IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 18 February 2024, accepted 2 March 2024, date of publication 12 March 2024, date of current version 18 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3374777

RESEARCH ARTICLE

Low-Rank Active Learning for Generating Speech-Drive Human Face Animation

HUI XU[1], XIAOYANG YU[2], YU CHENG[3], MENGQIONG XIAO[4], AND YUE YU[5]
[1]School of Electronic and Information Engineering, Wuhan Technical College of Communications, Wuhan 430071, China
[2]School of Transportation Information, Hubei Communications Technical College, Wuhan 430050, China
[3]Wuhan Intelligent Connectivity Technologies Company Ltd., Wuhan 430074, China
[4]China University of Geosciences, Wuhan 430079, China
[5]College of Computer Sciences, Anhui University, Hefei 230093, China

Corresponding author: Xiaoyang Yu (XiaoyangYu@tom.com)

ABSTRACT Emotion&speech-based human facial animation technique can be considered as a useful application in many artificial intelligent systems. Given a speech signal, the recognizer output a sequence of the phoneme and emotion pairs. Thereby, we calculate the sequence of viseme and expression pairs accordingly, which are subsequently transformed to a consistent and synchronous video describing facial animation. This article introduces a novel facial animation technique that can intelligently generates real human face animation videos by leveraging an emotional speech. More specifically, we first extract acoustic features sufficiently discriminative to the emotion and phoneme pairs. And the corresponding sequence of phoneme and emotion pairs are computed. Next, we propose a low-rank active learning paradigm for discovering multiple key facial frames that can best represent the above phoneme and emotion pairs in the feature subspace. We associate each phoneme and emotion pair with a key facial frame, based on which the well-known morphing technique fits the associated key facial frames to a smooth animated facial video. We focus on generating multiple transitional facial frames between pairwise temporally adjacent key ones. Experiments demonstrated that the synthesized facial videos look real, smooth, and synchronous with different male/female speeches.

INDEX TERMS Facial animation, low-rank, feature selection, morphing, active learning.

I. INTRODUCTION

Synthesizing facial animation video using human speech [1] is an important technique that is pervasively applied in modern AI systems. As an example, this technique is helpful for fully/partially hearing impaired people recognizing speech in noisy environments. Besides, it is significant for synthesizing human lip movements, which is a widely-used technique in virtual reality. Further, as a computer-assisted multi-person communication tool, speech-guided human facial animation (*e.g.*, Apple Memoji) is becoming a useful interface for online chatting in remote collaborative circumstances.

In the literature, a rich variety of facial animation frameworks have been proposed. We can boardly categorize facial modeling and animation techniques into two classes: geometric manipulations-guided techniques and image manipulations-guided techniques. Geometric manipulations include the following techniques: key-framing and geometric interpolations [2], [3], parameterizations [4], finite element methods [5], modeling using facial muscless [6], pseudo-muscle-based facial animation [7], spline-guided approaches [8], and free-way deformations [9]. Comparatively, image manipulations denote techniques like image morphing between pairwise photographic images [10], texture manipulations [11], image blending [12], and vascular expressions [13], [14]. These video animation techniques are practically guided by tracking/localizing visual features or animation driven by performance [15]. In spite of the various aforementioned methods, there are still challenges to implement them into an emotional speech animation system satisfying real-world requirements:

- Many approaches need complicated human intervention in the model training stage. For example, the system designers have to determine the key frames

The associate editor coordinating the review of this manuscript and approving it for publication was Christian Pilato.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

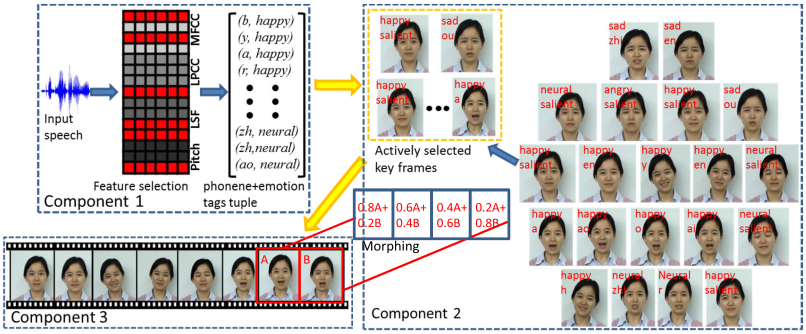


FIGURE 1. **Pipeline of the proposed speech-driven real human facial animation system.**

corresponding to each phoneme/emotion tag and how many key frames need to be employed. Such human intervention makes the accuracy of synthesized facial video intolerantly dependent on the domain knowledge of system designers. In practice, we expect a fully automatic training stage of the facial animation system, wherein no strong domain knowledge is needed.

