UNIVERSITY OF CALIFORNIA

Los Angeles

Neuromuscular Animation and FACS-Based Expression Transfer Via Deep Learning

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Computer Science

by

Surya Dwarakanath

2021

ABSTRACT OF THE THESIS

Neuromuscular Animation and FACS-Based
Expression Transfer Via Deep Learning

by

Surya Dwarakanath
Master of Science in Computer Science
University of California, Los Angeles, 2021
Professor Demetri Terzopoulos, Chair

The transfer of facial expressions from people to 3D face models is a classic computer graphics problem. In this paper, we present a novel, learning-based approach to transferring facial expressions and head movements from images and videos to a biomechanical model of the face-head-neck musculoskeletal complex. Specifically, leveraging the Facial Action Coding System (FACS) as an intermediate representation of the expression space, we train a deep neural network to take in FACS Action Units (AUs) and output suitable facial muscle and jaw activations for the biomechanical model. Through biomechanical simulation, the activations deform the face, thereby transferring the expression to the model. Our approach has advantages over previous approaches. First, the facial expressions are anatomically consistent as our biomechanical model emulates the relevant anatomy of the head, neck, and face. Second, by training the neural network using data generated from the biomechanical model itself, we eliminate the manual effort of data collection for expression transfer. The success of our approach is demonstrated through experiments involving the transfer of a range of expressive facial images and videos onto our biomechanical face-head-neck model.

The thesis of Surya Dwarakanath is approved.

Kai-Wei Chang

Song-Chun Zhu

Demetri Terzopoulos, Committee Chair

University of California, Los Angeles

2021

To my family . . .

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Professor Demetri Terzopoulos for the opportunity to work on this project in the UCLA Computer Graphics & Vision Laboratory. I am also thankful to him for his support, encouragement, and ideas throughout the duration of this project. I thank Professors Song Chun-Zhu and Kai-Wei Chang for their feedback, which guided me on improving this thesis. I would also like to thank my amazing colleagues, Xiao Zeng, Wuyue Lu, and Masaki Nakada, who worked with me on this project. Their hard work contributed greatly to the success of the project.

I want to express my gratitude to the faculty and students of the UCLA Computer Science Department from whom I have learnt a lot and improved as a graduate student. Additionally, I would to thank my parents, T. A. Dwarakanath and Anupama Dwarakanath, and my brother, Harsha Dwarakanath, for their constant support during these trying times. Finally, I would also like to thank a dear friend, R. S. Nikhil Krishna, for introducing me to the wonderful world of Computer Science.

CHAPTER 1

Introduction

Biomechanical human musculoskeletal models aim to realistically capture the anatomy and physics underlying human motion generation. Much of the research in this domain of computer animation has focused on modeling the body (e.g., (Lee et al., 2009), (Si et al., 2014), (Rajagopal et al., 2016), (Lee et al., 2019)). Only a few such models have focused on simulating the human face (e.g., (Lee et al., 1995),(Sagar et al., 2014)). These have succeeded in realistically emulating facial expression generation; however, they require significant effort in parameter setting and tuning to produce realistic results.

In this thesis, we show how to endow a biomechanical, musculoskeletal model of the human face with the ability to perform facial expressions by machine learning from real-world reference images and videos (Figure 1.1). To this end, we introduce a deep-neural-network-based method for learning the representation of human facial expressions through Ekman's Facial Action Coding System (FACS) (Cohn et al., 2007) in the context of the muscle actuators that drive the musculoskeletal face model. We furthermore augment the face animation system with a musculoskeletal cervicocephalic (neck-head) system to animate head movement during facial expression synthesis.

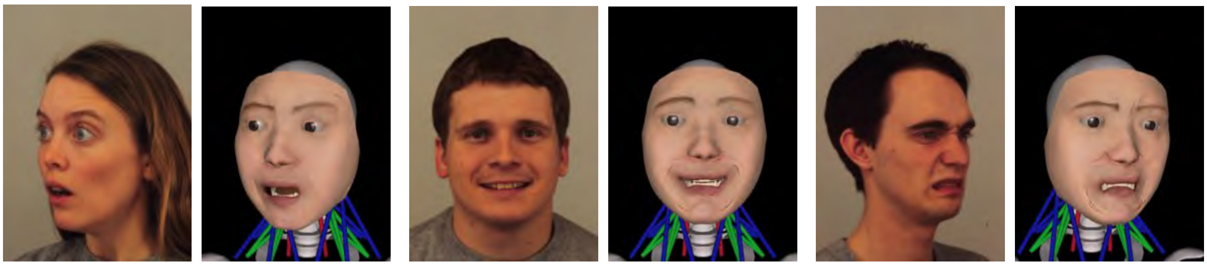


Figure 1.1: Our deep learning framework transfers facial expression and head pose from a real human image or video to a biomechanical face-head-neck model.

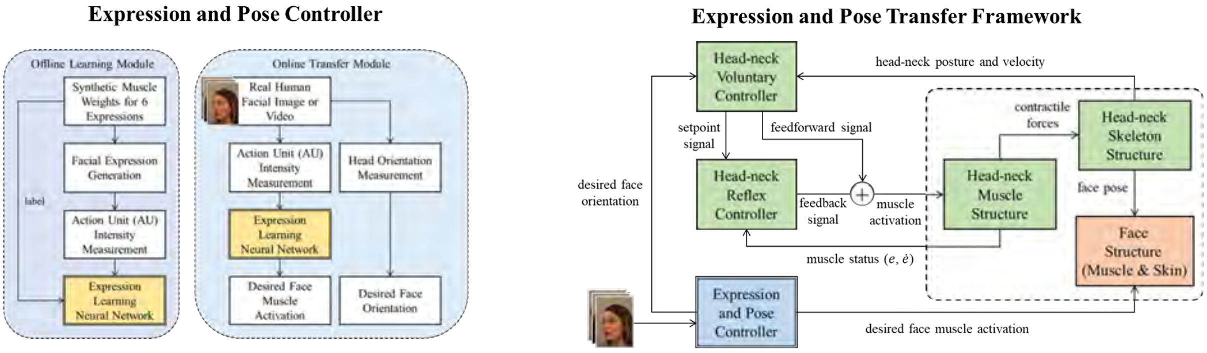

Expression and Pose Controller
Offline Learning Module
Synthetic Muscle Weights for 6 Expressions
Facial Expression Generation
Action Unit (AU) Intensity Measurement
Expression Learning Neural Network
label
Online Transfer Module
Real Human Facial Image or Video
Action Unit (AU) Intensity Measurement
Head Orientation Measurement
Expression Learning Neural Network
Desired Face Muscle Activation
Desired Face Orientation
Expression and Pose Transfer Framework
Head-neck Voluntary Controller
head-neck posture and velocity
setpoint signal
feedforward signal
contractile forces
Head-neck Skeleton Structure
desired face orientation
Head-neck Reflex Controller
feedback signal
muscle activation
Head-neck Muscle Structure
face pose
Face Structure (Muscle & Skin)
muscle status (e, $\dot{e}$)
Expression and Pose Controller
desired face muscle activation

Figure 1.2: Components of the expression and head pose controller (left) and the overall structure of our expression and head pose transfer framework (right). The expression learning neural network (yellow) is first trained offline. In a transfer task, an image or a video sequence of a real face is fed into the online transfer module of the expression and pose controller to output the desired facial muscle activations and head orientation information. The obtained muscle activations are input to the biomechanical face model (orange) to perform the corresponding expression. The head orientation is provided to the head-neck biomechanical neuromuscular system (green) to produce the desired head pose.

1.1 Contributions

The novelty of our framework lies in the following features:

- We propose the first biomechanical face-head-neck animation system that is capable of learning to reproduce expressions and head orientations through neuromuscular control.
- Our novel deep neuromuscular motor controller learns to map between FACS Action Units (AUs) extracted from human facial images and videos and the activations of the muscle actuators that drive the biomechanical system.
- As a proof of concept, we demonstrate an automated processing pipeline (Figure 1.2) for animating expressions and head poses using an improved version of the physics-based face-head-neck animation system developed by Lee and Terzopoulos (2006), but which can potentially be applied to any physics-based, muscle-driven model.

1.2 Thesis Overview

The remainder of this thesis is organized as follows:

Chapter 2 reviews the relevant background literature and related work on biomechanical musculoskeletal modeling in computer graphics, expression representation and transfer, and muscle-based facial animation.

Chapter 3 and Appendix A present the details of our muscle-actuated biomechanical model of the face and its associated control mechanisms.

Chapter 4 develops our novel expression transfer pipeline, including its network architecture, training data generation, network training, and expression transfer components.

Chapter 5 presents our experiments and discusses the results involving facial expression and head orientation transfer from both still images and video sequences.

Chapter 6 presents our conclusions and suggestions for future work.

