Received: 06 September 2021 | Revised: 15 November 2021 | Accepted: 18 November 2021
DOI: 10.37917/ijeee.18.1.4 Vol. 18| Issue 1| June 2022

Open Access

Iraqi Journal for Electrical and Electronic Engineering
Review Article

Facial Modelling and Animation: An Overview of The State-of-The Art

Samia Dawood Shakir*, Ali A. Al-Azza
Department of Electrical Engineering, University of Basrah, Basra, Iraq

Correspondence
* Samia Dawood Shakir
Department of Electrical Engineering,
University of Basrah, Basra, Iraq
Email: samea.shaker@uobasrah.edu.iq

Abstract
Animating human face presents interesting challenges because of its familiarity as the face is the part utilized to recognize individuals. This paper reviewed the approaches used in facial modeling and animation and described their strengths and weaknesses. Realistic face animation of computer graphic models of human faces can be hard to achieve as a result of the many details that should be approximated in producing realistic facial expressions. Many methods have been researched to create more and more accurate animations that can efficiently represent human faces. We described the techniques that have been utilized to produce realistic facial animation. In this survey, we roughly categorized the facial modeling and animation approach into the following classes: blendshape or shape interpolation, parameterizations, facial action coding system-based approaches, moving pictures experts group-4 facial animation, physics-based muscle modeling, performance driven facial animation, visual speech animation.

KEYWORDS: Blendshape, MPEG-4, Facial animation, Visual speech, Physics-based.

I. INTRODUCTION

Facial modelling and animation denote to methods of representing the face graphically on a computer system and animating the face in a manner consistent with real human movement. This is considered one of the most difficult functions undertaken in the area of animation, because of many issues. Firstly, because most of people experience several natural human interactions daily, humans are capable at recognizing unnatural face activities. Therefore, the smallest changeability in an animated face directly notifies the observer and the animation loses its naturalism. The human facial is a complicated system of a large number of muscles that require to be skillfully coordinated to consider actual. Another issue that shares to the difficulty of modelling and animation of the human facial is its variety. Diverse people have various face features, produced by various bone structures and muscle proportions. Facial animation has been widely used in medicine, education, military, entertainment and many other fields.

Considerable advancement has been occurred by the researchers from the computer graphics community, which have developed a large number of methods to generate high quality facial models. But, despite of this advancement, the computer synthesised human face animation still needs costly resources. This paper goals at providing a comprehensive review for the existing methods in the area of face modelling and animation, giving analysis to their strength and weakness. In this survey we pay attention to more recent methods, which permit the construction of highly realistic results. Also, the survey provides a historical view on the progress of these methods.

There are limited studies or detailed historical treatments of the topic. This review described facial modelling and animation methods. Categorizing face modelling and animation methods is a hard task, because groupings are complex by the lack of exact boundaries between techniques and the fact that current methods often integrate several approaches to produce better results. In this review, we categorize facial modelling and animation methods into the following classes: blendshape or shape interpolation, parameterisations, facial action coding system-based approaches, moving pictures experts group-4 facial animation, physics-based muscle modelling, performance driven facial animation, visual speech animation. Fig. 1 displays the overview of facial animation methods that begins from input, facial modelling, facial animation and output as simulation and rendering.

II. FACIAL MODELLING AND ANIMATION METHODS

A. Blendshape

There are a number of methods to facial animation. These include key framing utilising blendshapes or shape interpolation, performance-driven animation, parameterised

models, muscle-based. Blendshapes is a simple linear model of facial expression. It has driven animated characters in Hollywood movies, and is a standard feature of commercial animation packages. The beginning of the blendshape method is in the computer graphics industry by the 1980s. In 1972, Frederick I. Parke first utilised the shape interpolations and animated the face utilising cosine interpolation. By the late 1980s the offset blendshape scheme became widespread and exist in commercial software [1]. In this variant a neutral facial shape is designated and the residual shapes are substituted by the differences between those shapes and the neutral one. The differences between the target shape and the neutral face are restricted to a small region, although it relies on the modeller to generate shapes with this property. This concept was expanded to a segmented face where separate parts are blended independently [2], therefore promising local control. A standard instance is the segmentation of a face into an upper part and a lower part: the upper part is utilised for expressing emotions, whereas the lower part expresses speech [3]. Despite the blendshape method is conceptually simple, developing such a blendshape face model is a labor intensive effort. In order to express a whole collection of realistic expressions, digital modellers often have to produce large libraries of blendshape targets. For instance, the character of Gollum in The Lord of the Rings had 675 targets [4]. Generating a realistically detailed model can be as much as a year of effort for a skilled modeller, including many iterations of enhancement. The Blendshape facial animation is the general choice for realistic human characters in the films production. The method has been utilised for lead characters in films such as King Kong [5], The Curious Case of Benjamin Button [6]. The simplest case is an interpolation between two key-frames at extreme positions over a time interval as shown in Fig. 2.