- Owing to the popularity of portable devices like Apple Watch and Google Pixel, more and more communication Apps are developed on mobile platforms, *e.g.*, Skype and Facetime. This stimulates the demand of developing facial animation systems on mobile devices. However, due to the limited computational capability, it is difficult to develop a real-time mobile facial animation App. Besides, no optimization have been proposed to transform an off-the-shelf desktop animation system onto mobile platforms.
- Most of the previous facial animations are based on 2D/3D cartoon figures. Toward a more natural human-computer interface, animation based on real human faces is preferred. This is a challenging task because illumination, expression, and head position are difficult to control when synthesizing a real human face. This factors may lead to an unnatural face as shown in practical speech animation systems.

To tackle these difficulties, we design an emotional speech driven facial animation system that is trained in a fully automatic way. Moreover, the animation system is based on a real human face and can execute in real-time on mobile devices. The flowchart of our designed animation system in Fig. 1, which can be divided into three main components. **Part 1:** For each recorded human speech with emotion, the six well-known acoustic features [16] (*e.g.*, MFCC and LSF) are extracted in the first place. Thereafter, our phoneme and emotion pairs from an emotional speech can be rapidly and accurately calculated by a multi-label classifier. **Part 2:** To match one phoneme emotion pair with a selected representative faces, a low-rank active learning technique is leveraged to discover multiple key facial images from the recorded videos during training. In our implementation, these videos are recorded by a volunteer from our Department. She is a native Mandarin/English speaker. Herein, we divide each video into multiple sentences, each associated with an specific emotion (*i.e.*, ''happy'', ''surprise'', ''sad'', ''angry'', and ''neutral'') are used to speak the sentence. Our proposed low-rank active learning algorithm is effective since it exploits the underlying distribution of facial frames from a video. **Part 3:** After matching the key facial frame to each phoneme and emotion pair, toward a smooth synthesized video, the morphing [35] technique is adopted to produce a set of intermediate frames between key facial frames that are temporally adjacent. To make our synthesized facial expressions seemingly natural, illumination compensation is applied to each facial frame.

Totally, our work has the following advantages: 1) an intelligent platform for real human facial animation, which is trained with little human intervention; 2) leveraging an active learning paradigm for calculating key facial frames from multiple recorded training videos; and 3) our system is a general that can be trained from an arbitrary human face.

II. RELATED WORK

The proposed system is basically relevant to two topics in modern artificial intelligence systems: 1) recognizing emotion and phoneme using human speech, and 2) speech-driven facial animation technique.

A. EMOTION AND PHONEME RECOGNITION BY SPEECH

Identifying emotion and phoneme pairs based on human speech [16], [17] aims to understand human affective attributes of each utterance by analyzing the acoustic features engineered from human speeches. Practically, we can formulate this task as a speech clip categorization problem. To accurately and fast categorize different speeches into emotion and phoneme pairs, researchers proposed a couple of acoustic features. In the literature, machine learning researchers proposed probabilistic generative models, *e.g.*, Latent Dirichlet Allocation (LDA) and Long Short-Term Memory (LSTM), to exploit the underlying distribution of the aforementioned acoustic features. Afterward, they deployed the softmax layer or the maximum posterior probability estimation to recognize different emotion and phoneme pairs [18], [19]. Another line of research focused on deriving the so-called background models from the acoustic representations, based on which the supervectors are calculated for categorization [20], [21]. Such categorization pipeline has been pervasively utilized in domains like speakers localization. Some researchers designed statistical algorithms to learn the distribution of the acoustic representations. Herein, the globally calculated statistical distributions are leveraged to classify each emotion and phoneme pair. In practice, support vector machine is treated as the most popular tool for classify such global acoustic representations [22], [23]. Meanwhile, different classifiers, *e.g.*, random forest [24] and softmax [25], are also pervasively applied in speech-based emotion and phoneme understanding. Noticeably, however, the above methods largely rely on the possibly high-dimension and manually-designed acoustic features that are selected by some prior knowledge.

B. FACIAL ANIMATION VIDEO DRIVEN BY SPEECH

In the literature, the synthesization of an aesthetically pleasing facial video based on a the input human speech was investigated comprehensively. Herein, an extensive review of the previously published speech-guided facial video synthesization is provided in [26]. The authors [27] attempted to transform the two dimensional facial frames into a natural facial video by rebuilding the 3D facial frames using a morphing technique. Thereafter, they calculated a so-called expression+viseme feature space using the above synthesized 3D faces. The authors [28] proposed a speech-driven-lips framework that simultaneously constructs human speech co-articulation as well as the expression-guided eigenspaces. A rich set of other methods [29] were designed so as to produce expression-guided speech videos. In [30], the authors proposed to intelligently predict lip-based movement trajectory using human speech. The designed system accurately calculates human lip movements from the original human speech. Simultaneously, it can optimally produce video animation trajectory by leveraging the well-known HMM. The authors established a real-time framework for automatically generating speech-guided facial gestures in virtual contexts. More specifically, the method can produce gestures such as multiple nods/ head movements and eye blinks. The system is practically realized by incorporating HMM, multiple pre-defined crteria, as well as some statistic distributions. In conclusion, the above discussed facial animation pipelines are not particularly designed toward mobile platforms. Besides, to our best knowledge, only a few animation pipelines can synthesize real-world human faces. Even worse, they cannot rapidly reduce the sub-optimal illumination.

