10.2352/EI.2022.34.8.IMAGE-255

VR Facial Expression Tracking via Action Unit Intensity Regression Model

Xiaoyu Ji[1] , *Justin Yang*[1] , *Jishang Wei*[2] , *Yvonne Huang*[2] , *Qian Lin*[2] , *Jan P Allebach*[1] , *Fengqing Zhu*[1] ;
1 Elmore School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, U.S.A;
2 HP Labs, Palo Alto, CA 94304, U.S.A.

Abstract

Virtual Reality (VR) Head-Mounted Displays (HMDs), also known as VR headsets, are powerful devices that provide interaction between people and the virtual 3D world generated by a computer. For an immersive VR experience, the realistic facial animation of the participant is crucial. However, facial expression tracking has been one of the major challenges of facial animation. Existing face tracking methods often rely on a statistical model of the entire face, which is not feasible as occlusions arising from HMDs are inevitable.

In this paper, we provide an overview of the current state of VR facial expression tracking and discuss bottlenecks for VR expression re-targeting. We introduce a baseline method for expression tracking from single view, partially occluded facial infrared (IR) images, which are captured by the HP reverb G2 VR headset camera. The experiment shows good visual prediction results for mouth region expressions from a single person.

Introduction

Virtual Reality (VR) technology overcomes the limitations of 2D animation and realizes 3D animation with functionalities to enhance the spaciousness of objects. 3D Facial animation combined with 3D rendering technology can generate a high-quality real-time animation video which is now widely used in applications such as human-computer interaction (HCI), cartooning, and virtual communications [1]. For the facial animation, it is critical to keep the facial expressions natural, adhere to human facial anatomy and correctly match the muscle movement of the person. Facial expression tracking is the process of extracting facial expression features from data captured by HMDs. The design of HMDs influences the type and quality of the collected data, while the selection of facial expression features affects the robustness of the predictions and the computational complexity.

Head-Mounted Displays (HMDs) include special VR cameras to capture facial expressions located around the eyes and mouth. Although different VR products each have their own settings for the cameras, they only capture partial face images due to the occlusion caused by the HMD device itself [2]–[4]. Hence, most real-world image or video expression recognition methods requiring full face input will not work, as partially occluded face data has limited information. Research about occluded facial expression recognition has also been conducted, including real occlusion and synthetic occlusion. However, the occluded region is smaller compared to images captured by the HMD device and most existing works focus on expression classification [5], [6]. Customized methods are developed for facial expression tracking based on specific HMD devices.

The three main types of headset-mounted cameras (HMCs) are IR cameras, RGB cameras, and RGBD cameras. The advantage of IR cameras is that the IR image has a more stable performance under different lighting conditions than RGB cameras. However, the IR images are single-channel images without color information. The depth camera is another type of camera to add distance information for each pixel in the image, with additional computation needed to calculate depth. Apart from the type of camera, the number of cameras and their position and angle are also important factors for facial expression tracking. Hence, the design of HMDs is a trade-off between memory, computation cost, and performance.

A widely used tool for coding facial expressions known as the Facial Action Coding System (FACS) was introduced in [7]. It breaks down facial expressions into individual components of muscle movements, called Action Units (AUs). One action unit (AU) represents a facial mesh deformation from a neutral face to a specific semantically meaningful expression, and the intensity of the AU indicates the extent of the expression. In total, 46 AUs are decomposed from facial behaviors in the FACs. There are complicated correlations among different AUs due to the diversity of facial expressions. Meanwhile, restraint also exists for the combination of AUs as the AU space is sparse, and any random combination of AUs may not produce realistic facial expressions [8]. For real-life expression tracking, we need to ensure the reconstructed facial expressions are realistic. Therefore, our work adopts this framework and focuses on realistic facial expressions prediction of related AUs.

In this paper, we propose a baseline method for facial expression tracking from IR images captured by the HMC of HP reverb G2 HMDs. We aim to predict facial action unit intensities to estimate the facial muscle movement of the wearer. We select seven expressions including mouth smile, smile left, smile right, frown, pucker, move left and move right. We use a regression network to predict facial action unit intensities and compare the predicted facial expressions with the ground truth expressions. The prediction results are also evaluated quantitatively using Mean Absolute Error (MAE) and Pearson's correlation coefficient (PCC).

