Neuromuscular Control of the Face-Head-Neck Biomechanical Complex with Learning-Based Expression Transfer from Images and Videos

Xiao S. Zeng[1(✉)], Surya Dwarakanath[1], Wuyue Lu[1], Masaki Nakada[2], and Demetri Terzopoulos[1]

[1] Computer Science Department, University of California, Los Angeles, CA, USA
stevennz@cs.ucla.edu
[2] NeuralX, Inc., Los Angeles, CA, USA

Abstract. 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 complex. 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 activation signals for the musculoskeletal model. Through biomechanical simulation, the activations deform the facial soft tissues, 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 face, head, and neck. 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 onto our face-head-neck model of facial expressions and head poses from a range of facial images and videos.

Keywords: Animation · Physical simulation · Computer vision · Deep learning

1 Introduction

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. Biomechanical human musculoskeletal models realistically capture the anatomy and physics underlying human

---

X. S. Zeng and S. Dwarakanath—Co-first author.

(a) Subject 1   (b) Subject 2   (c) Subject 3
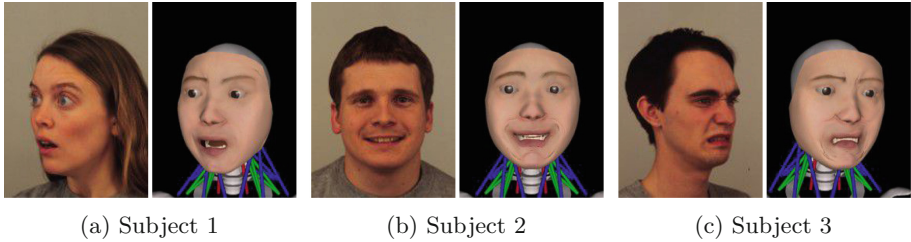
Fig. 1. Our deep learning framework enables the transfer of facial expressions and head poses from images or videos to a biomechanical face-head-neck model.

motion generation. In particular, those pertaining to the simulation of the human face have succeeded in convincingly emulating facial expressiveness; however, they require significant effort in parameter tuning to produce realistic results.

In this paper, we show how to endow a biomechanical, musculoskeletal model of the human face with the ability to produce facial expressions via machine learning from real-world reference images and videos (Fig. 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) [5] in the context of the muscle actuators that drive the musculoskeletal face model augmented with a musculoskeletal cervicocephalic (neck-head) system to animate head movement during facial expression synthesis.

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 (Fig. 2) for animating expressions and head poses using an improved version of the physics-based neck-head-face animation system developed by Lee and Terzopoulos [10], but which can potentially be applied to any muscle-driven model.

2   Related Work

Muscle-based facial animation has been studied for decades [19]. A series of studies have endeavored to build increasingly anatomically accurate musculoskeletal face models. Using such models, Terzopoulos and Waters [17] and Sifakis et al. [15], among others, have addressed the automatic determination of facial muscle activations from videos, but these methods require nonrigid motion trackers or motion capture markers. 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. Most relevant to our work in the present paper is an existing biomechanical model of the cervicocephalic complex [10], which incorporates an improved version of the face model developed by Lee et al. [11].
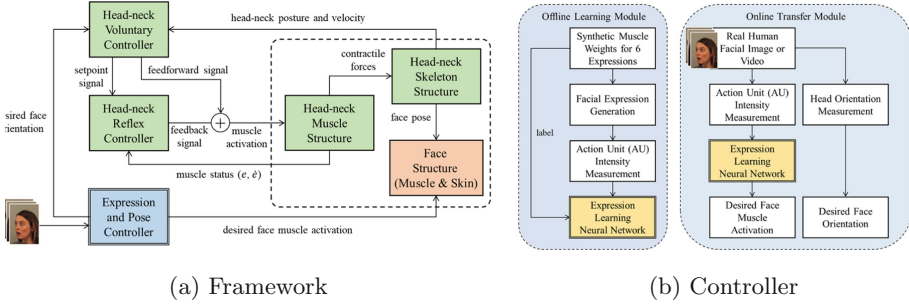
(a) Framework   (b) Controller

Fig. 2. The structure of our expression and head pose transfer framework (a) and the components of the facial expression and head pose controller (b). 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 muscle activations are input to the biomechanical face model (orange) to perform the corresponding expression and the head orientation is provided to the head-neck biomechanical neuromuscular system (green) to produce the desired head pose. (Color figure online)

Many facial animation researchers have made use of the well-known FACS [8], a quantitative phenomenological abstraction that 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 [5].

