Rigid Stabilization of Facial Expressions

Thabo Beeler
Disney Research Zurich

Derek Bradley
Disney Research Zurich

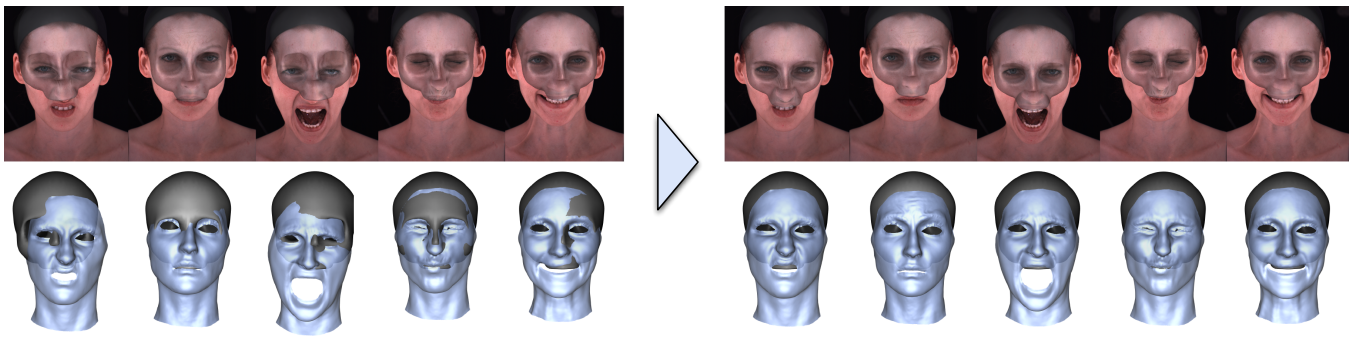

Figure 1: *Captured facial expressions always contain a superposition of rigid transformation due to head motion on top of the non-rigid deformation caused by the expression (left: captured shapes). Rigid stabilization estimates and removes this rigid transformation given only observations of the non-rigidly deforming skin, which allows to automatically extract the pure facial expression (right: stabilized shapes).*

Abstract

Facial scanning has become the industry-standard approach for creating digital doubles in movies and video games. This involves capturing an actor while they perform different expressions that span their range of facial motion. Unfortunately, the scans typically contain a superposition of the desired expression on top of un-wanted rigid head movement. In order to extract true expression deformations, it is essential to factor out the rigid head movement for each expression, a process referred to as *rigid stabilization*. In order to achieve production-quality in industry, face stabilization is usually performed through a tedious and error-prone manual process. In this paper we present the first automatic face stabilization method that achieves professional-quality results on large sets of facial expressions. Since human faces can undergo a wide range of deformation, there is not a single point on the skin surface that moves rigidly with the underlying skull. Consequently, computing the rigid transformation from direct observation, a common approach in previous methods, is error prone and leads to inaccurate results. Instead, we propose to indirectly stabilize the expressions by explicitly aligning them to an estimate of the underlying skull using anatomically-motivated constraints. We show that the proposed method not only outperforms existing techniques but is also on par with manual stabilization, yet requires less than a second of computation time.

CR Categories: I.3.3 [Computer Graphics]: Picture/Image Generation—Digitizing and scanning;

Keywords: Rigid Stabilization, Face Scanning.

Links: DL PDF

1 Introduction

Human facial animation is one of the most important and widespread topics of computer graphics. However, it is also one of the most challenging tasks, since audiences are well-trained to identify even the slightest inaccuracies in facial performances, which can lead to strong feelings of unfamiliarity and the infamous *uncanny valley* effect. Despite several decades of research, we continue to strive for perfectly realistic digital 3D face animation. In current practices, facial animation is typically performed using a blendshape face rig, which consists of a set of face shapes that span the range of desired expressions of the character. Using this rig, new poses for animation can be created by blending different amounts of the expression shapes together. Clearly, the quality of the final animation depends highly on the quality of the underlying blendshapes.

Face blendshapes can be created through manual sculpting, which is common for creatures and other fictional characters. However, for human-like characters the blendshapes are usually reconstructed by scanning real actors performing the expressions. High-resolution digital face scanning is a growing trend in the entertainment industry. This can be attributed to the increasing demand for photorealistic digital actors, coupled with recent advances in high-quality facial reconstruction [Ma et al. 2007; Beeler et al. 2010; Bradley et al. 2010; Alexander et al. 2010; Beeler et al. 2011; Huang et al. 2011; Ghosh et al. 2011; Fyffe et al. 2011]. In addition to entertainment demands, facial expression capture is a key element of statistical face analysis, for example in the recent FaceWarehouse database [Cao et al. 2013a], and the latest trend is to capture actor-specific blendshape rigs for real-time facial animation [Weise et al. 2011; Bouaziz et al. 2013; Li et al. 2013; Cao et al. 2013b].

A major problem that arises when scanning actors performing different expressions is that the resulting scans contain both expression movement as well as rigid head movement, since the actor cannot keep their head perfectly still while performing a wide range of expressions (Figure 1, left). If the expression shapes contain this ”baked-in” rigid motion, then any facial animation or statistical analysis constructed from the expressions will also contain un-wanted rigid head motion. Therefore, the scanned expressions must be rigidly aligned to a common frame of reference - which is

essentially the skull. This alignment process is referred to as *stabilization*. In order to achieve production-quality in industry, face stabilization is usually performed through a tedious and error-prone manual process, during which an artist iteratively translates and rotates the captured expression, trying to infer the change in shape with respect to a reference pose. This 6-DOF alignment process requires a lot of experience, and may take several man-months of work for current movie productions. With the demand for realistic digital doubles likely to increase in the coming years, manual stabilization will quickly become a bottleneck in production.

Proper stabilization is very important in order to avoid artifacts in facial animation, retargeting, anatomical simulation, eye tracking and statistical analysis, yet so far the problem has received very little attention. Despite the growing field of work that relies on scanned facial expressions, stabilization has been typically only approximated or ignored completely. In this work, using anatomically-motivated constraints, we propose the first automatic face stabilization method that achieves professional-quality results on large sets of facial expressions (Figure 1, right).

Contrary to previous methods which try to estimate the rigid transformation directly from the observed skin, we propose to indirectly stabilize expressions by explicitly aligning them to a common actor-specific skull. This actor-specific skull is computed by deforming a generic skull given a set of facial landmarks. The same set of landmarks is further used in conjunction with the skull to establish actor-specific anatomical constraints, which guide the automatic stabilization.

