Sys Rev Pharm 2020;11(12):907-915
A multifaceted review journal in the field of pharmacy

Survey Interactive Synthesis of Animation Model Based on Tracking User's Utterance and Facial Expression

Rana Ali salim

Fine art institute, Iraq, raname802@gmail.com

ABSTRACT

Synthesizing expressive facial animation is a very challenging topic within the graphics community. Where Two-dimensional techniques have been used before, which lose a person's expressions, identity and embodiment in the image, thus unrealistic images that show an image without any expressions as a sad face when simulated from a distance, as well as generate lip and face movements of an unreal character by simulating a real character. The purpose of this paper is to review some of the ways through which to generate face animations and lip movements based on simulating the person's real face to arrive at a conclusion which of these methods are best for the purpose of adopting them and making some developments on them. The process of combining the AU with the deep learning using CNN algorithm gives excellent results and this is indicated by the 2019 research.

Keywords: Action Unit (AU) Annotation, 3D facial animation, blend shape, 2D facial animation.

Correspondence:
Rana Ali salim
Fine art institute, Iraq, raname802@gmail.com

INTRODUCTION

In the context of computer graphics, virtual reality, and human user interaction, having intuitive control over three-dimensional facial expressions is a significant issue. Such a program will be useful for a range of applications including video games, electronically mediated communication, user interface agents and teleconferencing. Nonetheless, there are two difficulties in animating a computer-generated face model: creating a rich collection of plausible acts for the virtual character and giving the user intuitive control of those acts. One approach to animating movement is to mimic the muscle and skin dynamics. Creating and simulating such a model has proved to be an extremely difficult activity due to the subtlety of the motions of facial skin [1]. An alternative approach is the use of data collection by motion. Recent technical advancements in motion capture equipment allow high fidelity, resolution, and accuracy to record the three-dimensional facial expressions. While motion capture data is a reliable way of capturing the information and complexity of live action, reuse and alteration remains a challenging task for a specific purpose. It is difficult to provide the user with an intuitive interface to monitor a wide variety of facial expressions, because character expressions are high-dimensional but most input devices available are not Precise motion capture, however, requires the subject's expensive hardware and extensive instrument, and is therefore not widely available or practical. A vision-based interface will provide an inexpensive and non-intrusive alternative to interactively controlling the avatar, though precise, high-resolution, and real-time facial tracking systems have not yet been developed [2]. A lot of techniques and tools can be used to create facial animations. These tools can be as simple as a pencil and sketchpad, or as complicated as a 3D face model based on physics. In either case creating credible facial animations is a very delicate task that can only be accomplished by talented and well-trained animators. An alternative is the development of facial animations based on video images of the face of a real person. In the video the face can be monitored by retrieving the location and expression at each frame. Then this information can be further modified optionally to create a new animation. It is a difficult problem to recover the face location and facial expression automatically from a video. The challenge stems from the wide variety of appearances that the human face can produce under various orientations, conditions of lighting, and facial expressions. It is therefore a very challenging task to create a model that encompasses these variations and allows the robust estimation of the face parameters [3].

Facial animation is one alternative to allow natural interaction of human computers. Computer facial animation has many uses. Realistic simulated humans with facial expressions are increasingly being used for example in the entertainment industry. Interactive talking faces in communication applications not only make user-machine interaction more enjoyable, but also provide a friendly interface and help attract users [4].

Humanlike speech is important among the questions concerning the realism of synthesized facial animation. Despite the need for synthesis of expressive facial animation in these various applications, the computer graphics community still has a very challenging topic to deal with. It is because the deformation of a moving face is complex and we humans have an innate sensitivity to the subtleties of facial expression but also because human emotion is a very complicated interdisciplinary research subject studied by researchers in computer graphics, artificial intelligence, communication, psychology, etc.[5].

Due to its many practical applications ranging from language instruction for hearing impaired people to film and game productions, animated agents for human – computer interaction, virtual avatars, model-based image coding in MPEG4 and online commerce, animating correct visible speech is one of the most relevant research areas of face animation [6].

Motion capture technologies have been used successfully in body animation of the characters. Realist motions can be generated efficiently by recording motions directly from real actors and mapping them to character models. While techniques based on performance capture for 3D face animation have been suggested in the last decade, accurate visible speech recognition for any text is still a challenge due to complex facial muscles in regions of the lip and significant presence of co-articulation in recognizable speech Some of the techniques proposed in previous work include the adaptation of a standardized 3D face model with recorded facial movements to the face of the subject. Such methods cannot be extended to the situation where the face of the subject varies dramatically from the 3D face model. Recent research focuses on the issues of automated visible speech synthesis for 3D face-models with specific meshes [7].

Highlight in this paper the benefits of three-dimensional methods on two-dimensional methods, which are commonly assumed to have the ability to improve precision and intensity in recognition. Our poll for the newcomer offers the most current scenario and the best results in the field of fully understanding facial expressions as compared to previous research focusing on this subject. We include and evaluate the 3D methods available, concentrating primarily on facial expressions and lip movements.

Related works

Although there have many agent facial synthesis methods, the facial emotion expression of 3D animation might be an acceptable model for human perception. To date, there are varies techniques has been proposed to synthesize naturalistic emotion expression from 3-D facial animation.

