Real-time Facial Animation for 3D Stylized Character with Emotion Dynamics

Ye Pan*
Shanghai Jiao Tong University
Shanghai, China

Ruisi Zhang
Shanghai Jiao Tong University
Shanghai, China

Jingying Wang
Shanghai Jiao Tong University
Shanghai, China

Yu Ding
Virtual Human Group, Netease Fuxi AI Lab
Hangzhou, China

Kenny Mitchell
Roblox & Edinburgh Napier University
CA, USA

ABSTRACT

Our aim is to improve animation production techniques' efficiency and effectiveness. We present two real-time solutions which drive character expressions in a geometrically consistent and perceptually valid way. Our first solution combines keyframe animation techniques with machine learning models. We propose a 3D emotion transfer network makes use of a 2D human image to generate a stylized 3D rig parameter. Our second solution combines blendshape-based motion capture animation techniques with machine learning models. We propose a blendshape adaption network which generates the character rig parameter motions with geometric consistency and temporally stability. We demonstrate the effectiveness of our system by comparing it to a commercial product Faceware. Results reveal that ratings of the recognition, intensity, and attractiveness of expressions depicted for animated characters via our systems are statistically higher than Faceware. Our results may be implemented into the animation pipeline, supporting animators to create expressions more rapidly and precisely.

CCS CONCEPTS

• **Computing methodologies** → **Motion capture**; • **Human-centered computing** → **User studies**.

KEYWORDS

motion capture, virtual characters, emotion

ACM Reference Format:
Ye Pan, Ruisi Zhang, Jingying Wang, Yu Ding, and Kenny Mitchell. 2023. Real-time Facial Animation for 3D Stylized Character with Emotion Dynamics. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3613803

*Corresponding author. whitneypanye@sjtu.edu.cn

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3613803

1 INTRODUCTION

Keyframe animation and motion capture are practical approaches for producing animation, and both have their pros and cons [11, 31]. Keyframe animation generates expressive and artistic animation based on the twelve principles of animation, but it is very tedious work. Motion capture captures actor performance, which is limited to realistic animation. It is difficult to produce perceptual valid expressions for stylized characters because the geometric features of human faces are different from that of stylized characters. According to the storyline, characters must have readily discernible facial expressions that are congruent with their emotional condition [21, 22, 24]. Recently, there has been a couple of solutions [2, 34] that could generate perception-valid character expressions; however, these systems are not real-time.

We start with proposing our first solution generating stylized character expressions from human performances that is both perceptually and geometrically consistent, by using lightweight methods and interpolation techniques. We train the 3D expression transfer network that takes images of human faces and generates the rig parameters or controller values of the character that best match the human's facial expression. To increase performance, we develop a two-step filtering strategy to learn the mapping between human and character feature space. We used a lightweight method, called a multi-character adaption network, transfers character expressions to secondary characters.

We then propose our second solution by combining traditional blendshape animation techniques with a machine learning model. To begin, we train the blendshape adaption network that generates the character rig parameters based on the corresponding blendshape weights. In order to produce temporally stable, flicker-free, and geometrically consistent results, we take the rig parameters over the last three previous frames, together with the blendshape weights at the current frame, as the input to our network. The multi-character adaption network is then used to drive secondary characters, allowing us to reuse a principal character rig that we trained in the previous steps.

We investigated the effectiveness of using our methods to animate the characters. We compared three different tracking methods: first solution (Interpolation), second solution (Blendshape), and Faceware on the expression recognition, emotion intensity, and overall attractiveness, as these are crucial factors for audience engagement [10, 32]. Results show that both our methods significantly improved ratings of the expression recognition, thus validating the

effectiveness of the 3D expression transfer network and the blendshape adaption network, respectively.

The following are the key contributions of our work: (1) For the first time, we contributed two real-time methods transferring human facial expressions to multiple 3D stylized characters in a geometrically consistent and perceptually correct way. (2) The amalgamation of data sources (e.g., human expression video database, character blendshape database, controller value database, etc.) motivates further study in the domains of character rigging and animation. In particular, we constructed a high-quality emotional audio-visual dataset as materials for our user study: a set of video clips featuring two male and two female stylized characters talking with seven basic emotions. (3) We systematically conducted a user study to validate the effectiveness of our solutions in terms of expression recognition, intensity, and appeal. This grows the existing knowledge of how animated characters' facial expressions can influence our perception.

2 RELATED WORK

2.1 Blendshape facial animation

An industrial standard for rigging facial animation is the use of blendshapes and may be broadly classed among morphable models [5]. The sum of weighted blendshape models can represent facial expressions quickly and compactly [12]. The neutral phase of an avatar is denoted as B_0, a set of its blendshapes are $\{B_1, B_2, ..., B_N\}$ and a expression can be expressed as $B = B_0 + \sum_{i=1}^{N} w_i B_i$, where w_i are blendshape weights. Several software tools for markerless face motion capture have already been developed [5, 35]. Faceware and Faceshift/ARkit [1], for instance, collect the blendshapes related to a set of standard expressions given by a human source and map them into stylized characters.

Despite the ease of use of the blendshape representation, there are a few matters to consider. To begin, in order to depict a wide range of emotions, digital artists must frequently construct vast libraries of blendshape targets. For a professional artist, creating a suitably detailed model can take up to a week of labor and numerous cycles of refinement. There have been attempts to automate blendshape construction with individualized comprehensive human facial geometry capture and subsequent optimization procedures [28]. However, generating blendshapes for styled avatars still necessitates a time-consuming and labor-intensive modeling process [9, 20]. At the moment, each blendshape is typically handcrafted by experienced artists using professional applications like Blender or MAYA.

Second, in some circumstances, the generally linear structure of blendshapes impacts the quality of the animation [15, 27]. Exaggerated expressions or specific expressions would be impossible to portray outside of the linear span [25]. In fact, animators must sometimes account for these shortcomings by sculpting new key shapes or adding new correctives per 3 to 5 frames [13].

In light of the success of previous data-driven shape analytical techniques [2, 4, 26], we propose a real-time blendshape-based system that combines blendshape animations with our machine learning techniques to improve expression representation from linear to nonlinear.