Besides, In [31], Yuan et al. presented crucial insights into active learning applied in a visual context, particularly in tracking applications. Based on active learning, the proposed CNN-guided visual tracker can be conveniently trained by leveraging a highly diverse set of training video frames. In [32], Ren et al. systematically summarized the existing deep active learning algorithms, associated with a comprehensive overview. They also presented the development of deep active learning in different vision applications.

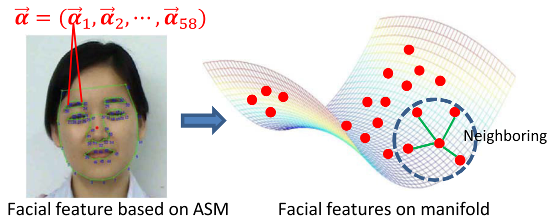

$\vec{\alpha} = (\vec{\alpha}_1, \vec{\alpha}_2, \cdots, \vec{\alpha}_{58})$
Neighboring
Facial feature based on ASM
Facial features on manifold

FIGURE 2. Left: the active shape model (ASM) model of a human face; Right: projecting ASM facial features from all the facial frames (red dots) onto manifold.

III. OUR APPROACH

A. ACOUSTIC FEATURES EXTRACTION

In our implementation, for a male/female speech set, the entire feature combination is constructed by multiple well-known acoustic feature dimensions, that is, pitch, log energy, 3 format frequencies, 11 MFCCs, 16 PLCCs, and 9 LSFs. We choose these acoustic features by cross validation. The above 41-dimensional features are utilized to train a multi-label classifier to classify each speech sentence into the corresponding phoneme and emotion pairs. Such pairs are utilized for synthesizing the speech-drive facial animation video subsequently.

B. ACTIVE LEARNING FOR KEY FACES SELECTION

In order to build an optimal facial animation framework, we typically record facial videos of a male/female speaking English or Chinese during the system training stage. It is observable that each video practically has large number of facial frames. Practically, it is non-trivial to detect facial frames which can best associate the phoneme and emotion pairs. Previous AI systems typically employ pre-specified key facial frames, which might be sub-optimal. Herein, we select the key faces by leveraging a novel active learning paradigm that are conducted in a completely automatic way. In our implementation, the speech videos are captured by a Mandrin speaker in a well-established recording studio. Totally, we obtain 105 recorded speech videos, each lasts about 420 seconds.

Theoretically, we treat active learning as a sample selection paradigm, wherein multiple criteria were developed to select highly representative sample. For our system, we discover multiple key facial frames based on the aforementioned recorded speech videos. Herein, the objective is that the discovered key facial frames are best representative to frames from the recorded speech videos.

Denote $\mathbf{A} = [\vec{\alpha}_1, \cdots, \vec{\alpha}_n] \in \mathbb{R}^{58 \times N}$ as a collection of facial video frames distributed on the underlying subspace. Herein, N counts the training video frames. The objective is to conduct subspace learning and active frames selection jointly. We denote $\mathbf{B} \in \mathbb{R}^{58 \times K}$ as the selected K representative frames.

In theory, we still adopt the strategy of minimizing the overall reconstruction loss in the original space to select the most representative samples. To this end, we take advantage of the following objective function:

$$\min_{\mathbf{R}\in\mathbb{R}^{N\times N}} ||\mathbf{A}-\mathbf{A}\mathbf{R}|| + \lambda||\mathbf{R}||_l, \tag{1}$$

Herein, $\lambda \geq 0$ measures the significance of our designed regularizer. For the above objective function, the left term attempts to maximally rebuild the input facial frames, wherein $\mathbf{R}$ is a matrix containing the rebuilding parameters. Meanwhile, Meanwhile, the right term represents a pre-defined regularizer with a particular matrix norm. Herein, the objective is to acquire the top K key facial frames, and thus the rebuilding terms toward the top K key frames should be heavily weighted. In contrast, the remaining NK unselected facial frames should be lightly weighted. Taking a very particular case as an example, when all elements of one row in $\mathbf{R}$ become zeros, that means these facial frames are not recognized as the key facial frames. This is because they are considered to have no contribution to rebuild the rest facial frames. In this way, $\mathbf{R}$ is a matrix that is sparse in row, as each row measures the importance of each facial frame in rebuilding the remaining ones.

