HP Omnicept Face Tracking in VR

The HP Omnicept face tracking solution was developed based on an HP Reverb G2 VR headset Omnicept edition. The HP Reverb G2 VR headset is equipped with a single infrared mouth camera to capture wearer's lower face movement. In this technical report, we present a novel and robust machine learning solution for tracking face movement in real time using a single infrared mouth camera. To develop the machine learning solution, we collected quantifiable measures of real people's facial movements according to the Facial Action Coding System (FACS) definition. These quantifiable measures are used to animate avatars in a virtual environment, and we capture the avatars facial movements with a virtual camera. These captured synthetic avatar facial data are used to train our multi-regression deep neural network face tracking model. We have demonstrated the robustness of our solution through both quantitative and qualitative analysis.

Full technical report and downloadable dataset available December 31, 2022

Wei, J., Zhang, S., Ji X., Sundaramoorthy P., Kelley N., Rawlings B., Yraguen M., Huang Y., Siegel, E.H., Ballagas, R., Yang J., Lin Q., Zhu M., Allebach J., "HP Omnicept Face Tracking in VR", *Technical Report*, HP Labs, Palo Alto, CA.

1. Introduction

1.1 Face Tracking

A real time face tracking solution in the virtual reality (VR) environment has the potential to enable a wide range of applications [1] [2]. From a user standpoint, retargeting natural expressions in the real world to an avatar in VR increases the sense of "reality" in the virtual world and improves the user's experience in interacting with others. From an enterprise perspective, facial tracking data can give businesses insights into the emotional state of the user for marketing research purposes. In line with HP's previous work on determining the mental workload of users from biometric sensors, facial tracking could provide additional information on the user's true mental state and work in tandem with the cognitive load inference. For example, understanding the arousal level of a hard skilled worker undergoing training in a high-stakes task may suggest their confidence in completing the task. Alternatively, facial tracking in VR could uncover a user's true emotional response to an advertisement. These are a few examples out of the plethora of business challenges that could be addressed with an AI model designed to track facial movements.

1.2 Facial Action Coding System (FACS)

A widely used tool for coding facial expressions known as the Facial Action Coding System (FACS) was introduced in [3]. 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 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 a random combination of AUs may not produce realistic facial expressions [4].

1.3 Review of Past Research

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

1.3.1 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 sections, we discuss face tracking works using either HMC or EMG input.

(1) 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 a face tracking model with 9 IR HMCs for training and 3 IR HMCs for real-time testing [5]. In addition, the training device collects two more eye region views and four more mouth region views. While research in [6] 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 a face tracking model. The same research group proposed another RGB-based HMC design in 2016 [7], 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.

(2) 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 [8]. Lou *et al.* proposed an approach to recover facial action unit intensities from EMG signals in [9]. FACETEQ hardware [10] 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.

1.3.2 3D Modeling

Reconstructing 3D facial expressions from a single 2D image is an active research topic in image processing and computer graphics. 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.

(1) 3D Morphable Model (3DMM)

Different methods have been proposed for face tracking from monocular inputs. Some focus on using statistical models for facial texture and shape. A popular method is the 3D Morphable Model (3DMM) [11]. 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 [12].

(2) 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 tracking, AAM is able to successfully disentangle the shape and appearance using Principal Component Analysis (PCA) [13]. 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 [14]. 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.

(3) 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 in modeling human faces in 3D [15]. Using images captured by 40 cameras from different viewing angles, a Variational Autoencoder (VAE) [16] is used to model a data-driven avatar that learns a joint latent representation of face geometry and appearance. With view-point conditioning, VAE can disentangle the viewpoint-specific information from the latent representation of face geometry and appearance. By manipulating the latent variable, it can perform controllable a synthesis of the facial expression without modifying the facial geometry and identity of the given avatar.

(4) Landmark Fitting Approach

Incorporating landmark detection into 3D face modeling can add constraints for synthesis [17] [18]. Given a set of landmark points with the correspondence in the 3D face model, some techniques fit the 3D surface with the detected landmarks [19]. These approaches are accurate in terms of face tracking. However, it is unclear how well they perform if the face is partially occluded, especially when the occlusion affects the detection of facial landmarks.

(5) 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) [20]. As more detailed datasets such as BP4D [21] and CK+ [22] 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 [23] [24]. 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 [25].

