文章编号：1004-4353(2020)03-0215-06

基于 Bi-LSTM 的面部特征与语音特征的映射模型

刘奕， 金小峰*
（延边大学 工学院，吉林 延吉 133002）

摘要：针对人脸动画技术中的面部特征与语音特征的映射问题，提出了一种基于双向长短时记忆网络(Bi-LSTM)的映射模型学习方法. 首先，在训练视频中同步地分别提取语音信号的 MFCC 参数和视频帧序列中的人脸特征点参数. 其次，训练映射模型过程中将 MFCC 参数作为 Bi-LSTM 网络的输入，将面部特征参数作为网络的期望输出，并引入参数调优机制对迭代次数、隐层单元数、批处理大小、优化器类型等进行实验调优，以此得到最优的映射模型. 对最优映射模型进行实验结果表明，采用双向Bi-LSTM 网络明显优于单向的 LSTM 网络，而且经过参数调优后映射准确率达到 0.895；因此，本文方法可以为后续的基于语音驱动的人脸视频合成应用提供有效的人脸特征预测参数.

关键词：人脸动画；梅尔频率倒谱系数；双向长短时记忆网络；参数调优
中图分类号：TP391　　**文献标识码**：A

A mapping model of facial features and speech features based on Bi-LSTM

LIU Yi， JIN Xiaofeng*
(*College of Engineering*, *Yanbian University*, *Yanji* 133002, *China*)

Abstract: Aiming the issue of mapping model between facial features and speech features in face animation technology, a mapping model learning method based on Bi-LSTM is proposed. Firstly, both MFCC parameters of speech and the facial landmark parameters of video frame are extracted concurrently from training video clips. Secondly, the mapping model is converged gradually by iterative training process with inputting MFCC parameters to Bi-LSTM network and expecting the corresponding facial landmark parameters as output. In the meantime, approaches of fine-tune is applied to obtain best mapping model by experimental method, such as best epoch times, number of hidden layers, batch size and type of optimizer. The best mapping model experimental results show that Bi-LSTM is significantly better than LSTM, and the mapping accuracy reaches 0.895 after parameter fine-tuning. Therefore, the proposed method can provide effective facial predictive landmark parameters for applications of speech-driven face video synthesis.

Keywords: facial animation; MFCC; Bi-LSTM; fine-tuning

0 引言

人脸动画技术是一种与指定输入的语音具有一致性且逼真的虚拟人脸动画合成技术，目前该技术被广泛应用在电影、游戏等娱乐产业[1]. 研究证明，人对视听双模态输入获取的信息比单一模

收稿日期：2020-03-21
* **通信作者**：金小峰(1970—)，男，教授，研究方向为机器感知、图像及音频处理.
基金项目：吉林省教育厅“十三五”科学技术项目(JJKH20191126KJ)；延边大学世界一流学科建设培育项目(18YLPY14)

态多,且更能理解感知对象,因此人脸动画技术在人机交互等计算机技术领域也具有广阔的应用前景[2].目前,在人脸动画技术的应用中,仍需要解决的两个主要问题是:一是语音特征与人脸特征之间的映射方法,二是针对指定语音的人脸视频帧的生成与视频合成方法.其中,映射方法是人脸动画合成的前提与基础,但因语音和人脸特征为异质的模态信息,因此语音和人脸特征之间对应的唯一标准是严格的时序对齐关系.Luo 等[3]提出了一种基于双高斯混合模型(Gaussian mixture model,GMM)的音频到视觉的映射模型,初步获取了视觉与音频特征之间的映射关系.赵晖[4]采用多阶隐马尔可夫模型(hidden Markov model,HMM)在双模态语料中对视素单元与语音单元进行了建模并建立了两者的映射模型,解决了可视语音合成中逼真度低的问题.张贺等[5]提出了一种基于主动外观模型(active appearance model,AAM)特征和异步发音特征的动态贝叶斯模型(dynamic Bayesian network,DBN)的人脸视频合成方法,该方法可有效改善原视频人脸唇部和音频不同步的现象.S.Taylor 等[6]采用深度神经网络(deep neural network,DNN)实现了音频特征到人脸特征的映射,并通过后期处理得到了高质量的人脸视频帧.肖磊[2]采用基于长短时记忆网络(long short-term memory,LSTM)得到了一种音频特征到人脸特征的映射模型,该模型得到的视频流畅、逼真.阳珊等[7]提出了一种基于双向长短时记忆网络(bi-directional long short term memory,Bi-LSTM)的人脸和语音特征映射模型,但该方法在参数化的合成过程中由于引入了PCA 的方法对参数进行降维,因此合成出的面部图像在细节方面存在不足.本文基于Bi-LSTM 方法,提出一种采用人脸约束局部模型(constrained local model,CLM)的面部特征与 MFCC 语音特征的映射模型,并通过测试验证该模型的有效性和准确性.