CHAPTER 2

Background and Related Work

2.1 Related Work

The face actuated by the muscles of facial expression, and head movements actuated by the cervical muscles, are a powerful mode of nonverbal communication between humans. The simulation of the face-head-neck musculoskeletal complex is of importance in understanding how we convey thinking and feeling in fields from affective science to 3D computer animation. Extensive research has been directed at reconstructing the geometric details of the face (Richardson et al., 2017; Moschoglou et al., 2020) and synthesizing expressive facial shapes (Chen et al., 2019; Wang et al., 2020). Despite the high-quality 3D meshes and plausible facial expressions reproduced by existing approaches, the causal relationship between the musculoskeletal system and facial deformation has often been ignored. Methods for transferring expressive movements from humans to stylized characters (Aneja et al., 2016) or avatars (Zhang et al., 2020) are a different avenue towards studying facial expressions that has also been successful in recent years.

2.1.1 Biomechanical Musculoskeletal Models

Musculoskeletal systems model the human body and synthesize body movements by simulating the biomechanics of hard and soft tissues. They are gradually gaining popularity in computer graphics since pioneering efforts in modeling the dynamics of bones (Hodgins et al., 1995; Faloutsos et al., 2001) and muscles (Chen and Zeltzer, 1992). Sophisticated full-body musculoskeletal models have emerged (Lee et al., 2009; Van den Bogert et al., 2013; Si et al., 2014; Nakada et al., 2018; Lee et al., 2019), which can perform complex motor tasks such as swimming, walking, and sensorimotor control. Although these systems achieve realistic movements with high anatomical accuracy, they often require manual parameter tuning and lots of domain-specific knowledge due to their remarkable complexity. Other efforts have focused on developing biomechanical models of specific body parts, such as the upper extremities (Sueda et al., 2008; Sachdeva et al., 2015), the lower extremities (Rajagopal et al., 2016; Cardona and Cena, 2019), and the head-neck complex (Lee and Terzopoulos, 2006). The latter, which is most relevant to our work in the present paper, is driven by Hill-type muscle actuators. Its neuromuscular motor controller was trained, using data synthesized by the biomechanical model itself, to synthesize natural head movements. Moreover, it incorporated the biomechanical face model of Lee et al. (1995) for facial expression synthesis.

2.1.2 Expression Representation and Transfer

Numerous facial animation researchers have made use of the well-known FACS (Ekman, 2002), which is a quantitative phenomenological abstraction of facial muscle action developed by Ekman and Friesen (1978). It decomposes facial expressions into the intensity levels of Action Units (AUs), each of which corresponds to the actions of one or more facial muscles. An advantage of using the FACS is that it encodes anatomical and psychological semantics (Cohn et al., 2007). It has been combined with blendshape interpolation to create facial animation for 3D avatars (Alkawaz et al., 2015) and has inspired the development of a FACS-based blendshape database (Cao et al., 2013). Chen et al. (2019) integrated FACS with a 3D morphable model (Blanz and Vetter, 1999) to estimate expressions with facial semantic information in facial geometry synthesis.

Expression transfer or retargeting is a trending topic and recent approaches often use FACS-based blendshapes as the basic parametric representation (Weise et al., 2011; Thies et al., 2016; Zhang et al., 2020). Other works used techniques such as interactive mesh deformation control (Xu et al., 2014) and deep neural network based perceptual models (Aneja et al., 2016) to represent expressions in blendshapes. However, transferring expressions using a musculoskeletal system is more natural since facial actions are the result of muscle contraction and tissue deformation. A musculoskeletal approach also enables the automatic generation of motions under the influences of external forces (Sifakis et al., 2005). Although Ishikawa et al. (2000) introduced an estimation strategy that approximates facial muscle parameters from 2D images for muscle-based face models, there remains a deficiency of work in transferring expressions to musculoskeletal systems.

2.1.3 Muscle-Based Facial Animation

Muscle-based facial animation has been studied for decades (Platt and Badler, 1981; Waters, 1987). A series of studies have endeavored to build anatomically accurate musculoskeletal face models. Terzopoulos and Waters (1990) proposed a 3D model of human face that contains a physical approximation of facial tissues and anatomically-motivated muscle actuators. Lee et al. (1995) augmented and improved the biomechanical components of this model and incorporated an expression control system. Subsequent work further developed the model by increasing the number of embedded muscles, adding an interactive behavior model, and distributing simulation of multiple face instances (Terzopoulos and Lee, 2004). Terzopoulos and Waters (1993) and, with a more detailed biomechanical model of the face, Sifakis et al. (2005) studied the automatic determination of facial muscle activations from videos, but these methods require nonrigid motion trackers or motion capture markers. Sagar et al. (2014) introduced another realistic biomechanical face that deforms based on finite element elasticity as a component of their psychobiological system simulating an infant, but its use of a piecewise linear expression manifold of muscle actions in expression synthesis is restrictive.

To address these shortcomings, we present a markerless, real-time biomechanical face-head-neck simulation system that can automatically learn muscle activation parameters for a variety of expressions from reference images and videos.

CHAPTER 3

Biomechanical Model

3.1 Musculoskeletal Model

Our real-time musculoskeletal model is based on that of Lee and Terzopoulos (2006), but both the underlying face-head-neck control system and the facial expression system are significantly improved. Its overall architecture is controlled in a hierarchical manner as illustrated in Figure 1.2. Specifically, the skeletal structure is an articulated multibody dynamics system, with bones and joints consistent with human anatomy. The skeletal model is driven by a Hill-type muscle actuator model.

The biomechanical face component consists of epidermis, dermal-fatty, fascia and muscle layers together with a skull beneath them. The first three of the aforementioned layers comprise the synthetic skin model, which is constructed based on the work by Lee et al. (1995). There are 26 pairs of primary facial muscles embedded in the biomechanical face, including frontalis, corrugator, levator labii, orbicularis oculi, mentalis, and orbicularis oris groups. The contractions of these muscles apply forces to the facial tissue layers, which deform to produce meaningful facial expressions. We significantly augment the expressive details such as wrinkles on the face model by applying multiple levels of subdivision to increase the number of facial nodes that can be activated by muscle forces. We also adapt a high resolution texture image to our generic face mesh to provide a natural look.

Our musculoskeletal head-neck model is based on that of Lee and Terzopoulos (2006). Its overall architecture is controlled in a hierarchical manner as illustrated in Figure 1.2. Specifically, the skeletal structure is an articulated multibody dynamics system, with bones and joints consistent with human anatomy. The skeletal model is driven by a

Hill-type muscle actuators.

Appendix A explains the biomechanical components of our model in greater detail.

3.2 Control

To control the face-head-neck system, our novel neural network-based expression and pose controller generates facial muscle activations that produce recognizable expressions. It simultaneously outputs head pose estimates to the head-neck complex, where voluntary and reflex neuromuscular control layers generate muscle activation signals to achieve the desired head orientations.

The higher-level voluntary controller receives the current head-neck posture and velocity information, as well as the desired adjustment of the posture, then generates a *feedforward muscle activation signal* and a *setpoint control signal*, which will be passed to the lower-level reflex controller. The setpoint control signal contains the desired muscle strains and strain rates. The reflex controller takes in these desired muscle status, and generates a *feedback muscle activation signal* and adds them to the feedforward signal generated by the voluntary controller. As a result, each muscle receives an activation level signal a and generates a contractile muscle force f_C accordingly. Together with the external forces and gravity from the environment, the whole system is simulated through the time to generate a physics-based animation. The voluntary controller runs at 25Hz (in simulation time), while the reflex controller runs at 4KHz along with the physical simulation steps.

CHAPTER 4

Expression Transfer

We next explain our machine learning approach of using a deep neural network to transfer facial expressions to our biomechanical face model. We leverage the FACS and synthesize the muscle activations for the model using the trained deep neural network. The following sections describe the architecture of the network, the pipeline for using the biomechanical face model to generate training data, and the process of training the neural network.

4.1 Network Architecture

The function of the neural network is to generate activation signals for the facial muscles in the biomechanical face model and thus produce contractile forces that deform the synthetic facial tissues to produce facial expressions. We use a fully connected deep neural network architecture for our purpose (Figure 4.1). The input layer consists of a total of 17 neurons representing the important AUs that are involved in the majority of facial movements, with each neuron corresponding to a single normalized AU. We employ 4 hidden layers, with 100 neurons in each layer. The output layer consists of 56 neurons. The first 52 neurons encode the activations a_i, with $1 \leq i \leq 52$, for each of the 26 pairs of facial muscles. The last 4 neurons encode the jaw rotation, jaw slide, jaw twist, and an auxiliary value. Given its architecture, the network has a total of 37,300 weights. It is implemented in Keras with a TensorFlow backend. Each hidden layer employs ReLU activation.

