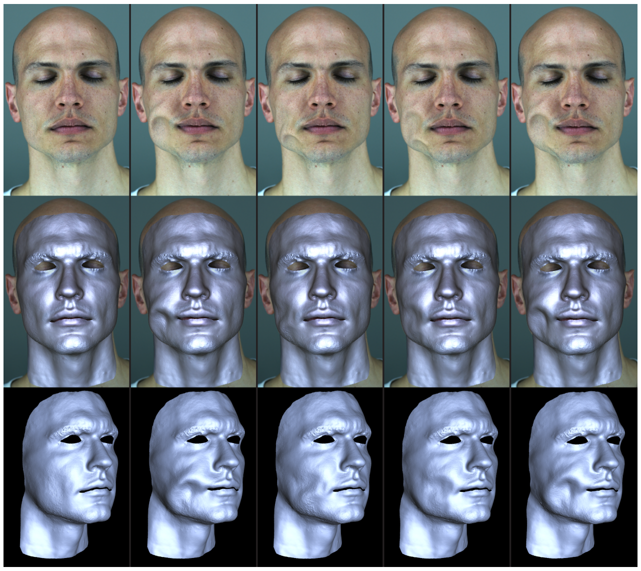

Figure 13: ***High Speed Ripples*** *– Our method successfully recovers the shape of ripples propagating over the face caused by a stream of compressed air. Recovering these shapes is possible thanks to the local deformation model, which allows extrapolation outside of the pre-captured shapes.*

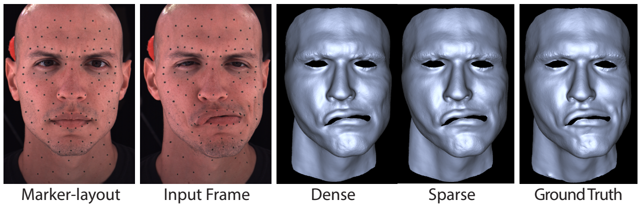

Figure 14: ***Marker Input*** *– To test the performance of our method in a sparse MoCap scenario, we synthetically applied a set of ∼110 markers on the face as shown on the left. At the marker locations we provide 2D motion vectors from ground truth, to mimic the fact that markers can typically be tracked very accurately and robustly. The recovered mesh (fourth column) is visually almost indistinguishable from the mesh computed with dense input (third column) and also very similar to ground truth (right column).*

The anatomical constraints proposed in this paper are only the beginning of what could be done. Adding in additional constraints, for example teeth to better constrain the lips, jaw limits or even more sophisticated tissue models could further improve the quality of the approach.

Finally, our method currently requires about 1-2 minutes to reconstruct one frame on a modern desktop computer for our high resolution meshes (∼700k vertices). We expect the method to perform similarly well on lower resolution meshes and would like to improve the overall implementation to achieve interactive rates. Then, sketch based manipulation as shown in Fig. 15 could become an interesting means of manipulating facial rigs in the future.

9 Conclusion

We present a novel anatomically-constrained local deformation model for facial performance capture from monocular input data. The proposed model is much more expressive than the traditionally employed global blendshape models and requires many fewer ex-

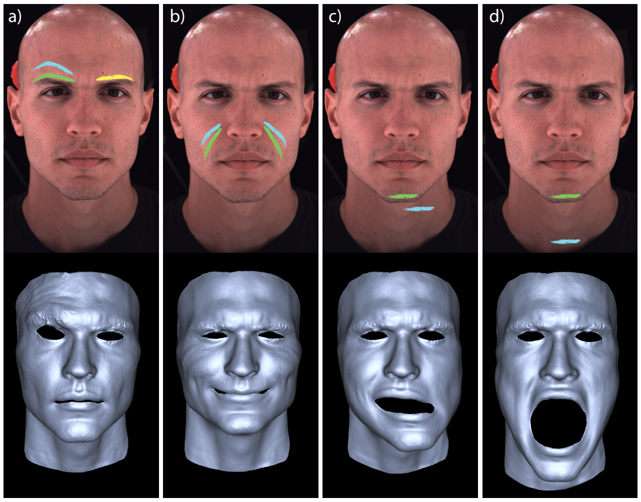

Figure 15: ***Artistic Control***– *As our method can handle very sparse input constraints, it opens up new possibilities for direct control. In this example, a user draws a few source (green) and target (blue) strokes on the 2D image and the method plausibly deforms the actors face according to these strokes, adding in large scale expression change and skin wrinkling, and even moves the jaw. The method also extrapolates well to non-physical shapes (d) and could be used e.g. to create stylized and cartoony facial animations.*

pressions to be pre-acquired. Unlike local blendshape models, our local deformation model explicitly decouples the rigid and non-rigid local motion, allowing to recover face shapes at very high accuracy. Furthermore, the proposed combination with anatomical constraints renders it extremely robust, and suitable for single-view reconstruction. In addition, the method simultaneously provides an estimate of the underlying skull bone, which allows to stabilize the captured performance and to only extract the motion caused by the expression itself without superposition of the head motion. We conduct an in-depth analysis of the model parameters, namely the number of shapes to use when building the subspace and the number of patches which determine the locality of the model.