1.4 Overview of Our Work

We designed and developed a face tracking deep regression model which can predict AU intensities from a single facial image captured by a camera on the HP Reverb G2 VR headset.

To train a deep learning model to predict the user's AU intensities requires the AU intensity ground truth. Existing solutions involve labor-intensive AU labeling by annotators. It is even harder to collect such AU intensity ground truth considering that only half the face is visible when the user puts on a VR headset. In our work, we leverage synthetic data to train the deep learning model. We collect real people's full face AU intensities, which are used to drive 3D virtual avatars in a virtual environment. Synthetic partial facial images of these virtual avatars are captured. Using these synthetic data, we train a deep regression model to predict multiple AU intensities simultaneously. After model training, we take the video stream as input data to our model and receive the AU intensities as output. Then, we retarget the predicted expression onto an avatar to reenact expressions. We compared the visual result of each output expression with the ground truth input expressions. For the mouth region expressions, our method shows good prediction results both qualitatively and quantitatively.

2. Data Collection and Generation

2.1 Real Facial Movement Data

To establish a ground truth of facial movement for our synthetic dataset, we collected facial action units as a quantifiable measure of expression intensities. Variations of facial AUs, or categorized muscle movements, moving together indicate meaningful expressions, such as a smile or frown. We designed a study, where participants were asked to demonstrate and record 16 mouth expressions using a full-face tracking solution. The face tracking solution records video data and outputs the corresponding AU intensities, which aligns closely with the facial AUs defined by the Facial Action Coding System [29]. Participants were asked to demonstrate a variety of mouth expressions, including closed mouth smile, open mouth smile, closed mouth frown, open mouth frown, pucker/pursed lips, anger teeth, mouth right, mouth left, mouth funnel, smile right, smile left, cheek squint right, cheek squint left, upper up right, upper up left, and cheek puff expressions, each for two times.

2.2 Avatar Generation

We create a large collection of 3D virtual avatar faces with various facial attributes. The collection includes a diverse set of avatars with a balance of gender (male/female), ethnicity (white/black/Asian/Latin), and age (young/senior). The avatars we created cover a large variety of face colors, face shapes, and other face attributes (e.g., facial hair).

2.3 Synthetic Data Generation

In a virtual environment, an avatar wears a simulated HP Reverb G2 VR headset. IR images of the avatars are taken from the view angle of the camera on the VR headset while the avatars are animated using the collected AUs intensities of different expressions. Specifically, 23 mouth region AUs were utilized to recreate expressions on avatars and build a synthetic dataset. The 23 mouth AUs include mouth close, mouth funnel, mouth pucker, mouth left, mouth right, mouth smile (left & right), mouth frown (left & right), mouth dimple (left & right), mouth stretch (left & right), mouth roll (upper & lower), mouth shrug (upper & lower), mouth press (left & right), mouth lower down (left & right), and mouth upper up (left & right).

To this end, we have built a synthetic data generation pipeline that supports large-scale data generation to train powerful deep learning models. With this paper, we make a representative set of our data available to the public.

3. Deep Regression Model

3.1 Model Design

Since deep neural networks have been proven effective in learning latent patterns in high-complexity data (such as images) in the presence of a huge number of training data, we propose a deep regression network to predict the AU intensity of facial expressions. The deep regression network contains two parts. The first part is a backbone network to extract facial features and the second part is a regression network. We have compared a few neural network structures, such as EffNet [26], VGG-11 [27], and Inception-ResNet-V1 [28], to choose from as the backbone of our regression model. We balance model performance and model size during the model selection process. Inception-ResNet-V1 has a relatively large model size which does not fit our design requirement of deployment on an edge device. Among the network structures with large model sizes, VGG-11 performs the best, and EffNet has a comparable result. While comparing the Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC) among the network structures with a small model size, our experiment results show that MobileNetV3 [31] has similar performance as EffNet. Although MobileNetV2 and ShuffleNet have smaller model sizes, MobileNetV3 is chosen as the backbone network for our baseline due to better prediction results. A MobileNetV3 model pre-trained on ImageNet [32] dataset is used in our neural network structure. The output of the first part is a facial feature vector. The second part is a regression network which consists of five fully connected layers and four activation layers, to predict the AU intensity from the feature vector that has been extracted.