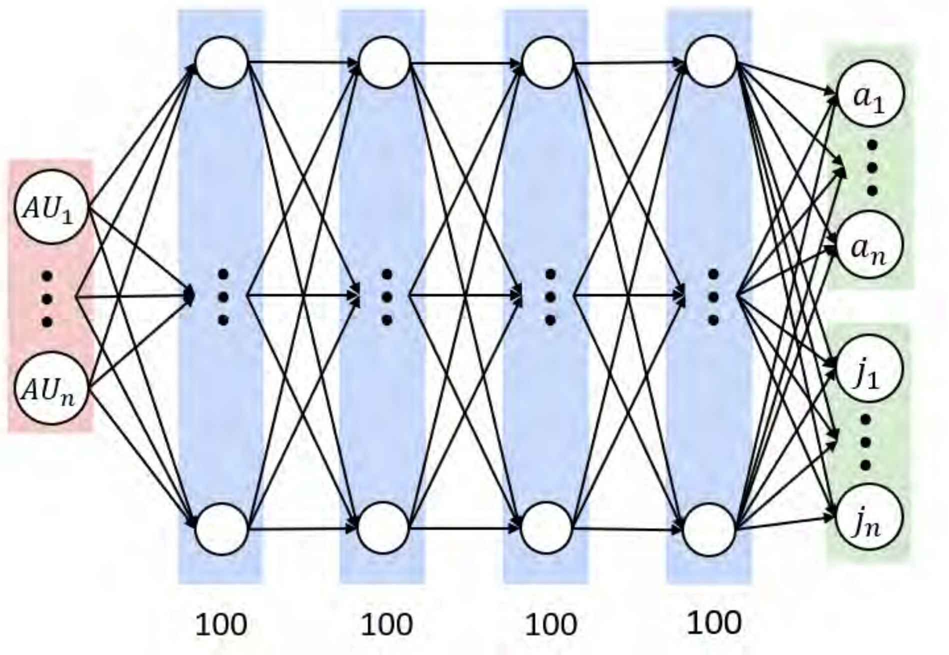


Figure 4.1: Neural Network Architecture

4.2 Training Data Generation

The training data generation process is divided into two steps:

1. generation of muscle activations, and
2. generation of AUs for the corresponding muscle activations.

Each basic expression requires a combination of muscles to be activated. Given n muscles, we define W_e as a set of weights w_i, with $1 \leq i \leq n$, which determine the effect each muscle will have on an expression e. The activation for each muscle a_i for an expression is then defined as $a_i = w_i s$, where $w_i \in W_e$ and s denotes a scale term. For a single expression, we determine the weights in a set W_e manually by visually analyzing the facial expressions formed by the face model. For each set, we also add variations by changing the non-zero weights by a small amount. We repeat the process for all the basic expressions, namely (1) Joy, (2) Sadness, (3) Anger, (4) Fear, (5) Disgust, and (6) Surprise to generate the sets W_{Joy}, W_{Sadness}, W_{Anger}, W_{Fear}, W_{Disgust} and W_{Surprise}, respectively.

We can then sample the value of s randomly in the range $[0.0, 1.0]$ to generate all the muscle activations a_i. We also assign a random value between 0 and 1 for the jaw rotation. For the purpose of our experiments, we maintain the jaw twist and jaw slide at a value of 0.5. We further form a set A consisting of all the muscle activations a_i, where $1 \leq i \leq n$, and the jaw activations including jaw rotation, jaw twist and jaw slide. We iterate the aforementioned process, generating the set A by repeatedly sampling the value of s for a single expression. We finally extend this to all basic expressions and obtain multiple sets A_{Joy}, A_{Sadness}, A_{Anger}, A_{Fear}, A_{Disgust} and A_{Surprise} for each expression.

We leverage the functionality of OpenFace 2.0 (Baltrusaitis et al., 2018) for facial expression recognition and head pose estimation. OpenFace is a computer vision tool capable of facial landmark detection, head pose estimation, facial AU recognition, and eye-gaze estimation. OpenFace employs Convolutional Experts Constrained Local Models (CE-CLMs) for facial landmark detection and tracking. CE-CLMs use a 3D representation of the detected facial landmarks by which OpenFace estimates the head pose. Eye gaze estimation in OpenFace is done using a Constrained Local Neural Fields (CLNF) landmark detector. The final task, facial expression recognition is performed using Support Vector Machines and Support Vector Regression.

We use a single set A formed by the above described procedure to activate the muscles and jaw of the biomechanical face model. We then render the model to form an image. The image is input to OpenFace, which performs facial expression recognition and head pose estimation, outputting the AUs and head orientation associated with the input image. We repeat this process for each set A formed (as described previously) to get the corresponding AUs and head orientations.

Thus, we synthesize a large quantity of training data pairs each consisting of (i) muscle and jaw activations A and (ii) the associated AUs and head orientations.

dataset	MSE_{AU1}	MSE_{AU4}	MSE_{AU6}	MSE_{AU9}	MSE_{AU12}	MSE_{AU15}	MSE_{AU20}	MSE_{AU25}	Average
KDEF (unnormalized)	0.191	0.088	0.186	0.133	0.011	0.155	0.129	0.092	0.157
KDEF (normalized)	0.118	0.064	0.186	0.068	0.028	0.109	0.146	0.052	0.129
RAVDESS (video)	0.012	0.222	0.107	0.044	0.083	0.016	0.077	0.036	0.086
CK (video)	0.038	0.054	0.059	0.015	0.008	0.023	0.057	0.034	0.048

Table 4.1: Average MSE of selected AUs in original data and transfer results using different experimental settings. Note that we randomly select 8 out of 17 AUs output by OpenFace. Last column is the average MSE over all AUs in each setting.

4.3 Network Training

We use the aforementioned data to train our deep neural network to input AUs and output corresponding muscle and jaw activations.

The AUs from each training pair generated as described in the previous section, are passed to the network as input. We then compare the corresponding muscle and jaw activations; i.e., the ground truth to the predictions of muscle and jaw activations given by the neural network. We use a Mean Square Error training loss between the predictions and the ground truth. The loss is then backpropagated to update the weights, thus training the neural network.

We normalize the intensity values of each AU class across all pairs to remove the disparity of intensity values between synthetic faces and real faces. We use a total of 6,000 pairs, with about 1,000 pairs for each basic expression.

To train the neural network, we use the ADAM stochastic gradient descent optimizer with a Mean Squared Error loss, a batch size of 32, and a learning rate of 0.01. We train the network in a total of 100 epochs, running on an NVIDIA GeForce GTX 1650 GPU installed on a Windows 10 Machine with a 2.6GHz Intel Core i7-9750H CPU. The convergence of the training error is presented in Figure 4.2.

4.4 Expression Transfer Pipeline

We use a pipeline similar to the offline training module for the transfer of real facial expressions on the fly. We once again leverage OpenFace for facial expression recognition

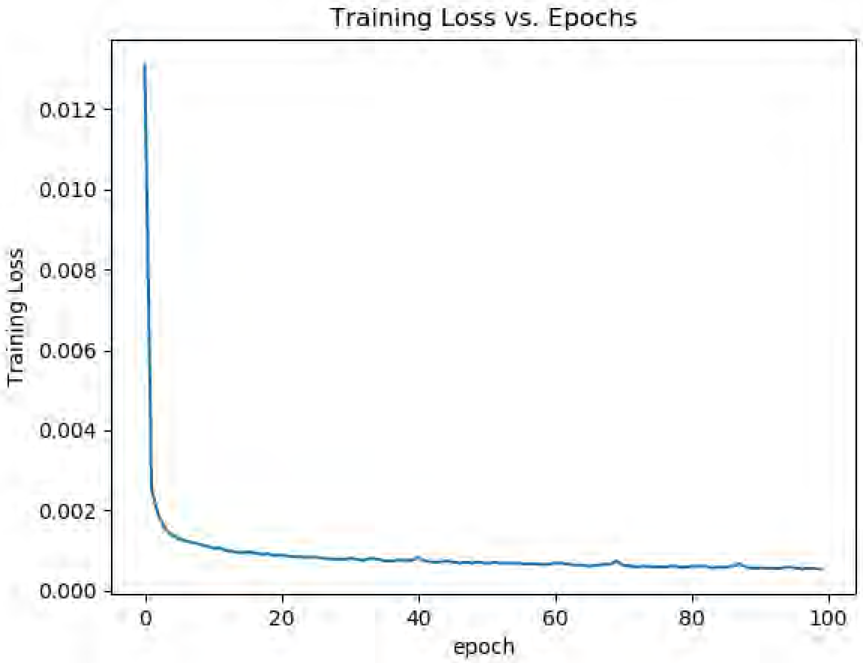


Figure 4.2: Convergence of the neural network

and head pose estimation. We input an image of an expressive face into OpenFace to obtain all of the corresponding AUs and head orientations. The AUs are then passed into the trained neural network which outputs predictions of the muscle and jaw activations, driving the biomechanical face to deform the muscles and transfer the expressions onto it.

We transfer both image and video inputs. Each frame in a video is processed independently and a resulting video is created using the transferred frames. The intensity values for each AU class are normalized across all the images or frames as in the case of the training pipeline.

