EUROGRAPHICS 2017 / L. Barthe and B. Benes
(Guest Editors)

Volume 36 (2017), Number 2

Multi-View Stereo on Consistent Face Topology

G. Fyffe[1*†], K. Nagano[1*], L. Huynh[1], S. Saito[2], J. Busch[1†], A. Jones[1], H. Li[1,2], and P. Debevec[1,2]

[1]USC Institute for Creative Technologies, USA
[2]University of Southern California, USA

(a) (b) (c) (d)

Figure 1: *Our method automatically reconstructs dynamic facial models from multi-view stereo with consistent parameterization. (a) Facial reconstruction with artist-quality mesh topology. (b) Reconstructed facial mesh overlaid on the input video. (c) Reconstructed face model with a displacement map estimated from details in the images. (d) Close-up of fine details, such as pores and dynamic wrinkles from (c).*

Abstract

We present a multi-view stereo reconstruction technique that directly produces a complete high-fidelity head model with consistent facial mesh topology. While existing techniques decouple shape estimation and facial tracking, our framework jointly optimizes for stereo constraints and consistent mesh parameterization. Our method is therefore free from drift and fully parallelizable for dynamic facial performance capture. We produce highly detailed facial geometries with artist-quality UV parameterization, including secondary elements such as eyeballs, mouth pockets, nostrils, and the back of the head. Our approach consists of deforming a common template model to match multi-view input images of the subject, while satisfying cross-view, cross-subject, and cross-pose consistencies using a combination of 2D landmark detection, optical flow, and surface and volumetric Laplacian regularization. Since the flow is never computed between frames, our method is trivially parallelized by processing each frame independently. Accurate rigid head pose is extracted using a PCA-based dimension reduction and denoising scheme. We demonstrate high-fidelity performance capture results with challenging head motion and complex facial expressions around eye and mouth regions. While the quality of our results is on par with the current state-of-the-art, our approach can be fully parallelized, does not suffer from drift, and produces face models with production-quality mesh topologies.

Categories and Subject Descriptors (according to ACM CCS): I.4.1 [Computer Vision]: Image Processing and Computer Vision—ScanningI.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation

1. Introduction

Video-based facial performance capture has become a widely established technique for the digitization and animation of realistic virtual characters in high-end film and game production. While recent advances in facial tracking research are pushing the boundaries of real-time performance and robustness in unconstrained capture settings, professional studios still rely on computationally demanding offline solutions with high resolution imaging. To further avoid the uncanny valley, time-consuming and expensive artist input, such as tracking clean-up or key-framing, is often required to

* G. Fyffe, and K. Nagano are joint first authors.
† G. Fyffe and J. Busch are now at Google.

fine-tune the automated tracking results and ensure consistent UV parameterization across the input frames.

State-of-the-art facial performance capture pipelines are mostly based on a multi-view stereo setup to capture fine geometric details, and generally decouple the process of model building and facial tracking. The facial model (often a parametric blendshape model) is designed to reflect the expressiveness of the actor but also to ensure that any deformation stays within the shape and expression space during tracking. Because of the complexity of facial expressions and potentially large deformations, most trackers are initialized from the previous input frames. However, such sequential approaches cannot be parallelized and naturally result in drift, which requires either artist-assisted tracking corrections or ad-hoc segmentation of the performance into short temporal clips.

We show in this work, that it is possible to directly obtain, for any frame, a high-resolution facial model with consistent mesh topology using a passive multi-view capture system with flat illumination and high-resolution input images. We propose a framework that can accurately warp a reference template model with existing texture parameterization to the face of any person, and demonstrate successful results on a wide range of subjects and challenging expressions. While existing multi-view methods either explicitly compute the geometry [BHB*11, KH12] or implicitly encode stereo constraints [VWB*12], they rely on optical flow or scene-flow to track a face model, for which computation is only possible sequentially. Breaking up the performance into short clips using anchor frames or key frames with a common appearance is only a partial solution, as it requires the subject to return to a common expression repeatedly throughout the performance.

Our objective is to warp a common template model to a different person in arbitrary poses and different expressions while ensuring consistent anatomical matches between subjects and accurate tracking across frames. The key challenge is to handle the large variations of facial appearances and geometries, as well as the complexity of facial expression and large deformations. We propose an appearance-driven mesh deformation approach that produces intermediate warped photographs for reliable and accurate optical flow computation. Our approach effectively avoids image discontinuities and artifacts often caused by methods based on synthetic renderings or texture reprojection.

In a first pass, we compute temporally consistent animations, that are produced from independently computed frames, by deforming a template model to the expressions of each frame while enforcing consistent cross-subject correspondences. To initialize our face fitting, we leverage recent work in facial landmark detection. In each subsequent phase of our method, the appearance-based mesh warping is driven by the mesh estimate from the previous phase. We show that even where the reference and target images exhibit significant differences in appearance (due to significant head rotation, different subjects, or expression changes), our warping approach progressively converges to a high-quality correspondence. Our method does not require a complex facial rig or blendshape priors. Instead, we deform the full head topology according to the multi-view optical flow correspondences, and use a combination of surface and volumetric Laplacian regularization to produce a well-behaved shape, which helps especially in regions that are prone to occlusion and inter-penetration such as the eyes and mouth pocket.

As the unobserved regions such as the back of the head are inferred from the Laplacian deformation, these regions may be temporally inconsistent in the presence of significant head motion or expression changes. Thus we introduce a PCA based technique for general deformable surface alignment to align and denoise the facial meshes over the entire performance, which improves temporal consistency around the top and back of the head and reduces high-frequency "sizzling" noise. Unlike [BB14, WBGB16], our method does not employ any prior knowledge of anatomy. We then compute a subject-specific template and refine the performance capture in a second pass to achieve pore-level tracking accuracy.

Our method never computes optical flow between neighboring frames, and never compares a synthetic rendering to a photograph. Thus, our method does not suffer from drift, and accurately corresponds regions that are difficult to render synthetically such as around the eyes. Our method can be applied equally well to a set of single-frame expression captures with no temporal continuity, bringing a wide variety of facial expressions into (u,v) correspondence with pore-level accuracy. Furthermore, our joint optimization for stereo and fitting constraints also improves the digitization quality around highly occluded regions such as mouth, eyes, and nostrils as they provide additional reconstruction cues in the form of shape priors. We report timings for each step of our method, most of which are trivially parallelizable across multiple computers. In summary, our contributions include:

- a fully parallelizable multi-view stereo facial performance capture pipeline that produces a high-quality facial reconstruction with consistent mesh topology.
- an appearance-driven mesh deformation algorithm using optical flow on high-resolution imaging data combined with volumetric Laplacian regularization.
- a PCA-based pose estimation and denoising technique for general deformable surfaces.

2. Related Work

Driving the motion of digital characters with real actor performances has become a common and effective process for creating realistic facial animation. Performance-driven facial animation dates back as least as far as Williams [Wil90] who used facial markers in monocular video to animate and deform a 3D scanned facial model. Guenter et al. [GGW*98] drove a digital character from multi-view video by 3D tracking a few hundred facial markers seen in six video cameras. Yet even a dense set of facial markers can miss subtle facial motion details necessary for conveying the entire meaning of a performance. Addressing this, Disney's Human Face Project [Yea02] was perhaps the first to use dense optical flow on multi-view video of a facial performance to obtain dense facial motion for animation, setting the stage for the markerless multi-view facial capture system used to animate realistic digital characters in the "Matrix" sequels [BPL*05]. In our work, we use a multi-view video setup to record facial motion, but fit a model to the images per time instant rather than temporally tracking the performance.

Realistic facial animation may also be generated though physical simulation as in [PB81, TW93]. The computer animation "The

Jester" [CSDV99] tracked the performer's face with a standard set of mocap markers but used finite element simulation to simulate higher-resolution performance details such as the skin wrinkling around the eyes. [SNF05] used a bone, flesh, and muscle model of a face to reverse-engineer the muscle activations which generate the same motion of the face as recorded with mocap markers. Recent work showed a way to automatically construct personalized anatomical model for volumetric facial tissue simulations [CBE*15]. In our work, we use a volumetric facial model to enable robust model fitting solutions including occluded regions.

Faces assume many shapes but have the same features in similar positions, and as a result can be modeled with generic templates such morphable models [BV99]. Such models have proven useful in recent work for real-time performance tracking [LWP10, WBLP11,LYYB13,BWP13,CHZ14,HMYL15,CWW*16,SLL16], expression transfer [WLVGP09, BGY*13, TZN*15, TZS*16], and performance reconstruction from monocular video [GVWT13, SKS14, SWTC14, GZC*16]. Recent work demonstrated reconstruction of a personalized avatar from mobile free-form videos [IBP15], medium scale dynamic wrinkles from RGB monocular video [CBZB15], and high fidelity mouth animation for an head mounted display [OLSL16]. These template based approaches can provide facial animation in a common artist-friendly topology with blendshape animations. But they cannot capture shape details outside of the assumed linear deformation subspace, which may be important for high quality expressive facial animation. On the other hand, our technique captures accurate 3D shapes comparable to multi-view stereo, on a common head topology without the need for complex facial rigs.

Multi-view stereo approaches [FP10, BBB*10, BHB*11] remain popular since they yield verifiable and accurate geometry even though they require offline computation. Our dynamic performance reconstruction technique differs from techniques such as [FP09,BHPS10,BHB*11] in that we do not begin by solving for independent multi-view stereo geometry at each time instant. In fact our method does not require a set of high-resolution facial scans (or even a single facial scan) of the subject to assist performance tracking as in [ARL*09, HCTW11, GVWT13, AFB*13, FJA*14]. Instead, we employ optical flow and surface/volume Laplacian priors to constrain 3D vertex estimates based on a template.