Expression transfer or retargeting is a trending topic and recent approaches often use FACS-based blendshapes as the basic parametric representation [18,20, 24]. Others have used techniques such as interactive mesh deformation control [22] and neural-network-based perceptual models [1] to represent expressions in blendshapes. However, transferring expressions using a musculoskeletal system is more natural since facial actions are the result of coordinated muscle contractions inducing soft tissue deformation. Despite the existing literature, there remains a deficiency of work in transferring expressions to musculoskeletal facial systems.

3   Musculoskeletal Model

Our real-time musculoskeletal model is based on the one of Lee and Terzopoulos [10], 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 Fig. 2a. Specifically, the skeletal structure is an articulated multibody dynamics system, with bones and joints consistent with human anatomy. The skeletal system is driven by a Hill-type muscle actuator model.

The biomechanical face component consists of a facial soft tissue model comprising epidermis, dermal-fatty, fascia, and muscle layers supported by an underlying skull, which is constructed based on the work of Lee et al. [11]. The soft tissue is a deformable model assembled from discrete uniaxial finite elements, which simulates dynamic facial deformation in an anatomically consistent yet simplified way compared to the models described in [15,21], and [6], thus maintaining a low computational cost that affords real-time simulation performance on readily available hardware.

There are 26 pairs of primary facial muscles embedded in the biomechanical face, including the frontalis, corrugator, levator labii, orbicularis oculi, mentalis, and orbicularis oris muscle groups. The contractions of these muscles apply forces to the facial soft tissue layers, inducing deformations that produce meaningful facial expressions. We have augmented the expressive details, such as wrinkles on the face model, by applying multiple levels of subdivision to increase significantly the number of surface nodes that can be activated by muscle forces, and applied a high resolution texture map to the surface mesh for a natural appearance.

The appendix of reference [23] explains the biomechanical components of our model in greater detail.

3.1   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 concurrently outputs head pose estimates to the head-neck musculoskeletal 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, and generates a feedforward muscle activation signal and a setpoint control signal. The latter, which encodes the desired muscle strains and strain rates, is input to the reflex controller that then generates a feedback muscle activation signal and adds it to the feedforward signal generated by the voluntary controller. As a result, each cervical muscle receives an activation signal a and, through simulation, generates a contractile muscle force accordingly. Together with the external environmental forces and gravity, the whole system is simulated through time and rendered as a physics-based animation. The voluntary controller runs 25 Hz (in simulation time), while the reflex controller runs at 4KHz along with the physical simulation time steps.

4   Expression Learning

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,
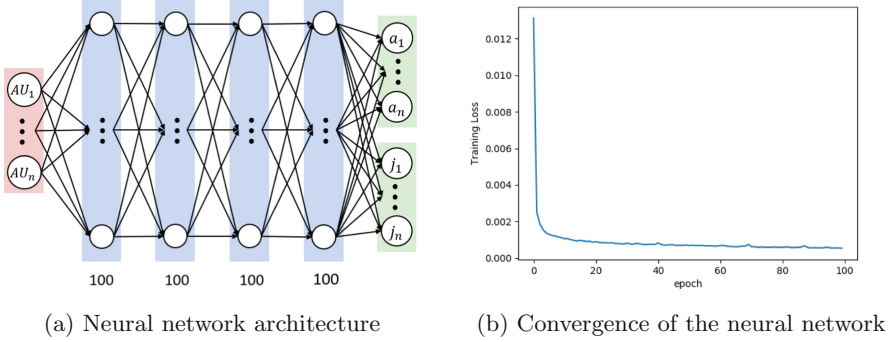
(a) Neural network architecture   (b) Convergence of the neural network

