Corrective 3D Reconstruction of Lips from Monocular Video

Pablo Garrido[1] Michael Zollhöfer[1] Chenglei Wu[2] Derek Bradley[3]
Patrick Pérez[4] Thabo Beeler[3] Christian Theobalt[1]

[1]Max Planck Institute for Informatics [2]ETH Zurich [3]Disney Research [4]Technicolor

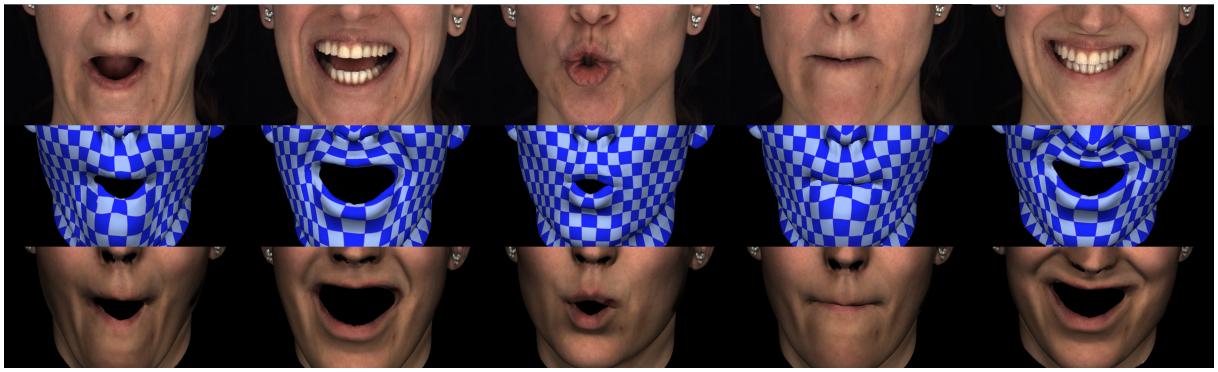


Figure 1: *We present a novel approach to extract high quality lip shapes (bottom) from just monocular video footage (top) based on a learned regression function. Our approach reconstructs expressive mouth motions, such as a kiss or expressions with rolling lips, with high fidelity.*

Abstract

In facial animation, the accurate shape and motion of the lips of virtual humans is of paramount importance, since subtle nuances in mouth expression strongly influence the interpretation of speech and the conveyed emotion. Unfortunately, passive photometric reconstruction of expressive lip motions, such as a kiss or rolling lips, is fundamentally hard even with multi-view methods in controlled studios. To alleviate this problem, we present a novel approach for fully automatic reconstruction of detailed and expressive lip shapes along with the dense geometry of the entire face, from just monocular RGB video. To this end, we learn the difference between inaccurate lip shapes found by a state-of-the-art monocular facial performance capture approach, and the true 3D lip shapes reconstructed using a high-quality multi-view system in combination with applied lip tattoos that are easy to track. A robust gradient domain regressor is trained to infer accurate lip shapes from coarse monocular reconstructions, with the additional help of automatically extracted inner and outer 2D lip contours. We quantitatively and qualitatively show that our monocular approach reconstructs higher quality lip shapes, even for complex shapes like a kiss or lip rolling, than previous monocular approaches. Furthermore, we compare the performance of person-specific and multi-person generic regression strategies and show that our approach generalizes to new individuals and general scenes, enabling high-fidelity reconstruction even from commodity video footage.

Keywords: Lip Shape Reconstruction, Radial Basis Function Networks, Face Modeling, Facial Performance Capture

Concepts: •**Computing methodologies** → **Reconstruction;** *Shape modeling; Supervised learning by regression;*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

1 Introduction

When designing virtual humans and creatures, animation artists pay particular attention to the quality and realism of the facial animation. In recent years, approaches to capture dense shape, motion and appearance of real faces have been developed in computer graphics and vision. Nowadays, animation artists can use captured facial performances as a baseline when creating facial animations, which drastically simplifies their workflow.

Many state-of-the-art passive facial performance capture methods enable high-quality, dense, static [Beeler et al. 2010; Ghosh et al. 2011] and dynamic [Bradley et al. 2010; Beeler et al. 2011; Klaudiny and Hilton 2012] reconstruction of the human face from multi-view data. They capture the geometry of the entire face, or specifically the eyelids [Bermano et al. 2015], the eyeball [Bérard et al. 2014], facial hair [Beeler et al. 2012], or scalp hair [Luo et al. 2012; Echevarria et al. 2014; Hu et al. 2015]. More recently, even monocular methods were developed that capture dense face geometry from monocular RGB [Suwajanakorn et al. 2014; Shi et al. 2014; Cao et al. 2015; Garrido et al. 2016; Thies et al. 2016] or RGB-D [Hsieh et al. 2015; Thies et al. 2015] video.

Unfortunately, none of these methods accurately captures the incredible range of shapes and deformations of moving lips. In particular, expressive mouth motions, such as a kiss or expressions with rolling lips, are almost impossible to reconstruct, even with multi-view methods in controlled studios. Further still, subtle lip shape differences that may disambiguate a friendly smile from a smirk, are very hard to capture. Passive photogrammetric reconstruction of lips is fundamentally hard due to several of their intrinsic properties: Lips are almost featureless in appearance, specular, show subsurface scattering, and exhibit very quick and shape-dependent changes of blood flow. They are highly deformable, their skin strongly stretches

Request permissions from permissions@acm.org. © 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM.
SA '16 TechnicalPapers,, December 05 - 08, 2016, , Macao
ISBN: 978-1-4503-4514-9/16/12
DOI: http://dx.doi.org/10.1145/2980179.2982419

and compresses, and they exhibit strong self-occlusions complicating surface tracking. Contour-based tracking is another option to estimate lip shapes. However, while the outer contour of the lips corresponds to a fixed ring on the face, the inner contour is an *occlusion boundary* (or so-called "rolling contour") and is not associated to any fixed location on the lips, making contour-based shape tracking very challenging.

Yet, accurate animation of the lip motion of virtual humans is of paramount importance. Face-to-face communication is multi-modal and combines the visual and auditory channels. Subtle visible nuances in face and mouth expressions can strongly influence interpretation of speech and intent, and exact mouth motion is essential for the hearing-impaired relying on lip reading. A video with a purposefully modified lip motion can even make us hear a different vowel, even if the audio channel is playing the correct one - an effect known as the McGurk effect [Nath and Beauchamp 2012]. Thus, animation artists spend a lot of time and effort to adjust incorrectly captured lips.

Only few passive methods so far explicitly addressed lip shape reconstruction. Bradley et al. [2010] use a dense multi-view passive performance capture system with controlled lighting and improve lip reconstruction with detected contour constraints. However, they ignore the rolling nature of the inner contour and thus sacrifice accuracy. Bhat et al. [2013] combine blend-shape tracking and out-of-model deformation for improved lip shape tracking from a multi-camera face helmet, however they use markers and manual labeling of occluded contours. Anderson et al. [2013] use a multi-camera photometric stereo system, and match lip contours with predefined iso-lines on the surface mesh. All of these approaches require professional camera setups and the lip shapes that can be acquired are still limited, e.g. lip rolling remains a challenge to capture.

We therefore present the first automatic method to passively capture detailed expressive lip geometry along with the dense geometry of the entire face, from just monocular RGB video. Our first contribution is the adaptation of a state-of-the-art multi-view face performance capture system such that it reconstructs high-quality 3D lip geometry, including rolling and skin stretching (Section 3). This is accomplished by adding additional features to the lips through the application of an artificial color pattern. Using this setup, we record a training set of high-quality 3D face and mouth motions of several individuals, along with RGB video. We believe that our work is the first to generate an accurate lip shape dataset of this quality.

