Data-driven Extraction and Composition of Secondary Dynamics in Facial Performance Capture

GASPARD ZOSS, DisneyResearch|Studios and ETH Zurich
EFTYCHIOS SIFAKIS, University of Wisconsin-Madison and DisneyResearch|Studios
MARKUS GROSS, DisneyResearch|Studios and ETH Zurich
THABO BEELER, DisneyResearch|Studios
DEREK BRADLEY, DisneyResearch|Studios

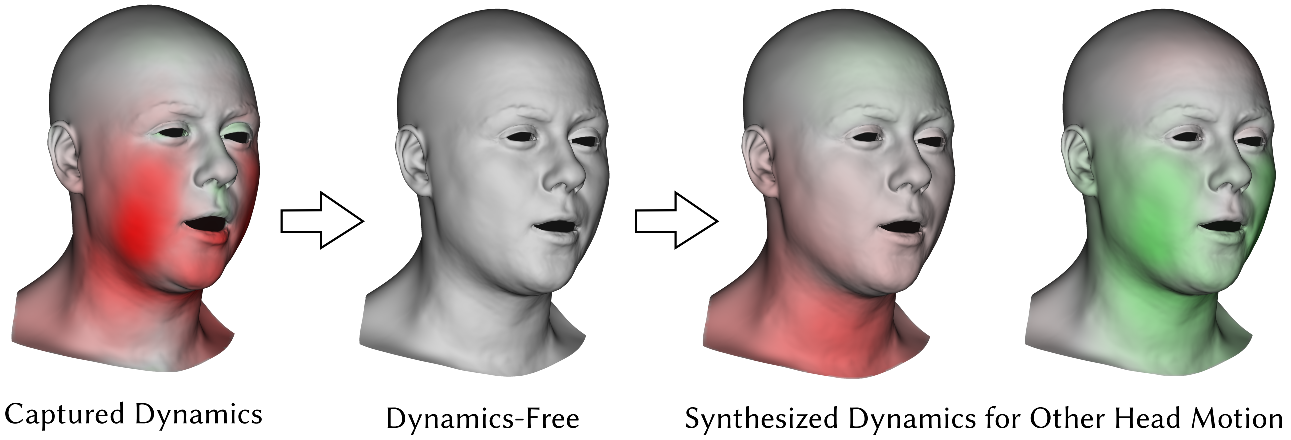

Fig. 1. We present a data-driven method to remove secondary dynamic effects from performance capture data, such as jiggling skin due to root skull motion, and a complementary method to synthesize dynamics under different root motion. Here, the color (red/green) represents the signed distance from the dynamics-free performance (gray).

Performance capture of expressive subjects, particularly facial performances acquired with high spatial resolution, will inevitably incorporate some fraction of motion that is due to inertial effects and dynamic overshoot due to ballistic motion. This is true in most natural capture environments where the actor is able to move freely during their performance, rather than being tethered to a fixed position. Normally these secondary dynamic effects are unwanted, as the captured facial performance is often retargeted to different head motion, and sometimes to completely different characters, and in both cases the captured dynamic effects should be removed and new secondary effects should be added. This paper advances the hypothesis that for a highly constrained elastic medium such as the human face, these secondary inertial effects are predominantly due to the motion of the underlying bony structures (cranium and mandible). Our work aims to compute and characterize the difference between the captured dynamic facial performance, and a speculative quasistatic variant of the same motion should the inertial effects have been absent. This is used to either subtract parasitic secondary dynamics that resulted from unintentional motion during capture, or compose such effects on top of a quasistatic performance to simulate a new dynamic motion of the actor's body and skull, either artist-prescribed or acquired via motion capture. We propose a data-driven technique that comprises complementary removal and synthesis networks for secondary dynamics in facial performance capture. We show how such a system can be effectively trained from a collection of acquired dynamic deformations under varying expressions where the actor induces rigid head motion from walking and running, as well as forced oscillatory body motion in a controlled setting by external actuators.

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

Additional Key Words and Phrases: Secondary Dynamics Prediction, Soft-Tissue Motion, Facial Performance Capture, Data-Driven Animation

ACM Reference Format:
Gaspard Zoss, Eftychios Sifakis, Markus Gross, Thabo Beeler, and Derek Bradley. 2020. Data-driven Extraction and Composition of Secondary Dynamics in Facial Performance Capture. *ACM Trans. Graph.* 39, 4, Article 1 (July 2020), 10 pages. https://doi.org/10.1145/3386569.3392463

Authors' addresses: Gaspard Zoss, DisneyResearch|Studios, ETH Zurich, gaspard.zoss@disneyresearch.com; Eftychios Sifakis, University of Wisconsin-Madison, DisneyResearch|Studios, sifakis@cs.wisc.edu; Markus Gross, DisneyResearch|Studios, ETH Zurich, gross@disneyresearch.com; Thabo Beeler, DisneyResearch|Studios, thabo.beeler@gmail.com; Derek Bradley, DisneyResearch|Studios, derek.bradley@disneyresearch.com.

0730-0301/2020/7-ART1 $15.00
https://doi.org/10.1145/3386569.3392463

1 INTRODUCTION

Facial performance capture is an industry-standard method for generating facial animation of digital characters. More and more, actors are donning helmet-mounted cameras that track their facial movements as they run and jump around film sets, and the resulting performances are reconstructed in 3D and transferred onto virtual characters. Our field has seen tremendous advances in perfecting the 3D shape recovery and tracking of the face, all the way down to separating the the performance into meaningful layers, such as the rigid head motion and the non-rigid facial deformation - a crucial step in order to retarget the performance to a virtual character. However, an important aspect that has been ignored thus far in facial capture is that the recovered facial motion will contain both the desired expression deformation as well as undesired inertial deformation such as jiggling, especially during fast and jerky motions, but even visible during simple motions like walking. In typical retargeting scenarios, the performance is mapped to a virtual character that often has different face proportions and almost always different rigid head motion. Thus, the secondary dynamic motion that is present in the captured performance will not match the character after retargeting. In this work, we present the first method to explicitly model and remove secondary dynamic effects from facial performance capture, and provide the ability to compose new dynamic effects to artist-scripted head movements of the character.

Even though human soft tissue is largely constrained to follow the skeletal motion, flesh stiffness is soft enough to manifest inertial deformation when the underlying bones accelerate, and can cause jiggly motion when the acceleration changes direction. It is easier to appreciate and quantify the magnitude of the "jiggly" components of motion when contrasted with a *quasistatic* model of skin deformation: We can intuitively think of quasistatic animation as taking the keyframed sequence of all factors that drive skin deformation (e.g. skeletal motion, active muscle contraction), and time-scaling those keyframes so that the motion occurs much more slowly, and over a much longer period of time. The slower we allow this procession to take place, the more we suppress inertial motion by means of elastic damping; at the limit of infinite time scaling, the motion can be regarded as devoid of inertial dynamics. Scaling back the animation to the original duration produces what we define as the corresponding quasistatic performance. Our work seeks to identify in a given performance capture sequence the difference between the captured dynamic motion and what the corresponding dynamics-free, quasistatic motion would have been. This idea can lead to either a filter that removes secondary dynamics from a jiggly capture caused by unintentional motion, or a means to compose artist-directed root skeletal motion on top of a quasistatic baseline by synthesizing the secondary dynamics that this motion would incur.