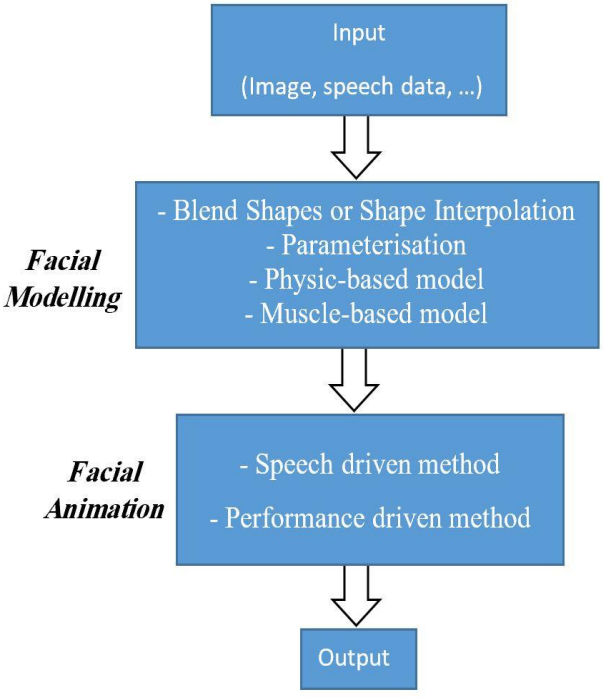

Fig. 1: Overview flows of facial animation.

Fig. 2: Linear interpolation is performed on blendshapes.

B. *Parameterisation*

The essential concept for a parameterised model is to build a desired face or facial expression relied on some number of controlling parameter values. If the model supports conformation parameters, it can be utilised to generate a range of individual faces.

Parameterisations identify any face and expression by a grouping of independent parameter values [7]. Different than interpolation approaches, parameterisations permit explicit control of particular facial configurations. Groupings of parameters offer a large range of facial expressions with low computational costs. Frederic I. Parke produced the first 3D parametric model of a human face [8]. The facial geometry is broken into parts and controlled by parameters, for instance, the rotation of the jaw or the direction of an eye's gaze. The parameters affecting the various nodes are showed. A parametric facial animation system describes a set of parameters for the face. These are mostly the expression parameters for different fragments of the face, for example forehead, eyes, and nose and the conformation parameters that globally applied to the entire face. The main parameters for the forehead are scale and shape of forehead. The main parameters for the eyes are pupil size, eyelid opening, eyebrow colour and separation, etc. The conformation parameters are aspect ratio of the face, colour of the skin, etc. Each expression parameter affects a set of vertices of the face model. In this way, key frames can be defined easily. A complete generic parameterisation is not possible, because that the parameter set relies on the facial mesh topology. Moreover, boring manual tuning is required to set parameter values. The limitations of parameterisation led to the development of different approaches such as morphing between images and geometry, physically muscle-based animation, and performance driven animation.

C. *Facial Action Coding System*

The most common expression coding systems are facial action coding system (FACS) and moving pictures experts' group-4 (MPEG-4) Facial animation models. FACS was suggested by Ekman and Friesen in 1978 [9] and has been improved in 2002 [10]. It defines all the movements that can be seen in the face based on face anatomy. It has been utilised extensively in facial animation during the last few decades. FACS has become a typical in interpretation facial behaviour in science research such as psychophysiology [11], and in fields such as video games [12], movies [6], robots [13] and facial expression recognition, mapping, generation, [14–16]. FACS is an anatomically based system for defining all observable face movements. Each component of a facial

movement is called an Action Unit (AU). Each AU is identified by a number (AU1, AU3, AU20, . . . etc.). Samples of these action units are presented in Table I. Facial expressions are produced by combining the action units. For instance, combining AU1 (Cheeks raiser), AU4 (Lip Corner Puller), and AU15 (Lips Part) produces Happiness expression.

Recently, the interest in utilising the FACS for producing visual speech has declined. This is because two reasons. Firstly, nowadays most of the face models designed for visual speech synthesis purposes are not relied on human anatomy, but consist of high detailed polygon meshes and textures which are generally automatically computed by 3D scanning approaches. Normal mesh deformations are learned by advanced 3D motion capture methods, which is faster and easier than a manual explanation of the numerous muscles of the face and their effect on the facial appearance. Secondly, FACS offers many Action Units that can be utilised to precisely mimic certain expressions, but, these Action Units are less appropriate to simulate all the detailed gestures of the face corresponding to speech production. Such that FACS is not optimised for modelling visual speech. Facial expression generation or synthesis has recently received increasing attention in the facial expression modelling domain. Ekman and Friesen [17] developed the FACS for describing facial expressions with some basic face action units (AUs), each of which represents a basic face muscle movement or expression change.

Kumar and Sharma [18] suggested an improved Waters facial model utilised as an avatar for research published in [19], which discussed a facial animation system driven by the FACS in a low-bandwidth video streaming setting. To build facial Expressions, FACS defines 32 single Action Units (AUs) which are created by an underlying muscle action that interact in various methods. In this work enhancements were provided to the Waters facial model by enhancing its UI, adding sheet muscles, providing an alternative implementation to the jaw rotation function, introducing a new sphincter muscle model that can be utilised around the eyes and alterations to operation of the sphincter muscle utilised around the mouth. Zhou et al. [20] introduced a conditional difference adversarial autoencoder (CDAAE) to transfer AUs from absence to presence on the global face. This approach uses the low-resolution images, which could lose facial details vital for AU production. Pumarola et al. [21] proposed GANimation which transfers AUs on the whole face and generate a co-generated phenomenon between different AUs. For this approach it is difficult to generate a single AU respectively without touched the other AU.

