Blendshape-augmented Facial Action Units Detection

Zijun Cui
Rensselaer Polytechnic Institute
cuiz3@rpi.edu

Qiang Ji
Rensselaer Polytechnic Institute
jiq@rpi.edu

Abstract

Action units (AUs) represent descriptions of the facial muscle activation. Given the activation of facial muscles, the facial skin is deformed by following the underlying facial muscle mechanics. This paper proposes to apply the generic knowledge from the principles of the facial anatomy and 3D projection models. And the generic knowledge is combined with the data information through the proposed joint top-down and bottom-up framework. Specifically, we first propose a top-down AU model where 3D AU blendshapes and the 3D projection models are applied as the representation of the generic knowledge. A data-driven bottom-up AU model is constructed from the observed 2D images. Finally, an integration model is proposed to perform final AU predictions by combining the estimations from both the top-down model and the bottom-up model. The final AU predictions are thus based on both generic knowledge and data information. Evaluations on a benchmark dataset show that by leveraging the generic knowledge, the proposed top-down and bottom-up framework can improve the AU detection performance.

1 Introduction

Conventional methods infer AU activation from input images [8, 4, 2]. These conventional methods work in a bottom-up manner which are data-driven, and may not generalize well to different datasets. Some existing AU detection models explore the usage of the prior knowledge, and such prior knowledge is mainly about the AU relationships. The AU relationships can be captured by structure models such as tree [20, 7] or graphical models [18, 19, 14, 16, 6, 12]. And most of the existing works learn AU relationships from training data as prior knowledge and the knowledge is thus data-based. To the best of our knowledge, only limited papers tried to explore the generic domain knowledge, i.e., data-free knowledge based on both the underlying facial anatomy and 3D projection models. The current data-free knowledge mainly comes from field experience [20, 10, 24, 25]. As AU activation is caused by facial muscle movement, we propose to exploit the generic knowledge from both the facial anatomic and physical principles of facial muscles for AU detection.

Usually, the prior knowledge model works in two different ways. In one way, the prior knowledge is used to model structural outputs(e.g., [19]), where the exploited prior knowledge still works in a bottom-up framework. On the other hand, the prior knowledge works as a regularization for the learning of the conventional bottom-up models(e.g., [7]). Different from existing approaches, we propose a joint top-down and bottom-up framework for AU detection.

2 Proposed Method

Specifically, the proposed framework includes a top-down projection module and a bottom-up estimation module as shown in Figure 1. Given a 2D facial image, the top-down model can predict the active levels of AUs based on both facial anatomy-driven 3D AU blendshapes and 3D projection models. A bottom-up deep model is learned given the 2D facial images with corresponding AU annotations. In the end, we produce the final AU predictions by combining the top-down projections with the bottom-up estimates. The overview of our proposed approach is shown in Figure 1.

Workshop on Differentiable Vision, Graphics, and Physics in Machine Learning at NeurIPS 2020.

2.1 Top-down AU Projection

We propose the top-down AU model to predict AU activation based on generic knowledge on the underlying facial muscle mechanics as well as the 3D projection models. The top-down AU model contains two components: 3D blendshape engine based on facial anatomy and 3D facial landmark estimator with through inverse 3D projection model. We firstly introduce the facial muscle and the modelling of its underlying mechanics. We then introduce the proposed 3D blendshape engine and the 3D facial landmark estimator.

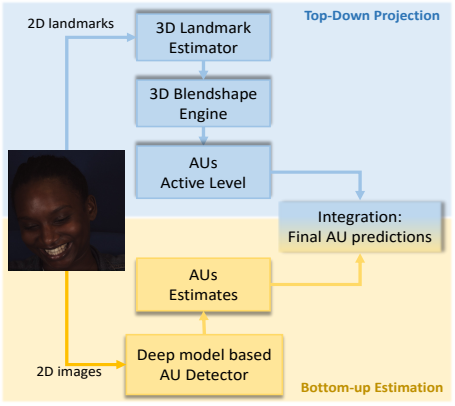

Figure 1: Overview of the proposed method. The top-down model combines anatomic knowledge with projection models to predict AUs, and the bottom-up model performs data-based AU prediction in 2D. The top-down and bottom-up prediction are combined through the integration model.