In summary, our main contributions include the following. First, we provide an overview of current VR expression retargeting works and discuss existing challenges. Second, we propose a facial expression tracking baseline method for the challenging case where faces are partially occluded by the HMDs. Third, our method is evaluated on a single-user dataset both quantitatively and qualitatively.

In the following sections, we will first briefly review related expression tracking works. Then,we introduce our baseline structure and the regression model. Finally, we show the detailed result of the performance experiment based on the proposed model, including quality comparison and quantity evaluation.

Related Work

In this section, we review two related topics: Head-Mounted Displays (HMDs) and 3D modeling. We will introduce the design details of different HMDs inputs and 3D face modeling from these inputs.

Head-Mounted Displays (HMDs) Inputs

Expression tracking relies on the type of data captured by the HMDs. Although most research focuses on image or video input, the electromyography (EMG) signal is another input form that supports a good prediction of facial expressions. In the following subsections, we discuss facial expression tracking works using either HMC or EMG input.

Headset Mounted Cameras (HMCs)

Image or video input of HMCs is the most common method to detect facial muscle movement. Because the inference process requires an efficient expression tracking pipeline, the number and computational cost of HMCs need to be small. A single-view IR camera is the most lightweight choice for HMCs but it provides limited information about the facial features when only a single view is used. Several approaches have been proposed to compensate for the limitation and occlusion of HMC videos or images. For example, Wei *et al.* proposed to enrich the dataset by designing a training device that is different from the testing device. They generated an expression tracking model with 9 IR HMCs for training and 3 IR HMCs for real-time testing [3]. In addition, the training device collects two more eye region views and four more mouth region views. While research in [4] used depth cameras and included 3 RGBD single view cameras. The design of HMDs for testing and training is also different. Here, full-face four-channel videos were used as training data and partially occluded face videos were used in testing. This method has comprehensive information to build an expression tracking model. Another RGB-based HMC design was proposed by the same research group in 2016 [9], where a monocular RGB HMC is placed in front of the mouth to track speech and mouth movement while two IR cameras are used to track eye region motions. The combination of speech and eye animation shows good prediction results on specific users with selected expressions.

Electromyography (EMG)

Integrated EMG sensors are more lightweight and ergonomically comfortable compared to HMCs, while the disadvantage comes from direct contact with the skin is required. With electrodes directly placed around facial muscles, the EMG signal has a large signal-to-noise ratio [10]. Lou *et al.* proposed an approach to recover facial action unit intensities from EMG signals in [11]. FACETEQ hardware [12] can be placed on the emotionally salient facial part (ESFP), which is around the eyes and nose region. Consequently, the sensors cover the forehead, cheek, and outer-eye-corner regions. Seven action units could be predicted, including eyebrow movement, eye, and mouth openness, cheek raising, lip corner puller, and lip pucker.

3D Modeling

Reconstructing 3D facial expressions from a single 2D image is an active research topic in image processing and computer graphics. In order to construct a fully-rigged 3D mesh, it is necessary to obtain a parameterized representation for 3D face synthesis. Given an input 2D image, one can obtain a representation that contains information related to face shape and appearance. Here, we introduce several commonly used 3D modeling methods, including 3D morphable modeling, active appearance modeling, deep appearance modeling, landmark fitting, and AU-related approaches.

3D Morphable Model (3DMM)

Different methods have been proposed to track facial expressions from monocular inputs. Some focus on using statistical models for facial texture and shape. A popular method is the 3D Morphable Model (3DMM) [13]. 3DMM is a statistical model that builds face shape and appearance based on the facial image data from 200 people. It parameterizes the human face into high dimensional subspaces to represent the 3D facial mesh in terms of face shape and texture. The authors proposed this statistical model as a parametric linear subspace with point-to-point correspondence that enables 3D face reconstruction from 2D images. Given a 2D face image, the method finds a point in this high dimensional subspace that represents a similar face. This task can be achieved by regressing the 3DMM face shape parameters using the 2D input image [14]–[17].