Toward a row-wise sparse matrix $\mathbf{R}$, it is straightforward to upgrade term $||\mathbf{R}||_l$ into term $||\mathbf{R}|| + 2, 1$ or term $||\mathbf{R}||_\infty$. In our implementation, term $||\mathbf{R}||_{2,1}$ is deployed. In practice, we notice that term $||\mathbf{R}||_\infty$ is also an appropriate choice. In theory, $\mathbf{R}$ has two key contribution in the above objective function. i) a matrix containing the rebuilding parameters and each column functions as the linear combination of the key facial frames to rebuild a new one; and ii) a matrix for representing itself. That is, each column $\mathbf{r}_i \in \mathbb{R}^N$ is considered as a feature for representing $\vec{\alpha}_i$. Herein, we treat $\mathbf{A}$ as an unknown dictionary.

As we mentioned, the facial frames are practically distributed on the underlying subspace hidden in a high-order feature space. In this way, $\mathbf{R}$ is constrained to be a low-rank matrix, based on which the above objection function is updated as follows:

$$\min_{\mathbf{R}\in\mathbb{R}^{N\times N}} ||\mathbf{A}-\mathbf{A}\mathbf{R}|| + \lambda||\mathbf{R}||_{2,1} + \eta\,\mathrm{rank}(\mathbf{R}), \tag{2}$$

where $\eta \geq 0$ denotes a weight to the corresponding term. rank$(\cdot)$ calculate the matrix rank. We minimize term rank$(\mathbf{R})$ to achieve a low-rank matrix $\mathbf{R}$. Therefore, we can recover the low-rank geometry from the input matrix. Practically, we notice that the above objective function is NP-hard. Instead, we update rank$(\mathbf{R})$ to the well-known nuclear norm of matrix $\mathbf{R}$ [33]. This makes the problem a convex one, that is,

$$\min_{\mathbf{R}\in\mathbb{R}^{N\times N}} ||\mathbf{A}-\mathbf{A}\mathbf{R}|| + \lambda||\mathbf{R}||_{2,1} + \eta||\mathbf{R}||_*, \tag{3}$$

Herein, $||\mathbf{R}||_*$ denotes the nuclear norm implemented for the aforementioned rank function. Details of the solution is presented in [34]. By leveraging the calculated $\mathbf{R}$, we acquire K representative frames to represent each facial video.

C. ANIMATION VIDEO GENERATION BY MORPHING

For one second, we practically generate three phoneme and emotion pairs. The three pairs have three corresponding key frames accordingly. In practice, three frames for each second cannot ensure a smooth and natural synthesized facial video, *i.e.*, 24 frames for each second. Herein, the well-known morphing [35] algorithm is leveraged for calculating the intermediate faces for pairwise temporally adjacent key frames.

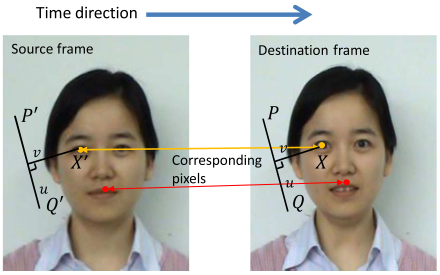


FIGURE 3. Coordinates mapping from the source image to the destination one.

Given two key facial frames as shown in Fig. 3, morphing combines them by cross-dissolving their corresponding image pixels (*e.g.*, pixels from the lips in the two key facial frames in Fig. 3). Before this, we have to locate the corresponding pixels between pairwise key facial frames. Given a pair of corresponding lines PQ and $P'Q'$ from the destination and the source frames respectively, a mapping can be derived from the coordinate of the destination frame pixel X to that of the source frame pixel X':

$$\vec{\mathbf{u}} = \frac{(X-P)\cdot(X-P)}{||Q-P||^2}, \tag{4}$$

$$\vec{\mathbf{v}} = \frac{(X-P)\cdot Pen(X-P)}{||Q-P||}, \tag{5}$$

$$X' = \vec{\mathbf{v}} \cdot \frac{Pen(Q'-P')}{||Q'-P'||} + \vec{\mathbf{u}}(Q'-P') + P' \tag{6}$$

Herein, $Pen(\cdot)$ returns the vector that is perpendicular to, as well as the same length to the input vector. $\vec{\mathbf{u}}$ is the direction along the line PQ or $P'Q'$. $\vec{\mathbf{v}}$ calculates the distance between X (a pixel) and PQ (a line) (or the distance from X' to $P'Q'$).

Denoting O and O' as the origins of the destination and the source frames respectively, we can obtain $X = O + dX$. By putting (4) and (5) into (6), we obtain:

$$\begin{aligned} =\!\!\!\!X'O' &+ \frac{dX\cdot(Q-P)}{||Q-P||^2}\cdot(Q'-P') \\ &+ \frac{Pen(Q-P)\cdot Pen(Q'-P')\cdot dX}{||Q-P||\cdot||Q'-P'||}. \end{aligned} \tag{7}$$

Based on the above derivation, given a destination frame, we start from its origin O and map each of its pixel coordinates to that of the source frame. Two directions of increments are used: $dX_1 = (1, 0)$ and $dX_2 = (0, 1)$.