Inertial dynamics can be triggered or modulated by a number of factors. In the human face those would be primarily the motion of the skeletal components (cranium and mandible) and the contractile action of muscles that form expressions or articulate the jaw. We would consider these as intrinsic drivers of skin deformation; extrinsic factors such as contact with objects or the action of external forces (e.g. wind) will be excluded from our investigation. We also consciously omit gravity as a separately parameterized influencing factor, with the understanding that a change in pose that causes gravity to act in a different direction relative to the frame of the face would be encoded in the motion of the skull itself.

We will use the term *expression* to collectively refer to muscle action that either triggers facial expressions, or moves the jaw relative to the cranium. Expression has a particular effect on secondary dynamics, that is qualitatively distinct from the influence that head motion incurs. Although it is conceivable that an exceptionally fast twitch of a muscle can trigger some overshoot in skin deformation all by itself, the magnitude of such muscle forces relative to the damping capacity of the flesh renders dynamics of expression rather negligible as pure triggers of secondary dynamics. However, the formation of expression has a much more prominent *modulating* effect on jiggly motion that is instigated by head kinematics, by making certain parts of the face more stiff, while making other areas loosen up and more susceptible to secondary dynamics. The influence of the kinematic motion of the skeletal bones is the more direct contributor to the observed secondary dynamics; in fact, the observed dynamic deformation depends on the *history* of kinematic motion, as momentum builds up over time. We claim nevertheless that only a finite-length history of skeletal kinematics is required to infer secondary motion, due to the dissipatory damping behavior of soft tissue, along with a history of the local dynamic skin behavior.

Our work proposes and tests three central hypotheses. In the absence of extrinsic influencing factors (e.g. forces or collision):

- The difference between a given dynamic motion acquired by a performance capture system and a quasistatic dynamics-free version of the same performance can be adequately inferred from (a) the kinematic history of the underlying bone within a short window of time (forward and backward), and (b) the corresponding sequence of the dynamic skin motion that includes secondary effects, but without requiring knowledge of the performed expression.
- The difference between a given quasistatic skin motion performance and the corresponding dynamic performance exhibiting secondary effects due to rigid head motion can be inferred from (a) the kinematic history of the underlying bone within a short window of time (forward and backward), and (b) a representation of the current facial expression, parameterized by surface stretch.
- A data-driven model performing the aforementioned dynamics removal and synthesis can be constructed using a motion corpus that exemplifies (a) a range of distinctive facial expressions, combined with (b) a range of amplitudes and frequencies of secondary motion, induced by walking, running, jumping, and for more controlled input data: forcing an oscillatory motion of the head via a multi-speed vibrating actuator.

Based on these hypotheses we present a data-driven deep learning approach for the tasks of removal and synthesis of secondary dynamics in performance capture. Note that even though both tasks aim to predict the delta between quasistatic and dynamic motion, the processes are decoupled, since fundamentally different input data is available for each task. Specifically, for dynamics removal we know the kinematic history of the skin with secondary motion, but cannot infer an exact facial expression due to the parasitic secondary effects, whereas for dynamics synthesis we can determine the facial expression from the quasistatic performance but have no information about the dynamic skin history. Therefore, we construct and train two separate but complementary networks, one for dynamics removal and one for synthesis. Our methodology does not presume knowledge of the volumetric geometry and physical properties of the underlying subject, and operates purely on sequences of animated surface meshes, and without requiring simulation in the processing loop.

We demonstrate the efficacy of our networks in two complementary motion processing tasks: removal of secondary dynamics in capture data to estimate the corresponding quasistatic skin deformation, and data-driven emulation of secondary dynamics in a synthetic skeletal motion sequence (distinct from the one in the performance capture). We apply our method on three distinct individuals, showing robustness of the algorithm across facial structure and material composition.

2 RELATED WORK

We briefly review related prior research in the domains of performance capture, human character animation and simulation, on which our work draws inspiration or has notable thematic affinity.

Simulation-based animation. Simulation from first principles is one possibility for generating high-quality facial animation [Cong et al. 2016]. Muscle activations are an intuitive expression descriptor, while collisions and external forces are handled naturally. However, a prerequisite for quality in these techniques is accurate knowledge of anatomical geometry and material properties, neither of which is trivial to acquire or model. Unreduced simulation is also costly, with highly detailed approaches [Sifakis et al. 2006] often opting for quasistatic approximations in lieu of full dynamic simulation. Anatomical models that admit a lower-dimensional parametric approximation of bulk motion offer opportunities for peformance acclerations; skeleton-driven full body models with an underlying rig enable cost-saving simplifications of the governing equations [Capell et al. 2007] or open the possibility of the elastic deformation problem to be solved directly in the deformation space defined by the rig [Hahn et al. 2012, 2013]. Subspace deformation schemes can also be hybridized with rig-actuated character models to enrich them with secondary dynamics [Xu and Barbič 2016]. For facial simulation, in particular, the burden of generating subject-specific models of musculature can be alleviated by inferring blendshape-like deformation descriptors from input scans [Ichim et al. 2016] with an associated simulation that tracks the generated shape targets. These shape descriptors can be used as replacement of traditional contractile muscles and selectively activated [Ichim et al. 2017]. The descriptors may also include expression-specific material parameters [Kozlov et al. 2017] or forces [Barrielle et al. 2016], creating so-called blendmaterial and blendforce analogs to blendshape animation.

The aforementioned techniques all require, to varying degrees, some structural knowledge about the underlying physical object: a mesh representation of the flesh volume, an animation rig, muscle geometries etc. The distribution of material properties is also often a requirement, although those can often be inferred from an anatomical template, or learned from data, even for damping behaviors [Xu and Barbič 2017]. Our goal in this paper is to remove any such implicit requirements, and synthesize secondary dynamics for high-resolution animated face meshes, using only the moving meshes and skull motion obtained by performance capture as the input to our method, along with a very small subset (≈ 10) of precaptured extreme expression scans of the actor. Finally, methods that define dynamic elastic deformations of space without an explicit volumetric description of a material body [De Goes and James 2017] can be used to craft secondary dynamics, but our method also can modulate the shape and amplitude of such secondary motion in a pose/expression dependent fashion.

Data-driven techniques. A significant segment of prior research tackled dynamic animation problems by leveraging collections of performance capture data. Some of the most influential early advances in the domain of dynamic human body motion leveraged marker-based motion capture [Park and Hodgins 2006], and used to synthesize new motion with realistic secondary dynamics [Park and Hodgins 2008]. Our approach to synthesizing secondary dynamics parallels their approach in that we estimate the dynamic motion of discrete locations on a moving model, although we furnish that estimate as a direct function of the kinematic history of the underlying bones at any given time instance, rather than integrating forward in time an elementary oscillator; furthermore, in their work constant parameters were inferred for said oscillators from data, while the output of our data-driven technique is modulated via an additional input of the local deformation, as a proxy to a local expression descriptor. Combining motion capture performance data with 3D body scans for distinct poses and human subjects allowed the generation of models that capture body shape change due to both pose and identity (i.e. body type) [Anguelov et al. 2005]. Such models laid the foundation for follow-up work that incorporated the ability for synthesis of dynamic motion, that were in turn highly influential for our own work: The *Dyna* system [Pons-Moll et al. 2015] used a second-order auto-regressive model to synthesize dynamic motion using the kinematics of the root coordinate system, as well as the velocities and accelerations of pose parameters. The Dynamic-SMPL model [Loper et al. 2015] demonstrates how analogous dynamic deformations can be produced with computations on a per-vertex basis, rather than using triangle deformations. Follow up work [Kim et al. 2017] demonstrated how such a statistical body model can be used to animate all but the top layers of a tetrahedralized body model, driving a physics-based simulation using the Finite Element Method for the topmost layer of the body mesh that produces realistic elastic, dynamic response. The recently introduced *SoftSMPL* model [Santesteban et al. 2020] combines a deep recurrent regressor with a non-linear deformation subspace to synthesize soft-tissue dynamics from body shape and motion inputs.