In 2013, Terissi et al, this audio-visual information is extracted from audio-visual training data and then used to measure the parameters of a single audio-visual hidden Markov model (AV-HMM). The trained AV-HMM provides a compact representation of audio-visual data without the need for phoneme (word) segmentation, making it adaptable to different languages. Visual features are measured based on the AV-HMM's inversion from the speech signal. The predicted features of visual speech are used to animate a graphical model of the face. Animation of a more complex head model is then obtained by mapping the deformation of the simple model to it automatically, using a small number of control points for the interpolation. The algorithm proposed allows for the animation of arbitrary complexity 3D head models through a simple setup process. The resulting animation is tested, showing a promising output, in terms of visual speech intelligibility via perceptual tests. The computational complexity of the proposed method is analysed, demonstrating the viability of applying it in real time.

In 2014, Zhigang Deng and Ulrich Neumann, Present the novel motion capture mining technique that "learns" voice co-articulation models from the captured data for diphones and triphones. A Phoneme-Independent Speech Eigenspace (PIEES) that encompasses the signals of dynamic expression is constructed by movement signal processing (phoneme-based time-warping and subtraction) and reduction of the main component analysis (PCA). New expressive facial animations are synthesized as continues to follow: First, the learned co-articulation models are concatenated to synthesize neutral visual speech according to novel speech input, then a texture-synthesis method is used to produce a new dynamic expression signal from the PIEES model, Finally, the synthesized voice signal is combined with the synthesized visual neutral speech to create the final expressive animation of the face. Their studies show that the device can effectively synthesize practical facial expressiveness [8].

In 2017, Lingxiao Song et.al Proposes a G2-GAN (Geometry-Guided Generative Adversarial Network) for photo-realistic and identity-preserving synthesis of face expression. Therefore, using facial geometry (fiducial points) as a controllable condition to direct the synthesis of facial texture with a particular expression. A pair of generative adversarial subnetworks are trained jointly for opposing tasks: elimination of expression and synthesis of expression. The paired networks form a mapping loop between neutral speech and arbitrary expressions, which also facilitates other applications such as invariant face recognition and face transfer and speech. Experimental results show that the method can produce convincing perceptional results on various facial expression synthesis databases and the accuracy 97% [9].

In 2018 Hu Ni et.al, To apply to human-robot contact, created a new 3-D virtual speaking head device with complex emotional facial expression. In addition, the eye-tracking experiment and subjective appraisal tests were used to investigate the emotional experience of the virtual 3-D talk brain. The results showed that there was no substantial difference in observation mode between 3-D virtual speaking head videos (AV3D) and human face audiovisual (AVHF) animation. In addition, the accuracy of recognition of human face videos (HF) was higher than 3D, and almost all of the emotional accuracy had been enhanced by adding audio to images. Finally, the results demonstrated that happiness was identified the best whether watching 3-D virtual talking head videos (3D) or human face videos (HF). These results implied that the 3-D talking head has potentially been as a suitable natural communication form in human-computer interaction [10].

In 2019, Ekmen, Beste, and Hazım Kemal Ekenel Offer a three-stage approach creating realistic facial animations by recording expressions of the human face in 2D and moving them in real-time to a human-like 3D model. No training is required for their calibration-free process, which is based on an average human face. Tracking is achieved using a single camera to allow multiple realistic applications where the expressions are transported with a joint-based system to enhance animation quality and persuasiveness. Firstly, the 3D model is attached to a joint-based facial rig which provides mobility to pseudo-muscles. The second stage includes the analysis of the facial landmarks' 2D locations from a single camera view and the transfer of relative 3D motion data to move the respective joints on the model. The final step is to capture the animation using a partially automated key-framing technique. Experiments using maximum frames in facial-view videos on the extended Cohn-Kanade dataset have shown that the presented approach generates visually satisfactory facial animations [11].

Facial Animation

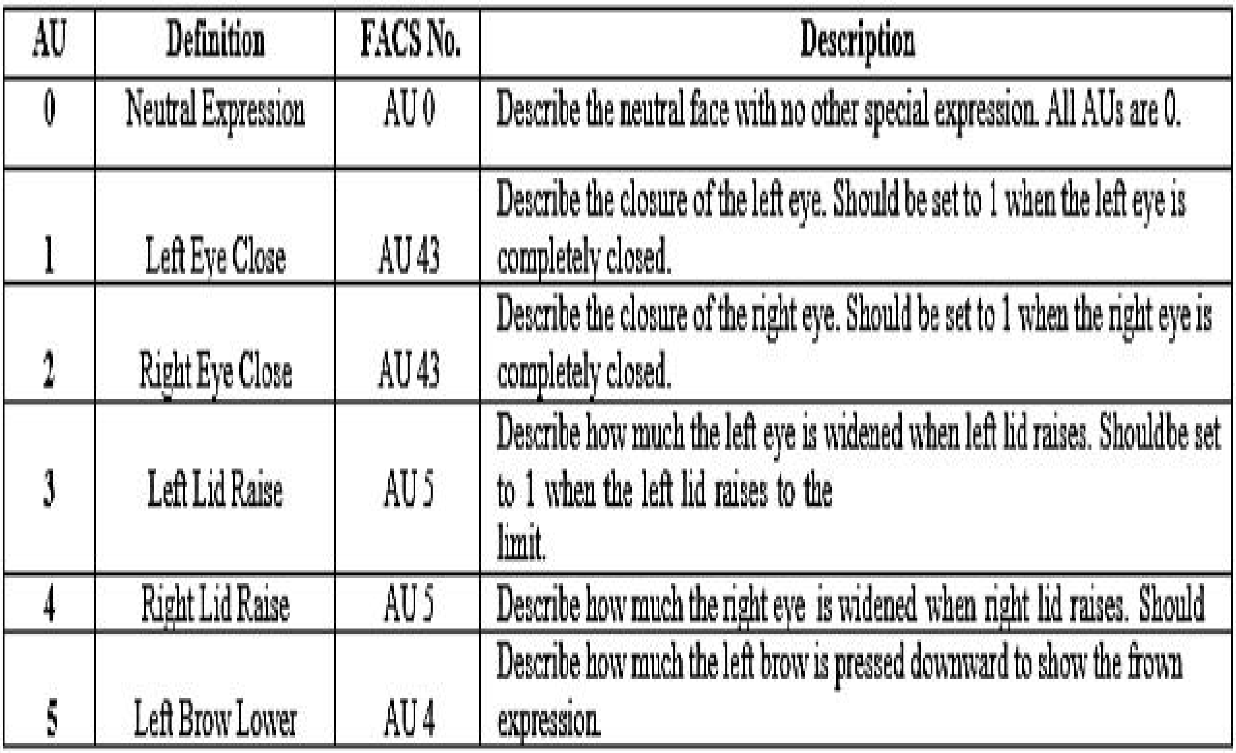
Modeling a real human, i.e. modeling the 3D geometry of an individual face, is a major problem in facial animation. Three-dimensional coordinates may be determined by a range scanner, digitizer probe or stereo disparity. Often the models produced through these methods are poorly suited for facial animation. Information on the facial structures is missing; measuring noise creates distracting artifacts; and poorly distributed vertices of the model. Also, several methods of measurement produce incomplete models which lack hair, ears, eyes, etc. Therefore, it is always appropriate to post-process the calculated data [12].