Video-based facial performance capture is susceptible to "drift", meaning inconsistencies in the relationship between facial features and the mesh parameterization across different instants in time. For example, a naive algorithm that tracks vertices from one frame to the next will accumulate error over the duration of a performance. Previous works have taken measures to mitigate drift in a single performance. However, none of these approaches lends itself to multiple performance clips, or collections of single-frame captures. Furthermore, previous works addressing fine-scale consistency involve at least one manual step if high-quality topology is desired. [BHB*11] employs a manually selected reference frame and geometry obtained from stereo reconstruction. If a clean topology is desired, is it edited manually. The method locates "anchor frames" similar to the reference frame to segment the performance into short clips, and optical flow tracking is performed within each clip and across clip seams. The main drawback of this method is that all captures must contain well distributed anchor frames that resemble the reference, which limits the expressive freedom of the performer. [KH12] constructs a minimum spanning tree in appearance space and employs non-sequential tracking to reduce drift, combined with temporal tracking to reduce temporal seams. The user must manually create a mesh for the frame at the root of the tree, based on geometry obtained from stereo reconstruction. Despite the minimum spanning tree, expressions far from the root expression still require concatenation of multiple flow fields, accumulating drift. If single-frame captures are included in the data, it may fail altogether. [VWB*12] employs the first frame of a performance as a template. If clean topology is desired, it must be manually edited. Sequences are processed from the first frame to the last. Synthetic renderings of the template are employed to reduce drift via optical flow correction. This method cannot handle multiple performance clips or single frame captures in correspondence. [GVWT13] employs a neutral facial scan as a template, and requires manually refined alignment of the neutral scan to the starting frame of the performance. The method locates "key frames" resembling the neutral scan (much like anchor frames) to segment the performance into short clips that are tracked via temporal optical flow, and employs synthetic renderings of the neutral scan to reduce drift via optical flow correction. This method is unsuitable for collections of multiple performances, as the manual initial alignment required for each performance would be prohibitive and error-prone. [FJA*14] employs multiple facial scans, with one neutral scan serving as a template. The neutral scan topology is produced manually. The neutral scan is tracked directly to all performance frames and all other scans using optical flow, which is combined with temporal optical flow and flows between a sparse set of frames and automatically selected facial scans to minimize drift. This method handles multiple performance clips and multiple single-frame captures in correspondence, but requires multiple facial scans spanning the appearance space of the subject's face, one of which is manually processed.

3. Shared Template Mesh

Rather than requiring a manually constructed personalized template for each subject, our method automatically customizes a generic template including the eyes and mouth interior. We maintain a consistent representation of this face mesh throughout our process: a shared template mesh with its deformation parameterized on the vertices. The original template can be any high-quality artist mesh with associated multi-view photographs. To enable volumetric regularization, we construct a tetrahedral mesh for the template using TetGen [Si15] (Fig. 3). We also symmetrize the template mesh by averaging each vertex position with that of the mirrored position of the vertex bilaterally opposite it. This is because we do not want to introduce any facial feature asymmetries of the template into the Laplacian shape prior. For operations relating the template back to its multi-view photographs, we use the original vertex positions. For operations employing the template as a Laplacian shape prior, we employ the symmetrized vertex positions.

We demonstrate initialization using a high-quality artist mesh template constructed from multi-view photography. We use the freely available "Digital Emily" mesh, photographs, and camera

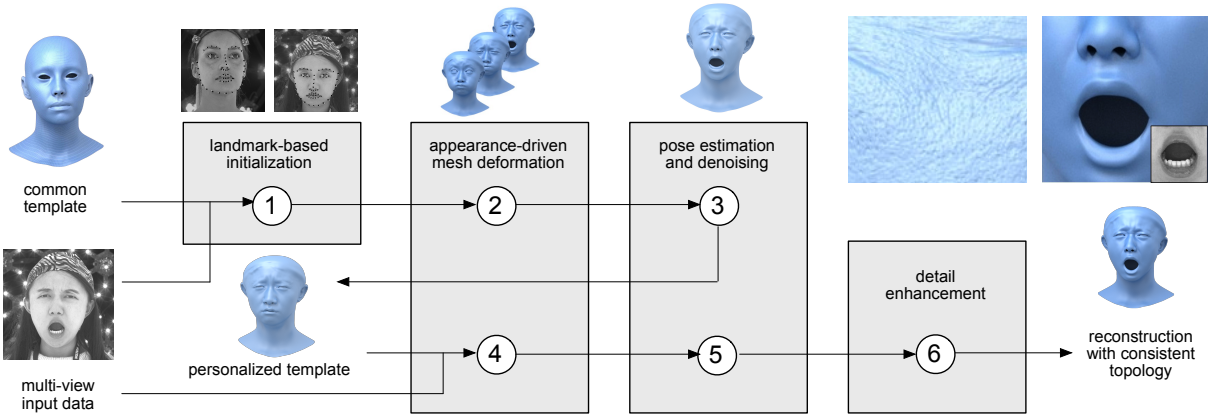
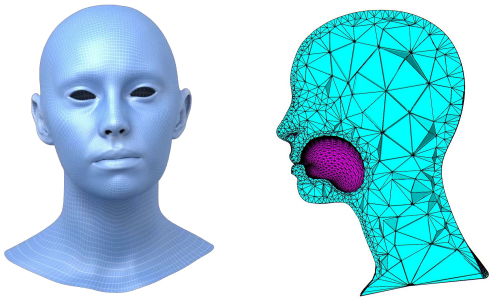
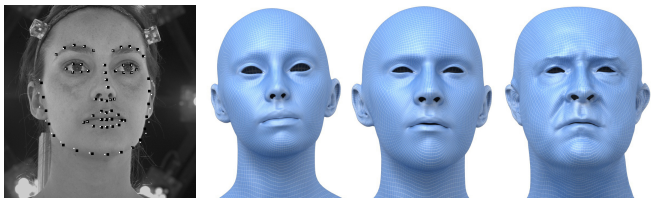


Figure 2: *Our pipeline proceeds in six phases, illustrated as numbered circles. 1) A common template is fitted to multi-view imagery of a subject using landmark-based fitting (Section 4.1). 2) The mesh is refined for every frame using optical flow for coarse-scale consistency and stereo (Section 4.2). 3) The meshes of all frames are aligned and denoised using a PCA scheme (Section 4.3). 4) A personalized template is extracted and employed to refine the meshes for fine-scale consistency (Section 4.4). 5) Final pose estimation and denoising reduces "sizzling" (Section 4.5). 6) Details are estimated from the imagery (Section 4.6).*



(a) (b)

Figure 3: *Production-quality mesh template and the cross-section of the volumetric template constructed from the surface.*



(a) (b) (c) (d)

Figure 4: *(a) Facial landmarks detected on the source subject and (b) the corresponding template; (c) The template deformed based on detected landmarks on the template and subject photographs; (d) Detailed template fitting based on optical flow between the template and subject, and between views.*

calibration from [Lea15]. The identity of the template is of no significance, though we purposely chose a template with no extreme unique facial features. A single template can be reused for many recordings of different subjects. We also compare to results obtained from a morphable model [TZN*15] with synthetic renderings in place of multi-view photographs.

4. Method Overview

Given an existing template mesh, we can reconstruct multiple video performances by optimizing photoconsistency cues between different views, across different expressions, and across different subjects. Our method consists of six sequential phases, illustrated in Fig. 2. Some phases share the same underlying algorithm, therefore in this section we provide a short overview of each phase, and then provide further technical details in Sections 5 and 6. We report run times for each phase based on computers with dual 8-core Intel E5620 processors and NVidia GTX980 graphics cards. All phases except for rigid alignment are trivially parallelizable across frames.

4.1. Landmark-Based Initialization

First, we leverage 2D facial landmark detection to deform the common template and compute an initial mesh for each frame of the performance. Subsequent optical flow steps require a mesh estimate which is reasonably close to the true shape. We estimate facial landmark positions on all frames and views using the method of [KS14] implemented in the *DLib* library [Kin09]. We then triangulate 3D positions with outlier rejection, as the landmark detection can be noisy. We use the same procedure for the template photographs to locate the template landmark positions. Fig. 4(a) shows an example with detected landmarks as black dots, and triangulated landmarks after outlier rejection as white dots. We transform the 3D landmarks of all poses to a common coordinate system using an approximate rigid registration to the template landmarks. We perform PCA-based denoising per subject in the registered space to remove any isolated errors, and then transform the landmarks back into world space. We additionally apply Gaussian smoothing to the landmark trajectories in each performance sequence. We finally compute a smooth deformation of the template to non-rigidly register it to the world space 3D landmarks of each captured facial pose, using Laplacian mesh deformation. Fig. 4(c) shows an exam-

Figure 5: *Facial expressions reconstructed without temporal flow.*

ple of the template deformed to a subject using only the landmarks. These deformed template meshes form the initial estimates in our pipeline, and are not required to be entirely accurate. This phase takes only a few seconds per frame, which are processed in parallel except for the PCA step.

4.2. Coarse-Scale Template Warping

Starting from the landmark-based initialization in Section 4.1, we employ an appearance-driven mesh deformation scheme to propagate the shared template mesh onto the performance frames. More details on this algorithm are provided in Section 5. This phase takes 25 minutes per frame, which are processed in parallel. (Most time is spent in the volumetric Laplacian solve.) After this phase, the processed facial meshes are all high quality 3D scans with the same topology, and are consistent at the level of coarse features such as the eyebrows, eyes, nostrils, and corners of the mouth. Fig. 5 shows the results at this phase directly deforming the template to multiple poses of the same individual without using any temporal information. Despite significant facial motion, the mesh topology remains consistent with the template. If only a single pose is desired for each subject, we can stop here. If sequences or multiple poses were captured, we continue with the remaining phases to improve consistency across poses.

4.3. Pose Estimation, Denoising, and Template Personalization

The face mesh estimates from Section 4.2 are reasonably good facial scans, but they exhibit two sources of distracting temporal noise. First, they lack fine-scale consistency in the UV domain, and second, any vertices that are extrapolated in place of missing data may differ considerably from frame to frame (for example, around the back of the head). The primary purpose of this phase is to produce a mesh sequence that is temporally smooth, with a plausible deformation basis, and closer to the true face sequence than the original estimate in Section 4.1. We wish to project the meshes into a reduced dimensional deformation basis to remove some of the temporal noise, which requires the meshes to be registered to a rigidly aligned head pose space, rather than roaming free in world space. Typically this is accomplished through iterative schemes, alternating between pose estimation and deformation basis estimation. In Section 6 we describe a method to decouple the pose from the deformation basis, allowing us to first remove the relative rotation from the meshes without knowledge of the deformation basis, then remove the relative translation, and finally compute the deformation basis via PCA. We truncate the basis retaining 95% of the variance, which reduces temporal noise without requiring frame-to-frame smoothing. Finally, we identify the frame whose shape is closest to the mean shape in the PCA basis, and let this frame be the personalized template frame for the proceeding phases. This phase, which is not easily parallelized, takes about 8 seconds per frame.

4.4. Fine-Scale Template Warping

This phase is nearly identical to Section 4.2, except that we propagate the shared template only to the personalized template frame identified in Section 4.3 (per subject), after which the updated personalized template becomes the new template for the remaining frames (again, per subject). This enables fine-scale consistency to be obtained from optical flow, as the pores, blemishes, and fine wrinkles on a subject's skin provide ample registration markers across poses. Further, we start from the denoised estimates from Section 4.3 instead of the landmark based estimates of Section 4.1, which are much closer to the actual face shape of each frame, reducing the likelihood of false matches in the optical flow.

