

Accurate Markerless Jaw Tracking for Facial Performance Capture

GASPARD ZOSS, DisneyResearch|Studios, ETH Zurich
THABO BEELER, DisneyResearch|Studios
MARKUS GROSS, DisneyResearch|Studios, ETH Zurich
DEREK BRADLEY, DisneyResearch|Studios



Fig. 1. We present a method to accurately track the jaw during facial performance capture, without the need for attaching markers or tracking teeth.

We present the first method to accurately track the invisible jaw based solely on the visible skin surface, without the need for any markers or augmentation of the actor. As such, the method can readily be integrated with off-the-shelf facial performance capture systems. The core idea is to learn a non-linear mapping from the skin deformation to the underlying jaw motion on a dataset where ground-truth jaw poses have been acquired, and then to retarget the mapping to new subjects. Solving for the jaw pose plays a central role in visual effects pipelines, since accurate jaw motion is required when retargeting to fantasy characters and for physical simulation. Currently, this task is performed mostly manually to achieve the desired level of accuracy, and the presented method has the potential to fully automate this labour intense and error prone process.

Authors' addresses: Gaspard Zoss, DisneyResearch|Studios, ETH Zurich, gaspard.zoss@disneyresearch.com; Thabo Beeler, DisneyResearch|Studios, thabo.beeler@disneyresearch.com; Markus Gross, DisneyResearch|Studios, ETH Zurich, gross@disneyresearch.com; Derek Bradley, DisneyResearch|Studios, derek.bradley@disneyresearch.com.

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. Request permissions from permissions@acm.org.

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
0730-0301/2019/7-ART50 \$15.00
<https://doi.org/10.1145/3306346.3323044>

CCS Concepts: • **Computing methodologies** → *Motion capture; Motion processing.*

Additional Key Words and Phrases: Jaw Tracking, Markerless, Data Driven Animation, Facial Performance Capture, Acquisition

ACM Reference Format:

Gaspard Zoss, Thabo Beeler, Markus Gross, and Derek Bradley. 2019. Accurate Markerless Jaw Tracking for Facial Performance Capture. *ACM Trans. Graph.* 38, 4, Article 50 (July 2019), 8 pages. <https://doi.org/10.1145/3306346.3323044>

1 INTRODUCTION

Generating realistic facial animation has always been a central ingredient in the creation of digital characters for computer games, visual effects for film and other virtual experiences. A very important component of the face is the underlying jaw, as the jaw's motion is often used to control the deformation of the face surface using methods like skinning. For this reason, most facial animation rigs contain an explicit jaw rig. While oftentimes the jaw is rigged with a few simple rotation and transformation controls, we have also seen recent advances in the design and control mechanisms of jaw rigs built from real-world capture data [Yang et al. 2018; Zoss et al. 2018].

Advances in jaw rigging are largely designed to help animators create facial animation through key-frame interpolation. However, a significant amount of character facial animation is created through performance capture, in a process where a real actor is recorded and the facial surface is reconstructed and tracked to provide a digital performance that exactly matches the real one. The digital performance is then typically provided to the animator as a starting point for further manipulation, which often involves a retargeting step to a fantasy creature for final display. During this process, it is essential to know the jaw motion that corresponds to the captured facial performance of the actor. This yields a major challenge - how can we accurately track the jaw of an actor during performance capture?

Accurate jaw tracking has been studied in several fields such as dentistry and speech analysis, however typically at the cost of complicated setups and invasive instruments like electromagnetic sensors or optical fiducials attached to the teeth. Recently in the graphics community, researchers have become more interested in accurate jaw tracking, for example the data-driven rigging methods mentioned earlier required accurate jaw motion for ground truth evaluation [Yang et al. 2018; Zoss et al. 2018]. However, the tracking methods they used are still cumbersome, requiring fiducial markers or accurate dental scans of the teeth, and are thus not suitable for the general widespread application of facial performance capture involving dialog.

In this work, we present the first method for accurate jaw tracking for facial performances given only the tracked facial surface as input. We achieve this by learning a non-linear mapping from the skin surface to the underlying jaw pose for several subjects for which we obtain ground-truth jaw motion using the marker based approach presented by Zoss et al. [2018] in a one-time database creation step. We then demonstrate how this mapping can be transferred onto new subjects, for which marker based jaw tracking is not available, rendering the method applicable to traditional facial performance acquisition.

In production, jaw tracking is a very manual and labour intensive step, where artists leverage the teeth wherever visible and if not try to guesstimate the jaw pose from the perceived surface. Aside from the required effort, which is very substantial, manual tracking offers a lot of potential for human error, which will manifest itself in the final character. The proposed approach allows to fully automate jaw tracking at high accuracy and removes human error, without imposing any additional requirements onto the actor or the setup during acquisition.

2 RELATED WORK

Our work is most related to other methods for jaw tracking, and has applications in the field of facial performance capture.

2.1 Jaw Tracking

Jaw tracking has been well-studied in fields outside of computer graphics, for example dentistry and speech analysis. Several rather invasive methods have been proposed, including the application of electromagnetic sensors inside the mouth [Piancino et al. 2013; Santos et al. 2006], ultrasonic emitters [Prinz 1997], fiducial markers

attached to the teeth [Eriksson et al. 2000; Ferrario et al. 2005; Kinuta et al. 2005; Mostashiri et al. 2018; Pinheiro et al. 2008], and LED or optical reflective markers attached to the skin [Buschang et al. 2000, 2007; Wiesinger et al. 2014; Wilson and Weismer 2012; Wintergerst et al. 2004; Zafar et al. 2000] (which only provide approximate jaw motion as we will show later). An overview of different approaches can be found in Bando et al. [2009]. While these methods may provide highly accurate jaw motion, the acquisition setups are too cumbersome or uncomfortable for widespread application in computer graphics.

In the field of forensics, a related topic is craniofacial superimposition, where a given skull is superimposed onto face images for the purpose of identification. While the jaw is often overlooked in this procedure, Bermejo et al. [2017] propose a genetic algorithm for estimating the 1-DOF jaw aperture that matches the image, allowing the jaw to contribute to the identity. Such an approach could be considered a rudimentary jaw tracking method.