The main contributions of this paper are:

1. We propose the first method to stabilize facial expressions at production-level quality.
2. Our method is fast and fully automatic, after a simple one-time actor initialization.
3. We automatically infer anatomical properties, such as the underlying skull, from facial scans.
4. We demonstrate the quality of the results both quantitatively and qualitatively on several actors, and show that the proposed method not only outperforms existing techniques but is also on par with manual stabilization.

2 Related Work

Despite the importance of separating rigid and non-rigid components of facial scans, the problem of facial expression stabilization has received little attention to date, and is generally performed through a tedious manual process in industry.

One of the main applications for expression stabilization is in the construction of actor-specific blendshape models for facial animation, which is a growing trend in recent years. Li et al. [2010] propose to build a full set of blendshapes from a reduced set of facial expressions. In their work, the expressions are aligned using their previous algorithm for non-rigid registration of shapes [Li et al. 2009], however they do not consider the stabilization problem at all. Cao et al. [2013a] also use example-based facial rigging [Li et al. 2010] to build blendshapes for their FaceWarehouse dataset. Huang et al. [2011] find the minimum set of scans required to build a blendshape rig that is capable of reproducing a given low-resolution motion capture sequence at high resolution. Again, stabilization is ignored.

Actor-specific blendshapes are also popular in real-time facial animation, but related work does not focus on the stabilization problem. Weise et al. [2009] aim to remove the rigid motion component of multiple expressions using ICP registration [Besl and McKay 1992], but this does not perform well on non-rigidly deforming faces. Bouaziz et al. [2013] also use ICP but attempt to improve on its shortcomings by restricting the alignment to forehead and nose regions. Still, this approach assumes those face regions will never deform in a non-rigid way, and thus fails when the forehead or nose is deformed. In the 3D shape regression work of Cao et al. [2013b], blendshapes are created with the help of their FaceWarehouse dataset, which, as mentioned, is not properly stabilized. Finally, Li et al. [2013] remove the need for stabilization since they do not scan multiple expressions, but instead roughly approximate them through deformation transfer [Sumner and Popović 2004] using a generic face rig. This approach comes at the cost of losing important actor-specific geometric details, which can only be obtained by face scanning. Thus far, the approaches for stabilizing actor-specific blendshapes for real-time animation are insufficient for high-quality facial animation rigs.

In the area of statistical analysis, related work aims to build linear models of scanned faces in correspondence. In order to generate accurate statistics across different scans, the faces should first be stabilized. Blanz and Vetter [1999] build a 3D morphable model of faces, but stabilize the geometry using simple normalization of poses. Vlasic et al. [2005] compute multilinear face models, which are further used for video face replacement [Dale et al. 2011]. They stabilize the rigid motion using Procrustes alignment [Gower 1975], which computes the least-squares fit of the neutral face to each expression using vertex correspondences. As before, these approaches only approximate the true stabilization.

As we will show in comparison, previous methods are all insufficient for generating production-quality blendshapes, since stabilization is either only approximated or ignored altogether. With existing techniques, the rigid component of the face motion cannot be fully removed, which entails tedious manual stabilization by production artists in industry. Our approach is the first to perform automatic rigid stabilization of extreme facial expressions, which we accomplish by considering anatomical constraints, aiming to separate the rigid motion of the skull from the non-rigid deformation of the face tissue.

3 Method Overview

Since human faces can undergo a wide range of deformation, there is not a single point on the surface that moves rigidly with the underlying skull. Consequently, computing the rigid transformation from direct observation, a common approach in previous methods, is error prone and leads to inaccurate results. However, the relative motion of the skin to the underlying skull as well as the change in shape is constrained by human anatomy. Skin slides over the skull and buckles as a consequence of underlying muscular activity. We argue therefore that it is advantageous to explicitly fit the skull to the expressions considering these anatomical constraints. The rigid transformation between any two given expressions can then be computed from the transformations of the skull. More formally, given a reference shape $\hat{\mathcal{F}}$ with corresponding skull $\hat{\mathcal{S}}$ and a deformed shape $\mathcal{F}$, the goal is to determine the underlying rigid transformation T of the skull such that it fits $\mathcal{F}$. Transforming $\mathcal{F}$ by the inverse T^{-1} yields the desired stabilization with respect to the reference shape $\hat{\mathcal{F}}$.

The proposed pipeline consists of two main stages as depicted in Figure 2. In a first stage, described in Section 4, we register and non-rigidly deform a generic skull to the neutral shape of the actor, and initialize anatomically-motivated constraints such as volume or shape preservation. With these constraints we then stabilize

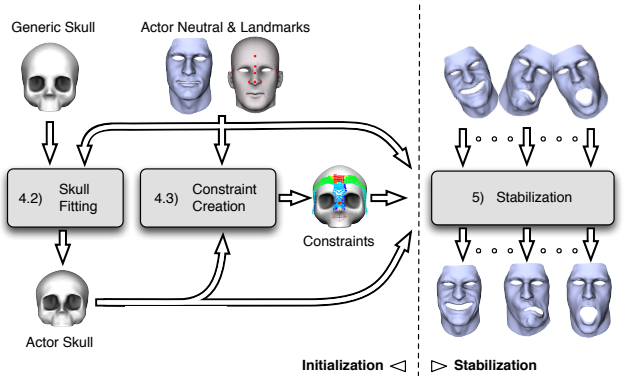

Figure 2: Method Overview - *During initialization (Section 4) we fit a generic skull to the actor's neutral shape (Section 4.2) and then establish actor-specific constraints (Section 4.3). After initialization, the skull and constraints are employed to automatically stabilize input shapes (Section 5).*

each deformed shape in a non-linear optimization framework as described in Section 5.

4 Actor Initialization

The purpose of the first stage is to set up the underlying anatomically-motivated constraints that will drive the stabilization. This involves annotating a sparse set of facial landmarks (Section 4.1), fitting a generic skull to the actor's face (Section 4.2) and establishing the constraints (Section 4.3). These steps are performed only once per actor.

In this work, all facial scans were captured using the method of Beeler et al. [2011], however our algorithm can stabilize any set of facial expressions in correspondence.

4.1 Face Annotation

Our method requires a sparse set of facial landmarks annotated on the neutral shape of an actor. Figure 3 depicts the landmarks, which consist of the sides of the head, the forehead, between-eyes, as well as the bridge, tip and sides of the nose. As the annotation is only required once per actor, we specify the landmarks manually, however this step could be automated using a landmark detection algorithm such as Amberg and Vetter [2011].