Facial muscle and the modeling of its mechanics Human face is a soft tissue organ complex, with a large investing network of musculature. It consists of several anatomically distinct layers: the skin, subcutis, fascia and muscles. It is evident that the skin, being supported by bone and multiple layers of muscles, exhibits different facial expressions, activated by the underlying facial muscles. Different muscle activation can result in different facial expressions. For example, the *corrugator supercilii* muscle is used to compress the skin between the eyebrows, which are drawn downwards and inwards to create such expressions as anger and disgust.

Modelling the mechanics for facial expression generation is important for this research. And such modelling can be very challenging as the facial soft tissues are structurally complex and exhibit nonlinear constitutive behaviour. The constitutive law is usually carefully defined to obtain the stress tensor for each facial point based on the strain at that point [11]. In order to simulate dynamic facial deformations, Newton's second law of motion is applied. Given this understanding, the human face is usually represented by the conforming tetrahedralized flesh mesh [5, 15, 1] in computer graphics. The finite element method [15] or the finite volume method [17, 5, 1] are widely employed to compute the deformation for each mesh vertex. Though performing realistic simulations, existing facial muscle models in computer graphics are very computationally challenging. To evaluate muscle contribution from anatomically accurate geometries while maintaining the computational complexity at a tractable level, we instead consider 3D facial blendshapes, which implicitly capture the principles of the face muscles and their motions.

3D blendshape engine Facial blendshapes are widely used for realistic animation due to its simplicity and effectiveness [9, 22, 13]. Facial blendshaps are a set of 3D surface morphs where each morph corresponds to a specific semantic meaning, such as an expression [3] or an AU[21]. Different morphs have different vertex positions capturing their deformations and the topologies of all the morphs capture the holistic facial expression. An expression can be synthesized through the linear combination of particular blendshapes together with the neutral face. In this paper, we apply the 3D AU blendshapes provided by the FaceGen[1], which provides morphs at three different levels: phonemes, action units, and expressions. Given the AU blendshapes, we define our 3D AU basis, where each basis corresponds to one AU action together with a neutral face, as shown in Figure 2(a). And we manually label the 3D facial landmarks on the surface mesh to represent facial skin motions. Particularly, we collected K AU bases $\boldsymbol{b}_k \in R^{M\times 3}$ with $k = 1, .., K$ and M being the number of landmarks. Given 3D landmarks $\bar{\boldsymbol{p}}$ for the neutral face, we have 3D landmarks $\boldsymbol{p}$ for an expressed face as

$$\boldsymbol{p} = \bar{\boldsymbol{p}} + \sum_{k=1}^{K} a_k \boldsymbol{b}_k \tag{1}$$

where $\boldsymbol{a} = \{a_k\}_{k=1}^{K}$ represent the active level for each AU and $a_k \geq 0$.

Instead of generating an expressed face in a forward way as shown in Eq. 1, we are aimed at the inverse process, i.e., inferring AU activation given observed facial skin distortions. To achieve this, we

[1] https://facegen.com/

introduce the 3D blendshape engine as shown in Figure 2. Specifically, given observed 3D landmarks