Recently, the graphics community has been interested in improving jaw animation through accurate jaw tracking. Zoss et al. [2018] employed the invasive approach of attaching fiducial markers to the teeth and then used the resulting motion as ground truth for constructing a novel jaw rigging method. In this work we will use similar ground truth data in training our jaw motion predictor. Yang et al. [2018] avoid the use of markers but require an accurate 3D model of the actors' teeth using a dental scanner, and can only track the jaw for performances where the mouth and lips are fully open to reveal the teeth. These approaches are designed for short term use in order to generate ground truth data, and, like the dental tracking methods above, are unsuited for widespread use in graphics and entertainment.

Relatively little work has attempted to perform automatic and accurate jaw tracking from just images or meshes of a face. In the monocular face capture method of Wu et al. [2016], a facial performance is reconstructed from video, including the motion of the underlying bones. This approach involves building a person-specific local anatomical model where the jaw must be prescribed for a set of approximately a dozen input poses. This approach would not extend to cases where such a specialized model is not available. Beeler and Bradley [2014] propose a method to rigidly align an actor-specific skull to facial scans using skin tissue thickness analysis, for the purpose of rigid stabilization. While also dealing with aligning bones under the face, their method is limited to skull fitting and is not applicable to the jaw, due to their specialized forehead skin sliding and nose constraints. Finally, Tanaka et al. [Tanaka et al. 2016] and Zafar et al. [Zafar et al. 2000] discuss the potential for a trivial correlation between skin motion at the chin and the corresponding mandible motion, suggesting that markerless jaw tracking based on the 3D skin surface might be feasible. In this work we also demonstrate the feasibility of such a correlation (see *baseline* experiments in Section 7), and further propose a novel jaw prediction algorithm with learned regression that provides more accurate jaw tracking given only the skin surface as input.

2.2 Facial Capture

We believe our method for accurate jaw tracking is most applicable for the application of facial performance capture. While the facial capture method of Wu et al. [2016] described above is one of the few that can already solve for the jaw pose, most facial tracking methods focus only on the surface of the skin and ignore the underlying bones, leaving the jaw tracking problem as a manual task for artists. An extensive overview of facial capture methods is beyond the scope of this paper, but since our method relies on the facial motion as input and our results would highly complement facial capture methods, a short overview of recent methods is warranted. Facial performance capture approaches come in several flavors, including multi-view [Beeler et al. 2011; Bradley et al. 2010; Fyffe et al. 2011, 2017] or binocular stereo [Valgaerts et al. 2012], model fitting to monocular video [Garrido et al. 2013; Shi et al. 2014; Suwajanakorn et al. 2014; Wu et al. 2016], deep learning [Laine et al. 2017; Tewari et al. 2017], real-time capture and puppeteering [Bouaziz et al. 2013; Cao et al. 2015, 2014; Hsieh et al. 2015; Li et al. 2013; Weise et al. 2011], or expression transfer for facial re-enactment [Thies et al. 2015, 2016, 2018]. In our work, we use the multi-view method of Beeler et al. [2011] to capture facial performance sequences, since it provides dense and accurate facial motion. However, our technique could be easily applied to any facial capture method that provides consistent surface topology, including the real-time approaches as our jaw tracking algorithm solves at real-time rates. This flexibility makes our method immediately applicable in several research and industry scenarios.

3 METHOD OVERVIEW

We propose to learn a non-linear mapping from the skin surface deformation to the motion of the underlying, invisible jaw. The mapping, proposed in Section 5, is learned from a corpus of data for which both skin surface and ground truth jaw poses are available (Section 4). Once learned, we show how to retarget the mapping to new actors in Section 6, for which ground truth jaw poses are not available, using a small set of provided calibration poses. In Section 7 we demonstrate our novel jaw tracking method on several sequences of different actors, and evaluate the accuracy of jaw prediction for both new expressions of a training actor as well as performances of new actors that were not part of the training set. An illustration of our method is shown in Fig. 2.

4 DATA ACQUISITION

To learn a mapping from the skin surface to the jaw pose we need to capture both at the same time. We follow the approach suggested by Zoss et al. [2018] and glue a set of markers onto the actors teeth (Fig. 3.a), which will allow us to accurately estimate the 6-DoF pose of both the skull and the jaw. We refer the reader to their paper for details. To capture the skin surface we employ the system of Beeler et al. [2011]. Since the markers are obstructing the skin and cause artifacts in the reconstruction (Fig. 3.b), we automatically mask them from the input images, which leads to reasonable reconstruction of the geometry despite their presence (Fig. 3.c).

We capture a total of five subjects undergoing a large range of jaw motion. For reference, we will refer to them as S_1 through S_5 in the

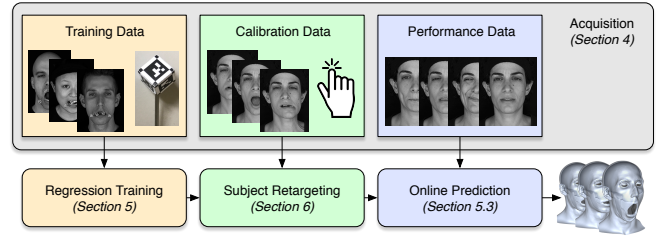


Fig. 2. **Overview.** *Training (yellow):* We train a non-linear mapping (Section 5) that predicts jaw pose from skin surface deformation. The training subjects are captured using fiducial markers to acquire ground truth jaw poses in addition to their facial geometry (Section 4). *Retargeting (green):* For a new subject without markers we acquire a sparse set of calibration poses (~ 5), where we solve for the jaw pose by manually labelling a few points on the teeth in the input imagery. Using these calibration poses we retarget the mapping to the new subject as described in Section 6. *Prediction (blue):* Once retargeted, the proposed method predicts the jaw pose on a per-frame basis in real-time using the fitting described in Section 5.3.