3.2 Model Training

Figure 1 shows the avatar-independent face tracking model training process. We collected expressional AU intensity ground truth from participants, which are used to drive the 3D avatar's facial movement. Then, synthetic images of the avatars (with different expressions) rendered from the VR camera viewpoint are provided as input to the model. The model output are AU intensities, the ground truth of which are the AU intensities collected from participants.

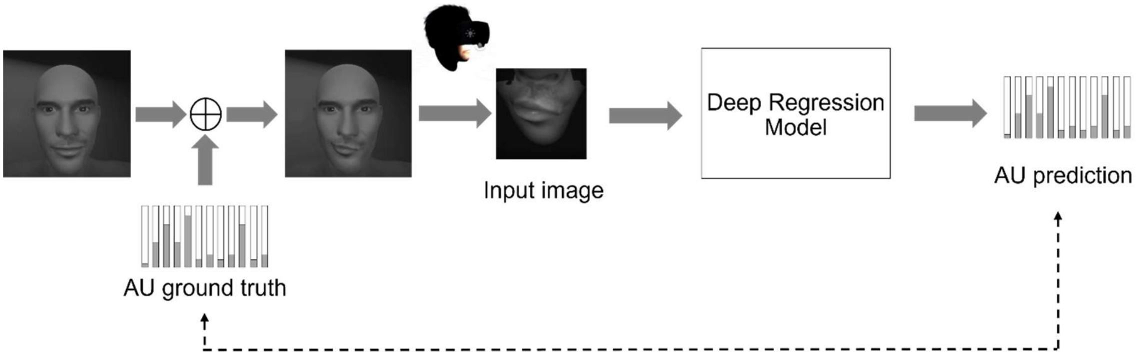

Figure 1. Avatar-independent face tracking model training pipeline. We train the deep regression model to predict facial action units.

The deep regression model is trained on the synthetic expression images of avatars and the corresponding 23 target AU intensities. Each of the target AU intensities has a range of 0 to 100. The network output is a 23 AU intensity vector prediction. We use the Mean Absolute Error (MAE) of the 23 AU intensity ground truth and prediction as the loss function. The data synthesis process is described in Section 2.3. The model is trained on a diverse set of 116 virtual avatars with expressions from data collection, including 26 aged female avatars, 30 young female avatars, 26 aged male avatars, and 34 male avatars. We use 164 synthetic frames from each avatar for model training.

3.3 Evaluation

We evaluated the quantitative result of the model trained on 116 avatars and tested on 6 avatars (3 female avatars and 3 male avatars). The test dataset includes 162 images for each avatar. The evaluation of AU intensities employs the metrics of Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC). The AU ground truth and prediction ranges are both 0 to 100, and we have achieved a test MAE value of 2.39. The PCC value is 0.964. The quantitative result demonstrates the effectiveness of our proposed deep regression model.

4. Expression Reenactment via Face Tracking

We further built a facial expression reenactment application based on our trained face tracking model. Figure 2 shows the overall system, which is composed of four parts: data preprocessing, deep learning model, postprocessing, and avatar reenactment.

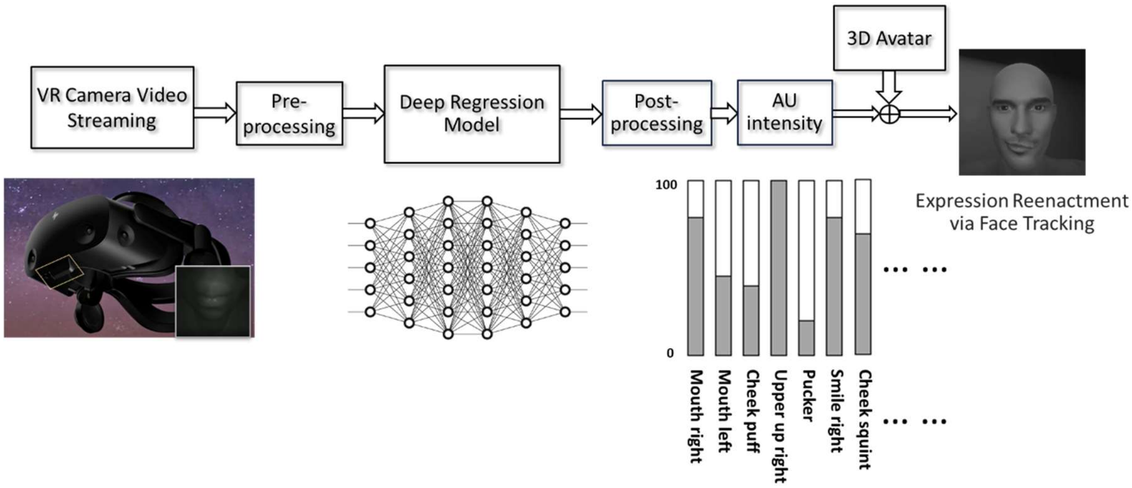

Figure 2. Overall system of user's expression reenactment on an avatar via face tracking