1 面部和语音的特征提取

1.1 基于 CLM 的面部特征提取

目前,定位面部特征点的主要方法包括基于可变形模板的定位方法[8]、点分布算法[9]、图算法[10]等,这些方法都是基于主动外观模型(AAM)或主动形状模型(active shape model,ASM)对特征点进行定位.2006 年,D.Cristinacce 等[11]提出了一种采用有约束的局部模型的 CLM 人脸特征提取方法,因该方法融合了 AAM 和 ASM 两个模型的优点,具有更高的检测准确性,因此本文采用该方法对人脸面部特征进行提取.图 1 是 68 个 CLM 特征点的空间分布图.

图 1 68 个 CLM 特征点的空间分布图

1.2 语音特征的提取方法

目前,常用的频谱特征有梅尔频率倒谱系数(Mel frequency cepstrum coefficients,MFCC)和线性预测倒谱系数(linear prediction cepstrum coefficient,LPCC)等[12].其中静态的 13 维 MFCC 参数虽然能够感知不同的语音频带,但是不能描述语音信号的时变特性,即静态的 MFCC 参数无法体现相邻帧之间和帧内部参数之间的动态相关性.为此,J.Ahmad 等[13]提出了一种改进算法,即在静态的 MFCC 参数上通过增加一阶和二阶差分形成了一个 39 维的 MFCC 参数组.本文采用文献[13]中的改进算法提取语音的 MFCC 特征参数.

2 基于 Bi-LSTM 的面部特征与语音特征的映射方法

2.1 Bi-LSTM 模型

由于 LSTM 不受固定输入序列长度的限制,而且还能灵活地根据需要选择保留的信息或者遗忘的信息,因此 LSTM 在处理不等长的序列化数据时具有明显的优势[14].LSTM 的循环体由遗忘门、输入门和输出门构成,如图 2 所示.通过遗忘

门可使信息有选择性地改变网络中每个时刻的输出状态，从而降低损失函数值. 图 2 中 $\boldsymbol{C}_t$ 表示细胞状态，$\boldsymbol{h}_{t-1}$ 和 $\boldsymbol{h}_t$ 分别表示不同时刻的隐层状态. 每个门结构都包含一个全连接层及其 sigmoid 激活函数(图 2 中标记为 σ)，其中 sigmoid 函数输出的是一个 0 到 1 之间的值(不包括 0 和 1)，该值表示当前输入结构中的信息量.

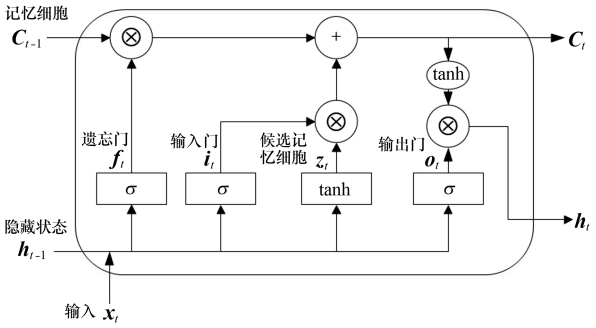

图 2　LSTM 模型的循环体结构

在 LSTM 中，遗忘门将前一时刻隐层状态 $\boldsymbol{h}_{t-1}$ 和当前时刻的输入 $\boldsymbol{x}_t$ 经 σ 输出 $\boldsymbol{f}_t$，$\boldsymbol{f}_t$ 的每一维值介于(0,1) 区间，因此遗忘门能够决定从细胞状态中丢弃哪些信息. 遗忘门的计算公式为：

