A Comparative Study of Four 3D Facial Animation Methods: Skeleton, Blendshape, Audio-Driven, and Vision-Based Capture

Mingzhu Wei(✉), Nicoletta Adamo, Nandhini Giri, and Yingjie Chen

Purdue University, West Lafayette, IN, USA
wei376@purdue.edu

Abstract. In this paper, the authors explore different approaches to animating 3D facial emotions, some of which use manual keyframe animation and some of which use machine learning. To compare approaches the authors conducted an experiment consisting of side-by-side comparisons of animation clips generated by skeleton, blendshape, audio-driven, and vision-based capture facial animation techniques. Ninety-five participants viewed twenty face animation clips of characters expressing five distinct emotions (anger, sadness, happiness, fear, neutral), which were created using the four different facial animation techniques. After viewing each clip, the participants were asked to identify the emotions that the characters appeared to be conveying and rate their naturalness. Findings showed that the naturalness ratings of the happy emotion produced by the four methods tended to be consistent, whereas the naturalness ratings of the fear emotion created with skeletal animation were significantly higher than the other methods. Recognition of sad and neutral emotions were very low for all methods as compared to the other emotions. Overall, the skeleton approach had significantly higher ratings for naturalness and higher recognition rate than the other methods.

Keywords: Facial Animation · Emotion Recognition · Animation Perception · Affective Animated Agents

1 Introduction

In computer graphics, the ability to represent the human face and animate the nuances of facial emotions remains a serious challenge. Computer facial animation has become more and more important, as affective animated agents capable of expressing believable facial emotions are increasingly being used in many areas and industries, such as games, film, medicine, education, social media and more.

This work is supported by NSF-IIS award #1821894: Multimodal Affective Pedagogical Agents for Different Types of Learners.

A. L. Brooks (Ed.): ArtsIT 2022, LNICST 479, pp. 36–50, 2023.
https://doi.org/10.1007/978-3-031-28993-4_3

Current 3D facial animation creation approaches can be divided into three categories, namely traditional keyframe animation, machine-learning based auto-facial-animation, and performance facial animation. Facial animation techniques have been discussed extensively in the existing literature [12,16], but most of the published works primarily talk about concept and theory. To date, there is not enough research that compares outputs generated by different facial animation methods with the goal of providing guidelines for animation researchers and practitioners. The work reported in the paper aimed to fill this research gap. It compared current facial animation methods to examine the differences in people perception of the facial emotions generated by each method. In particular, the work reported in the paper sought to answer the following research questions:

1. R_1: Which of the four facial animation methods (Skeleton, Blendshape, Audio-Driven or Visual-Driven) creates the most natural and clearly recognizable facial emotion animations as perceived by participants?
2. R_2: Are there any differences in perceived emotion naturalness based on participants' gender?
3. R_3: Are there any differences in perceived emotion naturalness based on participants' animation experience?
4. R_4: Are there any differences in perceived naturalness based on emotion type?

2 Related Work

2.1 Keyframe Skeleton Facial Animation

The principle behind key-frame or key-pose animation is that the intended facial expressions are specified at different key points in time (e.g., the keyframes) and the frames in between these key frames are generated by a computer algorithm. According to [14], the intermediate forms of the surface are achieved by interpolating each vertex between its two extreme positions. Parke [15] was the first person to demonstrate the use of this approach to produce viable facial animation. From an animator's perspective, facial animation can be thought of as the manipulation of the facial rig over time. The rig includes the facial skeleton, the animation controls for the facial model, and the animator's interface to those controls [24]. For example, Hanrahan and Sturman's system [6] is an early example. The goal is to manipulate the rig to create geometric data describing the face in at least two different expressions at two different key points in time. Then, as a function of time, a single control parameter, the interpolation coefficient, is employed to alter the face from one expression to the other and hence generate the facial animation [14].

Keyframe skeleton facial animation has some limitations. First, the amount and variety of expression poses are directly proportional to the range of expression control offered by the facial skeletal rig. Second, key-frame animation necessitates a thorough geometry specification for each key facial expression, which is a time-consuming task that requires high level of expertise and artistic talent [14].

