DOI: 10.3979/j.issn.1673-825X.2020.05.020

# 基于语音与人脸参数化表示的跨模态稠密深度网络学习方法

唐　俊[1],牟海明[1],冷　洁[1],李清都[2],刘　娜[1]
(1.上海理工大学 机器智能研究院,上海 200093;2.重庆邮电大学 自动化学院,重庆 400065)

**摘　要:** 为了提高跨模态人脸表示与合成的性能,针对语音与人脸图像 2 种模态数据,提出一种基于人脸参数化表示与稠密深度网络相结合的面部生成方法。针对输入语音模态,通过对信号进行频谱变换,将一维时域信号转换到二维频率域,可提取频域上稳健的特征描述;针对输出图像模态,利用主动外观模型对不同面部区域独立建模以降低区域间的相关性,并提取紧凑的人脸参数化特征;为了获得有效的跨模态学习性能,提出采用稠密连接的深度卷积神经网络学习语音、图像 2 种模态的回归预测,并通过预测的人脸参数进行面部重构,所采用的深度网络模型可以加强特征传播与特征复用,有利于增强面部细节的合成。在 2 组音视频数据集上验证了提出方法的有效性。

**关键词:** 跨模态学习;深度学习;卷积神经网络;参数化表示;语音;图像
**中图分类号:** TP391　　**文献标志码:** A　　**文章编号:** 1673-825X(2020)05-0867-07

# Cross-modal learning based on speech and parameterized face representation using densely deep networks

TANG Jun[1], MOU Haiming[1], LENG Jie[1], LI Qingdu[2], LIU Na[1]
(1. University of Shanghai for Science and Technology, Institute of Machine Intelligence, Shanghai, 200093, P. R. China;
2. Chongqing University of Posts and Telecommunications, Institute of Automation, Chongqing, 400065, P. R. China)

**Abstract:** To improve the performance of cross-modal face representation and synthesis, a facial synthesis method is proposed based on two modalities of speech and face image using densely deep network and cross-modal learning. First, frequency domain feature description is obtained by performing spectral transformation on the speech modal, which transforms one-dimensional time domain signal into two-dimensional frequency domain. Secondly, the active appearance model is applied to different facial regions to reduce the region correlations of the output image modal; the compact face parameterized features can then be extracted. Finally, in order to obtain effective cross-modal performance, a densely connected deep convolutional neural network is proposed to learn the regression prediction between speech and image modalities, followed by face reconstruction on the predicted parameters. The proposed deep learning model helps enhance feature communication and feature reuse, which is conducive to enhance the synthesis of facial details. Experiments on two audio and video datasets demonstrate the effectiveness of the proposed method.

**Keywords:** cross-modal learning; deep leaning; convolutional neural network; parameters description; speech; image

## 0　引　言

语音驱动的人脸合成涉及语音、图像以及文本分析领域内的相关重要技术,在生活中有着诸多应用场景,包括智能家电、声控游戏、智能穿戴设备、虚拟现实等,在人机交互的发展过程中占据着非常重



**收稿日期:** 2020-08-02　**修订日期:** 2020-09-18　**通讯作者:** 李清都　liqd@cqupt.edu.cn

要的地位。基于语音的人脸合成不仅需要对说话人识别,还需要在合成人脸的同时,保持语音、唇形、面部表情等信息的同步,其中,最关键的步骤在于人脸合成时唇形和语音之间的同步。如何对语音进行准确的唇形预测并合成逼真的连续人脸图像序列,是图像合成技术中的一个热门方向,也是本文重点讨论的问题。

目前常用的人脸合成方法包括基于视频的方法[1-3]和基于模型的方法[4]两大类:①利用现有数据集中的人脸图像生成视频人脸的方式进行人脸合成,该方法的优势是包括更多的纹理特性,而局限性在于人脸图像仅限于数据集中的人脸图像,数据的多样性和丰富性较少;②通过人脸形变模型对图像进行处理以扩充数据的多样性。主动外观模型(active appearance models, AAM)[5]是目前常用的人脸合成模型,该方法通过对图像中的形状和纹理信息进行统计,包括对数据进行建模和模型匹配2个步骤。