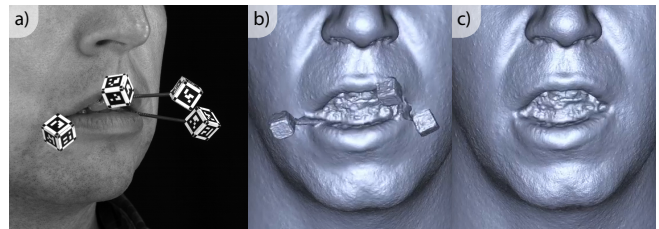


Fig. 3. **Data Acquisition.** To accurately estimate the jaw pose we follow the approach of Zoss et al. [2018] and glue markers to the actor's teeth (a). To avoid degradation of the reconstruction quality due to these markers (b) we automatically mask them in the images, which leads to reasonable reconstruction of the geometry despite their presence (c).

text. The reconstructed performances are stabilized to remove rigid head motion using the upper teeth markers, and for markerless acquisition using the method of Beeler and Bradley [2014]. The result of the data acquisition is per-frame tracked 3D geometry of the face, with corresponding per-frame jaw poses, all registered to a canonical head pose with consistent vertex topology across subjects.

5 JAW POSE PREDICTION

The core of our idea is to exploit the relation between skin and jaw motion. Unfortunately, there is no unique relation between a point on the skin and the pose of the underlying jaw since, on the one hand, skin slides over the bone when actuating the jaw, and on the other hand skin can be deformed without actually moving the jaw. However, our observation is that when considering several skin points at the same time to predict several jaw points, it is possible to disambiguate bespoken ambiguities. We also rely on a thorough captured dataset that spans a significant range of jaw and skin motion in order to robustly train our predictor. More formally, we seek to find a mapping $\Phi(\mathcal{F}) \rightarrow \mathcal{B}$, which predicts jaw features

\mathcal{B} from observed skin features \mathcal{F} , and then the final jaw pose will be computed from the estimated jaw features.

We start by defining the features (Section 5.1), and then introduce our mapping as a regression (Section 5.2). Finally, from the predicted jaw features we fit the jaw pose (Section 5.3).

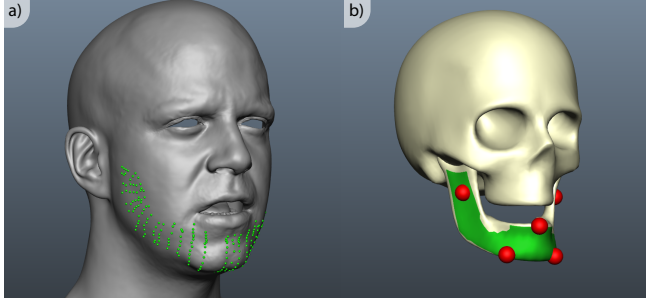


Fig. 4. **Features.** To compute the skin feature \mathcal{F} a set of 242 vertices is selected on the skin surface (a, green spheres) by sampling vertices on the jaw within a manually painted mask (b, green mask) and computing the closest point on the skin along the jaw normals. As bone feature \mathcal{B} we select five points distributed over the jaw (b, red spheres).

5.1 Features

Skin features \mathcal{F} . In order to render our feature space invariant to translation and global rotation of the head, we define a skin feature $\mathcal{F}_j \in \mathbb{R}^{|\mathcal{V}| \times 3}$ to be the displacements $\{\mathbf{d}_v\}_j$ of a set of sample vertices $v \in \mathcal{V}$ from reference neutral to expression j , relative to the coordinate frame of the skull. The feature vertices \mathcal{V} on the face are computed by randomly sampling jaw vertices and finding the closest point on the face along the normal of the jaw. We sample only jaw points within a manually defined mask (Fig. 4.b, green), which excludes vertices on the inside of the jaw, leading to a total of 242 skin feature vertices over the chin area of the face (Fig. 4.a, green spheres), hence $\mathcal{F}_j \in \mathbb{R}^{242 \times 3}$.

Bone features \mathcal{B} . Analog to the skin features we seek target bone features $\mathcal{B}_j \in \mathbb{R}^{|\mathcal{W}| \times 3}$ that are also invariant to translation and global rotation, and hence again express them as displacements $\{\mathbf{b}_w\}_j$ of a set of sample vertices $w \in \mathcal{W}$ from the reference bone pose to the pose in expression j , relative to the coordinate frame of the skull. Since these features will be used to define the jaw pose as a rigid transformation, a minimum of three non-collinear bone samples are required, and more samples can increase robustness to noise. We found empirically that five samples, selected manually as shown in Fig. 4.b as red spheres, produce compelling results while keeping the feature space compact, yielding $\mathcal{B}_j \in \mathbb{R}^{5 \times 3}$.

5.2 Regression

For every bone sample w we train a support vector machine (SVR) regressor $\phi_w(\{\mathbf{d}_v\}) \rightarrow \mathbf{b}_w$, which predicts the displacement \mathbf{b}_w for the bone sample w from the skin displacements $\{\mathbf{d}_v\}$ of all skin sample vertices $v \in \mathcal{V}$, trained over all captured expressions j in the training data. We employ DLib's SVR implementation [King 2009], which only supports predicting a single scalar, and hence train three

different regressors per bone target, one for each dimension. The final mapping $\Phi(\mathcal{F}) \rightarrow \mathcal{B}$ is given by aggregating the regressors ϕ_w for the individual jaw bone samples w .

5.3 Jaw Pose Fitting

For any new expression k of the face, we can now compute the skin feature \mathcal{F}_k and evaluate the regression to obtain a corresponding bone feature \mathcal{B}_k . From these predicted displacements $\{\mathbf{b}_w\}_k$ for the jaw bone samples w we can compute absolute locations $\{\mathbf{x}_w\}_k$ and solve for the rigid transformation \mathbf{T}_k that optimally fits the jaw bone to these locations in a least-squares sense, by minimizing

$$E_{fit}(\mathbf{T}) = \sum_{w \in \mathcal{W}} \|\mathbf{T}\hat{\mathbf{x}}_w - \mathbf{x}_w\|_2^2, \quad (1)$$

where $\hat{\mathbf{x}}_w$ denotes the location of bone feature w in the reference pose, and note that we removed the pose subscript k for simplicity. We minimize E_{fit} using the ceres solver [Agarwal et al. 2016].