We demonstrate the versatility of the proposed method on a number of different inputs. We show results on footage acquired with different cameras, ranging from studio setups with well controlled illumination to unconstrained outdoor acquisition using GoPro and iPhone devices, which are small and readily available. We also demonstrate, for the first time, spatio-temporal reconstructions from high-speed video footage of a face deforming due to external forces. Specifically, we reconstruct the ripples forming on the face when blown at with compressed air. The method works not only on dense input data but generalizes also well to sparse data, such as marker based MoCap or even artist created sketches.

We believe that the anatomically-constrained local deformation model introduced in this work will have a substantial impact on different areas of facial performance capture and animation, as it combines the robustness of the traditionally employed global models with the flexibility of local models. It also requires substantially less expressions to be acquired, processed and integrated, which reduces the effort required from actors and artists alike.

Acknowledgements

We gratefully acknowledge Roland Mörzinger for his suggestions and discussions. We also thank the actors (Seth Fry, Michael Erikson, and José Echevarría) as well as Max Grosse for his help with obtaining the video footage.

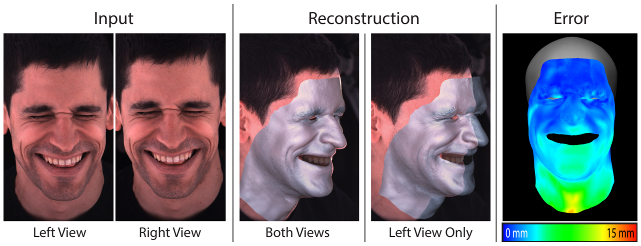

Figure 16: ***Multi view extension***– *Even though designed for the single view case, incorporating additional views into our method is straightforward. By adding a second view, the absolute depth can be recovered better and also the relative skin deformation is improved a bit as shown on the right when compared to the monocular result shown in the fourth column of Fig. 10. Thanks to the anatomical constraints the discrepancy in the relative skin deformation is much less than for the absolute depth.*

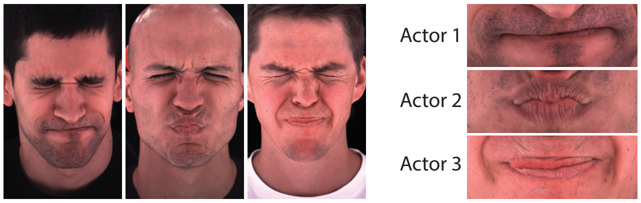

Figure 17: *One caveat is that actors might perform the same expression very differently. On the left we show the second most significant shape according to our analysis, which the three different actors interpreted in three very different ways, ranging from kiss shaped to fully compressed lips (right).*

References

Beeler, T., and Bradley, D. 2014. Rigid stabilization of facial expressions. *ACM Trans. Graphics (Proc. SIGGRAPH) 33*, 4, 44:1–44:9.

Beeler, T., Bickel, B., Sumner, R., Beardsley, P., and Gross, M. 2010. High-quality single-shot capture of facial geometry. *ACM Trans. Graphics (Proc. SIGGRAPH).*

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 Trans. Graphics (Proc. SIGGRAPH) 30*, 75:1–75:10.

Black, M., and Yacoob, Y. 1995. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *ICCV*, 374–381.

Blanz, V., and Vetter, T. 1999. A morphable model for the synthesis of 3d faces. In *Proc. SIGGRAPH*, 187–194.

Bouaziz, S., Wang, Y., and Pauly, M. 2013. Online modeling for realtime facial animation. *ACM Trans. Graphics (Proc. SIGGRAPH) 32*, 4, 40:1–40:10.

Bradley, D., Heidrich, W., Popa, T., and Sheffer, A. 2010. High resolution passive facial performance capture. *ACM Trans. Graphics (Proc. SIGGRAPH) 29*, 41:1–41:10.

BREGLER, C., MALIK, J., AND PULLEN, K. 2004. Twist based acquisition and tracking of animal and human kinematics. *IJCV 56*, 3, 179–194.

BROX, T., BRUHN, A., PAPENBERG, N., AND WEICKERT, J. 2004. High accuracy optical flow estimation based on a theory for warping. In *ECCV*. 25–36.

BRUNTON, A., BOLKART, T., AND WUHRER, S. 2014. Multilinear wavelets: A statistical shape space for human faces. In *ECCV*.

CAO, C., WENG, Y., LIN, S., AND ZHOU, K. 2013. 3d shape regression for real-time facial animation. *ACM Trans. Graphics (Proc. SIGGRAPH) 32*, 4, 41:1–41:10.