4.5. Final Pose Estimation and Denoising

After the consistent mesh has been computed for all frames, we perform a final step of rigid registration to the personalized template and PCA denoising, similar to Section 4.3 but retaining 99% of the variance. We found this helps remove "sizzling" noise produced by variations in the optical flow. We also denoise the eye gaze animation using a simple Gaussian filter. More details on this phase are provided in Section 6.

4.6. Detail Enhancement

Finally we extract texture maps for each frame, and employ the high frequency information to enhance the surface detail already computed on the dense mesh in Section 5.3, in a similar manner as [BBB*10]. We make the additional observation that the sequence of texture maps holds an additional cue: when wrinkles appear on the face, they tend to make the surface shading darker relative to the neutral state. To exploit this, we compute the difference between the texture of each frame and the texture of the personalized template, and then filter it with an orientation-sensitive filter to remove fine pores but retain wrinkles. We call this the *wrinkle map*, and we employ it as a medium-frequency displacement, in addition to the high-frequency displacement obtained from a high-pass filter of all texture details. We call this scheme "darker is deeper", as opposed to the "dark is deep" schemes from the literature. Fig. 6 shows the dense mesh, enhanced details captured by our proposed technique, and details using a method similar to [BHB*11].This step including texture extraction and mesh displacement takes 10 minutes per frame and is trivially parallelizable.

5. Appearance-Driven Mesh Deformation

We now describe in detail the deformation algorithm mentioned in Sections 4.2 and 4.4. Suppose we have a known reference mesh with vertices represented as $y_i \in Y$ and a set of photographs $I_j^Y \in \mathcal{I}^Y$ corresponding to the reference mesh along with camera calibrations. Now suppose we also have photographs $I_k^X \in \mathcal{I}^X$ and camera calibrations for some other, unknown mesh with vertices represented as $x_i \in X$. Our goal is to estimate X given $Y, \mathcal{I}^Y, \mathcal{I}^X$. In other

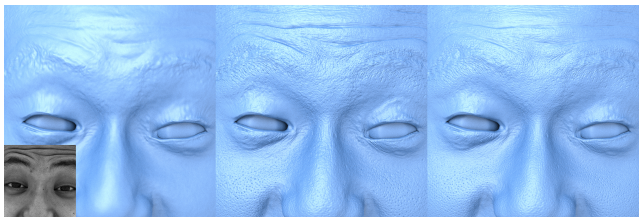


Figure 6: *(a) Dense base mesh; (b) Proposed detail enhancement; (c) "Dark is deep" detail enhancement.*

words, we propagate the known reference mesh Y to the unknown configuration X using evidence from the photographs of both. Suppose we have a previous estimate $\hat{X}$ somewhat close to the true X. (We explain how to obtain an initial estimate in Section 4.1.) We can improve the estimate $\hat{X}$ by first updating each vertex estimate $\hat{x}_i \in \hat{X}$ using optical flow (described in Section 5.2), then updating the entire mesh estimate $\hat{X}$ using Laplacian shape regularization with Y as a reference shape (described in Section 5.4). Finally we position the eyeballs based on flow vectors and geometric evidence from the eyelid region (described in Section 5.5).

5.1. Image Warping

Before further discussion, we must address the difficult challenge of computing meaningful optical flow between pairs of photographs that may differ in viewpoint, in facial expression, in subject identity, or any combination of the three. We assume high-resolution images and flat illumination, so different poses of a subject will have generally similar shading and enough fine details for good registration. Still, if the pose varies significantly or if the subject differs, naive optical flow estimation will generally fail. For example, Fig. 7(b, f) shows the result of naively warping one subject to another using optical flow, which would not be useful for facial correspondence since the flow mostly fails. Even in these cases, we desire a flow field that aligns coarse facial features even though the fine-scale features will lack meaningful matches, as in Fig. 7(d, h).

Our solution is to warp the image of one face to resemble the other face before computing optical flow (and vice-versa to compute optical flow in the other direction). We warp the images based on the current 3D mesh estimates (first obtained via the initialization in Section 4.1.) One might try rendering a synthetic image in the first camera view using the first mesh and texture sourced from the second image via the second mesh, to produce a warped version of the second image in a similar configuration to the first. However this approach would introduce artificial discontinuities wherever the current mesh estimates are not in precise alignment with the photographs, and such discontinuities would confuse the optical flow algorithm. Thus we instead construct a smooth vector field to serve as an image-space warp that is free of discontinuities. We compute this vector field by rasterizing the first mesh into the first camera view, but instead of storing pixel colors we write the second camera's projected image plane coordinates (obtained via the second mesh). We skip pixels that are occluded in either view using a z buffer for each camera, and smoothly interpolate the missing vectors across the entire frame. We then apply a small Gaussian blur

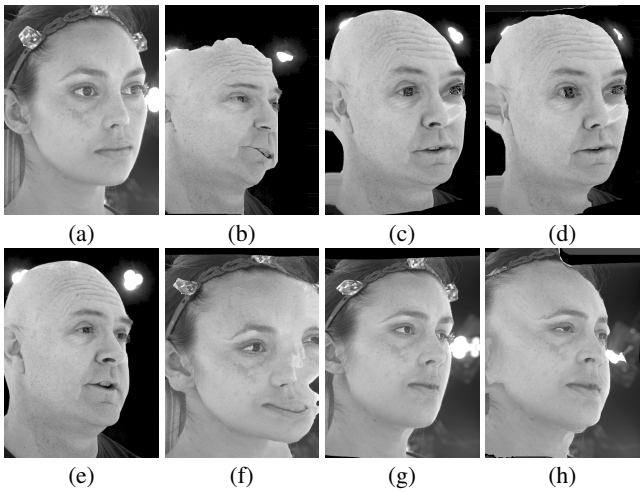


Figure 7: *Optical flow between photographs of different subjects (a) and (e) performs poorly, producing the warped images (b) and (f). Using 3D mesh estimates (e.g. a template deformed based on facial landmarks), we compute a smooth vector field to produce the warped images (c) and (g). Optical flow between the original images (a, e) and the warped images (c, g) produces the relatively successful final warped images (d) and (h).*

to slightly smooth any discontinuities, and finally warp the image using the smooth vector field. Examples using our warping scheme are shown in Fig. 7(c, g), and the shape is close enough to the true shape to produce a relatively successful optical flow result using the method of [WTP*09], shown in Fig. 7(d, h). After computing flow between the warped image and the target image, we concatenate the vector field warp and the optical flow vector field to produce the complete flow field. (This is implemented simply by warping the vector field using the optical flow field.)

5.2. Optical Flow Based Update

Within the set of images $\mathcal{I}^Y, \mathcal{I}^X$, we may find cues about the shape of the unknown mesh X and the relationship between X and the reference mesh Y. We employ a *stereo cue* between images from the same time instant, and a *reference cue* between images from different time instants or different subjects, using optical flow and triangulation (Fig. 8). First, consider a *stereo cue*. Given an estimate $p_i^k = P^k(\hat{x}_i) \approx P^k(x_i)$ with P^k representing the projection from world space to the image plane coordinates of view k of X, we can employ an optical flow field F_k^l between views k and l of X, to estimate $p_i^l = F_k^l(p_i^k) \approx P^l(x_i)$. Defining $\mathrm{D}^k(p_i^k) = \mathrm{I} - d^k(p_i^k)d^k(p_i^k)^\mathsf{T}$ where $d^k(p_i^k)$ is the world space view vector passing through image plane coordinate p_i^k of the camera of view k of X, and c^k the center of projection of the lens of the same camera (and likewise for l), we may employ p_i^k and p_i^l together to triangulate $\hat{x}_i$ as:

$$\hat{x}_i \leftarrow \underset{x_i}{\operatorname{argmin}} \left\| \mathrm{D}^k(p_i^k)(x_i - c^k) \right\|^2 + \left\| \mathrm{D}^l(p_i^l)(x_i - c^l) \right\|^2, \quad (1)$$

which may be solved in closed form. Next, consider a *reference cue*, which is a cue involving the reference mesh Y, being either a shared template or another pose of the same subject. Given a known or estimated $q_i^j = Q^j(y_i)$ with Q^j representing the projection from

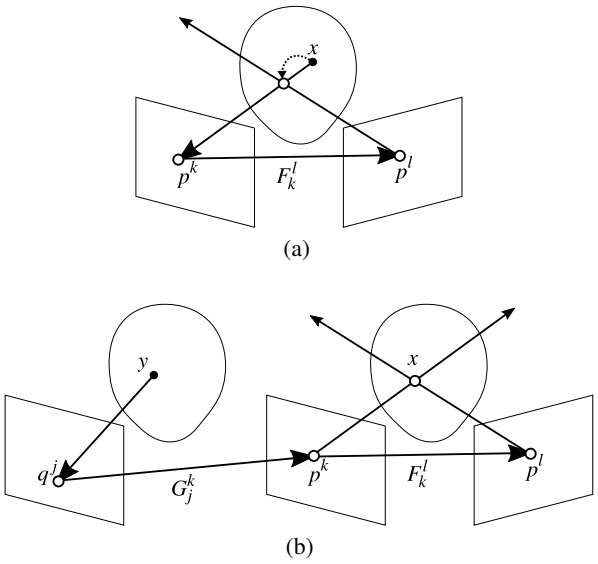


Figure 8: *(a) A stereo cue. An estimated point x is projected to the 2D point p^k in view k. A flow field F_k^l transfers the 2D point to p^l in a second view l. The point x is updated by triangulating the rays through p^k and p^l. (b) A reference cue. An estimated point y is projected to the 2D point q^j in view j. A flow field G_j^k transfers the 2D point to p^k in view k of a different subject or different time. A second flow field F_k^l transfers the 2D point to p^l in view l and then point x is estimated by triangulating the rays through p^k and p^l.*

world space to the image plane coordinates of view j of Y, we can employ an optical flow field G_j^k between view j of Y and view k of X, to estimate $p_i^k = G_j^k(q_i^j) \approx P^k(x_i)$. Next, we use F_k^l to obtain p_i^l from p_i^k as we did for the stereo cue, and triangulate $\hat{x}_i$ as before. However, instead of triangulating all these different cues separately, we combine them into a single triangulation, introducing a scalar field r_j^k representing the optical flow confidence for flow field G_j^k, and s_k^l representing the optical flow confidence for flow field F_k^l. Including one more parameter γ to balance between stereo cues and reference cues, the whole thing looks like this:

$$\begin{aligned}\hat{x}_i \leftarrow \operatorname*{argmin}_{x_i} \gamma \sum_{k,l} {s_k^l}^2(p_i^k)\Big(\Big\|\mathrm{D}^k(p_i^k)(x_i - c^k)\Big\|^2 \\ + \Big\|\mathrm{D}^l(F_k^l(p_i^k))(x_i - c^l)\Big\|^2\Big) \\ + (1-\gamma)\sum_{j,k,l} r_j^k(q_i^j) s_k^l(G_j^k(q_i^j))\Big(\Big\|\mathrm{D}^k(G_j^k(q_i^j))(x_i - c^k)\Big\|^2 \\ + \Big\|\mathrm{D}^l(F_k^l(G_j^k(q_i^j)))(x_i - c^l)\Big\|^2\Big). \quad (2)\end{aligned}$$

This differs from [FJA*14] in that reference flows are employed as a lookup into stereo flows instead of attempting to triangulate pairs of reference flows, and differs from [BHB*11] in that the geometry is not computed beforehand; rather the stereo and consistency are satisfied together. While (2) can be trivially solved in closed form, the flow field is dependent on the previous estimate $\hat{x}_i$, and hence we perform several iterations of optical flow updates interleaved with Laplacian regularization for the entire face. We schedule the parameter γ to range from 0 in the first iteration to 1 in the last iteration, so that the solution respects the reference most at the beginning, and respects stereo most at the end. We find five iterations to generally be sufficient, and we recompute the optical flow fields after the second iteration as the mesh will be closer to the true shape, and a better flow may be obtained.