4.2 Skull Fitting

Five of the landmarks introduced in the previous step - *forehead*, *between-eyes*, *nose-bridge*, *head-left*, and *head-right* - have correspondences on the generic skull. As the skull lies underneath the skin, the corresponding features of the skull can be found along the normal direction, at a distance δ equal to the tissue thickness at these points. Typical thicknesses were retrieved from a human CT scan, and are listed in Figure 3. We compute the rigid transformation $T^{\mathcal{S}}$ to align the generic skull to the neutral face by minimizing the sum of euclidean distances between the correspondences [Arun et al. 1987]. We account for differences in scale by scaling the skull landmarks such that the average distance to the barycenter corresponds to that of the facial landmarks.

After transforming the skull by $T^{\mathcal{S}}$ we deform it non-rigidly to fit the face using iterative linear shell deformation [Botsch and Sorkine 2008] employing the aforementioned correspondences $\mathcal{C}$ as hard

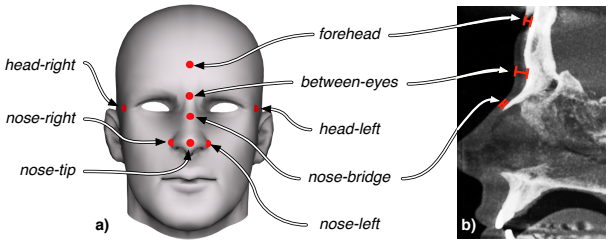

Figure 3: Facial Landmarks - *We manually annotate a set of landmarks on the face (a) which are used to fit the skull and create the nose constraint. To fit the skull, we empirically determined the tissue thickness at a subset of these landmarks based on CT scans (b), namely forehead (∼4.5mm), between-eyes (∼7mm), nose-bridge (∼2mm), head-left (∼3.5mm), and head-right (∼3.5mm).*

constraints. The linear shell minimizes bending and stretching energies for the displacements $\mathbf{d}$:

$$\begin{aligned} \min_{\mathbf{d}} \quad & E_{shell}(\mathbf{d}) + \sum_{i \in \mathcal{S}} \omega_i \left\| \delta_i \mathbf{n}_i - \mathbf{d}_i \right\|^2 \\ \text{s.t.} \quad & \mathbf{d}_c = \mathbf{x}_c - \mathbf{x}_c^s - \mathbf{n}_c^s \delta_c; \; c \in \mathcal{C}. \end{aligned} \tag{1}$$

Initially, we set all ω_i to zero, suppressing soft constraints. These will be introduced in subsequent iterations. The hard deformation constraints $\mathbf{d}_c$ are given by the difference of the position $\mathbf{x}_c$ on the face to the corresponding position $\mathbf{x}_c^s$ on the skull, offset along its normal $\mathbf{n}_c^s$ by the typical tissue thickness δ_c at this point. The deformed skull is only guaranteed to fit the face at the sparse constraint points and might still penetrate the skin in other areas. However, it yields a good initialization for the following iterative optimization, which will ensure a good fit everywhere. In every iteration, we compute for every vertex $\mathbf{x}_i^s$ on the skull the distance $\delta_i^\star$ along the normal $\mathbf{n}_i^s$ to the surface of the face. Depending on $\delta_i^\star$ we define soft deformation constraints δ_i and the according weights ω_i as

1. if $\delta_i^\star \equiv \inf$ $\quad \rightarrow \delta_i = 0, \quad \omega_i = 0,$
2. if $\delta_i^\star < \delta_{min}$ $\quad \rightarrow \delta_i = \delta_{min}, \; \omega_i = \lambda[(\delta_i^\star - \delta_{min})^2 + 1],$
3. if $\delta_i^\star > \delta_{max}$ $\quad \rightarrow \delta_i = \delta_{max}, \; \omega_i = \lambda/[(\delta_i^\star - \delta_{max})^2 + 1],$
4. otherwise $\quad \rightarrow \delta_i = \delta_i^\star, \quad \omega_i = \lambda.$

Intuitively, the function softly constrains the skin thickness to lie within $[\delta_{min}, \delta_{max}]$. While it puts a lot of emphasis on preventing the skin thickness from becoming too thin (2), its influence quickly decays the thicker the tissue becomes (3) allowing the skull to retain its original shape in these areas. Given these soft constraints we apply again the linear shell deformation to update the skull deformation. These steps are repeated until the the deformation converges, which is typically achieved after only a few iterations (∼2-3). Figure 4 shows the generic and deformed skulls for three actors with different facial anatomy. We use $\delta_{min} = 2$mm, $\delta_{max} = 7$mm and $\lambda = 1$ for all results in this paper.

This method produces a tighter skull fit than purely affine tranfomation [Ali-Hamadi et al. 2013] while still ensuring that the skull will not penetrate the skin.

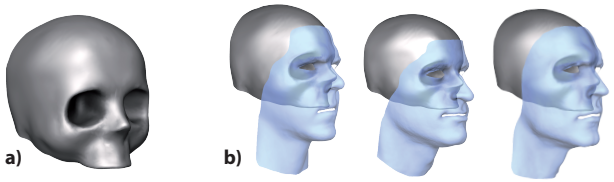

Figure 4: Skull Fitting - *A generic skull (a) is deformed to fit underneath the facial anatomy of three different actors (b).*

4.3 Constraint Creation

We propose two anatomical constraints that should be satisfied when fitting a skull to facial expressions. The first is a *skin constraint*, which attempts to maintain a certain distance between the skull and the skin, while incorporating changing tissue thickness due to deformation. The second is a *nose constraint*, which constrains the distance between the tip of the nose and the skull, considering the amount of strain on the nose. Our framework could include any number of other constraints, such as the position of the upper teeth (when available), which transform rigidly with the skull. However, as shown in Section 6, the proposed two types of constraints are sufficient for high-quality stabilization.

4.3.1 Skin Constraint

During facial deformation, skin is stretched and compressed as a consequence of muscular activity while it slides over the skull (Figure 5 (b)). If we assume that the volume within a small patch of skin remains constant, we can predict the skin thickness $h(\mathbf{x})$ at a given position $\mathbf{x}$ as function of the surface area ratio $\xi(\mathbf{x})$ and rest state skin thickness $\hat{h}(\mathbf{x})$ as

$$h(\mathbf{x}) = \xi(\mathbf{x})\hat{h}(\mathbf{x}) = \frac{\hat{A}(\mathbf{x})}{A(\mathbf{x})}\hat{h}(\mathbf{x}), \tag{2}$$

where $A(\mathbf{x})$ is the surface area, approximated by the weighted average of discs centered at $\mathbf{x}$ through neighboring vertices $\mathbf{x}_i$. We can then rewrite the area ratio as

$$\xi(\mathbf{x}) = \frac{1}{\sum_{i\in\mathcal{N}(\mathbf{x})} w_i} \sum_{i\in\mathcal{N}(\mathbf{x})} w_i \frac{\|\mathbf{x}_i - \mathbf{x}\|}{\|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}\|}, \tag{3}$$

where the weights w_i are computed according to

$$w_i = \frac{(\ell - \|\mathbf{x}_i - \mathbf{x}\|)^2}{\ell^2} \tag{4}$$

for vertices in the neighborhood $\mathcal{N}(\mathbf{x})$ of $\mathbf{x}$ that are closer than 2ℓ to $\mathbf{x}$. Equation 4 favors vertices at distance ℓ from $\mathbf{x}$ and smoothly attenuates the influence of vertices that are closer or farther away, rendering it robust against small positional noise. We set $\ell = 1$cm for all actors in this paper.