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f)\,, \tag{1}$$

$$\sigma(x)=\frac{1}{1+\mathrm{e}^{x}}. \tag{2}$$

其中，$\boldsymbol{x}_t$ 表示当前时刻输入的音频特征向量，$\boldsymbol{h}_{t-1}$ 表示上一个时刻隐藏层的状态.

输入门 $\boldsymbol{i}_t$ 决定哪些输入层传入的信息可以被输入到细胞状态 $\boldsymbol{C}_t$ 里. 细胞根据由输入门输出的信息创建一个通过 tanh 层输出的新候选值向量 $\boldsymbol{z}_t$，此过程可表示为：

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_i\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i)\,, \tag{3}$$

$$\boldsymbol{z}_t=\tanh(\boldsymbol{W}_z\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_z)\,, \tag{4}$$

$$\tanh(x)=\frac{\mathrm{e}^{x}-\mathrm{e}^{-x}}{\mathrm{e}^{x}+\mathrm{e}^{-x}}. \tag{5}$$

其中，由 tanh 层输出的候选值向量 $\boldsymbol{z}_t$ 和 $\boldsymbol{i}_t$ 决定哪些值需要更新.

更新细胞状态 $\boldsymbol{C}_t$ 的步骤为：首先将前一时刻的 $\boldsymbol{C}_{t-1}$ 和 $\boldsymbol{f}_t$ 做 Hadamard 乘法操作(以此丢弃细胞状态中不需要的信息)，然后加上 $\boldsymbol{i}_t$ 和 $\boldsymbol{z}_t$ 的 hadamard 乘积. 上述过程可表示为

$$\boldsymbol{C}_t=\boldsymbol{f}_t\otimes\boldsymbol{C}_{t-1}+\boldsymbol{i}_t\otimes\boldsymbol{z}_t\,, \tag{6}$$

其中，$\otimes$ 表示 Hadamard 乘法操作.

输出门 $\boldsymbol{O}_t$ 确定输出什么值，其计算公式为

$$\boldsymbol{O}_t=\sigma(\boldsymbol{W}_o\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_o)\,. \tag{7}$$

t 时刻的 $\boldsymbol{h}_t$ 由公式 $\boldsymbol{h}_t=\boldsymbol{O}_t\otimes\tanh(\boldsymbol{C}_t)$ 计算得到. 上述公式中，$\boldsymbol{W}_f$、$\boldsymbol{W}_i$、$\boldsymbol{W}_z$、$\boldsymbol{W}_o$ 是可学习的参数矩阵，$\boldsymbol{b}_f$、$\boldsymbol{b}_i$、$\boldsymbol{b}_z$、$\boldsymbol{b}_o$ 是可学习的偏置.

由图 2 可以看出，LSTM 通过遗忘门忘掉过去的一些历史信息，然后再通过输入门决定当前时刻的信息是否被添加到细胞状态中，最后根据当前时刻的细胞状态进行选择性的输出. LSTM 模型是假设当前时刻的输出与之前的序列有关，即信息是通过隐藏状态从前向后传递的，但由于当前时刻的输出也有可能与后续的序列有关，因此有学者提出了双向 LSTM 模型(Bi-LSTM). Bi-LSTM 模型由一条前向和反向的循环神经网络组成，即每一时刻的输出由前向和后向的神经网络在该时刻的输出拼接组成[15]，因此 Bi-LSTM 模型更符合描述视频帧和语音帧序列的前后时序相关性. Bi-LSTM 模型的结构如图 3 所示.