The optical flow confidence fields r_j^k and s_k^l are vitally important to the success of the method. The optical flow implementation we use provides an estimate of flow confidence [WTP*09] based on the optical flow matching term, which we extend in a few ways. First, we compute flows both ways between each pair of images, and multiply the confidence by an exponentially decaying function of the round-trip distance. Specifically, in $s_k^l(p_i^k)$ we include a factor $\exp\left(-\kappa\left\|p_i^k - F_l^k(F_k^l(p_i^k))\right\|^2\right)$ (with normalized image coordinates), where $\kappa = 20$ is a parameter controlling round-trip strictness, and analogously for r_j^k. Since we utilize both directions of the flow fields anyway, this adds little computational overhead. For stereo flows (i.e. flows between views of the same pose) we include an additional factor penalizing epipolar disagreement, including in $s_k^l(p_i^k)$ the factor $\exp\left(-\lambda\ \mathrm{dl}^2(c^k, d^k(p_i^k), c^l, d^l(F_k^l(p_i^k)))\right)$, where $\mathrm{dl}(o_1, d_1, o_2, d_2)$ is the closest distance between the ray defined by origin o_1 and direction d_1 and the ray defined by origin o_2 and direction d_2, and $\lambda = 500$ is a parameter controlling epipolar strictness. Penalizing epipolar disagreement, rather than searching strictly on epipolar lines, allows our method to find correspondences even in the presence of noise in the camera calibrations. In consideration of visibility and occlusion, we employ the current estimate $\hat{X}$ to compute per-vertex visibility in each view of X using a z-buffer and back-face culling on the GPU, and likewise for each view of Y. If vertex i is not visible in view k, we set $s_k^l(p_i^k)$ to 0. Otherwise, we include a factor of $(-n_i \cdot d^k(p_i^k))^2$ to soften the visibility based on the current surface normal estimate n_i. We include a similar factor for view l in $s_k^l(p_i^k)$, and for view j of Y and view k of X in r_j^k. As an optimization, we omit flow fields altogether if the current estimated head pose relative to the camera differs significantly between the two views to be flowed. We compute the closest rigid transform between the two mesh estimates in their respective camera coordinates, and skip the flow field computation if the relative transform includes a rotation of more than twenty degrees.

5.3. Dense Mesh Representation

The optical flow fields used in Section 5.2 contain dense information about the facial shape, yet we compute our solution only on the vertices of an artist-quality mesh. It would be a shame to waste the unused information in the dense flow fields. Indeed we note that sampling the flow fields only at the artist mesh vertices in Section 5.2 introduces some amount of aliasing, as flow field values in between vertices are ignored. So we compute an auxiliary dense mesh with 262,144 vertices parameterized on a 512×512 vertex grid in UV space. Optical flow updates are applied to all vertices of the dense mesh as in Section 5.2. We then regularize the dense mesh using the surface Laplacian terms from Section 5.4, but omit the volumetric terms as they are prohibitive for such a dense mesh. The surface Laplacian terms, on the other hand, are easily expressed and solved on the dense grid parameterization.

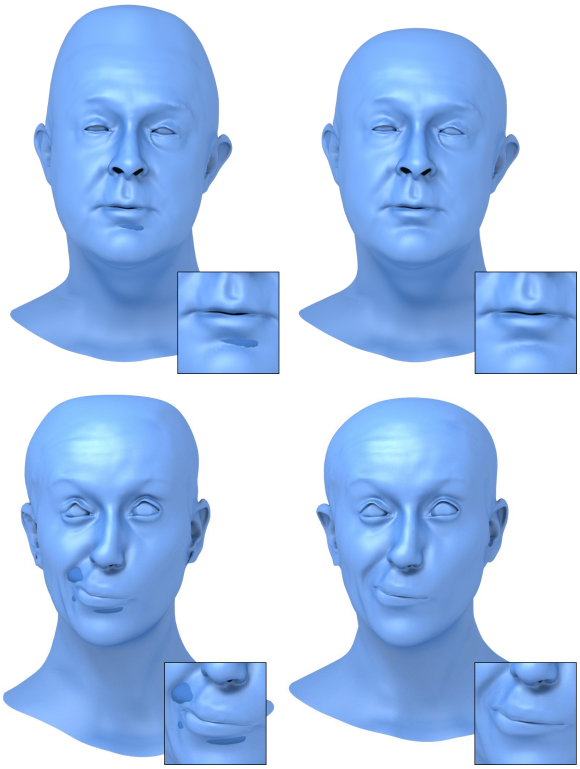


Figure 9: *Laplacian regularization results. Left: surface regularization only. Right: surface and volumetric regularization.*

This dense mesh provides two benefits. First, it provides an intermediate estimate that is free of aliasing, which we utilize by looking up the dense vertex position at the same UV coordinate as each artist mesh vertex. This estimate lacks volumetric regularization, but that is applied next in Section 5.4. Second, the dense mesh contains surface detail at finer scales than the artist mesh vertices, and so we employ the dense mesh in Section 4.6 as a base for detailed displacement map estimation.

5.4. Laplacian Regularization

After the optical flow update, we update the entire face mesh using Laplacian regularization, using the position estimates from Section 5.3 as a target constraint. We use the framework of [ZHS*05], wherein we update the mesh estimate as follows:

$$\hat{X} \leftarrow \underset{X}{\operatorname{argmin}} \sum_{i \in S} \alpha_i \|x_i - \hat{x}_i\|^2 + \sum_{i \in S} \|\mathscr{L}_S(x_i) - \varepsilon_i\|^2 + \beta \sum_{i \in V} \|\mathscr{L}_V(x_i) - \delta_i\|^2, \quad (3)$$

where $\alpha_i = \sigma(\gamma \sum_{k,l} s_k^{l\,2}(p_i^k) + (1-\gamma)\sum_{j,k,l} r_j^k(q_i^j) s_k^l(G_j^k(q_i^j)))$ is the constraint strength for vertex i derived from the optical flow confidence with $\sigma = 15$ being an overall constraint weight, S is the set of surface vertices, $\mathscr{L}_S$ is the surface Laplace operator, ε_i is the surface Laplacian coordinate of vertex i in the rest pose, V is the set of volume vertices, $\mathscr{L}_V$ is the volume Laplace operator, δ_i is the volume Laplacian coordinate of vertex i in the rest pose, and β is a parameter balancing $\mathscr{L}_S$ and $\mathscr{L}_V$ as in [ZHS*05]. In our framework, the rest pose is the common template in early phases, or the personalized template in later phases. We solve this sparse linear problem using the sparse normal Cholesky routines implemented in the Ceres solver [AMO]. We also estimate local rotation to approximate as-rigid-as-possible deformation [SA07]. We locally rotate the Laplacian coordinate frame to fit the current mesh estimate in the neighborhood of $\hat{x}_i$, and iterate the solve ten times to allow the local rotations to converge. Fig. 9 illustrates the effect of including the volumetric Laplacian term. While previous works employ only the surface Laplacian term [VWB*12], we find the volumetric term is vitally important for producing good results in regions with missing or occluded data such as the back of the head or the interior of the mouth, which are otherwise prone to exaggerated extrapolation or interpenetration.

5.5. Updating Eyeballs and Eye Socket Interiors

Our template represents eyeballs as separate objects from the face mesh, with their own UV texture parameterizations. We include the eyeball vertices in the optical flow based update, but not the Laplacian regularization update. Instead, the eyes are treated as rigid objects, using the closest rigid transform to the updated positions to place the entire eyeball. This alone does not produce very good results, as the optical flow in the eye region tends to be noisy, partly due to specular highlights in the eyes. To mitigate this problem, we do two things. First, we apply a 3×3 median filter to the face images in the region of the eye, using the current mesh estimate to rasterize a mask. Second, after each Laplacian regularization step, we consider distance constraints connecting the eye pivot points e_0 and e_1 to the vertices on the entire outer eyelid surfaces, and additional distance constraints connecting the eye pivot points to the vertices lining the inside of the eye socket. These additional constraints appear as: $\sum_{j=0}^{1} \left[\sum_{i \in E_j \cup O_j} (\|x_i - e_j\| - \|y_i - e_j^Y\|)^2 + \sum_{i \in E_j} \phi(\|x_i - e_j\|, r)\right]$, where E_0 and E_1 are the set of left and right eyelid vertices, O_0 and O_1 are the set of left and right socket vertices, e_0^Y and e_1^Y are the eye pivot positions in the reference mesh, ϕ is a distance constraint with a non-penetration barrier defined as $\phi(a,b) = (a-b)^2 \exp(\rho(b-a))$ with $\rho = 10$ being a parameter controlling barrier fall-off, and r is a hard constraint distance representing the radius of the eyeball plus the minimum allowed thickness of the eyelid. The target distance of each constraint is obtained from the reference mesh. We minimize an energy function including both (3) and the eye pivot distance constraints, but update only the eye pivots, interior volume vertices, and vertices lining the eye sockets, leaving the outer facial surface vertices constant. The distance constraints render this a nonlinear problem, which we solve using the sparse Levenberg-Marquardt routines implemented in the Ceres solver [AMO]. After the eye pivots are computed, we compute a rotation that points the pupil towards the centroid of the iris vertex positions obtained in the optical flow update. Although this scheme does not personalize the size and shape of the eyeball, there is less variation between individuals in the eye than in the rest of the face, and we obtain plausible results.