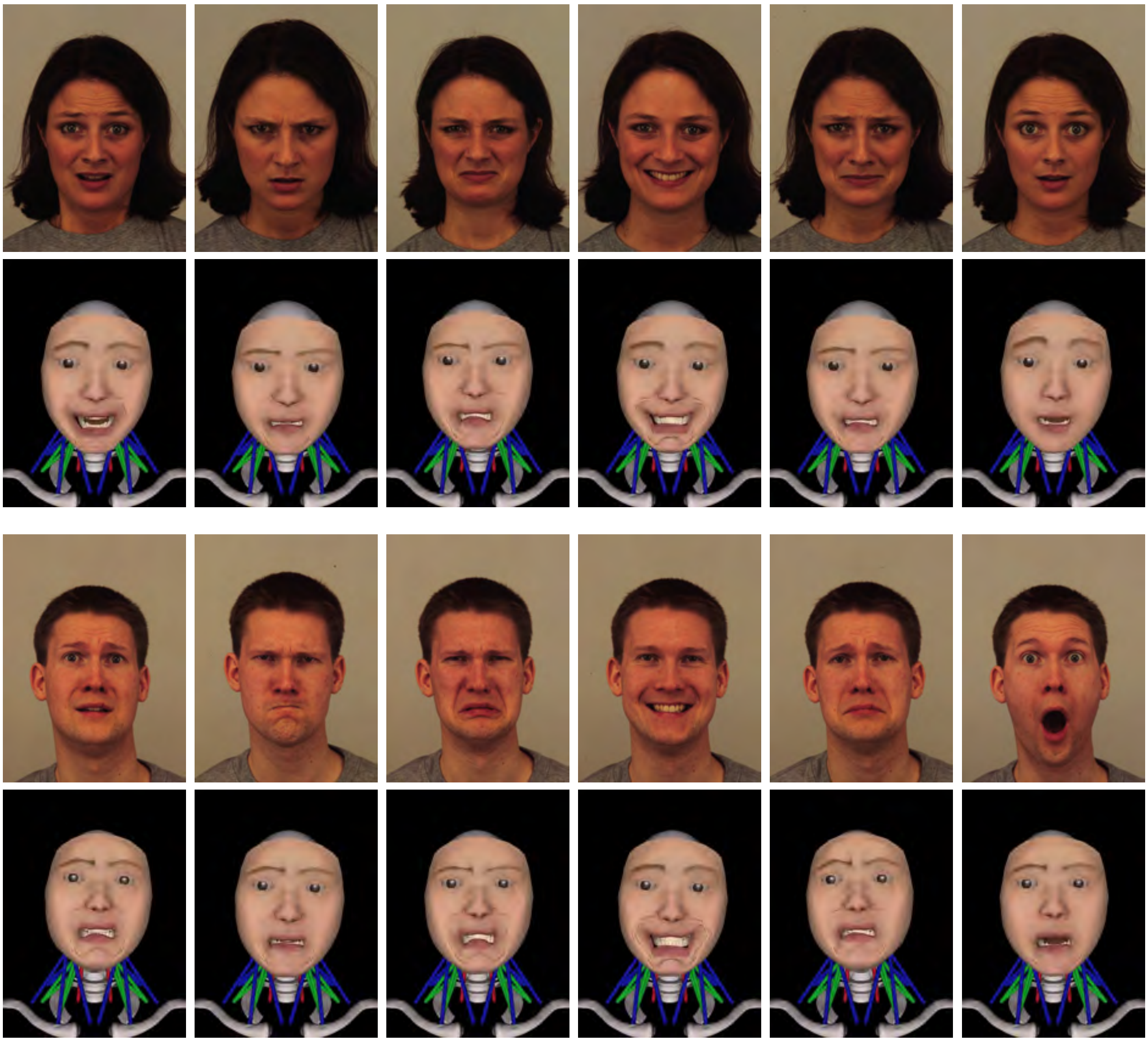


Figure 4.3: Transfer of Faces from the KDEF dataset. The first and third row consist of original images of two subjects. The second and fourth row are transferred versions of the first and third rows, respectively. The transferred expressions in order are (1) Fear, (2) Anger, (3) Disgust, (4) Joy, (5) Sadness, and (6) Surprise

CHAPTER 5

Experiments and Results

We next present the results of transferring facial expressions from the wild using our trained neural network. We evaluate our expression transfer pipeline on different expressions while using a variation of AUs and muscles in the biomechanical face model.

5.1 Facial Expression Datasets

Several facial expression datasets are available online. The datasets that we used in experiments are as follows:

Karolinska Directed Emotional Faces (KDEF) (Calvo and Lundqvist, 2008). The KDEF dataset consists of 4,900 pictures of human facial expressions. It covers 70 subjects (35 female and 35 male) enacting all the basic facial expressions, namely Neutral, Joy, Sadness, Anger, Fear, Disgust, and Surprise. Each expression enacted by the subject is captured from multiple directions. We use the dataset to transfer facial expressions onto the biomechanical face model and visually analyze the performance of our trained neural network in this paper.

Cohn Kanade Dataset (CK) and **Extended Cohn Kanade Dataset (CK+)** (Kanade et al., 2000; Lucey et al., 2010). The CK and the CK+ dataset combined consist of 593 video sequences of 123 subjects. Each sequence consists of images from a neutral expression (first frame) to a peak expression (last frame). The peak expressions are FACS coded for AUs. We use the sequences in the CK+ dataset to transfer videos of expression transitions onto the biomechanical face.

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

(Livingstone and Russo, 2018). The original RAVDESS dataset consists of 24 actors vocalizing two lexically-matched statements. An extension of the dataset named **RAVDESS Facial Landmark Tracking** contains tracked facial landmark movements from the original RAVDESS datasets (Swanson et al., 2019). Each actor performs 60 speech trials and about 44 song trials. This yields a total of 2,452 video files for the complete dataset. We leverage this dataset to test the transfer of actor faces in speech videos onto the biomechanical face.

5.2 Results

5.2.1 Actions Units and Muscle Activations

There exist a total of 30 Action Units corresponding to facial expressions. OpenFace provides occurrence predictions for 18 out the 30 Action Units and measures intensity values for 17 out of the 30 Action Units. We consider the 17 Action Units for which the intensity values are present as the super-set of the Action Units for our use case.

Due to the correlation between the Action Units and muscles in the face, there also exists a correlation between a basic facial expression and the Actions Units activated by it. Our initial experiments focused on training the neural network for each expression in an isolated manner. The neural network was trained to output muscle activation for muscles corresponding to a single expression using Actions Units which pertained to the same expression. In further trials, we observed that the usage of all the 17 Action Units and all facial muscles improved the performance and the scalability of the expression transfer pipeline.

5.2.2 Normalization

Figure 5.1 presents a comparison of transfer results with and without normalization of intensity values of each AU class across the complete dataset. Due to limitations in the biomechanical model, the range of each AU class differs from that of the real faces. Hence, we use normalization to overcome the bias and better transfer real facial expressions onto the biomechanical model. The expression intensities in transferred expressions with normalization better represent the real faces than those without normalization.

5.2.3 Jaw Activation

In Figure 5.2, we compare the transfer results with and without jaw activation. We choose to activate only jaw rotation so as to maintain the symmetry of the expression for our use case. We observe, that without jaw activations, expressions such as surprise are not well synthesized by the biomechanical face model.

5.2.4 Head Orientation

We leverage OpenFace for head pose estimation. We pass the estimated orientation of the head into the trained neck controller to activate the neck muscles. This in turn actuates the neck to adjust the head orientation in accordance with the input image. We present transferred results with the head orientations from the KDEF dataset in Figure 5.3.

5.2.5 Facial Expression Transfer

The results for the facial expression transfer for each of the expressions is shown in Figure 4.3. We present transfer results of a small sample of the KDEF dataset. We pick 2 subjects (1 male and 1 female), enacting all the basic expressions.

We also evaluate the transfer results by comparing the mean square error (MSE) of selected AUs in Table 4.1. We calculate the average MSE by AUs over all transferred expressions and their corresponding reference data, then compute the mean value of the average MSEs over all AUs in each experimental setting. We report such mean values together with 8 randomly chosen AUs. The normalization step decreases the MSEs of most selected AUs and yields results with higher AU similarity, which is consistent with the better transfer of expressions illustrated in Figure 5.1. The transfer results with the RAVDESS and CK datasets are best seen in the supplemental video. The more