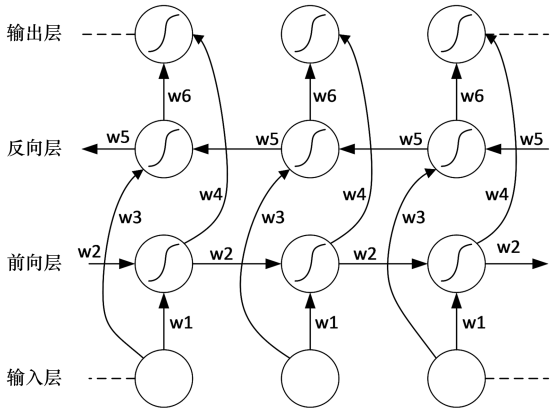

图 3　Bi-LSTM 模型的网络结构

2.2　映射模型的算法

本文提出的音频特征与人脸特征的映射模型的学习算法如下：

输入：视频训练样本，Bi-LSTM 网络参数初值，最大迭代次数

输出：音频特征与人脸特征的映射模型

Step1　提取语音 MFCC 特征参数. 首先从视频文件中分离出音频，然后提取语音的 MFCC 特征参数，并将提取的音频 MFCC 特征参数序列作为 Bi-LSTM 网络的输入. 分帧时，窗类型为矩形窗，窗长为 15 ms，窗移为 10 ms.

Step2 提取视频帧中的人脸特征点. 由于音频分帧与视频帧率不同,因此采用插值的方式来获取与音频帧数相同的视频帧数.

Step3 变换人脸特征点坐标. 将检出的人脸区域作为新的坐标变换空间,然后将提取的人脸特征点坐标变换到新的空间得到新坐标. 变换后的新坐标作为 Bi-LSTM 的输出期望.

Step4 训练神经网络. 视频训练样本经 Step1—Step3 步骤后即可得到人脸和语音特征的序列对. 利用 Bi-LSTM 网络训练这些序列对后即可得到拟合误差最小的映射模型,即 Bi-LSTM 网络的参数.

音频特征和人脸关键点的映射模型结构如图 4 所示.

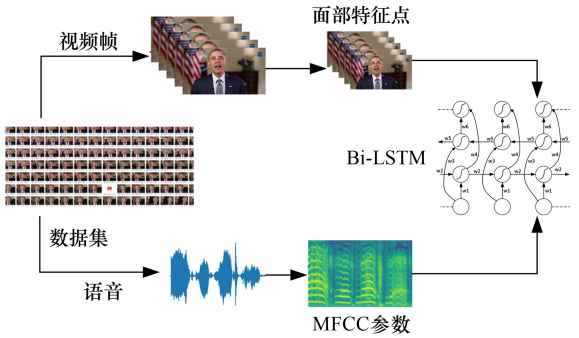

图 4 音频特征和人脸关键点的映射模型结构

采用人脸特征点坐标的误差平方和(SSE)作为本文算法映射模型的损失函数,SSE 的计算公式为

$$SSE=\sum_{i=1}^{n}(\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i)^2. \tag{8}$$

其中:n 表示每一帧人脸特征点坐标数,因本文采用 CLM 进行人脸特征点提取,所以 $n=68\times 2=136$;$\boldsymbol{x}_i$ 和 $\hat{\boldsymbol{x}}_i$ 分别表示特征点位置的真实值和预测值. 由式(8)可知,SSE 越小,模型的拟合精度越高.

计算单帧预测是否正确的公式为

$$c_i=\begin{cases}1, & \sum_{j=1}^{J}\|\boldsymbol{p}_j-\hat{\boldsymbol{p}}_j\|_2\leqslant\theta;\\ 0, & \text{其他}.\end{cases} \tag{9}$$

其中,c_i 表示第 i 帧的正确与否,θ 为阈值,$\boldsymbol{p}_j$ 和 $\hat{\boldsymbol{p}}_j$ 分别表示第 i 帧中第 j 个特征点的真实坐标值和预测坐标值. 累加所有测试视频帧中被正确预测的帧数后,按式(10)进行计算即可得到映射模型的准确率.

$$acc=\frac{1}{N}\sum_{i=1}^{N}c_i, \tag{10}$$

其中,N 是测试集的总帧数.