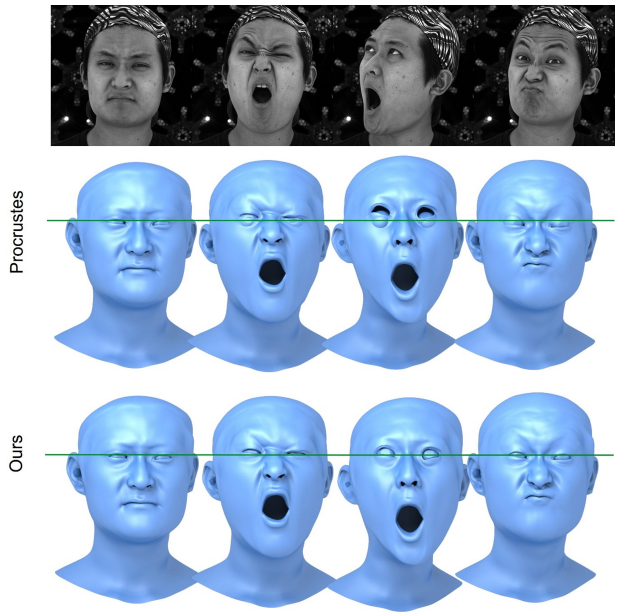


Figure 10: *Comparison of rigid mesh alignment techniques on a sequence with significant head motion and extreme expressions. Top: center view. Middle: Procrustes alignment. Bottom: our proposed method. The green horizontal strike-through lines indicate the vertical position of the globally consistent eyeball pivots.*

6. PCA-Based Pose Estimation and Denoising

We next describe the pose estimation and denoising algorithm mentioned in Sections 4.3 and 4.5. Estimating the rigid transformation for each frame of a performance serves several purposes. First, it is useful for representing the animation results in standard animation packages. Second, it allows consistent global 3D localization of the eyeballs, which should not move with respect to the skull. Third, it enables PCA-based mesh denoising techniques that reduce high-frequency temporal noise and improve consistency of occluded or unseen regions that are inferred from the Laplacian regularization. A rigid stabilization technique could be employed such as [BB14] (also employed in [WBGB16]), which involves fitting an anatomical skull template and skin thickness models. Instead we estimate rigid rotation without any anatomical knowledge, and then estimate rigid translation using a simple eyelid thickness constraint. Section 6.1 describes our novel rotation alignment algorithm for deformable meshes which does not require knowledge of the deformation basis. Section 6.2 describes our novel translation alignment algorithm that simultaneously estimates per-mesh translation and globally consistent eyeball pivot placement. Section 6.3 then describes a straightforward PCA dimension reduction scheme.

We compare our rigid alignment technique to Procrustes analysis [Gow75] as a baseline. Fig. 10 shows several frames from the evaluation of our technique on a sequence with significant head motion and extreme facial expressions including wide open mouth. The baseline method exhibits significant misalignment especially on wide open mouth expressions, which becomes more apparent when globally consistent eye pivots are included. Our proposed technique stabilizes the rigid head motion well, enabling globally consistent eye pivots to be employed without interpenetration.

6.1. Rotation Alignment

We represent the set of facial meshes as a $3N \times M$ matrix $\mathbf{X}$, where each column of $\mathbf{X}$ is a $3N$ dimensional vector $X^t, t = 1 \ldots M$, representing the interleaved x, y, and z coordinates of the N vertices of mesh t. We assume a low-dimensional deformation basis, and hence $\mathbf{X} = \mathbf{B}\mathbf{W}$ in the absence of rigid transformations, where $\mathbf{B}$ is a $3N \times K$ basis matrix with $K \ll M$, and $\mathbf{W}$ is a $K \times M$ weight matrix with columns W^t corresponding to the basis activations of each X^t. (We do not separately add the mean mesh in the deformation model, so it will be included as the first column of $\mathbf{B}$.) The trouble is that each X^t may actually have a different rigid transform, so that $X^t = \mathbf{R}^t \mathbf{B} W^t + T^t$ for some unknown rotation $\mathbf{R}^t$ and translation T^t, rendering $\mathbf{B}$ and W^t difficult to discover. Translation can be factored out of the problem by analyzing the mesh edges $\vec{X}^t$ rather than the vertices X^t, defining $\tilde{\mathbf{B}}$ appropriately so that $\vec{X}^t = \mathbf{R}^t \tilde{\mathbf{B}} W^t$, and defining a matrix $\tilde{\mathbf{X}}$ having columns $\vec{X}^t$. This eliminates T^t, however the rotations $\mathbf{R}^t$ still obfuscate the solution.

To solve this, we first roughly align the meshes using Procrustes method so that any remaining relative rotations are small. Recall that for rotations of magnitude $O(\varepsilon)$, composing rotations is equal to summing rotations up to an $O(\varepsilon^2)$ error, as in the exponential map $\mathbf{R}^t = \mathbf{I} + r_x^t \mathbf{G}_x + r_y^t \mathbf{G}_y + r_z^t \mathbf{G}_z + O(\left\| r^t \right\|^2)$, where r^t is the Rodrigues vector corresponding to $\mathbf{R}^t$ and $\mathbf{G}_x$, $\mathbf{G}_y$, $\mathbf{G}_z$ are the generator functions for the SO_3 matrix lie group:

$$\mathbf{G}_x = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}, \mathbf{G}_y = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}, \mathbf{G}_z = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{4}$$

Defining $\mathbf{G}_x^*$ as a block diagonal matrix with $\mathbf{G}_x$ repeated along the diagonal (and likewise for y, z), we construct the $3N \times 4M$ matrix $\hat{\mathbf{X}} = \begin{bmatrix} \tilde{\mathbf{X}} & \mathbf{G}_x^* \tilde{\mathbf{X}} & \mathbf{G}_y^* \tilde{\mathbf{X}} & \mathbf{G}_z^* \tilde{\mathbf{X}} \end{bmatrix}$, forming a basis that spans the set of meshes (and hence deformation) as well as the local rotation neighborhood, up to the $O(\left\| r^t \right\|^2)$ error mentioned previously. Because the last three block columns of this matrix represent small rotation differentials, and composition of small rotations is linear, they lie in the same subspace as any rotational components of the first block column, and therefore performing column-wise principal component analysis on this matrix without mean subtraction separates deformation and rotation bases, as $\hat{\mathbf{X}} = \hat{\mathbf{B}}\hat{\mathbf{W}}$. There are at most M deformation bases in $\hat{\mathbf{B}}$ and three times as many rotation bases, so we can assume that 3 out of 4 bases are rotational but need to identify which ones. To do this, we score each basis with a *data weight* and a *rotation weight*. The data weight for the basis at column b in $\hat{\mathbf{B}}$ is the sum of the squares of the coefficients in the first M columns of row b of $\hat{\mathbf{W}}$, and the rotation weight is the sum of the squares of the coefficients in the last $3M$ columns of row b of $\hat{\mathbf{W}}$. The rotational score of column b is then the rotation weight divided by the sum of the data weight and rotation weight. We assume the $3M$ columns with the greatest such score represent rotations of meshes and rotations of deformation bases. The remaining M columns form a basis with deformation only (up to $O(\left\| r^t \right\|^2)$), which we call the rotation suppressed basis $\tilde{\mathbf{B}}_s$, and use it to suppress rotation by projecting $\tilde{\mathbf{X}} \leftarrow \tilde{\mathbf{B}}_s \tilde{\mathbf{B}}_s^{\mathsf{T}} \tilde{\mathbf{X}}$. This may also contain a residual global rotation, so we compute another rigid alignment between the mean over all $\vec{X}^t$ and (the edges of) our template mesh,

and apply this rotation to update all $\vec{X}^t$, making convergence possible. We iterate the entire procedure starting from the construction of $\hat{\mathbf{X}}$ until convergence, which we usually observed in 10 to 20 iterations. Finally, we compute the closest rigid rotation between each original $\vec{X}^t$ and the final rotation suppressed $\vec{\hat{X}}^t$ to discover $\mathbf{R}^t$.

6.2. Translation Alignment

After $\mathbf{R}^t$ and W^t are computed using the method in Section 6.1, there remains a translation ambiguity in obtaining $\mathbf{B}$. We define the *rotationally aligned mesh* $A^t = {\mathbf{R}^t}^\mathsf{T} X^t$, and we define the aligned-space translation $\tau^t = {\mathbf{R}^t}^\mathsf{T} T^t$, thus $A^t = \mathbf{B}W^t + \tau^t$. We compute the mean of A^t and compute and apply the closest rigid translation to align it with the template mesh, denoting the result $\bar{A}^t$. We then wish to discover $\tau_t, t = 1 \ldots M$ such that each $A^t - \tau_t$ is well aligned to $\bar{A}^t$. Since our model has eyes, we also wish to discover globally consistent eye pivot points in an aligned head pose space, which we call $\bar{e}_0$ and $\bar{e}_1$, and allow the eyes to move around slightly relative to the facial surface, while being constrained to the eyelid vertices using the same distance constraints as in Section 5.5, in order to achieve globally consistent pivot locations. We find $\bar{e}_0$, $\bar{e}_1$, and $\tau_t, t = 1 \ldots M$ minimizing the following energy function using the Ceres solver [AMO]:

$$\sum_{t=1}^{M} \Big[\sum_{i \in S} \psi(\left\| a_i^t - \tau_t - \bar{a}_i^t \right\|) + \sum_{j=0}^{1} \sum_{i \in E_j} \phi(\left\| a_i^t - \tau_t - \bar{e}_j \right\|, \left\| y_i - e_j^Y \right\|) \Big], \quad (5)$$

where a_i^t is a vertex in mesh A^t (and $\bar{a}_i^t$ in $\bar{A}^t$), and ψ is the Tukey biweight loss function tuned to ignore cumulative error past 1 cm. With τ^t computed, we let $T^t = \mathbf{R}^t \tau^t$.

6.3. Dimension Reduction

With $\mathbf{R^t}$ and T^t computed for all meshes in Sections 6.1 and 6.2, we may remove the relative rigid transforms from all meshes to place them into an aligned pose space. We perform a weighted principle component analysis, with vertices weighted by the mean of the confidence α_i (see Section 5.4), producing the basis $\mathbf{B}$ and weight matrix $\mathbf{W}$. We truncate the basis to reduce noise and inconsistencies across poses in areas of ambiguous matching to the shared template, and in areas of insufficient data that are essentially inferred by the Laplacian prior, such as the back and top of the head.

