ACM SIGGRAPH / Eurographics Symposium on Computer Animation 2018
T. Beeler and N. Thuerey
(Guest Editors)

Volume 37 (2018), Number 8

User-Guided Lip Correction for Facial Performance Capture

D. Dinev[1,2], T. Beeler[2], D. Bradley[2], M. Bächer[2], H. Xu[2], and L. Kavan[1]

[1]University of Utah [2]Disney Research

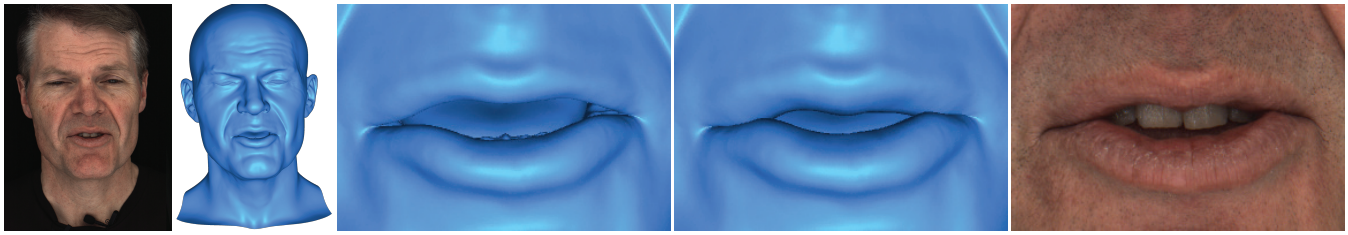

Figure 1: *We present a user-guided method for correcting lips in facial performance capture. From left to right: state-of-the-art facial capture methods can achieve high quality 3D face results but often struggle in the lip region. Our regression-based lip correction method is easy to use and can quickly improve the lip shapes for a whole performance, increasing the fidelity with respect to the true motion.*

Abstract

Facial performance capture is the primary method for generating facial animation in video games, feature films and virtual environments, and recent advances have produced very compelling results. Still, one of the most challenging regions is the mouth, which often contains systematic errors due to the complex appearance and occlusion/dis-occlusion of the lips. We present a novel user-guided approach to correcting these common lip shape errors present in traditional capture systems. Our approach is to allow a user to manually correct a small number of problematic frames, and then our system learns the types of corrections desired and automatically corrects the entire performance. As correcting even a single frame using traditional 3D sculpting tools can be time consuming and require great skill, we also propose a simple and fast 2D sketch-based method for generating plausible lip corrections for the problematic key frames. We demonstrate our results on captured performances of three different subjects, and validate our method with an additional sequence that contains ground truth lip reconstructions.

1. Introduction

Facial performance capture has become the industry standard for generating facial animation for virtual characters, especially in the case of digital doubles of real actors. Despite great progress, one area that remains a challenge is the lip region, which poses problems because of the complex appearance properties caused by lip wetness, shadows, and continuous occlusion and dis-occlusion. The reconstructed geometry in this region is often plausible, capturing most of the gross lip motion, but effects such as bulging, precise lip curl and adhesion (so-called "sticky lips") are often missed. To make matters worse, this region is extremely important for realistic facial motion, in particular when it comes to speech, and omitting those subtle effects can be the difference between realistic and uncanny-looking characters. For this reason, for production quality digital humans, typically many hours of artist time go into manually sculpting 3D lip corrections, even on top of the highest quality performance capture results.

In this work we aim to alleviate the problem by providing a user-guided approach for correcting lips in facial performances. The main idea is to choose a small set of important frames in a performance sequence that identify problematic lip shapes that were not reconstructed well, and then manually correct these few frames. Our system will then learn the 3D correction and automatically propagate the sparse manual corrections to the full sequence, using a gradient-based regression framework. Our approach lends itself well to incremental shot production, where the user can start with only a few corrected frames and gradually add corrections only where needed. Furthermore, since it can be time-consuming and cumbersome to manually sculpt 3D corrections for even a small number of problem frames, we additionally propose a simple method to apply lip corrections through an intuitive 2D sketching interface, which can generate plausible 3D lip corrections in little time without the need for artistic skill.

Our approach draws on related ideas for facial capture that pay

particular attention to the lips. In particular, Bhat et al. [BGY*13] show a production application of marker-based facial capture that uses hand-drawn lip contours to obtain improved lip shapes, however they must annotate each and every frame of the performance. The key idea of our method is to learn the underlying corrections in order to reduce the amount of manual effort. With a similar goal, Garrido et al. [GZW*16] capture ground truth data of lip motions using high-frequency lip tattoos and apply automatic corrections in a monocular facial performance capture scenario. However, their goal is to create a generalized regression framework for basic improvements to low-resolution lip shapes, targeting more a mass-consumer audience, while our work is designed for shot-specific production-quality lip refinement at high quality, with an easy-to-use interface.

The key idea of our approach is to predict the inner shape of the lips from the shape of the surrounding area of the face, which is assumed to be captured with high accuracy because it is not affected by the artifacts due to wetness and occlusions. Inspired by anatomically-based facial animation [SNF05, CBF16, LCF17, IKKP17], where the facial deformation is determined from underlying muscle activations, we argue that the inner lip shape is inherently linked to the outer lip shape since the overall lip deformation is determined by the activation of the lip muscles underneath. In our work, we avoid the complexity of establishing an explicit anatomical facial model but instead train a regressor to directly predict the inner lip shape from the surrounding region, which we assume to be captured well. Our lip shape regressor operates in the gradient domain, which makes it robust to global translations of the mouth. Also, our method is designed specifically to operate only with small amounts of training data in order to minimize the burden on the user.

We demonstrate our user-guided lip correction method on several performance capture sequences of three different actors. Additionally, we provide an evaluation of our approach on previously presented ground truth lip performance data [GZW*16]. We believe this work can have a large impact on high-quality facial performance applications by providing improved reconstructions of a very important part of the face at little cost to the user.

2. Related Work

In the following we discuss related work in face capture and animation, as well as capture methods specifically dedicated to the lip and mouth region.

Face Capture and Animation. There are several approaches for creating facial animation, including blendshape animation [LAR*14], physically-based animation [TW90, SNF05, CBF16, LCF17, KBB*17, IKKP17], marker-based motion capture [BGY*13], and dense markerless performance capture [BHPS10, BHB*11, FJA*14, LKA*17]. Capturing a dense surface performance currently offers the highest fidelity as it can faithfully reproduce all the subtle nonlinear effects of a skin, and is thus the method of choice for production-level facial animation.