3 实验结果与分析

3.1 实验数据

实验所用的数据集来源于奥巴马时任美国总统期间的每周电台演讲视频,每期视频时长为 2~4 min. 由于很多期的内容是奥巴马外出考察或接受记者采访时录制的视频,因此部分视频中会出现没有人脸或多个人脸的情况. 剔除没有人脸或出现多个人脸的视频后,本文实际选取了其中 212 个视频作为样本(总时长约 1 000 min),并将样本集划分为训练集、验证集和测试集.

3.2 实验

3.2.1 LSTM 和 Bi-LSTM 映射算法的准确率对比 实验过程中 LSTM 和 Bi-LSTM 算法采用相同的测试集数据,并按式(10)计算准确率. 计算时采用的初始网络参数如表 1 所示,得到的结果见表 2.

表 1 LSTM 和 Bi-LSTM 算法的初始网络参数

初始参数	LSTM	Bi-LSTM
步长/ms	100	100
优化器	SGD	SGD
学习率	1e-6	1e-6
批处理大小	32	32
隐单元个数	10	10
训练迭代次数	50	50

表 2 不同阈值下 LSTM 和 Bi-LSTM 算法的准确率

阈值	LSTM 算法的准确率	Bi-LSTM 算法的准确率
1	0.384	0.412
4	0.551	0.566
9	0.641	0.652
16	0.716	0.734
25	0.765	0.790
50	0.921	0.943

从表 2 可以看出,在不同阈值下 Bi-LSTM 的准确率均优于 LSTM 的准确率,因此本文将

Bi-LSTM作为优选网络结构. 另外，从表 2 还可以看出，Bi-LSTM 和 LSTM 的预测准确率均随阈值的升高而提高. 因阈值越高预测值会更加偏离真实值，进而导致合成的人脸动画帧的逼真度下降，因此本文对最佳阈值的选取进行了实验. 经实验发现，当阈值 $\theta=25$ 时生成的人脸动画视觉效果最佳.

3.2.2　Bi-LSTM 网络参数的调优　调优参数包括训练次数、隐单元个数、批处理大小、优化器以及正则化等，调优步骤为：

1)选取最优迭代次数. 迭代次数分别取 1、10、20、50、100、200、300 进行训练，结果如表 3 所示. 从表 3 可以看出，当最大迭代次数为 100 和 200 时，模型的准确率最高(0.795)，因此本文选取 100 次作为最优迭代次数.

表 3　不同最大迭代次数下的模型准确率

最大迭代次数	准确率
1	0.314
10	0.439
20	0.537
50	0.721
100	0.795
200	0.795
300	0.781

2)选取最优隐单元个数. 隐单元个数决定模型的复杂度，其数量越多网络越复杂. 在最大迭代次数为 100 的情况下，分别取 32、64、128、256、512 个隐单元数量进行训练，结果如表 4 所示. 从表 4 可以看出，隐单元数量超过 256 后模型的准确率开始衰减，因此本文选取 256 作为最优隐单元数量.

表 4　不同隐单元数量下的模型准确率

隐单元数量	准确率
32	0.795
64	0.811
128	0.813
256	0.821
512	0.819

3) 选取最佳批处理大小. 分别取 8、16、32、64、128、256、512 等不同的批处理大小进行训练，结果如表 5 和图 5 所示. 从表 5 和图 5 可以看出，当批处理大小为 128 时模型的准确率最高，因此本文选取 128 作为最优的批处理大小.

表 5　不同批处理大小下的模型准确率

批处理大小	准确率
32	0.821
64	0.832
128	0.840
256	0.815
512	0.550

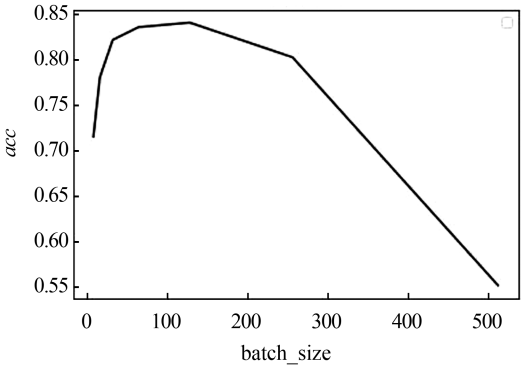

图 5　不同批处理大小下的模型准确率

4)选取最优优化器. 在不同迭代次数下，计算 SGD、RMSprop 和 Adam 等 3 种典型的优化器的准确率，结果如表 6 所示. 由表 6 可以看出，Adam 的准确率始终优于其他两种优化器，并在迭代次数为 100 时达到最高的准确率(0.885)，因此本文选取 Adlam 作为最优优化器.

