计算机科学
COMPUTER SCIENCE

基于时空图卷积网络的语音驱动个人风格手势生成方法

张斌，刘长红，曾胜，揭安全

引用本文

张斌，刘长红，曾胜，揭安全.基于时空图卷积网络的语音驱动个人风格手势生成方法[J]．计算机科学，2022，49(11A)：210900094-5.

ZHANG Bin，LIU Chang-hong，ZENG Sheng，JIE An-quan. Speech-driven Personal Style Gesture Generation Method Based on Spatio-Temporal GraphConvolutional Networks [J]. Computer Science，2022，49(11A)：210900094-5.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900094

基于时空图卷积网络的语音驱动个人风格手势生成方法

张 斌 刘长红 曾 胜 揭安全
江西师范大学 南昌 330022
(zhangbin@jxnu.edu.cn)

摘 要 人们在发言时的手势动作往往具有自己独特的个人风格,研究者们提出了基于生成式对抗网络的语音驱动个人风格手势生成的方法,然而所生成的动作不自然,存在时序上动作不连贯的问题。针对该问题,文中提出了一种基于时空图卷积网络的语音驱动个人风格手势生成的方法,引入以时空图卷积网络为基础的时序动态性判别器,构建手势动作关节点之间空间和时间上的结构关系,并通过时空图卷积网络捕获手势动作关节点在空间上的相关性和提取时序上的动态性特征,使所生成的手势动作保持时序上的连贯性,以更符合真实手势的行为和结构。在Ginosar等构建的语音手势数据集上进行实验验证,与相关方法相比,正确关键点百分比指标提高了2%~5%,所生成的手势动作更自然。

关键词:跨模态生成;手势生成;个人风格学习;时空图卷积网络;时序动态性
中图法分类号 TP391.1

Speech-driven Personal Style Gesture Generation Method Based on Spatio-Temporal Graph Convolutional Networks

ZHANG Bin,LIU Chang-hong,ZENG Sheng and JIE An-quan
School of Computer & Information Engineering,Jiangxi Normal University,Nanchang 330022,China

Abstract People's gestures in speaking often have their own unique personal style. Researchers have proposed a speech-driven personal style gesture generation method based on generative adversarial networks. However,the generated actions are unnatural for temporal discontinuity. To solve this problem,this paper proposes a speech-driven personal style gesture generation method based on the spatio-temporal graph convolutional networks,which adds the temporal dynamic discriminator based on spatio-temporal graph convolutional network. The spatial and temporal structural relationships between gesture joint points is firstly constructed,and then the spatial correlation of gesture joint points is captured and the dynamic characteristics in time sequence are extracted through the spatio-temporal graph convolution network(STGCN),so that the generated gestures maintain the consistency in time sequenceand are more consistent with the behavior and structure of real gestures. The proposed method is verified on the speech and gesture dataset constructed by Ginosar et al. Compared with relevant methods,the percentage of correct keypoints improves by about 2%~5%,and the generated gestures are more natural.

Keywords Cross-modal generation,Gesture generation,Personal style learning,Spatio-Temporal graph convolutional networks,Temporal dynamics

1 引言

近年来,虚拟人物的应用场景越来越广泛,如老年助手[1]、视频教学[2]、儿童情绪调节[3]、电子商务[4]、机器人[5]以及虚拟治疗[6]等。甚至与真实视频中的人物相比,人们可能更多地被虚拟人物所吸引[7],而手势动作的生成是虚拟人物构建的关键技术之一。

当一个人全神贯注进行发言时,总是会不自觉地做出习惯性的手势动作,这些手势动作作为语言之外的一种重要补充[8],在沟通中发挥着不可忽视的作用。即使在进行打电话这种看不见对方的活动时,人们仍然会做出手势动作[9],这意味着手势不仅仅是为了丰富视觉上的内容,而且是个人风格的独特行为表现。Pouw等[10]的研究表明,人类说话和手势动作的相关性具有生物学和物理学的基础,手势对呼吸-发声系统有重要的影响,这为语音驱动个人风格的手势动作生成的研究提供了有力的依据。