By locating the pixels in the destination key facial frame to those in the source one, we use cross-dissolve to obtain each intermediate facial frame. Denote $g_1(x_1, y_1)$ and $g_2(x_2, y_2)$ as the RGB values of the corresponding pixels (x_1, y_1) and (x_2, y_2) in the source and the destination frames respectively,

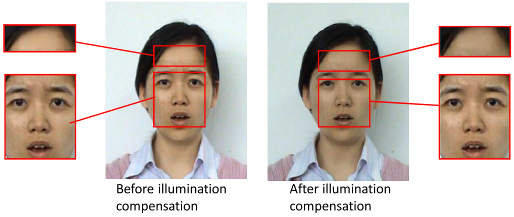

Before illumination compensation
After illumination compensation

FIGURE 4. An example of illumination compensation for the intermediate faces.

the RGB value of a pixel in the intermediate frame is:

$$g(x, y) = k \cdot g_1(x_1, y_1) + (1 - k) \cdot g_2(x_2, y_2), \tag{8}$$

where $k \in [0, 1]$ is the interpolation coefficient. We set $k = 0.4$ according to our implementation.

For our built animation system, we notice that our synthesized facial skins might be visually inconsistent. The inherent reason is the illumination discrepancy from the original and target human faces. Practically, to tackle such shortcoming, we adopt a lighting compensation scheme during our pixel cross-dissolve stage, *i.e.*,

$$g(x, y) = \frac{g_2(x, y)}{\eta \cdot (g_1(x, y) - \bar{g}_1) + \bar{g}_2}, \tag{9}$$

where $\eta = \sigma(g_2)/\sigma(g_1)$, $\sigma(\cdot)$ is the variance of the RGB color in a frame, and $\bar{g}$ is the average RGB color of a frame. As shown in Fig. 4, the illumination compensation scheme makes the facial skin in the animation video more consistent.

IV. EXPERIMENTS

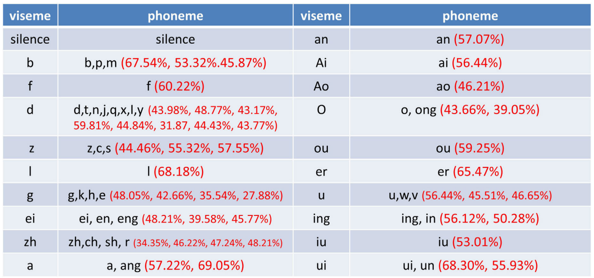
In this section, we test our designed animation system using three empirical validations. The first set of experiments step-by-step evaluates the important modules in our animation system. The second set of experiments evaluates the performance our system under different parameter settings. The third set of experiments visualizes the synthesized animation video and some intermediate results.

Our facial animation system for testing is briefed as follows. During the training stage, we collected 5,600 English speech sentences. These sentences are recorded by five males and three females, whom are from our Computer Sciences Department. Each sentence lasts approximately 42 ∼ 550 seconds. To accurately describe each speech sentence, 41 well-known acoustic features are calculated. To label the emotion of each speech sentence, we employ five different emotion labels (“anger”, “happiness”, “neutrality”, “sadness”, and “surprise”) and the pre-defined 44 phonemes (as shown in Fig. 5). To refine the speech sentences, a pre-emphasizing stage is deployed, including blocking and Hamming windowing.

A. IMPORTANT MODULES EVALUATION

1) LOW-RANK ACTIVE KEY FACIAL FRAMES DISCOVERY

Here, our adopted key facial frames selection algorithm is compared with multiple well-known frame selection algorithms, that is, online clustering key frames extraction (OCFE) [36] using the same ASM [37]-based facial feature as ours, dictionary selection based key frame selection (DSVS) [38], and motion-based key frame extraction (MKFE) [39]. OCFE first leverages a clustering algorithm to categorize the frames to different centers. Thereby, the remaining frames are progressively integrated to cluster. DSVS formulates video frame selection as a dictionary selection by seeking sparsity. A key-frame-based dictionary is calculated, wherein the training facial videos can best rebuild the calculated dictionary. For MKFE, we predict camera as well as object motion features for extracting the descriptors. Each video is subsequently decomposed into multiple clips based on different motion types. Accordingly, multiple rules are leveraged for calculating the key frames.