The difficulties in controlling facial animations led to the performance driven approach where tracked human actors drive the animation. Real time video processing allows interactive animations where the actors observe the animations they create with their motions and expressions. Accurate tracking of feature points or edges is important to maintain a consistent and quality of animation. Often the tracked 2D or 3D feature motions are filtered or transformed to generate the motion data needed for driving a specific animation system. Motion data can be used to directly generate facial animation [13].

Action Units (AU) Annotation

This approach provides a parameter structure for quantizing facial expression with 25 redefined action units (AUs) to explain the facial expression in great detail. This approach picked nine symmetric FACS AUs, 10 asymmetric (unilateral) FACS AUs, two symmetric FACS ADs, and 2 asymmetric FACS ADs to classify most of human face expressions. The FACS AUs were reorganized and renumbered to promote facial expression analysis based on our database, in particular the blend formation method for 3D facial animation .This method offers a system of parameters to quantize facial expression with 25 redefined units of action (AUs) to describe the facial expression in great detail. This method selected nine symmetric FACS AUs, 10 asymmetric (unilateral) FACS AUs, two symmetric FACS ADs, and 2 asymmetric FACS ADs to describe most of the expressions in human face. Based on our database, the FACS AUs were reorganized and renumbered to encourage facial expression analysis, in particular the blend forming method for facial animation in 3D. For AU annotation, labels need to be more precise than those of FACS which has only five levels for each unit of action. Therefore, use floating point numbers from 0 to 1 and correctly calculate each AU to two decimal places. A facial action state similar to a neutral state is given a corresponding AU value close to 0; a corresponding AU value closer to 1. is given greater deviations of the facial action state from the neutral state; This unique method of annotation is instrumental in the identification of AU values with coefficient of expression. The meaning and description for each AU are given in Table 1.

Table 1. AU Definition and Description

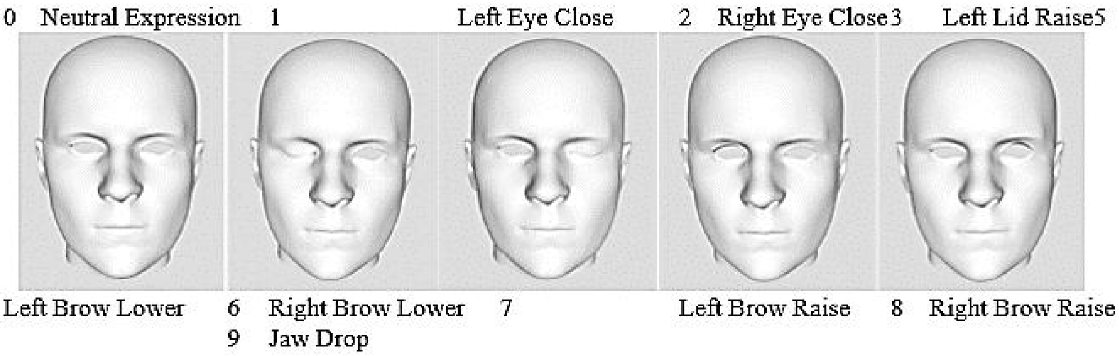


AU	Definition	FACS No.	Description
0	Neutral Expression	AU 0	Describe the neutral face with no other special expression. All AUs are 0.
1	Left Eye Close	AU 43	Describe the closure of the left eye. Should be set to 1 when the left eye is completely closed.
2	Right Eye Close	AU 43	Describe the closure of the right eye. Should be set to 1 when the right eye is completely closed.
3	Left Lid Raise	AU 5	Describe how much the left eye is widened when left lid raises. Shouldbe set to 1 when the left lid raises to the limit.
4	Right Lid Raise	AU 5	Describe how much the right eye is widened when right lid raises. Should
5	Left Brow Lower	AU 4	Describe how much the left brow is pressed downward to show the frown expression.

