International Conference on Artificial Reality and Telexistence
Eurographics Symposium on Virtual Environments (2024)
T. Tanabe and V. Yem (Editors)

*Poster*

# Mapping of Facial Action Units to Virtual Avatar Blend Shape Facial Movement

Tony Wolff[1], Felix Dollack[1], Monica Perusquia-Hernandez[1]†, Hideaki Uchiyama[1], Kiyoshi Kiyokawa[1]

[1]Nara Institute of Science and Technology (NAIST), Japan

**Abstract**

*Action Units and blend shapes are two frameworks to describe facial movement. However, mappings between the two frameworks are underinvestigated. We present an automated mapping technique using machine learning. Our model infers ARKit-compatible blend shape weights from action unit intensities extracted with OpenFace. We use a GRU architecture to retain time-dependent information leveraging the particularities of Recurrent Neural Networks while still permitting fast, real-time inference. Our generalized model yields an activation precision of 90% and an activation recall of 85%.*

**CCS Concepts**

• ***Human-centered computing*** → ***Virtual reality;*** • ***Computing methodologies*** → ***Machine learning algorithms;***

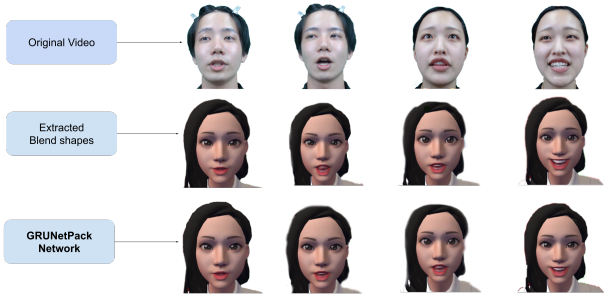


**Figure 1:** *Comparison of our results with the ground truth blend shapes and actual facial expression.*

## 1. Introduction

In animation, facial expressions are often represented in units of blend shapes [LAR*14]. A blend shape defines a set of vertices to be deformed with a weight going from zero (no deformation at all) to one (exaggerated deformation). Blend shapes are intuitive and allow for freedom compared to bone-based animation. However, much of the literature describing facial expressions of emotion uses the Facial Action Coding System (FACS) [EF82], which describes a set of Action Units (AUs). Therefore, a precise mapping between the two would enable animators to use the body of knowledge about facial expression production already described in the FACS and transfer it directly to facial animations. Recently, simultaneous speech and facial expression animation has been mostly voice-driven using Deep Learning (DL) approaches. However, few works try to add affective visual information to speech-based animations. JALI [ELFS16] is a method that combines speech animation with the FACS. However, it only supports audio input. The FACS is used to construct their template facial rig. Therefore, we propose a model that maps speech-related facial movement from the FACS to blend shape weights using a Gated Recurrent Unit (GRU) architecture. This architecture has already been proven to have fewer parameters than Long short-term memory (LSTM) networks, achieving equally good results [YYZ20]. With this, we avoid using embeddings as the data are directly usable inside our networks.

## 2. Dataset

We used the facial videos and blend shape tracking ground truth by Miyawaki et al. [MPHI*22]. Nine Japanese participants read 100 emotion-related sentences in four volume levels. This makes for 3,780 RGB facial videos with their corresponding 52 blend shape ground truth values tracked in real-time using ARKit.

## 3. Data Analysis

The code is open-source and freely available †. AUs were extracted with OpenFace 2.0 [BZLM18], a toolkit capable of facial landmark detection, and facial action unit recognition. AU intensities

† m.perusquia@is.naist.jp

† https://github.com/tony-wolff/AU_BS_model



DOI: 10.2312/egve.20241392

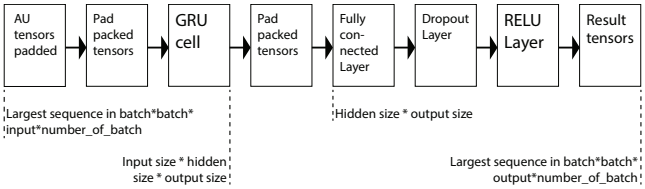


**Figure 2:** *Network architecture, used in our GRU models.*

**Table 1:** *Comparing different training configurations on the test set*

| network | scheduler | batch size | hidden size | patience | R2 | RMSE | activation precision | activation recall |
|---|---|---|---|---|---|---|---|---|
| GRUNetPack | None | 8 | 256 | 10 | 0.74 | 0.031 | 0.91 | 0.84 |
| GRUNetPack | StepLR | 8 | 256 | 10 | 0.74 | 0.032 | 0.91 | 0.84 |
| GRUNetPack | MultiStepLR | 4 | 128 | 10 | 0.74 | 0.032 | 0.90 | 0.84 |
| GRUNetPack | MultiStepLR | 8 | 128 | 10 | 0.72 | 0.033 | 0.90 | 0.84 |
| GRUNetPack | CosineAnnealingLR | 8 | 256 | 30 | 0.75 | 0.031 | 0.90 | 0.85 |

are derived by OpenFace in the range of zero to five. We normalized these intensities to have values between zero and one. We synchronized video and blendshape recordings by cross-correlating the blend shape tracking with the OpenFace output for AU26 (Jaw drop). AU26 was chosen because it is a large facial movement involved in speech. A Savitzky-Golay filter was applied before cross-correlation to ease the matching. We excluded 30 videos where the differences between the video and the blend shapes were greater than one second. This represents less than 1% of the initial dataset. The resulting dataset is composed of 3,750 videos. A low pass filter of 100 ms was applied to both the extracted OpenFace features; then the lag was corrected.