2.2 Morph Target or BlendShape Facial Animation

In blendshape animation, the "morph target" (also called blendshape) is a deformed version of a "base" shape. When applied to the human face, the face is first modeled with a neutral expression and a "target deformation" is then created (e.g. a raised eyebrow). When the face is being animated, the animator can smoothly morph (or "blend") between the base shape and one or several morph targets, which represent different facial articulations. More specifically, the neutral (base) facial mesh is blended with, or summed linearly with topologically conforming shape primitives of target faces that reflect user-defined facial articulations. In other words, blendshape animation linearly combines over time a set of artist-created facial expressions.

Although the blendshape approach is conceptually straightforward, creating a blendshape face model is a time-consuming task which requires expertise in sculpting and 3D modeling. For example, hundreds of facial blendshape targets may be required to create a facial blend shape model for professional animation [11]. If the model topology needs to be altered, all morph targets must be recreated [16]. Further, the quality of the final animation produced with blendshapes depends to a great extent on the talent of the artist that models the morph targets.

2.3 Machine Learning Face Animation Approaches

There are two popular machine learning methods for creating facial animation: audio-driven facial animation, and computer vision based facial animation. One of the most difficult aspects of machine learning is describing the learning task. Tasks such as finding appropriate inputs and defining the form of outputs and training set directly affect the outcome. However, machine learning methods can speed up content creation significantly compared to manual methods.

Audio-Driven Facial Animation. The goal of audio-driven facial animation is to automatically generate production quality facial animation given only audio (speech) as input. Various methods for synthesizing a face model based on spoken recordings [4,19,29] have been developed. Hidden Markov Models (HMMs) were utilized in early studies on talking heads [26,27].

Nowadays, deep neural networks (DNNs) have been used to improve speech synthesis and facial animation [10,22,30]. The advantage of DNNs is that they can learn extremely nonlinear input-output mappings, while traditional methods were completely linear. The deformation of the human face is very nonlinear, and it is widely recognized that a simple mixing of static expressions frequently produces unacceptable results. Long short-term memory recurrent neural networks (LSTM-RNN) were used by Suwajanakorn [21] to learn photorealistic speech animation, which exhibited some moderate improvement over HMMs.

Nvidia's Audio2Face is a fairly recent successful Audio-Driven Facial Animation 3D software tool. The architecture of the Audio2Face deep neural network consists of three conceptual parts. The network starts with a formant analysis

network, which extracts raw formant information and outputs abstract, time-varying features related to facial animation such as intonation, emphasis, and specific phonemes. Next, the articulation network outputs abstract features representing the desired facial pose. The third network system, output network, produces the final 3D position of the control vertices in the facial mesh. The software allows for creating characters that not only speak but also convey different facial expressions.

Visual-Driven Facial Animation. Most production pipelines now include vision-based performance capture. In facial puppetry the performance of a real actor is captured and transfered (retargeted) to a 3D virtual character. Vision-based performance capture facial animation uses a variety of techniques for generating the 3D facial model and the animation data and for retargeting such data to a character. 2D Face Alignment and 3D Face Reconstruction are often used to generate the 3D facial model. Convolutional neural networks are commonly used in 2D face alignment and 3D face reconstruction to learn 3D morphable model (3DMM) parameters from 2D face photos. Face alignment, which seeks to identify a specific 2D fiducial point [13,31,33] is often employed as a requirement for other facial tasks, including face recognition, and greatly aids 3D face reconstruction [7,35]. However, studies have discovered that 2D alignment has trouble coping with hug poses or occlusion [8,32]. Therefore, some studies of regressing 3DMM parameters using CNNs and fitting 3DMM to 2D pictures have recently gained popularity in reconstructing the corresponding 3D information from a single 2D facial image [9,18,34].

After the 3D face mesh has been reconstructed, facial retargeting methods are used for generating the facial animation. Expressions can be cloned by mapping dense motion fields between the actor and the target rig or by using semantically significant animation parameters. Parameter-based approaches transfer parameters directly from source to destination. The approaches either assume that the source and destination models have the same repression basis [2,3,25,28] or train a linear subspace between them to compensate for global shape changes during transfer [20,23].