人脸合成的关键是如何实现唇形同步,即语音和人脸图像的同步,而唇形同步中的难点是解决协同发音问题,也就是使得每句话中每个音素同步。由于协同发音过程中相邻音素会对某个音素的发音产生影响,因此,相互之间产生了某种依赖关系。目前,处理该问题的方法包括回归方法和分类方法2种。回归方法可以将输入特征映射到连续轨迹[6-7],这有助于在连续帧之间获得平滑的唇形。分类方法是获取唇形和音素之间的特定映射关系,我们把它称之为视位[8]。前期的分类工作[9]对人脸图像进行了变形来解决唇形同步的问题。而多数分类方法[8,10-11]获取数据集上的视位映射是通过统计学习的方式来实现的。

当前,多数关于人脸合成的研究都是在可控环境下进行的,包括人脸的姿态和光照[1,12]。同时,测试对象是以中性或其他情感方式来表达固定句子[4]。为了获得训练数据,需要对视频进行目标分割[13-14]以获取感兴趣区域。在早期工作中,Bregler等[15]通过跟踪、音素检测和唇形合成等方法来生成人的嘴唇运动进而匹配音素序列。随后的研究多数也是基于可控环境下进行的[16-17]。最近,Taylor等[7]利用全连接网络生成自然流畅的语音动画,这是一种基于可控数据集的语音合成模型。为了训练网络模型,他们使用单个演员的音频与视频同步的数据集,其中包含2 543个中性句子,按照每秒29.97帧采集人脸正面视频,并对数据集中所有句子进行手动标注。除上述可控环境外,Suwajanakorn等[6]提出了一种递归神经网络,在不受控制的环境下合成奥巴马的视频。通过对每周数小时的奥巴马演讲录像进行学习,模型可以将原始的音频特征合成至唇形,从而生成具备精确口型同步的高质量演讲视频。

本文提出一种基于序列到序列的跨模态回归学习模型。模型包含3个主要内容,输入语音的特征预处理、人脸图像的参数化表示以及从语音特征到人脸参数化表示的跨模态深度学习,所提出的方法可以获取时态上的局部上下文信息,同时忽略输入的长距离依赖性。

## 1 本文方法

给定音视频同步的语音与图像数据,提出方法的主要目的是学习从语音到人脸图像的预测。模型学习之前,需对2种不同模态的数据进行处理,分别是对输入语音的特征提取以及对输出人脸图像的参数化表示。在2种模态基础上,进一步学习跨模态的深度网络模型。本文所提出方法的处理流程如图1。


图1 模型框架图

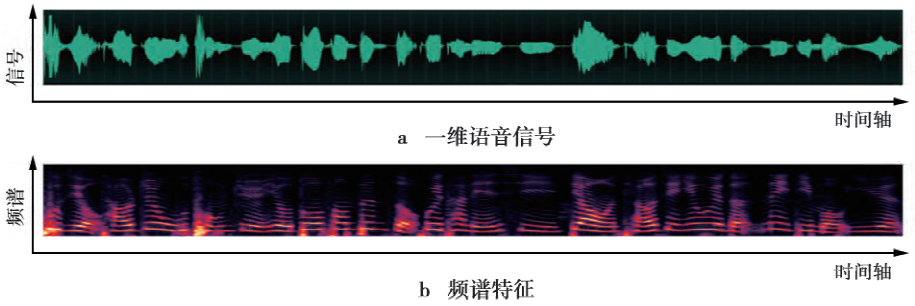
Fig.1 Framework of the model

### 1.1 语音特征提取

语音信号的预处理主要是将原始音频信号转换到频谱空间,然后通过梅尔频谱系数将频域内的信号进行转换,进而获得目标域的特征,如图2。总体上,语音信号的预处理大致包括6个步骤:信号预加重处理、语音分帧、帧加窗、离散傅里叶变换、三角带通滤波器组、计算对数能量。