Figure 2 illustrates the baseline architecture for our proposed Gated Recurrent Units (GRU)-based neural network. It is composed of one GRU cell taking packed tensors. Packed tensors are a trick used with RNNs to save computations when dealing with batches of variable lengths. We then pad the data before entering the fully connected layer, and obtain the final result tensors: the blend shape coefficients. The unpadding is realized later during output reconstruction. Note that a batch here is composed of full-size scripts. We implemented Early Stopping during training for faster training and better generalization.

We evaluated the performance independently of the volume levels used when recording the dataset (volume-independent). We tested the seq2seq performance compared to models inferring the sentences in one go instead of sequential processing with typical RNNs. In the proposed architecture, packed tensors are fed to the GRU layer. The result of the GRU layer is padded as the forward layer does not accept packed tensors. Forward results are passed to a dropout layer before entering the ReLU layer. During loss computation, the padding is removed. This model allows for variable-sized batches, as each batch is padded to match the length of the longest script. Volume-independent models were trained by splitting the dataset into three parts: 2625 training scripts (70% of the data), 750 testing scripts (20%), and 375 validation scripts (10%).

## 4. Results

Results can be seen in this YouTube video, and are displayed in Figure 1. After experimentation with multiple loss functions, and hyperparameters (see Table 1), our best configuration resulted in R2 of 0.75, activation precision of 0.90, and activation recall of 0.85. We also provide a comparison with Emotalk-generated blend shapes. Table 2 shows that Emotalk's precision-recall is 0.63-0.97, but compared to our model (precision-recall: 0.9-0.85), the performance to extracted ground truth blend shapes was lower. Our R2 score and RMSE metrics show a high correlation (0.75 for R2, 0.031 for RMSE) between predictions and ground truth. Our activation precision score indicates that we activate the right blend shape 90% of the time. Our activation recall score states that we predict an 85% activation for the ground truth blend shapes.

**Table 2:** *Comparison of Emotalk and our method on the test set*

| Methods | Inference Time (sec) | RMSE | R2 | Activation Precision | Activation Recall |
|---|---|---|---|---|---|
| Emotalk | 10,080 | 0.15 | -4.93 | 0.63 | 0.97 |
| GRUnetPack | 8.84 | 0.031 | 0.75 | 0.9 | 0.85 |

## 5. Discussion and future work

We proposed a method to map FACS-based visual information into animator-friendly blend shape weights. The first process involves extracting and pre-processing data for 17 action units normalized between zero and one. The second step generates correlated blend shape weights using a GRU-based model. The outcomes show better performance for activation precision and reasonably high recall scores (0.9, 0.85) compared to Emotalk (0.63, 0.97). A limitation of this work is that AU and Blendshape estimation methods are not perfect, and that the performance of those methods limits our results. Future work should verify the perceptual effectiveness of this model to automatically translate facial expression characteristics described in the FACS to blend shapes.

## References

[BZLM18] BALTRUSAITIS T., ZADEH A., LIM Y. C., MORENCY L.: OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)* (May 2018), pp. 59–66. 1

[EF82] EKMAN P., FRIESEN W. P.: Measuring facial movement with the Facial Action Coding System. In *Emotion in the human face*, Ekman P., (Ed.), second ed. Cambridge University Press, 1982, pp. 178–211. 1

[ELFS16] EDWARDS P., LANDRETH C., FIUME E., SINGH K.: JALI: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics 35*, 4 (July 2016), 1–11. 1

[LAR*14] LEWIS J. P., ANJYO K., RHEE T., ZHANG M., PIGHIN F. H., DENG Z.: Practice and theory of blendshape facial models. *Eurographics (State of the Art Reports) 1*, 8 (2014), 2. 1

[MPHI*22] MIYAWAKI R., PERUSQUIA-HERNANDEZ M., ISOYAMA N., UCHIYAMA H., KIYOKAWA K.: A Data Collection Protocol, Tool and Analysis for the Mapping of Speech Volume to Avatar Facial Animation. In *ICAT-EGVE 2022 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments* (2022), Uchiyama H., Normand J.-M., (Eds.), The Eurographics Association. 1

[YYZ20] YANG S., YU X., ZHOU Y.: Lstm and gru neural network performance comparison study: Taking yelp review dataset as an example. In *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)* (2020), pp. 98–101. 1