With the recent rise of deep learning, CNN have been widely used to extract AU features. Zhao et al. [22] suggested a deep region and multi-label learning (DRML) system to partition the face images into 8 - 8 blocks and utilised individual convolutional kernels to convolve each block. Despite this method treats each face as a set of individual parts, it partitions blocks uniformly and does not reflect the FACS knowledge, thus leading to poor performance. Zhilei Liu [23], proposes an Action Unit (AU) level facial expression synthesis approach named Local Attentive Conditional Generative Adversarial Network (LAC-GAN) relied on face action units annotations. They build a model for facial action unit synthesis with more local texture details. In this approach local AU regions is integrated with conditional generative adversarial network. The proposed method manipulates AUs between various states, which learns a mapping between a facial manifold related to AU manipulation. Moreover, the key point of this approach is to do the manipulation module concentrate only on the generate of local AU region without touching the remainder identity information and the other AUs. The development of deep graph networks modelling has recently attracted increasing attention. Zhilei Liu [24] introduces an end-to-end deep learning framework for facial AU detection with graph convolutional network (GCN) for AU relation modelling. They use the graph convolutional network (GCN) [25] for AU relation modelling to support the facial AU detection. AU related areas are extracted; these AU areas are feed into some specific AU auto-encoder for deep representation extraction. In addition, each latent representation is pull into GCN as a node.

Table I
Sample single facial action units.

AU	FACS Name
1	Inner Brow Raiser
14	Dimpler
5	Upper Lid Raiser
17	Chin Raiser

D. Moving Pictures Experts Group-4

Moving pictures experts' group-4 (MPEG-4) is an object-based multimedia compression standard that permits encoding independently diverse scene's audiovisual objects (AVO). MPEG-4 has facial definition parameter set (FDP) and the facial animation parameter set (FAP) which were designed to describe the facial shape and texture, as well as regenerating the animation of faces for instance speech pronunciation, expressions, and emotions. MPEG-4 facial animation outlines many parameters of a talking face in a standardised approach. It identifies and animates 3D face models by describing face definition parameters (FDP) and facial animation parameters (FAP). FDPs enclose information for building particular 3D face geometry, while FAPs encode motion parameters of key feature points on the face over time. In MPEG-4, the head is grouped into 84 feature points (FPs), every point defines the shape of an area. Fig. 3 demonstrates part of the MPEG-4 feature points. After excluding the feature points that are not simulated by FAPs, 68 FAPs are classified into collections. Samples of these collections are presented in Table II. The FAPs are two groups, one represents the facial expressions which consist of six basic emotions, i.e., surprise, anger, sadness, joy, disgust, and fear. The second one concentrates on facial areas such as the left mouth corner, the chin bottom and the right eyebrow. Refer to the MPEG- 4 facial animation book for more details about MPEG-4 facial animation standard [26].

El Rhalibi et al. [27] presented a method relied on 3D Homura that integrate MPEG-4 standards to realistic

animation streams identified as Charisma, which can be applied for animation systems utilised with games and virtual characters in the web.

E. Physics-Based Muscle Modelling

Many attempts have been set on physics-based muscle modelling to model anatomical facial behaviour. These are classified into three classes; mass-spring systems, vector representation, and layered spring meshes. Mass-spring approaches propagate muscle forces in an elastic spring mesh which models skin deformation [28]. The vector method deforms a facial mesh utilising motion fields in delineated regions of influence [29]. A mass-spring structure was extended into three connected mesh layers by a layered spring mesh [30].

The limitation of blendshapes is that they provide only the linear subspace. Recently, researchers have tended to use physical simulation to achieve more expressive, non-linear facial animation. One of the first approaches for physics-based facial animation was suggested by Sifakis et al. [31], who construct a detailed face rig comprising of a complete, anatomically muscle structure, generated manually from the actor's medical data. Constructing the muscle structure for an actor is a time-consuming process. Cong et al. [32] enhanced an automatic method to transfer anatomy pattern to target input faces. Ichim et al. [33] fit a template model of muscles, bones and flesh to face scans. This approach succeeds by resolving for the muscle activation parameters that best appropriate the input scans through forward simulation, and generates an actor physical face mesh for animation. Ma et al. [34] use a mass-spring system to construct a blendshape model which incorporates physical interaction. Kozlov et al. [35] concentrates on the production of expression-specific physical effects, however the drawback is that spatially-varying material parameters require to be painted and set manually for each expression.

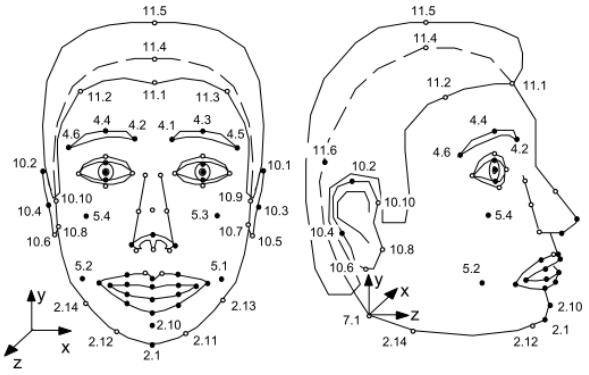

Fig. 3: The facial feature points defined in the MPEG-4 standard [36].

Table II
FAP groups in MPEG-4.