7. Results

We demonstrate our method for dynamic facial reconstruction with five subjects - three male and two female. The first four subjects were recorded in a LED sphere under flat-lit static blue lighting. The blue light gives us excellent texture cues for optical flow. We used 12 Ximea monochrome 2048×2048 machine vision cameras as seen in Fig. 11. We synchronized the LEDs with cameras at 72Hz, and only exposed the camera shutter for 2ms to eliminate motion blur as much as possible. Also pulsing the LEDs for a shorter period of time reduces the perceived brightness to the subject, and is more suitable for recording natural facial performance.

Figure 11: *12 views of one frame of a performance sequence. Using flat blue lighting provides sharp imagery.*

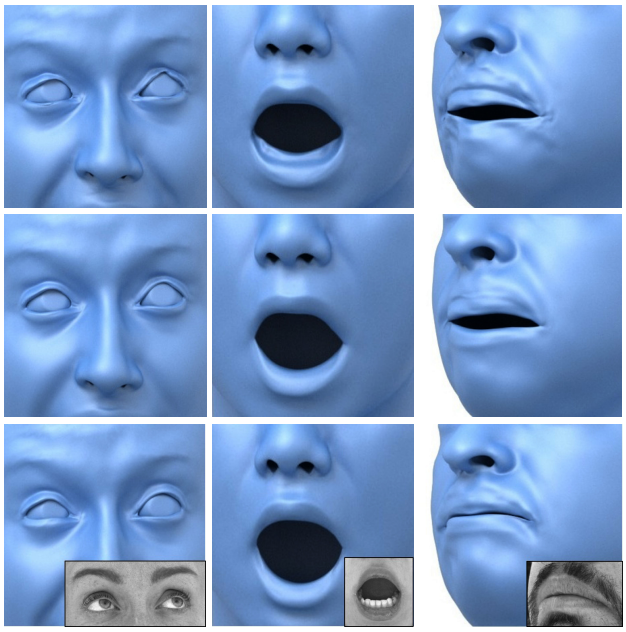


Figure 12: *Zoomed renderings of different facial regions of three subjects. Top: results after coarse-scale fitting using the shared template (Section 4.2). The landmarks incorrectly located the eyes or mouth in some frames. Middle: results after pose estimation and denoising (4.3). Bottom: results after fine-scale consistent mesh propagation (4.4) showing the recovery of correct shapes.*

Though we captured 72Hz, we only processed 24Hz in the results, to reduce computation time.

Fig. 12 shows intermediate results from each reconstruction step described in Sections 4.2 through 4.4. This is a particularly challenging case as the initial facial landmark detector matches large-scale facial proportions but struggles in the presence of facial hair that partially occludes the lips and teeth. Artifacts remain even after dense optical flow in Section 4.2. PCA based denoising (4.3) and fine-scale consistent mesh propagation (4.4) fill in more accurate mouth and eye contours that agree with inset photographs.

An alternative approach would be to initialize image-based mesh warping with a morphable model [TZN*15] in place of the Digi-

tal Emily template (Fig. 13). To perform this comparison, we fitted a morphable model to multi-view imagery of a male subject (Fig. 14(a)) and filled in the rest of the head topology using Laplacian mesh deformation (Fig. 13(b)). We then generated synthetic renderings of the morphable model using estimated morphable albedo and inpainted textures. Our technique reconstructs more geometric details such as the nasolabial fold on Fig. 13(c) which are not captured well by the linear deformation model in Fig. 13(a), perhaps because our method employs stereo cues from multi-view data. Fig. 14(b) illustrates the warped image to match (e) in the same camera, compared to the warped image (d) starting from the Digital Emily model (c) in a similar camera view.

Our final geometry closely matches fine details in the original photographs. Fig. 15 shows directly deformed artist quality topology (a) and a captured detail layer (b) in a calibrated camera view. Overlaying half the face onto the original photograph shows good agreement for details such as forehead wrinkles.

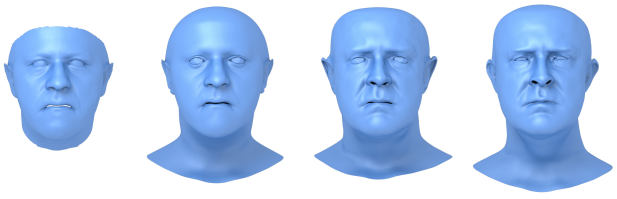
Fig. 16 shows several frames from a highly expressive facial performance reconstructed on a high-quality template mesh using our pipeline. The reconstructed meshes shown in wire-frame with and without texture mapping indicate good agreement with the actual performance as well as texture consistency across frames. The detail enhancement from Section 4.6 produces high-resolution dynamic details such as pores, forehead wrinkles, and crows feet, adding greater fidelity to the geometry.

We directly compare our method with [BHB*11] based on publicly available video datasets as shown in Fig. 17. Our method is able to recover significantly greater skin detail and realistic facial features particularly around the mouth, eyes, and nose, as well as completing the full head. Our system also does not rely on temporal flow making it easier to parallelize each frame independently for faster processing times. Fig. 18 also illustrates the accuracy of our method compared to the multi-view reconstruction method of [FP10]. Though we never explicitly compute a point cloud or depth map, the optical flow computation is closely related to stereo correspondence and our result is a very close match to the multi-view stereo result (as indicated by the speckle pattern apparent when overlaying the two meshes with different colors). Unlike [FP10] our technique naturally fills in occluded or missing regions such as the back of the head and provides consistent topology across subjects and dynamic sequences. We can reconstruct a single static frame as shown in Fig. 5 or an entire consistent sequence.

Since our method reconstructs the shape and deformation on a consistent UV space and topology, we can transfer attributes such as appearance or deformation between subjects. Fig. 19 shows morphing between facial performances of three subjects, with smooth transition from one subject to the next. Unlike previous performance transfer techniques, the recovered topology is inherent to the reconstruction and does not require any post processing.

8. Limitations

While our technique yields a robust system and provides several benefits compared to existing techniques, it has several limitations. Initial landmark detection may incorrectly locate a landmark, for example sometimes facial hair will be interpreted as a mouth. Improvements in landmark detection would help here. The coarse-scale template alignment fails in some areas when the appearance of the subject and the template differ significantly, which can happen in the presence of facial hair or when the tongue and teeth become visible as they are not part of the template (see Fig. 20). While these errors are often mitigated by our denoising technique, in the future it would be of interest to improve tracking in such regions by providing additional semantics such as more detailed facial feature segmentation and classification, or by combining tracking from more than one template to cover a larger appearance space. While our appearance-driven mesh deformation warps the image and deforms the personalized template progressively closer to the solu-



(a) (b) (c) (d)

Figure 13: *Comparison using a morphable model of [TZN*15] as a initial template. (a) Front face region captured by the previous technique, (b) stitched on our full head topology. (c) Resulting geometry from 4.2 deformed using our method with (b) as a template, compared to (d) the result of using the Digital Emily template. The linear morphable model misses details in the nasolabial fold.*

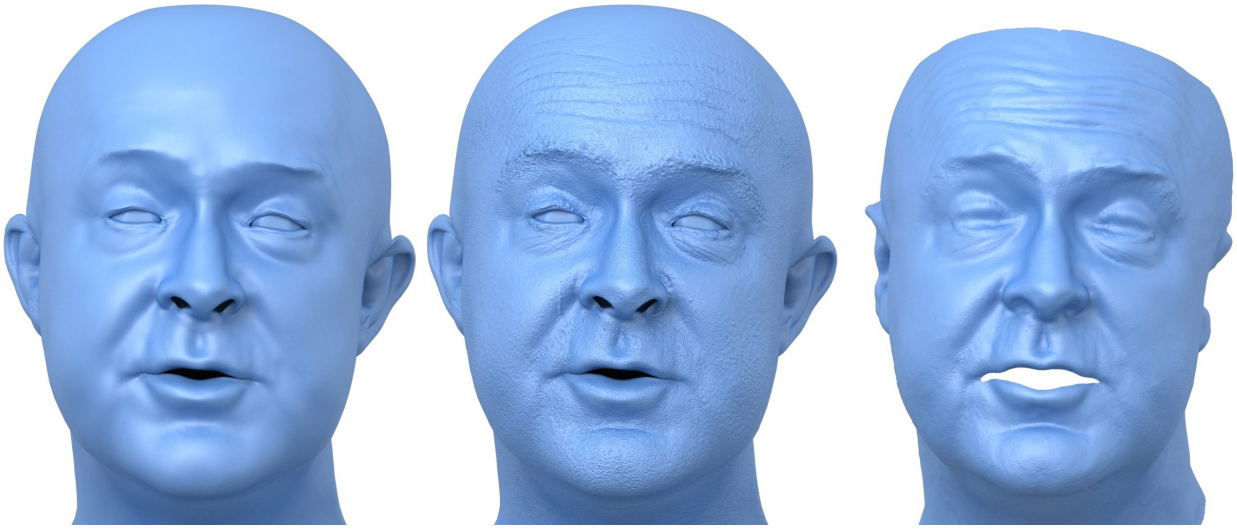
(a) (b) (c) (d) (e)

Figure 14: *(a) Synthetic rendering of the morphable model from Fig. 13(b). (b) Result using our image warping method to warp (a) to match real photograph (e). Similarly the common template image (c) is warped to match (e), producing plausible coarse-scale facial feature matching in (d).*

(a) (b)

Figure 15: *(a) High quality mesh deformed using our technique; (b) high-resolution displacement details. Half-face overlay visualization indicates good agreement and topology flow around geometric details within the facial expression.*

Figure 16: *Dynamic face reconstruction from multi-view dataset of a male subject shown from one of the calibrated cameras (top). Wireframe rendering (second) and per frame texture rendering (third) from the same camera. Enhanced details captured with our technique (bottom) shows high quality agreement with the fine-scale details in the photograph.*



(a) (b) (c)

Figure 17: *Reconstructed mesh (a), enhanced displacement details with our technique (b), and comparison to previous work (c). Our method automatically captures whole head topology including nostrils, back of the head, mouth interior, and eyes, as well as skin details.*

(a) (b) (c)

Figure 18: *Comparison of our result (a) to PMVS2 [FP10] (b). Overlaying the meshes in (c) indicates a good geometric match.*

tion, registration error could still occur under significant appearance change. This could particularly occur around the eyes and the mouth due to occlusion. Previous works are also susceptible to such occlusion artifacts. Our facial surface details come from dynamic high-frequency appearance changes in the flat-lit video, but as with other passive illumination techniques, they miss some of the facial texture realism obtainable with active photometric stereo processes such as in [GFT*11]. If such an active-illumination scan of the subject could be used as the template mesh, our technique could propagate its high-frequency details to the entire performance, and dynamic skin microgeometry could be simulated as as in [NFA*15]. Furthermore, it would be of interest to allow the animator to conveniently modify the captured performances; this could be facilitated by identifying sparse localized deformation components as in [NVW*13] or performance morphing techniques as in [MBW*15].