Our second contribution is a new model-based facial performance capture method. At its core is a new regression method based on a radial basis function (RBF) network trained on the aforementioned database, see Section 5 and Fig. 2. It learns the difference between inaccurate shapes of lips found with a state-of-the-art monocular face performance capture method [Garrido et al. 2016], and the true 3D shapes of lips (and the surrounding face region) reconstructed by the high-quality multi-view system (Section 4). Rather than resorting to unreliable photometric features, we use shape features computed from extracted inner and outer lip contours as input to a robust gradient domain regression strategy. We thus indirectly exploit the relation between lip contour shape and 3D lip shape, without having to explicitly assign 2D-3D correspondences during model-based face tracking. We quantitatively and qualitatively show that our monocular approach can capture detailed shape and motion of lips, even rolling lips and kiss shapes, at much higher quality than with previous monocular methods. We compare the performance of person-specific and multi-person generic regression models and show that our approach generalizes to unseen individuals and general scenes, enabling high-fidelity reconstruction even from mobile phone videos (Section 6).

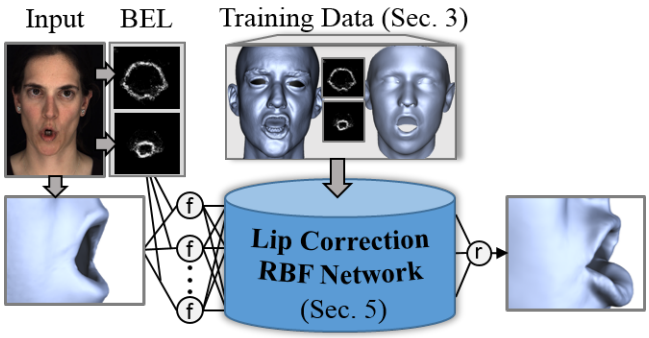


Figure 2: ***Pipeline*** *- Lip correction radial basis function network.*

2 Related Work

Capturing the geometry, appearance and motion of the human head has received a lot of attention. In the following, we highlight selected related work.

Digital Head Models and Head Reconstruction. Photo-realistic digital faces can be realized by a combination of high-resolution 3D scanning and video-based animation [Alexander et al. 2010; Alexander et al. 2013]. In general, facial identity and appearance is often represented based on a 3D parameteric shape prior [Blanz and Vetter 1999] or a 2D active appearance model (AAM) [Cootes et al. 2001]. Expression changes are parameterized based on blendshapes [Lewis and Anjyo 2010] or physics-based muscle simulation [Sifakis et al. 2005]. The combination of both the identity and expression dimension in one holistic multi-linear model [Vlasic et al. 2005], has also been proposed.

High-quality static [Beeler et al. 2010; Ghosh et al. 2011] and dynamic [Bradley et al. 2010; Beeler et al. 2011; Klaudiny and Hilton 2012] 3D face models can be passively reconstructed from multi-view imagery. In addition to the facial geometry, eyelids [Bermano et al. 2015], the eyeball [Bérard et al. 2014], facial hair [Beeler et al. 2012], skin reflectance properties [Wenger et al. 2005; Weyrich et al. 2006] and micro-structure detail [Graham et al. 2013; Nagano et al. 2015], or scalp hair [Luo et al. 2012; Echevarria et al. 2014; Hu et al. 2015] can be captured at high fidelity using such professional setups. Other approaches aim for light-weight acquisition of game-quality avatars, capturing the entire head and its animation from multi-view imagery and a video captured with a mobile device [Ichim et al. 2015]. Garrido et al. [2016] build a personalized animatable 3D face rig that models person-specific idiosyncrasies from just monocular video. All these approaches struggle with accurate reconstruction of expressive lip motion.

Face Performance Capture and Applications. Facial performances can be captured offline in controlled studios [Borshukov et al. 2003; Pighin and Lewis 2006], optionally with invisible makeup [Williams 1990] or facial markers [Guenter et al. 1998; Bickel et al. 2007; Huang et al. 2011]. Many markerless approaches use multi-view video data [Bradley et al. 2010; Beeler et al. 2011; Valgaerts et al. 2012; Fyffe et al. 2014] or input from active triangulation scanners [Wang et al. 2004b; Weise et al. 2009]. Multi-view video performance capture has been demonstrated using a combination of dense flow and a sparse set of correspondences between the input and static scans [Fyffe et al. 2014]. Generic blendshape models can also be tracked from captured stereo data [Valgaerts et al. 2012] and single-view RGB-D video [Weise et al. 2011]. High-fidelity performances have been reconstructed by combining marker-based

motion capture with a minimal set of face scans [Huang et al. 2011]. Some recent methods obtain high-quality facial animations by fitting a 3D template model to monocular video data and estimating fine-scale detail from shading cues [Garrido et al. 2013; Shi et al. 2014; Suwajanakorn et al. 2014]. Again, these approaches do not succeed at reconstructing accurate lip shapes.

In contrast to these offline approaches, several real-time techniques were developed. Weise et al. [2011] track a parametric blendshape model with a commodity depth camera. Chen et al. [2013] propose a tracking approach based on non-rigid mesh deformation. 3D regression can be used to infer a sparse set of 3D landmarks just from monocular data [Cao et al. 2014], or to regress fine-scale transient surface detail [Cao et al. 2015], i.e. expression wrinkles. Tracking and shape accuracy can be improved based on corrective layers [Bouaziz et al. 2013; Li et al. 2013], and face occlusions, e.g. from hair strands [Hsieh et al. 2015], can be resolved.

Face reconstruction enables face reenactment [Suwajanakorn et al. 2015; Vlasic et al. 2005], image-based puppetry [Kemelmacher-Shlizerman et al. 2010] and face replacement [Dale et al. 2011]. Garrido et al. [2015] propose a method for virtual dubbing. They explicitly enforce lip closure based on detected phonemes, but do not obtain high-quality lip shapes. Thies et al. [2015; 2016] reenact RGB-(D) videos at real-time rates. They do not explicitly handle the tracking of lips leading to non-photorealistic results if they undergo strong shape changes. The quality of all these approaches would greatly improve if high-quality 3D lip shapes could be reconstructed from just monocular data.

Lip Tracking and Occluding Contours. The challenging appearance and deformation range of lips makes lip detection and tracking from monocular video hard. Many approaches focus on 2D contour lines and do not recover dense 3D information. Nguyen et al. [2009] use a semi-adaptive appearance model to track lips in image data. Lip contours are often tracked based on 2D snakes and pattern matching [Barnard et al. 2002]. Also semi-automatic approaches based on snakes and a parametric contour model have been proposed [Eveno et al. 2004] to obtain accurate results. Tian et al. [2000] propose a model-based approach to lip tracking. They use a multi-state mouth prior and their alignment strategy is based on shape, color and motion. Tracked lips can improve speech recognition accuracy [Kaucic and Blake 1998; Wang et al. 2004a].

In contrast to these 2D tracking approaches, recent multi-view reconstruction methods extract and model the dense 3D shape of the lips in controlled studio conditions. Kawai et al. [2014] propose a photo-realistic approach for inner mouth restoration by fitting animations to speech signals. Bradley et al. [2010] do high-resolution passive facial performance capture and improve lip tracking based on edge detection and silhouette alignment contraints. Bhat et al. [2013] record with a head mounted multi-camera rig and face markers, and use a combination of blendshape fitting and out-of-model deformation with manually annotated occluding lip contours. Anderson et al. [2013] propose an automatic approach for tracking lips using an iterative alignment strategy. The alignment constraints make the occluding contour of the lips match to automatically selected predefined isolines on the mesh's surface. All these methods require multi-camera input and controlled recording conditions, yet many of them still struggle with strongly occluded and very expressive mouth shapes. Liu et al. [2015] propose a data-driven approach that fuses RGB-D video and audio for real-time mouth shape refinement. However, the reconstructed mouth shapes lack expresiveness and fall short in accuracy. Hence, our method is the first to capture highly expressive lip shapes from monocular video in general surroundings.

3 Data Collection

In this work, our goal is to enhance lightweight face capture methods, in particular by improved reconstruction of the lips (and the adjacent mouth region). Lips tend to be one of the most challenging facial regions, especially for under-constrained capture approaches such as monocular reconstruction. Our approach constructs a training database of high-quality lip shapes and learns a regression function that explicitly maps approximate lip shapes from a lightweight capture method to high-quality and accurate shapes.