Active Appearance Modeling (AAM)

Active Appearance Modeling (AAM) aims to match a given image to a statistical shape model that parameterizes the shape and appearance of an object. For facial expression tracking, AAM is able to successfully disentangle the shape and appearance using Principle Component Analysis (PCA) [18]. Similar to 3DMM-based approaches, given a 2D image, AAM-based methods focus on fitting AAM parameters accurately to reconstruct a 3D face with an expression [19]. However, AAM uses the entire face region to localize facial landmarks in order to establish correspondence between the training 3D face mesh and the input 2D image for more accurate estimation. In this case, the performance of such approaches would be restricted when the face is partially occluded, for example when wearing a VR headset.

Deep Appearance Modeling (DAM)

Recently, with the emerging interest in the intersection between deep learning and computer graphics, Deep Appearance Modeling (DAM) has shown success for modeling human faces in 3D [20]. Using images captured by 40 cameras from different viewing angles, a Variational Autoencoder (VAE) [21] is used to model a data-driven avatar that learns a joint latent representation of face geometry and appearance. With view-point conditioning, VAE is able to disentangle the viewpoint-specific information from the latent representation of face geometry and appearance. By manipulating the latent variable, it is able to perform controllable synthesis of the facial expression without modifying the facial geometry and identity of the given avatar.

Landmark Fitting Approach

Incorporating landmark detection into 3D face modeling can add constraints for synthesis [22], [23]. Given a set of landmark points with the correspondence in the 3D face model, some techniques fit the 3D surface with the detected landmarks [24]. These approaches are accurate in terms of keeping the facial expression. However, it is unclear how well they perform if the face is partially occluded, especially when the occlusion affects the detection of facial landmarks.

Action unit (AU) approach

Research about the facial action unit is initially a classification problem to detect whether the action unit has been activated. For instance, Gwen *et al.* proposed a toolbox for facial expression recognition and action unit intensity estimation using Gabor filters and support vector machine (SVM) [25]. As more detailed datasets such as BP4D [26] and CK+ [27] became available including annotated intensity of action units, more researchers have been exploring AU intensity estimation. The regression problem associated with AU intensity estimation is typically based on six levels of intensities, ranging from 0 to 5. A heatmap-based hourglass network is proposed as the estimation model which is jointly combined with landmark detection in [28], [29]. To support high precision AU intensity estimation, a dataset with higher precision annotations was proposed which included two decimal points within the range 0 to 1 [30].

Method

In this section, we illustrate the baseline structure of VR expression tracking for HP Reverb G2 HMDs[1]. As shown in Figure 1, the input to the regression model is an IR image, and the output expression is reenacted on the neutral 3D model of the target avatar using the predicted AU intensities from the regression model. The phrase reenact here means transferring a facial expression from a source face to a target face while preserving the appearance and the identity of the target face. The quantity evaluation uses Mean Absolute Error (MAE) and Pearson's Correlation Coefficient (PCC) between the ground truth and predicted AU intensities. The quality evaluation assesses the difference between the expression of the input image and that of the reenacted image visually.

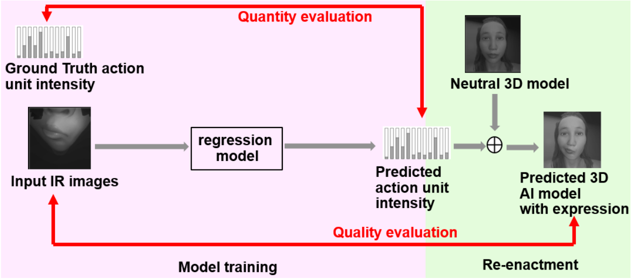
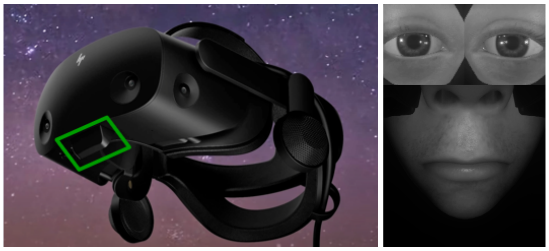


Figure 1: Overview of the baseline structure. The left side is the model training part, and the right side is for expression reenactment.