9. Discussion

We have presented an entirely automatic method to accurately track facial performance geometry from multi-view video, producing consistent results on an artist-friendly mesh for multiple subjects from a single template. Unlike previous works that employ temporal optical flow, our approach simultaneously optimizes stereo and consistency objectives independently for each instant in time. We demonstrated an appearance-driven mesh deformation algorithm that leverages landmark detection and optical flow techniques, which produces coarse-scale facial feature consistency across subjects and fine-scale consistency across frames of the same subject. We also demonstrated a displacement map estimation scheme that compares the uv-space texture of each frame against an automatically selected neutral frame to produce stronger displacements in dynamic facial wrinkles. Our method operates solely in the desired artist mesh domain and does not rely on complex facial rigs or morphable models. While performance retargeting is beyond the scope of this work, performances captured using our proposed pipeline could be employed as high-quality inputs into retargeting systems such as [LYYB13]. To our knowledge, this is the first method to produce facial performance capture results with detail on par with multi-view stereo and pore-level consistent parameterization without temporal optical flow, and could lead to interesting applications in building databases of morphable characters and simpler facial performance capture pipelines.

Acknowledgements

The authors thank the following people: Adair Liu and Chloe LeGendre for sitting as subjects, Xueming Yu for Light Stage programming, Chris Ellis, Etienne Danvoye, and Javier von der Pahlen for the Zoe camera software, and Kathleen Haase, Christina Trejo, and Randall Hill for their support and assistance. We thank the authors of Beeler et al. [BHB*11] for graciously providing data for comparisons. This research is supported by the Google PhD Fellowship, and in part by the Funai Foundation for Information Technology, Adobe, Oculus, Facebook, Huawei, the Google Faculty Research Award, the Okawa Foundation Research Grant, the Office of Naval Research (ONR) / U.S. Navy under award number N00014-15-1-2639, the Office of the Director of National Intelligence (ODNI) and Intelligence Advanced Research Projects Activity (IARPA) under contract number 2014-14071600010, and the U.S. Army Research Laboratory (ARL) under contract W911NF-14-D-0005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ONR, ODNI, IARPA, ARL, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

References

[AFB*13] ALEXANDER O., FYFFE G., BUSCH J., YU X., ICHIKARI R., GRAHAM P., NAGANO K., JONES A., DEBEVEC P., ALTER J., JIMENEZ J., DANVOYE E., ANTIONAZZI B., EHELER M., KYSELA Z., WU X.-C., VON DER PAHLEN J.: Digital Ira: High-resolution facial performance playback. In *ACM SIGGRAPH 2013 Computer Animation Festival* (New York, NY, USA, 2013), SIGGRAPH '13, ACM, pp. 1–1. URL: http://doi.acm.org/10.1145/2503541.2503641, doi:10.1145/2503541.2503641. 3

[AMO] AGARWAL S., MIERLE K., OTHERS: Ceres solver. http://ceres-solver.org. 8, 10

[ARL*09] ALEXANDER O., ROGERS M., LAMBETH W., CHIANG M., DEBEVEC P.: Creating a photoreal digital actor: The digital emily project. In *Visual Media Production, 2009. CVMP '09. Conference for* (Nov 2009), pp. 176–187. doi:10.1109/CVMP.2009.29. 3

[BB14] BEELER T., BRADLEY D.: Rigid stabilization of facial expressions. *ACM Trans. Graph. 33*, 4 (July 2014), 44:1–44:9. URL: http://doi.acm.org/10.1145/2601097.2601182, doi:10.1145/2601097.2601182. 2, 9

[BBB*10] BEELER T., BICKEL B., BEARDSLEY P., SUMNER B., GROSS M.: High-quality single-shot capture of facial geometry. *ACM Trans. Graph. 29*, 4 (July 2010), 40:1–40:9. URL: http://doi.acm.org/10.1145/1778765.1778777, doi:10.1145/1778765.1778777. 3, 5

[BGY*13] BHAT K. S., GOLDENTHAL R., YE Y., MALLET R., KOPERWAS M.: High fidelity facial animation capture and retargeting with contours. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (New York, NY, USA, 2013), SCA '13, ACM, pp. 7–14. URL: http://doi.acm.org/10.1145/2485895.2485915, doi:10.1145/2485895.2485915. 3

[BHB*11] BEELER T., HAHN F., BRADLEY D., BICKEL B., BEARDSLEY P., GOTSMAN C., SUMNER R. W., GROSS M.: High-quality passive facial performance capture using anchor frames. In *ACM SIGGRAPH 2011 papers* (New York, NY, USA, 2011), SIGGRAPH '11, ACM, pp. 75:1–75:10. URL: http://doi.acm.org/10.1145/1964921.1964970, doi:10.1145/1964921.1964970. 2, 3, 5, 7, 11, 13

[BHPS10] BRADLEY D., HEIDRICH W., POPA T., SHEFFER A.: High resolution passive facial performance capture. *ACM Trans. Graph. 29*, 4 (July 2010), 41:1–41:10. URL: http://doi.acm.org/10.1145/1778765.1778778, doi:10.1145/1778765.1778778. 3

Figure 19: *Since our method reconstructs the face on a common head topology with coarse-scale feature consistency across subjects, blending between different facial performances is easy. Here we transition between facial performances from three different subjects.*

(a) (b) (c) (d)

Figure 20: *Our system struggles to reconstruct features that are not represented by the template. For example, visible facial hair or tongue (a) may cause misplacement of the landmarks employed in 4.2 (b), which the denoising in 4.3 may not be able to recover (c), and remain as artifacts after fine-scale warping in 4.4 (d).*

[BPL*05] BORSHUKOV G., PIPONI D., LARSEN O., LEWIS J. P., TEMPELAAR-LIETZ C.: Universal capture - image-based facial animation for "the matrix reloaded". In *ACM SIGGRAPH 2005 Courses* (New York, NY, USA, 2005), SIGGRAPH '05, ACM. URL: http://doi.acm.org/10.1145/1198555.1198596, doi:10.1145/1198555.1198596. 2

[BV99] BLANZ V., VETTER T.: A morphable model for the synthesis of 3d faces. In *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (New York, NY, USA, 1999), ACM Press/Addison-Wesley Publishing Co., pp. 187–194. doi:http://doi.acm.org/10.1145/311535.311556. 3

[BWP13] BOUAZIZ S., WANG Y., PAULY M.: Online modeling for realtime facial animation. *ACM Trans. Graph. 32*, 4 (July 2013), 40:1–40:10. URL: http://doi.acm.org/10.1145/2461912.2461976, doi:10.1145/2461912.2461976. 3

[CBE*15] CONG M., BAO M., E J. L., BHAT K. S., FEDKIW R.: Fully automatic generation of anatomical face simulation models. In *Proceedings of the 14th ACM SIGGRAPH / Eurographics Symposium on Computer Animation* (New York, NY, USA, 2015), SCA '15, ACM, pp. 175–183. URL: http://doi.acm.org/10.1145/2786784.2786786, doi:10.1145/2786784.2786786. 3

[CBZB15] CAO C., BRADLEY D., ZHOU K., BEELER T.: Real-time high-fidelity facial performance capture. *ACM Trans. Graph. 34*, 4 (July 2015), 46:1–46:9. URL: http://doi.acm.org/10.1145/2766943, doi:10.1145/2766943. 3

[CHZ14] CAO C., HOU Q., ZHOU K.: Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph. 33*, 4 (July 2014), 43:1–43:10. URL: http://doi.acm.org/10.1145/2601097.2601204, doi:10.1145/2601097.2601204. 3

[CSDV99] CHARETTE P., SAGAR M., DECAMP G., VALLOT J.: The jester. In *ACM SIGGRAPH 99 Electronic Art and Animation Catalog* (New York, NY, USA, 1999), SIGGRAPH '99, ACM, pp. 151–. URL: http://doi.acm.org/10.1145/312379.312968, doi:10.1145/312379.312968. 3

[CWW*16] CAO C., WU H., WENG Y., SHAO T., ZHOU K.: Real-time facial animation with image-based dynamic avatars. *ACM Trans. Graph. 35*, 4 (July 2016), 126:1–126:12. URL: http://doi.acm.org/10.1145/2897824.2925873, doi:10.1145/2897824.2925873. 3

[FJA*14] FYFFE G., JONES A., ALEXANDER O., ICHIKARI R., DEBEVEC P.: Driving high-resolution facial scans with video performance capture. *ACM Trans. Graph. 34*, 1 (Dec. 2014), 8:1–8:14. URL: http://doi.acm.org/10.1145/2638549, doi:10.1145/2638549. 3, 7

[FP09] FURUKAWA Y., PONCE J.: Dense 3d motion capture for human faces. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA* (2009), pp. 1674–1681. URL: http://dx.doi.org/10.1109/CVPRW.2009.5206868, doi:10.1109/CVPRW.2009.5206868. 3

[FP10] FURUKAWA Y., PONCE J.: Accurate, dense, and robust multiview stereopsis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 32*, 8 (Aug 2010), 1362–1376. doi:10.1109/TPAMI.2009.161. 3, 11, 13

[GFT*11] GHOSH A., FYFFE G., TUNWATTANAPONG B., BUSCH J., YU X., DEBEVEC P.: Multiview face capture using polarized spherical gradient illumination. *ACM Trans. Graph. 30*, 6 (Dec. 2011), 129:1–129:10. URL: http://doi.acm.org/10.1145/2070781.2024163, doi:10.1145/2070781.2024163. 13

[GGW*98] GUENTER B., GRIMM C., WOOD D., MALVAR H., PIGHIN F.: Making faces. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1998), SIGGRAPH '98, ACM, pp. 55–66. URL: http://doi.acm.org/10.1145/280814.280822, doi:10.1145/280814.280822. 2

[Gow75] GOWER J. C.: Generalized procrustes analysis. *Psychometrika 40*, 1 (1975), 33–51. URL: http://dx.doi.org/10.1007/BF02291478, doi:10.1007/BF02291478. 9

[GVWT13] GARRIDO P., VALGAERT L., WU C., THEOBALT C.: Reconstructing detailed dynamic face geometry from monocular video. *ACM Trans. Graph. 32*, 6 (Nov. 2013), 158:1–158:10. URL: http://doi.acm.org/10.1145/2508363.2508380, doi:10.1145/2508363.2508380. 3