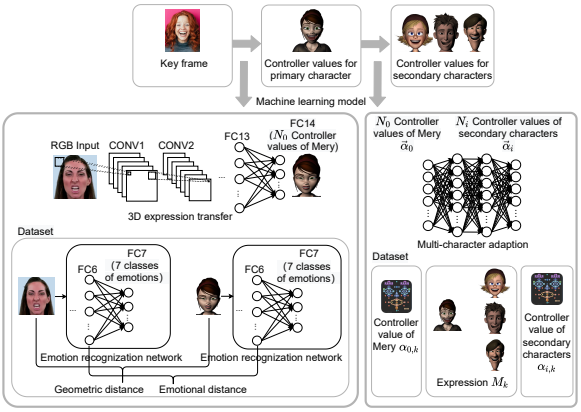


Figure 1: Overview of our first solution (Interpolation), which takes the human facial images as input and predicts the character's controller values

2.2 Data-based animation

The most related previous work to ours are DeepExpr [3], and ExprGen [2].

To begin, DeepExpr [3] presented a retrieval approach for locating the closest 2D expression image in the present database to a specific human image, while our Interpolation proposed a way for generating for human image a 3D stylized character expression. While DeepExpr inspired us, we employed alternative distance measurements. DeepExpr trained a Convolutional Neural Network (CNN) on a huge human expressions dataset to input a human expression and output the seven classes' probabilities, then trained a similar character model on a character expression dataset, and finally used transfer learning approaches to learn a mapping between the human and character feature space. Alternatively, we propose a two-step filtering approach to find the best-matched primary character image with the human face.

Second, our method shares a similar goal as ExprGen to learn 3D character expressions from individuals in a geometrically consistent and perceptually valid way, but our lightweight methods support live animation. ExprGen detects human and character face expressions to provide a perceptual metric for expression generation, and then develops a joint embedding to transfer human expressions to character expressions. In particular, ExprGen created triplets of training images to solve the issue of incorrect geometry matches within the same expression class. The triplets are generated by selecting the hard positive/negative exemplars from within a mini-batch. We did, however, present a novel method for learning the mapping between the human and character feature space.

Finally and most importantly, we integrate our solution with animation techniques to present two real-time methods. Different from static images, we need to take the time dimension into consideration to avoid flickering output. We further introduced a predictive model producing temporal stability and geometric consistency for the character rig parameter motions.

3 FIRST SOLUTION (INTERPOLATION)

We begin with the first solution (see Figure 1) that takes human facial image as input and predicts character's controller values.

3.1 Data Acquisition

Our first solution framework employs four databases: (1) Human expression database (HED), (2) Character Expression Database-3D (CED-3D), (3) Character Expression Database-2D (CED-2D), and (4) Human Expression Video Database (HEVD). The specifics of these datasets are as follows:

Human Expression Database (HED): We created the HED by combining four publicly accessible labeled face expression datasets: (a) the Extended Cohn-Kanade database (CK+)[18], (b) the Denver Intensity of Spontaneous Facial Actions (DISFA) database[19], (c) the Karolinska Directed Emotional Faces (KDEF)[8] , and (d) the MMI database[23]. The HED database contains about 100K images with seven labeled expressions: anger, sadness, joy, neutral, disgust, fear, and surprise.

Character Expression Database-3D (CED-3D): We use FERG-3D-DB [2], which has about 40000 annotated examples for four stylized characters. Each example is a set of controller values that, when applied to the 3D rig, produce a certain facial emotion. Each character's expressions are classified into seven subgroups: anger, sadness, joy, neutral, disgust, fear, and surprise.

Character Expression Database-2D (CED-2D): We render the 3D character rigs in the CED-3D into 2D images. When rendering frames, we mark 49 landmarks on the characters' texture and save the geometric information for each image. After the characters' faces are cropped and registered with 49 facial landmarks [33], the images are resized to 256 by 256 pixels for analysis.

Human Expression Video Database (HEVD): For training, we make use of the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [17], which includes expressions from 24 professional performers. Anger, sadness, joy, neutral, disgust, fear, and surprise are all expressions. Each expression has two levels of emotional intensity, with an additional neutral expression.

3.2 Human-Characters Match

3.2.1 Human-Primary Character Match. To find the best-matched primary character image and HED human face pairs, we proposed a two-step filtering approach: given an input human face query, first, retrieve the rendered primary character dataset and find the top 30 character images with the closest emotional distance; then retrieve matched character image with closest geometric distance among the 30 candidate images.

Emotional Distance We use the 512 dimension vector from the fully connected layers of the expression classification network proposed in DeepExpr[3] as the Emotion Feature Vector. The expression classification network is first trained on HED and then fine-tuned on CED-2D. For each human-primary character pair, we measure the Jensen—Shannon divergence in Eqn.1 as their Emotion Distance.

$$JSD(H||C) = \frac{1}{2}D(H||M) + \frac{1}{2}D(C||M) \tag{1}$$

where $M = \frac{1}{2}(H + C)$, $D(H||M)$ and $D(C||M)$ represents the Kullback—Leibler divergence.

Geometric Distance We register 49 facial landmarks from average frontal faces using an affine transformation. Then, we normalize the following geometric distance as Geometric Feature Vector: mouth width (left mouth corner to right mouth corner distance), closed mouth height (distance is vertical between the upper and the lower lip), nose width (distance is horizontal between leftmost and rightmost nose landmarks), left/right eyebrow height (distance is vertical between the top of the eyebrow and center of the eye), left/right eyelid height (distance is vertical between the top of an eye and bottom of the eye), and left/right lip height (distance is vertical between the lip corner from the lower eyelid). For each human-primary character pair, the L2 norm distance between their Geometric Feature Vectors is used as Geometric Distance.

The highlight of our two-step filtering strategy is to further improve the efficiency and accuracy of the retrieved results in a perceptually valid and geometrically consistent way. For example, some expressions, e.g., sad and disgust, the emotional distance is close, while others, e.g., fear and surprise, the geometric distance is close. Thus, the one-step solution combining the emotional distance and the geometric distance together could result in retrieving images with incorrect emotion or far geometry distance.