Facial performance capture has received a tremendous amount of attention over the past decades. In recent years we have witnessed significant improvements in acquisition of high-fidelity facial expressions and performances [ARL*09, BBB*10, BHPS10, GFT*11, BHB*11, FJA*14, LKA*17]. In addition to the facial geometry, high fidelity eyelids [BBK*15], eyes [BBN*14, BBGB16], skin microstructure [GTB*13, NFA*15], facial hair [BBN*12] and scalp hair [LLP*12, HMLL15] can also be captured at high accuracy.

Expressive and detailed lip reconstruction, however, remains an unsolved challenge, due to the extreme deformation of the lip tissue, dramatic color changes caused by blood flow, challenging specular reflectance due to saliva, as well as recurring occlusion and dis-occlusion of the inner lips. Recently, Garrido et al. [GZW*16] succeeded at capturing high quality lip shapes by placing high-frequency tattoos on the lips, circumventing some of the aforementioned challenges. While suitable for generating a shape dataset, the tattoos destroy the natural appearance of the lips and are thus less attractive for creating digital doubles, and complicated tattooing is not always an option for actors in a high-end production setting. As a consequence, VFX workflows still rely on manually correcting 3D lip shapes, often on a per-frame basis, which is a lengthy and cumbersome task that requires talented artists. Our work is designed to dramatically reduce the artist work spent on 3D lip refinement for high-fidelity performance capture, by requiring only a handful of corrected frames and then filling in the remaining performance with a regression-based learning approach.

Lip Capture. High-fidelity lip appearance and deformation are especially important for high quality facial animation [KB98, WLL04] and mouth/lip tracking and reconstruction has been a long-studied problem. Many approaches use 2D contour lines to track lips in the image space [NM09, TKC00, ECC04, BHO02] but do not reconstruct the dense 3D geometry. In our paper, we provide a 2D image-based editing tool to refine the tracked lip shapes but with the goal of generating corrections of 3D geometry. Similar to previous sketch-based interfaces [ZNA07, LCXS09, MAO*11], our 2D editing tool allows for intuitive control of lip shapes and is easy to use even for non-expert users. This idea for simple facial animation editing is akin to keyframe-based performance editing tools, such as those used during expression retargeting [SLS*12, XCLT14].

Recently, there have been several works on modeling and reconstruction of 3D shape of the lips in multiview capture setups [BHPS10, ASC13, BGY*13, KIMM14] or even just from monocular video [LYYB, LXC*15, GZW*16]. In particular, Bradley et al. [BHPS10] propose edge-based mouth tracking which improves the lip reconstruction with detected contour constraints. Anderson et al. [ASC13] track the 2D lip contours and automatically register the mouth region to 3D geometry by aligning the lip contours with predefined isolines on the surface. A data-driven approach was also proposed by Liu et al. [LXC*15] for real-time lip shape refinement, and landmark-based lip correction is proposed in the real-time system of Li et al. [LYYB]. Olszewski et al. are able to demonstrate lip animations such as“sticky-lips" using deep learning in a head-mounted capture system [OLSL16]. Even though impressive results have been achieved with these prior works, expressive lip shapes and subtle effects on the lips are still missing for

production-level reconstruction. Our method is orthogonal to all of these 3D lip tracking approaches and enables corrections of the entire performance with minimal artist input. Bhat et al. [BGY*13] show a production application of marker-based facial capture and improve lip tracking with *per-frame* hand-drawn lip contours. With corrections required only for a sparse set of frames, our method automates the editing of the remaining facial animation frames, which significantly reduces the amount of manual effort. Furthermore, our approach provides temporal smoothness and editing consistency, which is challenging to achieve with per-frame correction.

Garrido et al. [GZW*16] learn the difference between high-quality lip shapes and coarse monocular lip reconstructions based on a regression function and automatically apply the corrections to low-resolution lip shapes from monocular video. Our approach shares a similar goal to their work. Unlike their approach, however, our method learns the lip corrections directly from the surrounding facial deformation and is designed for shot-specific, production-quality lip refinement on top of highest quality facial performance capture.

3. Overview

We take a user-guided approach for improving lip shapes in performance capture applications. Given a reconstructed mesh sequence, a user can select a sparse number of frames that contain systematic reconstruction errors caused by the challenges inherent to 3D lip capture, and manually provide the corrected shapes. Our system then learns the type of corrections the user desires, and applies a corresponding refinement to the entire sequence (Section 4). The process can then be repeated incrementally if some frames remain problematic. Since manually sculpting correct lip shapes is a tedious task that requires great artistic skill, we also propose a simple sketch-based interface for refining the lip shapes of the chosen key frames (Section 5). Finally, we show results for various performances of three different actors, as well as a quantitative evaluation of our method using ground truth lip data (Section 6).

4. Automatic Lip Correction Learning

Many facial capture systems can provide accurate and high-resolution reconstructions of the actor with only a few problematic areas, one of which are the inner lips. The key idea behind our method is to use the accurate portions of the mesh to predict the inaccurate parts. Since the inner lip shape is heavily influenced by the underlying muscles that also dictate the shape of the mouth area, we believe that this relationship can be learned. In this work we employ the facial capture method of Beeler et al. [BHB*11], which is commonly used in feature film productions. We first divide the mesh (Figure 2) into a mouth region where we trust the capture data (blue) and an inner lip region where the capture is challenged and might yield less accurate results (red) with the goal to learn a mapping between the two.

We have several choices in how to represent the regions in Figure 2. An obvious representation is to use the Cartesian coordinates of each vertex and learn a mapping for the positions directly. However, this approach is not invariant to global translation, which is problematic since head and body motion directly impact the absolute position of the lips during a performance. Rigid stabilization [BB14] can alleviate the problem by removing the global head motion, but the jaw motion remains a problem. To remedy this, we represent the shapes of our regions differentially. Specifically, for every triangle, we construct a deformation gradient $\mathbf{F} \in \mathbb{R}^{3\times 3}$ by first creating a phantom vertex [SP04] in the neutral pose and the deformed pose. We use this phantom vertex to create shape matrices for the rest shape ($\mathbf{D}_r$) and the deformed shape ($\mathbf{D}_d$). The deformation gradient is then defined as $\mathbf{F} = \mathbf{D}_d\mathbf{D}_r^{-1}$ (we refer to [SB12] for more details on linear finite elements). We save the deformation gradients of every triangle in the mouth region in a vector $\mathbf{y} \in \mathbb{R}^{9m}$, and save the inner lip deformation gradients in a vector $\mathbf{x} \in \mathbb{R}^{9n}$ where m and n are the number of mouth and inner lip triangles, respectively.