2.4 Perception of Facial Expressions

Ongoing research suggests that the human vision system has dedicated mechanisms to perceive facial expressions [17] and categorizes facial perception into three types: holistic, componential and configural perception. Holistic perception models the face as a single unit whose parts cannot be isolated. Componential perception assumes that the human vision system processes different facial features individually. Configural perception models the spatial relations among different facial components (e.g. left eye-right eye, mouth-nose). It is possible that we use all these models when we perceive facial expressions [1]. Ekman and Friesen [5] suggest that there are three types of signals produced by the face: Static, Slow and Rapid. The static signals are the permanent or semi-permanent

aspects of the face such as skin pigmentation, shape, bone structure. The slow signals include facial changes that occur gradually over time, such as permanent wrinkles, changes in muscle tone, skin texture, and even skin coloration. The rapid signals are the temporary changes in facial appearance caused by the movement of facial muscles. The rapid signals are what the majority of people consider when thinking of emotion, for instance, the physical movement of the face to a smile or a frown. All three of these signals play an important role in how a viewer perceives the facial emotion of another being or character. In our study we are concerned with how the rapid signals of the face (e.g. facial animation) affect perception of emotions, as in our experiment the static and slow signals (the character models) were kept the same for each animation method.

3 Framework and Methodology

The study reported in the paper aimed to examine the effects of four facial animation methods on people perception of animated agents facial emotions. The audio-driven and the visual-driven facial animations were generated automatically by software tools with machine learning algorithms, whereas the skeleton and blendshape animations were created by a skilled 3D animator who used commercially available facial rigs. The study used a within-subject design and collected quantitative data. The main hypotheses of the study were the following:

Hypothesis H_{01}**:** All four types of facial animations will receive the same naturalness rating and same recognition rate for each of the five emotions.
Hypothesis H_{11}**:** At least one type of facial animation will receive different emotion recognition rate and different naturalness rating for one or several of the five emotions, implying that participants perceive facial emotions differently depending on the method used to create the facial animations, and that certain methods create facial emotions that are easier to recognize than others and that are perceived as more natural than others.
Hypothesis H_{02}**:** There will be no differences in emotion naturalness rating based on participants' gender.
Hypothesis H_{12}**:** There will be differences in emotion naturalness rating based on participants' gender.
Hypothesis H_{03}**:** There will be no differences in emotion naturalness rating based on participants' animation experience.
Hypothesis H_{13} There will be differences in emotion naturalness rating based on participants' animation experience.
Hypothesis H_{04}**:** There will be no differences in emotion naturalness rating and recognition rate based on emotion type.
Hypothesis H_{14}**:** There will be differences in emotion naturalness rating and recognition rate based on emotion type.

The study had one categorical independent variable (e.g., the facial animation creation method) with four levels (e.g., skeleton, blendshape, audio-driven

method, and visual-driven method). Each method was used to create four distinct animated emotions and one state of no emotion: anger, fear, sad, happy and neutral. The experiment included two dependent variables: emotion naturalness rating and emotion recognition. The naturalness rating used a 5-point Likert scale: 1-Very unnatural, 2-Unnatural, 3-Moderate, 4-Natural, 5-Very natural. Emotion Recognition was designed as a multiple-choice question.

3.1 Study Materials

Software Tools, 3D Facial Models and Animation References. AutoDesk MAYA[1] software was used to implement the skeleton and blendshape face animations. The audio-driven facial animations were created using Nvidia Audio2Face[2] software. The Visual-driven facial animations were created by Faceit[3] (Blender Plug-in) with Mocapx[4]. The blendshape and skeleton animations were created by a skilled animator from Purdue Computer Graphics Technology department. The 3D male and female facial models (Fig. 1) that all methods used were generated by Autodesk Character Generator[5] and had pre-made blendshape and skeleton rigs. In order to have real human emotions, reference videos were collected from acting students (one male and one female) from Purdue Theatre Department.

Fig. 1. Character Models

Stimuli. Twenty facial animation clips[6] demonstrating different types of emotion were presented to participants. Each video clip had a length of 10 s and a frame rate of 24fps. The virtual characters in the clips spoke the sentence "What this book says" in four basic emotions and one state of no emotion and had corresponding facial expressions of those five states: anger, happy, fear, sad, and neutral. The clips were muted to prevent participants form recognizing the emotions from the tone of voice. Figure 2 illustrates the entire process of generating the stimuli animations.