4.1 Data Preprocessing

As the user may wear the VR headset in varying lighting conditions, the captured mouth image is possible to suffer from overexposure or underexposure. We preprocess the facial images of the VR headset user before applying the model to make sure our model works well in all kinds of working environments. In our work, Adaptive Histogram Equalization [30] is used to process captured IR images to improve the lighting conditions. Specifically, we adopt the contrast-limited adaptive histogram equalization (CLAHE) algorithm, which improves contrast within images. The algorithm is adaptive in that different images (and more specifically, different regions of an image) may have their contrast amplified by different amounts. In a contrast-limited implementation, it limits the amount of contrast amplification in any region of an image.

4.2 Deep Learning Model

After image preprocessing, we employ our trained deep learning model to track the user's facial AU intensities. We converted the trained model into ONNX format and integrated the ONNX model into our real-time facial tracking application for an HP Reverb G2 Omnicept headset.

4.3 Postprocessing and Expression Reenactment

Once we get the prediction of the deep regression model, we apply a moving average filter to the predicted AU intensity with a window size of 5 (previous 5 predictions) to smoothen the output. The window size is chosen as a trade-off between latency and filtering effect. A larger window size will increase the latency of the overall model, and a smaller window size will give a weak filtering effect. After the model outputs the AU intensity prediction, we apply the prediction on a 3D avatar in a game engine to reenact the user's facial expression for visual evaluation and assessment.

4.4 Evaluation

Our current model has been validated on a certain cohort of avatars. We rendered 3D virtual avatars with tracked facial expressions to visually evaluate our face tracking model. We also tested our face tracking model on a user wearing the headset and performing the same set of 16 expressions in the training process. As we aim to test the user-independent performance of the model, we did not include the user's avatar in the training set to eliminate the model's exposure to the face geometry and attributes of the test user beforehand. We present example test results on the real user in Figure 3. From the figure, we can see all the 16 expressions were recognized and transferred successfully with high fidelity onto the virtual avatar. During online testing, we have not witnessed any observable latency due to the small footprint and acceptable computing consumption of the inference model. To this end, we have validated the feasibility and usability of the facial tracking and reenactment system using our face tracking model.