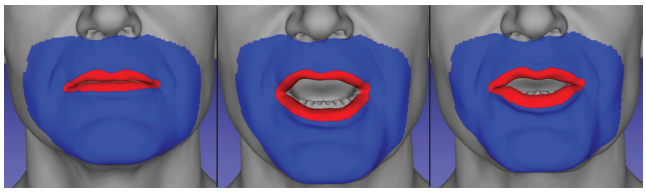

Figure 2: *We segment the mesh into a mouth region (blue) and an inner lip region (red). The rest of the mesh (gray) is ignored.*

Our goal is to learn a linear function $\mathbf{f}(\mathbf{y}) = \theta\mathbf{y} = \mathbf{x}$ where matrix $\theta \in \mathbb{R}^{9n\times 9m}$ maps input vectors $\mathbf{y}$ to output vectors $\mathbf{x}$. We create a training data set that consists of k meshes with corrected lips, and for every mesh we generate training tuples $(\mathbf{y},\mathbf{x})$. θ is then the minimizer of a linear regression problem

$$\min_{\theta} \frac{1}{2k}\sum_i ||\theta\mathbf{y}^{(i)} - \mathbf{x}^{(i)}||^2.$$

We then create the matrix $\mathbf{Y} \in \mathbb{R}^{9m\times k}$, whose ith column is $\mathbf{y}^{(i)}$, and an analogous matrix $\mathbf{X} \in \mathbb{R}^{9n\times k}$ for the $\mathbf{x}$ vectors, recasting the regression problem in matrix form

$$\min_{\theta} \frac{1}{2k}||\theta\mathbf{Y} - \mathbf{X}||_F^2$$

with corresponding normal equations

$$\theta\mathbf{Y}\mathbf{Y}^T = \mathbf{X}\mathbf{Y}^T.$$

We observe that the resulting system is under-determined as $k \ll n,m$. Due to the highly detailed facial geometry that is encoded in the meshes, the triangle counts are high (e.g., $m = 19315$) and the matrix θ is large, suggesting high memory costs and the risks of overfitting. To avoid these issues, in the following we propose a reduced regression approach, where the number of unknowns depends on the number of corrected frames k (typically under 20) rather than the resolution-dependent quantities m and n.

The first step towards this reduction is forming the singular value decompositions (SVDs) of $\mathbf{Y}$ and $\mathbf{X}$

$$\mathbf{Y} = \mathbf{U}_\mathbf{Y}\mathbf{C}_\mathbf{Y} \text{ with } \mathbf{C}_\mathbf{Y} = \Sigma_\mathbf{Y}\mathbf{V}_\mathbf{Y}^T \text{ and } \mathbf{X} = \mathbf{U}_\mathbf{X}\mathbf{C}_\mathbf{X} \text{ with } \mathbf{C}_\mathbf{X} = \Sigma_\mathbf{X}\mathbf{V}_\mathbf{X}^T,$$

where we introduce coefficient matrices $\mathbf{C}_\mathbf{Y}$ and $\mathbf{C}_\mathbf{X}$, respectively.

Using the invariance of the Frobenius norm under orthonormal transformations and plugging in the coefficient matrices, we obtain

$$||\theta\mathbf{Y}-\mathbf{X}||_F^2 = ||\mathbf{U}_\mathbf{X}^T(\theta\mathbf{Y}-\mathbf{X})||_F^2 = ||\mathbf{U}_\mathbf{X}^T\theta\mathbf{U}_\mathbf{Y}\mathbf{C}_\mathbf{Y}-\mathbf{C}_\mathbf{X}||_F^2.$$

We then define the reduced unknown matrix $\bar{\theta} = \mathbf{U}_\mathbf{X}^T\theta\mathbf{U}_\mathbf{Y}$, solving the problem

$$\min_{\bar{\theta}} \frac{1}{2k}||\bar{\theta}\mathbf{C}_\mathbf{Y}-\mathbf{C}_\mathbf{X}||_F^2$$

which leads to normal equations

$$\bar{\theta}\mathbf{C}_\mathbf{Y}\mathbf{C}_\mathbf{Y}^T = \mathbf{C}_\mathbf{X}\mathbf{C}_\mathbf{Y}^T$$

where the matrix $\bar{\theta}$ has size $k \times k$ and can be therefore computed very quickly.

For any new shape j, we can reconstruct the inner lip deformation gradient vector $\mathbf{x}_j$ by computing $\mathbf{y}_j$ as above and applying our reduced regression model

$$\mathbf{x}_j = \theta\mathbf{y}_j \;\text{ with }\; \theta = \mathbf{U}_\mathbf{X}\bar{\theta}\mathbf{U}_\mathbf{Y}^T.$$

Note that the coefficients of the large matrix θ are never explicitly evaluated and instead, the matrix is compactly represented in its factorized form $\mathbf{U}_\mathbf{X}\bar{\theta}\mathbf{U}_\mathbf{Y}^T$.

Finally, we convert the resulting vector of deformation gradients $\mathbf{x}_j$ back to vertex positions. This corresponds to solving a sparse linear system, using the positions of the vertices at the border between the inner lips and outer mouth region as Dirichlet boundary conditions.

5. Sketch-Based Lip Refinement

Our automatic lip correction method described in the previous section requires a set of user-provided corrected lip shapes for a small set of frames, which are used as training data. Naturally, the quality of the results depends on the quality of this training data, and so the best option is to have a digital artist manually sculpt the correct 3D lip shapes. However, this can be very time consuming and typically such a skilled artist is not available. For this reason, we propose a simple 2D sketch-based method for correcting individual lip shapes. While not as accurate as hand-sculpted corrections, this approach is much more widely suited for inexperienced users and yet produces visually very plausible lip shapes.

