Supplementary Material for Emotion Guided Speech-Driven Facial Animation

KEYWORDS

speech-driven facial animation, emotion recognition, cross modality, deep neural networks

ACM Reference Format:

. 2021. Supplementary Material for Emotion Guided Speech-Driven Facial Animation. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 SUPPLEMENTARY MATERIAL

In the following supplementary documentation, we will explain the procedure and mechanism of our method, emotion guided speech-driven facial animation (EG-SDFA), in detail. The content includes descriptions of emotion guided projection and multi-functional fusion from our pipeline. In addition, supplementary video demonstrate the animation result from our method.

2 EMOTION GUIDED PROJECTION

Emotion guided projection is a solution to leverage from the integration of classification and regression done by an emotion recognition network (ERN) and a facial expression network (FEN). From given speech input data, ERN classifies which emotion (e.g. happy, sad, etc.) the speech contains. This classified emotion loads a master blendshape of the corresponding emotion. This master shape can be predefined as an average of target blendshapes (label data) of the emotion from the training session of FEN, or an average of predictions from the training session of FEN, or a custom-defined blendshape representing the emotional features. Once the master blendshape is loaded, FEN predicts the facial animation from the speech input in a blendshape form. As the predicted blendshapes are generated, we proceed linear interpolation between each unit vector of the predicted blendshapes and the unit vector of the loaded master blendshape. Then, each predicted blendshape is projected to the interpolated vectors, forming an emotion-enhance blendshapes. This way, the predicted blendshapes transforms into blendshapes with reflections of the master blendshape, which should contain the ideal condition of the emotion. The illustration for the procedure of the emotion guided projection is presented (Figure 1).

Assuming that ERN has predicted the correct emotion for the speech and the master blendshape weights represent the ideal condition of the emotion, it can provide evident guidance to the predicted blendshapes. Depending on the coefficient used in the linear interpolation, the degree of emotional reflection can be controlled.

3 MULTI-FUNCTIONAL FUSION

Regardless of how much the emotion-enhanced blendshapes represent the evidence in emotional features, it suffers from expressiveness in mouth movement. Therefore, we leverage the result from a network of speech-driven facial animation network of specialty in mouth expressiveness to enhance the quality of overall facial animation. For a raw combination of the two types of expressions (emotional and mouth) produces awkward features, multiple parametric functions to mitigate the deficiencies and to enhance the quality are included.

3.1 Linearly Weighted Regional Integration

To create a suitable combination of the two different expressions, we perform linear weighted integration. We assign an importance weight for the two types of expressions (from emotion-guided projection and the mouth expression network) to each blendshape type. While the middle and upper regions of facial blendshape types show a drastic range of displacement for emotional features, blendshape types of the central lower part of the face are mainly involved with mouth movement (Figure 3). According to such regional characteristics, importance weights are generated, which are used for the linear weighted summation of blendshapes from the two types of expressions.

3.2 Intensity Control

To provide more flexible handling of the emotional expression, intensity is to be controlled to extract a desirable level of result. The intensity control of the emotion is done by scaling the weights of the blendshapes. Linear multiplication has shown to be an effective way to control the intensity monotonously, increasing and decreasing the blendshape weights with the same scaling coefficient for all the values in the blendshape (Figure 2). On the other hand, non-linear gamma transformation has been shown to change the ratio of the blendshape to simulate non-uniform variation in emotional expression. Gamma value of greater than 1 can weaken a blendshape weights with a relatively high value to leave a smaller number of features to represent the emotion, while gamma value of less than 1 and greater than 0 can strengthen the blendshape with relatively small value to leave more features to represent the emotion. Such non-linearity could simulate drastic variation for the emotional expression.

3.3 Dynamic Range Control

For further flexibility of the final animation, we provide the capability to control the range of movement of animation within the time frame. While the result may contain emotional features, the movement may not be dynamic enough to simulate the intensiveness of the emotion. Dynamic range can be increased to simulate drastic movement or be decreased to weaken the movement rate.

First, We calculate the magnitude of each blendshape for each frame with L2 norm. These extracted frame number of magnitude

Conference'17, July 2017, Washington, DC, USA
© 2021 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