The goal for the model training part is to accurately predict AU intensity given only a single IR image from an HMC. We will first demonstrate the process of capturing the input IR camera images for data collection. As shown in Figure 2a, the HP reverb headset has a total of three IR cameras installed on the device. One camera is placed in front of the nose as marked by the green box and the other two are in front of the eyes. The eye cameras focus on only the eyeball region, and most of the eyebrow is excluded from the camera view. The eyeball region captured from the eye camera may deviate from the center of the image or video as the participant moves when wearing the HMD. The mouth IR camera captures the mouth region and part of the nose region of the participant. This view includes the area under the nose except the partial upper cheek regions. Figure 2b shows an example of the combined IR image of the eyes and mouth regions.



(a) (b)

Figure 2: (a) The HP Reverb G2 HMD. (b) IR images from HP reverb camera view. The eyes and mouth images are separately rendered and combined as one image. The size of eye images is 200 × 200, and that of mouth image is 400 × 400.

Data imbalance is a major issue while collecting facial expressions for model training. The difficulty comes from ensuring the proportions of different expressions are similar in order to keep the dataset balanced. An imbalanced dataset typically contains more neutral faces, which results in a bias toward the prediction of a neutral face.

Since a much larger number of AUs are involved in eye and mouth combined expressions than only mouth expressions, currently, we focus on designing models to track mouth movements in terms of AU intensity prediction. We select 20 AUs corresponding to the mouth region as the regression model output. They are associated with 7 mouth expressions including smile, smile left, smile right, frown, pucker, mouth move left, and mouth move right. These expressions are used to implement expression tracking for a single person.

The goal of our model is to track facial expressions using a regression model. The selection of the regression model is based on the following considerations. For potential deployment on the device, we need to consider the model size and device compatibility. We also need to consider the prediction accuracy of facial expressions. As a result, we evaluated several deep neural network architectures including Inception-ResNet-V1 [31], EffNet [32], ShuffleNet [33], VGG-11 [34], MobileNetV2 [35] and MobileNetV3 [36] in an initial experiment. This initial evaluation allowed us to test the accuracy of expression prediction using the regression model trained on one person's expression data.

Table 1 shows the detailed configuration for each network structure. Some networks use pre-trained model weights from ImageNet [37], which is indicated in the second column. The epoch number column shows the speed of convergence of each network. Each network is set to train for a maximum of 400 epochs, and

[1] https://www.hp.com/us-en/vr/reverb-g2-vr-headset.html

the number in the table represents the actual epoch number when the training converges. Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC) are used as the metrics to compare the mean result of 7 expressions. From the comparison result in Table 1, EffNet, VGG-11, and Inception-ResNet-V1 have a relatively large model size which does not fit our design requirement. Among the network structures with large model sizes, VGG-11 performs the best, and EffNet has a comparable result. While comparing the mean MAE loss and mean PCC among the network structures with a small model size, the result of MobileNetV3 has about the same result as EffNet. Although MobileNetV2 and ShuffleNet have smaller model sizes, MobileNetV3 is selected as the backbone network for our baseline due to better prediction results.

Table 1: Backbone model structure comparison

Network	pre-trained (Y/N)	# of Epochs	MAE	PCC	Model Size (M)
EffNet [32]	N	167	0.63	0.995	1,878
VGG-11 [34]	Y	366	0.62	0.994	1,487
Inception-ResNet-V1 [31]	N	349	0.92	0.984	272
MobileNetV3 [36]	**Y**	**370**	**0.64**	**0.990**	**53**
MobileNetV2 [35]	Y	388	0.87	0.986	29
ShuffleNet [33]	N	377	0.84	0.987	15

Mean Squared Error (MSE) (Eq. (1)) is used as the loss function for each model, where G represents ground truth AU intensities and P represents prediction AU intensities. We aim to predict a total of 20 AUs with our model.

$$\text{Loss} = \frac{\text{mean}(\mid G - P \mid^2)}{\text{batch size} \times 20} \tag{1}$$

Data augmentation is also used during the model training. To maximize the diversity of the dataset and balance the left and right expressions, we randomly picked half of the images and flipped the images from left to right. Meanwhile, the left AU intensities are also exchanged with the right AU intensities. For example, the mouth smile left intensity is exchanged with the mouth smile right intensity.