As may be expected, our assumption of constant local tissue volume will not be satisfied everywhere on the face. In particular, when muscles bulge, local volume can increase. To account for this, we define a spatially varying weight mask for enforcing the skin constraint higher or lower in different facial regions, guided by anatomy. We observed that thin tissue areas without underlying muscles, such as the bridge of the nose, fulfill the assumption best. The mask, as shown in Figure 5, contains the values ρ, which will be used to weight the influence of the skin constraints in Section 5.

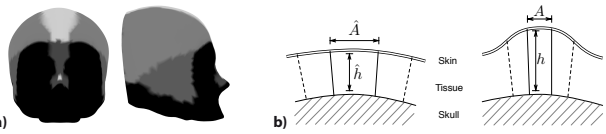

Figure 5: Skin Constraint - *(a) Shows the weights ρ for the skin constraints (the brighter, the higher the weight). These weights are designed once on the generic skull using anatomical considerations, such as muscle and tissue distribution. (b) The underlying assumption for the skin constraint is that the volume of the tissue remains constant at a point on the skull. Thus, if the area A changes due to stretch or compression, the height h should be adjusted accordingly.*

4.3.2 Nose Constraint

Our second constraint considers the nose. The lower part of the nose consists of cartilage covered by a thin layer of skin tissue. Especially around the nose tip, skin sliding is minimal and we can consider the skin attached to the underlying cartilage. Unfortunately, cartilage is not rigid and thus any motion causes an elastic deformation of the nose. This deformation, however, is relatively well-defined. In particular, we observed that the point $\mathbf{x}_t$ at the tip of the nose primarily exhibits a rotation around the nose tip $\mathbf{x}_t^s$ on the skull, with only little compression and stretching. This is illustrated in Figure 6 (a). The small spheres mark the position of the tip of the nose for several ground truth shapes, which were manually stabilized to the skull (refer to Section 6.2). The lines and coloring indicate the discrepancy of the points to the predicted distance $\hat{\ell}_{nose}$. We design the nose constraint to preserve the distance

$$\ell_{nose} = \nu\hat{\ell}_{nose} = \nu \left\|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^s\right\|, \tag{5}$$

where ν is an estimation of the compression of the nose. The compression is estimated from the Cauchy strains between a subset of the landmarks specified in Section 4.1. The Cauchy strain measures the change in length of a vector with respect to its rest length. The strains between *nose-bridge* and *nose-left* (e^{b-l}), *nose-bridge* and *nose-right* (e^{b-r}), as well as *nose-bridge* and *nose-tip* (e^{b-t}) are proportionally related to the nose compression. For example, a snarl will compress the nose and consequently decrease the strain e^{b-t}. The strains between *nose-tip* and *nose-left* (e^{t-l}), and *nose-tip* and *nose-right* (e^{t-r}) are considered to be inversely proportional to the compression, due to the way the underlying muscles affect the skin in this area. For example, raising the cheeks will pull back the nose, increasing the strains e^{t-l} and e^{t-r} while the nose itself compresses. We therefore estimate the nose compression as

$$\nu = 1 + 0.2\left(e^{b-l} + e^{b-r} + e^{b-t} - e^{t-l} - e^{t-r}\right). \tag{6}$$

While this heuristic is not an anatomically accurate representation, it predicts the nose compression well in most cases as can be seen in Figure 6 (b).

5 Stabilization

Given the anatomical constraints defined in Section 4, we now present our method for automatically stabilizing facial expressions. We start by pre-stabilizing an input expression $\mathcal{F}$ by computing the rigid transformation that best aligns the same subset of the landmarks as used for skull fitting in Section 4.2. This rough alignment provides a good initialization for the following optimization.

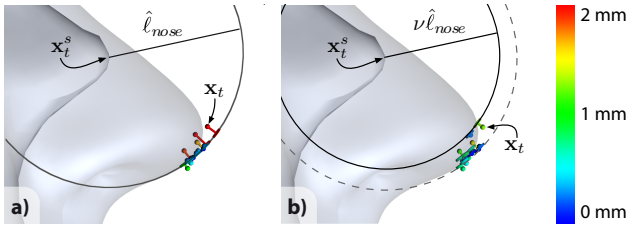

Figure 6: Nose Constraint - *(a) The small spheres mark the position of the tip of the nose for several ground truth shapes. The lines and coloring indicate the discrepancy to the predicted distance $\hat{\ell}_{nose}$. (b) Incorporating a model for the nose compression ν improves the predicted distance $\nu\hat{\ell}_{nose}$ and reduces the discrepancy further.*

Stabilization is cast as a non-linear optimization, minimizing the energy

$$E_{tot} = \lambda_{skin} E_{skin} + \lambda_{nose} E_{nose}, \tag{7}$$

over the translation $\mathbf{t}$ and the rotation $\mathbf{r}$, a total of six degrees of freedom. We could either align the face to the skull or the skull to the face. Both approaches are similar in complexity and performance, but we found empirically that aligning the skull to the face is advantageous, because sampling density and distribution of the skin constraints over the skull surface remain constant even if the face exhibits extreme deformations. In the following we describe the individual energy terms, which are weighted equally for all results in this paper ($\lambda_{skin} = \lambda_{nose} = 1$).