Group	Number of FAPs
Viseme and expressions	2
Cheeks	4
Eyebrow	8
Tongue	5
Lip, Chin and Jaw	26

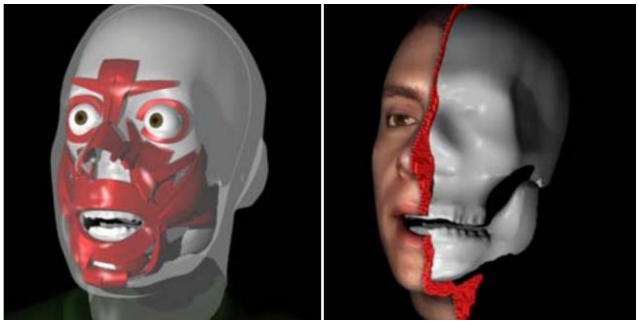

Fig. 4: Example of physics-based facial animation [31].

F. Performance Driven Facial Animation

Performance capture is a method that uses motion capture technology to represent the performance of the character. In conventional motion capture, the face and the body are recorded at different times and then blended together. Using performance capture, the face and the body are captured at the same time to describe the entire performance of the performer. The Polar Express [37] was the first film to successfully use facial motion capture for an entire computer graphics (CG) movie as shown in Fig. 5.

Fig. 5: Example of performance-driven facial capture utilising markers, used in the Polar Express [37].

Artist driven manual key-frame animations may never capture the subtleties of a human face. The trend in facial animation has moved towards utilising the human face itself as the driver and input device for facial animation. Extracting information from an actual performance of facial movements is natural, easy, and fast which lead to the concept of performance-driven facial animation. A 'performance' can be understood as a visual capture of an actor's face talking and emoting which is utilised to extract information then retarget the motion onto a digital character. Williams [38] presented the term performance driven facial animation to the computer graphics society in Siggraph 1990. Since then there have been many studies that have extended the main concept. Hardware motion capture systems were familiar in the mid-90s and were utilised regularly in short demos [38].

The process of performance driven facial animation can be divided into three stages: modelling, capture, and retargeting. The modelling stage has to do with the model of the human face such that it could be digitally stored, displayed and modified. The choice of representation does have an effect on the final animation as the model inherently limits the expressive abilities of the face. modelling approaches variety from mesh propagation-based approaches where a single 3D mesh is deformed over the performance [39, 40] as shown in Fig. 6, 2D and 3D statistical models based on PCA [41, 42], blendshape models

[43] and muscle based anatomical representations combined with deformable skin [44]. The capture stage of the performance driven facial animation could be supposed as the extraction of relevant valuable information from the input video such that this information could then be applied onto the underlying face representation to synthesise the animation. This capture could be achieved using approaches that are active or passive. Active approaches include Marker-based capture where physical markers are located on the actor's face and tracked through the performance [45]. Passive approaches include approaches that utilise video inputs of the actor's face without any markers placed on the actor's face [43]. The aim of the Retargeting stage is to adapt the parameters acquired from the capture step and animating the virtual target character. The parameters utilised to drive this character could be different from the obtained capture parameters. This is not easy task especially when the target character has proportions different from the actor's face. Fig. 7 shows an example of Retargeting an actor's facial expression onto multiple target characters.

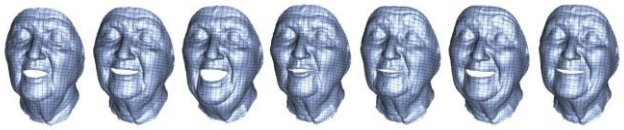

Fig. 6: Example of mesh propagation being utilised for the underlying representation for the animation [40].

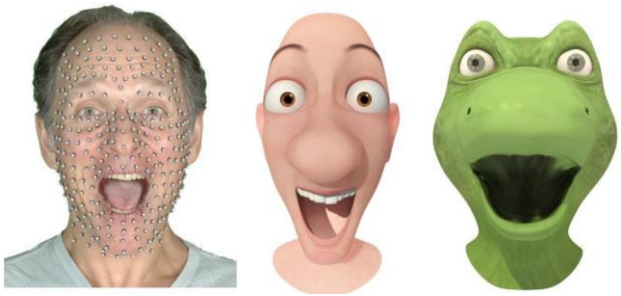

Fig. 7: Retargeting an actor's facial expression onto multiple target characters [53].

A technique for computing the suitable mesh deformations for simulating human facial expressions is called performance-driven strategy [38]. This method requires a 3D polygon mesh illustrating a human face. This mesh could be hand-crafted or it could be automatically constructed utilising a 3D scanner. Afterward, facial gestures from a human speaker are captured. Speech-related deformations of the mesh are learned when these original gestures are mapped on the polygon mesh. Then, these mesh deformations can be utilised to animate the virtual speaker to synthesize new visual speech. Recently, hybrid models have joined physically-based methods with other approaches [46, 35]. The recent industry standard approaches are motion capture, marker less [40] or marker-based [47], and blendshape animation, relied on extreme poses named blend targets, where each blend target encodes one action, i.e., raising an eyebrow, then multiple blend targets can be linearly joined.

Thies et al. [48] suggested a real-time photo-realistic facial monocular reenactment method. They track facial landmarks depending on a dense photometric consistency measure and utilise GPU-based iteratively reweighted least squares solver to achieve real-time frame rates. Recently, some commercial facial performance capture software have been released, for example Apple's iPhone X application to animate a virtual character with its depth camera [49]. Barros et al. [50] introduces a method for real-time performance-driven facial animation from monocular videos. facial expressions are transferred from 2D images to a 3D virtual character, by determining the rigid head pose and non-rigid face deformation from detected and tracked face landmarks. Blendshape models are used to map the input face into the facial expression space of the 3D head model.