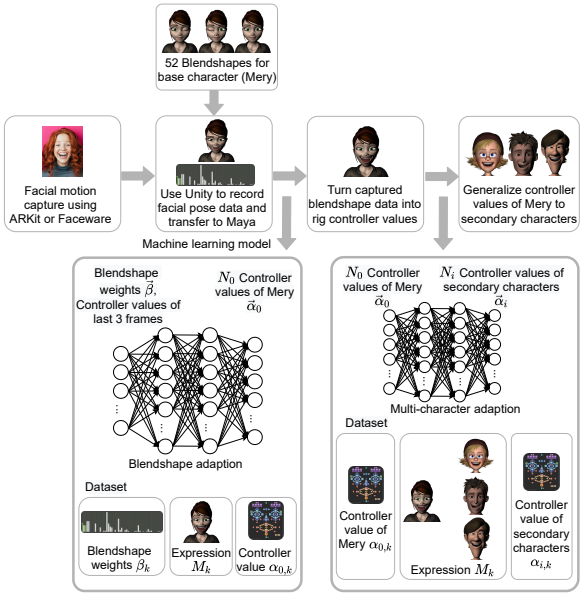


Figure 2: Second solution (Blendshape), which takes blendshape weights as input and produces characters' expressive controller values.

3.2.2 Character-Character Match. Primary character to secondary character matching pairs are retrieved in a similar manner: given an input primary character face query, first retrieve rendered secondary character dataset and find the top 30 character images with the closest emotional distance; then retrieve the matched character image with closest geometric distance among the 30 candidate images.

3.3 3D Expression Transfer Network

We train a 3D expression transfer network, which takes human faces in HED as input and generates controller values for the primary character. The human-primary character pairs are used as ground truth during training. The loss function is formulated in eqn. 2,

where $\boldsymbol{\alpha}$ is the controller value generated by 3D expression transfer network, $\boldsymbol{\alpha}^{'}$ is the controller value in human-primary character pair and a_i and $a_i^{'}$ are the i-th item in $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^{'}$.

$$HH\left(\boldsymbol{\alpha}, \boldsymbol{\alpha}^{'}\right) = -\sum_{i} a_i' \log\left(\text{softmax}\left(a_i\right)\right) \tag{2}$$

3.4 Character to Character Transfer

Our multi-character adaptation model is designed to learn to map the controller values of the primary character to the secondary character automatically. We used a lightweight method instead of training a new network for each additional secondary character. We create a separate multilayer perceptron (MLP) for each secondary character, which refers to a two-hidden-layer neural network with N output nodes, M input nodes, and a two-hidden-layer neural network with ReLU activation, where N and M are the number of controller values of the secondary and primary characters, respectively. These networks (together referred to as the Multi-character adaption network) are trained in parallel and then enhanced at the conclusion of the 3D expression transfer network to imprint the input human expression on numerous stylized characters at the same time.

Note that our multi-character adaptation model is inspired by the C-MLP model of the ExprGen [2] system. The key difference is the methodology of producing a set of matching primary and secondary character pairs. Our two-step filtering strategy further ensures retrieval results that are both perceptually and geometrically consistent. The minor difference is our model is two hidden-layer neural networks, whereas C-MLP model is one hidden-layer neural network.

3.5 Frame Interpolation

In order to create smooth transitions between images, the convention in animation requires at least 24 frames per second (FPS). However, our first solution only generates at 3 to 4 FPS, since processing each individual frame in deep neural networks takes a considerable amount of time. Thus, the animation would no longer look live or realistic and the user would see images jumping from one expression to another. We simply use linear interpolation to inbetween frames and create more frames to fill the spaces between the original. This allowed to increase this frame rate to 24 FPS without fully solving for the additional frames.

4 SECOND SOLUTION (BLENDSHAPE)

The first solution described above is based on static images. It has several drawbacks: (1) The deep neural network leads to a delay when generating keyframes. (2) Mapping facial geometry features to emotion space without constraint often produces flickering results. In our second method, we designed a lightweight network which takes blendshape weights as input and generates expressive controller values.

4.1 Data Collection

Our blendshape-based framework makes use of two databases: (1) Character Blendshape Database (CBD), and (2) Character Controller Value Database (CCVD).

Character Blendshape Database (CBD): We collect the blendshape weights of weak emotion intensity videoes in HEVD using Faceware[7]. Faceware is a real-time face tracking system that can effectively caption the geometry feature of human facial poses and provide weights of a range of shapes. We first calibrate human faces in neutral expression pose and recover the optimized blendshape coefficients (eg. brow down left, brow up left) frame by frame with animation tuning.

Character Controller Value Database (CCVD): We collect controller value of weak emotion intensity videos in HEVD using the solution discussed in Section 3.

4.2 Human Expression to Primary Character

Both the blendshape weights and controller values can animate the characters' expression and some of the parameters are correlated. For example, "mouthClose" is a coefficient describing closure of the lips independent of jaw position in the blendshape, and controller values provide "up_lf_lip_inout" and "up_rf_lip_inout" to enable more precise manipulation. However, the frame-by-frame mapping method can cause flickering results. To overcome the inconsistency between frames, we train a blendshape adaption network, which takes blendshape weights and controller values over the last three frames as input and generates controller values in the current frame. The controller values in CCVD are used as ground truth during training. The training process can be formulated as follows:

Given an input vector $\boldsymbol{\beta}$ which consists of blendshape weights in the current frame and controller value from the last three frames, the blendshape adaption network $\boldsymbol{B}$ outputs the controller value in the current frame vector is $\boldsymbol{\alpha}^{'} = \boldsymbol{B}(\boldsymbol{\beta})$.

The loss function of the blendshape adaption network can be formulated as eqn.3, where $\boldsymbol{\alpha}$ is the controller value from CCVD, $\boldsymbol{\alpha}^{'}$ is the controller value generated from our blendshape adaption network and, a_i and $a_i^{'}$ are the i-th item in $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^{'}$.