Our approach requires the user to sketch two contours, one for the upper lip and one for the lower lip, both representing the inner occluding contour (see Figure 3 (a)). Given these sketches, the original reconstructed mesh and the camera projection matrices we automatically correct the lip shapes to match the user drawings. This is accomplished using iterative Laplacian mesh deformation [SCOL*04]. Specifically, let the face mesh $\mathbf{M}$ consist of a set of vertex positions $\{\boldsymbol{v}_i\}$, then we solve for new vertex positions $\{\bar{\boldsymbol{v}}_i\}$ which satisfy the following energy in a least-squares sense,

$$E = \lambda_{pos}\cdot E_{pos} + \lambda_{silh}\cdot E_{silh} + \lambda_{reg}\cdot E_{reg}. \quad (1)$$

where $E_{pos} = \sum_i ||\boldsymbol{v}_i - \bar{\boldsymbol{v}}_i||_2^2$ is a position constraint designed to minimize vertex deviation. $E_{reg} = ||L(\{\boldsymbol{v}\}) - L(\{\bar{\boldsymbol{v}}\})||_2^2$ is a regularization constraint that attempts to preserve local shape, where L is the discrete mesh Laplacian [SCOL*04]. E_{silh} is a silhouette constraint that aims to deform the mesh to match the user sketches, and is defined as follows. We first pre-filter the set of vertices to consider only those corresponding to silhouettes with respect to the camera, and then identify which silhouette vertices correspond to the upper lip versus lower lip by projecting the vertex normals onto the image plane (the y-component of the projected normal is positive for lower lip vertices and negative for the upper lip). Then, for each pixel p of the sketched contours we compute the corresponding ray $\boldsymbol{r}_p$ in 3D and find the closest silhouette vertex $\boldsymbol{v}_p$ for the corresponding lip. The silhouette constraint minimizes the distance of the vertex to the ray, for all contour pixels as $E_{silh} = \sum_p ||\boldsymbol{r}_p - \boldsymbol{v}_p||_2^2$.

Since the lip silhouettes can change during deformation, we solve Eq. 1 iteratively, recomputing the silhouette vertices and the corresponding constraints each iteration. This approach works well if the mesh deforms slowly into position, which we accomplish using a higher rigidity weight ($\lambda_{reg} = 3$) compared to the silhouette term ($\lambda_{silh} = 1$). We set a low position weight ($\lambda_{pos} = 0.1$) for all vertices except those that we want to guarantee not to move, for example hidden ones like the back side of the head and the inner mouth ($\lambda_{pos} = 1$). The mesh is deformed over 5 iterations, and the result is a plausible correction of the lips matching the user-given sketches, as shown in Figure 3 (c) for starting mesh (b).

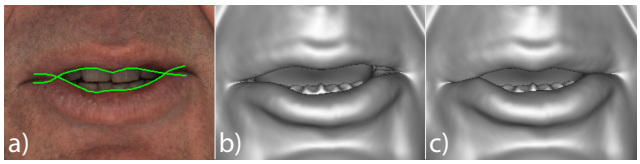

Figure 3: *We propose a simple sketch-based lip refinement method for correcting a small number of problem frames. The user draws contours for the upper and lower inner lip silhouettes (a), and the original reconstruction (b) is automatically corrected in a plausible way (c).*

One thing to note is that we detect when the lower lip contour is drawn above the upper lip contour and consequently constrain them to be at the same 2D position, which allows us to exaggerate the contour overlap at the mouth corners and guarantee that the lips will be compressed together when desired.

Our proposed sketch-based lip refinement tool is fast and user friendly. The entire process of drawing the contours and solving the mesh deformation takes less than one minute per frame, and requires no particular artistic skill.

6. Results

We first ran our method on dialogue sequences from three different actors, using the 2D contour tool to selectively correct frames that we add to our linear regressor. Then, to evaluate the accuracy of our method we applied our method to the the data set from Garrido et al. [GZW*16]. This data set uses high-frequency patterns to capture the inner lip region more accurately, so we can use this as a ground-truth data set to evaluate our method.

6.1. Lip Correction Results

Figure 4 shows our method applied to a dialogue sequence from Actor 1. With only 11 frames corrected, we were able to greatly improve the quality of the capture. In the original capture, some expressions where the mouth should be closed have a small gap between the lips (top row) and the corners of the lips lose their fleshiness (second and third rows). After using our tool to correct a few of these frames, similar issues in other frames are automatically corrected. In expressions where the mouth is open, the capture system can sometimes produce plausible results (bottom row). For such frames, our method does not introduce errors when compared to the initial capture result. The frames shown in Figure 4 were not part of the training set; they were corrected by our method. The full training set is shown in Figure 5. Our method does not have a noticeable effect on the structure of the individual triangles, as shown in Figure 6. This is because the regularization term E_{reg} in Eq. 1 ensures that our training data preserves the local shape of the original captured mesh.

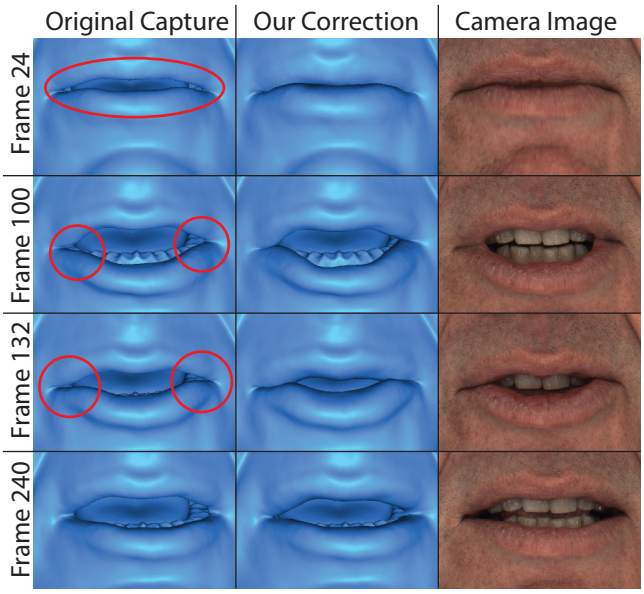

Figure 4: *The performance was trained using the samples in Figure 5; the frames shown here are not in the training data set. We are able to correct many artifacts that occur during capture: the lips not being properly closed (top row) and loss of fleshiness in the corners (second and third column). In situations where the original capture is feasible (bottom row), our method does not downgrade the results.*

In Figure 7 we applied our method to a different dialogue sequence from a different actor than Figure 4. In this sequence, the original capture exhibited artifacts due to interference with the teeth (note that neither the original capture system nor our method is attempting to reconstruct the motion of the teeth). These artifacts are circled in red in Figure 7. After correcting the lip shapes in only 8 meshes, our method is able to propagate these corrections throughout the performance. As before, the frames shown in Figure 7 are not part of the training data set.