Experiments

In this section, we describe the experimental results of our proposed VR expression tracking method. Based on the MobileNetV3 structure, we trained the model in the PyTorch framework. The regression model is trained on 7 expressions from one person, and the testing expressions are separately collected from the same person. The training parameters are set to the following: learning rate at 0.0001, weight decay at 0.0001, batch size at 64, and epoch number at 135. The learning rate scheduler is StepLR with a step size of 47 and a weight of 0.9. The evaluation metrics include Mean Squared Error (MSE), Mean Absolute Error (MAE), and Pearson's Correlation Coefficient (PCC).

To evaluate the visual quality of predicted facial expressions, we transfer the predicted expression to a target virtual avatar. The process of transferring source expression to the target character is called facial reenactment. Figure 3 and Figure 4 show the visual result of predicted expressions after reenactment. We evaluate the results by comparing the predicted expression with the ground truth expression and assess whether the two expressions have the same semantic meaning. Figure 3 shows the output expression with smile-related input expressions. The first row indicates a good prediction of mouth openness and intense smile. The second row shows two smile left expressions (to the left side of the person) with a slight difference. From the output, we observe different muscle movements around the right mouth corner. In Figure 4, the first row shows results of relaxed and stretched mouth corner prediction. The second row shows the visually correct prediction for mouth pucker and mouth move right expressions.

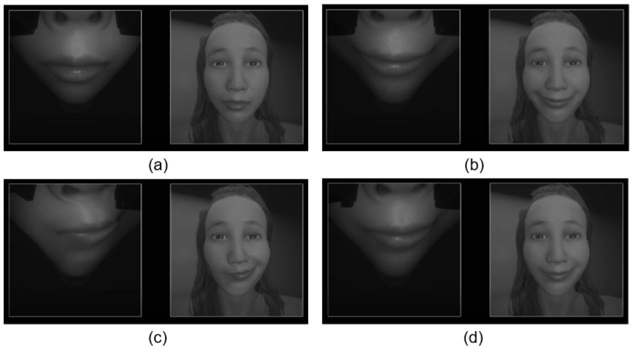


Figure 3: Predicted smile expressions. For each result image, the left partial face is the input mouth IR image, the right image is the predicted expression. Subfigures (a) - (d) correspond to the results of four expressions: neutral face, smile, smile left, and slight smile left.

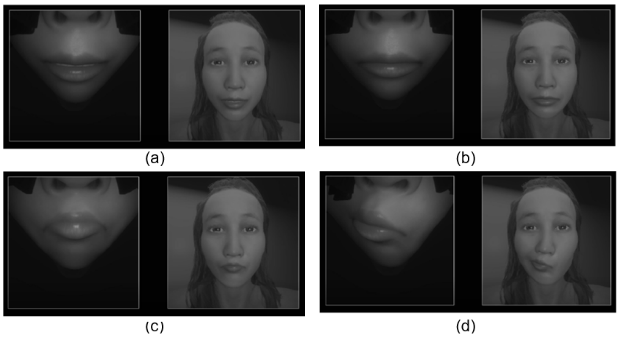
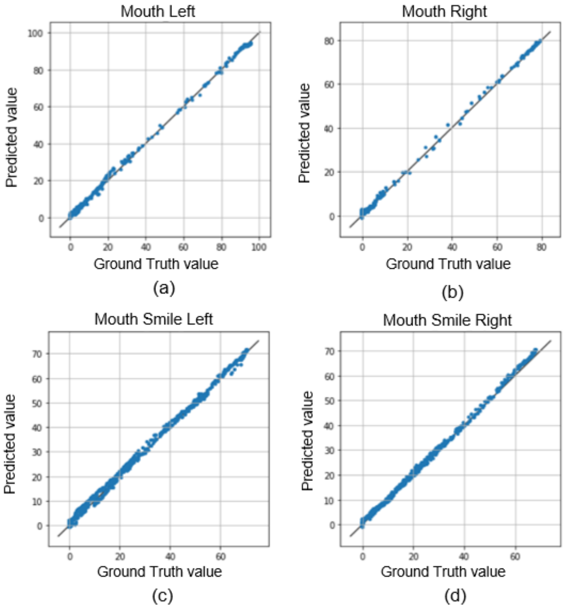


Figure 4: Other predicted expressions. For each result image, the left partial face is the input mouth IR image, the right image is the predicted expression. Subfigures (a) - (d) correspond to the results of four expressions: mouth open, slight frown, pucker, mouth move right.