1)信号预加重处理。消除在发声过程中因声带和唇形产生的效应,补偿该过程中引发的高频信号抑制。通过高通滤波器对抑制掉的高频部分进行补偿,增强高频对应的共振峰,以保证语音信号整体的平稳性。

2)语音分帧。由于语音信号具有短时稳定性,所以对语音信号进行采样时,选取语音帧长为20~30 ms。同时,为了减少帧之间的变化,确保采样时相邻帧之间有重叠交叉区域,从而获取到平稳的语音信号。




a　一维语音信号

b　频谱特征

**图 2　语音特征提取**

Fig.2　Speech feature extraction

3) 帧加窗。为了保证信号的全局连续性、减少语音信号产生的吉布斯效应,采用汉明窗函数对语音信号进行处理。

4) 离散傅里叶变换。在语音信号处理中常用的信号分析方法是离散傅里叶变换,它是将语音信号从时域变换至频域,相应的描述为

$$x(q)=\sum_{m=0}^{M-1}x[m]\,\mathrm{e}^{-\mathrm{j}(2\pi/M)qm},0\leqslant q\leqslant M \tag{1}$$

(1) 式中: $M$ 为采样点个数; $x(q)$ 为采样得到的语音信号。

5) 三角带通滤波器组: 将傅里叶变换求得的功率谱通过 $Q$ 个三角滤波器的滤波器组,以消除谐波影响,得到更为平滑的功率谱。

6) 计算每个带通滤波器的对数能量输出,表示为

$$s(m)=\ln\left(\sum_{i=0}^{N-1}|X_a(i)|^2H_m(i)\right),0\leqslant m\leqslant Q \tag{2}$$

(2) 式中,$H_m(i)$ 是第 $m$ 个三角带通滤波器,频谱特征矩阵的可视化见图 2b,其中的行表示频率分布,列表示时间轴。

### 1.2　人脸参数化表示

采用 AAM 对人脸图像进行参数化表示,如图 3。AAM 的形状由面部特征点决定,关于面部特征点的提取,采用开源框架 Dlib [18] 实现,面部特征点的表达为

$$\boldsymbol{s}=\{x_1,y_1,x_2,y_2,\cdots,x_N,y_N\}^{\mathrm{T}} \tag{3}$$

给定人脸图像序列,根据面部特征点的定位结果,利用普氏分析将人脸图像序列进行对齐。对齐的目的是将所有人脸图像校正到同一姿态,以保证对应的面部特征点处于同一空间位置。校正完成后,可通过主成分分析对人脸图像进行参数化表示为

$$\boldsymbol{s}=\boldsymbol{s}_0+\sum_{i=1}^{m}\boldsymbol{s}_ip_i \tag{4}$$

(4) 式中: $\boldsymbol{s}_0$ 是平均形状; $\boldsymbol{s}_i$ 是形状基础向量; 系数 $p_i$ 是形状参数。

采用 PCA 对面部图像的外观进行参数化表示为

$$\boldsymbol{A}(X)=\boldsymbol{A}_0(X)+\sum_{i=1}^{n}\boldsymbol{\lambda}_i\boldsymbol{A}_i(X),\forall X\in s_0 \tag{5}$$

(5) 式中: 系数 $\boldsymbol{\lambda}_i$ 是外观参数; $\boldsymbol{A}_0(X)$ 是平均外观,而 $\boldsymbol{A}_i(X)$ 是外观基础向量。

与传统对整幅人脸图像进行 AAM 表示的方法相比,图 3 展示的本文方法利用面部特征点的空间位置分布,对人脸图像进行多个区域的划分,主要将人脸图像划分成 3 个部分,分别是左上、右上、底部 3 个部分的面部区域,然后,每个面部区域分别进行 *AAM* 建模表示,描述如下

$$\boldsymbol{c}=\begin{pmatrix}\boldsymbol{p}\\ \boldsymbol{\lambda}_1\\ \boldsymbol{\lambda}_2\\ \boldsymbol{\lambda}_3\end{pmatrix} \tag{6}$$