Our proposed approach, similar to prior work recognizes the kinematics of the model root as the primary instigator of secondary dynamics. A significant difference however with similar papers, especially those that model whole body motion [Kim et al. 2017; Loper et al. 2015; Pons-Moll et al. 2015; Santesteban et al. 2020] is that those works explicitly utilize velocities and accelerations of pose/joint parameters (beyond those of the body root) to infer dynamic secondary motion. For facial animation, the most direct analogue to pose or joint parameters would be descriptors of expressions, such as muscle activations, blendshape weights, etc. Our proposed approach has conceptual differences with this paradigm, both for its removal and its synthesis stage. For removal, we only rely on the kinematic history of the root head frame, as well as a brief temporal window of the dynamic trajectory of any model vertex, which is not associated with any specific joint or collection of joints (or their temporal derivatives). For synthesis, the only modifier that is used as input to our system – in addition to the head kinematics – is a local measure of surface deformation, which is evaluated only at the present point in time without reliance on its history. One can see that the feasibility of this hypothesis on faces would not extend to other types of secondary dynamics that are only loosely bound to the underlying skeletal motion, such as clothing [de Aguiar et al. 2010] where authors rightly incorporate a description of the inertial state of the cloth surface in their evolution scheme. In the domain of faces, however, our hypothesis is motivated by the observation that (a) even extremely fast twitch of muscles is unlikely to inherently trigger ballistic overshoot, and (b) the facial flesh tissue is a highly damped, highly constrained layer that typically subdues oscillatory displacements within a fraction of a second.

Facial performance capture. Performance-based 3D face tracking has become an industry-standard for character facial animation. Highest-quality facial capture methods typically use multi-view camera setups and sophisticated reconstruction algorithms [Beeler et al. 2011; Bradley et al. 2010; Fyffe et al. 2011, 2017], although advances in model-based face fitting [Garrido et al. 2013; Shi et al. 2014; Suwajanakorn et al. 2014; Wu et al. 2016] and deep learning [Laine et al. 2017; Tewari et al. 2017] can reduce the hardware burden. All of the previous performance capture methods suffer from the same limitation - none consider the separation of voluntary deformation through expression and involuntary motion due to inertial dynamics. Our work is the first to explicitly address this important aspect. As we take a data-driven approach, our method relies on accurate performance capture input. For this we use the local anatomical model fitting method of Wu et al. [2016], which is particularly good at recovering local deformations caused by dynamics and can perform high quality shape tracking from just a small number of input videos.

3 OUTLINE

We present a data-driven approach for modeling the difference between a facial performance exhibiting inertial, secondary dynamics and the same performance in the absence of inertial effects, with the goal of transforming animation sequences in either direction. To this end, we propose a deep learning approach to predict a per-frame, per-point offset vector that removes secondary dynamics from a performance capture sequence (Section 5), and a complementary network that predicts the inverse offset given a quasistatic performance sequence allowing to synthesize secondary dynamics (Section 6). The input to the removal network is a short time window of bone velocity values, both past and future, together with a corresponding window of stabilized skin velocities for the given point. Stabilization here means that the head motion has been removed, i.e. the velocities are computed in a canonical reference frame. The input to the synthesis network is a short time window of the new desired skull velocities and the skin surface stretch as a signature of the current expression. The networks are trained on a large corpus of motion patterns, including walking, running, and jumping, as well as controlled, repeatable actuation using a vibrating platform with variable frequency to instigate rapid skeletal motion.

Next we describe how the data used for training has been captured and preprocessed (Section 4). We will then introduce our removal (Section 5) and synthesis networks (Section 6) followed by a thorough evaluation (Section 7) and discussion (Section 8).

4 PERFORMANCE CAPTURE

As our method is data-driven, we require high-quality 3D facial performance data captured with varying amounts of secondary dynamic effects. To this end, we built a custom multi-view camera setup and use the facial tracking method of Wu et al. [2016] (see Fig. 2, a). This reconstruction algorithm is chosen since it demonstrates exceptional accuracy, both in terms of expression recovery but also in capturing the local effects caused by dynamic motion, and the algorithm produces a time sequence of face meshes in vertex correspondence complete with the rigid motion of the underlying skull, which will be essential for modeling the dynamics in our work.

In this work we focus on secondary effects caused by the root node of the head, i.e. the skull motion. We capture multiple subjects undergoing several typical motions, such as walking, running, or jumping. In order to provide more control and provoke more extreme dynamics we further capture the subjects using an oscillating "step" platform designed for fitness (see Fig. 2, b). Standing on one end of the vibrating platform creates repeatable up-down motion that is translated throughout the body to the skull, and varying the speed of oscillation creates a range of secondary dynamic effects on the face.

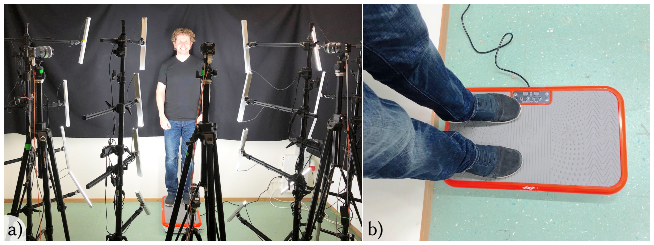

Fig. 2. **Capture Setup:** We use a multi-view capture setup for reconstructing facial performances (a). In addition to capturing motions such as walking or jumping, we also employ a vibrating "step" platform that oscillates at varying speeds in order to induce controllable and repeatable dynamic effects (b).

As we wish to explore how dynamics vary with expression, we capture the subject under a number of different facial expressions, each one held for approximately 3 seconds while performing the above-mentioned motion patterns, plus various oscillatory speeds on the vibrating platform, ranging from 6Hz to 11Hz. We also obtain a static reconstruction of each expression using the facial scanning technique of Beeler et al. [2011], from which we build an Anatomical Local Model as proposed by Wu et al. [2016]. We use the 10 expressions suggested in their paper plus an expression with relaxed, slightly parted lips, which we found important since for the other expressions the lip muscles are always stiffened. As we ultimately want to model the dynamics for generic performances, we also capture speech sequences while performing the same root motion sequences, plus once without skull motion for reference. An overview of the captured data is shown in Fig. 3 and more details are provided in Section 7.1.