6	Right Brow Lower	AU 4	Describe how much the right brow is pressed downward to show the frown expression.
7	Left Brow Raise	AU 2	Describe how much the left brow raises with left lid raising, to illustrate surprise.
8	Right Brow Raise	AU 2	Describe how much the right brow raises , with right lid raising, to illustrate the surprise expression.
9	Jaw Drop	AU 26	Describe how much the mouth opens driven by the jaw. Should be set to 1 when the mouth is open to the limit.
10	Lip Slide Left	AD 30	Describe how much the lower lip slides left driven by the lower jaw.
11	Lip Slide Right	AD 30	Describe how much the lower lip slides right driven by the lower jaw.
12	Left Lip Corner Pull	AU 12	Describe how much the left lip corner raises, which causes the left cheek to raise as well, and AU1 may be involved.
13	Right Lip Corner Pull	AU 12	Describe how much the right lip corner raises, which causes the right cheek to raise as well, and AU2 may be involved.
14	Left Lip Corner Stretch	AU 20	Describe how much the left lip corner stretches to the left, to illustrate the action of lip corner and smile face in daily chat. Should be set to 1 when stretching to the limit.
15	Right Lip Corner Stretch	AU 20	Describe how much the left lip corner stretches to the right, to illustrate the action of lip corner and smile face in daily chat. Should be set to 1 when stretching to the limit.
16	Upper Lip Suck	AU 28	Describe how much the upper lip purses to illustrate the pursing expression.
17	Lower Lip Suck	AU 28	Describe how much the lower lip purses to illustrate the pursing
18	Jaw Thrust	AD 29	Describe how much the lower lip moves outward. The variation of this expression is tiny.
19	Upper Lip Raise	AU 10	Describe how much the upper lip raises, which causes the wing of nose to raise. This expression is not driven by palate.
20	Upper Lip Raise	AU 16	Describe how much the lower lip moves down with the jaw moving down as well. This expression is not driven by the palate. Notice that the combined effect of AU19 and AU20 is not equal to the effect of AU9.
21	Lower Lip Raise	AU 17	Describe how much the left and right lip corners move down driven by the lower lip raising.
22	Lip Pucker	AU 18	Describe how much the left and right lip corners move toward each other with lip wrinkling and pouting.
23	Cheeks Puff	AD 34	Describe how much the cheeks puff by filling them with the air.
24	Nose Wrinkle	AD 9	Describe how much the nose raises with wrinkling, which usually illustrates disgust.

Relationship between AUs and expression blend shapes

Similar to several techniques of facial animation, they reflect types of blend of facial expressions in terms of language. In order to generate any possible expression of the source actors, the shape of a neutral face blend and 24 types of blend of expression are required. An example of the expression blend shapes selected from Face Warehouse is shown in Figure 1 and the participants can express themselves according to this model of facial expression.