Fig. 3. The expression learning neural 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 muscles of the biomechanical face model such that they generate contractile forces that deform the synthetic facial tissues to produce facial expressions. We employ a fully connected deep neural network architecture for this purpose (Fig. 3a). The input layer consists of a total of 17 neurons representing the important AUs that are involved in the majority of facial movements, each neuron corresponding to a single normalized AU. We include 4 hidden layers with 100 neurons in each layer and ReLU activation.[1] The output layer consists of 56 neurons, 52 of which encode the activations a_i, with $1 \leq i \leq 52$, for each of the 26 pairs of facial muscles, and the remaining 4 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.

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

---

[1] This choice was based on observed Mean Squared Errors (MSE) on test data of networks with different numbers of hidden layers and different numbers of neurons per hidden layer.

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 is 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. 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 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 above 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 [2] 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 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 estimated AUs and head orientation associated with the input image. We repeat this process for each set A formed (as described previously) to obtain 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.

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 compared to the predictions of muscle and jaw activations given by the neural network. We use a Mean Square Error (MSE) training loss between the predictions and the ground truth, which is 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 an MSE 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 2.6 GHz Intel Core i7-9750H CPU. Figure 3b shows the convergence of the training error.

4.4   Expression Transfer Pipeline

To transfer real facial expressions on the fly, we use a pipeline similar to the offline training module, again leveraging OpenFace for facial expression recognition 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.

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 our experiments are as follows:

Karolinska Directed Emotional Faces (KDEF) [3]. 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 performed by the subject is imaged 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+)* [9,13]. 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

Table 1. Comparison of the Mean Squared Errors for training individual neural networks for each expression and for training a single neural network for all.

MSE_{Joy}	$MSE_{Sadness}$	MSE_{Fear}	MSE_{Anger}	$MSE_{Surprise}$	$MSE_{Disgust}$	MSE_{All}
0.000591	0.002925	0.001611	0.009689	0.000266	0.002516	0.000729

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) [12]. 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 [16]. 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   Action Units and Muscle Activations

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

Due to the correlation between the AUs and muscles in the face, there also exists a correlation between a basic facial expression and the AUs 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 AUs which pertained to the same expression. In further trials, we observed that usage of all 17 AUs and all facial muscles improved the performance and the scalability of the expression transfer pipeline.

Table 1 provides a comparison between training a single neural network for all expressions and training individual neural networks for each expression, where each network is trained only from the AUs and muscles relevant to its expression (using the mappings of expressions to AUs and AUs to muscles presented in [4,7] and [14]). We observe better performance when training a single neural network for all the expressions, suggesting that AUs not directly relevant to an expression also play a role in expression transfer.

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.

Table 2. Average MSE of selected AUs in the original data and transfer results using different settings. Note that we randomly select 8 out of 17 AUs output by OpenFace. The last column is the average MSE over all the AUs in each setting.

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

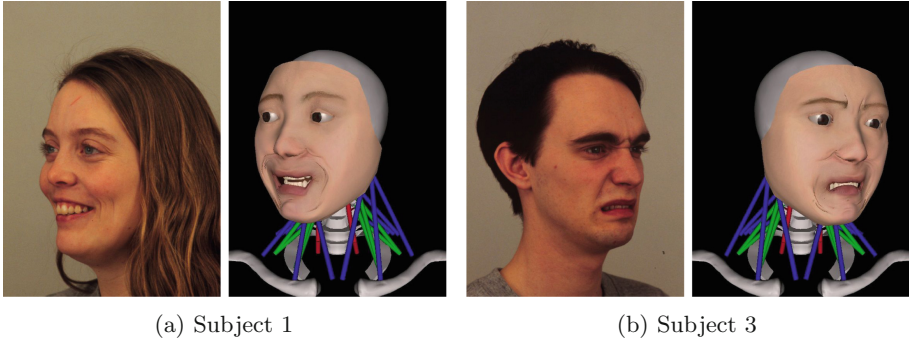(a) Subject 1   (b) Subject 3

Fig. 4. Transfer of head orientation along with facial expression.