6 MAPPING RETARGETING

The approach presented in Section 5 succeeds at predicting novel jaw poses for a subject where ground-truth jaw poses have been acquired for a large number of expressions. Typically, however, such ground truth poses are not available and hence we present a method to retarget the mappings learned for our captured training subjects to new target subjects.

The underlying assumption for retargeting is that the relation between skin and jaw motion generalizes across people, which is a reasonable assumption to make since the underlying anatomy is similar across humans. Still, the skin of different people will exhibit slightly different motion, which is why we propose a retargeting method which refines the mapping learned on source actors with respect to a new target subject.

6.1 New Target Subject

To accomplish the retargeting, we capture a small set of calibration poses \mathcal{P} for which lower teeth are visible. Typically this set includes only 5 poses: mouth wide open, jaw to the left, to the right, forward and backward. We then use the method of Beeler and Bradley [2014] to align the skull to these poses and recover the rigid transformation of the jaw with respect to the skull for each frame by triangulating teeth outlines on images. For each calibration pose $p \in \mathcal{P}$, we extract the calibration skin features \mathcal{F}_p and corresponding calibration bone features \mathcal{B}_p . Our retargeting method assumes consistent vertex topology (as described in Section 4), thus the skin features \mathcal{F}_p are consistent across actors.

The idea behind our retargeting method is to transform the motion space of each feature vertex v on the target subject to best align with the mapping Φ that was computed from a source subject. To this end, we formulate an energy term for solving the optimal rigid transformations \mathbf{R}_v for every skin feature v , defined over all poses p of the calibration set

$$E_{ret}(\{\mathbf{R}_v\}) = \sum_{p \in \mathcal{P}} \|\Phi(\{\mathbf{R}_v\} \otimes \mathcal{F}_p) - \mathcal{B}_p\|_2^2. \quad (2)$$

We define $\{\mathbf{R}_v\} \otimes \mathcal{F}_p$ as the operator applying each transformation \mathbf{R}_v to the corresponding displacement $\mathbf{d}_v \in \mathcal{F}_p$.

Additionally we add regularization terms for both the translational (\mathbf{t}_v) and rotational (\mathbf{q}_v) components of each transformation $\mathbf{R}_v = \mathbf{T}(\mathbf{q}_v, \mathbf{t}_v)$. The optimal rigid transformations are then computed by solving

$$\min_{\{\mathbf{q}_v, \mathbf{t}_v\}} E_{ret}(\{\mathbf{T}(\mathbf{q}_v, \mathbf{t}_v)\}) + \lambda_1 \sum_{\mathbf{q}_v} \|\mathbf{q}_0 \cdot \mathbf{q}_v^{-1}\|_2^2 + \lambda_2 \sum_{\mathbf{t}_v} \|\mathbf{t}_v\|_2^2 \quad (3)$$

with \mathbf{q}_0 being the identity quaternion. We minimize the above equation using *ceres* [Agarwal et al. 2016], with $\lambda_1 = 1e-5$, and $\lambda_2 = 2.5e-5$. This optimally aligns the feature spaces of source and target subjects, allowing to use the source mapping on unseen target input shapes after applying the transformations $\{\mathbf{R}_v\}$ to the skin features. We show the effect of applying this calibration procedure to one subject in Fig. 5, compared to no calibration where the source mapping is used directly without transformations.

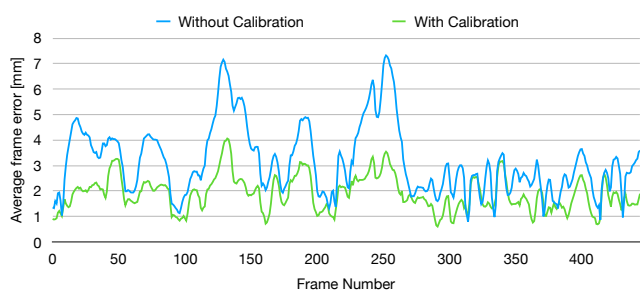


Fig. 5. **Calibration.** The proposed calibration procedure rigidly aligns the feature spaces of source and target subjects from a handful of poses, leading to substantially lower prediction errors when retargeting (green) compared to uncalibrated retargeting (blue), in particular for wider openings of the jaw.

6.2 Multi-Subject Mapping

The proposed retargeting method optimally aligns the feature spaces of source and target individuals, and produces good results if the overall facial anatomy is similar (Fig. 6 - red curve). However, if the facial morphology of the two subject differs substantially, e.g. one has a much higher BMI, then retargeting produces inferior results (Fig. 6 - purple curve). Hence, we propose to not retarget the mapping from a single subject but to combine the features of several people and learn a multi-subject mapping. We start by selecting one of the source subjects, further referred to as the primary, and follow the approach presented in Section 5 to learn a mapping specific to that subject. This mapping is then used to align the other subjects with the primary following the calibration method described in 6.1, but instead of solving for a rigid transformation per feature, we solve for a single rigid transformation for all features globally. This ensures that the skulls of the subjects are optimally aligned without

destroying their relative differences, which would be the case if every feature was aligned separately. Once all the subjects are aligned with the primary, we solve for a single, combined, multi-subject mapping, again following the method described in Section 5 but this time using training data from all (aligned) subjects. This new mapping, trained on several actors, can then be used when retargeting to unseen subjects using again the method described in Section 6.1. Figure 6 (blue curve) shows the effect of combining several source subjects for retargeting and demonstrates that the multi-subject mapping outperforms every single-subject mapping overall, indicating that the regression succeeds at interpolating between subjects. The final result is a powerful jaw predictor trained from all available capture subjects that can be retargeted to perform jaw tracking for any unseen target subject.

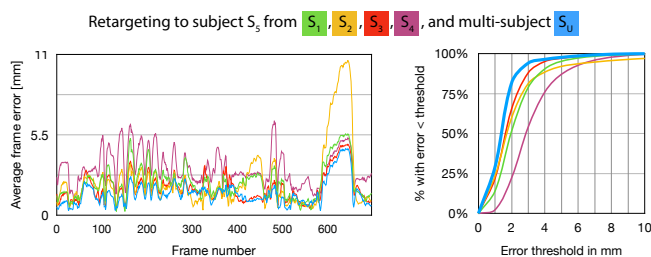


Fig. 6. **Single- vs. Multi-Subject.** Depending on the similarity of source and target subjects, the quality of the retargeting might vary (S_1 - S_4). Combining all source subjects into a single mapping yields a powerful jaw predictor (S_0) that consistently outperforms all the individual mappings.