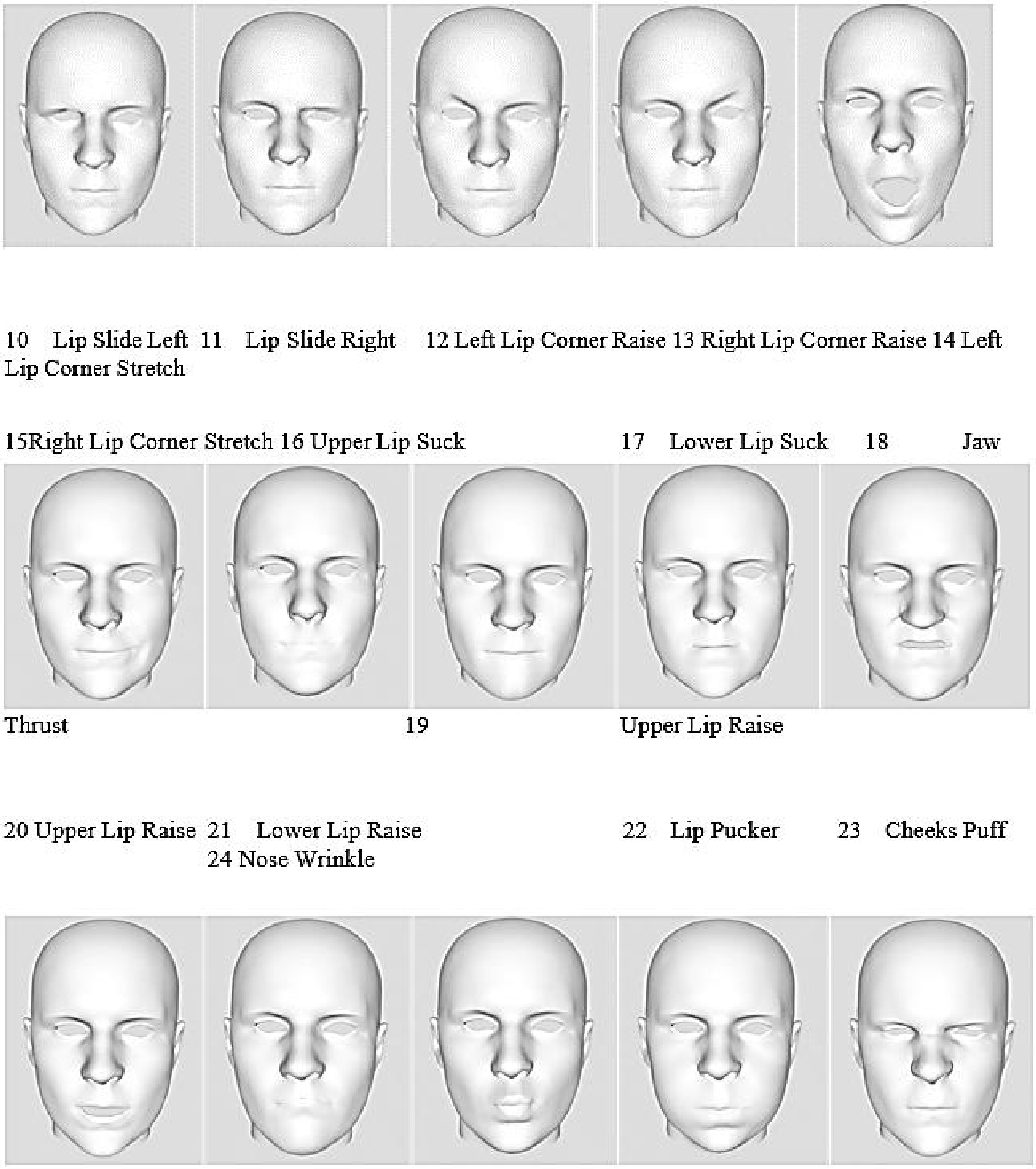

10 Lip Slide Left 11 Lip Slide Right 12 Left Lip Corner Raise 13 Right Lip Corner Raise 14 Left Lip Corner Stretch
15Right Lip Corner Stretch 16 Upper Lip Suck 17 Lower Lip Suck 18 Jaw
Thrust 19 Upper Lip Raise
20 Upper Lip Raise 21 Lower Lip Raise 22 Lip Pucker 23 Cheeks Puff
24 Nose Wrinkle

Figure 1. Relationships between AUs and Expression Blend shapes

Required Expression

When recording the videos, the expressions needed are as follows:

1. Closure of the arms, for AU1 and AU2.

2. Amazing. Drop the upper lid and lift the outer lip, AU3AU4AU7and AU8.

3. Frown ...... frown. Browse the AU5 and AU6 at the bottom and maybe the AU1 and AU2 at the eyes.

4. Mouth open, around AU9.

5. Stretcher to the neck. The left corner of the lip extends first, and then back to neutral language. Second, the right corner of the lip extends, and then returns to neutral language. Finally, both the left corner of the lip and the right corner of the lip extend at the same time as for AU14 and AU15. The variation of that phrase is relatively subtle.

6. Suck. Suck. Upper lip sucks, then lower lip sucks, and then both suck at the same time, for AU16 and AU17.
7. Smile. Smile. Grin without teeth, about AU12 and AU13 and then toothy smile, about AU12, AU13 and AU9.
8. Jaw lateral. Jaw slides left and then right, with respect to AU10 and AU11.
9. Puller to the Lip Corner. Left corner pull (AU12), maybe related to AU1; then right corner pull (AU13), maybe related to AU2; Others. Lip pucker (AU18). Upper lip raiser (AU19). Lower lip depressor (AU20). Chin Raiser (AU21). Lip Pucker (AU22). Cheeks Puff (AU23). Nose Wrinkler (AU24)
10. Mixture expressions.

a. Boost lids (AU3, AU4) and boost the outer brows (AU7, AU8) and lower jaw (AU9);
b. Jaw falls (AU9), and left and right slides (AU10, AU11).
c. Depressed by frown (AU5, AU6, AU1, AU2) and lip (AU21).

d. Close eyes (AU1 u2) and puffy cheeks (AU23).

e. Fall in the jaw (AU9) and upper lip (AU19) and lower lip depressions (AU20).

f. Left corner of the lip lifts (AU12) and left eye closes (AU1) and left depressions of the forehead (AU5), perhaps in comparison to AU2, AU6.

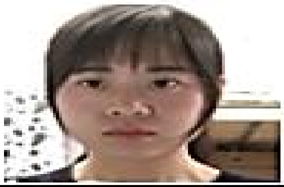
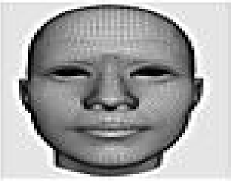
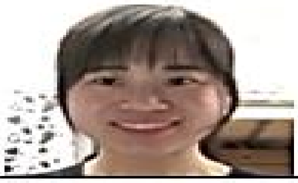
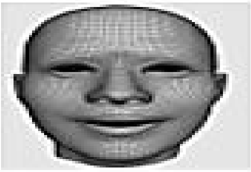
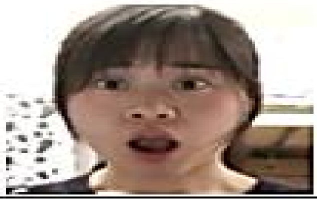
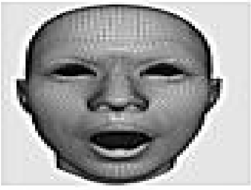
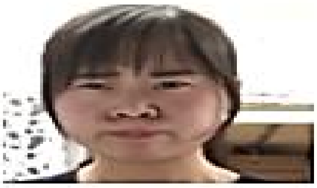
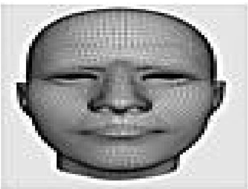
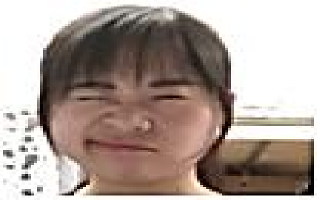
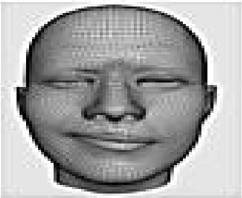
g. Right corner of the lip lifts (AU13) and right eye closes (AU2) and right brow depressions (AU6), perhaps in comparison to AU1, AU5.