3.1 High-Quality Lip Database

We start by building a database of high-resolution 3D lip shapes with the state-of-the-art reconstruction method of Beeler et al. [2011], which uses a multi-view camera setup and controlled studio lighting to produce high-resolution face meshes that are in full vertex correspondence over time. For our application, we configure the physical setup (see Fig. 3.a) such that four cameras are directly focused and zoomed in onto the lip region (one stereo pair from above and one from below), and six additional cameras (three stereo pairs) frame the entire face. Obtaining highly accurate lip reconstructions even in such a controlled environment can be very challenging, since the lips have very few features and change appearance over time. To overcome this and obtain the best possible 3D data, we apply patterns to the lips via temporary tattoos[1] (see Fig. 3.b), which provide surface disambiguation and consistency of appearance over time, without drastically altering the natural lip motions of the subject. Fig. 3.c shows a subset of reconstructed lip shapes. The ground truth training data is cleaned up as a pre-process, e.g. gums are masked out to remove penetrations. We assign correspondences between our base mesh and the ground truth reconstructions once per subject using the method described in Section 4.1. To ensure good correspondences of the occluding lip contour, we use a reconstruction of the neutral pose with slightly open mouth. The region used for correspondence association is shown in the supplemental document. No further assignment is needed as both meshes preserve temporal correspondence. Our training set, the first of its kind, contains a very high-resolution and accurate lip shape $\mathcal{H}_f$ for each frame f. The lip shapes span a wide range of lip motions including smiling, frowning, smirking, kissing, puffing, rolling in/out, sticky-lips and side-to-side mouth motions. The dataset consists of both transitions in and out of these complex shapes, as well as general speech animations. The complete database of 3289 total shapes captured from 4 different actors is a central contribution of this work.

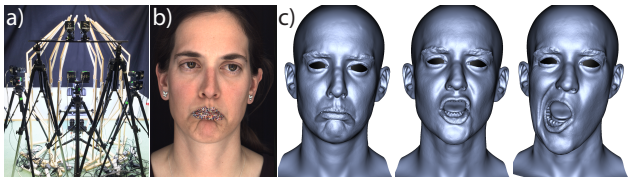


Figure 3: ***High-Quality Lip Database*** *- Using a controlled multi-view capture setup (a), and lip tattoos (b), we reconstruct high-quality lip shapes for training (c).*

3.2 Regression Training Data

Given the acquired high-quality shape $\mathcal{H}_f$ of frame f, we wish to learn the geometric difference between this shape and a coarse approximation $\mathcal{C}_f$ produced by a conventional facial reconstruction method. Our approach is generic and can be applied to enhance

[1] www.violentlips.com

any reconstruction technique. In this paper, we enhance an off-the-shelf lightweight monocular facial tracker [Garrido et al. 2016], but monocular results with other trackers could be similarly enhanced. This model-based face tracker is based on an *analysis-by-synthesis* approach that minimizes the dense photometric error between the model and each of the input frames. As a first step, it jointly computes a coarse estimate of the person's identity (geometry and dense face albedo), face expression, and scene lighting. This reconstruction step is based on a multi-linear face model that spans the identity and expression space and uses a spherical harmonics illumination model. A personalized albedo texture is also generated and used for tracking. The per-frame reconstructions are refined by an *out-of-space* step that estimates a medium scale person-specific corrective layer that allows for higher quality fits. Finally, shading-based geometry refinement (i.e. shape-from-shading under estimated lighting and albedo) can be applied to extract fine-scale static and transient surface details from the RGB input. The tracker captures dynamic face geometry at state-of-the-art quality (see comparisons in [Garrido et al. 2016]), but, like all related monocular methods, struggles to capture expressive lip shapes (Section 6).

In order to compute the shape difference for training, we need to also run the lightweight tracker on the input training data. In this case, we choose one of the frontal cameras of the multi-view setup. However, the applied lip tattoos would lead to a bias when training the regression function, since during testing the tracker will be applied to monocular data without such artificially added features. To alleviate this problem, we digitally inpaint the sequences to remove the tattoos, as described in Section 3.2.1.

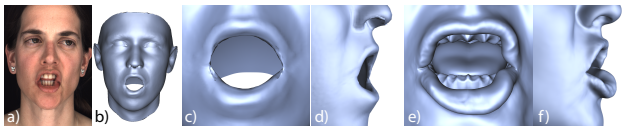


Figure 4: ***Monocular Training Data*** *- After inpainting the lips (a), we apply the monocular face tracker of Garrido et al. [2016] (b). The approximate lip shapes are shown in (c) and (d), while corresponding high-quality reconstructions are given in (e) and (f).*

The difference in lip shape between the monocular $\mathcal{C}_f$ and the high-quality reconstructions $\mathcal{H}_f$ can be seen for one pose in Fig. 4. These differences will be used to train a regression-based lip enhancement algorithm (see Section 4). However, we found that the coarse lip shapes alone are an insufficient feature for robust regression due to the amount of possible ambiguity. For this reason, we add additional image-based constraints, namely lip contour curves, which we detect using semi-supervised learning, as described in Section 3.2.2.

3.2.1 Lip Tattoo Inpainting

Traditional digital inpainting involves replacing corrupt or unwanted image pixels in a semantically meaningful way, typically using surrounding pixels for context. Removing the unwanted lip tattoos is a special case where neither interpolation nor copy operations can generate plausible appearance since the tattoos cover the entire lip region. Fortunately, we have more information available, namely the reconstructed 3D geometry. We can therefore apply a geometry-guided inpainting process by capturing and reconstructing each actor one additional time without tattoos, and copying the lip region from the un-tattooed image to the tattooed sequences. To this end, we record the actor in the high-resolution setup of Beeler et al. [2011] with the mouth slightly open (to avoid occlusions) and without wearing the lip tattoo. During reconstruction of this pose, we use the same mesh topology as in the lip shape database, putting the un-tattooed shape in a dense vertex correspondence with the training data. We construct a UV texture for the un-tattooed lips by projecting the geometry into the camera images. Then, inpainting each tattooed image can proceed by rendering the lip geometry from the viewpoint of the camera using the un-tattooed texture and compositing the output with the image using a feathering operation at the boundaries. The drawback of this approach is that the inpainted lips will always exhibit the same appearance and lack dynamic effects such as shape-dependent shading. However, we can compensate for such effects through a shading-equalization scheme. Specifically, we compute the pixel-wise intensity difference of the lip region between each frame and a reference pose, chosen to be similar to the un-tattooed pose, and then add this frame-dependent appearance change to the un-tattooed texture. An example inpainting is shown in Fig. 5.

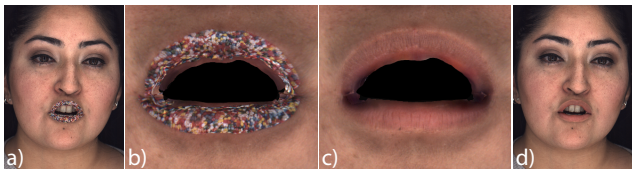


Figure 5: ***Lip Inpainting*** *- Geometry-guided inpainting is performed in UV space to remove the lip tattoos. (a) One image of the training set, (b) the corresponding UV texture, (c) the inpainted UV texture, (d) the final inpainted image after compositing.*

3.2.2 Lip Contour Detection

We aim to increase the robustness of our regression function by adding 2D lip contour features in addition to the 3D geometric features. Lips are almost featureless, highly deformable and their appearance changes due to shape-dependent blood flow patterns. The most reliable visual features of the lips are the inner and outer contours, of which the inner is an occluding contour. We employ the Boosted Edge Learning (BEL) algorithm proposed by Dollar et al. [2006] to automatically detect the contours. BEL is a general-purpose supervised learning algorithm for boundary detection, and we train it separately on a few hand-labeled inner and outer contours for each of the different illumination conditions, respectively. It is based on a large set of generic fast features, which are evaluated over a small image patch, including: gradients, histograms of filter responses, and Haar wavelets at different scales. We train two separate detectors to regress the inner lip contour $\mathcal{B}_f^{\mathcal{I}}$ and outer lip contour $\mathcal{B}_f^{\mathcal{O}}$ likelihood maps (see Fig. 6) for each input frame f. In summary, the collected training data includes high quality 3D lip shapes and detected inner and outer 2D lip contours, and corresponding approximate lip shapes from a lightweight face tracker.