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\alpha}^{'}) = \sum_{i=0}^{n} \left(a_i - a_i^{'}\right)^2 \tag{3}$$

4.3 Character to Character Transfer

Our multi-character adaptation model for our system is designed to learn to map the controller values of the primary character to the secondary character automatically. We create a separate multilayer perceptron (MLP) for each secondary character, which refers to a two-hidden-layer neural network with N output nodes, M input nodes, and a two-hidden-layer neural network with ReLU activation, where N and M are the number of controller values of the secondary and primary characters, respectively. We use the Gradient descent with a mini-batch size of 10 and a learning rate of 0.01 to minimize the square loss between the input and output parameters. These networks are trained in parallel and then enhanced at the conclusion of the 3D expression transfer network to imprint the input human expression on numerous stylized characters at the same time.

5 EVALUATION

By comparing the expression recognition accuracy of the interpolation-based solution , blendshape-based solutions to that of the commercial product Faceware, we were able to assess their effectiveness.

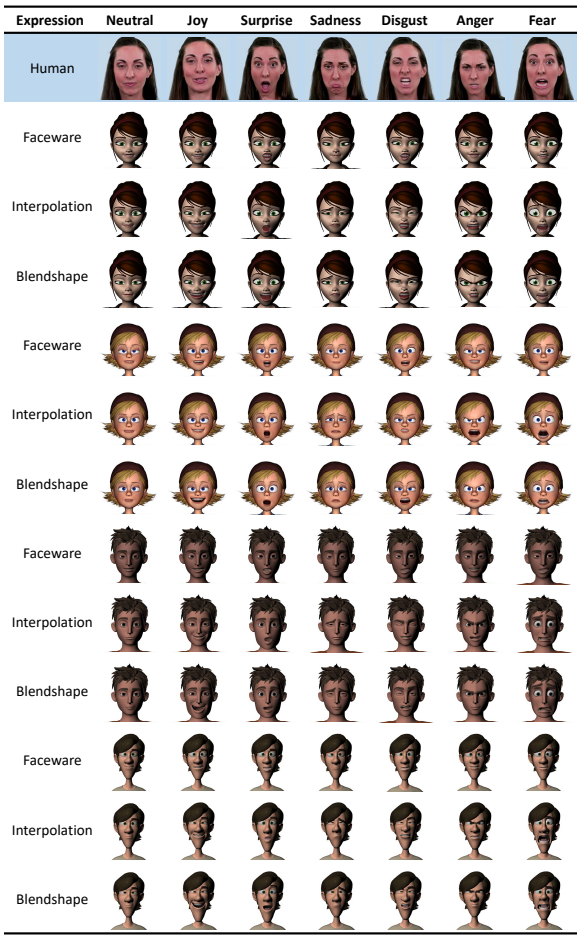


Figure 3: Resulting images generated from Faceware, Interpolation, and Blendshape for all four characters. The uppermost image in each column is the input image.

Faceware was selected for two purposes: (1) It is the only viable and equivalent system with the same input and output modalities as our platforms. Because the results of Faceshift Studio/ARkit require depth sensors to catch human facial motion, we did not compare them. (2) We also used Faceware Live Client for Unity to produce the blendshape weight sequence for our second solution.

In all the subsequent figures, we use the 2D rendered images to represent the 3D characters and employ the expression categories in the following manner: N = neutral, A = anger, Sa = sadness, F = fear, D = disgust, J = joy, Su = surprise.

5.1 Participant

We recruited 24 participants from Shanghai Jiao Tong University to complete all conditions of this study. The average age of the participants was 21 years, ranging between 19 and 24 years old; 12 were men. They were naïve to the purposes of the experiment.

5.2 Material

5.2.1 Animation clips. We first took alternative 4 sets of recordings from 2 male and 2 female actors from HEVD. The dialogue was recorded with 7 basic emotions. The dialogue used for each set of recordings is the same (e.g., "Dogs are sitting by the door"), and we used a face-only format (face, but no voice). Each recording lasted about 3 or 4 seconds. Then, we ran Faceware, Interpolation, and Blendshape solutions to create 4×7 animation clips for 4 characters (Mery, Bonnie, Ray & Malcolm).

5.2.2 Images. We used one frame in each recording discussed above and retrieved the same frame via Faceware, Interpolation, and Blendshape solutions to create 4×7 images for 4 characters (Mery, Bonnie, Ray & Malcolm).

5.3 Design

The experiment utilized 4 characters (Mery, Bonnie, Ray & Malcolm) $\times$ 7 emotions (Neutral, Anger, Sadness, Fear, Disgust, Happiness, & Surprise) $\times$ 3 capturing methods (Faceware, Interpolation & Blendshape) $\times$ 2 media (Image & Video) in a mixed design, with a between-subject design for media, but a within-subject design regarding characters, emotions, and tracking methods.

Each participant took part in 91 trials to evaluate the input human expression (7 emotions = 7 trials), the generated primary character expression (7 emotions $\times$ 3 capturing methods = 21 trials), and the expression transfer results on different three stylized characters (63 trials). Thus, there were 2184 trials in total. To avoid fatigue or carry-over effects, images or video clips were presented to the participants in random order.