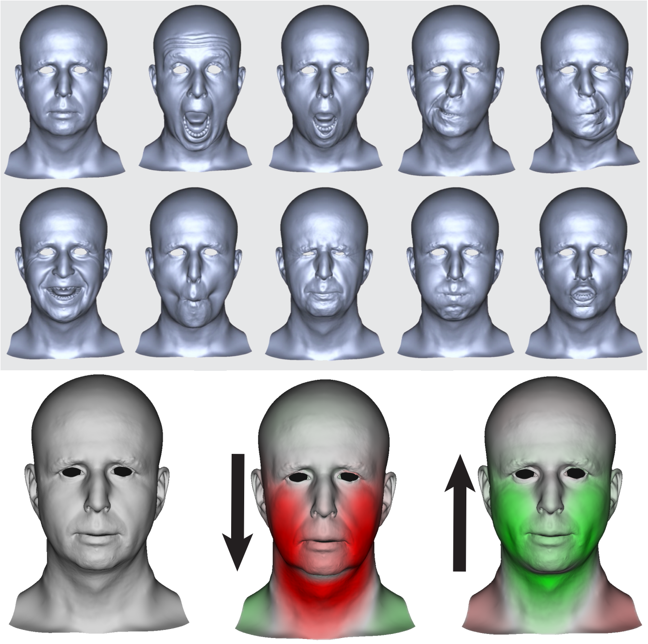

Fig. 3. **Data Acquisition:** We capture neutral plus 10 expressions and dialog under various motions, such as running or jumping, as well as under 6 different up-down vibration frequencies. Top: An overview of the expressions. Bottom: The dynamic motion induced by the highest oscillation frequency significantly deforms the neutral mesh (left) , in particular around the cheeks and corners of the mouth as shown at the extremes of the down (middle) and up (right) positions just after reversing direction. The coloring red/green encodes the signed difference with respect to the quasistatic performance, with a scale ranging from +/- 5mm

5 MODELING AND REMOVAL OF SECONDARY DYNAMICS

Our objective is to define a mapping $\mathcal{F}(\mathcal{B}(t), \mathcal{X}_i(t)) \longmapsto \delta\mathbf{x}_i(t)$ to predict the dynamic offsets $\delta\mathbf{x}_i(t)$ of individual vertices i for performance capture sequences over time t, as shown schematically in Fig. 4.c. We will subsequently define $\mathcal{B}(t)$ as a descriptor of a short-term kinematic window of the skull and $\mathcal{X}_i(t)$ as a descriptor of a short-term kinematic window of an individual vertex i on the face. Since this prediction is not the result of a temporal evolution, the prediction can be computed directly on any given frame of the performance sequence, considering both past and future.

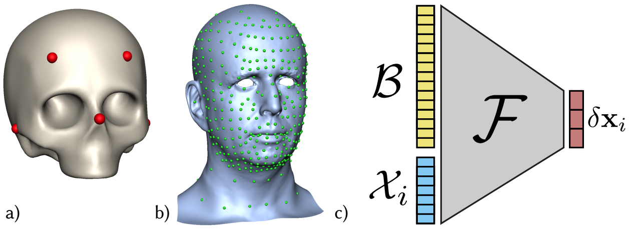

Fig. 4. **Removal Overview:** For five sample points $\mathbf{b}_{1-5}$ on the skull (a) and 344 sample points $\mathbf{x}_i$ on the skin (b), feature vectors $\mathcal{B}$ and $\mathcal{X}_i$ are computed from the velocity of the points within a time window (c). From those features, the neural network $\mathcal{F}$ predicts the displacement vector $\delta\mathbf{x}_i$ due to secondary dynamics.

5.1 Feature Modeling

The head motion history descriptor $\mathcal{B}(t)$ is defined as:

$$\begin{aligned}\mathcal{B}(t) \quad = \quad & \Big\{\big\{\dot{\mathbf{b}}_1(t-w_b), \ldots, \dot{\mathbf{b}}_1(t), \ldots, \dot{\mathbf{b}}_1(t+w_b)\big\}, \\ & \big\{\dot{\mathbf{b}}_2(t-w_b), \ldots, \dot{\mathbf{b}}_2(t), \ldots, \dot{\mathbf{b}}_2(t+w_b)\big\}, \\ & \quad \vdots \\ & \big\{\dot{\mathbf{b}}_5(t-w_b), \ldots, \dot{\mathbf{b}}_5(t), \ldots, \dot{\mathbf{b}}_5(t+w_b)\big\}\Big\},\end{aligned} \tag{1}$$

where w_b denotes the half-size of the temporal window, which we set to 10 frames in our implementation (corresponding to approximately 150ms in a 128fps capture sequence). The components of $\mathcal{B}(t)$ correspond to finite difference approximations of the linear velocities of the five selected skull landmarks $\mathbf{b}_{1-5}$ (see Fig. 4.a), defined as

$$\dot{\mathbf{b}}_i(t) = \frac{\left[\mathbf{I} - \mathbf{T}(t)^{-1}\mathbf{T}(t-1)\right]\mathbf{b}_i}{\Delta t}, \tag{2}$$

where the 4x4 matrix $\mathbf{T}(t)$ denotes the rigid head transformation at time t and $\mathbf{I}$ is the identity matrix. Applying the composed transform $\mathbf{I} - \mathbf{T}(t)^{-1}\mathbf{T}(t-1)$ to $\mathbf{b}_i$ instead of computing the finite difference directly from $\mathbf{b}_i(t)$ and $\mathbf{b}_i(t-1)$ ensures that the velocity is computed relative to a canonical coordinate frame, factoring out the absolute head pose.

Similarly, the skin motion descriptor $\mathcal{X}_i(t)$ is defined as

$$\mathcal{X}_i(t) = \{\dot{\mathbf{x}}_i(t-w_x), \ldots, \dot{\mathbf{x}}_i(t), \ldots, \dot{\mathbf{x}}_i(t+w_x)\}, \tag{3}$$

where w_x denotes again the half-size of the temporal window. In our experiments we set $w_x = w_b$. The components of $\mathcal{X}_i(t)$ correspond to finite difference approximation of the linear velocities of a selected set of skin landmarks $\mathbf{x}_i$, again expressed relative to a canonical coordinate frame. We define those finite differences as

$$\dot{\mathbf{x}}_i(t) = \frac{\mathbf{T}(t)^{-1}\mathbf{x}_i(t) - \mathbf{T}(t-1)^{-1}\mathbf{x}_i(t-1)}{\Delta t}. \tag{4}$$

Since the effect of secondary dynamics on skin is spatially smooth we use a subset of 344 samples distributed over the face as shown in Fig. 4.b.

5.2 Training and inference

As stated in Section 4, our performance capture sequences includes footage of 11 distinct expressions, where the captured subject was asked to sustain the same facial expression while performing various tasks such as walking or running in place. Additionally, the captured subject was asked to sustain those expressions while standing on a vibrating fitness platform, such that the head was subject to externally induced oscillations. In this setting, the corresponding idealized quasistatic performance would be the fixed reference expression, transforming rigidly from frame to frame bare from any deformation. Of course, we cannot expect the captured subjects to perfectly hold an expression for several seconds while experiencing induced head motion. However, deformation induced by expression change will take place at a much larger time-scale compared to secondary dynamics coming from impact impulse forces or oscillatory actuation. Hence we use a centered moving average to generate the quasistatic performance

$$\mathbf{y}_i(t) = \mathbf{T}(t)\left(\frac{1}{2w+1}\sum_{k=t-w}^{t+w}\mathbf{T}(k)^{-1}\mathbf{x}_i(k)\right), \quad (5)$$

with the expectation that, after factoring out the rigid head motion, dynamic deformation away from the quasistatic sustained expression average out over the time window. In our experiments, we used $w = 10$ frames, leading to an averaging window of ~ 300ms. We can now define the dynamic offset $\delta\mathbf{x}_i(t)$ in the canonical coordinate frame as

$$\delta\mathbf{x}_i(t) = \mathbf{T}(t)^{-1}\left(\mathbf{x}_i(t) - \mathbf{y}_i(t)\right). \quad (6)$$