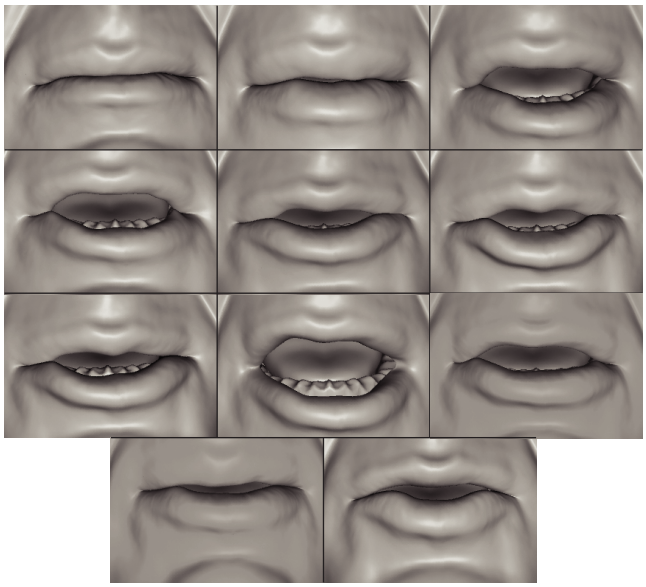

Figure 5: *These 11 meshes are the entire training data set used to generate the corrected lips for the sequence shown in Figure 4. They were created by applying our sketch-based corrections on the corresponding captured meshes.*

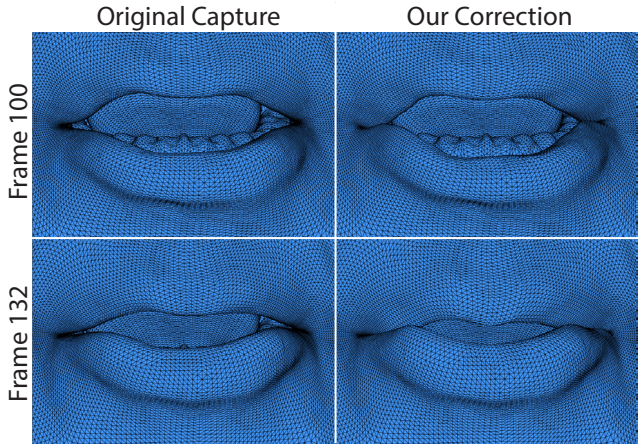

Figure 6: *A wireframe rendering of two results from Figure 4, showing the influence our method has on the individual triangles.*

We show the versatility of our method by applying lip corrections to a third actor, performing extreme lip shapes, as shown in Figure 8. Again, only 8 frames are corrected from a sequence of 350, and the frames from Figure 8 are not part of the training set.

6.2. Evaluation on Lip Tattoo Data

To test the accuracy of our method, we used a performance from the lip tattoos data set from [GZW*16] as a ground truth (276 frames in total) and tested how accurately we were able to reconstruct the inner lip region. We selected a handful of expressions for our training data set and applied our correction to the rest of the frames

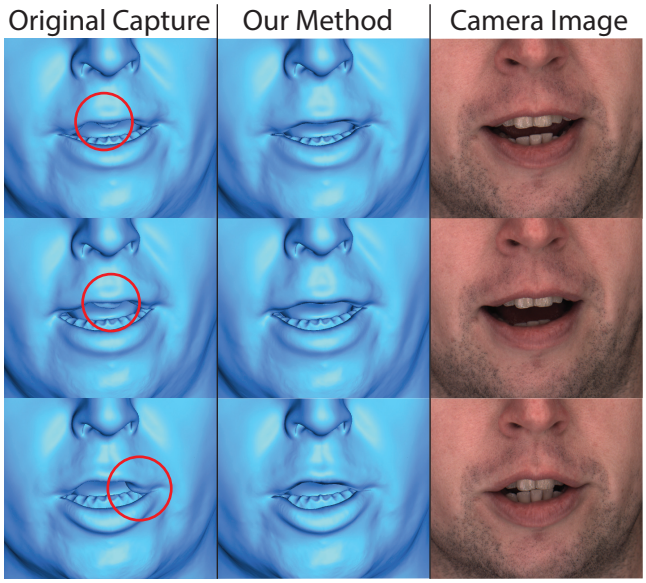

Figure 7: *In this dialogue capture, the actor's teeth interfered with the tracking abilities of the system, causing strange geometric artifacts to appear in the lip regions (circled in red, left column). Our method is able to remove these artifacts from all frames by only correcting a few example frames. The frames depicted in this figure were not part of the training set.*

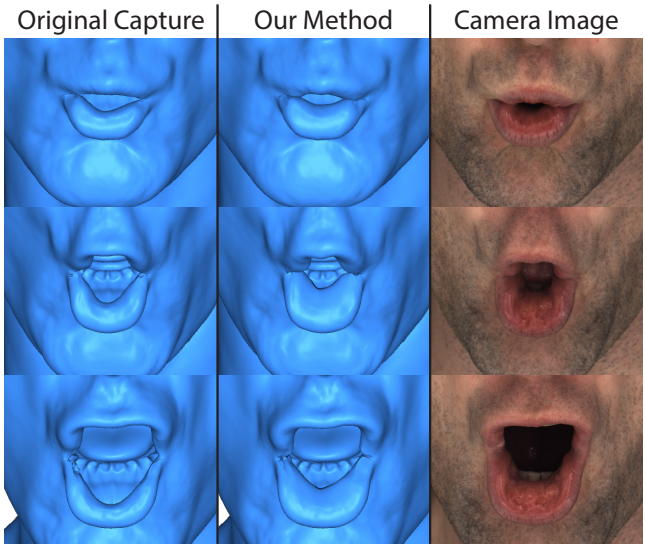

Figure 8: *We show the versatility of our method by applying lip correction to a third actor, performing extreme expressions.*

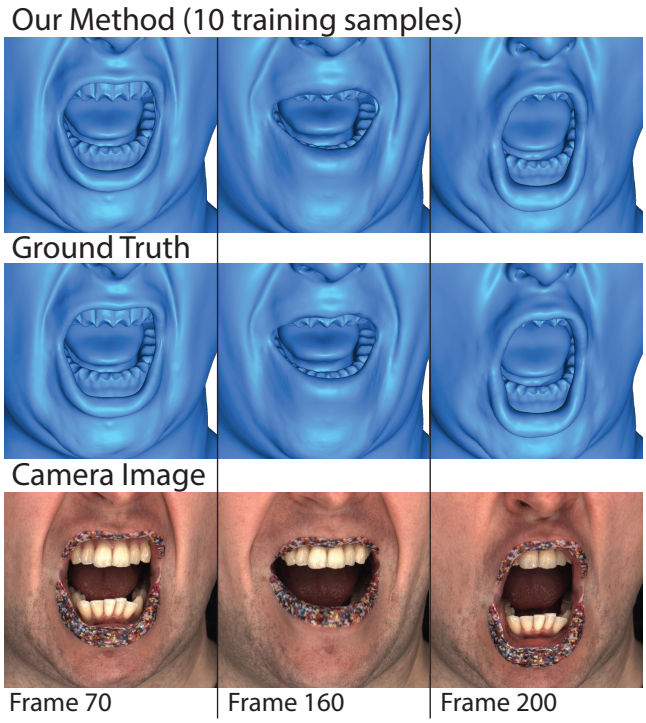

Figure 9: *Using the data set from [GZW*16], we can evaluate how accurately our method reconstructs the inner lip region. With 10 training samples, our method can reconstruct the inner lip region with only subtle differences from the ground truth.*