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.

5.3   Head Orientation

Leveraging 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. Figure 4 presents sample transferred results including head orientations from the KDEF dataset.
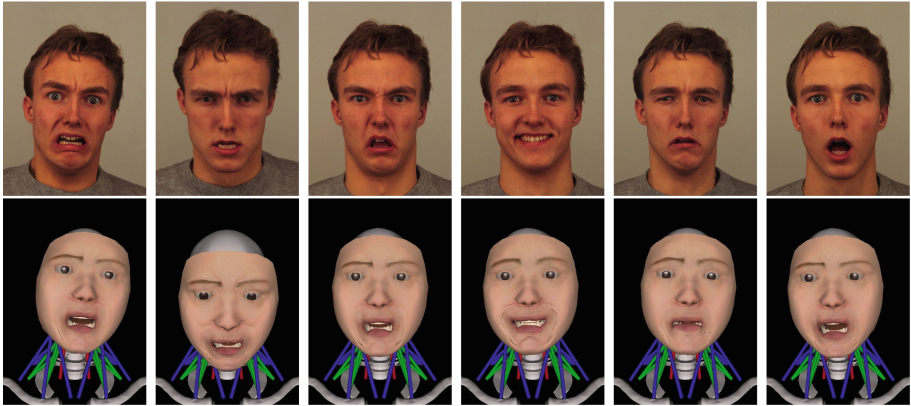
5.4   Facial Expression Transfer

The results for the facial expression transfer for each of the expressions is shown in Fig. 5. We present transfer results of a small sample of the KDEF dataset, selecting two subjects (a female and a male) enacting all the basic expressions.

We also evaluate the transfer results by comparing the MSE of selected AUs in Table 2. 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

(a) Subject 4

(b) Subject 5

Fig. 5. Transfer of facial expressions (fear, anger, disgust, joy, sadness, surprise) and head poses from KDEF images to the biomechanical face-head-neck model.

values together with 8 randomly chosen AUs. The normalization step decreases the MSEs of most selected AUs and yields results with higher AU similarity. The more subtle expressions performed by subjects in the videos result in the lower average MSEs compared with those of the image dataset.

6   Conclusion and Future Work

Our expression transfer approach is uniquely advantageous. First, it is anatomically consistent as the biomechanical model emulates the human cervicocephalic musculoskeletal system. Further, our approach is based on the Facial Action Coding System, which is a widely adopted representation of facial expressions, including in the computer animation field. Finally, our approach is based on

deep learning, which can capture the complex, non-linear relation between the Action Units of the FACS and associated 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 itself. This only needs to take place once, in advance. Once trained offline using the synthesized training data, the neural network can quickly transfer a large number of facial expressions. Unlike popular face tracking software such as Faceware, which requires manual calibration steps to transfer expressions of different subjects, our proposed approach needs no additional parameter adjustments to perform the same task.

Although we have achieved satisfactory transfer results for all the basic facial expressions, note that the simulation and transfer of complex mouth and eye movements lies outside the scope of our current work. To improve the fidelity of the results, we plan to extend our musculoskeletal model (e.g., by adding more muscles to help activate facial AUs in a more anatomically accurate manner) and its associated control system so as to explicitly control the lips and eyes, as well as to transfer mixtures of expressions, subtle expressions, and microexpressions, which the human face is capable of producing.