5.1 Skin Energy

The skin energy is chosen such that it tolerates sliding over the skull but penalizes deviation from the predicted tissue thickness. It is defined over all points on the skull, as

$$E_{skin} = \sum_{i \in \mathcal{S}} w_{skin}\left(\mathbf{x}_i, \rho_i\right) \left\| \left(\mathbf{x}_i - \mathbf{x}_i^s\right) \mathbf{n}_i^s - \xi(\mathbf{x}_i)\hat{h}_i \right\|^2. \tag{8}$$

The terms in Equation 8 are given as

$$\begin{aligned} \mathbf{x}_i^s &= T(\mathbf{r}, \mathbf{t})\hat{\mathbf{x}}_i^s \\ \mathbf{n}_i^s &= T(\mathbf{r}, \mathbf{0})\hat{\mathbf{n}}_i^s \\ \mathbf{x}_i &= \chi\left(\mathcal{F}, \mathbf{x}_i^s, \mathbf{n}_i^s\right) \\ \hat{h}_i &= \left\| \hat{\mathbf{x}}_i^s - \chi\left(\hat{\mathcal{F}}, \hat{\mathbf{x}}_i^s, \hat{\mathbf{n}}_i^s\right) \right\|, \end{aligned}$$

where $T(\mathbf{r}, \mathbf{t})$ denotes the transformation given rotation $\mathbf{r}$ and translation $\mathbf{t}$ vectors, $\hat{\mathbf{x}}^s$ and $\hat{\mathbf{n}}^s$ are the skull position and normal in the reference frame, $\xi\left(\mathbf{x}\right)$ computes the stretch at position $\mathbf{x}$ as defined in Equation 3, and $\chi(\mathcal{F}, \mathbf{x}, \mathbf{n})$ computes the first intersection with the shape $\mathcal{F}$ of a ray starting at the point $\mathbf{x}$ in direction $\mathbf{n}$.

Constant volume does not hold in general, e.g. due to muscle bulging. The more skin compresses or stretches, the less accurate this assumption will be. We therefore reduce the weight w_{skin} of a skin constraint depending on the stretch $\xi(\mathbf{x})$ as

$$w_{skin}(\mathbf{x}, \rho) = \frac{\rho}{\kappa_{skin}\left(\xi(\mathbf{x}) - 1\right)^2 + 1} \tag{9}$$

where ρ is the user-specified weight as defined in Section 4.3.1 and κ_{skin} is a user provided parameter that controls how quickly the weight decays with increasing stretch. We set $\kappa_{skin} = 1$ for all results in this paper.

5.2 Nose Energy

The nose energy penalizes deviation from the predicted nose length. It is defined as

$$E_{nose} = w_{nose}\left(\nu\right) \left\| \left(\left\| \mathbf{x}_t - T(\mathbf{r}, \mathbf{t})\hat{\mathbf{x}}_t^s \right\| \right) - \ell_{nose} \right\|^2 \tag{10}$$

where $\mathbf{x}_t$ denotes the tip of the nose on the deformed shape, $\hat{\mathbf{x}}_t^s$ is the position of the nose tip on the skull at the reference frame and as for the skin constraints, $T(\mathbf{r}, \mathbf{t})$ denotes the transformation given rotation $\mathbf{r}$ and translation $\mathbf{t}$ vectors. The estimated nose length ℓ_{nose} and compression ν are computed as described in Equations 5 and 6, respectively. The predicted nose length ℓ_{nose} is just an approximation and will be less accurate the more the nose compresses or stretches. Therefore, we reduce the influence of the nose constraint based on the estimated compression ν as

$$w_{nose}\left(\nu\right) = \frac{1}{\kappa_{nose}(\nu - 1)^2 + 1} \tag{11}$$

where κ_{nose} is a user provided parameter that controls how quickly the weight decays with increasing compression. We set $\kappa_{nose} = 1$ for all results in this paper.

The resulting combination of energy terms yields a non-linear optimization problem, which we solve using the Levenberg-Marquart algorithm. Convergence is quick (typically ~10-20 iterations) and since every shape is stabilized independently, the proposed method is well suited to efficiently stabilize large datasets. Even though the method does not incorporate any temporal continuity, stabilizations are temporally consistent as shown in Section 6.

6 Results

In this Section we present both quantitative and qualitative results of our rigid stabilization method.

6.1 The Upper Teeth Indicator

Assessing the results qualitatively poses a significant challenge, since even millimeter inaccuracies in stabilization will be visible in dynamic facial animations, but are difficult to visualize in print. Fortunately, since the upper teeth are rigidly attached to the skull, they can be used to assess the performance of our method (and others) whenever they are visible in the images. If a stabilization method successfully aligns a model of the teeth to the images for each expression, it provides a good indication that the stabilization is accurate. Furthermore, since we do not incorporate the teeth as constraints in our optimization, the quality achieved on expressions where they are visible can be considered representative for all expressions, whether teeth are visible or not.

To reconstruct a model of the teeth in 3D, we manually draw the outlines of the eight frontal upper teeth in four cameras for one of the expressions where they are visible, which we will refer to as Teeth-Frame $\mathcal{F}_T$ in the following. These outlines are then triangulated to produce the outline in 3D, which we transform rigidly into the reference frame. Figure 7 shows the drawn outlines (a) and

the triangulated shape (b). This teeth model can now be used for qualitative evaluation of our results.

Figure 7: Teeth Reconstruction - *To be able to qualitatively assess the stabilization, we reconstruct the upper teeth by manually drawing their outlines in four different views (a). These outlines are then triangulated and reconstructed in 3D (b).*

6.2 Ground Truth Generation

We quantitatively evaluate our method by comparing to ground truth data, which consists of a subset of 15 manually-stabilized shapes for one actor. Since the alignment of the upper teeth gives additional queues for the stabilization, manual fitting was performed only on those shapes where the upper teeth were visible in the images, and our teeth model was made available to the artist.

The artist stabilized the same set of shapes a second time without using the teeth model, which provides a measurable indication of the quality achievable by manual stabilization on expressions where the teeth are not visible.

6.3 Evaluation

Given our ground truth data and the upper teeth as an indicator of quality, we now evaluate our method in comparison to previous work and manual stabilization when no teeth are used. For previous methods, we consider ICP, used by Weise et al. [2009] and Bouaziz et al. [2013], and also Procrustes alignment, used by Vlasic et al. [Vlasic et al. 2005] and Dale et al. [Dale et al. 2011]. As can be seen in Table 1, the proposed method performs significantly better than previous techniques and even outperforms the human operator. For both ICP and Procrustes, we consider only the upper part of the face as suggested by [Bouaziz et al. 2013] to avoid negative influence of the jaw and neck motion. Without this masking, those algorithms perform considerably worse, as we indicate in the table. Therefore, for the rest of this evaluation we will compare only to the masked versions of ICP and Procrustes, in order to provide the best possible comparison. We further evaluate the contribution of the two individual constraints. The skin constraint alone yields an average mean squared error (MSE) of 1.07. The nose constraint on its own is under-constrained and so we combined it with the Procrustes method to be able to assess its contribution, which yields an average MSE of 1.96. Both constraints on their own give a lower error than previous methods (all >2) but are not as effective as when combined (0.89).

To visualize the quality of the stabilizations, we apply the rigid transformations T to the teeth model and project the outline into the respective images. A comparison with the previous techniques is shown in Figure 8. Note that the manual annotation and reconstruction is not perfectly accurate in itself and that the results must therefore be assessed relative to the Teeth-Frame $\mathcal{F}_T$.