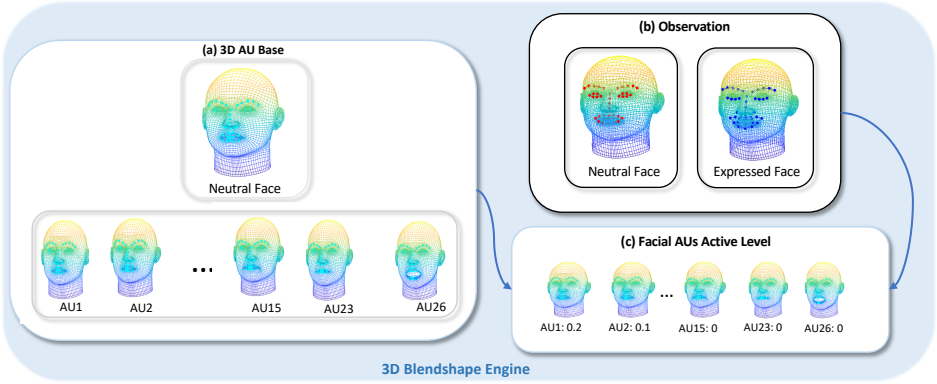

Figure 2: 3D Blendshape Engine

$\boldsymbol{p} \in R^{M\times 3}$ for an expressed face, the facial skin distortion of the expressed face is measured by the surface landmark displacements from their neutral positions. And the AUs' active levels of the expressed face is then automatically determined through the 3D blendshape engine as

$$\boldsymbol{a}^* = \arg\min_{a_k \geq 0} ||\boldsymbol{p} - \bar{\boldsymbol{p}} - \sum_{k=1}^{K} a_k \boldsymbol{b}_k|| \tag{2}$$

with $\boldsymbol{a^*} = [a_1^*, a_2^*, ..., a_K^*]$ and $a_k^* \geq 0$.

3D facial landmark estimator To use Eq. 2 for top-down AU prediction given only 2D images, we need first estimate 3D facial landmark positions from the 2D images. Here, we exploit the projective model in computer vision to infer 3D landmarks $\boldsymbol{p}$ from individual-specific 2D landmarks $\boldsymbol{q}$ in the image. Given an individual-specific 2D landmark point in the image frame $\boldsymbol{q} = (c, r)$ and the intrinsic camera matrix[2] W, we obtain its corresponding 3D individual-specific landmark point $\boldsymbol{p}^s = (x, y, z)$ in camera frame as

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \frac{1}{z} W^{-1} \begin{bmatrix} c \\ r \\ 1 \end{bmatrix} \tag{3}$$

We assume weak perspective projection, and z coordinate is Eq. 3 is replaced by $\bar{z}$, the average 3D facial landmark points to the camera. In practice, we treat $\bar{z}$ as a hyper-parameter to tune. Consider the reconstructed 3D landmarks for neutral face $\bar{\boldsymbol{p}}^s$ and for expressed face $\boldsymbol{p}^s$, the differences between $\bar{\boldsymbol{p}}^s$ and $\boldsymbol{p}^s$ are caused by both head pose and facial muscle actions. To accurately estimate facial muscle actions, we need to remove effects from rigid head poses. To achieve this, we select a subset of landmarks that always move rigidly, such as the landmark on the nose tip, and we solve for a rotation matrix and a translation vector via

$$R^A, T^A = \arg\min_{R,T} ||\bar{\boldsymbol{p}}^s - R\boldsymbol{p}^s - T|| + ||R^T\bar{\boldsymbol{p}}^s - R^T T - \boldsymbol{p}^s|| \tag{4}$$

Given estimated R^A and T^A, we have the aligned 3D landmarks for each $\boldsymbol{p}^s$ as $\boldsymbol{p}^{s+a} = R^A\boldsymbol{p}^s + T^A$. As the recovered 3D individual-specific landmarks $\boldsymbol{p}^{s+a}$ is specified with respect to the camera frame, we need to register the landmarks $\boldsymbol{p}^{s+a}$ to the blendshape frame, which is used for 3D AU bases by the 3D blendshape engine. To achieve this, we perform a rigid transformation, i.e., $\boldsymbol{p} = R^E\boldsymbol{p}^{s+a} + T^E$ where R^E and T^E can be estimated given a set of corresponding rigid facial landmark points in both the camera frame and the engine frame. We assume that different subjects' landmarks can be fitted into our AU bases through the rigid transformation with trivial errors.

[2] W can be recovered through a set of corresponding 2D points and 3D points.