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

CAO, C., BRADLEY, D., ZHOU, K., AND BEELER, T. 2015. Real-time high-fidelity facial performance capture. *ACM Trans. Graphics (Proc. SIGGRAPH)*.

CHEN, Y.-L., WU, H.-T., SHI, F., TONG, X., AND CHAI, J. 2013. Accurate and robust 3d facial capture using a single rgbd camera. In *ICCV*.

COOTES, T. F., EDWARDS, G. J., AND TAYLOR, C. J. 2001. Active appearance models. *IEEE TPAMI 23*, 6, 681–685.

DECARLO, D., AND METAXAS, D. 1996. The integration of optical flow and deformable models with applications to human face shape and motion estimation. In *CVPR*, 231.

EKMAN, P., AND FRIESEN, W. V. 1977. Facial action coding system.

ESSA, I., BASU, S., DARRELL, T., AND PENTLAND, A. 1996. Modeling, tracking and interactive animation of faces and heads using input from video. In *Proc. of Computer Animation*, 68.

FYFFE, G., JONES, A., ALEXANDER, O., ICHIKARI, R., AND DEBEVEC, P. 2014. Driving high-resolution facial scans with video performance capture. *ACM Trans. Graphics 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 Trans. Graphics (Proc. SIGGRAPH Asia) 32*, 6, 158:1–158:10.

GHOSH, A., FYFFE, G., TUNWATTANAPONG, B., BUSCH, J., YU, X., AND DEBEVEC, P. 2011. Multiview face capture using polarized spherical gradient illumination. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 30*, 6, 129:1–129:10.

GOWER, J. C. 1975. Generalized procrustes analysis. *Psychometrika 40*, 1.

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

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

KOBBELT, L., VORSATZ, J., AND SEIDEL, H.-P. 1999. Multiresolution hierarchies on unstructured triangle meshes. *Comput. Geom. Theory Appl. 14*, 1–3, 5–24.

LAU, M., CHAI, J., XU, Y.-Q., AND SHUM, H.-Y. 2009. Face poser: Interactive modeling of 3d facial expressions using facial priors. *ACM Trans. Graph. 29*, 1 (Dec.), 3:1–3:17.

LEWIS, J. P., ANJYO, K., RHEE, T., ZHANG, M., PIGHIN, F., AND DENG, Z. 2014. Practice and Theory of Blendshape Facial Models. In *Eurographics 2014 - State of the Art Reports*, The Eurographics Association, S. Lefebvre and M. Spagnuolo, Eds.

LI, H., ROIVAINEN, P., AND FORCHEIMER, R. 1993. 3-d motion estimation in model-based facial image coding. *IEEE TPAMI 15*, 6, 545–555.

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

NA, K.-G., AND JUNG, M.-R. 2011. Local shape blending using coherent weighted regions. *The Vis. Comp. 27*, 6-8, 575–584.

NEUMANN, T., VARANASI, K., WENGER, S., WACKER, M., MAGNOR, M., AND THEOBALT, C. 2013. Sparse localized deformation components. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 32*, 6, 179:1–179:10.

RHEE, T., HWANG, Y., KIM, J. D., AND KIM, C. 2011. Real-time facial animation from live video tracking. In *Proc. SCA*, 215–224.

SARAGIH, J. M., LUCEY, S., AND COHN, J. F. 2011. Deformable model fitting by regularized landmark mean-shift. *IJCV 91*, 2, 200–215.

SHI, F., WU, H.-T., TONG, X., AND CHAI, J. 2014. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 33*.

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

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

VALGAERTS, L., WU, C., BRUHN, A., SEIDEL, H.-P., AND THEOBALT, C. 2012. Lightweight binocular facial performance capture under uncontrolled lighting. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 31*, 6.

VLASIC, D., BRAND, M., PFISTER, H., AND POPOVIĆ, J. 2005. Face transfer with multilinear models. *ACM Trans. Graphics (Proc. SIGGRAPH) 24*, 3, 426–433.

WEISE, T., LI, H., VAN GOOL, L., AND PAULY, M. 2009. Face/off: live facial puppetry. In *Proc. SCA*, 7–16.

WEISE, T., BOUAZIZ, S., LI, H., AND PAULY, M. 2011. Realtime performance-based facial animation. *ACM Trans. Graphics (Proc. SIGGRAPH) 30*, 4, 77:1–77:10.

ZHANG, L., SNAVELY, N., CURLESS, B., AND SEITZ, S. M. 2004. Spacetime faces: high resolution capture for modeling and animation. *ACM Trans. Graphics (Proc. SIGGRAPH)*, 548–558.

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 Trans. Graphics (Proc. SIGGRAPH) 33*, 4, 156:1–156:12.