In general, ICP and Procrustes show similar average performance, and both exhibit problems for expressions where the shape changes substantially, for example when wrinkling the forehead or scrunching the nose (Figure 9, left and middle column). These algorithms do not perform well for such expressions since they estimate the transformation directly from the observed skin. In contrast, our method is able to make use of anatomical constraints and provides better results (Figure 9, right column).

Method	Mean [mm]	StdDev [mm]	Max [mm]
ICP	2.17	1.14	4.52
-no mask	3.20	2.51	8.54
Procrustes	2.16	1.13	4.52
-no mask	3.50	1.69	6.38
Manual (no teeth)	1.15	0.57	2.10
Ours	**0.89**	**0.49**	**2.06**

Table 1: Quantitative Results - *As an error measure we compute the mean squared error (MSE) for every expression and list mean, standard deviation and maximal MSEs over all expressions for the different methods. The shapes stabilized by hand using the teeth as reference are considered ground truth. Our method performs significantly better than existing techniques and even outperforms the same human operator when not using the teeth.*

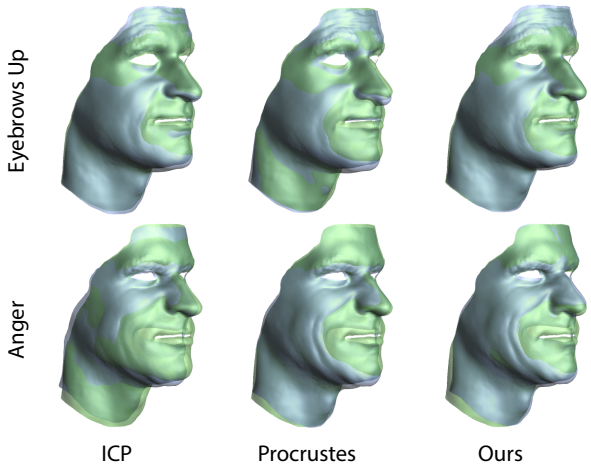

Figure 9: Shape Comparison - *We compare ICP, Procrustes and Our method for two challenging expressions ('Anger' and 'Eyebrows Up'). Results are shown by overlaying the stabilized expression (blue) on the reference pose (green). Both ICP and Procrustes suffer from large misalignments since they try to estimate the transformation directly from the observed skin, which is heavily deforming. The proposed method performs much better since it uses anatomical constraints to estimate the transformation indirectly. This can be verified by inspecting corresponding expressions in the top row of Figure 12.*

Figure 12 shows stabilizations for some extremal expressions for different actors. All actors were stabilized with the same set of parameters and exhibit comparable quality, which demonstrates the robustness of the proposed method. Figure 10 shows stabilizations for some frames out of a longer sequence. Results are temporally much more consistent than previous techniques even though the stabilization is performed on each frame independently without any explicit temporal continuity. All of our evaluation results are best seen in the accompanying video material.

6.4 Application

As a final result, we applied our algorithm to stabilize facial expressions that are used to build a blend-shape facial animation model. An artist then constructed an animation, and we directly transferred

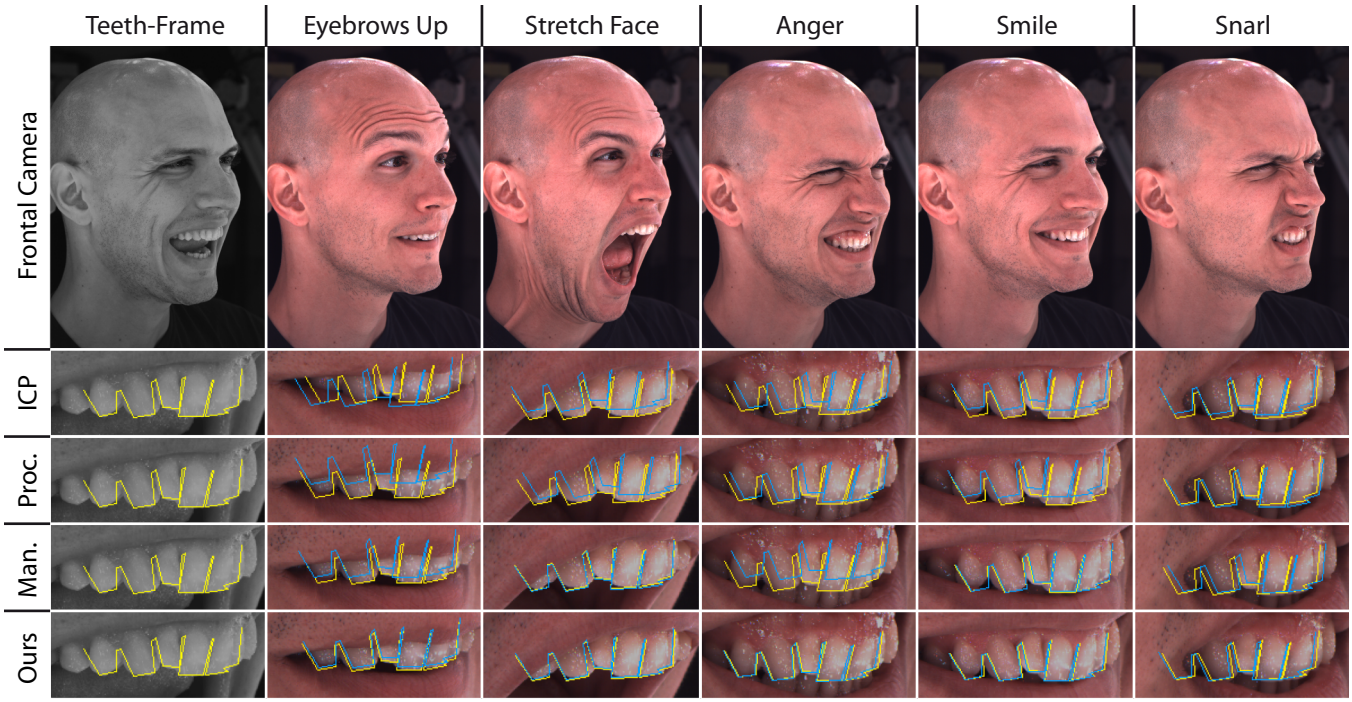

Figure 8: Comparisons - *We show qualitative comparisons of the results produced by ICP, Procrustes, Manual without teeth and Ours (from top to bottom) marked in blue with ground truth marked in yellow. Ground truth was generated by manually stabilizing the shapes using the teeth as additional queue. The first column shows the reconstructed teeth as reference, the others show results on various expressions. Our method visually outperforms previous techniques and even the human operator, which is quantitatively supported by the findings in Table 1.*

the blendshape weights to replica models built after stabilization using ICP and Procrustes, respectively. The resulting animations using the previous techniques contain un-wanted rigid motion that the artist would not be expecting, caused by errors in stabilization. We demonstrate one example of this in Figure 11, which shows two frames of the animation comparing the blendshape model stabilized by Procrustes and the one stabilized by our method. For Procrustes-stabilized blendshapes, the chin moves down whenever the character wrinkles his forehead. Our method, on the other hand, enables artifact-free animation. We would like to refer the reader to the accompanying video material, since the artifacts are much more apparent and extremely disturbing in animations.