7 RESULTS

In this section we evaluate the proposed jaw tracking method, including validating the simple case of training and testing on the same subject (Section 7.1), validating retargeting a trained model from a corpus of subjects to an unseen subject (Section 7.2), and demonstrating additional examples of unseen subjects qualitatively (Section 7.3).

Baseline Pose Prediction. Throughout the validation we will compare to a baseline result, often used by artists in the entertainment industry to obtain an initial estimate of the jaw pose automatically, using the skin surface deformation as a naïve proxy for the jaw transformation. Specifically, as baseline we compute the rigid transformation from the neutral pose to the input expression using Procrustes analysis [Gower 1975] on the skin feature vertices \mathcal{V} , and apply this transformation to the jaw bone. This baseline algorithm will typically predict the general direction of jaw motion correctly, but will fail to achieve high accuracy since it does not capture the effects of skin sliding over the bone, as we will see in the following evaluations.

7.1 Single Subject Validation

We start by validating the application of training on one subject, and testing on new expressions of the same subject. Fig. 7 illustrates the validation, where 3539 frames of various jaw motions from subject

S_2 were captured using fiducial markers to obtain ground truth jaw poses. Our predictor was trained on the first 2000 frames, and then tested on the remaining 1539 frames. The evaluation clearly shows that our method (blue curve) greatly improves over the baseline approach (green curve), and maintains a consistently low error despite the variety of jaw poses.

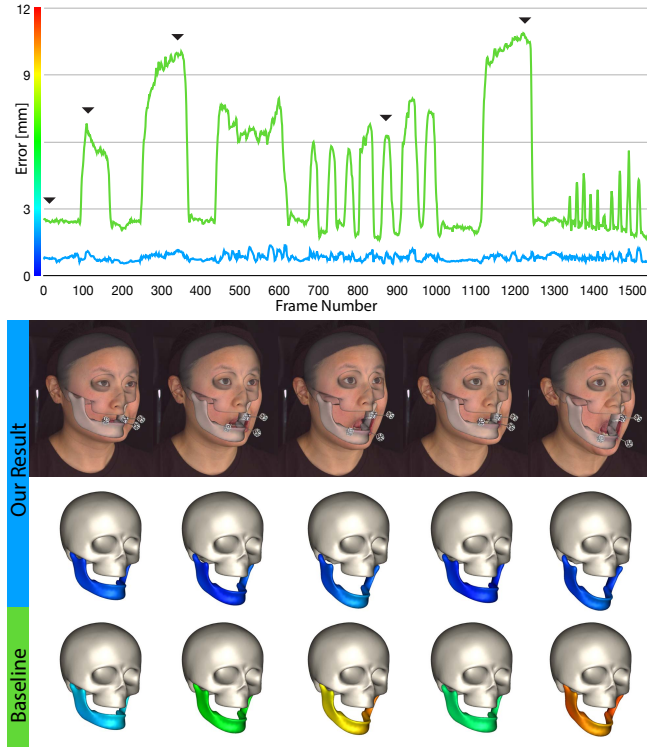







Fig. 7. **Single Actor Validation.** We evaluate our jaw tracking method compared to the naïve baseline approach. A sequence of 3539 frames of subject S_2 were captured, ground truth jaw poses were computed for the first 2000 frames and were used for training, and here we show the 1539 test frames. Our error is consistently below 1mm, where the baseline method can exceed 10mm.

7.2 Retargeting Validation

While the single subject validation is interesting, more practical is our method to retarget a trained predictor from several subjects to a new unseen subject. In order to validate such a retargeting, we take the four main subjects for which we have significant ground truth data (S_1 through S_4), and perform a *leave-one-out* cross validation. For example, we train a predictor on S_1 , S_2 , and S_3 , and then test on subject S_4 , comparing again to the baseline approach. The results of all combinations are shown in Table 1 (first four columns). We then created a single multi-subject predictor using all frames of S_1 through S_4 , and tested it on an additional subject S_5 (column five). As can be seen, in all cases our retargeting method quantitatively outperforms the baseline by a wide margin.

We demonstrate the multi-subject retargeting visually in Fig. 8, compared to the baseline as well as the single subject training

Table 1. This table show the average prediction error in mm of both the proposed and the baseline method. For the proposed method, we trained on three of the subjects to predict the fourth. As can be seen, our method outperforms the baseline for every subject. Subject 1 has the highest prediction error, which we discuss in Section 8. Subject S_5 was never used for training and is predicted using S_{1-4} .

					
Baseline	5.20	4.97	4.79	9.97	5.89
Ours	2.38	1.09	1.32	1.50	1.52

learned on ground truth of the same actor. For this figure, the target subject is S_4 , corresponding to the fourth column of Table 1. The single subject predictor was trained on 800 frames of jaw motion, and then both the single and retargeted predictors were tested on a sequence of 159 frames. Important to notice is that the retargeting results (yellow curve) are very similar in accuracy and consistency with respect to the single subject training (blue curve), indicating that the retargeting method works very well, and our method is highly applicable to subjects for which ground truth data is not available. Furthermore, both our single subject and retargeting results greatly outperform the baseline (green curve).

We show another visualization of the retargeting validation in Fig. 9, this time for subject S_3 corresponding to column three of Table 1. Again we see that our method (blue curve) is consistently quite accurate and outperforms the baseline approach (green curve).

7.3 Additional Results

We end this section with a qualitative demonstration of several frames showing different jaw fit results for different subjects, reconstructed under normal facial performance capture conditions (i.e. no fiducial markers glued to the teeth), as shown in Fig. 1. These results are representative of the jaw fitting performance one would expect in a motion capture scenario in the field of entertainment. We refer the reader to the accompanying video for additional examples, including larger variance in performance.