With these features, we train a feed-forward fully connected neural network with 2 hidden layers of size 128 and ReLU activation functions to learn the desired mapping $\mathcal{F}$ from the training data thus assembled. Note that we do not learn a mapping per skin landmark i, but a single mapping trained from all skin landmarks. Since the features do not contain any form of identity, the network pools information from all observations and can hence be expected to generalize better. The network is trained with L2 loss using the Adam optimizer [Kingma and Ba 2014], with a batch size of ten thousand. With a learning rate of $1e$-3 training converges after 250 epochs. We implemented the network in PyTorch [Paszke et al. 2019]. The depth and width of the network were determined through empirical evaluation, by selecting the smallest network capable of predicting reasonable vertex offsets in a validation set. In practice, we found that our dynamics removal results are not very susceptible to such hyper-parameters.

The offset vector predicted by the trained network can then be subtracted from the dynamic performance to recover the desired quasistatic positions of the skin landmarks as

$$\mathbf{y}_i(t) = \mathbf{x}_i(t) - \mathbf{T}(t)\mathcal{F}(\mathcal{B}(t), \mathcal{X}_i(t)). \quad (7)$$

Using $\mathbf{y}_i(t)$ as 3D positional constraints, we employ the Anatomical Local Model (ALM) proposed by Wu et al.[2016] to propagate the deformation over the entire surface. Other methods, such as Laplacian mesh deformation [Sorkine et al. 2004], could also be used, but the ALM has a more meaningful regularizer that takes the actual shape deformation subspace into account.

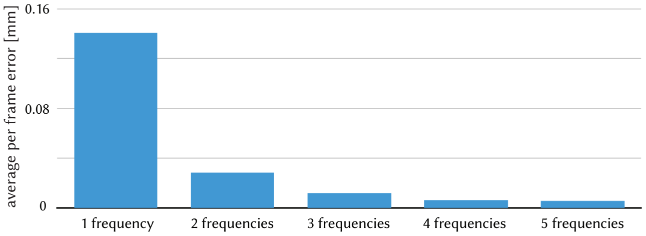

Fig. 5. **Removal Validation:** We validate our removal network by training it separately on five different subsets of the data captured using the vibrating platform, and testing removal of unseen motion. The first subset (1 frequency) contains all the sustained expressions captured only at the slowest frequency. The second subset (2 frequencies) adds the second slowest frequency (and so on). The error bars represent the average error of all predicted points over all the frames of the sequence. In all five experiments, the test sequence was a vibration frequency that was completely omitted during training.

5.3 Validation of Dynamics Removal

To validate the removal of dynamic motion, we select one of the sequences for which we can compute the quasistatic equivalent (neutral expression at the fastest frequency) and train multiple mappings $\mathcal{F}$ on a subset of the data from the vibrating platform. Using the reconstructed head motion of the the selected sequence, we predict $\delta\mathbf{x}_i(t)$ and compare them to the captured offsets. Fig. 5 shows the average error of all predicted $\delta\mathbf{x}_i(t)$ for all frames of the sequence. Each bar correspond to training a mapping $\mathcal{F}$ with the subset of the data captured at one to five frequencies. This experiment shows that the network can interpolate and even extrapolate to unseen motions.

6 SYNTHESIS AND COMPOSITION OF SECONDARY DYNAMICS

Our objective is to define a mapping $\mathcal{H}(\mathcal{B}(t), \mathcal{S}(t)) \longmapsto \delta\mathbf{Y}(t)$ to synthesize dynamic offsets $\delta\mathbf{Y}(t) = \{\delta\mathbf{y}_i(t)\}$ given a quasistatic animation of vertex positions $\mathbf{y}_i(t)$, as shown schematically in Fig. 6.b. The skull motion feature vector $\mathcal{B}(t)$ is defined analogously to the removal case (Section 5) over the same centered time window. In addition to the kinematic history of the head we also require a descriptor of the current expression, since expression will influence the secondary dynamics as the underlying anatomical structures change due to muscle activation.

6.1 Feature Modeling

We propose to leverage surface stretch as a local descriptor of how the surface changes, computed in uv-space as

$$\mathbf{s}_i(t) = \left[\frac{\|\Delta_u\mathbf{y}_i(t)\| - \|\Delta_u\mathbf{y}_i(0)\|}{\|\Delta_u\mathbf{y}_i(0)\|}, \frac{\|\Delta_v\mathbf{y}_i(t)\| - \|\Delta_v\mathbf{y}_i(0)\|}{\|\Delta_v\mathbf{y}_i(0)\|}\right], \quad (8)$$

where $\Delta u, \Delta v$ are chosen to be roughly 2mm. See Fig. 6.a for a visualization of the computed stretch values for a given expression. The stretch measurements of the individual samples are stacked to form the input feature vector $\mathcal{S}(t)$. Note that unlike the removal network, for synthesis we predict all displacements jointly, since

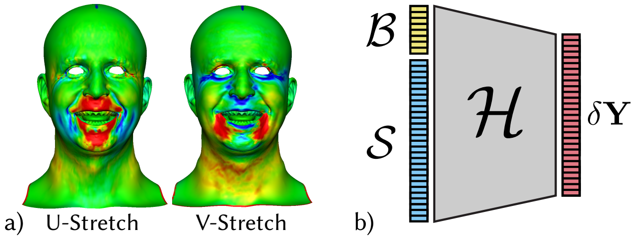

Fig. 6. **Synthesis Overview:** Given stretch $\mathcal{S}$ as local surface feature (a) and a kinematic history of the skull $\mathcal{B}$, the neural network $\mathcal{H}$ predicts the displacement vectors $\delta\mathbf{Y}$ for all sample points on the face jointly (b).

local stretch is ambiguous, but when considered globally it gives a good description of the current expression.

6.2 Training and inference

With these features, we train a feed-forward fully connected neural network with 2 hidden layers of size 256 and ReLu activation functions to learn the desired mapping $\mathcal{H}$. As for the removal network presented in Section 5, these hyper-parameters were selected empirically. The training data required, namely a quasistatic performance with corresponding secondary dynamics displacement vectors is generated from the corpus of data acquired in Section 4 using the removal network described in Section 5. The network is trained with L2 loss using the Adam optimizer [Kingma and Ba 2014], with a batch size of 50. With a learning rate of $1e$-3 training converges after 30 epochs. The network was implemented in PyTorch [Paszke et al. 2019].

6.3 Validation of Dynamics Synthesis

To validate the proposed synthesis approach we synthesize secondary dynamics for the quasistatic performance generated by removing the original secondary dynamics from a captured input performance, not used for training either of the networks. We can then compute the residual between the original performance and the resynthesized performance, as reported in Fig. 7.

7 RESULTS

In this section we first provide a detailed breakdown of the acquired data in Section 7.1. We then evaluate our entire pipeline by removing and re-synthesizing secondary dynamics on a performance (Section 7.2). Lastly, we evaluate motion retargeting in Section 7.3.