4 Lip Shape Correction Layer

We parametrize the difference between the high-quality reconstructions $\mathcal{H}_f$ and the corresponding coarse monocular reconstructions $\mathcal{C}_f$ in frame f using per-triangle deformation gradients [Sumner and Popovic 2004]. Later on, this differential lip correction layer $\boldsymbol{\ell}_f$ is used to improve on the monocular tracking results.

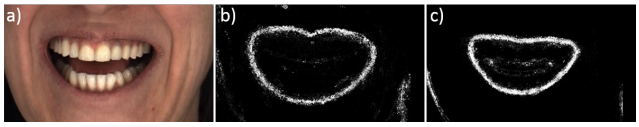


Figure 6: ***Contour Detection*** *- We apply BEL contour detection to the input image (a) to find the outer (b) and inner (c) lip contour.*



Figure 7: ***Lip Correction Layer*** *- The used part of the mesh in red.*

4.1 Dense Correspondence Association

The 3D reconstructions ($\mathcal{H}_f$ and $\mathcal{C}_f$) obtained by the two different reconstruction approaches are not in vertex correspondence, and in general may not even share the same coordinate system. As a first step, we compute a dense set of triangle-to-triangle correspondences based on a fully automatic Laplacian surface registration technique. Since the only common element of the two reconstructions is the input image, we use image-based landmarks as constraints for the deformation. To that end, we use a facial landmark tracker [Saragih et al. 2009; Saragih et al. 2011] to compute a set of 66 sparse landmarks on the inpainted image of the neutral face. Back-projecting the detected landmarks onto both the coarse mesh and the high-quality mesh provide an initial set of surface correspondences. Based on these constraints, we perform Laplacian surface deformation of the coarse mesh, followed by a dense correspondence search based on spatial proximity. Finally, a second Laplacian registration step is performed based on these dense constraints and the resulting alignment is used to establish dense triangle-to-triangle correspondences.

4.2 Gradient-based Lip Shapes

We formulate the shape correction layer in the gradient domain. This is preferable over position-based corrections because the high-quality and coarse meshes may differ by more than just lip shape, e.g. the monocular tracker may also have a slight error in depth, and gradient-based correction is ignorant of such global transformations. The gradient formulation is also advantageous since it allows us to regress improved shapes for only the confined region of the lips and the surrounding mouth area, yet to smoothly blend these improvements with the surrounding face. The gradient-based lip correction layer captures differences in surface orientation, scale and skew for all the T triangles in the lips and the local mouth region, as defined by the mask shown in Fig. 7. In a first step, per-triangle deformation gradients $\mathbf{G}_f^{(t)} \in \mathbb{R}^{3\times 3}$ between the T faces of the mesh $\mathcal{C}_f$ and the corresponding triangles in $\mathcal{H}_f$ are computed. We map from the monocular tracking results to the high-quality reconstructions using a neutral frame (first frame $f = 0$ of the sequence) as anchor point:

$$\mathbf{G}_f^{(t)} = \underbrace{\mathbf{H}_f^{(t)}}_{\mathcal{H}_0 \to \mathcal{H}_f} \cdot \underbrace{\hat{\mathbf{D}}^{(t)}}_{\mathcal{C}_0 \to \mathcal{H}_0} \cdot \underbrace{\left[\mathbf{C}_f^{(t)}\right]^{-1}}_{\mathcal{C}_f \to \mathcal{C}_0}. \quad (1)$$

The deformation gradients $\mathbf{C}_f^{(t)}$ and $\mathbf{H}_f^{(t)}$ model the expression of the monocular and high-quality reconstruction, respectively. The difference in identity, simply caused by the quality difference in the two trackers, is encoded using $\hat{\mathbf{D}}^{(t)}$. The deformation gradients jointly encode the rotation, scale and shear as a single matrix. This will be problematic for regression, as internally the correction layer will be interpolated linearly. For this reason, we extract the individual components as a set of scalar values which can be linearly interpolated. To this end, we compute the polar decomposition [Higham 1986] of the gradient matrix $\mathbf{G}_f^{(t)} = \mathbf{Q}_f^{(t)}\mathbf{S}_f^{(t)}$ factoring into rotation and shear. Following Alexa et al. [2002], we parametrize the rotations using the matrix exponential. Scale and skew are extracted from the shear component. In total, this leads to 9 parameters per triangle which allow for linear interpolation. The lip shape correction layer $\boldsymbol{\ell}_f \in \mathbb{R}^{9T}$ stacks the computed per-face deformation gradients,

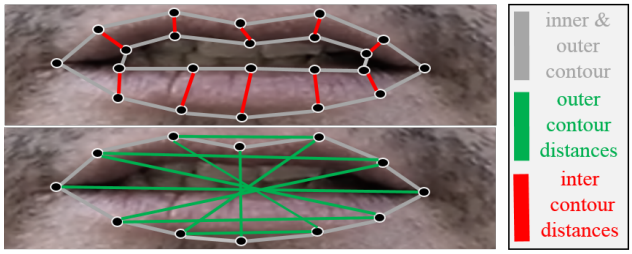


Figure 8: ***Relative Distance Features*** *- We use inner contour (red) and outer contour (green) distances to define the robust features used for lip shape regression.*

and will be the target of our regression framework described in the following section.

5 Lip Shape Regression

We learn the difference between the inaccurate and true 3D lip shape based on a regression function. The captured training data $\mathcal{T} = \{\mathcal{C}_f, \mathcal{H}_f, \mathcal{B}_f^{\mathcal{I}}, \mathcal{B}_f^{\mathcal{O}}, \boldsymbol{\ell}_f\}_{f=0}^{F}$ consists of the inaccurate monocular reconstructions $\mathcal{C}_f$, the accurate multi-view reconstructions $\mathcal{H}_f$, the computed inner $\mathcal{B}_f^{\mathcal{I}}$ and outer $\mathcal{B}_f^{\mathcal{O}}$ BEL contour maps and the corresponding ground truth output layer $\boldsymbol{\ell}_f$.

5.1 Robust Features for Lip Shape Correction

We use a set of discriminative features $\mathbf{f}$ that allow to robustly predict high-quality lip shapes given inaccurate monocular reconstructions. In the feature vector we jointly encode the inaccurate reconstruction result as well as the target contour constraints. We wish to encode the reconstruction in a compact manner, which is also independent of the particular reconstruction method in order to make our approach as general as possible. For this reason, we define a low-dimensional shape subspace ψ by computing Principle Components Analysis (PCA) on the 75 blendshapes used for monocular tracking and keep 99% of the variance. The inaccurate results are then projected to this subspace to obtain a shape vector of length $|\psi| = 33$. Target contour constraints are defined by a set of relative features that take the shape of the detected inner and outer lip contour into account. These features are normalized based on the inter-ocular distance, to make the regression results independent of global depth changes. We sample the inner and outer lip contour based on a search that starts from the monocular reconstruction result. To this end, in a preprocess, we specify iso-lines of the outer lip contour on the template geometry. Starting from sample points on the monocular reconstruction result of the outer contour, we search for the closest maxima in the BEL likelihood maps along the gradient of the iso-lines. The found maxima in the maps $\mathcal{B}_f^{\mathcal{I}}$ and $\mathcal{B}_f^{\mathcal{O}}$ are the corresponding points of the inner and outer contour, respectively. We use the obtained outer and inner contour points to define a set of relative features that encode 10 distances on the outer and 10 distances between the two contours, see Fig. 8. Note, this exploits the correlation between the 2D contours and the actual 3D lip shape. In total, together with the PCA coefficients, the lip feature vector $\mathbf{f}$ has $M = |\psi| + 20 = 53$ components.