8 DISCUSSION

We propose a new method for markerless jaw tracking in the context of facial performance capture, which uses only the deforming skin surface as input, yet achieves highly accurate jaw pose results. Furthermore, thanks to our convenient retargeting scheme, our method can be applied to new unseen actors for which ground truth training data is not available.

Limitations and Future Work. Like most learning-based approaches, the quality of our results depends on the quality and moreover the completeness of the training data, as seen in Fig. 10. We do not expect our jaw prediction to extrapolate well for expressions that are far outside the convex hull of the training set. In this case, one could detect these input poses which are outside the training data, project them onto the convex hull and evaluate the regression, then compensate for this projection in the resulting bone prediction. We

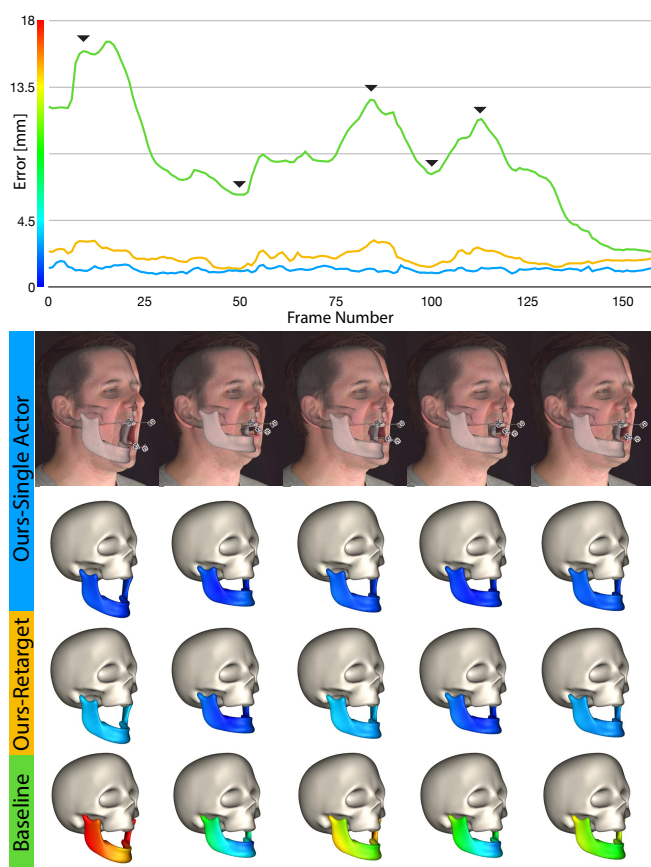


Fig. 8. **Single Actor vs. Retarget Validation.** We validate both single-actor training and multi-subject retargeting on subject S_4 , also comparing to the naïve baseline approach. A sequence of 959 frames were captured and the single actor predictor was trained on the first 800 frames. Here we show the remaining 159 test frames, together with a retargeting result using $S_{\{1,2,3\}}$ as training. As illustrated, retargeting from other subjects is almost as accurate as training on the specific subject, and both methods are significant improvements over the baseline.

leave this experimentation for future work. In practice, we have attempted to alleviate this limitation by building a large enough training set so that the need to extrapolate is minimal.

In summary, we believe our work has the potential for great impact on the entertainment industry, as current jaw tracking approaches in the fields of visual effects and computer games are largely manual and time-consuming. Our automatic approach will save significant artist time.

ACKNOWLEDGMENTS

We wish to thank our capture subjects Ariane Leo, Radek Daněček, Yeara Kozlov, Christian Schumacher, and Andreea Rădoescu.

REFERENCES

Sameer Agarwal, Keir Mierle, and Others. 2016. Ceres Solver. <http://ceres-solver.org>.

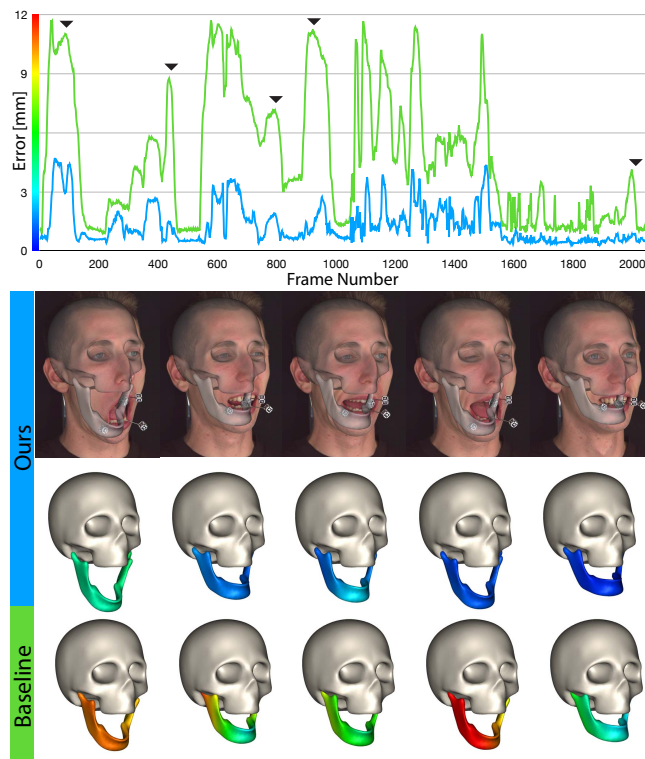


Fig. 9. **Retarget Validation.** We further validate our retargeted jaw tracking on subject S_3 , also comparing to the naïve baseline approach. Our approach provides consistently better accuracy than the baseline method.