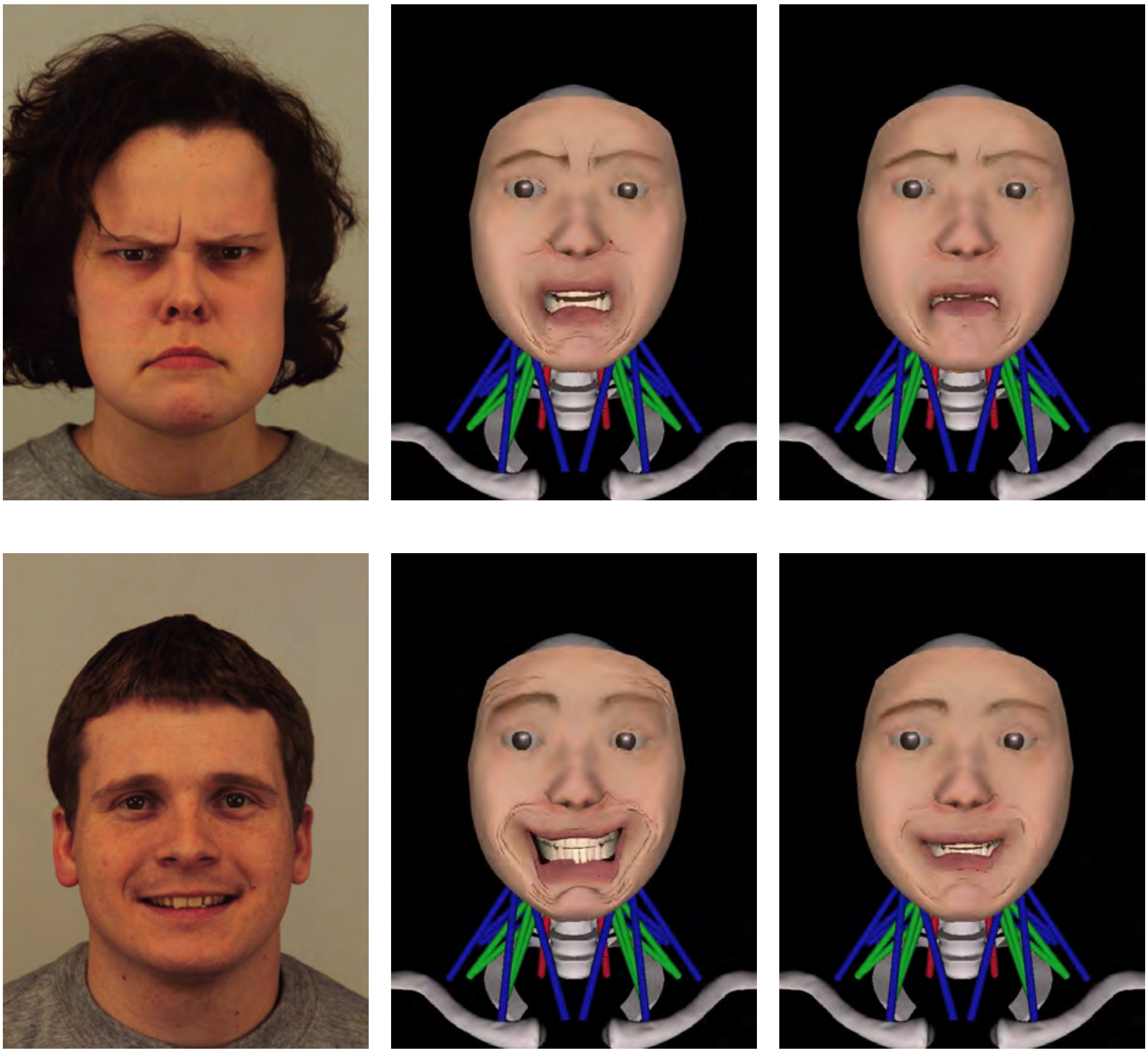


Figure 5.1: A comparison of transfer results with and without AU normalization. The three columns correspond to (i) The original image from the KDEF dataset, (ii) Result without normalization of AUs, and (iii) Result with normalization of AUs.

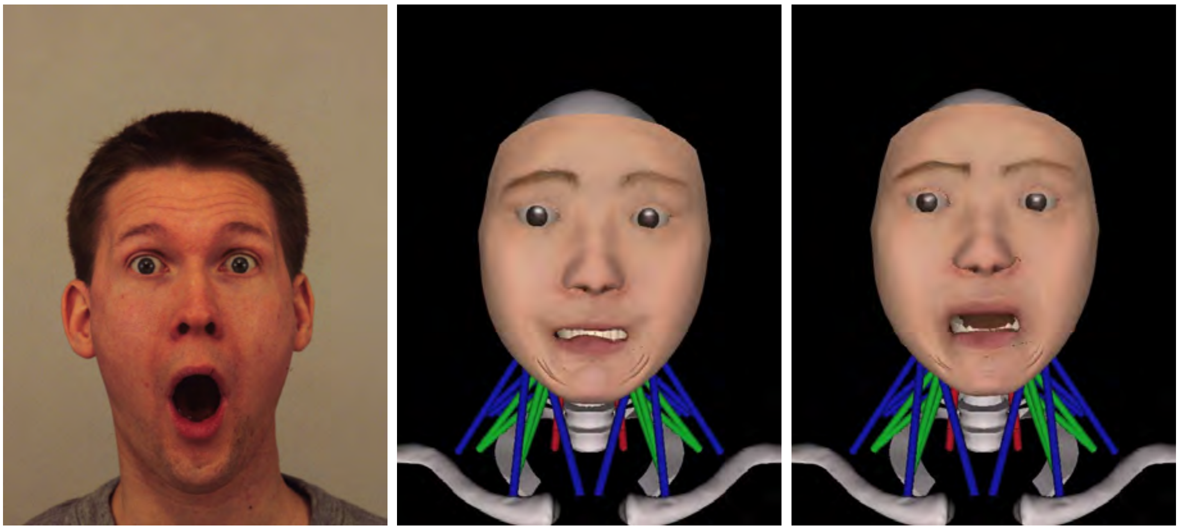


Figure 5.2: A comparison of transfer results with and without jaw activations. The three images are (i) The original image from the KDEF dataset, (ii) The transferred face without any jaw activations and (iii) The transferred face with jaw activations.

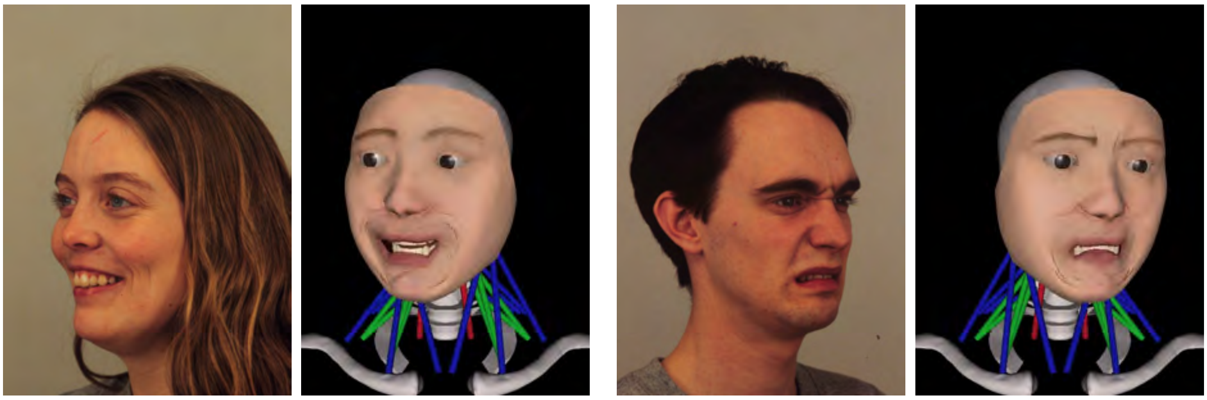


Figure 5.3: Head orientation applied to transfer task. The odd columns correspond to the original images and the even columns present the respective transferred faces along with orientation.

subtle expressions performed by subjects in the videos result in the lower average MSEs compared with those of image dataset.

5.2.6 Expression Transfer Involving Image Sequences

Our pipeline for expression transfer was further applied to sequences of images in which a subject is either forming an expression, speaking a sentence, or singing a song. We present expression transfer results for both the Cohn-Kanade (CK) dataset and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset.

5.2.6.1 Expression Transfer for the CK Dataset

As described in Section 5.1 the CK and CK+ datasets consist of sequences with subjects starting with a neutral expression and finally reaching a peak expression. Sample results of applying our expression transfer pipeline to two of these sequences are presented in Figure 5.4 and Figure 5.5. We observe that the pipeline is able to capture even subtle expression changes in the subjects.

5.2.6.2 Expression Transfer for the RAVDESS Dataset

As described in Section 5.1, the RAVDESS dataset has subjects either reading out a sentence or singing lyrics in a song. The results of applying our expression transfer pipeline to two sequences are presented in Figure 5.6 and Figure 5.7. We observe that our pipeline is able to transfer the subtle motion of the lips while talking and singing.

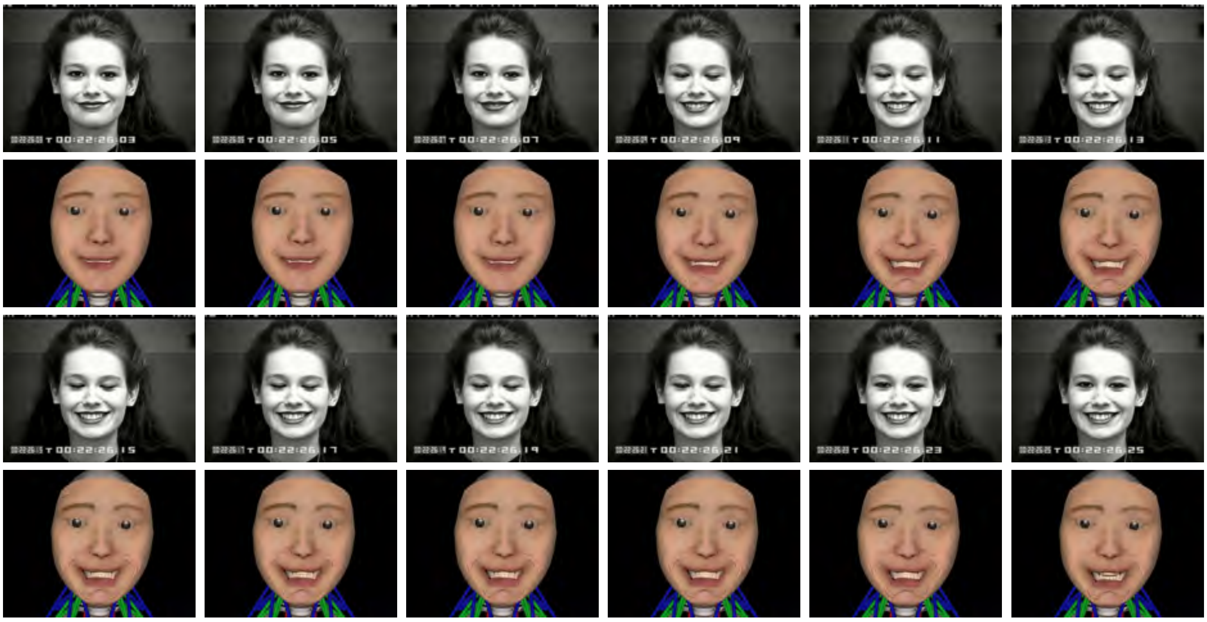


Figure 5.4: Transfer of Image sequences of a subject enacting Joy from the Cohn-Kanade dataset. The first and third row consist of original images of a subject in sequence. The second and fourth row are transferred versions of the first and third rows, respectively.