[1] https://www.autodesk.com.
[2] https://www.nvidia.com/en-us/omniverse/apps/audio2face/.
[3] https://blendermarket.com/products/faceit.
[4] https://www.mocapx.com/.
[5] https://charactergenerator.autodesk.com/.
[6] https://www.youtube.com/watch?v=4awHbEVvcjM&list=PL637gQB3PRrtQUUe-AeNDRRSiZ08a5il&index=12.

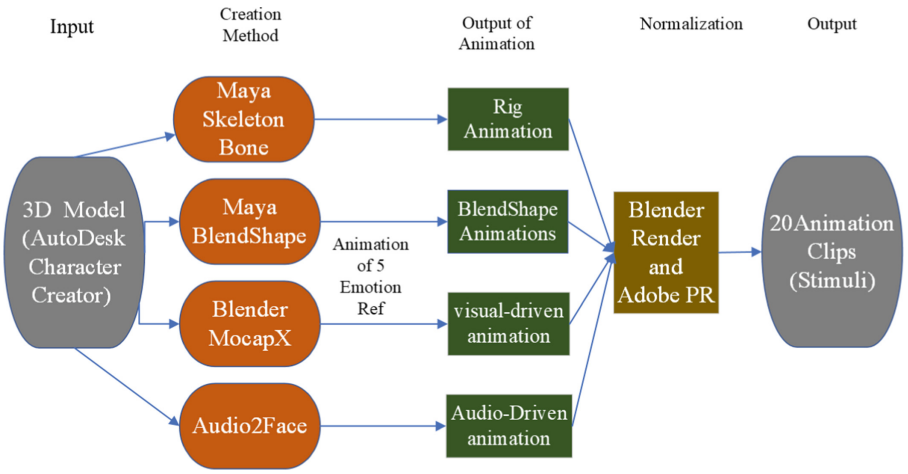

Fig. 2. Stimuli Generation

Participants. Ninety-eight participants were recruited through personal relationships, email notifications to Purdue University Computer Graphics Department and Prolific[7], an online platform for recruiting subjects.

Evaluation Instrument and Study Procedure. Participants were sent a link to an online survey which was hosted on Qualtrics[8], a web survey platform. The survey had two main sections. The first section had two components: detailed experimental instructions and IRB consent form and a set of demographics questions. The second section included 20 animation clips that were randomly presented to the subjects. Participants were asked to watch each clip, rate the naturalness of the character facial emotion on a 5-point Likert scale and identify the emotion by selecting it from 5 emotion choices.

4 Data Collection and Analysis

4.1 Pilot Study

Sample Size Calculation. ANOVA's power and sample size analysis was conducted to determine the appropriate sample size for the study. Power was set to 0.8 and significance level to 0.05. The Visual-Driven animation set had a mean score of 2.832 and standard deviation of 0.713. The Audio-Driven set had a mean score of 2.621 and a standard deviation of 0.700. The Skeleton set had a mean score of 3.074 and a standard deviation of 0.8072. The Blendshape set had a mean score of 2.763 and standard deviation of 0.927. The standard deviation

[7] https://www.prolific.co/.
[8] https://purdue.ca1.qualtrics.com/jfe/form/SV_77iAhuvDxAikaSG.

was set to the combined standard deviation of four groups, which was 0.8011. Results of overall F test for one-way ANOVA (Fig. 3) show that the total sample size of the study is around 344 or 86 per group, if the study has a power of 90%.

The POWER Procedure
Overall F Test for One-Way ANOVA

Fixed Scenario Elements	
Method	Exact
Alpha	0.05
Group Means	2.83 2.621 3.074 2.763
Standard Deviation	0.8011
Group Weights	1 1 1 1
Nominal Power	0.9

Computed N Total	
Actual Power	N Total
0.901	344

Fig. 3. Sample Size and Power Analysis

4.2 Main Study

Data Collection and Validation Checking. The study collected 98 responses. Most people completed the survey in 10–15 min; responses with a survey completion time lower than five minutes were discarded. Three responses had very short completion time and were not considered in the analysis. Ninety-five responses were used for analysis (power a of 94.5%). Twenty-nine responses were collected from Purdue and personal relationships, 64 responses were collected from Prolific. Each sample had 20 rating scores for four facial animation methods by five emotions. The naturalness ratings for five emotions of the four different methods were normally distributed.