Recently, deep learning approaches have shown interesting effort for high-quality facial performance capture. Olszewski et al. [51] utilised convolutional neural networks to recover blendshape weights corresponding to the mouth expression of virtual reality headset users. Laine et al. [52] used eep learning to learn a mapping from an actor's image to the corresponding performance captured mesh, permitting for the appropriate capture of extra high-quality data. These techniques can infer coherent data through lips contacts if such data was present in the training set.

G. Visual Speech Animation

Visual speech animation could be considered as visual motions of the face when humans are talking. Generating realistic visual speech animations corresponding to new text or pre-recorded audio speech input has been a hard task for decades. This is because human languages, generally, have a large vocabulary and a large number of phonemes (the basic units of speech), but also the phenomena of speech co-articulation that complicates the mappings between audio speech signals or phonemes and visual speech motions. Visual speech co-articulation can be defined as follows: The visual appearance of a phoneme depends on the phonemes that come before and after it.

Previous researchers classified visual speech animation synthesis into two different categories: viseme-driven and data-driven approaches. Viseme-driven methods need animators to design key mouth shapes for phonemes to synthesis new speech animations. While data-driven methods do not require pre-designed key shapes, but generally require a pre-recorded facial motion database for synthesis purposes. Viseme-driven methods typically utilise some form of hand-tuned dominance function to describe how visual parameters representing phone-level classes are blended to generate the animation [54]. This typically results in satisfactory animation. However, the need to hand-tune the blending functions for each character makes the utilising of this method unpractical. Instead, sample-based methods concatenate segments of pre-recorded visual speech, where the segments might correspond to fixed-sized units [55, 56] or variable length units [57, 58]. A limitation of sample-based methods is that good quality animation needs a large corpus from which units can be selected, and some form of smoothing is needed at the concatenation boundaries to

reduce the visual artifacts which result from discontinuities. The advantage of variable length units is that there are fewer concatenation boundaries, but the search is mostly more difficult. Moreover, the output animation is generally constrained to the talker and the environment of the original recording. Methods relied on parametric statistical models comprise switching linear dynamical systems [59], shared Gaussian process latent variable models [60], artificial neural networks [61], and hidden Markov models [62–64]. One of the noteworthy early work, Voice Puppetry [65], suggested an HMM-based talking face synthesis driven by speech signal. Xie et al. [66] suggested coupled HMMs (cHMMs) to model auditory-visual asynchrony. Choi et al. [67] and Terissi et. al [68] utilised HMM inversion (HMMI) to infer the visual parameters from speech signal. Zhang et al. [69] utilised a DNN to map speech features into HMM states, then further maps to synthesised faces.

Deep Learning is a recent direction in artificial intelligence and machine learning research. Lately, new deep learning frameworks are being born, outperforming state-of-the-art machine learning approaches. A few DNN-based methods have been investigated. Suwajanakorn et al. [70] designed an LSTM network to synthesis photo-realistic talking head videos of a target identity directly from speech feature. This system needs a number of hours of face videos of the target identity, which limits its application practically.

Chung et al. [71] introduced an encoder-decoder convolutional neural network (CNN) model to synthesis talking face video from speech feature and a single face image of the target identity. In this work, the reduction from a number of hours of face videos to a single face image to learn the target identity is a great improvement. The main limitation of end-to-end synthesis is the lack of freedom for further manipulation of the synthesised face video. For instance, within a synthesised video, one might need to vary the gestures, facial expressions, and lighting conditions, all of which could be independent of the content of the speech. These end-to-end frameworks could not provide such manipulations unless these factors could be taken as extra inputs. However, that would significantly increase the amount and diversity of data required for training the systems. For such manipulations a modular design which splits the generation of key parameters and the fine details of synthesised face images is more flexible. Pham et al. [72] adopted a modular design, speech features firstly mapped to 3D deformable shape and rotation parameters utilising an LSTM framework, and then a 3D animated face synthesised in realtime from the predicted parameters. This approach is improved in [73] by substituting speech features with raw waveforms as the input and substituting the LSTM framework with a convolutional architecture. Chen et al. [74] first transferred the auditory to face and then synthesised video frames conditioned on the landmarks. Song et al. [75] suggested a conditional recurrent adversarial network that integrated auditory and image features in recurrent units. But, the head pose generated by these 2D-based methods is almost fixed during talking. This drawback is caused by the defect inherent in 2D-based methods, since it is difficult to only use 2D information alone for naturally modelling the change of pose. They introduce 3D geometry information into the proposed system to simultaneously model personalised head pose, expression and lip synchronisation.

The current overview of generative adversarial network GANs [76] has shifted the motivation of the machine learning group to generative modelling. GANs contain two challenging networks: generative and discriminative networks. The generator's target is to generate realistic samples while the discriminator's target is to discriminate between the real and produced samples. This competition leads the generator to produce robustly realistic samples. Vougioukas [77–79] proposed an end-to-end model using temporal GANs for speech-driven facial animation, capable of generating a video of a talking head from an audio signal. Guo [80] introduce a GAN-based, end-to-end TTS training algorithm, which propose the generated sequence to GAN training to avoid exposure bias in autoregressive decoder. The suggested algorithm improves both output quality and generalisation of the model.