5.2 Local Radial Basis Function Networks

Given the per-frame lip features and corresponding lip correction layers $\{\mathbf{f}_f, \boldsymbol{\ell}_f\}_{f=1}^{F}$, we learn a per-triangle regression function $r^{(t)} : \mathbb{R}^M \to \mathbb{R}^9$ using a vector-valued radial basis function (RBF) network. We use a network architecture with a single hidden layer

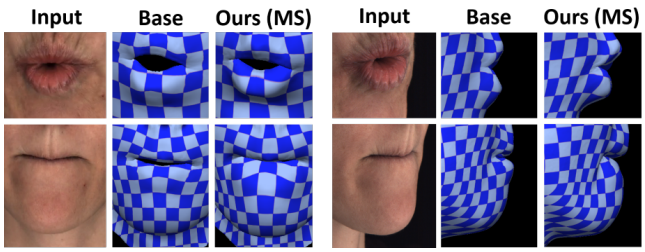


Figure 9: ***Reconstruction Quality*** *- Our RBF Network regression successfully improves a coarse base tracker [Garrido et al. 2016] to better handle stretching, bending and rolling of lips. (Subject: S1)*

(see [Bishop 2006]), the associated $N \ll F$ neurons $\Phi_n : \mathbb{R}^M \rightarrow \mathbb{R}$ have fixed prototypes $\mathbf{p}_n \in \mathbb{R}^M$ in feature space and share the same scale $\beta \in \mathbb{R}$:

$$\Phi_n(\mathbf{f}) = \exp\left(-\beta \cdot ||\mathbf{p}_n - \mathbf{f}||_2^2 \right). \tag{2}$$

Prototypes $\mathbf{p}_n$ are obtained by a temporally uniform sampling of the training sequences. The output node implements a linear weighted summation of the per-neuron activation levels and adds a constant bias parameter $\mathbf{b} \in \mathbb{R}^9$:

$$r^{(t)}(\mathbf{f}) = \left[\sum_{n=1}^{N} \mathbf{w}_n^{(t)} \Phi_n(\mathbf{f}) \right] + \mathbf{b}\,. \tag{3}$$

We tackle the problem of finding the N weights $\mathbf{w}_n^{(t)} \in \mathbb{R}^9$ using ridge regression [Hoerl and Kennard 2000]:

$$\min_{\{\mathbf{w}_n^{(t)}\}_{n=1}^N} \underbrace{\left[\sum_{f=1}^{F} \left\| r^{(t)}(\mathbf{f}_f) - \boldsymbol{\ell}_f^{(t)} \right\|_2^2 \right.}_{E_{data}} + \alpha \cdot \underbrace{\left. \sum_{n=1}^{N} \left\| \mathbf{w}_n^{(t)} \right\|_2^2 \right]}_{E_{reg}}. \tag{4}$$

Here, the data term E_{data} encodes how well the training data is reproduced and the ridge regularizer E_{reg} prevents overfitting. The importance of the regularizer is controlled by the ridge parameter α. Optimal values for α and the scale parameter β are found via cross-validation. Since the optimization problem is quadratic, the minimizer can be found by solving a linear system. Note, the linear system decomposes into 9 independent linear subproblems of size $N \times N$. Since all subproblems share the same system matrix (only the right-hand sides differ), the regression function can be efficiently computed.

Given a new input feature vector $\mathbf{f}$, the corresponding per-triangle correction can be obtained by $\boldsymbol{\ell}^{(t)} = r^{(t)}(\mathbf{f})$. Afterwards, the high-quality lip shape can be reconstructed by integrating the per-triangle deformation fields using deformation transfer [Sumner and Popovic 2004]. Note, we perform all steps in a canonical frame for rotation and translation invariance.

6 Results

We demonstrate the applicability of the proposed method on a variety of different datasets. In addition, we evaluate our design choices and compare with a model-based tracking approach using enhanced lip blend shapes and explicit lip contour alignment constraints. In total, we captured 3 female and 3 male subjects, henceforth referred to as S1-S6. S1-S5 were recorded indoors in the multi-view setup using 4MP machine vision cameras, S5 was not used for training; for S4 and S6 we also have outdoor sequences captured with an iPhone

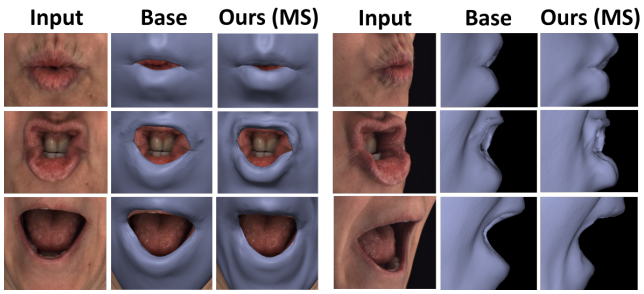


Figure 11: ***Lip Protrusion*** *- Our regressor is more resilient to the depth ambiguity inherently present in monocular tracking and can plausibly reconstruct protruding and rolling lips. (Subject: S1)*

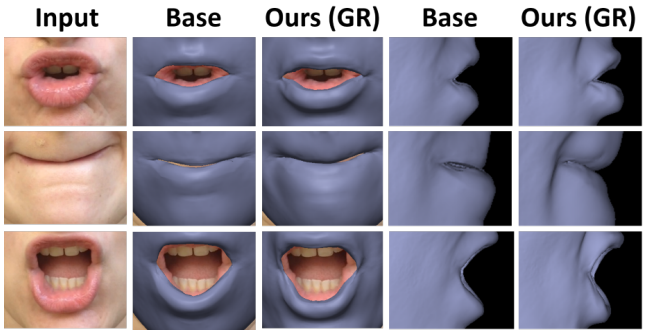


Figure 13: ***3D Lip Shape*** *- Our regressor captures inward and outward rolling of lips given just a monocular video. (Subject: S6)*

camera (resolution 1920×1080, 30 fps). We evaluate our approach on 9 sequences (4 captured using a controlled studio setup and 5 in a general uncontrolled environment). More results are available on the project website[2]. In average the runtime of our method (after training) is approximately 25 sec/frame on an Intel Xeon E5-2637 CPU (3.5 Ghz), where 20 seconds are spent on monocular tracking (previous work) and 5 seconds are added for our new lip correction approach. In all performed experiments, we use every tenth frame of the training set to define the prototype vectors of our RBF lip correction network. All parameters of our regressor remain constant during the experiments. In our evaluation, we use three different types of regressors:

- **PS:** A person-specific regressor trained for a specific subject. This regressor is only applied to sequences of the same subject. Note though that training and testing datasets are disjunct.
- **MS:** A multi person regressor trained on four different subjects (S1-S4). The test subject can be any of the four. Again, training and testing datasets are disjunct.
- **GR:** A generalization regressor trained on three or four subjects (out of S1-S4). The identity of the test subject is not included in the training set.

Improving Monocular Lip Reconstruction. First, we use our novel lip correction network to improve the reconstruction quality of a state-of-the-art monocular face tracker [Garrido et al. 2016], to which we will refer to as base tracker in the following. To this end, we use data captured by one of the frontal cameras of the multi-view setup as well as outdoor video footage captured under general uncontrolled illumination with an iPhone camera. A coarse base