Demographic Data. The survey collected participants' gender and animation experience. Twenty-five subjects were males, 69 were females, and one 'other'. Among 94 females and males, 24 had low animation experience, 68 had no prior animation experience, and two had high animation experience. Among 25 males, two had high animation experience, seven had low animation experience, and 16 had no prior experience. Among 69 females, 17 had low animation experience, and 52 had no prior experience.

4.3 Analysis of Perceived Naturalness of Facial Animations by Emotion

One-way ANOVA was used as the analysis approach to see whether there were statistically significant differences in naturalness ratings between skeleton, blendshape, visual-driven, and audio-driven facial animations. As shown in Table 1, anger emotion had $P = 0.0007, F = 5.79$, fear emotion had $P < 0.0001, F = 9.90$, happy emotion had $P = 0.0248, F = 3.16$, sad emotion had $P = 0.0081, F = 3.99$, and neutral emotion had $P < 0.0001, F = 19.19$, so statistically significant differences existed between the four facial animation methods across all emotions ($P < 0.05$). The study rejected the null hypothesis that all four types of facial animation would receive the same naturalness ratings. Happy had the highest P value. Although the P-value is lower than 0.05, happy P-value was the closest to 0.05, indicating that the naturalness scores of happy emotions produced by the four methods tended to be consistent. The neutral emotion had the highest F value and the fear emotion was the second highest, hence there is a significant difference between naturalness ratings of both neutral and sad emotions created by the four facial animation methods.

Table 1. ANOVA table of 5 emotions

Emotions	Anger	Fear	Happy	Sad	Neutral
DF (model)	3	3	3	3	3
DF (Error)	376	376	376	376	376
DF (Corrected Total)	379	379	379	379	379
Sum of Squares (Model)	19.589	33.713	10.863	12.534	61.968
Sum of Squares (Error)	424.316	426.695	431.095	393.937	404.632
Sum of Squares (Corrected Total)	443.905	460.408	441.958	406.471	466.589
Mean Square (Model)	6.530	11.238	3.621	4.178	20.653
Mean Square (Error)	1.128	1.135	1.147	1.048	1.076
F value	5.79	9.90	3.16	3.99	19.19
P value	0.0007	<0.0001	0.0248	0.0081	<0.0001
R-Square	0.0441	0.073	0.0246	0.031	0.133
Root MSE	1.062	1.065	1.071	1.024	1.037
Mean	2.616	2.461	2.411	2.818	2.905

The box-plots in Fig. 4[9] show the distribution of naturalness ratings by animation method. Most naturalness ratings of the audio-driven method were two or three points for all five emotions except for the neutral emotion, which ranged from one to three points. For anger, the majority of naturalness ratings for the

[9] https://github.com/weimingzhu101/A-Comparative-Study-of-Four-3D-Facial-Animation-Methods/blob/main/Fig4.png.

visual-driven method were lower than those of the audio-driven method, which was one to three points and had a median of two points. Blendshape and skeleton received relatively higher scores than the other two approaches. Most blendshape ratings fell in the range of one - three on fear and happy, and in the range of two - four on neutral, sad, and anger. The score distribution of the skeleton method in all emotions was in the range two - four, hence the skeleton facial animations were perceived as more natural than the animation produced with the other methods.

The ANOVA test yielded an overall significant difference. However, in order to identify which two groups were significantly different from each other, Tukey's Honest Significant Difference Test was implemented (see Fig. 5[10]). For the anger emotion, audio-driven and visual-driven naturalness ratings were statistically indistinguishable. The naturalness ratings for the pairs ('skeleton', 'blendshape'), ('skeleton', 'audio-driven'), and ('blendshape', 'audio') were not significantly different from each other. For the fear emotion, the naturalness rating of the skeleton method was statistically different (higher)from all other methods. The other pairs were not significantly different from each other. For the happiness emotion, all six pairs were not significantly different from each other. For the sadness emotion, the naturalness ratings of the pairs ('skeleton', 'audio-driven') and ('skeleton', 'visual-driven') were statistically different. For the neutral emotion, skeleton and visual-driven was the only pair for which the ratings of naturalness were not significantly different.