7.1 Data Acquisition

We captured three different actors following the procedure described in Section 4. The output of our data acquisition is more than seven thousand frames per sustained expression and more than twenty thousand frames of dialog performance, distributed 20% for the first actor and 40% for the other two (the discrepancy was due to actor availability). To properly capture the skin dynamic effects, the data was acquired at high frame-rates (between 96 and 128 fps, depending on the motion). Each sequence (held expression or dialog performance) was captured multiple times as follows: under no actuation, walking, running, jumping, and additionally under 6 oscillatory speeds of the vibrating platform. In order to reconstruct this data efficiently, we employ a recently introduced GPU solver designed for facial performance capture [Fratarcangeli et al. 2020], reducing reconstruction time by one order of magnitude. We encourage the reader to refer to the supplemental video for a more detailed view of the captured data in motion. Overall, our dataset consists of almost one hundred thousand frames with consistent mesh topology and tracked rigid skull motion.

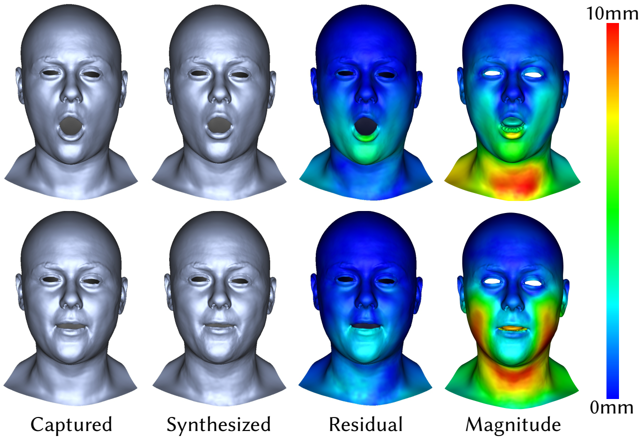

Fig. 7. **Synthesis Validation:** Performances with original (left column) and synthesized (second column) secondary dynamics deformation. Synthesis results coincide well with the original deformations (third column) despite the large deformation induced by secondary dynamics (fourth column).

We train the removal network $\mathcal{F}$ using strictly the sequences with held expressions as no quasistatic equivalent is available for the dialog performance sequences. When training the synthesis network $\mathcal{H}$, we also include some dialog performance data to help the network learn to interpolate between expressions. To generate the surface stretch feature of those sequences, we first use the removal network to generate a quasistatic frame and then compute the stretch feature vector $\mathcal{S}(t)$ with respect to the neutral face. We refer the reader to the supplemental video to appreciate the interpolation capabilities of both networks.

7.2 Complete Pipeline for Removal and Synthesis

Here we demonstrate the entire pipeline. Given a captured performance that contains undesired secondary dynamic effects (Fig. 8.a), we produce a quasistatic animation free from secondary dynamics (Fig. 8.b) using the proposed removal network $\mathcal{F}$ (Section 5). We then change the motion of the head simulating two different motion patterns for the subject, and add back appropriate synthesized secondary dynamics (Fig. 8.c and Fig. 8.d) using the proposed synthesis network $\mathcal{H}$ (Section 6). Note that obtaining similar results through physical simulation of skin dynamics would be extremely challenging, even with complete anatomical muscle models, as it can be difficult to obtain a suitable material model, deal with the bone collisions, and support the same high resolution that we can achieve with our data-driven method.

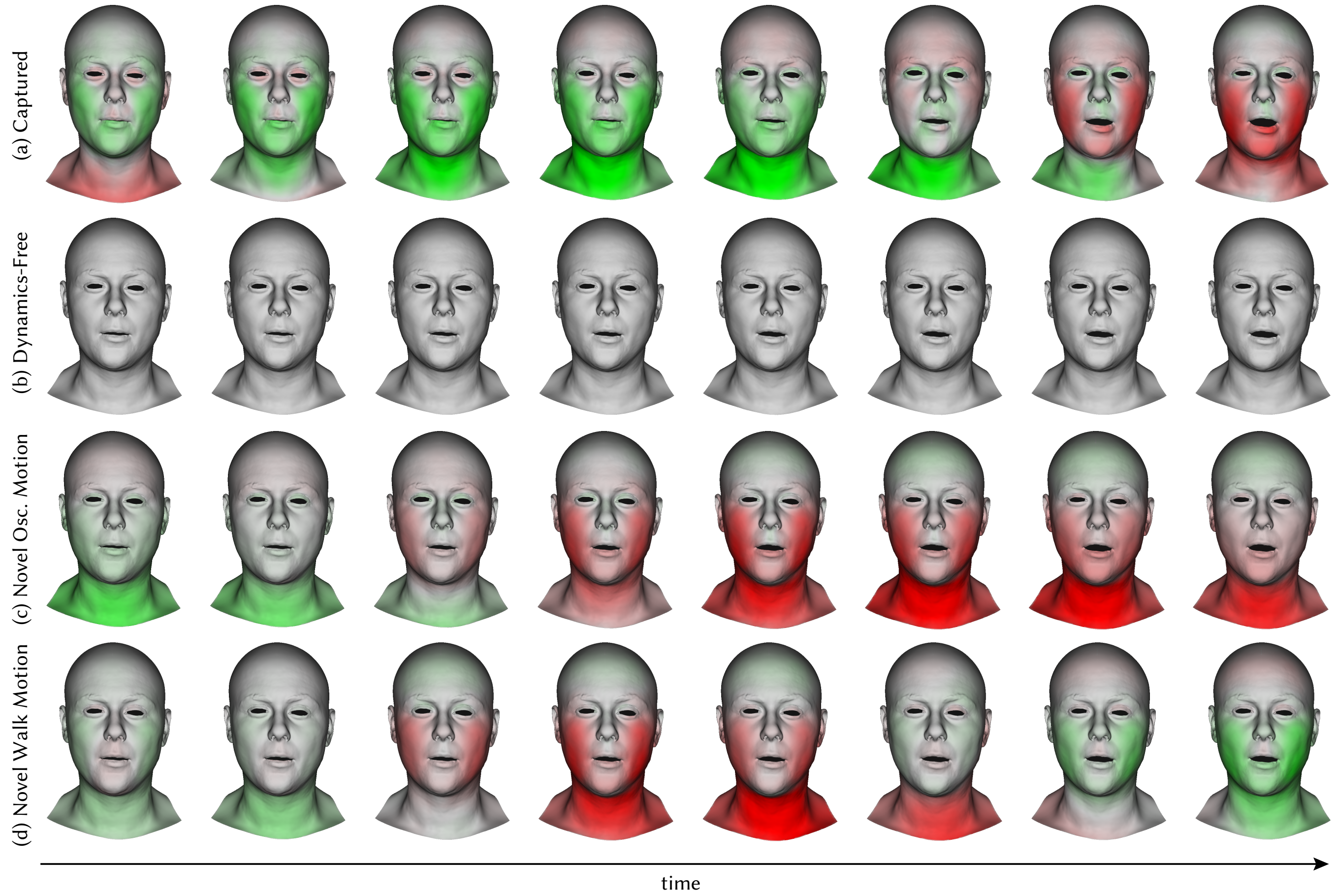

Fig. 8. **Entire Removal and Synthesis Pipeline:** The secondary dynamics present in 8 consecutive frames of the original capture (a) is removed to produce a quasistatic animation free from secondary dynamics (b). Using different input motions, novel secondary dynamics may be synthesized (an oscillating up-down motion (c) and a walk cycle (d)). The coloring red/green encodes the signed difference with respect to the quasistatic performance, with a scale ranging from +/- 5mm for (a), +/- 3mm for (c), and +/- 2mm for (d). Here you can clearly see the change in period of the oscillating dynamics between rows (a) and (c), and the peaky dynamics induced by the walk cycle in row (d).