viseme	phoneme	viseme	phoneme
silence	silence	an	an (57.07%)
b	b,p,m (67.54%, 53.32%.45.87%)	Ai	ai (56.44%)
f	f (60.22%)	Ao	ao (46.21%)
d	d,t,n,j,q,x,l,y (43.98%, 48.77%, 43.17%, 59.81%, 44.84%, 31.87, 44.43%, 43.77%)	O	o, ong (43.66%, 39.05%)
z	z,c,s (44.46%, 55.32%, 57.55%)	ou	ou (59.25%)
l	l (68.18%)	er	er (65.47%)
g	g,k,h,e (48.05%, 42.66%, 35.54%, 27.88%)	u	u,w,v (56.44%, 45.51%, 46.65%)
ei	ei, en, eng (48.21%, 39.58%, 45.77%)	ing	ing, in (56.12%, 50.28%)
zh	zh,ch, sh, r (34.35%, 46.22%, 47.24%, 48.21%)	iu	iu (53.01%)
a	a, ang (57.22%, 69.05%)	ui	ui, un (68.30%, 55.93%)

FIGURE 5. The viseme-phoneme for Chinese pronunciation (the red text denotes the recognition accuracy of the phonemes).

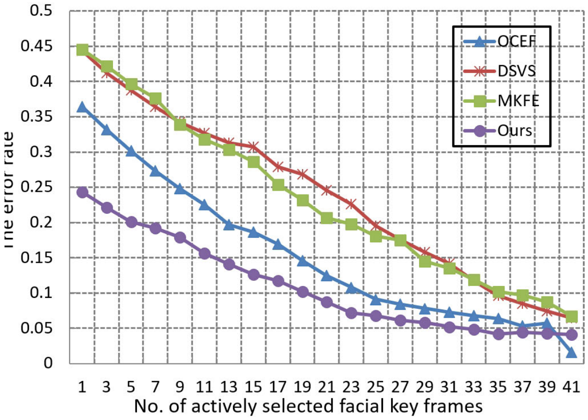


FIGURE 6. Reconstructing accuracy by leveraging different techniques as aforementioned (PM means the proposed method).

The key frames of the training facial videos are calculated in the first place. As shown in Fig. 6, the accuracy of key frames generated by different algorithms are reported. The accuracies indicate how the key frames can rebuild the entire facial frames during training. A high reconstruction accuracy means that key frames produced the method can optimally capture the training facial videos. Meanwhile, for each counterpart, we notice that some key frames capture each face with highly similar viseme and expression pairs. This observation is different from the principle that key facial frames must be evenly distributed and can effectively capture the facial videos.

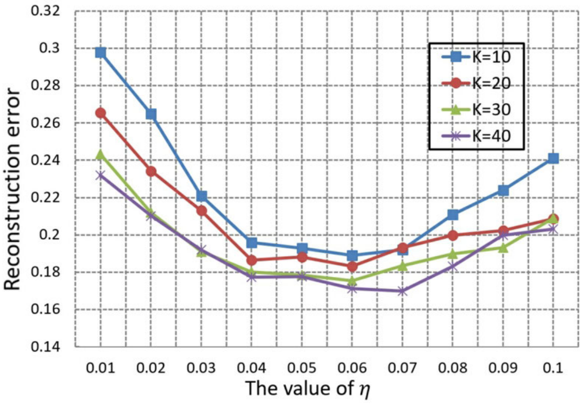


FIGURE 7. Key facial frames reconstruction error by leveraging different values of η.

2) MORPHING-BASED FACIAL ANIMATION VIDEO

In this subsection, our designed animation system is compared with the facial systems proposed by Deng et al. [28], Kshirsagar et al. [29], Hofer et al. [30], and Zoric et al. [40] respectively. Noticeably, either accuracy or ranking is an optimal choice for this task. The reason is that these methods are practically highly complicated for each observer to provide. In our implementation, we leverage the well-known paired comparison for user study. We use it to test the effectiveness of the proposed facial animation system. Paired comparison means, we present pairwise videos produced by two different facial animation systems to each subject, with the same input speech sentence. We preserve the above testing results in the so-called preference matrix. As displayed in Fig. 6, the element in row "Hofer" and the column "Zoric" is 12. This indicates that 12 subjects prefer the video produced from Hofer et al. than that produced by Zoric et al..

B. SYSTEM PERFORMANCE UNDER DIFFERENT PARAMETERS

This subsection evaluates the performance of our system under different parameter settings, that is, the parameter μ in the active key facial selection.

We evaluate the reconstruction error under different values of η in (2). We set the number of selected key facial frames K to 10, 20, 30, and 40 respectively. Then, we tune η from 0.01 to 0.1 with a step of 0.01. As shown in Fig. 7, the reconstruction error is minimal when $\eta = 0.05$. This is because η reflects the importance of preserving the distribution of the facial frames in the training videos. Emphasizing too much on this property will increase the reconstruction error.