5.4 Procedure

Participants were first presented with an information sheet and asked to sign a corresponding consent form. They were randomly assigned to either an image condition or a video condition. They were instructed to view an image or animation clip and then asked to answer three questions:

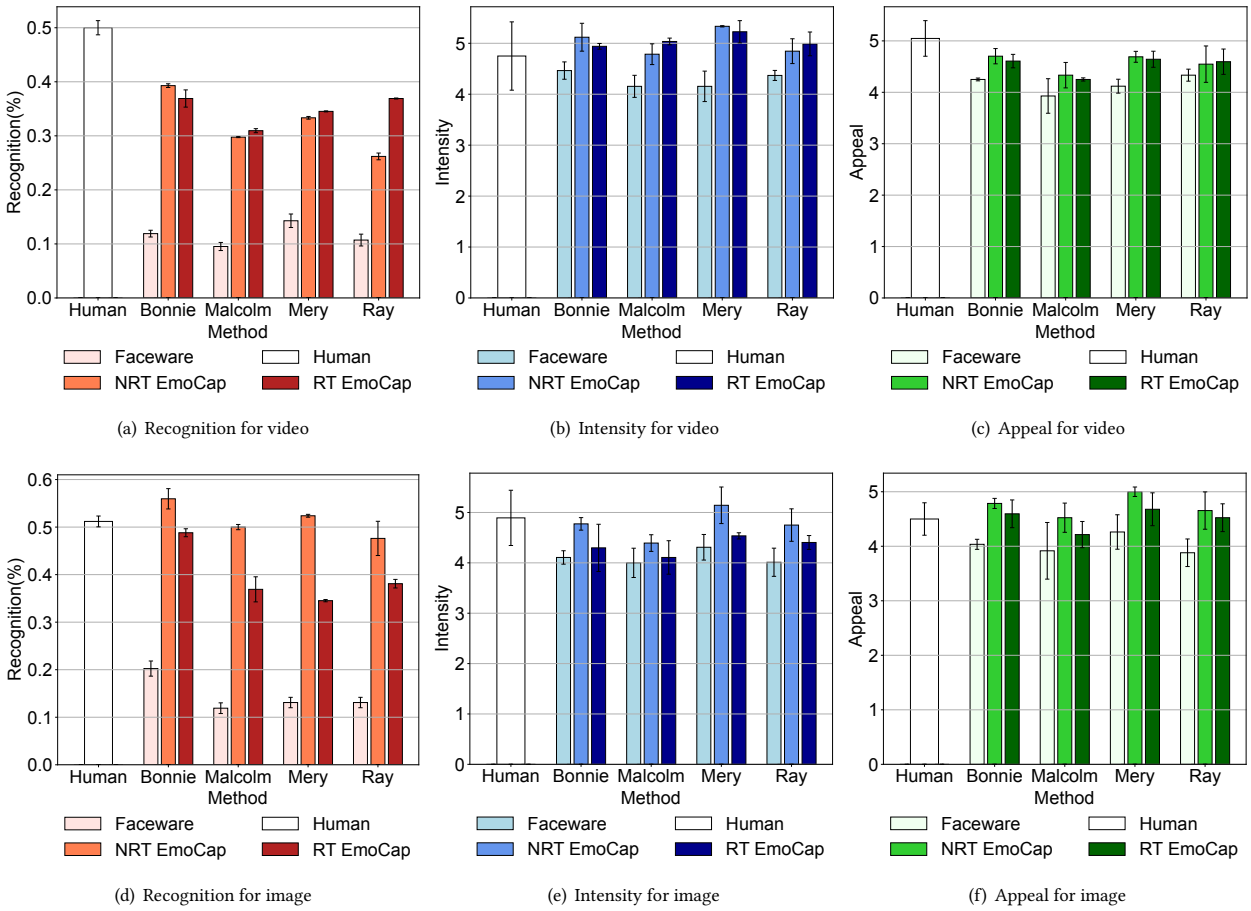
- "Which expression did the character depict?" Participants were asked to select one of the words: Neutral, Anger, Sadness, Fear, Disgust, Happiness, Surprise or Other.
- "How intense was the indicated emotion depicted by the character?" Participants rated the intensity on a scale from 1 to 7, where 1 represents a rating of "Not at all", and 7 represents "Extremely".
- "How attractive was the character overall?" Participants rated attractiveness on a scale from 1 to 7, where 1 represents a rating of "Not at all", and 7 represents "Extremely".

Each participant undertook one practice trial where they could ask questions, and then undertook 91 measured trials.

The participants were paid 10 GBP amount. The experiment took about 30 minutes. The experiment was approved by Shanghai Jiao Tong University Research Ethics Committee.

5.5 Scoring & results

We applied separate repeated measures Analysis of Variances (ANOVAs) for both video and image, looking at the results on recognition, intensity, and attractiveness. Each ANOVA had the within-participants factors character (4), emotion (7), and tracking methods (3). There were no outliers, and the data was normally distributed for each condition as assessed by boxplot and Shapiro–Wilk test ($p > 0.05$),



(a) Recognition for video (b) Intensity for video (c) Appeal for video

(d) Recognition for image (e) Intensity for image (f) Appeal for image

Figure 4: Mean for each tracking method and character on recognition, appeal, and intensity. Error bars show standard deviation.

respectively. We conducted Mauchly's test to assess the sphericity of the data, and whenever it is violated, we report results applied with Greenhouse-Geisser correction and marked with an asterisk "*". Bonferroni test was performed as Post Hoc test for multiple comparisons of means.

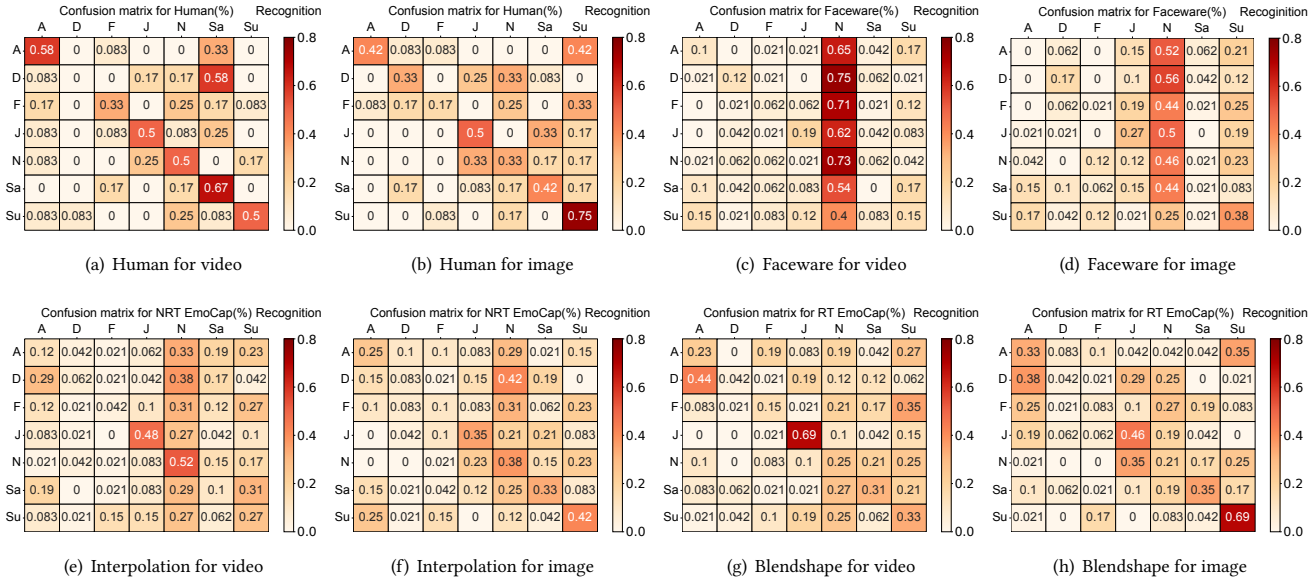
5.5.1 Recognition. For the recognition of expressions, responses were converted to scores, "1" for correct or "0" for incorrect, and then averaged over stimuli repetitions.