III. Methods Comparison and Evaluation

Facial blendshapes are the general option for realistic face animation in the film industry. They have driven animated characters in Hollywood movies and attracted many research attentions. Facial blendshapes can be classified into geometric and physics-based. Linear interpolation plays a principal role in geometric face blendshapes. Linear interpolation-based face blendshapes are general meanwhile they have the advantages of expressiveness, simplicity and interpretability. Despite of this, the following limitations have been identified in [81]. Facial blendshapes might be considered as samples from a hypothesized manifold of face expressions. Producing a new face shape needs enough target face shapes to sample the manifold and describe local linear interpolation functions. Generating sufficient target face shapes is typically an iterative and effort intensive procedure. Physics-based face blendshapes are to augment physics to facial blendshapes which has a prospect to tackle the above issue. Particularly, when physics-based simulations are combined with data-driven methods, truthful face animation can be generated.

The weaknesses with physically-based methods produced by their complexity. Acquiring physics-based rigs configured could be a very hard and boring task for performers. There has been effort achieved to automate some of the creation of these rigs for physics based facial animation [81], however there is still a noteworthy amount of effort that has to be achieved by hand to make the rigs look accurate and not fall into the uncanny valley. As well as the complexity causing the rigs to be hard to set up, these complex rigs need enormous computations to compute the animations. These large computations make physically-based animations inappropriate for real-time applications. Physics-based solutions to face animation have given the entertainment industry great animations that are becoming closer to being anatomically truthful, however the great amount of effort to get these rigs completed will keep the method from being utilised more widely.

MPEG-4 is could be seen as a formalisation of FACS, however what makes it different method is that, unlike

FACS, it lacks the direct correspondence between animation parameters and face muscles. FACS offers a very consistent description for the facial upper portions but it does not for the lower portions of the face. That restricts FACS from being the dominant method in the Face Animation area. MPEG-4 describes 66 low-level FAPs and two high-level FAPs. The low-level FAPs are based on the study of minimal face actions and are closely related to muscle actions. They denote a complete set of basic face actions, and therefore permit the representation of most natural face expressions. FACS clearly describe face expressions by combining the actions unites that are based on the face muscle.

Machine learning methods to face animation solve many of the problems found in traditional methods to face animation. If the data is available, precise models can be trained to obtain high-quality face animations. Machine learning needs an enormous of data to be able to train precise models. This type of data is not easily reachable, or is non-existent because the method is new and is not commonly utilised. Without sufficient data, models can be imprecise and produce results that would be restrict in the uncanny valley.

IV. Conclusion

Developing a facial animation comprises determining relevant geometric descriptions to represent the face model. The structured facial model should be capable to support the animation.

There are a lot of different methods to facial animation and it can be hard to get started with producing facial animation technologies. Since every approach is so exclusive, the amount of knowledge that transfers between methods is limited. Most workshops have their particular face animation pipelines and the lack of universal standards make it so there is no common method across industry. Even when attempting to find assistance for smaller projects, valuable information is difficult to come by because of the huge number of methods.

In this survey, we discuss and review several approaches utilised in driving the facial animation. In addition, we discuss the state-of-the art facial animation techniques. Within each method, the main ideas, and the strength and weakness of each approach are described in detail.

V. Future Directions

To be able to considerably advance the field of face animation with machine learning, future study will be required to address the problems that make machine learning ineligible to developers. Machine learning methods to facial animation have had favorable results, however complications with model formation have prohibited its wide-spread utilize. Generating machine learning models needs a great amount of labeled data and the data sets essential for training a model are hard and time-consuming to produce. To increase utilize of machine learning in face animation, future work will be required to research and produce solutions to permit for easier model production.

Conflict of Interest

The authors have no conflict of relevant interest to this article.

References

[1] M. Elson, "Displacement" facial animation techniques," *Vol 26: State of the Art in Facial Animation*, pp. 21–42, 1990.

[2] J. Kleiser, "A fast, efficient, accurate way to represent the human face," *SIGGRAPH'89 Course Notes 22: State of the Art in Facial Animation*, pp. 36–40, 1989.

[3] Z. Deng, J. Bailenson, J. P. Lewis, and U. Neumann, "Perceiving visual emotions with speech," in *International Workshop on Intelligent Virtual Agents. Springer*, 2006, pp. 107–120.

[4] J. Fordham, "Middle earth strikes back," *Cinefex*, vol. 92, pp. 71–142, 2003.

[5] M. Sagar1, "Facial performance capture and expressive translation for king kong," in *ACM SIGGRAPH 2006 Courses*.

[6] B. Flueckiger, "Computer-generated characters in avatar and benjamin button," *Digitalitat und Kino. Translation from German by B. Letzler*, vol. 1, 2011.

[7] F. I. Parke and K. Waters, *Computer facial animation*. CRC press, 2008.

[8] F. I. Parke, "A parametric model for human faces." UTAH UNIV SALT LAKE CITY DEPT OF COMPUTER SCIENCE, Tech. Rep., 1974.

[9] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, 1978.

[10] W. F. P. Ekman and J. Hager, "Facial action coding system," *The Manual on CD ROM, A Human Face, Salt Lake City, Tech. Rep.*, 2002.

[11] C. M. Haase and et al., "Short alleles, bigger smiles? The effect of 5-httlpr on positive emotional expressions." *Emotion*, vol. 15, no. 4, p. 438, 2015.

[12] A. A. Gunawan et al., "Face expression detection on kinect using active appearance model and fuzzy logic," *Procedia Computer Science*, vol. 59, pp. 268–274, 2015.

[13] I. M.Menne and B. Lugrin, "In the face of emotion: a behavioral study on emotions towards a robot using the facial action coding system," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 205–206.