[GZC*16] GARRIDO P., ZOLLHÖFER M., CASAS D., VALGAERTS L., VARANASI K., PÉREZ P., THEOBALT C.: Reconstruction of personalized 3d face rigs from monocular video. *ACM Trans. Graph. 35*, 3 (May 2016), 28:1–28:15. URL: http://doi.acm.org/10.1145/2890493, doi:10.1145/2890493. 3

[HCTW11] HUANG H., CHAI J., TONG X., WU H.-T.: Leveraging motion capture and 3d scanning for high-fidelity facial performance acquisition. *ACM Trans. Graph. 30*, 4 (July 2011), 74:1–74:10. URL: http://doi.acm.org/10.1145/2010324.1964969, doi:10.1145/2010324.1964969. 3

[HMYL15] HSIEH P.-L., MA C., YU J., LI H.: Unconstrained realtime facial performance capture. In *CVPR* (2015), IEEE Computer Society, pp. 1675–1683. URL: http://dblp.uni-trier.de/db/conf/cvpr/cvpr2015.html#HsiehMYL15. 3

[IBP15] ICHIM A. E., BOUAZIZ S., PAULY M.: Dynamic 3d avatar creation from hand-held video input. *ACM Trans. Graph. 34*, 4 (July 2015), 45:1–45:14. URL: http://doi.acm.org/10.1145/2766974, doi:10.1145/2766974. 3

[KH12] KLAUDINY M., HILTON A.: High-fidelity facial performance capture with non-sequential temporal alignment. In *Proceedings of the 3rd Symposium on Facial Analysis and Animation* (New York, NY, USA, 2012), FAA '12, ACM, pp. 3:1–3:1. URL: http://doi.acm.org/10.1145/2491599.2491602, doi:10.1145/2491599.2491602. 2, 3

[Kin09] KING D. E.: Dlib-ml: A machine learning toolkit. *J. Mach.*

Learn. Res. 10 (Dec. 2009), 1755–1758. URL: http://dl.acm.org/citation.cfm?id=1577069.1755843. 4

[KS14] KAZEMI V., SULLIVAN J.: One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition* (Washington, DC, USA, 2014), CVPR '14, IEEE Computer Society, pp. 1867–1874. URL: http://dx.doi.org/10.1109/CVPR.2014.241, doi:10.1109/CVPR.2014.241. 4

[Lea15] LEAGUE D. H.: The wikihuman project. http://gl.ict.usc.edu/Research/DigitalEmily2/, 2015. Accessed: 2015-12-01. 4

[LWP10] LI H., WEISE T., PAULY M.: Example-based facial rigging. *ACM Trans. Graph. 29*, 4 (July 2010), 32:1–32:6. URL: http://doi.acm.org/10.1145/1778765.1778769, doi:10.1145/1778765.1778769. 3

[LYYB13] LI H., YU J., YE Y., BREGLER C.: Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph. 32*, 4 (July 2013), 42:1–42:10. URL: http://doi.acm.org/10.1145/2461912.2462019, doi:10.1145/2461912.2462019. 3, 13

[MBW*15] MALLESON C., BAZIN J. C., WANG O., BRADLEY D., BEELER T., HILTON A., SORKINE-HORNUNG A.: Facedirector: Continuous control of facial performance in video. In *2015 IEEE International Conference on Computer Vision (ICCV)* (Dec 2015), pp. 3979–3987. doi:10.1109/ICCV.2015.453. 13

[NFA*15] NAGANO K., FYFFE G., ALEXANDER O., BARBIÇ J., LI H., GHOSH A., DEBEVEC P.: Skin microstructure deformation with displacement map convolution. *ACM Trans. Graph. 34*, 4 (July 2015), 109:1–109:10. URL: http://doi.acm.org/10.1145/2766894, doi:10.1145/2766894. 13

[NVW*13] NEUMANN T., VARANASI K., WENGER S., WACKER M., MAGNOR M., THEOBALT C.: Sparse localized deformation components. *ACM Trans. Graph. 32*, 6 (Nov. 2013), 179:1–179:10. URL: http://doi.acm.org/10.1145/2508363.2508417, doi:10.1145/2508363.2508417. 13

[OLSL16] OLSZEWSKI K., LIM J. J., SAITO S., LI H.: High-fidelity facial and speech animation for vr hmds. *ACM Trans. Graph. 35*, 6 (Nov. 2016), 221:1–221:14. URL: http://doi.acm.org/10.1145/2980179.2980252, doi:10.1145/2980179.2980252. 3

[PB81] PLATT S. M., BADLER N. I.: Animating facial expressions. *SIGGRAPH Comput. Graph. 15*, 3 (Aug. 1981), 245–252. URL: http://doi.acm.org/10.1145/965161.806812, doi:10.1145/965161.806812. 2

[SA07] SORKINE O., ALEXA M.: As-rigid-as-possible surface modeling. In *Proceedings of the Fifth Eurographics Symposium on Geometry Processing* (Aire-la-Ville, Switzerland, Switzerland, 2007), SGP '07, Eurographics Association, pp. 109–116. URL: http://dl.acm.org/citation.cfm?id=1281991.1282006. 8

[Si15] SI H.: Tetgen, a delaunay-based quality tetrahedral mesh generator. *ACM Trans. Math. Softw. 41*, 2 (Feb. 2015), 11:1–11:36. URL: http://doi.acm.org/10.1145/2629697, doi:10.1145/2629697. 3

[SKS14] SUWAJANAKORN S., KEMELMACHER-SHLIZERMAN I., SEITZ S. M.: Total moving face reconstruction. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV* (2014), pp. 796–812. URL: http://dx.doi.org/10.1007/978-3-319-10593-2_52, doi:10.1007/978-3-319-10593-2_52. 3

[SLL16] SAITO S., LI T., LI H.: Real-time facial segmentation and performance capture from RGB input. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII* (2016), pp. 244–261. URL: http://dx.doi.org/10.1007/978-3-319-46484-8_15, doi:10.1007/978-3-319-46484-8_15. 3

[SNF05] SIFAKIS E., NEVEROV I., FEDKIW R.: Automatic determination of facial muscle activations from sparse motion capture marker data. In *ACM SIGGRAPH 2005 Papers* (New York, NY, USA, 2005), SIGGRAPH '05, ACM, pp. 417–425. URL: http://doi.acm.org/10.1145/1186822.1073208, doi:10.1145/1186822.1073208. 3

[SWTC14] SHI F., WU H.-T., TONG X., CHAI J.: Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Trans. Graph. 33*, 6 (Nov. 2014), 222:1–222:13. URL: http://doi.acm.org/10.1145/2661229.2661290, doi:10.1145/2661229.2661290. 3

[TW93] TERZOPOULOS D., WATERS K.: Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Trans. Pattern Anal. Mach. Intell. 15*, 6 (June 1993), 569–579. URL: http://dx.doi.org/10.1109/34.216726, doi:10.1109/34.216726. 2

[TZN*15] THIES J., ZOLLHOFER M., NIESSNER M., VALGAERTS L., STAMMINGER M., THEOBALT C.: Real-time expression transfer for facial reenactment. *ACM Trans. Graph. 34*, 6 (Oct. 2015), 183:1–183:14. URL: http://doi.acm.org/10.1145/2816795.2818056, doi:10.1145/2816795.2818056. 3, 4, 10, 11

[TZS*16] THIES J., ZOLLHÖFER M., STAMMINGER M., THEOBALT C., NIESSNER M.: Face2face: Real-time face capture and reenactment of RGB videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016* (2016), pp. 2387–2395. URL: http://dx.doi.org/10.1109/CVPR.2016.262, doi:10.1109/CVPR.2016.262. 3

[VWB*12] VALGAERTS L., WU C., BRUHN A., SEIDEL H.-P., THEOBALT C.: Lightweight binocular facial performance capture under uncontrolled lighting. *ACM Trans. Graph. 31*, 6 (Nov. 2012), 187:1–187:11. URL: http://doi.acm.org/10.1145/2366145.2366206, doi:10.1145/2366145.2366206. 2, 3, 8

[WBGB16] WU C., BRADLEY D., GROSS M., BEELER T.: An anatomically-constrained local deformation model for monocular face capture. *ACM Trans. Graph. 35*, 4 (July 2016), 115:1–115:12. URL: http://doi.acm.org/10.1145/2897824.2925882, doi:10.1145/2897824.2925882. 2, 9

[WBLP11] WEISE T., BOUAZIZ S., LI H., PAULY M.: Realtime performance-based facial animation. *ACM Trans. Graph. 30*, 4 (July 2011), 77:1–77:10. URL: http://doi.acm.org/10.1145/2010324.1964972, doi:10.1145/2010324.1964972. 3

[Wil90] WILLIAMS L.: Performance-driven facial animation. In *Proceedings of the 17th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1990), SIGGRAPH '90, ACM, pp. 235–242. URL: http://doi.acm.org/10.1145/97879.97906, doi:10.1145/97879.97906. 2

[WLVGP09] WEISE T., LI H., VAN GOOL L., PAULY M.: Face/off: Live facial puppetry. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (New York, NY, USA, 2009), SCA '09, ACM, pp. 7–16. URL: http://doi.acm.org/10.1145/1599470.1599472, doi:10.1145/1599470.1599472. 3

[WTP*09] WERLBERGER M., TROBIN W., POCK T., WEDEL A., CREMERS D., BISCHOF H.: Anisotropic huber-l1 optical flow. In *British Machine Vision Conference, BMVC 2009, London, UK, September 7-10, 2009. Proceedings* (2009), pp. 1–11. URL: http://dx.doi.org/10.5244/C.23.108, doi:10.5244/C.23.108. 6, 7

[Yea02] YEATMAN H.: Human face project. In *Proceedings of the 29th International Conference on Computer Graphics and Interactive Techniques. Electronic Art and Animation Catalog.* (New York, NY, USA, 2002), SIGGRAPH '02, ACM, pp. 162–162. URL: http://doi.acm.org/10.1145/2931127.2931216, doi:10.1145/2931127.2931216. 2

[ZHS*05] ZHOU K., HUANG J., SNYDER J., LIU X., BAO H., GUO B., SHUM H.-Y.: Large mesh deformation using the volumetric graph laplacian. In *ACM SIGGRAPH 2005 Papers* (New York, NY, USA, 2005), SIGGRAPH '05, ACM, pp. 496–503. URL: http://doi.acm.org/10.1145/1186822.1073219, doi:10.1145/1186822.1073219. 8