Videos Figure 4(a) shows the comparison of average scores obtained for three tracking methods across four characters. The average score over all characters for Interpolation ($M = .321$) and Blendshape ($M = .348$) are significantly higher than the average score for Faceware ($M = .116$). Firstly, the main effect of the tracking method was significant, $F(2, 22) = 26.531, p < .001$. Bonferroni post-hoc comparisons indicated the mean recognition rates for Faceware is significantly lower than Interpolation, $p < .001$ and Blendshape, $p < .001$. However, the mean for Interpolation did not significantly differ from Blendshape, $p > .05$. Secondly, tracking methods × characters interaction, tracking methods × emotions interaction and tracking methods × characters × emotions interaction were not significant, $F(6, 66) = 1.004, p = .43$, $F(3.654, 40.192) = 1.717, p = .17^*$, $F(6.732, 74.057) = .999, p = .437^*$, respectively.

Images Figure 4(d) shows the results for images on the recognition scores. It confirms that the average score over all characters for Interpolation ($M = .515$) and Blendshape ($M = .396$) are significantly higher than the average score for Faceware ($M = .146$). Firstly, the main effect of tracking method was significant, $F(2, 22) = 42.094, p < .001$. Bonferroni post-hoc comparisons indicated the mean recognition rates for Faceware are significantly lower than Interpolation, $p < .001$ and Blendshape, $p < .001$. However, the mean for Interpolation did not significantly differ from Blendshape, $p = .069$. Secondly, tracking methods × characters interaction, tracking methods × emotions interaction were not significant, $F(6, 66) = .403, p = .875$, $F(3.217, 35.39) = 2.409, p = .079^*$, respectively. However, tracking methods × characters × emotions interaction was significant, $F(6.793, 74.727) = 2.639, p = .018^*$.

5.5.2 Intensity. As expected, intensity ratings for our Interpolation and Blendshape systems were high, because the facial expressions for stylized characters are generally exaggerated.

Videos Figure 4(b) shows the mean intensity ratings for three tracking methods across four characters. The average score over



(a) Human for video (b) Human for image (c) Faceware for video (d) Faceware for image

(e) Interpolation for video (f) Interpolation for image (g) Blendshape for video (h) Blendshape for image

Figure 5: Confusion matrix for perceived expression recognition (%) for basic expression classes.

all characters for Interpolation ($M = 5.021$) and Blendshape ($M = 5.048$) are significantly higher than the average score for Faceware ($M = 4.286$). Firstly, the main effect of tracking method was significant, $F(2, 22) = 62.811, p < .001$. Bonferroni post-hoc comparisons indicated the mean intensity ratings for Faceware is significantly lower than Interpolation, $p = .016$ and Blendshape, $p = .017$. However, the mean for Interpolation did not significantly differ from Blendshape, $p > .05$. Secondly, tracking methods × characters interaction, tracking methods × emotions interaction and tracking methods × characters × emotions interaction were not significant, $F(2.045, 22.496) = 1.745, p = .197^*$, $F(4.91, 54.009) = .412, p = .835^*$, $F(7.339, 80.725) = 1.13, p = .353^*$, respectively.

Images Figure 4(e) shows the results for images on the intensity ratings. The average score over all characters for Interpolation ($M = 4.765$) are significantly higher than the average score for Faceware ($M = 4.107$) and Blendshape ($M = 4.336$). Firstly, the main effect of tracking method was significant, $F(2, 22) = 6.259, p = .007$. Bonferroni post-hoc comparisons indicated the mean intensity ratings for Blendshape is significantly higher than Interpolation, $p = .026$ and Faceware, $p = .037$. However, the mean for Blendshape did not significantly differ from Faceware, $p = .83$. Secondly, tracking methods × characters interaction, tracking methods × emotions interaction and tracking methods × characters × emotions interaction were not significant, $F(2.827, 31.1) = .671, p = .568^*$, $F(4.387, 48.255) = .819, p = .529^*$, $F(6.855, 75.4) = 1.258, p = .283^*$, respectively.

5.5.3 Appeal. We look at the effect of tracking methods on appeal ratings across all characters.

Videos Figure 4(c) shows the mean appeal ratings for three tracking methods across four characters. The average score over all characters for Interpolation ($M = 4.568$) are significantly higher than the average score for Faceware ($M = 4.158$). However, the mean for Blendshape ($M = 4.524$) is not significantly different than either these two conditions. Firstly, the main effect of tracking method was significant, $F(2, 22) = 6.259, p = .007$. Bonferroni post-hoc comparisons indicated the mean appeal ratings for Interpolation is significantly higher than Faceware, $p = .036$. Secondly, tracking methods × characters interaction, tracking methods × emotions interaction and tracking methods × characters × emotions interaction were not significant, $F(3.211, 35.317) = .443, p = .737$, $F(4.88, 53.68) = 1.621, p = .172^*$, $F(7.702, 84.725) = 1.368, p = .224^*$, respectively.