然而,语音驱动个人风格手势生成是一项具有挑战性的任务。首先,手势与发言在发生时间上可能并不同步,手势可能出现在相应发言的同时,也可能出现在发言之前或之后[11],因此手势与语音在时间上不是对齐的。其次,这不是一个具有确定性的预测任务,因为发言者在不同场景下说同样的话时也可能做出不同的手势动作[12],这大大增加了任务的难度。最后,人体手势在结构上的自由度大[13],导致运动较为复杂,生成的手势应该是自然的,若要避免不流畅的手势动作,则必须考虑相邻关节的影响和时序上的连贯性[14]。

基金项目:国家自然科学基金(62067004,61662030)
This work was supported by the National Natural Science Foundation of China(62067004,61662030).
通信作者:刘长红(liuch@jxnu.edu.cn)

目前语音驱动手势动作生成已取得一定的研究成果[15-26]，加利福尼亚大学伯克利分校(University of California, Berkeley)的 Ginosar 等[12]从个人风格的角度，提出了语音至个人风格的动作生成，并公布了一个 144 h 时长视频的大型数据集，用于语音至个人风格手势动作生成研究。该模型能够根据个人语音生成个人独特风格的手势动作，但仅仅采用了真假判别器和 L_1 损失函数，从而导致所生成的动作不自然，存在时序上的不连贯性。为了解决该问题，本文引入时空图卷积网络[27]，提出了一种新的端到端的语音驱动个人风格手势生成模型。该模型构建关节点之间的空间结构关系，并通过时空图卷积网络捕获手势动作关节点在空间的相关性和提取时序上的动态性特征，使生成器所生成的手势动作保持时序上的连贯性，手势动作更自然。

2 相关工作

语音驱动手势动作生成主要是根据输入的语音数据生成对应的手势动作关节点数据，是跨模态生成研究领域的热点研究之一，目前已提出大量的方法，大致可分为 3 类：基于规则的方法、基于统计模型的方法和基于深度学习的方法。

(1)基于规则的方法。在早期的工作中，主要依赖规则生成手势动作。Marsella 等通过人们普遍使用的隐喻来驱动手势的选择[15]，如依次举起双手来表示两种事物的对比。Thiebaux 等则对点头等动作进行了具体的约束[16]。这些方法依赖明确的人为规则，局限性较大。

(2)基于统计模型的方法。该方法主要使用统计模型计算语音与动作的条件概率。譬如，Neff 等建立了一种针对特定员的统计模型[17]，Sadoughi 等使用概率图模型生成离散的手势[18]，而 Alexanderson 等提出的概率生成模型，解决了手势动作生成多样性问题，可在输入相同音频的情况下产生不同的手势动作[19]，能够对输出风格(如速度、范围等)进行一定程度的控制。

(3)基于深度学习的方法。自动编码器和生成式对抗网络等方法对深度特征的表达能力已成功被应用于跨模态生成研究领域[20]。Kucherenko 等[21]试图结合规则方法和数据驱动方法的优势，通过预测手势对某处发言的适合程度来优化生成模型。而 Hasegawa 等使用双向长短期记忆(Long Short-Term Memory, LSTM)生成三维的手势运动[22]。文献[23]首先提出学习运动的表示，然后训练网络，从语音中预测这种表示，而不是直接将语音映射到手势动作关节点。文献[24]着重关注声音信号与手势的时间关系，以眉毛动作作为节拍提升预测性能。Rebol 等在三维手势生成任务中固定了骨骼的长度，以生成更符合生物学的手势，并通过了图灵测试[25]。Yoon 等从 TED 演讲中收集且标注了一个用于协同语音手势研究的数据集，并提出了一种端到端的文本-动作生成方法[5]，通过机器人将这种动作带入现实世界。Habibie 等将音频信息编码后，分别使用 3 个解码器对脸部、身体和手进行解码，以区分身体不同部位在音频相关性上的差异[26]。最近 Ginosar 等则从个人风格的角度，提出了语音至个人风格的动作生成，并公布了 1 个时长为 144 h 的视频的大型数据集[12]，用于语音至个人风格手势动作生成研究。但该方法没有考虑到手势动作在时间和空间上的关系，从而导致所生成的动作不自然。