7.3 Retargeting

Fig. 9 shows root motion retargeting, allowing to combine the facial performance of an actor with the body performance of a stunt double, for example. This allows to mitigate the disconnect of facial and body performances, which oftentimes leads to uncanny valley effects. Due to the dynamic nature of the investigated problem, results are best appreciated in the accompanying video, but in this figure we depict a walking motion captured from Subject 1 and the corresponding root motion retargeted to Subject 2, followed by applying the dynamics synthesis trained on that subject. In order to validate that simply transferring the vertex delta motion between subjects is insufficient for retargeting, we illustrate another example of retargeting in Fig. 10, this time using data from the vibrating platform. The dynamics (Fig. 10.a) naïvely transferred as per-vertex deltas (Fig. 10.b) retain too many details from the source subject and appear uncanny on the target face (especially in motion, as illustrated in the supplemental video). Using our method, the synthesized dynamics given only the root motion of the source actor (Fig. 10.c) better match the target character's face, as illustrated with a captured frame from a sequence on the same vibrating platform frequency for reference (Fig. 10.d).

8 DISCUSSION

In this work we investigate the challenge of modeling secondary dynamic effects in performance-driven facial animation. Taking a data-driven approach, we propose new ideas for the prediction of deformation caused by dynamics, modeled as the difference between an input performance that contains secondary effects and its quasistatic counterpart that exhibits no extraneous dynamic motion. We propose a deep learning based framework to predict and remove the secondary dynamics based on a short kinematic history of skull and skin. Using the dynamics-free results, we train a second network to predict secondary dynamics based on a short kinematic history of

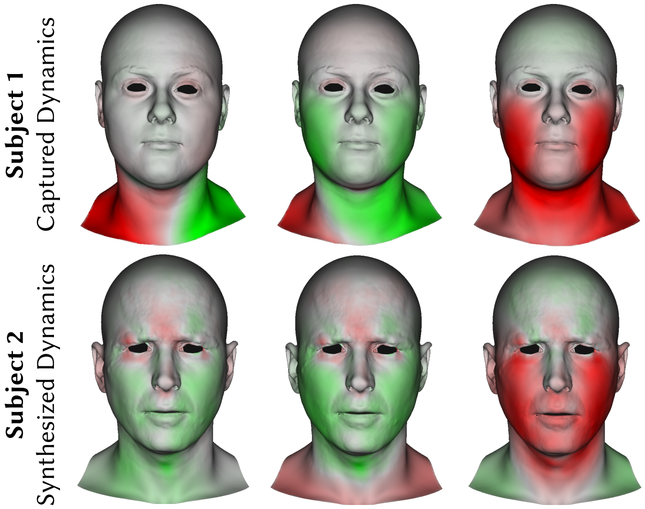

Fig. 9. **Root Motion Retargeting:** Walking motion captured from one subject (top row) can be transferred to a second subject (bottom row), which is for example useful to combine the facial performance of an actor with the body performance of a stunt double. The coloring encodes the signed difference with respect to the quasistatic performance, with a scale ranging from +/- 2mm. Note, the neck deformation is not transferred as this is not the aim of this work.

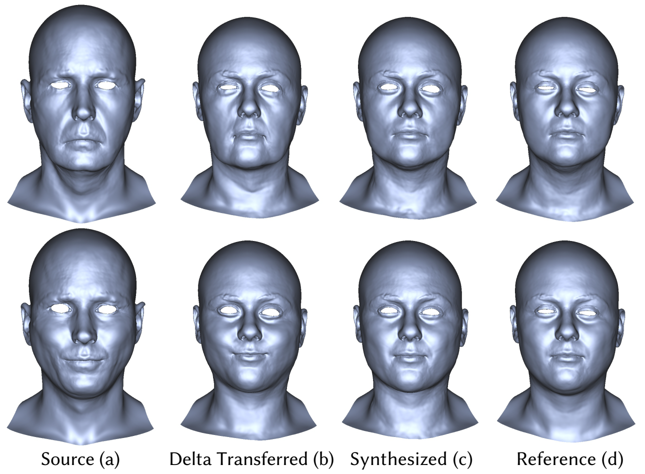

Fig. 10. **Retargeting Validation:** Naïvely transfering per-vertex delta motion as a means of retargeting dynamics from one subject (a) to another (b) leads to uncanny performances that too closely resemble the source actor. Our method more accurately synthesizes the dynamics (c) given only the root motion of the source, as compared to a reference performance captured under the same vibrating platform frequency (d). Top and bottom rows correspond to two extremal frames of an oscillatory sequence.

the skull and the current expression. Combined, the two networks allow to both remove parasitic dynamic motion from captured data, as well as synthesize new dynamic motion to be composed onto dynamics-free facial animation. As a result, the hypotheses and validations presented in this work have great potential to impact the widely-employed domain of performance-driven facial animation.

8.1 Limitations and future work

In order to practically validate our algorithms, we have made several simplifying assumptions. We have so far only investigate the case of training person specific removal and synthesis networks. The data required to train these, however, is significant and hence it would be extremely valuable to train person independent predictors. While we have not thoroughly looked at the generalization capabilities of the proposed architectures, we do believe that given sufficient training data from multiple subjects the approach can generalize. To test this hypothesis we applied the removal network trained on one subject to a different subject and the results are very promising as shown on Fig. 11 and in the accompanying video. Generalization of the synthesis network would require a disentanglement of the skin stretching feature and the subject's skin properties. Our current implementation will only generate dynamics similar to the subject it was trained on but we believe that including additional features such as BMI, age or other characteristics influencing the physical properties of the skin might allow generalization of the mapping given a sufficient amount of training data.

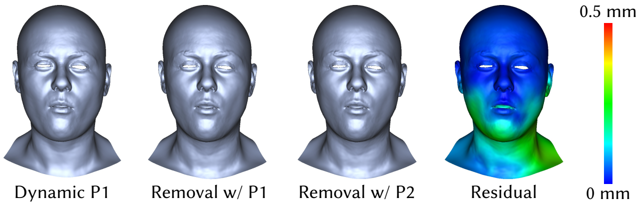

Fig. 11. **Removal Generalization:** Performance of subject P1 with secondary dynamics induced by the vibrating platform. Comparing a quasistatic frame produced using the mapping trained on data of the same subject and removal using a second mapping, trained exclusively on data from another actor. While the removal network trained on a different subject performs really well, some residual dynamics is still present. We refer the reader to the accompanying video for the full sequence and other examples of cross-actor generalization of the removal network.

Furthermore, while the captured data considers a lot of everyday motions, such as walking, running or jumping, extending the data capture to include also horizontal or angular motion patterns would be beneficial. Doing so, however, is rather challenging. An alternative to acquisition might be to augment the existing data with physical simulation, leveraging accurate anatomical models such as [Sifakis et al. 2006] to generate training data.

Our approach so far considers only dynamics that are induced by the moving skeleton. External forces, such as contact or even gravity, have not yet been considered. Measuring and quantifying these external actuators might be extremely challenging, and synthetic data augmentation might again be the route to go.