Analysis of Emotion Recognition. Results of a Chi-Square Test (shown in Fig. 6[11]) show that the recognition rates for the five different emotions created by the four methods differed significantly. For all five emotions the P value was lower than 0.05.

As shown in Fig. 7 (See footnote 11), the recognition rate for most emotions created by the skeleton facial animation method was higher than 55%, except for the neutral emotion which was 47.37%. Although the audio-driven facial animation method performed poorly on fear recognition, anger recognition, and happy recognition, it did rather well on neutral recognition and sad recognition compared to the other three approaches. The visual-driven method had the lowest recognition rate for the fear emotion. If we set 65% as an acceptable emotion recognition rate, then we can state that fear and anger emotions produced by the skeleton and blendshape methods can be recognized by people. The happy emotion created by the skeleton, visual-driven, and blendshape methods can also be recognized by people. Only the sad emotion created by the audio-driven method can be recognized.

[10] https://github.com/weimingzhu101/A-Comparative-Study-of-Four-3D-Facial-Animation-Methods/blob/main/Fig5.png.

[11] https://github.com/weimingzhu101/A-Comparative-Study-of-Four-3D-Facial-Animation-Methods/blob/main/Fig6-7.png.

Analysis of Gender. The study also explored whether there were differences in naturalness ratings based on participants' gender. A t-test was used to analyze the differences. Results (Tables 2, 3, 4 and 5) showed that all genders rated similarly for naturalness across all five emotions created by the four methods (all four methods' P values > 0.05), hence null hypothesis H02 could not be rejected.

Analysis of Animation Experience. The study collected the level of animation experience of the participants. Three levels were collected: no prior animation experience, low animation experience, and high animation experience. Since there were only two people who had high animation experience, the study could not provide meaningful insights on this group and the two subjects with high experience were removed from the data analysis. However, the study was able to compare naturalness ratings between people who had no animation experience and people who had some animation experience. T-tests were used to analyze the differences between them.

Results of the t-tests are reported in tables 6–9[12]. For the visual-driven and skeleton methods there were no significant differences between the two groups ($P > 0.05$). For the audio-driven method, the participants with no animation experience gave significantly lower naturalness ratings than the participants with low experience, ($t = -2.12$, $P = 0.035$). For the blendshape method, the participants with no animation experience gave significantly lower naturalness ratings than the participants with low animation experience, ($t = -3.76$, $P < 0.001$).

Table 2. Audio-Driven Method: Gender Differences in Naturalness Ratings

Gender	N	Mean	Std Dev
Male	125	2.4240	0.9181
Female	345	2.4754	0.9884
Diff(1-2)	0	-0.0514	0.9703

Method	Variances	DF	t Value	$P > \|t\|$
Pooled	Equal	468	-0.51	0.6123
Satterthwaite	Unequal	235.05	-0.52	0.6002

[12] https://github.com/weimingzhu101/A-Comparative-Study-of-Four-3D-Facial-Animation-Methods.

Table 3. Visual-Driven Method: Gender Differences in Naturalness Ratings

Gender	N	Mean	Std Dev
Male	125	2.6960	1.1861
Female	345	2.5768	1.0813
Diff(1-2)	0	0.1192	1.1100

Method	Variances	DF	t Value	$P > \|t\|$
Pooled	Equal	468	1.03	0.3042
Satterthwaite	Unequal	203.28	0.98	0.3258

Table 4. SKeleton Method: Gender Differences in Naturalness Ratings

Gender	N	Mean	Std Dev
Male	125	2.9840	0.9502
Female	345	2.9652	1.1407
Diff(1-2)	0	0.0188	1.0934

Method	Variances	DF	t Value	$P > \|t\|$
Pooled	Equal	468	0.16	0.8694
Satterthwaite	Unequal	261.57	0.18	0.8580

Table 5. Blenshape Method: Gender Differences in Naturalness Ratings

Gender	N	Mean	Std Dev
Male	125	2.4720	1.0669
Female	345	2.5855	1.1735
Diff(1-2)	0	-0.1135	1.1462