C. VISUALIZATION OF THE FACIAL ANIMATION RESULTS

In this subsection, we visualize the intermediate results of our facial animation system. First, we show the facial features extracted by the ASM [37] model in Fig. 8. We deliberately use left oriented faces and each face is not in the middle of the video. As can be seen, the ASM model can accurately locate the faces. Then, we present the intermediate faces generated by the morphing technique in our proposed system. As shown in Fig. 9, the leftmost and the rightmost facial frames are the original frames while the rest frames are generated by morphing. It is observed that these generated faces look very natural and quite real to human faces.

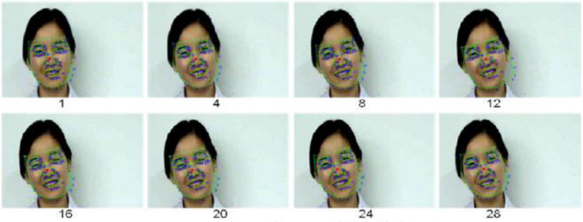


FIGURE 8. Human faces detected by the ASM model in the training facial videos.

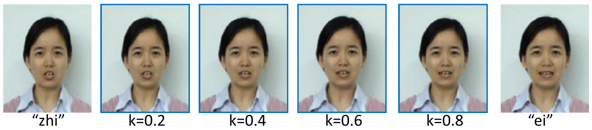


FIGURE 9. The intermediate faces (blue rectangles) generated by the morphing technique.

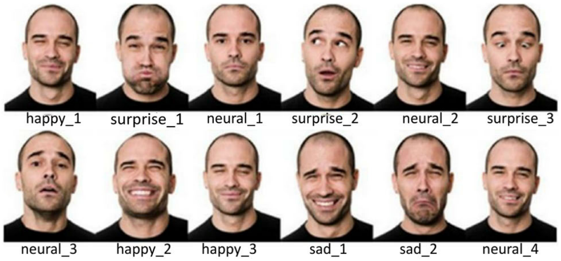


FIGURE 10. An exmaple of a male-face-based video animation framework.

V. CONCLUSION

In this work, we design a novel AI system to synthesize aesthetically pleasing facial videos by leveraging human speech sentences. More specifically, high quality acoustic features for recognizing phoneme and emotion pairs are identified using a multi-label SVM classifier. Afterward, we leverage a novel low-rank active learning algorithm to recognize the key faces from the large-scale training facial videos. By associating each emotion and phoneme pair with a key face, the well-known morphing algorithm fits the key frames into a smooth and natural synthesized facial video. Empirical results have shown that our method is efficient and effective. And it is learned in a completely automatic way.

REFERENCES

[1] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "Transductive face sketch-photo synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 9, pp. 1364–1376, Sep. 2013.

[2] A. Emmett, "Digital portfolio: Tony de peltrie," *Comput. Graph. World*, vol. 8, no. 10, pp. 72–77, 1985.

[3] F. I. Parke, "Techniques of facial animation," in *Model and Technique in Computer Animation*. Hoboken, NJ, USA: Wiley, ch. 16, 2019, pp. 229–241.

[4] M. Cohen and D. W. Massara, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*. Tokyo, Japan: Springer, 1993.

[5] B. Guenter, "A system for simulating human facial expression," in *State-of-the-Art in Computer Animation*. Tokyo, Japan: Springer, 1992.

[6] C. R. E. Pelachaud, "Communication and co-articulation in facial animation," Ph.D. thesis, Dept. Comput. Sci. Inf. Sci., School Eng. Appl. Sci., Univ. Pennsylvania, Philadelphia, PA, USA, Oct. 1992.

[7] P. Kalra, A. Mangili, N. M. Thalmann, and D. Thalmann, "Simulation of facial muscle actions based on rational free form deformations," *Comput. Graph. Forum*, vol. 11, no. 3, pp. 59–69, May 1992.

[8] M. Nahas, H. Huitric, M. Rioux, and J. Domey, "Facial image synthesis using skin texture recording," *Vis. Comput.*, vol. 6, no. 6, pp. 337–343, Nov. 1990.

[9] S. Coquillart, "Extended free-form deformation: A sculpturing tool for 3D geometric modeling," *ACM SIGGRAPH Comput. Graph.*, vol. 24, no. 4, pp. 187–196, Sep. 1990.

[10] T. Beier and S. Neely, "Feature-based image metamorphosi," *Comput. Graph.*, vol. 26, no. 2, pp. 35–42, 1992.

[11] M. Oka, K. Tsutsui, A. Ohba, Y. Kurauchi, and T. Tago, "Real-time manipulation of texture-mapped surfaces," in *Proc. 14th Annu. Conf. Comput. Graph. Interact. Techn.*, Aug. 1987, pp. 181–188.

[12] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin, "Synthesizing realistic facial expressions from photographs," in *Proc. ACM SIGGRAPH*, 1998, pp. 75–84.