Images Figure 4(f) shows the results for images on the appeal ratings. The average score over all characters for Interpolation ($M = 4.741$) and the Blendshape ($M = 4.503$) are significantly higher than the average score for Faceware ($M = 4.024$). Firstly, the main effect of the tracking method was significant, $F(1.209, 13.304) = 12.859, p = .002$. Bonferroni post-hoc comparisons indicated the mean appeal ratings for Faceware is significantly lower than Interpolation, $p = .005$ and Blendshape, $p = .045$. The mean for Blendshape is also significantly different from Interpolation, $p = .01$. Secondly, tracking methods × characters interaction, tracking methods × emotions interaction, and tracking methods × characters × emotions interaction were not significant, $F(2.776, 30.54) = .264, p = .836^*$, $F(4.62, 50.825) = 1.289, p = .285^*$, and $F(5.879, 64.673) = .791, p = .578^*$.

6 DISCUSSION

6.1 Expression recognition

The main effect of emotions was significant, $F(2.742, 60.329) = 5.531, p = .003$, according to our preliminary data on expression recognition. Thus, we look into participants' rating for seven expression classes. Figure 5 depicts the confusion matrix for each expression class's perceived expression recognition. For a specific

row (e.g., anger) in each sub-figure, the columns show the percentage (e.g., averaging nearly over all observed individual anger expressions) of respondents agreeing on the associated expression classes.

Human We included the human videos and images as the ground truth. The confusion matrix of observed expression recognition rate for humans is shown in Figures 5(a) and 5(b). Surprise and joy are highly accurate, while fear and disgust are extremely difficult for people to recognize and express. This is similar to Aneja et al.'s [2] prior result of evaluating human photos as input. We also noticed that character expression identification accuracies are sometimes higher than human, which could be due to the characters' simpler geometry and stylization, which makes the expressions simpler to discern.

Faceware, Interpolation & Blendshape Figure 5(c) and Figure 5(d) show that the majority of expressions are incorrectly perceived as neutral for Faceware condition. This indicates blendshape-based approaches often produce ambiguous expressions, due to the constraints of correspondence mapping. Figure 5(e) to Figure 4(d) show that our Interpolation and Blendshape system results in more precise expression transfer for the majority of expression classes when compared to Faceware. The most prevalent blunders are mixing up fear and surprise, as well as disgust and fury. Because the confounded statements have similar looking geometric arrangements, these errors are intuitively understandable. Disgust and fear were the least accurate results, as these reactions are difficult to discern in both human and character depictions.

6.2 Videos & images

We took videos from the RAVDESS [17] stimuli. We note our expression recognition ratings for human is different from the results reported by RAVDESS where their validation tasks were used with North American participants. There is strong evidence for a 'in-group' advantage in emotion recognition, with accuracy being higher for facial expressions and identified by people from the same cultural group [6].

6.3 Characters

We used 'Mery' as the base character, 'Bonnie', 'Ray' & 'Malcolm' as the secondary character. Our results show the tracking methods × emotions interaction were insignificant regarding expression recognition, intensity, and appeal. It indicates our solutions were effective irrespective of whether the character is primary or secondary.

We also benchmarked the retargeting ability and include the inference error per controller value on the test dataset. The error is calculated by the L2 distance between the predicted controller value and the ground truth. Error on Bonnie, Malcolm, and Ray test datasets are 0.033, 0.091, and 0.025, respectively.

6.4 Emotional audio-visual dataset

We built the stylized character emotional audio-visual dataset based on RAVDESS [17] via our real-time blendshape based system. This is the first video dataset with animated stylized characters (2 male and 2 female) talking with seven basic emotions. The set consists of synchronized 3D rig parameters, synchronized blendshape weights, synchronized RAVDESS input video and synchronized RAVDESS audio presented in North American English. All available in high-definition formats. Our user study revealed test-retest reliability and high rates of emotional validity. This set may be of interest to a wide range of technologists and researchers.

6.5 Comparison, limitation and future work

Compared to the current state-of-the-art technique, we present the first **real time** facial expressions capture approach for **stylized character** in **a geometrically consistent and perceptually valid** way. (1) Both DeepExpr [3], and ExprGen [2], similar to our step 1 Interpolation, include emotion recognition framework. However, they processing each individual frame in deep neural networks takes a lot of time and thus cannot perform in real-time. However, our interpolation solution is based on a lightweight method. Additionally, for the second blendshape-based method, we process 24 FPS in order to compare with the Faceware system. Note that the exact inference time for blendshape-based methods and Faceware is 9 ms and 24 ms, respectively. (2) Some real-time expression transfer frameworks [29, 30] focus on transferring a source human's facial expressions to a target human face, instead of stylized and expressive character expressions. (3) In other facial motion capture applications (e.g., [14, 16]), they would drive stylized characters, but lack of the expressive quality and perceptual validity created by professional artists.

The focus of our system is to generate rig parameters/ controller value to drive a stylized characters animation, without tedious handcrafted work by professional artists. However, we note producing high quality animation videos require a lot of aspects which is beyond the scope of this paper, such as, 3D modeling, texturing, and lighting etc. We plan to improve the rendering results in future work. Our future work also plans to add the concept of a universal base character rig that is powerful enough to create a full range of expressions and can be readily expanded to any new secondary characters.

7 CONCLUSION

In summary, for the first time, we contribute a real time system that captures human facial expressions to drive a stylized character in a perceptually correct and geometrically cohesive fashion. We conducted a user survey to show that our system creates more perceptually accurate expressions than popular commercially accessible software applications, such as, Faceware.

The ease of use of the our system and the expressiveness of our resulting animations can potentially improve the effectiveness of visual storytelling in areas of online marketing, gaming, animated films, and immersive experiences. Our system can also be used in real-time ("live") animation situations, where facial expression is a useful input modality and amateurs can communicate stories with expressive animation by capturing their own performances.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (NSFC, NO. 62102255), CCF-Tencent Open Research Fund (RAGR20220128), European Union's Horizon 2020 research and innovation programme (NO.101017779).