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

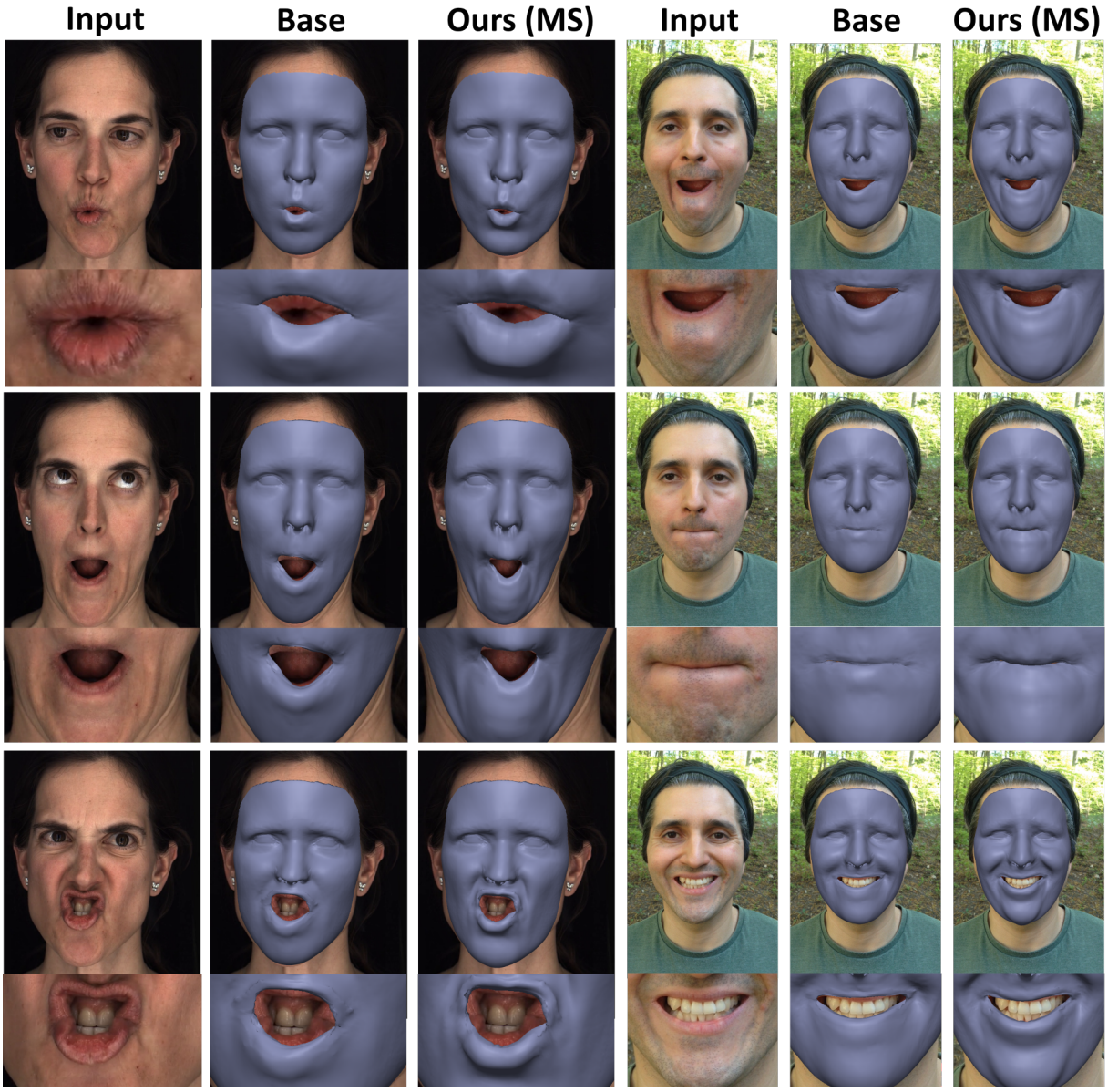


Figure 10: ***Results*** *- Our proposed regression framework substantially improves on the lip shapes reconstructed by the base tracker, the state-of-the-art monocular facial performance capture approach of Garrido et al. [2016]. Note how especially challenging lip motions, such as rolling or stretching, are better captured in the refined results. Our regressor is even able to improve the reconstruction quality of the surrounding area, such as nasolabial folds or the chin. (Subjects: S1 (left), S4 (right))*



Figure 12: ***Generalization to a Novel Subject*** *- Our approach generalizes well to novel subjects and general scenarios. (Subject: S6)*



Figure 14: ***Our RBF Regression Method vs. Augmented Model-based Tracking*** *- Our RBFN outperforms a model-based tracker that uses explicit contour alignment constraints and additional person-specific lip blend shapes. (Subject: S2)*

reconstruction is obtained and the regressed lip correction layer is applied. Fig. 10 shows the coarse base and refined reconstructions for both setups. In this experiment we used the **MS** regressor specified above. As can be seen, the regressor successfully improves the lip shapes of the monocular base reconstruction. Especially, inward and outward rolling of the lips, lip protrusions, and the kiss shape are nicely captured. This is further emphasized in Figs. 11 and 9, which visualize surface stretching and shape change from side views.

Please further note that shape improvements are also visible in the face region surrounding the lips, where the regression adds plausible bulging and folding of the skin, which supports the lip shapes (Fig. 10, middle left). In addition, we applied our generalized regressor (**GR**) to a novel subject captured outdoors under conditions that substantially differ from our training environment, shown in Fig. 12. Our approach generalizes nicely to such uncontrolled scenarios and different illumination conditions, since the shape-based features used for regression are less sensitive to changing environment conditions than photometric cues. Again, inspecting the lips closely and from the side (Fig. 13) clearly shows how the overall shape is improved by our regression strategy.

Generalization Properties of the Regressor. We evaluate the generalization properties of the proposed lip correction RBF network. To this end, we trained a person-specific regressor **PS**, a multi person regressor **MS** and a generalization regressor **GR**. We qualitatively and quantitatively evaluate the accuracy of these three architectures on a test sequence. Fig. 15 shows color coded error maps with respect to ground truth reconstructions. For the corresponding numbers see Table 1. The obtained reconstruction quality is largely independent of the regressor type. This shows that our approach generalizes well to novel subjects and does not require person-specific training data.

Comparison to Model-based Lip Tracking. In order to perform a baseline comparison, we extend the monocular face tracker [Garrido et al. 2016], which also serves as our base tracker, by incorporating explicit lip blendshapes and lip contour alignment constraints. Person-specific lip blenshapes have been computed based on the high-quality multi-view reconstructions. In particular, we transferred 30 user-selected expressive lip shapes to the 3D identity shape estimated by the monocular tracker using deformation transfer [Sumner and Popovic 2004]. Lip contours are detected using BEL and the optimization process tries to align the inner and outer contour of the model with the ridges in the likelihood maps. Since the inner contour is an occluding one, we perform this optimization in an iterative flip-flop fashion. This is similar to the approach used in

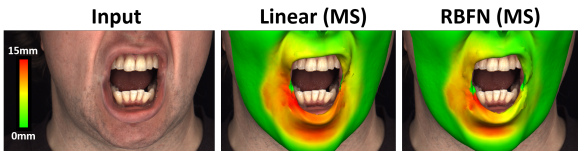


Figure 16: ***RBF Networks vs. Affine Regression*** *- Non-linear regression leads to smaller errors. (Subject: S2)*

Table 1: ***Quantitative Evaluation*** *- Tracking error in cm.*

	Base	Contour	Ours (PS)	Ours (MS)	Ours (GR)
μ	0.40	0.39	0.33	0.30	0.32
σ	0.14	0.12	0.12	0.10	0.11

Anderson et al. [2013]. As can be seen in Fig. 14, our approach obtains higher quality results, which better align to the ground truth. For the numbers see Table 1. This test shows that, in particular for capturing the true rolling and stretching of the lips, even enhancing previous model-based methods with additional image constraints is not sufficient. We obtain a better spatial alignment, more expressive lip shapes and better recover stretching and bending of the lips, without having to tediously augment a parametric 3D expression model per person.

Evaluation of the Regression Strategy. We compare our RBF network with a simple affine regressor, see Fig. 16. Especially surface dynamics are better handled by our non-linear approach. We also quantitatively show this improvement in a cross-validation experiment. To this end, we train both regressors on the same training data, while leaving out a set of validation clips (732 frames). In a first step, we select the best parameters for both regressors using cross-validation. The RBF network performs best for $\alpha = 0.1$ and $\beta = 0.1$. For the affine regressor, the Tikhonov regularization parameter $\alpha = 2.0$ leads to the best results. With these parameters, our RBF network obtains an average feature space error of 0.13 (0.04 standard deviation). In contrast, the affine regressor has a higher average feature error of 0.14 (0.05 standard deviation).