[13] P. Kalra and N. Magnenat-Thanmann, "Modeling of vascular expressions in facial animation," in *Proc. Comput. Animation*, 1994, pp. 50–58.

[14] T. Darrell and A. Pentland, "Space–time gestures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 1993, pp. 335–340.

[15] P. Bergeron and P. Lachapelle, "Controlling facial expressions and body movements in the computer generated animated short," in *Proc. SIGGRAPH Adv. Comput. Animation Seminar Notes*, Jul. 1985, pp. 1–5.

[16] L. Zhang, M. Song, N. Li, J. Bu, and C. Chen, "Feature selection for fast speech emotion recognition," in *Proc. 17th ACM Int. Conf. Multimedia*, Oct. 2009, pp. 753–756.

[17] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural netwoks," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[18] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Proceedings. (ICASSP)*, Apr. 2003, pp. 33–43.

[19] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *Proc. Interspeech*, Oct. 2004, pp. 77–87.

[20] H. Hu, M.-X. Xu, and W. Wu, "GMM supervector based SVM with spectral features for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2007, pp. 1–8.

[21] T. L. Nwe, N. T. Hieu, and D. K. Limbu, "Bhattacharyya distance based emotional dissimilarity measure for emotion classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7512–7516.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "OpenEAR—Introducing the Munich open-source emotion and affect recognition toolkit," in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops*, Sep. 2009, pp. 1–6.

[23] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 5, pp. 1057–1070, Jul. 2011.

[24] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," in *Proc. Interspeech*, Sep. 2009, pp. 22–27.

[25] Y. Kim and E. M. Provost, "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3677–3681.

[26] Z. Deng and J. Noh, *Data-Driven 3D Facial Animation*. London, U.K.: Springer, 2008.

[27] V. Blanz, C. Basso, T. Poggio, and T. Vetter, "Reanimating faces in images and video," *Comput. Graph. Forum*, vol. 22, no. 3, pp. 641–650, Sep. 2003.

[28] Z. Deng, U. Neumann, J. P. Lewis, T.-Y. Kim, M. Bulut, and S. Narayanan, "Expressive facial animation synthesis by learning speech coarticulation and expression spaces," *IEEE Trans. Vis. Comput. Graphics*, vol. 12, no. 6, pp. 1523–1534, Nov. 2006.

[29] S. Kshirsagar, T. Molet, and N. Magnenat-Thalmann, "Principal components of expressive speech animation," in *Proc. Comput. Graph. Int.*, 2001, pp. 38–46.

[30] G. Hofer, J. Yamagishi, and H. Shimodaira, "Speech-driven lip motion generation with a trajectory HMM," in *Proc. Interspeech*, Sep. 2008, pp. 2314–2317.

[31] D. Yuan, X. Chang, Q. Liu, Y. Yang, D. Wang, M. Shu, Z. He, and G. Shi, "Active learning for deep visual tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 4, no. 1, pp. 189–196, 2023.

[32] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, "A survey of deep active learning," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 180:1–180:40, 2022.

[33] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, Jan. 2010.

[34] C. Li, K. Mao, L. Liang, D. Ren, W. Zhang, Y. Yuan, and G. Wang, "Unsupervised active learning via subspace learning," in *Proc. AAAI*, 2021, pp. 154–162.

[35] T. Beier and S. Neely, "Feature-based image metamorphosis," in *Proc. 19th Annu. Conf. Comput. Graph. Interact. Techn.*, Jul. 1992, pp. 35–42.

[36] A. Bouguettaya, "On-line clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 8, no. 2, pp. 333–339, Apr. 1996.

[37] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, Jan. 1995.

[38] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 66–75, Feb. 2012.

[39] J. Luo, C. Papin, and K. Costello, "Towards extracting semantically meaningful key frames from personal video clips: From humans to computers," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 2, pp. 289–301, Feb. 2009.

[40] G. Zoric, R. Forchheimer, and I. S. Pandzic, "On creating multimodal virtual humans—Real time speech driven facial gesturing," *Multimedia Tools Appl.*, vol. 54, no. 1, pp. 165–179, Aug. 2011.

HUI XU is currently a Faculty Member with the School of Electronic and Information Engineering, Wuhan Technical College of Communications, Wuhan, China. His research interests include multimedia and image processing.

XIAOYANG YU is currently a Lecturer with the School of Transportation Information, Hubei Communications Technical College, Wuhan, China. His research interests include AI and NLP.

YU CHENG is currently a Researcher with Wuhan Intelligent Connectivity Technologies Company Ltd., Wuhan, China. His research interest includes computer vision.

MENGQIONG XIAO is currently a Faculty Member with China University of Geosciences, Wuhan, China. His research interests include image processing and language processing.

YUE YU is currently a Professor with the College of Computer Sciences, Anhui University, China. His research interests include computer vision and image processing.