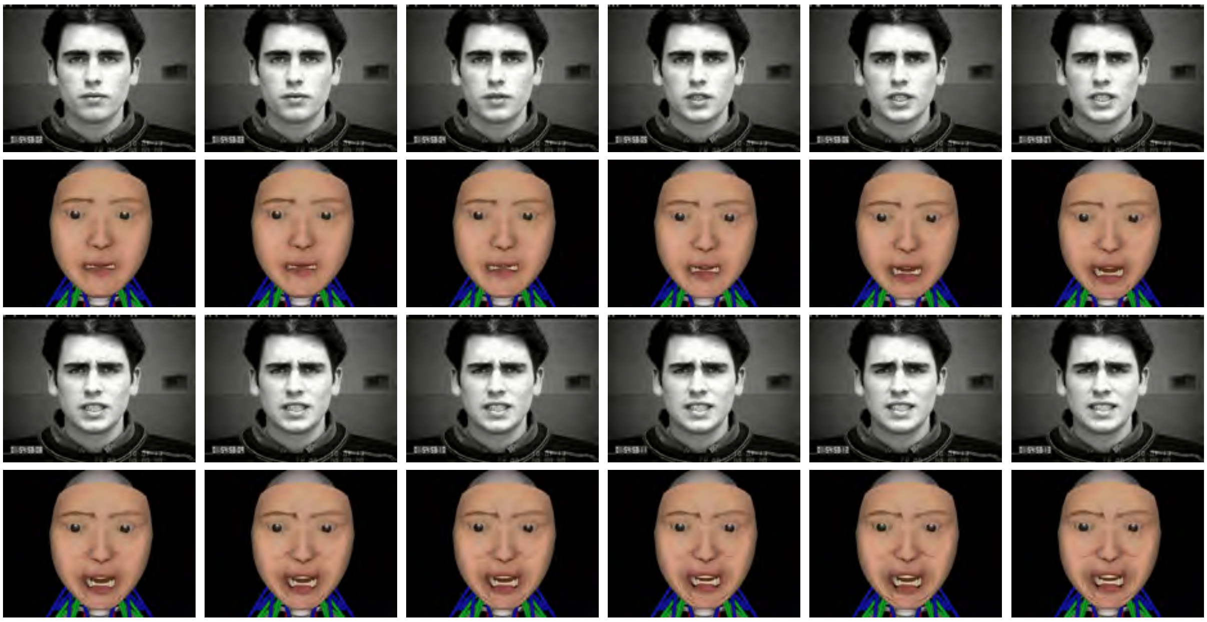


Figure 5.5: Transfer of Image sequences of a subject enacting Sadness from the Cohn-Kanade dataset. The first and third row consist of original images of a subject in sequence. The second and fourth row are transferred versions of the first and third rows, respectively.

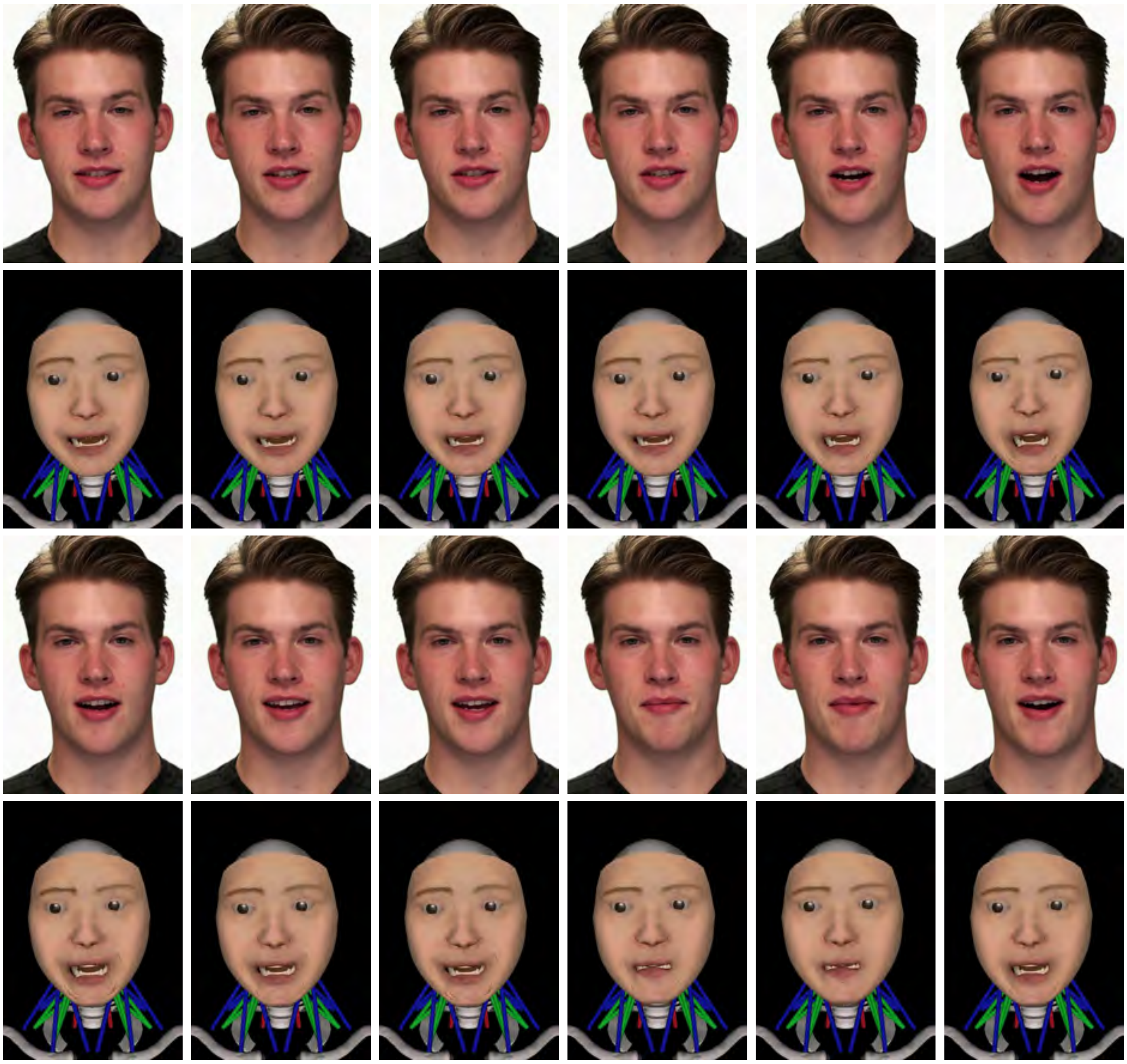


Figure 5.6: Transfer of Image sequences from the RAVDESS dataset of a subject speaking with Joy. The first and third row consist of original images of the subject. The second and fourth row are transferred versions of the first and third rows, respectively.