(6) 式中: $p$ 是形状参数的向量; $\boldsymbol{\lambda}_1$、$\boldsymbol{\lambda}_2$ 和 $\boldsymbol{\lambda}_3$ 是 3 个人脸区域的外观参数向量。各个外观和形状模型的维数分别为 $n_1$,$n_2$,$n_3$ 和 $m$。$\boldsymbol{c}$ 可以紧凑地表达对话过程中面部形状和外观的变化。各个面部区域单独参数化表示的目的是为了更加有效地进行唇形建模,这是因为: ①与人脸其他区域相比,嘴唇在说话过程中的变化相对剧烈,需要更加细化的建模描述; ②各个人脸区域独立建模有助于降低面部区域之间的相关性。

### 1.3　稠密连接的深度卷积神经网络

自 2016 年何恺明 [19] 提出的残差网络(ResNet)荣获 ImageNet 图像识别大赛第一名以来,使得层数越来越深的网络得到了迅猛发展。本文以稠密连接

的深度卷积神经网络 DenseNet[20] 为基本框架,主要改进在于:①原始 DenseNet 的学习对象为单一模态数据(输入图像,输出类别标签),本文则研究网络在跨模态学习(输入语音模态,输出人脸模态)上的有效性;②原始 DenseNet 是分类模型,即预测输入样本的类别标签,本文将网络作为回归问题进行求解,模型输出人脸表示系数;③对原始 DenseNet 网络结构进行精简,以适应在数据量较少情况下的模型学习,本文所采用的网络结构如图 4。此外,与其他网络(如 ResNet)相比,稠密连接的网络结构可以学习到更多的面部细节特征。

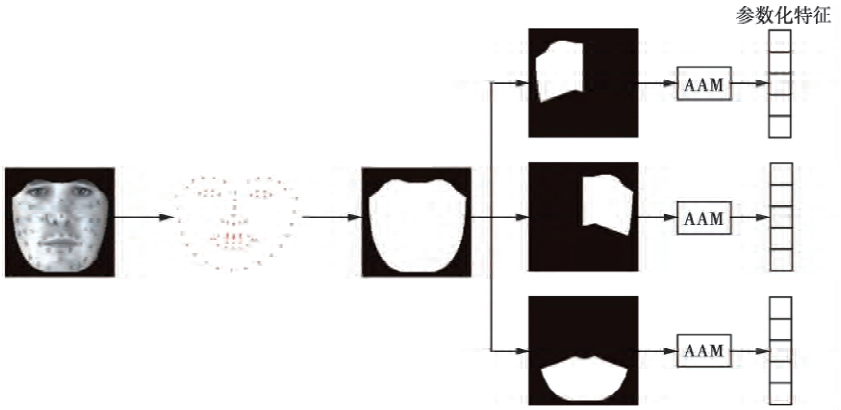



图 3 人脸参数化表示(从左至右:原始人脸,特征点定位,面部轮廓,面部区域,主动外观模型,参数化特征)

Fig.3 Face parameterized representation (from left to right: original face, feature point location, facial contour, facial area, active appearance model, parameterized features)