- Eiichi Bando, Keisuke Nishigawa, Masanori Nakano, Hisahiro Takeuchi, Shuji Shigemoto, Kazuo Okura, Toyoko Satsuma, and Takeshi Yamamoto. 2009. Current status of researches on jaw movement and occlusion for clinical application. *Japanese Dental Science Review* 45, 2 (2009), 83–97. <https://doi.org/10.1016/j.jdsr.2009.04.001>
- Thabo Beeler and Derek Bradley. 2014. Rigid stabilization of facial expressions. *ACM Transactions on Graphics* 33, 4 (2014), 1–9. <https://doi.org/10.1145/2601097.2601182>
- Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W. Sumner, and Markus Gross. 2011. High-quality passive facial performance capture using anchor frames. *ACM Transactions on Graphics* (2011), 1. <https://doi.org/10.1145/1964921.1964970> arXiv:arXiv:1011.1669v3
- Enrique Bermejo, Carmen Campomanes-Álvarez, Andrea Valsecchi, Oscar Ibáñez, Sergio Damas, and Oscar Cerdón. 2017. Genetic algorithms for skull-face overlay including mandible articulation. *Information Sciences* 420 (2017), 200–217. <https://doi.org/10.1016/j.ins.2017.08.029>
- Sofien Bouaziz, Yangang Wang, and Mark Pauly. 2013. Online modeling for realtime facial animation. *ACM Transactions on Graphics* 32, 4 (2013), 40:1–40:10. <https://doi.org/10.1145/2461912.2461976>
- Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. 2010. High Resolution Passive Facial Performance Capture. *ACM Transactions on Graphics* 29, 4 (2010), 41:1–41:10. <https://doi.org/10.1145/1778765.1778778>
- P. H. Buschang, H. Hayasaki, and G. S. Throckmorton. 2000. Quantification of human chewing-cycle kinematics. *Archives of Oral Biology* 45, 6 (2000), 461–474. [https://doi.org/10.1016/S0003-9969\(00\)00015-7](https://doi.org/10.1016/S0003-9969(00)00015-7)
- Peter H. Buschang, Gaylord S. Throckmorton, Dawn Austin, and Ana M. Wintergerst. 2007. Chewing cycle kinematics of subjects with deepbite malocclusion. *American Journal of Orthodontics and Dentofacial Orthopedics* 131, 5 (2007), 627–634. <https://doi.org/10.1016/j.ajodo.2005.06.037>
- Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. 2015. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics* 34, 4 (2015), 46:1–46:9. <https://doi.org/10.1145/2766943>
- Chen Cao, Qiming Hou, and Kun Zhou. 2014. Displaced Dynamic Expression Regression for Real-time Facial Tracking and Animation. *ACM Trans. Graph.* 33, 4 (2014), 43:1–43:10.

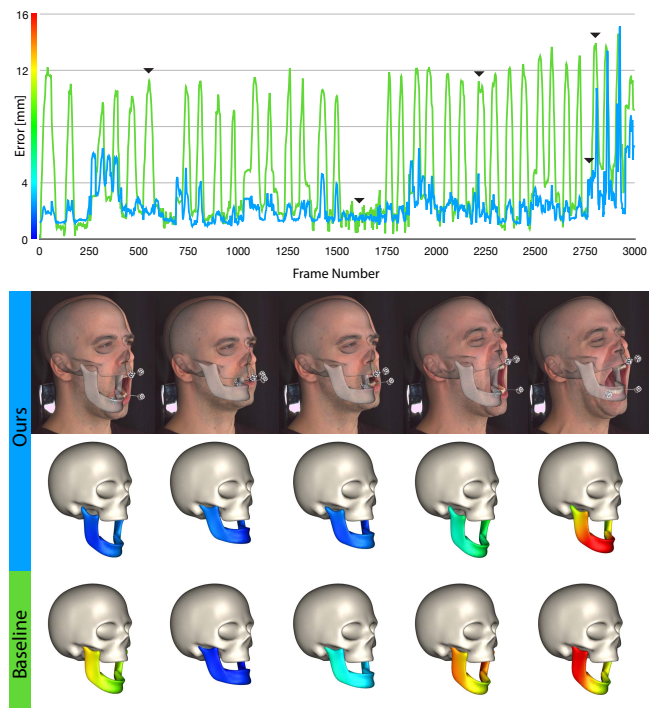


Fig. 10. **Limitations.** Subject S_1 exhibits the highest prediction error as listed in Table 1. Our explanation is that the subject plays a critical role for training, since it contributes not only the lion share of training data ($>30\%$), but also the best structured and most extreme data. This becomes particularly apparent towards the end, where the subject enters extreme poses that are outside of the training samples. We believe that by enriching our training pool with more subjects, these limitations would be overcome.

P. O. Eriksson, B. Häggman-Henrikson, E. Nordh, and H. Zafar. 2000. Co-ordinated mandibular and head-neck movements during rhythmic jaw activities in man. *Journal of Dental Research* 79, 6 (2000), 1378–1384. <https://doi.org/10.1177/0022034500079060501>

Virgilio F. Ferrario, Chiarella Sforza, Nicola Lovecchio, and Fabrizio Mian. 2005. Quantification of translational and gliding components in human temporomandibular joint during mouth opening. *Archives of Oral Biology* 50, 5 (2005), 507–515. <https://doi.org/10.1016/j.archoralbio.2004.10.002>

Graham Fyffe, Tim Hawkins, Chris Watts, Wan-Chun Ma, and Paul Debevec. 2011. Comprehensive Facial Performance Capture. In *Eurographics*.

G Fyffe, K Nagano, L Huynh, S Saito, J Busch, A Jones, H Li, and P Debevec. 2017. Multi-View Stereo on Consistent Face Topology. *Comput. Graph. Forum* 36, 2 (2017), 295–309.

Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. 2013. Reconstructing Detailed Dynamic Face Geometry from Monocular Video. In *[ACM] Trans. Graph. (Proceedings of SIGGRAPH Asia 2013)*, Vol. 32. 158:1–158:10.

John C Gower. 1975. Generalized procrustes analysis. *Psychometrika* 40, 1 (1975), 33–51.

Pei-Lun Hsieh, Chongyang Ma, Jihun Yu, and Hao Li. 2015. Unconstrained Realtime Facial Performance Capture. In *Computer Vision and Pattern Recognition (CVPR)*.