Method	Variances	DF	t Value	$P > \|t\|$
Pooled	Equal	468	-0.95	0.3433
Satterthwaite	Unequal	261.57	-0.99	0.3223

5 Discussion and Conclusion

The statistical data analysis provided significant evidence to reject the main null hypothesis (H_{01}) that Audio-Driven, Visual-Driven, Skeleton, and Blendshape facial animations would have the same naturalness ratings and recognition rate. For the happy emotion, although there were differences in naturalness ratings among the four methods, the P-value was close to 0.05, indicating that the naturalness scores of happy emotions produced by the four methods tended to be consistent. The naturalness ratings of the four methods for fear and neutral expressions varied greatly. It was found that the facial animations generated by

the skeleton method received significantly higher naturalness ratings across all five expressions.

In regard to emotion recognition, the skeleton method had the overall highest recognition rate across all five emotions. The blendshape was second, the visual-driven method was third, and the audio-driven was fourth. Hence, the results suggest that artist-created animated facial emotions can be recognized more accurately than the computer-generated ones. The sad and neutral emotions made by all methods were not easy to recognize compared to the other three emotions.

Further, the study found that men and women had similar perceptions of the naturalness of the animation. People with no animation experience tended to give lower naturalness scores than those with a little animation experience.

The study had some limitations, which could be addressed in future work. First, the study used a personal consumer level visual-driven method, which might not be the best among currently available visual-driven approaches. For instance, tools such as Live Link[13] with Unreal Engine[14] or Live Face with iClone[15] could produce better quality animations but they need a very specific 3D facial model which is very time-consuming to create. Second, the study utilized only two characters and it is possible that the findings are in part dependent on the intrinsic design and rig characteristics of the 3D models. In future studies, it would be interesting to examine whether the findings of our experiment hold true for characters with different design features and visual styles. Third, the keyframe facial animations were created by one animator and their quality is largely dependent on the animator's skills. Future studies should include facial animations produced by different animators with different sets of skills. Finally, the next version of the audio-driven facial animation creation tool, Audio2face, is expected to make a substantial improvement on emotion representation. Future research should continue to monitor and analyze the state of art of audio-driven methods and continue to perform comparative studies similar to the one reported in the paper.

The findings of the study have direct practical implications for character artists and media content designers, as they can help them make more informed animation decisions. The overall goal of our research is to develop an empirically grounded research base that will guide the design and animation of life-like affective animated agents that can be used in a variety of areas, including STEM education. Toward this goal, we will continue to conduct research studies to identify modeling an animation techniques that are the most effective at producing affective animated agents that express believable emotions.

Acknowledgments. This work is supported by NSF-Cyberlearning award 1821894: Multimodal Affective Animated Pedagogical Agents for Different Types of Learners.

[13] https://docs.unrealengine.com/5.0/en-US/live-link-in-unreal-engine/.
[14] https://www.unrealengine.com.
[15] https://www.reallusion.com/iclone.