表 6　不同迭代次数下各优化器的准确率

迭代次数	准确率		
	SGD	RMSprop	Adam
20	0.770	0.840	0.859
40	0.811	0.848	0.878
60	0.824	0.852	0.880
80	0.841	0.859	0.885
100	0.843	0.861	0.885

5)选取最优正则化方法. No-Regular、L1、L2 和 Combine-Regular 等 4 种正则化方法的模型准确率如表 7. 由表 7 可以看出，Combine-Regular 方法的准确率最高(0.895)，因此本文采用 Combine-Regular 作为最佳的正则化方法.

表7 不同正则方法下的模型准确率

正则化方法	准确率
No-Regular	0.885
L1	0.890
L2	0.889
Combine-Regular	0.895

4 结论

研究表明,本文提出的基于Bi-LSTM模型的面部特征与语音特征的映射模型在参数未调优且阈值取25的条件下,其准确率(0.790)明显优于LSTM模型的准确率(0.765);对本文提出的映射模型参数进行优化后,其准确率达到0.895;因此,本文提出的映射模型对人脸动画合成具有更好的实际应用价值.在今后的研究中我们拟将本文方法与高逼真度的人脸合成方法相结合,以此设计和开发基于语音驱动的人脸动画合成系统.

参考文献:

[1] 李欣怡,张志超.语音驱动的人脸动画研究现状综述[J].计算机工程与应用,2017,53(22):21-28.

[2] 肖磊.语音驱动的高自然度人脸动画[D].合肥:中国科学技术大学,2019.

[3] LUO C W, YU J, WANG Z F. Synthesizing real-time speech-driven facial animation[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence: IEEE, 2014: 4568-4572.

[4] 赵晖.真实感汉语可视语音合成关键技术研究[D].长沙:国防科学技术大学,2009.

[5] 张贺,蒋冬梅,吴鹏,等.基于AAM和异步发音特征DBN模型的逼真可视语音合成[C]//第十一届全国人机语音通讯学术会议论文集.西安:西北工业大学,2011.

[6] TAYLOR S, KATO A, MATTHEWS I A, et al. Audio-to-visual speech conversion using deep neural networks[C]//San Francisco: Interspeech. 2016: 1482-1486.

[7] 阳珊,樊博,谢磊,等.基于BLSTM-RNN的语音驱动逼真面部动画合成[J].清华大学学报(自然科学版),2017,57(3):250-256.

[8] 宋怀波,齐关锋,钱程.基于YUV颜色空间的脸部区域特征点定位方法[J].吉林大学学报(工学版),2013,43(S1):39-42.

[9] 潘翔,陈敖,周春燕,等.基于视图特征点分布的三维模型检索算法[J].浙江工业大学学报(自然科学版),2013,41(6):641-645.

[10] 贾海鹏,张云泉,徐建良.基于OpenCL的图像积分图算法优化研究[J].计算机科学,2013,40(2):1-7.

[11] CRISTINACCE D, COOTES T. Feature detection and tracking with constrained local models[C]//British Machine Vision Conference. Edinburgh: BMVA, 2006:929-938.

[12] 高庆吉,赵志华,徐达,等.语音情感识别研究综述[J].智能系统学报,2020,15(1):1-13.

[13] AHMAD J, FIAZ M, KWON S I, et al. Gender identification using MFCC for telephone applications-a comparative study[J]. International Journal of Computer Science and Electronics Engineering, 2015,3(5):351-355.

[14] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997,9(8):1735-1780.

[15] SCHUSTER M, PAILWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997,45(11):2673-2681.