Davis E King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

Soichiro Kinuta, Kazumichi Wakabayashi, Taiji Sohmura, Tetsuya Kojima, Takahiro Mizumori, Takashi Nakamura, Junzo Takahashi, and Hirofumi Yatani. 2005. Measurement of Masticatory Movement by a New Jaw Tracking System Using a Home Digital Camcorder. *Dental Materials Journal* 24, 4 (2005), 661–666. <https://doi.org/10.4012/dmj.24.661>

Samuli Laine, Tero Karras, Timo Aila, Antti Herva, Shunsuke Saito, Ronald Yu, Hao Li, and Jaakko Lehtinen. 2017. Production-level Facial Performance Capture Using Deep Convolutional Neural Networks. In *Proc. SCA*. 10:1–10:10.

Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. 2013. Realtime facial animation with on-the-fly correctives. *ACM Transactions on Graphics* 32, 4 (2013), 42:1–42:10. <https://doi.org/10.1145/2461912.2462019> arXiv:1111.6189v1

Naser Mostashiri, Jaspreet Dhupia, Alexander Verl, and Weiliang Xu. 2018. A Novel Spatial Mandibular Motion-Capture System Based on Planar Fiducial Markers. *IEEE Sensors Journal* 18, 24 (2018), 10096–10104. <https://doi.org/10.1109/JSEN.2018.2873349>

M. G. Piancino, T. Vallelonga, C. Debernardi, and P. Bracco. 2013. Deep bite: A case report with chewing pattern and electromyographic activity before and after therapy with function generating bite. *European Journal of Paediatric Dentistry* 14, 2 (2013), 156–159.

Al P Pinheiro, A O Andrade, A A Pereira, and D Bellomo. 2008. A computational method for recording and analysis of mandibular movements. *Journal of applied oral science : revista FOB* 16, 5 (2008), 321–7. <https://doi.org/10.1590/S1678-7752008000500004>

J. F. Prinz. 1997. The cybermouse: A simple method of describing the trajectory of the human mandible in three dimensions. *Journal of Biomechanics* 30, 6 (1997), 643–645. [https://doi.org/10.1016/S0021-9290\(97\)00012-2](https://doi.org/10.1016/S0021-9290(97)00012-2)

Isa C T Santos, João Manuel R S Tavares, Joaquim G Mendes, and Manuel P F Paulo. 2006. A System for Analysis of the 3D Mandibular Movement using Magnetic Sensors and Neuronal Networks. *Proceedings of the 2nd International Workshop on Artificial Neural Networks and Intelligent Information Processing 2006* (2006), 54–63. <https://doi.org/10.5220/0001208000540063>

Fuhao Shi, Hsiang-Tao Wu, Xin Tong, and Xinxiang Chai. 2014. Automatic Acquisition of High-fidelity Facial Performances Using Monocular Videos. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2014)* 33, 6 (2014).

Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M Seitz. 2014. Total Moving Face Reconstruction. In *ECCV*.

Yuto Tanaka, Takafumi Yamada, Yoshinobu Maeda, and Kazunori Ikebe. 2016. Markerless three-dimensional tracking of masticatory movement. *Journal of Biomechanics* 49, 3 (2016), 442–449. <https://doi.org/10.1016/j.jbiomech.2016.01.011>

Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. 2017. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *Proc. of IEEE ICCV*.

Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. 2015. Real-time Expression Transfer for Facial Reenactment. *ACM Trans. Graph.* 34, 6 (2015), 183:1–183:14.

J Thies, M Zollhöfer, M Stamminger, C Theobalt, and M Nießner. 2016. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *Proc. of IEEE CVPR*.

J Thies, M Zollhöfer, M Stamminger, C Theobalt, and M Nießner. 2018. HeadOn: Real-time Reenactment of Human Portrait Videos. *ACM Transactions on Graphics 2018 (TOG)* (2018).

Levi Valgaerts, Chenglei Wu, Andrés Bruhn, Hans-Peter Seidel, and Christian Theobalt. 2012. Lightweight Binocular Facial Performance Capture under Uncontrolled Lighting. *ACM Transactions on Graphics* 31, 6 (2012), 187:1–187:11. <https://doi.org/10.1145/2366145.2366206>

Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. 2011. Realtime Performance-Based Facial Animation. *ACM Trans. Graphics (Proc. SIGGRAPH)* 30, 4 (2011), 77:1–77:10.

B. Wiesinger, B. Häggman-Henrikson, A. Wänman, M. Lindkvist, and F. Hellström. 2014. Jaw-opening accuracy is not affected by masseter muscle vibration in healthy men. *Experimental Brain Research* 232, 11 (2014), 3501–3508. <https://doi.org/10.1007/s00221-014-4037-3>

Erin M Wilson and Gary Weismer. 2012. Motion for Early Chewing : Preliminary Findings. *Journal of Speech, Language, and Hearing Research* 55, 2 (2012), 626–638. [https://doi.org/10.1044/1092-4388\(2011/10-0236\).A](https://doi.org/10.1044/1092-4388(2011/10-0236).A)

A. M. Wintergerst, P. H. Buschang, and G. S. Throckmorton. 2004. Reducing within-subject variation in chewing cycle kinematics - A statistical approach. *Archives of Oral Biology* 49, 12 (2004), 991–1000. <https://doi.org/10.1016/j.archoralbio.2004.07.010>

Chenglei Wu, Derek Bradley, Markus Gross, and Thabo Beeler. 2016. An anatomically-constrained local deformation model for monocular face capture. *ACM Transactions on Graphics* 35, 4 (2016), 1–12.

Wenwu Yang, Nathan Marshak, Daniel Šykora, Srikumar Ramalingam, and Ladislav Kavan. 2018. Building Anatomically Realistic Jaw Kinematics Model from Data. *CoRR* abs/1805.0 (2018). arXiv:1805.05903 <http://arxiv.org/abs/1805.05903>

H. Zafar, P. O. Eriksson, E. Nordh, and B. Häggman-Henrikson. 2000. Wireless optoelectronic recordings of mandibular and associated head-neck movements in man: A methodological study. *Journal of Oral Rehabilitation* 27, 3 (2000), 227–238. <https://doi.org/10.1046/j.1365-2842.2000.00505.x>

Gaspard Zoss, Derek Bradley, Pascal Bérard, and Thabo Beeler. 2018. An Empirical Rig for Jaw Animation. *ACM Transactions on Graphics* 37, 4 (2018), 59:1–59:12. <https://doi.org/10.1145/3197517.3201382>