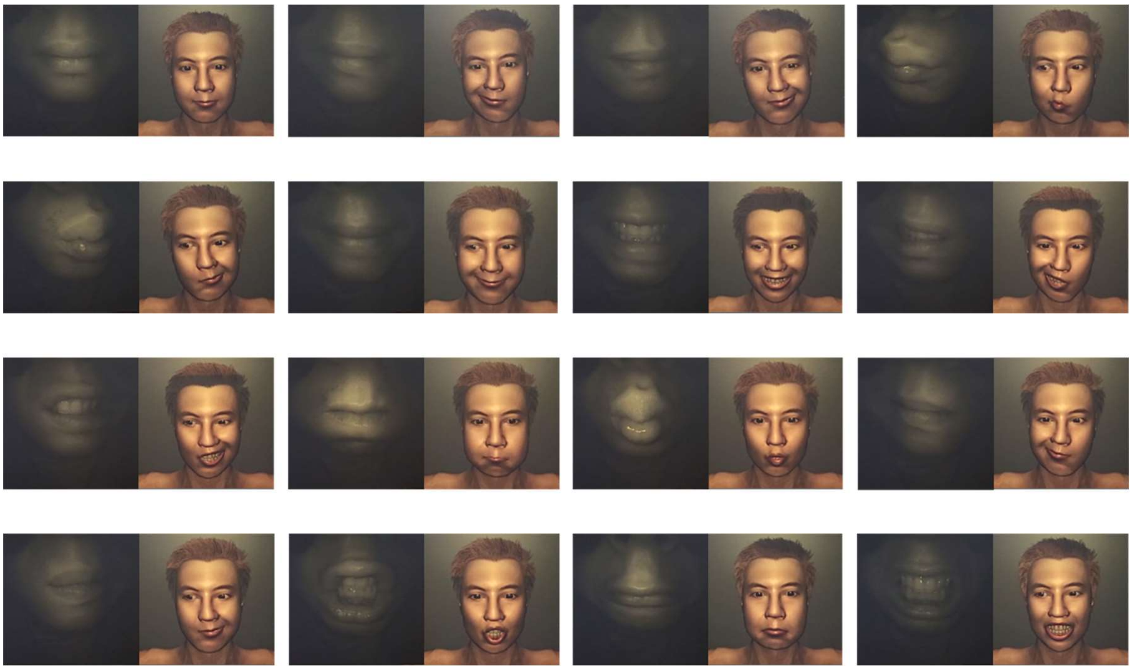

Figure 3. Results on a real user wearing the VR headset. For each pair, the left is the input IR image captured by the mouth IR camera on VR headset, and the right is the face tracking result reenacted on a virtual avatar. From left to right, top to bottom are respectively expressions: neutral face, mouth left, mouth right, closed mouth smile, open mouth smile, cheek squint left, cheek squint right, cheek puff, pucker, upper up left, upper up right, mouth funnel, frown, anger teeth.

5. Conclusion

VR headset-based facial tracking and reenactment will enable a more immersive user experience and open more opportunities in healthcare, education, and training domains. We designed and implemented the facial tracking solution using only one single VR headset infrared (IR) camera to track the user's facial expression. We built a prototype demo on an HP Reverb G2 Omnicept VR headset to track the user's face and reenact a virtual 3D avatar to imitate the user's expression in real time. In this technical report, we presented the deep learning face tracking approach and its training process and demonstrated an AI-based facial expression reenactment system using our deep learning model. Based on the quantitative and qualitative results on synthetic human faces and a real user, we have proven the usability and efficacy of our proposed method and the system.

References

[1] S. E. Wei et al., "VR facial animation via multiview image translation," ACM Transactions on Graphics, vol. 38, no. 4, 2019, doi: 10.1145/3306346.3323030.

[2] S. H. Wang, H. C. Sun, C. H. Lin, and H. Y. Yang. "Facial expression modeling method, apparatus and non-transitory computer readable medium of the same." U.S. Patent 10,885,702, issued January 5, 2021.

[3] 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.

[4] 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

[5] 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.

[6] 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

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

[8] 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.

[9] 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.

[10] 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.

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

[12] V. Blanz and T. Vetter, "Face recognition based on fit- ting a 3D morphable model," IEEE Transactions on Pat- tern Analysis and Machine Intelligence, vol. 25, no. 9, pp. 1063–1074, Jul. 2003

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] G. Littlewort, J. Whitehill, T. Wu, et al., “The computer ex- pression recognition toolbox (CERT),” 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops, pp. 298–305, Mar. 2011, Santa Barbara, USA.

[21] 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.

[22] 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.

[23] E. S ́anchez-Lozano, G. Tzimiropoulos, and M. F. Val- star, “Joint action unit localisation and intensity estimation through heatmap regression,” British Machine Vision Conference, May 2018, Newcastle, UK.

[24] 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.

[25] 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.

[26] I. Freeman, L. Roese-Koerner, and A. Kummert, “Effnet: An Efficient Structure for Convolutional Neural Networks,” in 2018 25th IEEE International Conference on Image Processing (ICIP), 2018.

[27] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv [cs.CV], 2014.

[28] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, Inception-ResNet and the impact of residual connections on learning,” Proc. Conf. AAAI Artif. Intell., vol. 31, no. 1, 2017.

[29] P. Ekman and E. L. Rosenberg, What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). 2012. doi: 10.1093/acprof:oso/9780195179644.001.0001.

[30] S. M. Pizer et al., “Adaptive Histogram Equalization And Its Variations,” Compute Vis Graph Image Process, vol. 39, no. 3, 1987, doi: 10.1016/S0734-189X(87)80186-X.

[31] A. Howard et al., “Searching for mobileNetV3,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, vol. 2019-October. doi: 10.1109/ICCV.2019.00140.

[32] O. Russakovsky et al., “ImageNet Large Scale Visual Recognition Challenge,” International Journal of Computer Vision, vol. 115, no. 3, 2015, doi: 10.1007/s11263-015-0816-y.