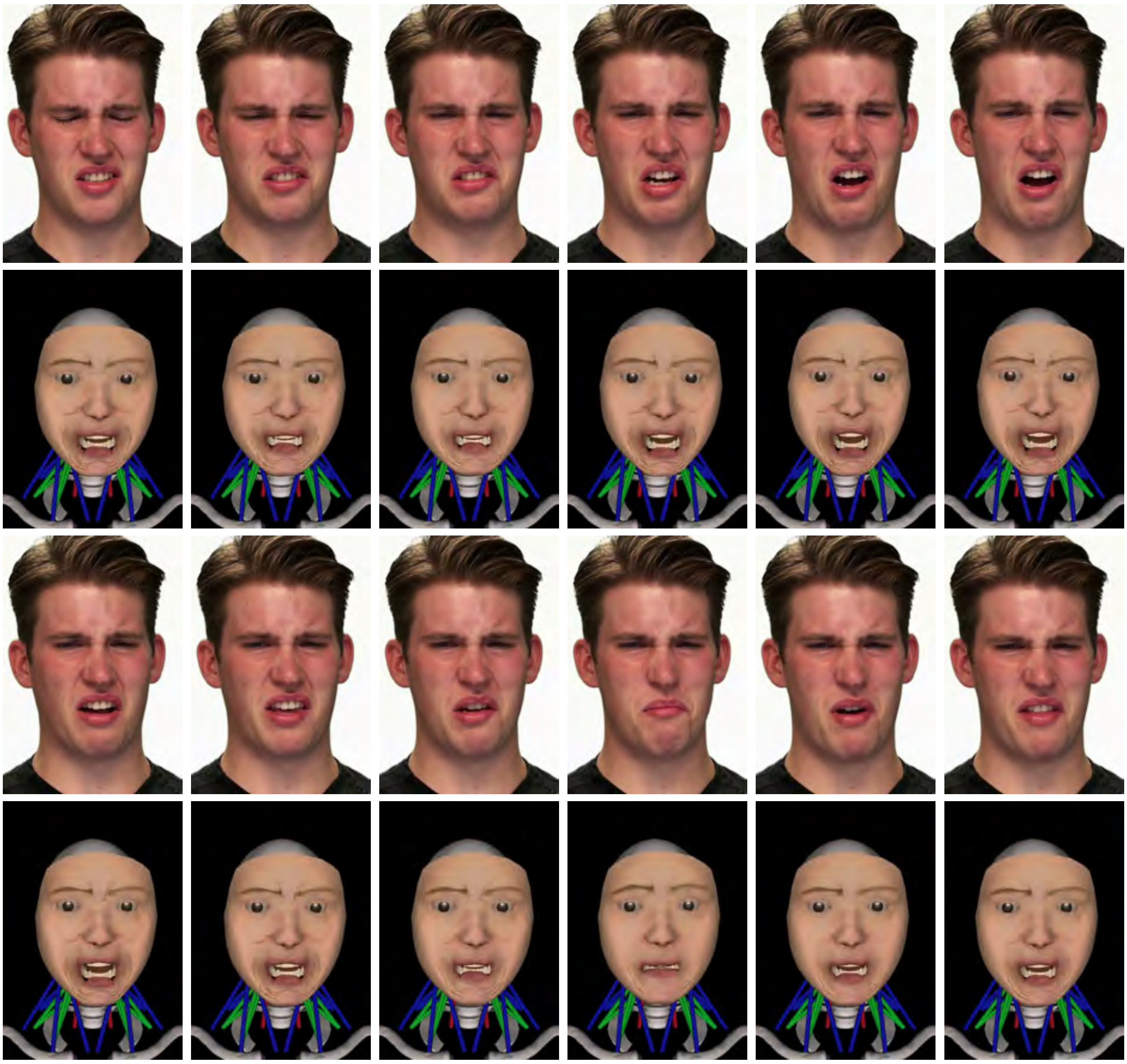


Figure 5.7: Transfer of Image sequences from the RAVDESS dataset of a subject speaking with Disgust. The first and third row consist of original images of the subject. The second and fourth row are transferred versions of the first and third rows, respectively.

CHAPTER 6

Conclusions and Future Work

6.1 Summary

Our expression transfer approach is uniquely advantageous. First, it is anatomically accurate as the biomechanical model emulates the musculoskeletal system of the head and neck of an actual human. Further, our approach is based on the Facial Action Coding System, which is a widely adopted representation of facial expressions in the computer animation field. Finally, our approach is based on deep learning, which can establish a non-linear relation between the Action Units of the FACS and facial muscle activations.

Additionally, our neural-network-based approach does not require any manual data collection as the training data is generated directly using the biomechanical model. This only needs to take place once, in advance. Once trained offline using the generated training data, the neural network can quickly transfer a large number of facial expressions.

6.2 Future Work

We demonstrated transfer results for all the basic expressions. As the human face is capable of creating a mixture of expressions, subtle expressions, and microexpressions, we would like to extend our work to transfer all of the above to our biomechanical model. We would also like to improve the biomechanical model by adding more muscles and wrinkle removal on specific areas of the face. We hope that this will help activate facial Action Units in a more anatomically accurate manner, which will further enable our approach to produce better transfer results.

APPENDIX A

Details of the Biomechanical Face Model

A.1 Skeleton

The skeletal structure of our model contains 9 links and is modeled as an articulated rigid-body system. The bone links are built on the base link, which is set to be immobile. Among the 8 movable bones, there are seven cervical bones, named C1 to C7, and the skull. The human neck skeleton system contains soft tissue between vertebrae and the cervical spine, which enables motion with 6 degrees of freedom (3 for translation and 3 for rotation) for the bones. We simplify each bone joint to have just 3 rotational degrees of freedom.

The equation of motion of the rigid-body system can be written as

$$\boldsymbol{M}(\boldsymbol{q})\ddot{\boldsymbol{q}} + \boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}}) + \boldsymbol{K}_s\boldsymbol{q} + \boldsymbol{K}_d\dot{\boldsymbol{q}} = \boldsymbol{P}(\boldsymbol{q})\boldsymbol{f}_m + \boldsymbol{J}(\boldsymbol{q})^T\boldsymbol{f}_{\text{ext}}, \tag{A.1}$$

where $\boldsymbol{q}$, $\dot{\boldsymbol{q}}$, and $\ddot{\boldsymbol{q}}$ are 24-dimensional vectors representing the joint angles, the angular velocities, and the angular accelerations, respectively. $\boldsymbol{M}$ is the mass matrix, and $\boldsymbol{C}$ represents the internal forces among the system, including Coriolis forces, gravity, and forces from connecting tissues. The moment arm matrix $\boldsymbol{P}$ maps the muscle force $\boldsymbol{f}_m$ (contractile muscle force $\boldsymbol{f}_C$ and passive muscle force $\boldsymbol{f}_P$) to the related joint torques, while the Jacobian matrix $\boldsymbol{J}$ transforms the applied external force $\boldsymbol{f}_{\text{ext}}$ to torques. The $\boldsymbol{K}_s\boldsymbol{q} + \boldsymbol{K}_d\dot{\boldsymbol{q}}$ term represents the rotational damping springs that we attach to the joints in order to simulate the stiffness of the inter-vertebral discs. We can alternatively write the torque from the spring as:

$$\tau_s = -k_s(q - q_0) - k_d\dot{q}, \tag{A.2}$$

where q is the current joint angle in the generalized coordinates, and q_0 is the joint angle in the natural pose (resting angle). The k_s and k_d are the stiffness and damping coefficients of the spring, respectively.

A.2 Muscles

We use a modified version of the Hill-type muscle model (Ng-Thow-Hing, 2001), which is a good balance of biomechanical accuracy and computational efficiency. The muscle force $f_m = f_P + f_C$ has two sources. The passive element f_P generates a restoring force due to the muscle elasticity, which constrains the muscle deformation passively. The passive muscle force is represented as

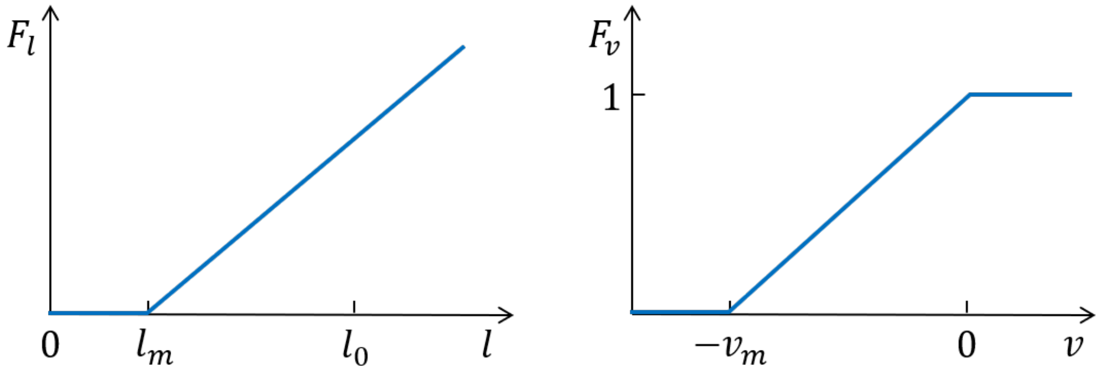
$$f_P = \max(0, k_s(\exp(k_c e) - 1) + k_d\dot{e}), \tag{A.3}$$

where k_s and k_d are the stiffness and damping coefficients of the above uni-axial exponential spring model. e is the strain of the muscle and $\dot{e}$ is the strain rate. We can calculate them using $e = (l - l_0)/l_0$ and $\dot{e} = \dot{l}/l_0$, respectively, where l and l_0 are the muscle length and muscle resting length.

The contractile element f_C generates the proactive contractile force of the muscle, which is proportionate to the *activation level* of the muscle:

$$\begin{aligned} f_C &= aF_l(l)F_v(\dot{l}), \\ F_l(l) &= \max(0, k_{max}(l - l_m)), \\ F_v(\dot{l}) &= \max(0, 1 + \min(\dot{l}, 0)/v_m), \end{aligned} \tag{A.4}$$

where $a \in [0, 1]$ is the muscle activation level, F_l is the force-length relation, and F_v is the



(a) Force-length relationship

(b) Force-velocity relationship

Figure A.1: The force relationships of Hill type muscle model.

force-velocity relation. During their calculation, we need the maximum muscle stiffness k_{max}, maximum muscle length l_m, and the maximum contractile velocity v_m. We present the plots of the two relations in Figure A.1. Additional details about the setting and the biomechanical background of the parameters can be found in (Lee and Terzopoulos, 2006; Ng-Thow-Hing, 2001).