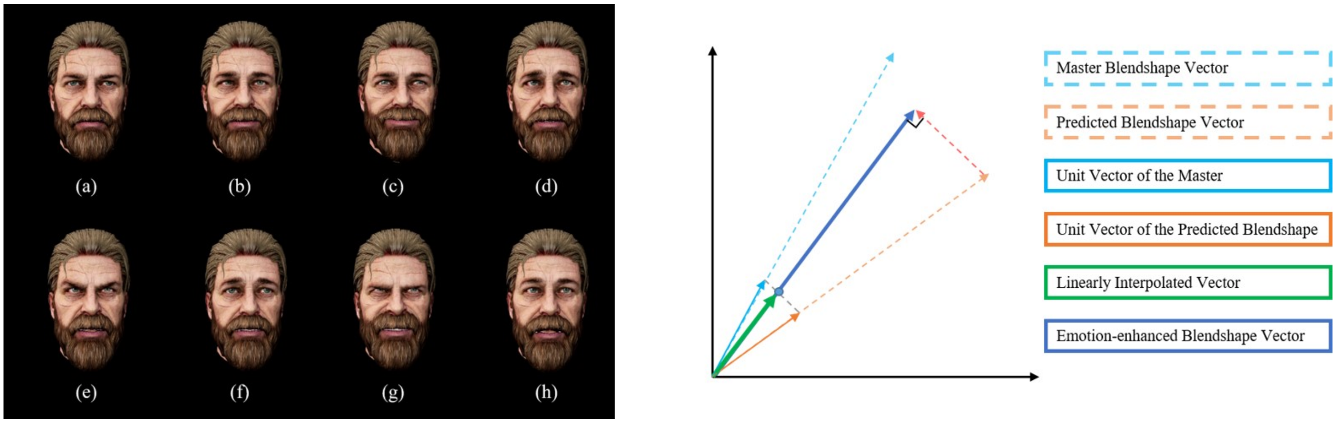
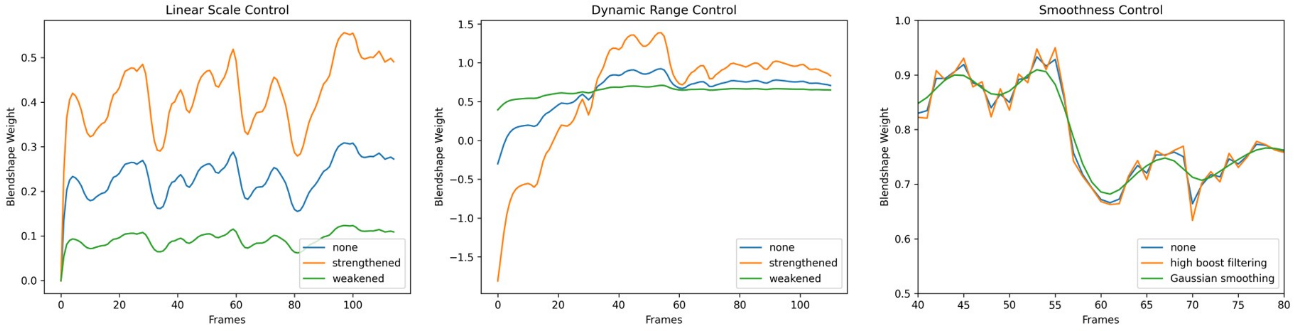

(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
Master Blendshape Vector
Predicted Blendshape Vector
Unit Vector of the Master
Unit Vector of the Predicted Blendshape
Linearly Interpolated Vector
Emotion-enhanced Blendshape Vector

Figure 1: Left: 8 emotional expression results by EG-SDFA: neutral (a), calm (b), happy (c), sad (d), angry (e), fearful (f), disgust (g), surprised (h). Right: extraction of emotion-enhanced blendshape vector from emotion guided projection procedure.



(a) (b) (c)

Figure 2: Blendshape Weight value change for the linear scale control (a), the dynamic range control (b), and the smoothness control (c).

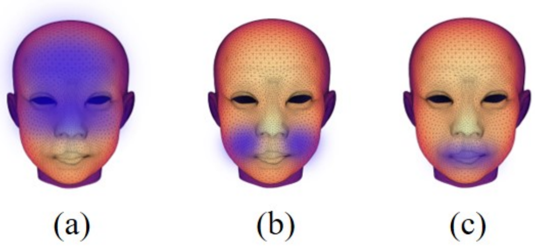


Figure 3: Regional importance highlight for emotional expression and mouth expression: emotional expression important (a), both expressions important (b), and mouth expression important (c).

values can form a probability distribution. Having an assumption that such distribution is a Gaussian distribution, we manipulate the distribution by forcefully stretching or contracting the distribution with a parameter of variance. After the manipulation, we can reassign the magnitude values to the corresponding blendshape to scale the blendshape to the magnitude. The result shows flexible control in the range of movement, preserving the mean value for the magnitude of blendshapes in the time frame (Figure 2).

3.4 Smoothness Control

Regardless of the recurrent structure of the FEN and the mouth expression network (MEN), seldom discontinuity and noisy transitions were observed. In order to mitigate such, simple 1-dimensional Gaussian smoothing was conducted on blendshapes in the time domain to smooth the transition. On the other hand, to simulate intentional "shaky" and noisy movement for the shivering movement of the character, a high boost filtering technique was used. By using adding the result from the 1-dimensional Laplacian of Gaussian filter, the overall movement was able to become noisy. Both methods were conducted iteratively to control the intensity of the smoothing or high boost filtering, that resulted in convincing result in the demonstration (Figure 2).