2.2 Bottom-up AU Estimation

We introduce the bottom-up model f_ψ that performs data-drived facial AU estimation from 2D images. Given input images $\{\boldsymbol{x}_n\}_{n=1}^N$ and the one-hot encoding of AU labels $\{\boldsymbol{y}_n^{GT}\}_{n=1}^N = \{\boldsymbol{y}_{1n}^{GT}, \boldsymbol{y}_{2n}^{GT}, ..., \boldsymbol{y}_{Mn}^{GT}\}_{n=1}^N$, where M is the total number of AUs and N is the total number of training samples, the AU detector f_ψ jointly detects M AUs and produces the probability distributions of M AUs. The model ψ is obtained as,

$$\psi^* = \arg\min_{\psi} \sum_{n=1}^{N} \sum_{m=1}^{M} l(\boldsymbol{y}_{mn}^{GT}, f_\psi^m(\boldsymbol{x}_n)) \tag{5}$$

where l is the cross entropy loss and $f_\psi^m(\boldsymbol{x}_n)$ is the predicted probability of m^{th} AU.

2.3 Top-down and Bottom-up Integration

We now introduce the integration model g_ϕ to combine the top-down AU projections with the bottom-up AU estimations. Given each image $\boldsymbol{x}_n$, the integration model takes the AU active level $\boldsymbol{a}_n^*$ from the top-down model and the estimated AU probabilities $f_\psi(\boldsymbol{x}_n)$ from the bottom-up model as input. Given the training images $\{\boldsymbol{x}_n\}_{n=1}^N$ and the AU annotations $\{\boldsymbol{y}_n^{GT}\}_{n=1}^N$, we obtain ϕ as:

$$\phi^* = \arg\min_{\phi} \sum_{n=1}^{N} \sum_{m=1}^{M} l(\boldsymbol{y}_{mn}^{GT}, g_\phi^m(f_\psi^m(\boldsymbol{x}_n), a_{mn}^*)) \tag{6}$$

where l is the cross entropy loss and $g_\phi^m(f_\psi^m(\boldsymbol{x}_n), a_{mn}^*)$ gives the probability of m^{th} AU.

3 Experiments

We apply BP4D-Spontaneous database[23] for the experiment. BP4D is a spontaneous database containing 2D and 3D expression data for 41 subjects. We collect 732 apex frames in total and employ 3-fold subject-independent cross-validation experiment. For the top-down model, intrinsic camera parameter W, rotation matrices R^A, R and translation vectors T^A, T are estimated given 2D/3D landmarks provided in the BP4D. For the bottom-up model, we employ the shallow three layer CNN model. The kernel size of each layer is 5x5, 5x5, 3x3 respectively. The integration model consists of three fully connected layers. We evaluate our proposed method with F1-score. We perform an ablation study of the propose model, where we show the performance of top-down model, bottom-up model and the integration model. Results are shown in Table 1. We can see that top-down model can

Table 1: Ablation study on BP4D (Measurement: F1-score)

AU Model	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Ave.
Top-down Model	47.6	**48.7**	**55.6**	73.7	86.4	78.3	76.1	64.6	37.5	70.9	55.2	44.7	61.6
Bottom-up Model	46.5	33.8	48.3	77.6	86.0	85.7	82.8	64.2	49.8	75.6	56.7	56.1	63.6
Integration Model	**48.6**	34.9	48.7	**78.4**	**87.4**	**86.3**	**83.2**	**66.2**	**51.8**	**76.7**	**57.7**	**58.6**	**64.9**

achieve comparable performance compared to the bottom-up model. By combining the top-down model with the bottom-up model, the integration model achieves better performance on most of the AUs, and achieves improved performance on average over all the AUs.

4 Discussion

In this work, we explore the usage of the generic knowledge from both the underlying facial muscle mechanism and projective models for AU detection. Particularly, we propose a top-down AU model where 3D AU blendshapes based on facial anatomy are applied as the representation of the generic knowledge on the physical principles of facial actions. We then employ the principles of projective models to infer 3D facial motion from their 2D images. The generic knowledge is then combined with data information through the proposed joint top-down and bottom-up framework. Our evaluation on the BP4D dataset shows that through the proposed joint top-down and bottom-up inference model, the performance of the AU prediction can be improved by leveraging the generic knowledge.