References

1. Aneja, D., Colburn, A., Faigin, G., Shapiro, L., Mones, B.: Modeling stylized character expressions via deep learning. In: Proceedings Asian Conference on Computer Vision, pp. 136–153 (2016)
2. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: OpenFace 2.0: facial behavior analysis toolkit. In: Proceedings IEEE Conference on Automatic Face and Gesture Recognition, pp. 59–66 (2018)
3. Calvo, M.G., Lundqvist, D.: Facial expressions of emotion (KDEF): identification under different display-duration conditions. Behav. Res. Methods **40**(1), 109–115 (2008)
4. Clark, E.A., et al.: The facial action coding system for characterization of human affective response to consumer product-based stimuli: a systematic review. Front. Psychol. **11**, 920 (2020)
5. Cohn, J.F., Ambadar, Z., Ekman, P.: Observer-based measurement of facial expression with the facial action coding system. Handbook Emotion Elicitat. Assessment **1**(3), 203–221 (2007)
6. Cong, M., Bao, M., E, J.L., Bhat, K.S., Fedkiw, R.: Fully automatic generation of anatomical face simulation models. In: Proceedings ACM SIGGRAPH/EG Symposium on Computer Animation, pp. 175–183 (2015)
7. Du, S., Tao, Y., Martinez, A.M.: Compound facial expressions of emotion. Proc. Natl. Acad. Sci. **111**(15), E1454–E1462 (2014)
8. Ekman, P., Friesen, W.V.: Manual for the Facial Action Coding System. Consulting Psychologists Press, Palo Alto (1978)

9. Kanade, T., Cohn, J., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53 (2000)
10. Lee, S.H., Terzopoulos, D.: Heads up! Biomechanical modeling and neuromuscular control of the neck. ACM Trans. Graph. **25**(3), 1188–1198 (2006)
11. Lee, Y., Terzopoulos, D., Waters, K.: Realistic modeling for facial animation. In: Proceedings ACM SIGGRAPH 95 Conference, pp. 55–62 (1995)
12. Livingstone, S.R., Russo, F.A.: 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 (2018)
13. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: Proceedings IEEE CVPR Workshops, pp. 94–101 (2010)
14. Scherer, K.R., Ellgring, H., Dieckmann, A., Unfried, M., Mortillaro, M.: Dynamic facial expression of emotion and observer inference. Front. Psychol. **10**, 508 (2019)
15. Sifakis, E., Neverov, I., Fedkiw, R.: Automatic determination of facial muscle activations from sparse motion capture marker data. In: Proceedings ACM SIGGRAPH 2005 Conference, pp. 417–425 (2005)
16. Swanson, R., Livingstone, S., Russo, F.: RAVDESS facial landmark tracking (version 1.0.0) [dataset] (2019). http://doi.org/10.5281/zenodo.3255102
17. Terzopoulos, D., Waters, K.: Analysis and synthesis of facial image sequences using physical and anatomical models. IEEE Trans. Pattern Anal. Mach. Intell. **15**(6), 569–579 (1993)
18. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: real-time face capture and reenactment of RGB videos. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition, pp. 2387–2395 (2016)
19. Waters, K.: A muscle model for animation three-dimensional facial expression. In: Proceedings ACM SIGGRAPH 1987 Conference, vol. 21, no. 4, pp. 17–24 (1987)
20. Weise, T., Bouaziz, S., Li, H., Pauly, M.: Realtime performance-based facial animation. ACM Trans. Graph. **30**(4), 1–10 (2011)
21. Wu, T., Hung, A., Mithraratne, K.: Generating facial expressions using an anatomically accurate biomechanical model. IEEE Trans. Visual Comput. Graph. **20**(11), 1519–1529 (2014)
22. Xu, F., Chai, J., Liu, Y., Tong, X.: Controllable high-fidelity facial performance transfer. ACM Trans. Graph. **33**(4), 1–11 (2014)
23. Zeng, X., Dwarakanath, S., Lu, W., Nakada, M., Terzopoulos, D.: Neuromuscular control of the face-head-neck biomechanical complex with learning-based expression transfer from images and videos. arXiv Preprint (2021)
24. Zhang, J., Chen, K., Zheng, J.: Facial expression retargeting from human to avatar made easy. IEEE Tran. Visualizat. Comput. Graph. (2020)