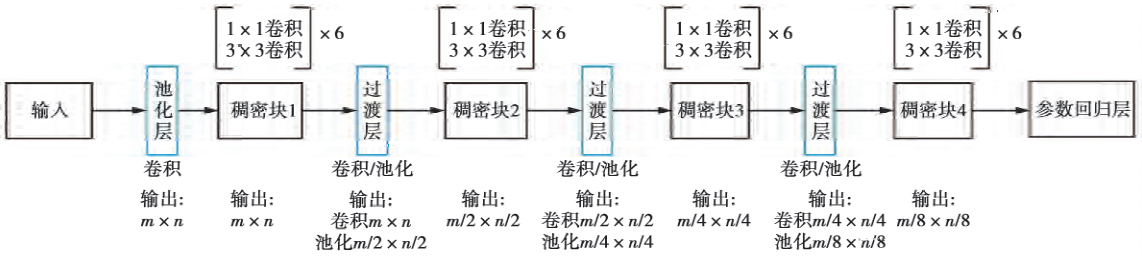



图 4 DenseNet 框架

Fig.4 Framework of DenseNet

本文采用的 DenseNet 结构包含 2 个部分:①稠密块:内部采用稠密连接的方式,定义了网络内部各层间的连接关系。各层均与前面层的输出有映射关系,多个稠密块的选取同时保证特征图和稠密块之间的大小一致性;②过渡层:2 个相邻稠密块之间的连接定义为过渡层,通过该层可以有效的控制网络内部的通道数量,使得网络结构更为紧凑。

稠密连接指的是各层与前层之间均有连接,网络内的任意 2 层之间都有连接存在。这样能够保证网络层数加深时的准确性和高效性。例如,对于传统 $N$ 层的卷积网络,网络的连接数为 $N$,而 DenseNet 网络的连接数则为 $\frac{N(N-1)}{2}$。

残差网络的输出为

$$x_n = F_n(x_{n-1}) + x_{n-1} \tag{7}$$

DenseNet 的输出为

$$x_n = F_n([x_0, x_1, \cdots, x_{n-1}]) \tag{8}$$

(7)—(8)式中,$F(\cdot)$ 为非线性变换函数;下标 $n$ 代表所在网络层数,共包含 $N$ 层;$x_n$ 代表第 $n$ 层的输出。

DenseNet 通过稠密连接的方式有效地缓解了更深层网络的梯度爆炸,降低了网络参数数量和过拟合情况,进而减少了网络的计算量、提高了特征重用性,可以获得理想的回归预测效果。

## 2 实验分析

### 2.1 数据集

实验采用开源的音视频数据集,第 1 组是英文数据集 SAVEE(surrey audio-visual expressed emotion)[21],该数据集包含 4 位说话者,每位说话者包

含 480 个英语句子、对应的文本内容、高质量音频和视频( 60 帧/秒) 数据。第 2 组是中文音视频数据集,考虑到不同语言发音的人脸合成验证,我们从央视新闻联播( CCTV) 收集了一位主播的音视频数据( 25 帧/秒) ,共包含 10 万帧人脸图像以及对应的中文语音。

### 2.2　跨模态学习的收敛性

从语音特征到人脸参数的跨模态学习过程中,DenseNet 模型的输入为语音特征图,具体来说,采用滑动窗口的方式对图 2 的频谱特征图截取固定大小的特征窗口,将二维矩阵形式的特征窗口作为网络输入,预测每个语音片段的网络输出。网络输出为人脸参数化向量,即图 3 中获取到的 AAM 参数向量( 3 个人脸区域所获得向量的拼接) 。

2 个数据集( 2 种不同语言) 在训练过程中的模型收敛效果如图 5。图 5 中横坐标表示模型学习的迭代次数,纵坐标表示迭代过程中的损失函数值。红色、蓝色曲线分别表示 SAVEE,CCTV 数据集的跨模态学习收敛效果。可以看出,随着迭代次数的增加,损失函数的值不断递减,大约迭代至 250 000 次时,损失值接近最小值。总体上,在 2 组音视频同步的多模态数据上,本文采用的深度网络均可以收敛至比较理想的状态。