A.3 Skin

Our facial skin model is automatically generated using the technique described in (Lee et al., 1995). After laser-scanning the individual's facial data, a range image and a reflectance image in cylindrical coordinates are adapted to a well-structured face mesh. Then the algorithm creates a dynamic skin and muscle model for the face mesh, which contains automatically generated facial soft tissues, estimated skull surface. Also, major muscles responsible for facial expression are inserted to the model.

The physical simulation of the muscle-actuated facial skin model is implemented as a discrete deformable model (DDM), where a network of fascia nodes are connected using uni-axial springs. The force exerts from spring j on node i can be written as

$$\boldsymbol{g}_i^j = c_j(l_j - l_j^r)\boldsymbol{s}_j, \tag{A.5}$$

where l_j and l_j^r are the current and resting length of spring j, and $\boldsymbol{s}_j = (\boldsymbol{x}_j - \boldsymbol{x}_i)/l_j$ is the spring direction vector.

The facial skin model is also actuated by the underlying muscles. We calculate the force exerts from muscle j on node i according to the length scaling function Θ_1 and the muscle-width scaling function Θ_2 as follows:

$$f_i^j = \Theta_1(\varepsilon_{j,i})\Theta_2(\omega_{j,i}). \tag{A.6}$$

In (Lee et al., 1995), the plots of the two scaling functions, as well as the definition of the length ratio ε and muscle width ω are explained in detail. Moreover, there are other aspects of the generated discrete deformable model; e.g., the volume preservation forces and skull penetration constraint forces.

REFERENCES

Alkawaz, M. H., Mohamad, D., Basori, A. H., and Saba, T. (2015). Blend shape interpolation and facs for realistic avatar. *3D Research*, 6(1):6. 5

Aneja, D., Colburn, A., Faigin, G., Shapiro, L., and Mones, B. (2016). Modeling stylized character expressions via deep learning. In *Asian conference on computer vision*, pages 136–153. Springer. 4, 5

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE. 11

Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. 5

Calvo, M. G. and Lundqvist, D. (2008). Facial expressions of emotion (kdef): Identification under different display-duration conditions. *Behavior Research Methods*, 40(1):109–115. 15

Cao, C., Weng, Y., Zhou, S., Tong, Y., and Zhou, K. (2013). Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425. 5

Cardona, M. and Cena, C. E. G. (2019). Biomechanical analysis of the lower limb: A full-body musculoskeletal model for muscle-driven simulation. *IEEE Access*, 7:92709–92723. 5

Chen, A., Chen, Z., Zhang, G., Mitchell, K., and Yu, J. (2019). Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9429–9439. 4, 5

Chen, D. T. and Zeltzer, D. (1992). Pump it up: Computer animation of a biomechanically based model of muscle using the finite element method. In *Proceedings of the 19th annual conference on computer graphics and interactive techniques*, pages 89–98. 4

Cohn, J. F., Ambadar, Z., and Ekman, P. (2007). Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, 1(3):203–221. 1, 5

Ekman, P. (2002). Facial action coding system (facs). *A human face.* 5

Ekman, P. and Friesen, W. V. (1978). *Manual for the facial action coding system.* Consulting Psychologists Press. 5

Faloutsos, P., Van de Panne, M., and Terzopoulos, D. (2001). Composable controllers for physics-based character animation. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 251–260. 4

Hodgins, J. K., Wooten, W. L., Brogan, D. C., and O'Brien, J. F. (1995). Animating human athletics. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 71–78. 4

Ishikawa, T., Morishima, S., and Terzopoulos, D. (2000). 3d face expression estimation and generation from 2d image based on a physically constraint model. *IEICE TRANSACTIONS on Information and Systems*, 83(2):251–258. 6

Kanade, T., Cohn, J., and Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 46–53. 15

Lee, S., Park, M., Lee, K., and Lee, J. (2019). Scalable muscle-actuated human simulation and control. *ACM Transactions On Graphics (TOG)*, 38(4):1–13. 1, 4

Lee, S.-H., Sifakis, E., and Terzopoulos, D. (2009). Comprehensive biomechanical modeling and simulation of the upper body. *ACM Transactions on Graphics (TOG)*, 28(4):1–17. 1, 4

Lee, S.-H. and Terzopoulos, D. (2006). Heads up! biomechanical modeling and neuromuscular control of the neck. In *ACM SIGGRAPH 2006 Papers*, pages 1188–1198. 2, 5, 7, 27

Lee, Y., Terzopoulos, D., and Waters, K. (1995). Realistic modeling for facial animation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 55–62. 1, 5, 6, 7, 27, 28

Livingstone, S. R. and Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):1–35. 16

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101. 15

Moschoglou, S., Ploumpis, S., Nicolaou, M. A., Papaioannou, A., and Zafeiriou, S. (2020). 3dfacegan: Adversarial nets for 3d face representation, generation, and translation. *International Journal of Computer Vision*, 128:2534–2551. 4

Nakada, M., Zhou, T., Chen, H., Weiss, T., and Terzopoulos, D. (2018). Deep learning of biomimetic sensorimotor control for biomechanical human animation. *ACM Transactions on Graphics (TOG)*, 37(4):1–15. 4

Ng-Thow-Hing, V. (2001). *Anatomically-based models for physical and geometric reconstruction of humans and other animals*. PhD thesis, University of Toronto. 26, 27

Platt, S. M. and Badler, N. I. (1981). Animating facial expressions. In *Proceedings of the 8th annual conference on Computer graphics and interactive techniques*, pages 245–252. 6

Rajagopal, A., Dembia, C. L., DeMers, M. S., Delp, D. D., Hicks, J. L., and Delp, S. L. (2016). Full-body musculoskeletal model for muscle-driven simulation of human gait. *IEEE transactions on biomedical engineering*, 63(10):2068–2079. 1, 5

Richardson, E., Sela, M., Or-El, R., and Kimmel, R. (2017). Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1259–1268. 4

Sachdeva, P., Sueda, S., Bradley, S., Fain, M., and Pai, D. K. (2015). Biomechanical simulation and control of hands and tendinous systems. *ACM Transactions on Graphics (TOG)*, 34(4):1–10. 5

Sagar, M., Bullivant, D., Robertson, P., Efimov, O., Jawed, K., Kalarot, R., and Wu, T. (2014). A neurobehavioural framework for autonomous animation of virtual human faces. In *SIGGRAPH Asia 2014 Autonomous Virtual Humans and Social Robot for Telepresence*, pages 1–10. 1, 6

Si, W., Lee, S.-H., Sifakis, E., and Terzopoulos, D. (2014). Realistic biomechanical simulation and control of human swimming. *ACM Transactions on Graphics (TOG)*, 34(1):1–15. 1, 4

Sifakis, E., Neverov, I., and Fedkiw, R. (2005). Automatic determination of facial muscle activations from sparse motion capture marker data. In *ACM SIGGRAPH 2005 Papers*, pages 417–425. 6

Sueda, S., Kaufman, A., and Pai, D. K. (2008). Musculotendon simulation for hand animation. In *ACM SIGGRAPH 2008 papers*, pages 1–8. 5

Swanson, R., Livingstone, S., and Russo, F. (2019). RAVDESS facial landmark tracking (version 1.0.0) [data set]. 16

Terzopoulos, D. and Lee, Y. (2004). Behavioral animation of faces: Parallel, distributed, and real-time. *FACIAL MODELING AND ANIMATION, ACM SIGGRAPH*, pages 119–128. 6

Terzopoulos, D. and Waters, K. (1990). Physically-based facial modelling, analysis, and animation. *The journal of visualization and computer animation*, 1(2):73–80. 6

Terzopoulos, D. and Waters, K. (1993). Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579. 6

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395. 5

Van den Bogert, A. J., Geijtenbeek, T., Even-Zohar, O., Steenbrink, F., and Hardin, E. C. (2013). A real-time system for biomechanical analysis of human movement and muscle function. *Medical & biological engineering & computing*, 51(10):1069–1077. 4

Wang, M., Bradley, D., Zafeiriou, S., and Beeler, T. (2020). Facial expression synthesis using a global-local multilinear framework. *Computer Graphics Forum*, 39(2):235–245. 4

Waters, K. (1987). A muscle model for animation three-dimensional facial expression. *Acm siggraph computer graphics*, 21(4):17–24. 6

Weise, T., Bouaziz, S., Li, H., and Pauly, M. (2011). Realtime performance-based facial animation. *ACM transactions on graphics (TOG)*, 30(4):1–10. 5

Xu, F., Chai, J., Liu, Y., and Tong, X. (2014). Controllable high-fidelity facial performance transfer. *ACM Transactions on Graphics (TOG)*, 33(4):1–11. 5

Zhang, J., Chen, K., and Zheng, J. (2020). Facial expression retargeting from human to avatar made easy. *IEEE Transactions on Visualization and Computer Graphics*. 4, 5