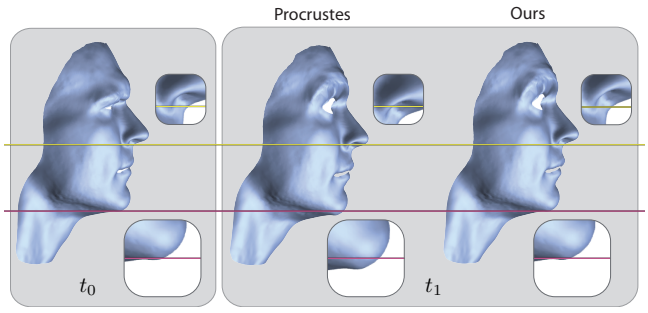

Figure 11: Blendshape Application - *The input expressions to a blendshape rig must be stabilized in order to avoid un-wanted rigid motion during animation. Here we show two frames of animation, t_0 and t_1. Previous methods, such as Procrustes (center), do not succeed in stabilizing and as a consequence the jaw will move down whenever this character wrinkles his forehead. Our method (right) allows for artifact-free animation. While these effects seem subtle on print media, they are extremely disturbing in animations and we would like to refer the reader to the accompanying video material.*

7 Discussion

We present the first method capable of automatically stabilizing facial expressions at a level of quality on par with human operators, which is currently the only viable solution for high-quality productions. Already, the time and effort required for rigid stabilization can be in the order of several man-months for a single production, and with the increasing demand for digital doubles, face stabilization will quickly become a bottleneck in coming years. The proposed method not only provides consistent high-quality results and major time savings, but will also facilitate other research directions such as anatomical simulation or simplified eye tracking.

Our method can be applied to stabilize entire performances, but note that we explicitly refrained from incorporating temporal continuity since it might not always be available, e.g. when capturing only extremal poses of an actor. Integrating temporal continuity could be an interesting extension.

Similarly, our approach leverages only two sets of constraints to explicitly avoid the use of constraints that are not visible in *every* expression, such as the upper teeth. The proposed framework, however, is general and any number of other constraints could be incorporated. For example, additionally aligning the upper teeth could provide even better quality for those shapes where they are present. For the proposed constraints to be effective, the method requires an accurate measure of the skin and nose deformation. In our current implementation, this requires dense correspondences, which might not be available in traditional marker-based facial motion capture. Extending the method to work with such sparse data would be a very valuable direction for future research. Also, since every face is different, learning strategies could be employed to adopt the constraints to the individual anatomy.

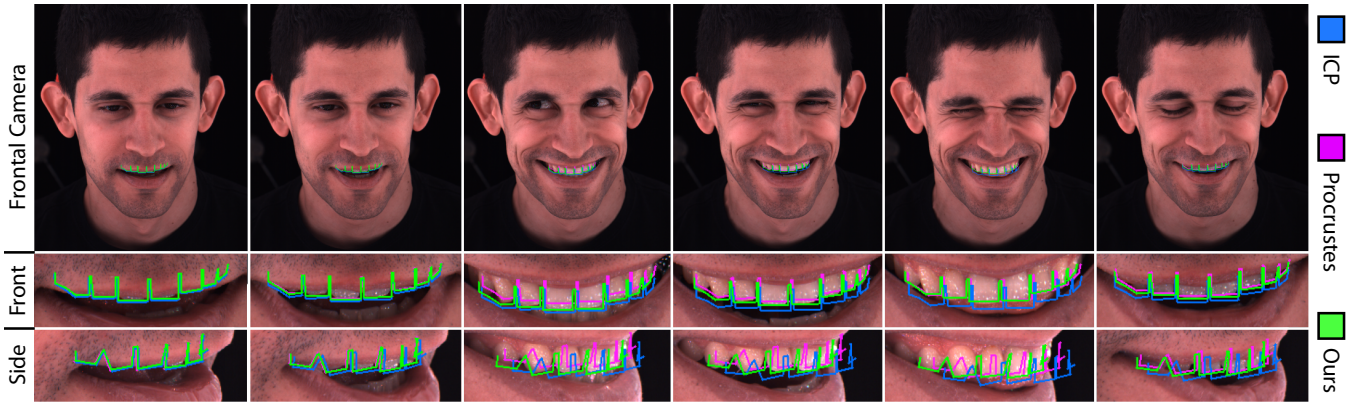

Figure 10: Performance - *The proposed method (green) may also be used to efficiently stabilize performance captures, since every frame is processed independently. Even without explicit temporal continuity, results are temporally smooth and accurate, which is not the case for ICP (blue) and Procrustes (magenta). Again, zoom regions are shown from both front and side views.*

Acknowledgements

We would like to thank our actors Lena, Seth, Michael and Cheve for letting us capture them. We further would like to thank PD Dr. med. Dr. med. dent. Heinz-Theo Lübbers for providing us with CT scans and anatomical background information and Michael Koperwas for sharing with us the industry viewpoint and manually stabilizing some shapes for reference.

References

ALEXANDER, O., ROGERS, M., LAMBETH, W., CHIANG, J.-Y., MA, W.-C., WANG, C.-C., AND DEBEVEC, P. 2010. The digital emily project: Achieving a photoreal digital actor. *IEEE Computer Graphics and Applications 30*, 4, 20–31.

ALI-HAMADI, D., LIU, T., GILLES, B., KAVAN, L., FAURE, F., PALOMBI, O., AND CANI, M. 2013. Anatomy transfer. *ACM Trans. Graph. 32*, 6 (Nov.), 188:1–188:8.

AMBERG, B., AND VETTER, T. 2011. Optimal landmark detection using shape models and branch and bound. In *Int. Conference on Computer Vision (ICCV)*.

ARUN, K. S., HUANG, T. S., AND BLOSTEIN, S. D. 1987. Least-squares fitting of two 3-d point sets. *IEEE Trans. PAMI 9*, 5, 698–700.