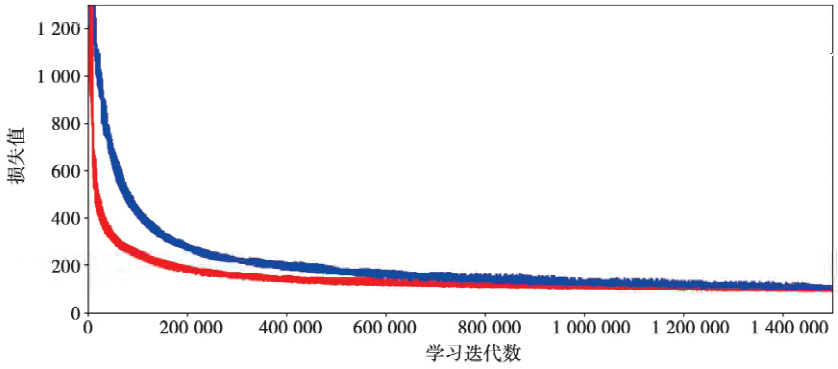



图 5　从语音到图像的跨模态学习收敛过程

Fig.5　Convergence process of cross-modal learning from speech to image

### 2.3　与其它深度模型的对比

实验对比了其他主流的深度网络模型,包括 Suwajanakorn 等使用的长短期记忆网络( long short-term memory, LSTM) [6] 模型、Taylor 等使用的全连接神经网络( fully-connected network, FCN) [7] 、以及当前流行的深度卷积神经网络框架( 如 VGG16 [22] , GoogLeNet [23] , ResNet [19] 、SENet [24] ) 。实验采用的 LSTM 模型包括隐藏层、全连接层和 LSTM 层; 全连接网络包含输入层、输出层,隐藏层均为全连接层,激活函数选用的是修正线性单元函数。关于其他网络,采用了对应文献的参考模型框架。

实验中,首先从数据集中随机选择一部分用于模型训练,剩余部分作为模型测试。实验采用均方根误差来比较上述方法与本文方法在人脸参数预测上的性能差异,均方根误差常用于衡量真实值和预测值之间的精度偏差,是机器学习当中常用的衡量指标。表 1 展示了各种方法的均方根误差值,均方根误差越大,人脸的重构效果越差。可以看出,相比于其他深度学习方法,本文所采用的稠密网络连接可以获得更小的参数重构误差,也即可以获得更加精确的人脸重构效果。

表 1　模型预测的均方误差定量对比

Tab.1　Quantitative comparison of the mean square error of model prediction

| 算法 | 左上人脸 | 右上人脸 | 下人脸 |
|---|---|---|---|
| FCN | 27.8 | 23.9 | 29.7 |
| LSTM | 26.3 | 22.7 | 27.6 |
| VGG16 | 25.4 | 21.8 | 25.8 |
| GoogLeNet | 23.6 | 20.8 | 25.2 |
| ResNet | 23.2 | 21.0 | 24.7 |
| SENet | 22.3 | 20.5 | 23.7 |
| 本文方法 | 21.9 | 20.6 | 23.2 |

### 2.4　人脸合成对比

经过稠密连接网络预测得到的人脸参数向量,当需要将其重构出人脸图像时,首先,需在预测参数的基础上进一步结合预训练得到的主成分基向量,将人脸参数重构出对应尺寸的人脸图像。图 6 展示了一些相关发音的人脸重构效果,图 6 中男性为

SAVEE 英文语音数据集的人脸合成结果，女性为 CCTV 中文语音数据集的人脸合成结果。总体上，与其他方法相比，本文采用的稠密连接网络可以获得更加逼真的人脸重构效果。