Participants should choose at least 3 expressions and make them trained by our members. Every expression should repeat two or three times

Illustration for LabellingAUs

Indeed, because the elicited expression gradually shifts frame by frame, AU Annotations should also be shifting smoothly. In the meantime, the elicited expression will be different because of individual variations, even if the participants render the same expression. Therefore, we need to set right AU values during labeling to make the expression as close as possible on source video frames and the expression on created 3D facial model. In addition, we need to calculate the deviations of the particular expression from neutral expression to set correct AU values for the current topic. Table 2 offers a few examples of how to mark AUs.

Table 2. Examples for labelling AUs.

Video Frames	Generated 3D Model	Labelling Advice
		All AUs should be set to 0.
		AU1: 0.10 AU2: 0.10 AU12: 0.65 AU13: 0.65 AU9: 0.28
		AU3： 0.88 AU4: 0.88 AU7： 0.85 AU8： 0.85 AU9： 0.90
		AU: 0.45 AU2: 0.35 AU5： 0.8 AU6： 0.8 AU21： 1 AU24： 0.2
		AU1： 0.90 AU2： 0.73 AU5： 0.55 AU6： 0.30 AU12： 0.95

3D Convolutional Neural Network

A 3D convolution is performed with 3D kernel and 3D data that 2D images are merged. A 3D convolution is expressed by the following Figure 2.

$$\begin{aligned} o_{(1,1,1)} = a_{(1,1,1)}b_{(1,1,1)} + a_{(1,2,1)}b_{(1,2,1)} + a_{(1,3,1)}b_{(1,3,1)} \\ + a_{(2,1,1)}b_{(2,1,1)} + a_{(2,2,1)}b_{(2,2,1)} \\ + a_{(2,3,1)}b_{(2,3,1)} + a_{(3,1,1)}b_{(3,1,1)} \\ + a_{(3,2,1)}b_{(3,2,1)} + a_{(3,3,1)}b_{(3,3,1)} \\ + a_{(1,1,2)}b_{(1,1,2)} + a_{(1,2,2)}b_{(1,2,2)} \\ + a_{(1,3,2)}b_{(1,3,2)} + a_{(2,1,2)}b_{(2,1,2)} \\ + a_{(2,2,2)}b_{(2,2,2)} + a_{(2,3,2)}b_{(2,3,2)} \\ + a_{(3,1,2)}b_{(3,1,2)} + a_{(3,2,2)}b_{(3,2,2)} \\ + a_{(3,3,2)}b_{(3,3,2)} + a_{(1,1,3)}b_{(1,1,3)} \\ + a_{(1,2,3)}b_{(1,2,3)} + a_{(1,3,3)}b_{(1,3,3)} \\ + a_{(2,1,3)}b_{(2,1,3)} + a_{(2,2,3)}b_{(2,2,3)} \\ + a_{(2,3,3)}b_{(2,3,3)} + a_{(3,1,3)}b_{(3,1,3)} \\ + a_{(3,2,3)}b_{(3,2,3)} + a_{(3,3,3)}b_{(3,3,3)} \end{aligned}$$

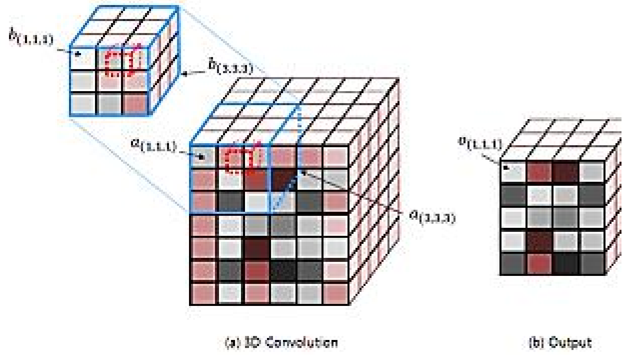


Figure 2. Example of a 3D convolution

Image 2. A 3D Convolution Example

A sub-sampling value of a pixel is determined by multiplying and accumulating each pixel of the kernel and image in a superimposed field. Here kernel values are for a mean measure. This process is carried out on the entire image, and the resulting image becomes small by dumping some pixels that overlap. A subsampling achieves a certain degree of invariance in change and deformation. A subsample is illustrated in Figure 3 below.