Influence of Input Features on Regression. We also quantitatively evaluate the influence of different input features. To this end, we compare the cross-validation error as well as the tracking error on our ground truth sequence (Subject S2) for different feature descriptors. Table 2 and Table 3 show that the use of both PCA coefficients and relative distance features improves upon descriptors that are only based on one of these two features. This can mainly be attributed

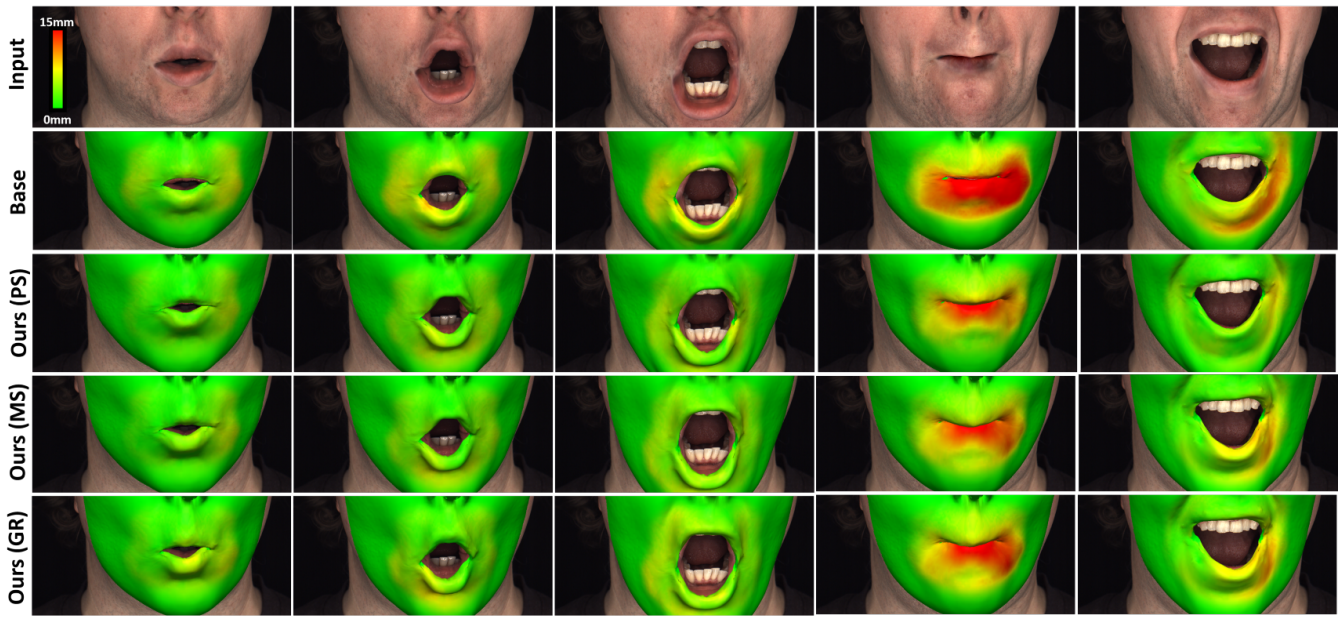


Figure 15: ***Comparison of Generalization*** *- Our RBF network generalizes well beyond the training set. As can be seen, the general regressor* ***GR*** *performs comparable to the* ***PS*** *and* ***MS*** *regressors, when the reconstruction error to the ground truth is evaluated. (Subject: S2)*

Table 2: ***Influence of Features*** *- Cross-validation error.*

	PCA Coeff.	Relative Distances	Combined
μ	0.14	0.17	0.13
σ	0.05	0.06	0.04

Table 3: ***Influence of Features*** *- Tracking error in cm.*

	PCA Coeff.	Relative Distances	Combined
μ	0.32	0.33	0.30
σ	0.12	0.13	0.10

to certain ambiguities that can not always be resolved by relative distances or lip shape geometry alone, e.g., symmetrically-consistent lip deformations or depth-involving lip shapes, respectively.

7 Limitations

We demonstrate high-quality lip shape reconstructions even for challenging and expressive mouth motions, such as a kiss or rolling lips. While we achieve these compelling results just based on monocular video, our approach still has some limitations. First, our approach is based on a set of collected training data, as such it shares the limitation of learning based approaches: It does not generalize well to situations that are drastically outside the span of used examples, i.e. for faster or more expressed motions than used for training. In such a case, the training set can be extended by capturing more data. Such an extension requires the availability of a high-quality multi-view setup, but has to be done only once. Also, while the employed feature descriptor is invariant to translation, it is not rotation invariant and handling different head rotations would thus require an extensive amount of additional training data. This problem could be alleviated by compensating for rigid head motion before computing the features, i.e. projecting the sample points back to a tracked plane or the template mesh and computing the contour distances in 3D. BEL is sensitive to lighting and strong color changes. Thus, we currently re-train the detector for the individual illumination conditions. This could in theory be overcome with a sufficiently large dataset containing these variations. Alternatively, this could be alleviated by using a different contour detection strategy that is more robust to these variations, which is easily possible since our algorithm does not directly depend on the chosen structure detector. Our shape features are only based on geometric properties and are therefore invariant to these situations. Drastic appearance changes (e.g., dark vs. pale skin color or beards) could be handled in a similar manner. Mild facial hair is normally captured as high-frequency detail by both the multi-view reconstruction and the baseline algorithm. As such it can be decoupled from the coarse lip motion estimation. Thick beards and occlusions can make our approach fail, since a robust detection of the lip contours would not always be possible. In general, we believe that the obtained reconstruction quality can be further improved by increasing the amount of training examples. As demonstrated, we are able to regress the shape of the lips very well, but since our features are translation invariant, an accurate alignment to the input data cannot be guaranteed. In many applications this is not of paramount importance, i.e. lip reading or movie dubbing, but future work could address this, for example, by incorporating the detected contours as reprojection constraints into the gradient based reconstruction strategy.

8 Conclusion

We present an approach to fully automatically reconstruct expressive lip shapes along with dense geometry of the entire face, from just monocular RGB data. At the core of our approach is a novel robust regression function that learns the difference between inaccurate lip shapes and true 3D lip shapes based on a captured database of high and low quality reconstructions. Rather than resorting to unreliable photometric features, we use shape features computed from extracted inner and outer lip contours. We show that our monocular approach reconstructs higher quality lip shapes, even for lip rolling or kiss shapes, than previous monocular approaches.

Since subtle visible nuances in face and mouth expression strongly influence the interpretation of speech and intent, we anticipate that our approach will be particularly useful for applications that deal

with audiovisual content i.e. movie dubbing and lip reading.

Acknowledgements

We thank all the reviewers for their valuable comments and our actors for their time and patience to capture the training/testing data. We are also grateful to Angiels Diaz for kindly labeling the 2D lip contours to train the BEL detector and Tobias Bertel for helping with the video. This work was supported by the ERC Starting Grant CapReal (335545) and by Technicolor.

References

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

ALEXANDER, O., ROGERS, M., LAMBETH, W., CHIANG, J., MA, W., WANG, C., AND DEBEVEC, P. E. 2010. The digital emily project: Achieving a photorealistic digital actor. *IEEE CGAA 30*, 4, 20–31.

ALEXANDER, O., FYFFE, G., BUSCH, J., YU, X., ICHIKARI, R., JONES, A., DEBEVEC, P., JIMENEZ, J., DANVOYE, E., ANTIONAZZI, B., EHELER, M., KYSELA, Z., AND VON DER PAHLEN, J. 2013. Digital Ira: Creating a real-time photoreal digital actor. In *ACM Siggrah Posters*.

ANDERSON, R., STENGER, B., AND CIPOLLA, R. 2013. Lip tracking for 3D face registration. In *Proc. MVA*, 145–148.

BARNARD, M., HOLDEN, E. J., AND OWENS, R. 2002. Lip tracking using pattern matching snakes. In *Proc. ACCV*, 1–6.

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

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

BEELER, T., BICKEL, B., NORIS, G., MARSCHNER, S., BEARDSLEY, P., SUMNER, R. W., AND GROSS, M. 2012. Coupled 3D reconstruction of sparse facial hair and skin. *ACM TOG 31*, 4, 117:1–117:10.