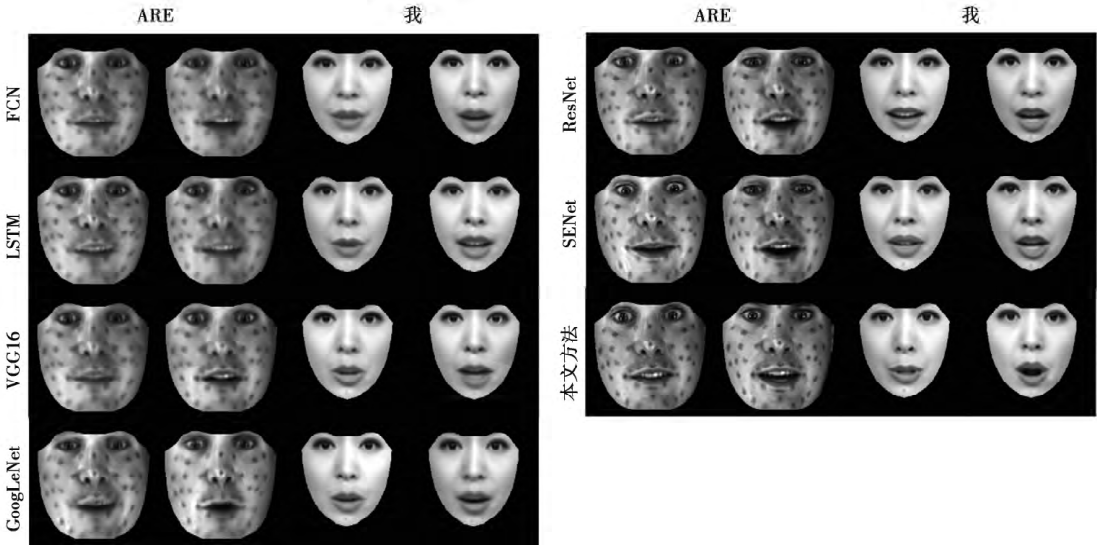



图 6 人脸合成对比

Fig.6 Comparison of face synthesis

## 3 总 结

本文提出了一种基于传统人脸参数化表示与深度网络相结合的跨模态学习方法。一方面，模型将一维语音信号经过频谱变换作为网络的输入；另一方面，模型将参数化表示的人脸图像作为网络的输出，在稠密连接的深度卷积网络 DenseNet 的基础上，进一步学习 2 种不同模态之间的非线性映射关系。在公开的音视频同步的数据集上，验证了所提出方法的有效性。

## 参考文献:

[1] WANG L, QIAN X, HAN W, et al. Synthesizing photo-real talking head via trajectory-guided sample selection [C] //Eleventh Annual Conference of the International Speech Communication Association. Makuhari, Chiba, Japan: DBLP,2010.

[2] LIU K, OSTERMANN J. Optimization of an image-based talking head system [J]. EURASIP journal on audio, speech, and music processing, 2009( 2009): 1-13.

[3] XIE L, LIU Z. Realistic Mouth-Synching for speech-driven talking face using articulatory modelling [J]. IEEE Transactions on Multimedia, 2007, 9( 3): 500-510.

[4] ANDERSON R, STENGER B, WAN V, et al. An expressive text-driven 3D talking head [M] //ACM SIGGRAPH 2013 Posters. Anaheim California, Association for Computing Machinery. New York, NY, United States: ACM SIGGRAPH 2013 Posters, 2013: 1-1.

[5] COOTES T F, EDWARDS G J, TAYLOR C J. Active appearance models [J]. IEEE Transactions on pattern analysis and machine intelligence, 2001, 23( 6): 681-685.

[6] SUWAJANAKORN S, SEITZ S M, Kemelmacher-Shlizerman I. Synthesizing obama: learning lip sync from audio [J]. ACM Transactions on Graphics ( TOG) , 2017, 36 ( 4): 1-13.

[7] TAYLOR S, KIM T, YUE Y, et al. A deep learning approach for generalized speech animation [J]. ACM Transactions on Graphics ( TOG) , 2017, 36( 4): 1-11.

[8] TAYLOR S L, MAHLER M, THEOBALD B J, et al. Dynamic units of visual speech [C] //Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation.Postfach 2926. Goslar, Germany: Eurographics Association, 2012: 275-284.

[9] EZZAT T, POGGIO T. Visual Speech synthesis by morphing visemes [J]. International Journal of Computer Vision, 2000, 38( 1): 45-57.

[10] ZHOU Y, XU Z, LANDRETH C, et al. Visemenet: Audio-driven animator-centric speech animation [J]. ACM Transactions on Graphics ( TOG) , 2018, 37( 4): 1-10.