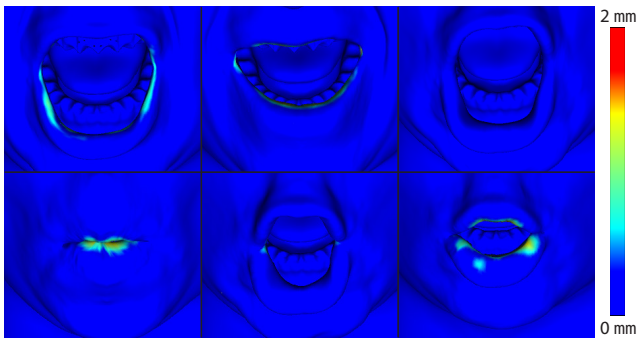

Figure 10: *An error map of our reconstruction compared to the ground truth using the data set from [GZW*16]. The top row shows the same expressions as Figure 9 and the bottom row contains additional lip shapes from the sequence.*

in the performance. Figure 9 shows the results of this reconstruction when we added 10 poses to the training set. Figure 10 shows an error map of the Hausdorff distance between our reconstructed meshes and the ground truth meshes. While we are not able to perfectly reconstruct the original inner lip shape with only 10 training samples, we are able to obtain a very good approximation.

To evaluate how the number of training points affects the final result, we selectively added meshes from this performance as training data and evaluated the absolute error of the deformation gradients: $\varepsilon = ||\mathbf{x}^* - \mathbf{x}||$. Figure 11 shows how the number of training samples affects the error of the reconstruction, taking the average ε over all frames in the sequence. The training samples were not chosen arbitrarily: we selected frames at the extreme poses where possible. This explains the diminishing returns observed in Figure 11: the sequence only has a limited number of extreme expressions, with

several of them being similar to each other. Once the extreme poses have been added to the training data, adding the in-between expressions does not have as dramatic of an effect. We can see what effect this error has visually in Figure 12. As we increase the number of training samples, the observable differences in this test frame become increasingly subtle.

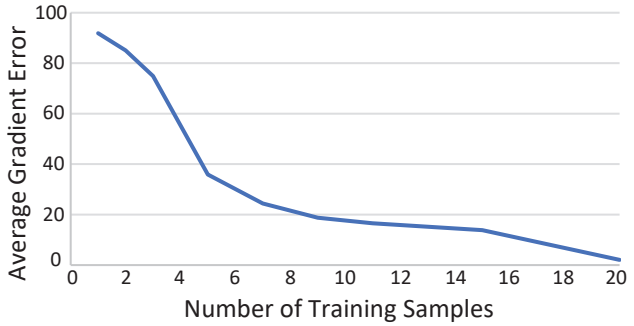

Figure 11: *The effect of the number of training samples on the average error of the sequence from Figure 9. Once all of the important poses are in the training set, adding additional poses yields diminishing returns.*

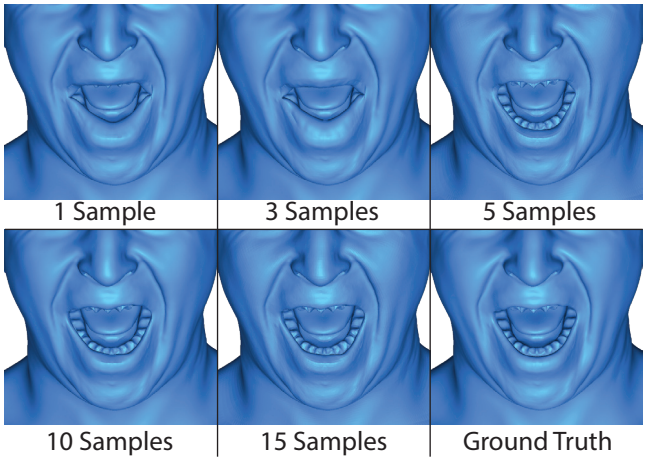

Figure 12: *The visual effect the number of training samples has on a test frame. The first several additions of new training samples make a significant impact, but larger numbers of training samples yield diminishing returns.*

One important parameter that affects our results is how we choose the inner lip/mouth regions (red and blue regions from Figure 2). If we choose an inner lip region that is too small, then we will end up keeping portions of the mesh that were not reliably captured (i.e., keeping bad corner shapes). Conversely, if the inner lip region is too large then we miss out important features that can help us predict the inner lip region (e.g., the outer lip can be important to accurately reconstructing the inner lip shape). It is important to select the regions such that they coincide with the boundary between the high/low confidence values of the capture system. We want to reconstruct the regions that are low confidence while maximally leveraging the regions that are high confidence.

6.3. Performance

The proposed method does not target interactive applications but high-quality VFX shot production and allows to automatically correct entire performances based on a few manually corrected shapes. With the proposed contour sketch tool it takes less than one minute for an inexperienced user to correct a shape, where the iterative solve takes about 20 seconds on the CPU and sketching takes approximately 30 seconds. For comparison we asked a professional digital artist to correct two sample frames and we timed how long it took to hand-correct the meshes in a 3D sculpting package. After spending approximately one minute to set up the scene, it took the artist 5 minutes 26 seconds (Figure 13, top) and 6 minutes 15 seconds (Figure 13, bottom) to correct the shapes. Our sketch tool is approximately 5 to 6 times faster and requires no artistic experience, yet provides similar results, as demonstrated in Figure 13.