Acknowledgement

The work described in this paper is supported in part by a DARPA grant FA8750-17-2-0132, and in part by the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons).

References

[1] Michael Bao, Matthew Cong, Stephane Grabli, and Ronald Fedkiw. High-quality face capture using anatomical muscles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[2] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, 2016.

[3] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013.

[4] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfermachine for personalized facial action unit detection. In *CVPR*, 2013.

[5] Matthew Cong, Michael Bao, Jane L E, Kiran S Bhat, and Ronald Fedkiw. Fully automatic generation of anatomical face simulation models. In *Proceedings of the 14th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 175–183, 2015.

[6] C. Corneanu, M. Madadi, and S. Escalera. Deep structure inference network for facial action unit recognition. In *ECCV*, 2019.

[7] S. Kaltwang, S. Todorovic, and M. Pantic. Latent trees for estimating intensity of facial action units. In *CVPR*, 2015.

[8] S. Koelstra, M. Pantic, and I. Y. Patras. A dynamic texture-basedapproach to recognition of facial actions and their temporal models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32:1940–1954, 2010.

[9] Yeara Kozlov, Derek Bradley, Moritz Bächer, Bernhard Thomaszewski, Thabo Beeler, and Markus Gross. Enriching facial blendshape rigs with physical simulation. In *Computer Graphics Forum*, volume 36, pages 75–84. Wiley Online Library, 2017.

[10] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In *AAAI*, 2019.

[11] Andrew Nealen, Matthias Müller, Richard Keiser, Eddy Boxerman, and Mark Carlson. Physically based deformable models in computer graphics. In *Computer graphics forum*, volume 25, pages 809–836. Wiley Online Library, 2006.

[12] X. Niu, H. Han, S. Yang, and S. Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *CVPR*, 2019.

[13] M Romeo and SC Schvartzman. Data-driven facial simulation. In *Computer Graphics Forum*, volume 39, pages 513–526. Wiley Online Library, 2020.

[14] G. Sandbach, S. Zafeiriou, and M. Pantic. Markov random field structures for facial action unit intensity estimation. In *ICCV*, 2013.

[15] Eftychios Sifakis, Igor Neverov, and Ronald Fedkiw. Automatic determination of facial muscle activations from sparse motion capture marker data. In *ACM SIGGRAPH 2005 Papers*, pages 417–425. 2005.

[16] Deep structured learning for facial action unit intensity estimation. Walecki, r. and rudovic, o. and pavlovic, v. and schuller and b., pantic, m. In *CVPR*, 2017.

[17] J. Teran, S. Blemker, V. Ng Thow Hing, and R. Fedkiw. Finite volume methods for the simulation of skeletal muscle. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '03, pages 68–74, Aire-la-Ville, Switzerland, Switzerland, 2003. Eurographics Association.

[18] Yan Tong and Qiang Ji. Learning bayesian networks with qualitative constraints. In *CVPR*, 2008.

[19] Z. Wang, Y. Li, S.Wang, and Q. Ji. Capturing global semantic relationships for facial action unit recognition. In *ICCV*, 2013.

[20] J. Chen Y. Li, Y. Zhao, and Q. Ji. Data-free prior model for facial action unit recognition. 2013.

[21] Yanfu Yan, Ke Lu, Jian Xue, Pengcheng Gao, and Jiayi Lyu. Feafa: A well-annotated dataset for facial expression analysis and 3d facial animation. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 96–101. IEEE, 2019.

[22] Xiangyu You, Feng Tian, and Wen Tang. Highly efficient facial blendshape animation with analytical dynamic deformations. *Multimedia Tools and Applications*, 78(18):25569–25590, 2019.

[23] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *FG workshop*, 2013.

[24] Yong Zhang, Weiming Dong, Baogang Hu, and Qiang Ji. Classifier learning with prior probabilities for facial action unit recognition. In *CVPR*, 2018.

[25] Yong Zhang, Weiming Dong, Baogang Hu, and Qiang Ji. Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In *CVPR*, 2018.