[11] FAN B, XIE L, YANG S, et al. A deep bidirectional LSTM approach for video-realistic talking head [J]. Multimedia Tools and Applications, 2016, 75( 9): 5287-5309.

[12] MATTHEYSES W, LATACZ L, VERHELST W. Comprehensive many-to-many phoneme-to-viseme mapping and

its application for concatenative visual speech synthesis [J]. Speech Communication, 2013, 55(7-8): 857-876.

[13] WANG W, SHEN J, YANG R, et al. Saliency-aware video object segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(1): 20-33.

[14] WANG W, SONG H, ZHAO S, et al. Learning unsupervised video object segmentation through visual attention [C] //Proceedings of the IEEE conference on computer vision and pattern recognition. Long Beach, California: IEEE, 2019: 3064-3074.

[15] BREGLER C, COVELL M, SLANEY M. Video rewrite: Driving visual speech with audio [C] //Proceedings of the 24th annual conference on Computer graphics and interactive techniques. 1515 Broadway, 17th Floor New York, NY, United States: ACM Press/ Addison-Wesley Publishing Co, 1997: 353-360.

[16] SHIMBA T, SAKURAI R, YAMAZOE H, et al. Talking heads synthesis from audio with deep neural networks [C] //2015 IEEE/SICE International Symposium on System Integration (SII). Nagoya, Japan: IEEE, 2015: 100-105.

[17] FAN B, WANG L, SOONG F K, et al. Photo-real talking head with deep bidirectional LSTM [C] //2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brisbane, QLD, Australia: IEEE, 2015: 4884-4888.

[18] KING D E. Dlib-ml: A machine learning toolkit [J]. The Journal of Machine Learning Research, 2009(10): 1755-1758.

[19] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C] //Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, Nevada: IEEE, 2016: 770-778.

[20] HUANG G, LIU Z, VAN D M L, et al. Densely connected convolutional networks [C] //Proceedings of the IEEE conference on computer vision and pattern recognition. Hawaii: IEEE, 2017: 4700-4708.

[21] HAQ S, JACKSON P J B, EDGE J. Speaker-dependent audio-visual emotion recognition [C] //AVSP. Norwich, UK: ISCA, 2009: 53-58.

[22] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [C] //International Conference on Learning Representation. San Diego, CA: IEEE, 2015: 3-5.

[23] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C] //Proceedings of the IEEE conference on computer vision and pattern recognition. Boston, MA, IEEE, 2015: 1-9.

[24] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C] //Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, UT: IEEE, 2018: 7132-7141.

**作者简介:**

**唐　俊**(1990-),广西桂林人,男,硕士,上海理工大学机器智能研究院实验员。分别在 2013 年和 2016 年毕业于重庆邮电大学,获学士学位和硕士学位。研究方向为机器人系统。E-mail: tjoint@sina.com。

**牟海明**(1989-),重庆市忠县人,男,博士研究生。分别在 2013 年和 2016 年毕业于重庆邮电大学,获学士学位和硕士学位。研究方向为机器人运动控制,深度学习。E-mail: mhming@126.com。

**冷　洁**(1992-),男,湖北武汉人,博士研究生。2018 毕业于重庆邮电大学,获硕士学位。研究方向为机器人运动控制与强化学习。E-mail: lengjie@163.com。

**李清都**(1980-),男,湖北襄阳人,教授,博士。分别在 2002 和 2004 毕业于重庆邮电大学,获学士学位和硕士学位。2004 年至今在重庆邮电大学任教。研究方向为双足机器人、机器人运动控制等。E-mail: liqd@cqupt.edu.cn。

**刘　娜**(1985-),山东泰安人,女,讲师,博士。研究方向为深度学习、机器视觉、图像检索与识别、目标跟踪、语音识别等。E-mail: liuna@usst.edu.cn。

(编辑: 刘　勇)