We also performed a quantitative analysis based on the evaluation metrics: MSE, MAE, and PCC. Figure 5 shows the scatter plots of predicted values. We list the results of four AUs: mouth move left, mouth move right, mouth smile left, and mouth smile right. The diagonal line (shown in black) indicates an exact match between the ground truth and our prediction. The closer the predicted values are to the diagonal line, the better the prediction. The PCC values for testing images are all higher than 0.96 except for several neutral face images as shown in Figure 6. The cases with small PCC values all have ground-truth AU intensities equal to zero for all 20 AUs, which result in a large variance of PCC value for the neutral face. However, the large variance does not necessarily indicate a bad prediction. In these cases, the predicted AU intensities for the neutral face are still very small, so the expressions are correctly predicted as a neutral face from visual assessment. For expressions other than the neutral face, the zoom-in plot shows high correlation values. The mean PCC value is 0.992 and the median PCC value is 0.995.



(a)

(b)

(c)

(d)

Figure 5: Scatter plots of predicted AU intensities vs. ground truth. Subfigures (a) - (d) correspond to the results of four expressions: mouth move left, mouth move right, mouth smile left, mouth smile right.

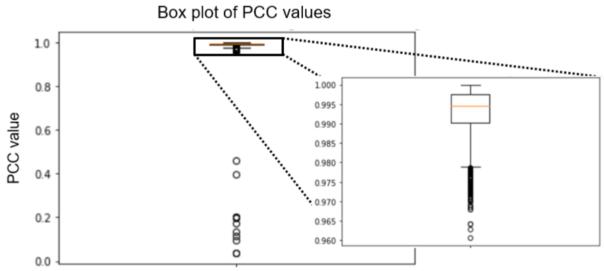


Figure 6: Box plot of the distribution of PCC values of the entire testing dataset. Each PCC value is the Pearson's correlation coefficient between the ground truth AU intensities and the predicted AU intensities of a testing image input. A zoom-in plot shows the region around the mean PCC value.

To conclude, we trained a regression model to predict 7 mouth region expressions from a single person. We compared the visual result of each output expression with the ground truth input expressions. For the selected mouth region expressions, our method shows good prediction results both visually and quantitatively.

Conclusion and Future Work

In this paper, we conducted a comprehensive survey of VR expression tracking and discussed current challenges. Next, we proposed a baseline expression tracking method for the HP Reverb G2 device based on a regression model trained on a single person. The regression model uses the MobileNetV3 network structure combined with data augmentation. Experimental results show good performance both visually and quantitatively for 7 mouth region expressions from the same person.

Our future work will focus on cross-person expression tracking modeling, which is more challenging due to the diversity of facial features. We are also interested in extending the current method to eye tracking, cheek expressions tracking, and tongue movement.

Another potential extension of our proposed work is to transfer from image input to video input, where temporal information can be taken into consideration to further enhance the prediction results. Other factors to be considered may include lighting variation in IR images, movement of the participant when wearing the headset, *etc.*. Mouth region expression tracking combined with speech animation is another important direction to explore, which would enable virtual communication for HMDs.

References

[1] R. Parent, *Computer Animation*. Boston, USA: Morgan Kaufmann, 2012, vol. 3.

[2] B. Houshmand and N. M. Khan, "Facial expression recognition under partial occlusion from virtual reality headsets based on transfer learning," *2020 IEEE 6th International Conference on Multimedia Big Data*, pp. 70–75, Sep. 2020, New Delhi, India.

[3] S. E. Wei, J. Saragih, T. Simon, *et al.*, "VR facial animation via multiview image translation," *ACM Transactions on Graphics*, vol. 38, no. 4, 67:1–67:16, Jul. 2019.