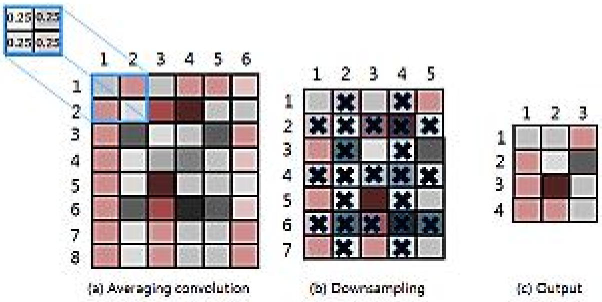


Figure 3. Example of a subsampling

Figure 4 shows a structure of 3D-CNN for facial expression recognition based on video. Here, the structure consists of 5 layers. First layer is for input, second layer is for convolution, third layer is for subsampling, forth layer is for convolution and fifth layer is for subsampling. Initial values of kernels are random in specific range. ***In a first layer***, A 3D data consisting of five video frames enters as device input. A convolution extracts features from the input in a second layer which has 3 maps. In a third layer, the size of the input image is decreased by a subsampling. A convolution extracts features from the output of previous layer in a fourth layer having 29 maps. A subsampling in a fifth layer decreases image size from previous layer. Finally, a vector of features is generated by having images on all maps organized to a single row. Data used for input is video-based representations of facial expression which are successively overlapped into five lines. The data size begins with 64x48x5 as the reference. In the second layer the data size will be 2x2x1 without time base. The data size transforms 26x18x1 at the fourth layer because a convolution is performed with a kernel of 5x5x3 size. The size of the data at the fifth layer turns 13x9x1 because a subsample is performed with a kernel whose size is 2x2x1 without time base. This 13x9x1 can be a vector by having a single row of it. That is, a map could have 117 values for features. So the last function vector size is 3393 since 29 maps are available.

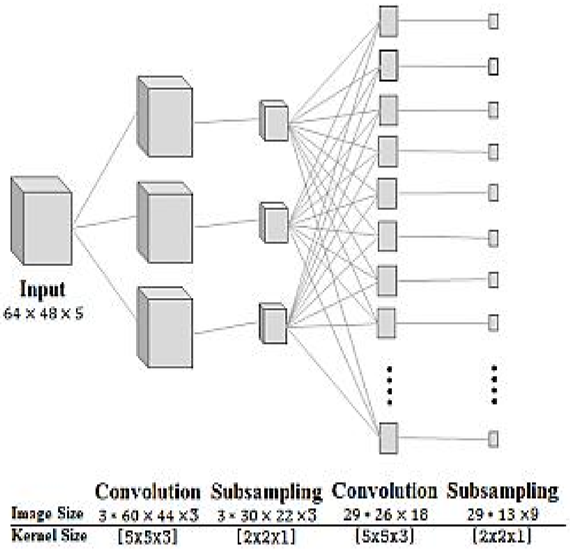


Figure 4. A structure of 3D-CNN

Speech Animation

Modeling speech articulation is an integral part of speech animation synthesis. In linguistics literature, speech articulation is described as follows: phonemes are not pronounced as an independent sequence of sounds but, instead, adjacent phonemes affect the sound of a particular phoneme. Coarticulation of visual expression is similar to that. Phoneme-driven methods enable animators to design key mouth shapes, and then use coarticulation rules for empirical smooth functions to produce voice animation.

Visible speech

Visible speech refers to the movements of the lips, tongue, and lower face during speech production by humans. According to the similarity measurement of an acoustic signal, a phoneme is the smallest identifiable unit in speech, while a viseme is a particular configuration of the lips, tongue, and lower face for a group of phonemes with similar visual outcomes. A viseme is an identifiable unit in visible speech. In English, there are many phonemes with visual ambiguity. For example, phonemes /p/, /b/, /m/ appear visually the same. These phonemes are grouped into the same viseme class. Each viseme is usually classified using 70–75% [15]. within-class recognition rate. Phonemes /p/, /b/, /m/ and /th/, /dh/ are universally recognized visemes; the remaining phonemes are not universally recognized across languages owing to variations of lip shape in different individuals. From a statistical point of view, a viseme is a random vector, because a viseme observed at different times or under different phonetic contexts may vary in its appearances. The basic underlying assumption of our visible speech synthesis approach is that the complete set of mouth shapes associated with human speech can be reasonably approximated by a linear combination of a set of visemes. Actually, this assumption has been proven to be acceptable in most authoring tools for 3D face animation. These systems use shape blending, a special example of linear combination, to synthesize visible speech. In this work, we chose 16 visemes from images of the human subject. Each viseme image was chosen at the point at which the mouth shape was judged to be at its extreme shape. Phonemes that look alike visually fall into the same viseme category. This was done in a subjective manner, by comparing the viseme images visually to assess their similarity. The 3D feature points for each viseme are reconstructed by our motion capture system. When synthesizing visible speech from text, we map each phoneme to a viseme to produce the visible speech. This ensures a unique viseme target is associated with each phoneme. We recorded sequences of nonsense words that contain all possible motion transitions from one viseme to another. After the whole corpus was recorded and digitized, the 3D facial feature points were reconstructed. Moreover, the motion trajectory of each diviseme was used as an instance of each diviseme. It should be noted that diphthongs need to be specially treated. Since a diphthong, such as /ay/ in ‘pie’, consists of two vowels with a transition between them, i.e., /aa/, /iy/, the diphthong transition is visually simulated by a diviseme corresponding to the two vowels. The mapping from phonemes to visemes is many-toone; for instance, in cases where two phonemes are visually identical, but differ only in sound, /p/, /b/, /m/. It is also important to note that the mapping from visemes to phonemes is also one-to-many. One phoneme may have different mouth shapes due to the coarticulation effect. In visible speech synthesis, a key challenge is to model coarticulation. Coarticulation20 relates to the observation that a speech segment is influenced by its neighbouring speech segments during speech production. The coarticulation effect from a phoneme’s adjacent two phonemes are referred to as the primary coarticulation effect of the phoneme. The coarticulation effect from a phoneme’s two second nearest neighbour phonemes is called the secondary coarticulation effect.