BEELER, T., BICKEL, B., SUMNER, R., BEARDSLEY, P., AND GROSS, M. 2010. High-quality single-shot capture of facial geometry. *ACM Trans. Graphics (Proc. SIGGRAPH) 29*, 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 Trans. Graphics (Proc. SIGGRAPH) 30*, 75:1–75:10.

BESL, P. J., AND MCKAY, N. D. 1992. A method for registration of 3-d shapes. *IEEE Trans. PAMI 14*, 2, 239–256.

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

BOTSCH, M., AND SORKINE, O. 2008. On linear variational surface deformation methods. *IEEE TVCG 14*, 1, 213–230.

BOUAZIZ, S., WANG, Y., AND PAULY, M. 2013. Online modeling for realtime facial animation. *ACM Trans. Graphics (Proc. SIGGRAPH) 32*, 4, 40:1–40:10.

BRADLEY, D., HEIDRICH, W., POPA, T., AND SHEFFER, A. 2010. High resolution passive facial performance capture. *ACM Trans. Graphics (Proc. SIGGRAPH) 29*, 41:1–41:10.

CAO, C., WENG, Y., ZHOU, S., TONG, Y., AND ZHOU, K. 2013. Facewarehouse: A 3d facial expression database for visual computing. *IEEE TVCG*.

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

DALE, K., SUNKAVALLI, K., JOHNSON, M. K., VLASIC, D., MATUSIK, W., AND PFISTER, H. 2011. Video face replacement. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 30*, 6, 130:1–130:10.

FYFFE, G., HAWKINS, T., WATTS, C., MA, W.-C., AND DEBEVEC, P. 2011. Comprehensive facial performance capture. In *Eurographics 2011*.

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

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

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

LI, H., ADAMS, B., GUIBAS, L. J., AND PAULY, M. 2009. Robust single-view geometry and motion reconstruction. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 28*, 5, 175:1–175:10.

LI, H., WEISE, T., AND PAULY, M. 2010. Example-based facial rigging. *ACM Trans. Graphics (Proc. SIGGRAPH) 29*, 4, 32:1–32:6.

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

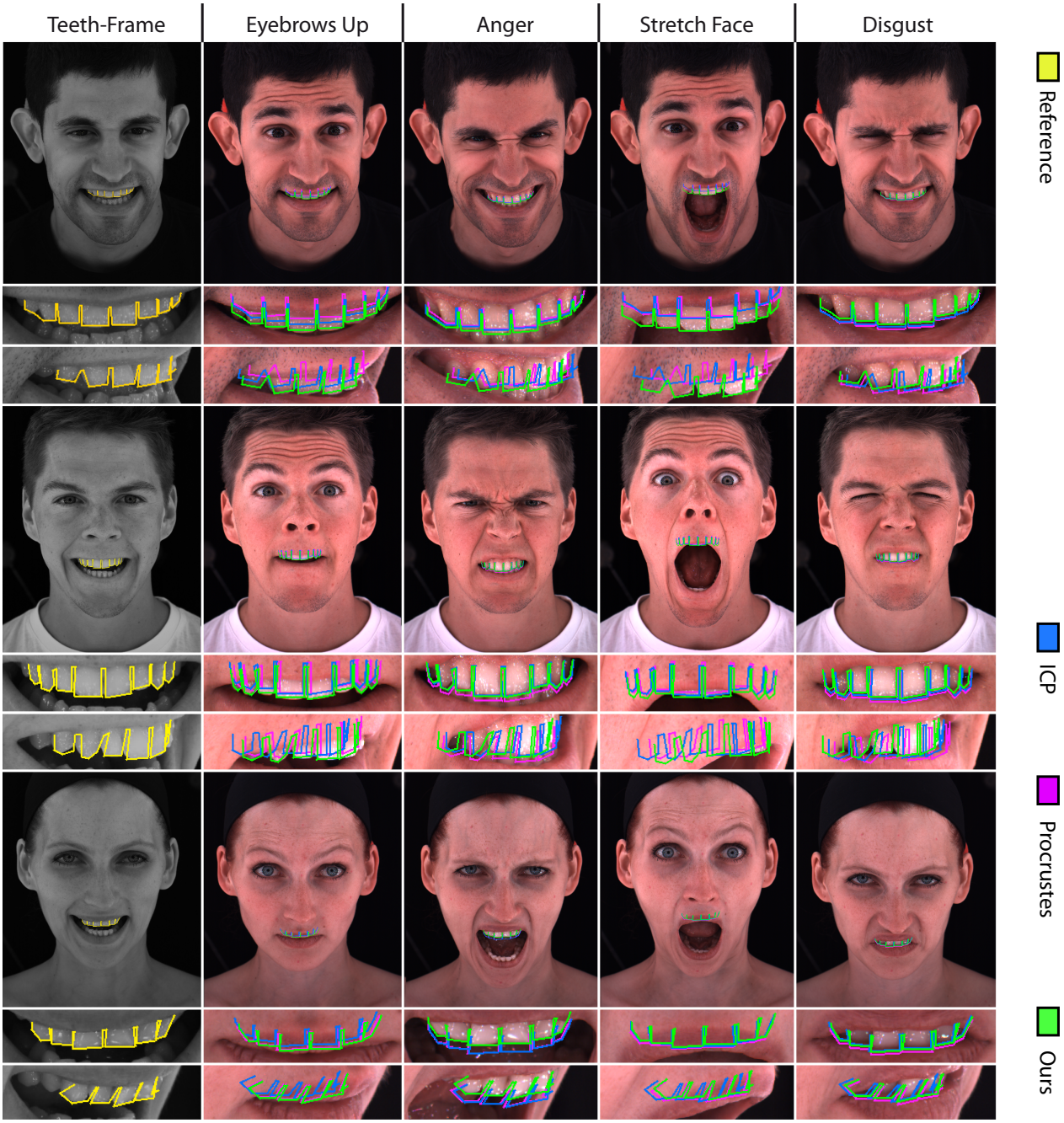

Figure 12: Robustness - *We demonstrate the performance on different actors using the same set of parameters. The leftmost column shows the outline of the teeth (yellow) in the frame they were reconstructed in (See Figure 7). The other columns show the stabilization results for ICP (blue), Procrustes (magenta), and Ours (green) for four different expressions. We show zoom regions from both front and side views.*

MA, W.-C., HAWKINS, T., PEERS, P., CHABERT, C.-F., WEISS, M., AND DEBEVEC, P. 2007. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Eurographics Symposium on Rendering*, 183–194.

SUMNER, R. W., AND POPOVIĆ, J. 2004. Deformation transfer for triangle meshes. *ACM Trans. Graphics (Proc. SIGGRAPH) 23*, 3, 399–405.

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

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

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