[4] H. Li, L. Trutoiu, K. Olszewski, *et al.*, "Facial performance sensing head-mounted display," *ACM Transactions on Graphics (Proceedings SIGGRAPH 2015)*, vol. 34, no. 4, 47:1–47:9, Aug. 2015.

[5] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2019.

[6] Y. Lu, S. Wang, W. Zhao, and Y. Zhao, "WGAN-based robust occluded facial expression recognition," *IEEE Access*, vol. 7, pp. 93 594–93 610, 2019.

[7] K. B. M. E. B. Prince and D. S. Messinger, "Facial action coding system," *The SAGE Encyclopedia of Communication Research Methods*, vol. 4, 2015, Thousand Oaks, USA.

[8] J. Lewis, K. Anjyo, T. Rhee, M. Zhang, F. Pighin, and Z. Deng, "Practice and theory of blendshape facial models," *Eurographics 2014 - State of the Art Reports*, Apr. 2014, Strasbourg, France.

[9] K. Olszewski, J. J. Lim, S. Saito, and H. Li, "High-fidelity facial and speech animation for VR HMDs," *ACM Transactions on Graphics*, vol. 35, no. 6, 221:1–221:14, Dec. 2016.

[10] A. Gruebler and K. Suzuki, "Design of a wearable device for reading positive expressions from facial EMG signals," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 227–237, Mar. 2014.

[11] J. Lou, Y. Wang, C. Nduka, *et al.*, "Realistic facial expression reconstruction for VR HMD users," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 1–1, Aug. 2019.

[12] I. Mavridou, L. McGhee, M. Hamedi, *et al.*, "FACETEQ interface demo for emotion expression in VR," *2017 IEEE Virtual Reality*, pp. 441–442, Jan. 2017, Los Angeles, USA.

[13] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 187–194, Jul. 1999, New York, USA.

[14] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, Jul. 2003.

[15] B. Chu, S. Romdhani, and L. Chen, "3D-aided face recognition robust to expression and pose variations," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1907–1914, 2014, Columbus, USA.

[16] P. Dou, S. K. Shah, and I. A. Kakadiaris, "End-to-end 3D face reconstruction with deep neural networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1503–1512, Jul. 2017, Honolulu, USA.

[17] S. Romdhani and T. Vetter, "Efficient, robust and accurate fitting of a 3D morphable model," *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 59–66, Oct. 2003, New York, USA.

[18] P. Martins, J. Sampaio, and J. Batista, "Facial expression recognition using active appearance models," *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, pp. 123–129, Jan. 2008, Funchal, Portugal.

[19] G. J. Edwards, T. F. Cootes, and C. J. Taylor, "Face recognition using active appearance models," *European Conference on Computer Vision*, pp. 581–595, Jun. 1998, Berlin, Germany.

[20] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh, "Deep appearance models for face rendering," *ACM Transactions on Graphics*, vol. 37, no. 4, 68:1–68:13, Jul. 2018.

[21] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *2nd International Conference on Learning Representations*, pp. 14–16, Apr. 2014, Banff, Canada.

[22] A. Jourabloo and X. Liu, "Large-pose face alignment via CNN-based dense 3D model fitting," *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4188–4196, Jun. 2016, Las Vegas, USA.

[23] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 146–155, Jun. 2016, Las Vegas, USA.

[24] P. Huber, G. Hu, J. R. Tena, *et al.*, "A multiresolution 3D morphable face model and fitting framework," *11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Feb. 2016, Rome, Italy.

[25] G. Littlewort, J. Whitehill, T. Wu, *et al.*, "The computer expression recognition toolbox (CERT)," *2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops*, pp. 298–305, Mar. 2011, Santa Barbara, USA.

[26] X. Zhang, L. Yin, J. F. Cohn, *et al.*, "BP4D-Spontaneous: A high-resolution spontaneous 3D dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, Jun. 2014.

[27] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 94–101, Jun. 2010, San Francisco, USA.

[28] E. Sánchez-Lozano, G. Tzimiropoulos, and M. F. Valstar, "Joint action unit localisation and intensity estimation through heatmap regression," *British Machine Vision Conference*, May 2018, Newcastle, UK.