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

BERMANO, A., BEELER, T., KOZLOV, Y., BRADLEY, D., BICKEL, B., AND GROSS, M. 2015. Detailed spatio-temporal reconstruction of eyelids. *ACM TOG 34*, 4, 44:1–44:11.

BHAT, K. S., GOLDENTHAL, R., YE, Y., MALLET, R., AND KOPERWAS, M. 2013. High fidelity facial animation capture and retargeting with contours. In *Proc. ACM SCA*, 7–14.

BICKEL, B., BOTSCH, M., ANGST, R., MATUSIK, W., OTADUY, M. A., PFISTER, H., AND GROSS, M. H. 2007. Multi-scale capture of facial geometry and motion. *ACM TOG 26*, 3, 33:1–33:10.

BISHOP, C. M. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

BLANZ, V., AND VETTER, T. 1999. A morphable model for the synthesis of 3D faces. In *Proc. ACM Siggraph*, 187–194.

BORSHUKOV, G., PIPONI, D., LARSEN, O., LEWIS, J. P., AND TEMPELAAR-LIETZ, C. 2003. Universal capture: Image-based facial animation for "The Matrix Reloaded". In *ACM SIGGRAPH 2003 Sketches & Applications*.

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

BRADLEY, D., HEIDRICH, W., POPA, T., AND SHEFFER, A. 2010. High resolution passive facial performance capture. *ACM TOG 29*, 4, 41:1–41:10.

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

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

CHEN, Y.-L., WU, H.-T., SHI, F., TONG, X., AND CHAI, J. 2013. Accurate and robust 3D facial capture using a single RGBD camera. In *Proc. ICCV*, 3615–3622.

COOTES, T. F., EDWARDS, G. J., AND TAYLOR, C. J. 2001. Active appearance models. *IEEE Trans. Pattern Anal. Machine Intell. 23*, 6, 681–685.

DALE, K., SUNKAVALLI, K., JOHNSON, M. K., VLASIC, D., MATUSIK, W., AND PFISTER, H. 2011. Video face replacement. *ACM TOG 30*, 6, 130:1–130:10.

DOLLÁR, P., TU, Z., AND BELONGIE, S. 2006. Supervised learning of edges and object boundaries. In *Proc. CVPR*, 1964–1971.

ECHEVARRIA, J. I., BRADLEY, D., GUTIERREZ, D., AND BEELER, T. 2014. Capturing and stylizing hair for 3D fabrication. *ACM TOG 33*, 4, 125:1–125:11.

EVENO, N., CAPLIER, A., AND COULON, P. Y. 2004. Accurate and quasi-automatic lip tracking. *IEEE Trans. Circuit and Systems for Video Tech. 14*, 5, 706–715.

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

GARRIDO, P., VALGAERTS, L., WU, C., AND THEOBALT, C. 2013. Reconstructing detailed dynamic face geometry from monocular video. *ACM TOG 32*, 6, 158:1–158:10.

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

GARRIDO, P., ZOLLHÖFER, M., CASAS, D., VALGAERTS, L., VARANASI, K., PÉREZ, P., AND THEOBALT, C. 2016. Reconstruction of personalized 3D face rigs from monocular video. *ACM TOG 35*, 3, 28:1–28:15.

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

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

GUENTER, B., GRIMM, C., WOOD, D., MALVAR, H., AND PIGHIN, F. 1998. Making faces. In *Proc. ACM Siggraph*, 55–66.

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

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

HSIEH, P.-L., MA, C., YU, J., AND LI, H. 2015. Unconstrained realtime facial performance capture. In *Proc. CVPR*, 1675–1683.

HU, L., MA, C., LUO, L., AND LI, H. 2015. Single-view hair modeling using a hairstyle database. *ACM TOG 34*, 4, 125:1–125:9.

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

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

KAUCIC, R., AND BLAKE, A. 1998. Accurate, real-time, unadorned lip tracking. In *Proc. ICCV*, 370–375.

KAWAI, M., IWAO, T., MAEJIMA, A., AND MORISHIMA, S. 2014. Automatic photorealistic 3D inner mouth restoration from frontal images. In *Proc. ISVC*, 51–62.

KEMELMACHER-SHLIZERMAN, I., SANKAR, A., SHECHTMAN, E., AND SEITZ, S. M. 2010. Being John Malkovich. In *Proc. ECCV*, 341–353.

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

LEWIS, J., AND ANJYO, K.-I. 2010. Direct manipulation blendshapes. *IEEE Comp. Graphics and Applications 30*, 4, 42–50.

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

LIU, Y., XU, F., CHAI, J., TONG, X., WANG, L., AND HUO, Q. 2015. Video-audio driven real-time facial animation. *ACM Trans. Graph. 34*, 6, 182:1–182:10.

LUO, L., LI, H., PARIS, S., WEISE, T., PAULY, M., AND RUSINKIEWICZ, S. 2012. Multi-view hair capture using orientation fields. In *Proc. CVPR*, 1490–1497.

NAGANO, K., FYFFE, G., ALEXANDER, O., BARBIČ, J., LI, H., GHOSH, A., AND DEBEVEC, P. 2015. Skin microstructure deformation with displacement map convolution. *ACM TOG 34*, 4, 109:1–109:10.

NATH, A. R., AND BEAUCHAMP, M. S. 2012. A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *NeuroImage 59*, 1, 781–787.

NGUYEN, Q. D., AND MILGRAM, M. 2009. Semi adaptive appearance models for lip tracking. In *Proc. ICIP*, 2437–2440.

PIGHIN, F., AND LEWIS, J. 2006. Performance-driven facial animation. In *ACM Siggraph Courses*.

SARAGIH, J. M., LUCEY, S., AND COHN, J. F. 2009. Face alignment through subspace constrained mean-shifts. In *Proc. ICCV*, 1034–1041.

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

SHI, F., WU, H.-T., TONG, X., AND CHAI, J. 2014. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM TOG 33*, 6, 222:1–222:13.

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

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

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

SUWAJANAKORN, S., SEITZ, S. M., AND KEMELMACHER-SHLIZERMAN, I. 2015. What makes Tom Hanks look like Tom Hanks. In *Proc. ICCV*, 3952–3960.

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

THIES, J., ZOLLHÖFER, M., STAMMINGER, M., THEOBALT, C., AND NIESSNER, M. 2016. Face2Face: Real-time face capture and reenactment of RGB videos. In *Proc. CVPR*.

TIAN, Y.-L., KANADE, T., AND COHN, J. F. 2000. Robust lip tracking by combining shape, color and motion. In *Proc. ACCV*, 1–6.

VALGAERTS, L., WU, C., BRUHN, A., SEIDEL, H.-P., AND THEOBALT, C. 2012. Lightweight binocular facial performance capture under uncontrolled lighting. *ACM TOG 31*, 6, 187:1–187:11.

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

WANG, S. L., LAU, W. H., AND LEUNG, S. H. 2004. Automatic lip contour extraction from color images. *Pattern Recogn. 37*, 12, 2375–2387.

WANG, Y., HUANG, X., SU LEE, C., ZHANG, S., LI, Z., SAMARAS, D., METAXAS, D., ELGAMMAL, A., AND HUANG, P. 2004. High resolution acquisition, learning and transfer of dynamic 3D facial expressions. *CGF 23*, 3, 677–686.

WEISE, T., LI, H., GOOL, L. J. V., AND PAULY, M. 2009. Face/Off: Live facial puppetry. In *Proc. ACM SCA*, 7–16.

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

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

WEYRICH, T., MATUSIK, W., PFISTER, H., BICKEL, B., DONNER, C., TU, C., MCANDLESS, J., LEE, J., NGAN, A., JENSEN, H. W., AND GROSS, M. 2006. Analysis of human faces using a measurement-based skin reflectance model. *ACM TOG 25*, 3, 1013–1024.

WILLIAMS, L. 1990. Performance-driven facial animation. In *Proc. ACM Siggraph*, 235–242.