Conclusion

This study discusses and attempts to group and organize all the works and the different approaches that attempted to solve the problem of identifying facial expression through the analysis of human emotions, concentrating on 3D solutions. Traditional and deep-learning methods for facial expression analysis are comprehensive, further distinguishing between data modality (2D, 3D, and multi-modal 2D+3D), granularity of expression (prototypical facial expression and facial action units), and temporal dynamics (still images and image sequences). With this analysis, we want to provide newcomers with guidance that will address this subject, and take stock of neural networks, taking advantage of AI's golden age. The most significant works in recent years have been published, outlining the pros and cons and the best findings in the whole field in facial expression recognition. The average accuracy of expression recognition is found in a range between 60% and 90%. Certain emotions, such as rage and fear, usually have the lowest rates of recognition. The motions of these expressions are in fact mild in comparison with happiness or surprise, and thus more challenging to recognize. Regarding the action units, the experiments reached recognition rates in a Wider variety, from 50 to 95 percent. The number of features to be identified with each AU will increase to obtain more reliable results and the neural networks will increasingly be used more and more in the field of recognition of facial expression. Despite the higher dimensional and computational costs and the greater complexity of operating in real time, 3D approaches have achieved better recognition rates than the more traditional 2D approaches. Predicting the expression of the human face in real time involves identification as accurately and as quickly as possible, but when contrasted with the static it becomes very complicated images because the video is a multi-frame set, not just a single frame. In the field of artificial intelligence science, identification of emotions in real-time will soon be useful, with the ability to identify as many different people's emotions in one frame and detect mixed emotions. Many researchers have developed algorithms that identify the six basic phrases, but fewer contributions have studied other types of facial expressions or units of action. We plan to use deep learning techniques in our future work, operating on a private database containing three-dimensional images, and psychological validation of labeled emotions, to perform emotion recognition in the wild.

Financial disclosure

There is no financial disclosure.

Conflict of interest

None to declare.

Ethical Clearance

All experimental protocols were approved under the Fine art institute and all experiments were carried out in accordance with approved guidelines.

REFERENCES

1. Blan. V, Basso. C, Poggio. T and Vetter. T. "Reanimating Faces in Images and Video." Computer Graphics Forum, vol.22, no.3, 2003, pp.641-650.
2. Tao. Jianhua, and Tieniu. Tan. "Emotional Chinese talking head system." International Conference on Multimodal Interfaces ACM, 2004, pp. 273-280.
3. Deng. Zhigang, Lewis. J. P and Lim. Tae-Yong. "Expressive facial animation synthesis by learning speech coarticulation and expression spaces." IEEE transactions on visualization and computer graphics, vol.12, no.6, 2001, pp.1523-1534.
4. J. Chen, D. Yi, J. Yang, G. Zhao, S. Z. Li, and M. Pietikainen. Learning mappings for face synthesis from near infrared to visual light images. In CVPR, 2009.
5. Nook. Erik C, Lindquist. Kristen A, and Zaki. Jamil. "A new look at emotion perception: Concepts speed and

shape facial emotion recognition." Emotion, vol.15, no. 5, 2015, pp.569-578.

6. Nirme, Jens & Haake, Magnus & Gulz, Agneta & Gullberg, Marianne. (2019). Motion capture-based animated characters for the study of speech–gesture integration. Behavior Research Methods. 52. 10.3758/s13428-019-01319-w.
7. Ma, Jiyong & Cole, Ronald & Pellom, Bryan & Ward, Wayne & Wise, Barbara. (2004). Accurate automatic visible speech synthesis of arbitrary 3D model based on concatenation of diviseme motion capture data. Journal of Visualization and Computer Animation. 15. 485-500. 10.1002/cav.11.
8. Deng. Zhigang, and U. Neumann. "eFASE:expressive facial animation synthesis and editing with phoneme-isomap controls." ACM Siggraph/eurographics Symposium on Computer Animation Eurographics Association, 2014, pp.251-260.
9. Song, Lingxiao & Lu, Zhihe & He, Ran & Sun, Zhenan & Tan, Tieniu. (2017). Geometry Guided Adversarial Facial Expression Synthesis.
10. H. Ni, J. Wang, L. Wang and N. Yan, "Track Your Emotional Perception of 3-D Virtual Talking Head in Human-computer Interaction," 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS), Shenzhen, 2018, pp. 298-303.
11. Ekmen, Beste, and Hazım Kemal Ekenel. "From 2D to 3D real-time expression transfer for facial animation." Multimedia Tools and Applications 78.9 (2019): 12519-12535.
12. Chai, Jin-xiang, Jing Xiao, and Jessica Hodgins. "Vision-based control of 3d facial animation." Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation. Eurographics Association, 2003.
13. Breton, Gaspard, Christian Bouville, and Danielle Pelé. "FaceEngine a 3D facial animation engine for real time applications." Proceedings of the sixth international conference on 3D Web technology. 2001.
14. Terissi. Lucas D, Cerda Mauricio, Gomez. Juan C, Hitschfeld-Kahler Nancy and Girau. Bernard. "A comprehensive system for facial animation of generic 3D head models driven by speech." EURASIP Journal on Audio, Speech, and Music Processing, vol.1, no. 2013, 2013: pp. 1-18.
15. Jackson PL. The theoretical minimal unit for visual speech perception: visemes and coarticulation. Volta Review 1988; 90(5): 99–115.