[29] Y. Fan and Z. Lin, "G2RL: Geometry-guided representation learning for facial action unit intensity estimation," *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 731–737, Jul. 2020, Yokohama, Japan.

[30] Y. Yan, K. Lu, J. Xue, P. Gao, and J. Lyu, "FEAFA: A well-annotated dataset for facial expression analysis and 3D facial animation," *2019 IEEE International Conference on Multimedia and Expo Workshops*, pp. 96–101, Jul. 2019, Shanghai, China.

[31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, pp. 4278–4284, Feb. 2016, San Francisco, USA.

[32] I. Freeman, L. Roese-Koerner, and A. Kummert, "EffNet: An efficient structure for convolutional neural networks," *25th IEEE International Conference on Image Processing*, pp. 6–10, Oct. 2018, Athens, Greece.

[33] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6848–6856, Jun. 2017, Salt Lake City, USA.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations*, Sep. 2015, San Diego, USA.

[35] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, Dec. 2018, Salt Lake City, USA.

[36] A. Howard, M. Sandler, B. Chen, *et al.*, “Searching for MobileNetV3,” *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1314–1324, May 2019, Long Beach, USA.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L.Fei-Fei, “ImageNet: A large-scale hierarchical image database,” *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009, Miami, USA.

Author Biography

Xiaoyu Ji *received her MS in Electrical and Computer Engineering from Purdue University (2021), received her BS from University of Electronic Science and Technology of China. She is pursuing a Ph.D. degree in Electrical and Computer Engineering at Purdue University. Her current work is focusing on facial image processing.*

Justin Yang *received his B.S. degree in Electrical and Computer Engineering from National Chiao Tung University in June 2018. He is currently a Ph.D. student at the School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA. His research interests include image processing, computer vision.*

Jishang Wei *is a principal scientist and research manager at HP Labs. His research focuses on human and environment sensing and understanding through multi-model machine learning. Jishang has successfully cultivated Omnicept from concept to production. He is currently leading a team to develop core AI capabilities, including cognitive load inference and expression recognition, for HP Reverb G2 Omnicept Edition at HP’s Advanced Computer and Solutions Group. Jishang received his Ph.D. degree in Computer Science from the University of California, Davis.*

Yvonne Huang *is a machine learning engineer in HP VR group. She holds B.S. and M.S. degree in computer science and has 3 years of experience in deep learning and image processing. Her current work is focusing on facial tracking.*

Qian Lin *is an HP Fellow working on computer vision and deep learning research. She is also an adjunct professor at Purdue University. She joined Hewlett-Packard Company in 1992. She received her BS from Xi’an Jiaotong University in China, her MSEE from Purdue University, and her Ph.D. in Electrical Engineering from Stanford University. She is inventor/co-inventor for 45 issued patents. She was awarded Fellowship by the Society of Imaging Science and Technology (IS&T) in 2012, Outstanding Electrical Engineer by the School of Electrical and Computer Engineering of Purdue University in 2013, and the Society of Women Engineers Achievement Award in 2021.*

Jan P. Allebach *is Hewlett-Packard Distinguished Professor of Electrical and Computer Engineering at Purdue University. He was named Electronic Imaging Scientist of the Year by IS&T and SPIE, and was named Honorary Member of IS&T, the highest award that IS&T bestows. He has received the IEEE Daniel E. Noble Award, the IS&T/OSA Edwin Land Medal, the IS&T Johann Gutenberg Prize, is a Fellow of the National Academy of Inventors, and is a member of the National Academy of Engineering.*

Fengqing Zhu *is an Assistant Professor of Electrical and Computer Engineering at Purdue University, West Lafayette, Indiana. Dr. Zhu received the B.S.E.E. (with highest distinction), M.S. and Ph.D. degrees in Electrical and Computer Engineering from Purdue University in 2004, 2006 and 2011, respectively. Her research interests include image processing, computer vision, video compression and digital health. Prior to joining Purdue in 2015, she was a Staff Researcher at Futurewei Technologies (USA). She is a senior member of the IEEE.*