References

1. Adolphs, R.: Perception and emotion: how we recognize facial expressions. Curr. Dir. Psychol. Sci. **15**(5), 222–226 (2006)
2. Bouaziz, S., Wang, Y., Pauly, M.: Online modeling for realtime facial animation. ACM Trans. Graph. (ToG) **32**(4), 1–10 (2013)
3. Cao, C., Hou, Q., Zhou, K.: Displaced dynamic expression regression for real-time facial tracking and animation. ACM Trans. Graph. (TOG) **33**(4), 1–10 (2014)
4. Fan, B., Xie, L., Yang, S., Wang, L., Soong, F.K.: A deep bidirectional LSTM approach for video-realistic talking head. Multimed. Tools Appl. **75**(9), 5287–5309 (2016)
5. Friesen, W.V.: Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues. Prentice-Hall (1975)
6. Hanrahan, P., Sturman, D.: Interactive animation of parametric models. Vis. Comput. **1**(4), 260–266 (1985)
7. Huber, P., et al.: A multiresolution 3D morphable face model and fitting framework. In: Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. University of Surrey (2016)
8. Jeni, L.A., Tulyakov, S., Yin, L., Sebe, N., Cohn, J.F.: The first 3D face alignment in the wild (3DFAW) challenge. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 511–520. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_35
9. Jourabloo, A., Liu, X.: Large-pose face alignment via CNN-based dense 3D model fitting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4188–4196 (2016)
10. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Trans. Graph. (TOG) **36**(4), 1–12 (2017)
11. Lewis, J.P., Anjyo, K., Rhee, T., Zhang, M., Pighin, F.H., Deng, Z.: Practice and theory of blendshape facial models. Eurograph. (State Art Rep.) **1**(8), 2 (2014)
12. Li, L., Liu, Y., Zhang, H.: A survey of computer facial animation techniques. In: 2012 International Conference on Computer Science and Electronics Engineering, vol. 3, pp. 434–438. IEEE (2012)
13. Liang, Z., Ding, S., Lin, L.: Unconstrained facial landmark localization with backbone-branches fully-convolutional networks. arXiv preprint arXiv:1507.03409 (2015)
14. Parke, F.I., Waters, K.: Computer Facial Animation. CRC Press, Boca Raton (2008)
15. Parke, F.I.: Computer generated animation of faces. In: Proceedings of the ACM annual conference, vol. 1, pp. 451–457 (1972)
16. Ping, H.Y., Abdullah, L.N., Sulaiman, P.S., Halin, A.A.: Computer facial animation: a review. Int. J. Comput. Theory Eng. **5**(4), 658 (2013)
17. Rhodes, G., Haxby, J.: Oxford Handbook of Face Perception. Oxford University Press, Oxford (2011)
18. Richardson, E., Sela, M., Kimmel, R.: 3D face reconstruction by learning from synthetic data. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 460–469. IEEE (2016)
19. Sako, S., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.: HMM-based text-to-audio-visual speech synthesis. In: Sixth International Conference on Spoken Language Processing (2000)

20. Saragih, J.M., Lucey, S., Cohn, J.F.: Real-time avatar animation from a single image. In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 117–124. IEEE (2011)
21. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing Obama: learning lip sync from audio. ACM Trans. Graph. (ToG) **36**(4), 1–13 (2017)
22. Taylor, S., et al.: A deep learning approach for generalized speech animation. ACM Trans. Graph. (TOG) **36**(4), 1–11 (2017)
23. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: real-time face capture and reenactment of RGB videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2387–2395 (2016)
24. Orvalho, V., Bastos, P., Parke, F., Oliveira, B., Alvarez, X.: A facial rigging survey. In: 2012 Eurographics Conference, pp. 182–204. EG Digital Library (2012)
25. Vlasic, D., Brand, M., Pfister, H., Popovic, J.: Face transfer with multilinear models. In: ACM SIGGRAPH 2006 Courses, p. 24-es (2006)
26. Wang, L., Han, W., Soong, F.K., Huo, Q.: Text driven 3D photo-realistic talking head. In: Twelfth Annual Conference of the International Speech Communication Association (2011)
27. Wang, L., Qian, X., Han, W., Soong, F.K.: Synthesizing photo-real talking head via trajectory-guided sample selection. In: Eleventh Annual Conference of the International Speech Communication Association (2010)
28. Weise, T., Bouaziz, S., Li, H., Pauly, M.: Realtime performance-based facial animation. ACM Trans. Graph. (TOG) **30**(4), 1–10 (2011)
29. Xie, L., Liu, Z.Q.: Realistic mouth-synching for speech-driven talking face using articulatory modelling. IEEE Trans. Multimed. **9**(3), 500–510 (2007)
30. Zhang, X., Wang, L., Li, G., Seide, F., Soong, F.K.: A new language independent, photo-realistic talking head driven by voice only. In: Interspeech, pp. 2743–2747 (2013)
31. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 94–108. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_7
32. Zhao, R., Wang, Y., Benitez-Quiroz, C.F., Liu, Y., Martinez, A.M.: Fast and precise face alignment and 3D shape reconstruction from a single 2D image. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 590–603. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_41
33. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 386–391 (2013)
34. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: a 3D solution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 146–155 (2016)
35. Zhu, X., Lei, Z., Yan, J., Yi, D., Li, S.Z.: High-fidelity pose and expression normalization for face recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 787–796 (2015)