Finally, since we only have supervised data from the static expressions to train the removal network, predicting the secondary dynamics from performances where expressions change is likely suboptimal. A solution to this problem would be to concatenate the two networks introducing cycle consistency, which would allow to also train on the performances end-to-end without requiring supervision.

Despite these limitations, we believe this work has provided important insight into the modeling of secondary dynamics for facial performance capture, and represents, to our knowledge, the first investigation into this challenging problem.

ACKNOWLEDGMENTS

We wish to thank Andreea Rădoescu for posing as a capture subject. E.S. was supported in part by NSF grants IIS-1763638, CCF-1812944.

REFERENCES

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: Shape Completion and Animation of People. *ACM Trans. Graph.* 24, 3 (July 2005), 408–416. https://doi.org/10.1145/1073204.1073207

Vincent Barrielle, Nicolas Stoiber, and Cédric Cagniart. 2016. BlendForces: A dynamic framework for facial animation. *Computer Graphics Forum* 35, 2 (2016), 341–352. https://doi.org/10.1111/cgf.12836

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

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

Steve Capell, Matthew Burkhart, Brian Curless, Tom Duchamp, and Zoran Popović. 2007. Physically based rigging for deformable characters. *Graphical Models (Proceedings of the 2005 ACM SIGGRAPH/Eurographics Symposium on Computer Animation)* 69, 1 (2007), 71–87. https://doi.org/10.1016/j.gmod.2006.09.001

Matthew Cong, Kiran S. Bhat, and Ronald Fedkiw. 2016. Art-directed Muscle Simulation for High-end Facial Animation. (2016), 119–127. http://dl.acm.org/citation.cfm?id=2982818.2982835

Edilson de Aguiar, Leonid Sigal, Adrien Treuille, and Jessica K. Hodgins. 2010. Stable spaces for real-time clothing. *ACM SIGGRAPH 2010 papers on - SIGGRAPH '10* 1, 212 (2010), 1. https://doi.org/10.1145/1833349.1778843

Fernando De Goes and Doug L. James. 2017. Regularized Kelvinlets: Sculpting Brushes Based on Fundamental Solutions of Elasticity. *ACM Trans. Graph.* 36, 4, Article 40 (July 2017), 11 pages. https://doi.org/10.1145/3072959.3073595

Marco Fratarcangeli, Derek Bradley, Aurel Gruber, Gaspard Zoss, and Thabo Beeler. 2020. Fast Nonlinear Least Squares Optimization of Large-Scale Semi-Sparse Problems. *Computer Graphics Forum (Proc. Eurographics), to appear* (2020).

Graham Fyffe, Tim Hawkins, Chris Watts, Wan-Chun Ma, and Paul E. Debevec. 2011. Comprehensive Facial Performance Capture. *Comput. Graph. Forum* 30, 2, 425–434. https://doi.org/10.1111/j.1467-8659.2011.01888.x

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 (May 2017), 295–309. https://doi.org/10.1111/cgf.13127

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. https://doi.org/10.1145/2508363.2508380

Fabian Hahn, Sebastian Martin, Bernhard Thomaszewski, Robert Sumner, Stelian Coros, and Markus Gross. 2012. Rig-space Physics. *ACM Trans. Graph.* 31, 4, Article 72 (July 2012), 8 pages. https://doi.org/10.1145/2185520.2185568

Fabian Hahn, Bernhard Thomaszewski, Stelian Coros, Robert W. Sumner, and Markus Gross. 2013. Efficient simulation of secondary motion in rig-space. *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation - SCA '13* (2013), 165. https://doi.org/10.1145/2485895.2485918

Alexandru-Eugen Ichim, Petr Kadleček, Ladislav Kavan, and Mark Pauly. 2017. Phace: Physics-based Face Modeling and Animation. *ACM Trans. Graph.* 36, 4, Article 153 (July 2017), 14 pages. https://doi.org/10.1145/3072959.3073664

Alexandru-Eugen Ichim, Ladislav Kavan, Merlin Nimier-David, and Mark Pauly. 2016. Building and Animating User-specific Volumetric Face Rigs. (2016), 107–117. http://dl.acm.org/citation.cfm?id=2982818.2982834

Meekyoung Kim, Gerard Pons-Moll, Sergi Pujades, Seungbae Bang, Jinwook Kim, Michael J. Black, and Sung-Hee Lee. 2017. Data-driven Physics for Human Soft Tissue Animation. *ACM Trans. Graph.* 36, 4, Article 54 (July 2017), 12 pages. https://doi.org/10.1145/3072959.3073685

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

Yeara Kozlov, Derek Bradley, Moritz Bächer, Bernhard Thomaszewski, Thabo Beeler, and Markus Gross. 2017. Enriching Facial Blendshape Rigs with Physical Simulation. *Computer Graphics Forum (Proc. Eurographics)* 36, 2 (2017).

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 *Proceedings of the ACM SIGGRAPH / Eurographics Symposium on Computer Animation.* ACM Press, Los Angeles, CA, 1–10. https://doi.org/10.1145/3099564.3099581

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-person Linear Model. *ACM Trans. Graph.* 34, 6, Article 248 (Oct. 2015), 16 pages. https://doi.org/10.1145/2816795.2818013

Sang Il Park and Jessica K. Hodgins. 2006. Capturing and Animating Skin Deformation in Human Motion. (2006), 881–889. https://doi.org/10.1145/1179352.1141970

Sang Il Park and Jessica K. Hodgins. 2008. Data-driven Modeling of Skin and Muscle Deformation. , Article 96 (2008), 6 pages. https://doi.org/10.1145/1399504.1360695

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. 2015. *Dyna: A Model of Dynamic Human Shape in Motion.* Vol. 34. 120:1–120:14 pages.

Igor Santesteban, Elena Garces, Miguel A. Otaduy, and Dan Casas. 2020. SoftSMPL: Data-driven Modeling of Nonlinear Soft-tissue Dynamics for Parametric Humans. *Computer Graphics Forum (Proc. Eurographics)* (2020).

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

Eftychios Sifakis, Andrew Selle, Avram Robinson-Mosher, and Ronald Fedkiw. 2006. Simulating Speech with a Physics-Based Facial Muscle Model. *Eurographics/ ACM SIGGRAPH Symposium on Computer Animation (2006)* (2006).

O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel. 2004. Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing.* 175–184.

Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M. Seitz. 2014. Total Moving Face Reconstruction. In *ECCV (4) (Lecture Notes in Computer Science)*, Vol. 8692. Springer, 796–812.

Ayush Tewari, Michael Zollöfer, 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*.

Chenglei Wu, Derek Bradley, Markus Gross, and Thabo Beeler. 2016. An Anatomically-constrained Local Deformation Model for Monocular Face Capture. *ACM Trans. Graph.* 35, 4, Article 115 (July 2016), 12 pages. https://doi.org/10.1145/2897824.2925882

Hongyi Xu and Jernej Barbič. 2016. Pose-space subspace dynamics. *ACM Transactions on Graphics* 35, 4 (2016), 1–14. https://doi.org/10.1145/2897824.2925916

Hongyi Xu and Jernej Barbič. 2017. Example-based Damping Design. *ACM Trans. Graph.* 36, 4, Article 53 (July 2017), 14 pages. https://doi.org/10.1145/3072959.3073631