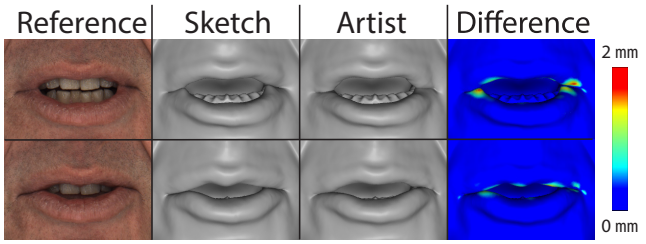

Figure 13: *We used our sketch-based refinement tool to correct two sample frames (left) and asked an artist to manually correct the same frames (middle). The difference is shown on the right. Our sketch-based refinement gives a similar result without requiring a skilled artist.*

Once the training data has been obtained, training the network is much faster, for example the network used for Figure 4 took 5.47 seconds on an Intel i7-4720HQ CPU at 2.6 GHz. Once the network has been trained, applying the regression to each mesh takes 5.15 seconds per frame (taken as an average over 10 frames). This timing includes converting between the position and gradient domains for every input mesh.

6.4. Limitations

Our sketch-based tool works best if used on a frontal view of the actor. However it is not necessary for the camera view to be exactly from the front. Drawing the lip contours on a side view, shown in Figure 14, can still produce an improved lip shape (Figure 14, c). However, if the view comes too far from the side, the correction is not as good on the opposite corner of the lip (Figure 14, e). Since we are targeting a multi-view capture scenario, if the head rotates too far from the frontal view there is typically a different camera view that can be used for lip correction annotations.

It should be noted that our method relies on a high-quality capture of the area surrounding the mouth. If this area is not captured accurately (e.g., if there are temporally inconsistent artifacts in this area) then the reconstruction cannot be trusted and the inner lip shapes will not be consistently predicted. This is not a real limitation, however, since we target the application of high-quality production-level facial capture.

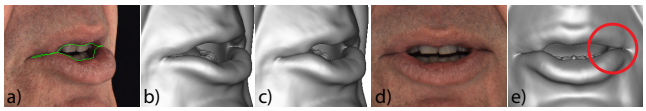

Figure 14: *Our sketch-based tool used from a side view angle (a) can improve the original mesh (b) to partially correct the lip shape (c). While plausible, the opposite corner of the mouth is not corrected quite as well, as seen from a front view (d,e). However, in this case the multi-view capture system would allow annotations to occur on a different camera image.*

7. Conclusion

We present a user-guided approach to correcting lips in facial performances. Starting by manually correcting a sparse set of training frames, our method learns the shape of the inner lip in relation to the surrounding mouth region, and then automatically fixes the lips in the entire facial performance. To perform the actual corrections, we propose a user-friendly, 2D sketch-based editing tool enabling even non-expert users to perform lip shape editing quickly and intuitively. This substantially reduces the amount of manual labour required to clean up captured facial performances. We apply our approach to captured performances of three subjects and validate it against ground truth lip reconstruction (high frequency lip tattoos). Our results demonstrate that the proposed method can reconstruct expressive lip shapes with subtle effects at high accuracy for production-level facial performances, while minimizing the required manual work.

Although it is typically easy for to select representative frames for training manually, it would be interesting to identify meaningful training frames automatically and suggest them to the user. In our work, we employ a simple but effective reduced linear regressor to predict the lip shapes from the surrounding facial deformation, which allows for fast correction but is not inherently physically correct. We are interested in incorporating stronger physical constraints which has the benefits of precise volume preservation and collision resolution. Furthermore, the learned predictor is currently actor-specific, but we believe that it could be generalized to different actors with sufficient training data, potentially yielding a fully automatic system to improve captured facial performances. Lastly, while the work focuses on user-guided lip corrections, we feel that similar concepts could also be employed to clean up other regions, such as the eye region which still remains a challenge for current facial performance capture systems.

References

[ARL*09] ALEXANDER O., ROGERS M., LAMBETH W., CHIANG M., DEBEVEC P.: The digital Emily project: Photoreal facial modeling and animation. In *ACM Siggraph Courses* (2009). 2

[ASC13] ANDERSON R., STENGER B., CIPOLLA R.: Lip tracking for 3d face registration. In *MVA* (2013), pp. 145–148. 2

[BB14] BEELER T., BRADLEY D.: Rigid stabilization of facial expressions. *ACM Transactions on Graphics (TOG) 33*, 4 (2014), 44. 3

[BBB*10] BEELER T., BICKEL B., BEARDSLEY P., SUMNER B., GROSS M.: High-quality single-shot capture of facial geometry. *ACM TOG 29*, 4 (2010), 40:1–40:9. 2

[BBGB16] BÉRARD P., BRADLEY D., GROSS M., BEELER T.: Lightweight eye capture using a parametric model. *ACM Transactions on Graphics (TOG) 35*, 4 (2016), 117. 2

[BBK*15] BERMANO A., BEELER T., KOZLOV Y., BRADLEY D., BICKEL B., GROSS M.: Detailed spatio-temporal reconstruction of eyelids. 44:1–44:11. 2

[BBN*12] BEELER T., BICKEL B., NORIS G., MARSCHNER S., BEARDSLEY P., SUMNER R. W., GROSS M.: Coupled 3D reconstruction of sparse facial hair and skin. 117:1–117:10. 2

[BBN*14] BÉRARD P., BRADLEY D., NITTI M., BEELER T., GROSS M.: High-quality capture of eyes. 223:1–223:12. 2

[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* (2013), ACM, pp. 7–14. 2, 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. *ACM TOG 30*, 4 (2011), 75:1–75:10. 2, 3

[BHO02] BARNARD M., HOLDEN E.-J., OWENS R.: Lip tracking using pattern matching snakes. In *Proc. of the Fifth Asian Conference on Computer Vision* (2002), vol. 1, Citeseer. 2

[BHPS10] BRADLEY D., HEIDRICH W., POPA T., SHEFFER A.: High resolution passive facial performance capture. In *ACM transactions on graphics (TOG)* (2010), vol. 29, ACM, p. 41. 2

[CBF16] CONG M., BHAT K. S., FEDKIW R.: Art-directed muscle simulation for high-end facial animation. Eurographics Association, pp. 119–127. 2

[ECC04] EVENO N., CAPLIER A., COULON P.-Y.: Accurate and quasi-automatic lip tracking. *IEEE Transactions on Circuits and Systems for Video technology 14*, 5 (2004), 706–715. 2

[FJA*14] FYFFE G., JONES A., ALEXANDER O., ICHIKARI R., DEBEVEC P.: Driving high-resolution facial scans with video performance capture. *ACM TOG 34*, 1 (2014), 8:1–8:14. 2