[14] P. Tripathi, K. Verma, L. Verma, and N. Parveen, "Facial expression recognition using data mining algorithm," *Journal of Economics, Business and Management*, vol. 1, no. 4, pp. 343–346, 2013.

[15] C. Butler, L. Subramanian, and S. Michalowicz, "Crowdsourced facial expression mapping using a 3d avatar," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2016, pp. 2798–2804.

[16] R. Amini, C. Lisetti, and G. Ruiz, "Hapfacs 3.0: Facs-based facial expression generator for 3d Speaking virtual characters," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 348–360, 2015.

[17] R. Ekman, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[18] D. Kumar and D. Sharma, "Enhanced waters 2d muscle model for facial expression generation." in *VISIGRAPP (1: GRAPP)*, 2019, pp. 262–269.

[19] D. Kumar and J. Vanualailai, "Low bandwidth video streaming using facs, facial expression and animation techniques." in *VISIGRAPP (1: GRAPP)*, 2016, pp. 226–235.

[20] Y. Zhou and B. E. Shi, "Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE*, 2017, pp. 370–376.

[21] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 818–833.

[22] K. Zhao.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3391–3399.

[23] Z. Liu, D. Liu, and Y. Wu, "Region based adversarial synthesis of facial action units," in *International Conference on Multimedia Modeling. Springer*, 2020, pp. 514–526.

[24] Z. Liu, J. Dong, C. Zhang, L. Wang, and J. Dang, "Relation modeling with graph convolutional networks for facial action unit detection," in *International Conference on Multimedia Modeling. Springer*, 2020, pp. 489–501.

[25] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv: 1609.02907*, 2016.

[26] A. Pakstas, R. Forchheimer, and I. S. Pandzic, *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. John Wiley & Sons, Inc., 2003.

[27] A. El Rhalibi, C. Carter, S. Cooper, and M. Merabti, "Highly realistic mpeg-4 compliant facial animation with charisma," in *2011 Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 2011, pp. 1–6.

[28] S. M. Platt and N. I. Badler, "Animating facial expressions," in *Proceedings of the 8th annual conference on Computer graphics and interactive techniques*, 1981, pp. 245–252.

[29] K. Waters, "A muscle model for animation three-dimensional facial expression," *Acm siggraph computer graphics*, vol. 21, no. 4, pp. 17–24, 1987.

[30] D. Terzopoulos and K. Waters, "Physically-based facial modelling, analysis, and animation," *The journal of visualization and computer animation*, vol. 1, no. 2, pp. 73–80, 1990.

[31] E. Sifakis, I. Neverov, and R. Fedkiw, "Automatic determination of facial muscle activations from sparse motion capture marker data," in *ACM SIGGRAPH 2005 Papers*, 2005, pp. 417–425.

[32] M. Cong, M. Bao, J. L. E, K. S. Bhat, and R. Fedkiw, "Fully automatic generation of anatomical face simulation models," in *Proceedings of the 14th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2015, pp. 175–183.

[33] A. E. Ichim, P. Kadleˇcek, L. Kavan, and M. Pauly, "Phace: physics-based face modeling and animation," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–14, 2017.

[34] W.-C. Ma, Y.-H. Wang, G. Fyffe, B.-Y. Chen, and P. Debevec, "A blendshape model that incorporates physical interaction," *Computer Animation and Virtual Worlds*, vol. 23, no. 3-4, pp. 235–243, 2012.

[35] Y. Kozlov, D. Bradley, M. Bacher, B. Thomaszewski, T. Beeler, and M. Gross, "Enriching facial blendshape rigs with physical simulation," in *Computer Graphics Forum*, vol. 36, no. 2. Wiley Online Library, 2017, pp. 75–84.

[36] ISO/IEC 14496-2:1999. *Information technology – Coding of audio-visual objects – Part 2: Visual*. ISO, Geneva, Switzerland. 2010.

[37] D. Bennett, "The faces of" the polar express"," in *ACM Siggraph 2005 Courses*.[38] L. Williams, "Performance-driven facial animation," in *Acm SIGGRAPH 2006 Courses*.

[39] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer, "High resolution passive facial performance capture," in *ACM SIGGRAPH 2010 papers*, 2010, pp. 1–10.

[40] T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R.W. Sumner, and M. Gross, "Highquality passive facial performance capture using anchor frames," in *ACM SIGGRAPH 2011 papers*, 2011, pp. 1–10.

[41] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[42] J. R. Tena, F. De la Torre, and I. Matthews, "Interactive region-based linear 3d face models," in *ACM SIGGRAPH 2011 papers*, 2011, pp. 1–10.

[43] C. Cao, D. Bradley, K. Zhou, and T. Beeler, "Real-time high-fidelity facial performance capture," *ACM Transactions on Graphics (ToG)*, vol. 34, no. 4, pp. 1–9, 2015.

[44] E. Sifakis, A. Selle, A. Robinson-Mosher, and R. Fedkiw, "Simulating speech with a physics-based facial muscle model," in *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, 2006, pp. 261–270.

[45] G. Borshukov, J. Montgomery, and W. Werner, "Playable universal capture: compression and real-time sequencing of image-based facial animation," in *ACM SIGGRAPH 2006 Courses*.

[46] V. Barrielle, N. Stoiber, and C. Cagniart, "Blendforces: A dynamic framework for facial animation," in *Computer Graphics Forum*, vol. 35, no. 2. Wiley Online Library, 2016, pp. 341–352.