REFERENCES

[1] 2019. Apple ARKit. https://developer.apple.com/arkit/. (2019).

[2] Deepali Aneja, Bindita Chaudhuri, Alex Colburn, Gary Faigin, Linda Shapiro, and Barbara Mones. 2018. Learning to generate 3D stylized character expressions from humans. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 160–169.

[3] Deepali Aneja, Alex Colburn, Gary Faigin, Linda Shapiro, and Barbara Mones. 2016. Modeling Stylized Character Expressions via Deep Learning. In *Asian Conference on Computer Vision*. Springer, 136–153.

[4] Keyu Chen, Jianmin Zheng, Jianfei Cai, and Juyong Zhang. 2020. Modeling Caricature Expressions by 3D Blendshape and Dynamic Texture. *arXiv preprint arXiv:2008.05714* (2020).

[5] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 2020. 3D Morphable Face Models—Past, Present, and Future. *ACM Trans. Graph.* 39, 5, Article 157 (June 2020), 38 pages. https://doi.org/10.1145/3395208

[6] Hillary Anger Elfenbein and Nalini Ambady. 2002. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin* 128, 2 (2002), 203.

[7] Inc. Faceware Technologies. 2021. Award-winning, Gold Standard Facial Motion Capture Solutions. https://facewaretech.com/

[8] Ellen Goeleven, Rudi De Raedt, Lemke Leyman, and Bruno Verschuere. 2008. The Karolinska Directed Emotional Faces: A validation study. *Cognition and Emotion* 22, 6 (2008), 1094–1118.

[9] Ju Hee Han, Jee-In Kim, Hyungseok Kim, and Jang Won Suh. 2021. Generate Individually Optimized Blendshapes. In *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 114–120.

[10] Jennifer Hyde, Elizabeth J Carter, Sara Kiesler, and Jessica K Hodgins. 2015. Using an interactive avatar's facial expressiveness to increase persuasiveness and socialness. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1719–1728.

[11] John Lasseter. 1987. Principles of traditional animation applied to 3D computer animation. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*. 35–44.

[12] J. P. Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Fred Pighin, and Zhigang Deng. 2014. Practice and Theory of Blendshape Facial Models. In *Eurographics 2014 - State of the Art Reports*, Sylvain Lefebvre and Michela Spagnuolo (Eds.). The Eurographics Association. https://doi.org/10.2312/egst.20141042

[13] John P Lewis, Matt Cordner, and Nickson Fong. 2000. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 165–172.

[14] Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. 2015. Facial performance sensing head-mounted display. *ACM Transactions on Graphics (ToG)* 34, 4 (2015), 1–9.

[15] Hao Li, Thibaut Weise, and Mark Pauly. 2010. Example-based facial rigging. *Acm transactions on graphics (tog)* 29, 4 (2010), 1–6.

[16] Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. 2013. Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph.* 32, 4 (2013), 42–1.

[17] Steven R. Livingstone and Frank A. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE* 13 (05 2018), 1–35.

[18] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 94–101.

[19] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. 2013. DISFA: A Spontaneous Facial Action Intensity Database. *IEEE Transactions on Affective Computing* 4, 2 (2013), 151–160.

[20] Hayato Onizuka, Diego Thomas, Hideaki Uchiyama, and Rin-ichiro Taniguchi. 2019. Landmark-guided deformation transfer of template facial expressions for automatic generation of avatar blendshapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

[21] Ye Pan, Ruisi Zhang, Shengran Cheng, Shuai Tan, Yu Ding, Kenny Mitchell, and Xubo Yang. 2023. Emotional Voice Puppetry. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (2023), 2527–2535.

[22] Ye Pan, Ruisi Zhang, Jingying Wang, Nengfu Chen, Yilin Qiu, Yu Ding, and Kenny Mitchell. 2022. MienCap: Performance-based Facial Animation with Live Mood Dynamics. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces*. IEEE, 654–655.

[23] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. 2005. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*.

[24] Tom Porter and Galyn Susman. 2000. On site: Creating lifelike characters in pixar movies. *Commun. ACM* 43, 1 (2000), 25.

[25] Sarah Radzihovsky, Fernando de Goes, and Mark Meyer. 2020. FaceBaker: Baking Character Facial Rigs with Machine Learning. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Talks*. 1–2.

[26] Roger Blanco i Ribera, Eduard Zell, John P Lewis, Junyong Noh, and Mario Botsch. 2017. Facial retargeting with automatic range of motion alignment. *ACM Transactions on graphics (TOG)* 36, 4 (2017), 1–12.

[27] Yeongho Seol, Jaewoo Seo, Paul Hyunjin Kim, John P Lewis, and Junyong Noh. 2011. Artist friendly facial animation retargeting. *ACM Transactions on Graphics (TOG)* 30, 6 (2011), 1–10.

[28] Robert W Sumner and Jovan Popović. 2004. Deformation transfer for triangle meshes. *ACM Transactions on graphics (TOG)* 23, 3 (2004), 399–405.

[29] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. 2015. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.* 34, 6 (2015), 183–1.

[30] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2387–2395.

[31] Frank Thomas, Ollie Johnston, and Frank Thomas. 1995. *The illusion of life: Disney animation*. Hyperion New York.

[32] Pisut Wisessing, Katja Zibrek, Douglas W Cunningham, John Dingliana, and Rachel McDonnell. 2020. Enlighten Me: Importance of Brightness and Shadow for Character Emotion and Appeal. *ACM Transactions on Graphics (TOG)* 39, 3 (2020), 1–12.

[33] X. Xiong and F. De la Torre. 2013. Supervised Descent Method and Its Applications to Face Alignment. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*. 532–539. https://doi.org/10.1109/CVPR.2013.75

[34] Juyong Zhang, Keyu Chen, and Jianmin Zheng. 2020. Facial Expression Retargeting from Human to Avatar Made Easy. *IEEE Transactions on Visualization and Computer Graphics* (2020).

[35] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. 2018. State of the art on monocular 3D face reconstruction, tracking, and applications. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 523–550.