[GFT*11] GHOSH A., FYFFE G., TUNWATTANAPONG B., BUSCH J., YU X., DEBEVEC P.: Multiview face capture using polarized spherical gradient illumination. 129:1–129:10. 2

[GTB*13] GRAHAM P., TUNWATTANAPONG B., BUSCH J., YU X., JONES A., DEBEVEC P. E., GHOSH A.: Measurement-based synthesis of facial microgeometry. 335–344. 2

[GZW*16] GARRIDO P., ZOLLHÖFER M., WU C., BRADLEY D., PÉREZ P., BEELER T., THEOBALT C.: Corrective 3d reconstruction of lips from monocular video. *ACM Transactions on Graphics (TOG) 35*, 6 (2016), 219. 2, 3, 4, 5, 6

[HMLL15] HU L., MA C., LUO L., LI H.: Single-view hair modeling using a hairstyle database. 125:1–125:9. 2

[IKKP17] ICHIM A.-E., KADLEČEK P., KAVAN L., PAULY M.: Phace: physics-based face modeling and animation. *ACM Transactions on Graphics (TOG) 36*, 4 (2017), 153. 2

[KB98] KAUCIC R., BLAKE A.: Accurate, real-time, unadorned lip tracking. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)* (Jan 1998), pp. 370–375. 2

[KBB*17] KOZLOV Y., BRADLEY D., BÄCHER M., THOMASZEWSKI B., BEELER T., GROSS M.: Enriching facial blendshape rigs with physical simulation. *Computer Graphics Forum (Proc. Eurographics) 36*, 2 (2017). 2

[KIMM14] KAWAI M., IWAO T., MAEJIMA A., MORISHIMA S.: Automatic photorealistic 3d inner mouth restoration from frontal images. In *International Symposium on Visual Computing* (2014), Springer, pp. 51–62. 2

[LAR*14] LEWIS J. P., ANJYO K., RHEE T., ZHANG M., PIGHIN F., DENG Z.: Practice and Theory of Blendshape Facial Models. In *Eurographics STARs* (2014), pp. 199–218. 2

[LCF17] LAN L., CONG M., FEDKIW R.: Lessons from the evolution of an anatomical facial muscle model. In *Proceedings of the ACM SIGGRAPH Digital Production Symposium* (2017), ACM, p. 11. 2

[LCXS09] LAU M., CHAI J., XU Y.-Q., SHUM H.-Y.: Face poser: Interactive modeling of 3d facial expressions using facial priors. *ACM Transactions on Graphics (TOG) 29*, 1 (2009), 3. 2

[LKA*17] LAINE S., KARRAS T., AILA T., HERVA A., SAITO S., YU R., LI H., LEHTINEN J.: Production-level facial performance capture using deep convolutional neural networks. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (2017), ACM, p. 10. 2

[LLP*12] LUO L., LI H., PARIS S., WEISE T., PAULY M., RUSINKIEWICZ S.: Multi-view hair capture using orientation fields. pp. 1490–1497. 2

[LXC*15] LIU Y., XU F., CHAI J., TONG X., WANG L., HUO Q.: Video-audio driven real-time facial animation. *ACM Transactions on Graphics (TOG) 34*, 6 (2015), 182. 2

[LYYB] LI H., YU J., YE Y., BREGLER C.: Realtime facial animation with on-the-fly correctives. 2

[MAO*11] MIRANDA J. C., ALVAREZ X., ORVALHO J., GUTIERREZ D., SOUSA A. A., ORVALHO V.: Sketch express: facial expressions made easy. In *Proceedings of the Eighth Eurographics Symposium on Sketch-Based Interfaces and Modeling* (2011), ACM, pp. 87–94. 2

[NFA*15] NAGANO K., FYFFE G., ALEXANDER O., BARBIČ J., LI H., GHOSH A., DEBEVEC P.: Skin microstructure deformation with displacement map convolution. 109:1–109:10. 2

[NM09] NGUYEN Q. D., MILGRAM M.: Semi adaptive appearance models for lip tracking. In *2009 16th IEEE International Conference on Image Processing (ICIP)* (Nov 2009), pp. 2437–2440. doi:10.1109/ICIP.2009.5414105. 2

[OLSL16] OLSZEWSKI K., LIM J. J., SAITO S., LI H.: High-fidelity facial and speech animation for vr hmds. *ACM Transactions on Graphics (TOG) 35*, 6 (2016), 221. 2

[SB12] SIFAKIS E., BARBIC J.: Fem simulation of 3d deformable solids: a practitioner's guide to theory, discretization and model reduction. In *ACM SIGGRAPH 2012 Courses* (2012), ACM, p. 20. 3

[SCOL*04] SORKINE O., COHEN-OR D., LIPMAN Y., ALEXA M., RÖSSL C., SEIDEL H.-P.: Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing* (2004), pp. 175–184. 4

[SLS*12] SEOL Y., LEWIS J., SEO J., CHOI B., ANJYO K., NOH J.: Spacetime expression cloning for blendshapes. *ACM Transactions on Graphics (TOG) 31*, 2 (2012), 14. 2

[SNF05] SIFAKIS E., NEVEROV I., FEDKIW R.: Automatic determination of facial muscle activations from sparse motion capture marker data. *ACM TOG 24*, 3 (2005), 417–425. 2

[SP04] SUMNER R. W., POPOVIĆ J.: Deformation transfer for triangle meshes. In *ACM Transactions on Graphics (TOG)* (2004), vol. 23, ACM, pp. 399–405. 3

[TKC00] TIAN Y.-L., KANADE T., COHN J.: Robust lip tracking by combining shape, color and motion. In *Proceedings of the 4th Asian Conference on Computer Vision (ACCV'00)* (January 2000). 2

[TW90] TERZOPOULOS D., WATERS K.: Physically-based facial modelling, analysis, and animation. *Computer Animation and Virtual Worlds 1*, 2 (1990), 73–80. 2

[WLL04] WANG S.-L., LAU W. H., LEUNG S. H.: Automatic lip contour extraction from color images. *Pattern Recognition 37* (2004), 2375–2387. 2

[XCLT14] XU F., CHAI J., LIU Y., TONG X.: Controllable high-fidelity facial performance transfer. *ACM Transactions on Graphics (TOG) 33*, 4 (2014), 42. 2

[ZNA07] ZIMMERMANN J., NEALEN A., ALEXA M.: Silsketch: automated sketch-based editing of surface meshes. In *Proceedings of the 4th Eurographics workshop on Sketch-based interfaces and modeling* (2007), ACM, pp. 23–30. 2