[47] Z. Deng, P.-Y. Chiang, P. Fox, and U. Neumann, "Animating blendshape faces by cross-mapping motion capture data," in *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, 2006, pp. 43–48.

[48] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.

[49] R. Ford. Use animoji on your iphone x and ipad pro. https://support.apple.com/engb/HT208190

[50] J. M. D. Barros, V. Golyanik, K. Varanasi, and D. Stricker, "Face it!: A pipeline for real-time Performance driven facial animation," in *2019 IEEE International Conference on Image Processing (ICIP). IEEE*, 2019, pp. 2209–2213.

[51] K. Olszewski, J. J. Lim, S. Saito, and H. Li, "High-fidelity facial and speech animation for vr hmds," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–14, 2016.

[52] S. Laine, T. Karras, T. Aila, A. Herva, S. Saito, R. Yu, H. Li, and J. Lehtinen, "Production-level facial performance capture using deep convolutional neural networks," in *Proceedings of the ACM SIGGRAPH/ Eurographics Symposium on Computer Animation*, 2017, pp. 1–10.

[53] N. Kholgade, I. Matthews, and Y. Sheikh, "Content retargeting using parameter-parallel facial layers," in *Proceedings of the 2011 ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, 2011, pp.195–204.

[54] M. M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and techniques in computer animation*. Springer, 1993, pp. 139–156.

[55] B.-J. Theobald and I. Matthews, "Relating objective and subjective performance measures for aam-based visual speech synthesis," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 8, pp. 2378–2387, 2012.

[56] W. Mattheyses, L. Latacz, andW. Verhelst, "Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis," *Speech Communication*, vol. 55, no. 7-8, pp. 857–876, 2013.

[57] S. L. Taylor, M. Mahler, B.-J. Theobald, and I. Matthews, "Dynamic units of visual speech," in *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, 2012, pp. 275–284.

[58] J. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise, "Accurate visible speech synthesis based on concatenating variable length motion capture data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 2, pp. 266–276, 2006.

[59] G. Englebienne, T. Cootes, and M. Rattray, "A probabilistic model for generating realistic lip movements from speech," in *Advances in neural information processing systems*, 2008, pp. 401–408.

[60] S. Deena, S. Hou, and A. Galata, "Visual speech synthesis using a variable-order switching shared gaussian process dynamical model," *IEEE transactions on multimedia*, vol. 15, no. 8, pp. 1755–1768, 2013.

[61] D.W. Massaro, J. Beskow, M. M. Cohen, C. L. Fry, and T. Rodgriguez, "Picture my voice: Audio to visual speech synthesis using artificial neural networks," in *AVSP'99-International Conference on Auditory-Visual Speech Processing*, 1999.

[62] M. Tamura, T. Masuko, T. Kobayashi, and K. Tokuda, "Visual speech synthesis based on parameter generation from hmm: Speech-driven and text-and-speech-driven approaches," in *AVSP'98 International Conference on Auditory-Visual Speech Processing*, 1998.

[63] D. Schabus, M. Pucher, and G. Hofer, "Joint audiovisual hidden semi-markov model-based speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 336–347, 2013.

[64] G. Hofer, J. Yamagishi, and H. Shimodaira, "Speech-driven lip motion generation with a trajectory hmm," 2008.

[65] M. Brand, "Voice puppetry," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 21–28.

[66] L. Xie and Z.-Q. Liu, "A coupled hmm approach to video-realistic speech animation," *Pattern Recognition*, vol. 40, no. 8, pp. 2325–2340, 2007.

[67] K. Choi, Y. Luo, and J.-N. Hwang, "Hidden markov model inversion for audio-to-visual conversion in an mpeg-4 facial animation system," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 29, no. 1-2, pp. 51–61, 2001.

[68] L. D. Terissi and J. C. Gomez, "Audio-to-visual conversion via hmm inversion for speech-driven facial animation," in *Brazilian Symposium on Artificial Intelligence*. Springer, 2008, pp. 33–42.

[69] X. Zhang, L.Wang, G. Li, F. Seide, and F. K. Soong, "A new language independent, photo-realistic talking head driven by voice only." in *Interspeech*, 2013, pp. 2743–2747.

[70] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync

from audio," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–13, 2017.
[71] J. S. Chung, A. Jamaludin, and A. Zisserman, "You said that?" *arXiv preprint arXiv: 1705.02966*, 2017.
[72] H. X. Pham, S. Cheung, and V. Pavlovic, "Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 80–88.
[73] H. X. Pham, Y.Wang, and V. Pavlovic, "End-to-end learning for 3d facial animation from raw waveforms of speech," *arXiv preprint arXiv: 1710.00920*, 2017.
[74] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7832–7841.
[75] Y. Song, J. Zhu, X. Wang, and H. Qi, "Talking face generation by conditional recurrent adversarial network," *arXiv preprint arXiv: 1804.04786*, 2018.
[76] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv: 1411.1784*, 2014.
[77] K. Vougioukas, S. Petridis, and M. Pantic, "End-to-end speech-driven realistic facial animation with temporal gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 37–40.
[78] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with gans," *International Journal of Computer Vision*, pp. 1–16, 2019.
[79] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9299–9306.
[80] H. Guo, F. K. Soong, L. He, and L. Xie, "A new gan-based end-to-end tts training algorithm," *arXiv preprint arXiv: 1904.04775*, 2019.
[81] O. Aina, and J. Zhang, "Automatic muscle generation for physically-based facial animation." *ACM SIGGRAPH 2010 Posters*. 2010.