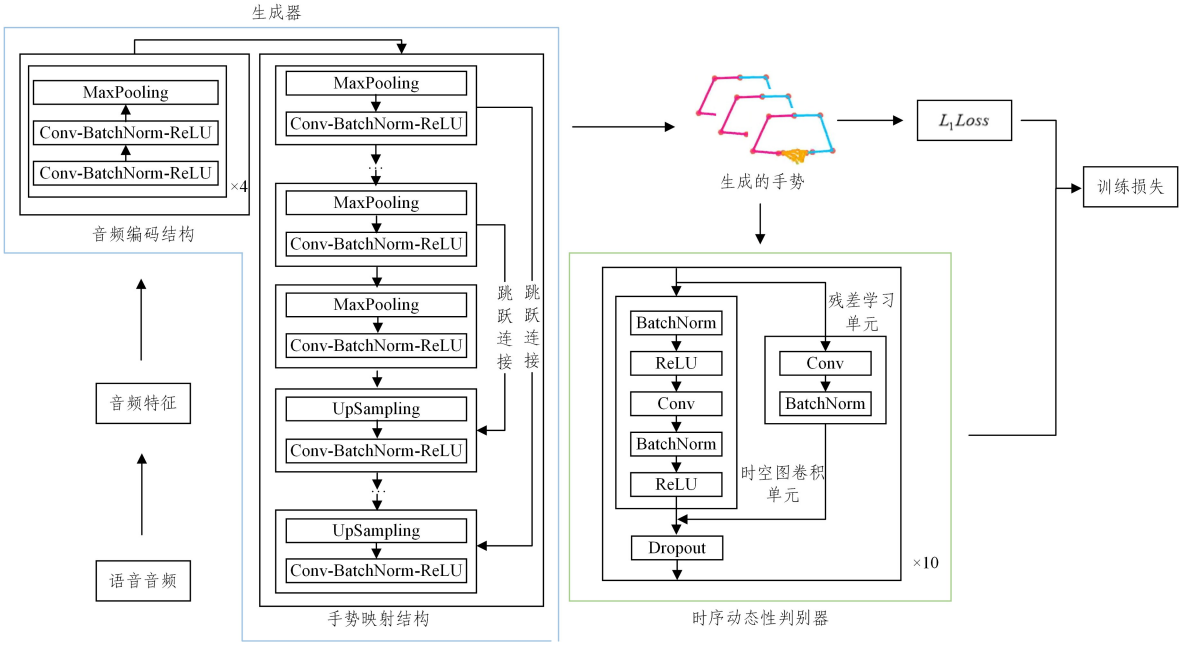


图 1 语音驱动个人风格手势生成模型示意图

Fig. 1 Schematic diagram of speech-driven personal style gesture generation model

3 基于时空图卷积网络的语音驱动个人风格手势生成模型

本文提出的语音驱动个人风格手势生成模型主要由两部分组成，即生成器和时序动态性判别器，如图 1 所示。首先提取发言者语音音频特征，如梅尔频谱(Mel Frequency Cepstrum Coefficient, MFCC)，然后输入生成器直接生成手势动作序列，手势动作采用人体关节点表示，接着将生成的手势动作序列与真实的手势动作序列输入预训练好的时序动态性判别器得到时序动态性损失，最后将生成的手势动作序列与真实的手势动作序列的 L_1 损失以及时序动态性损失结合并反馈给生成器，从而学习语音到手势动作序列的映射关系。

3.1　生成器

生成器 f_G 采用与文献[12]中相似的网络结构,由音频编码结构 f_{GE} 和手势映射结构 f_{GD} 两个部分组成。

音频编码结构 f_{GE} 由 4 个单元组成,每个单元为两个卷积层加上一个最大池化层,卷积层采用批归一化(Batch Normalization)和 ReLU 激活函数,每个单元的输入都是上一个模块的输出。将以梅尔频谱表示的语音特征 Y_T 输入 f_{GE} 进行下采样,从而得到低频语音表示 $f_{GE}(Y_T)$,其中 T 表示音频的时间跨度,维持一定长度的音频有利于保留手势动作在时序上的动态性。

手势映射结构 f_{GD} 由 5 个下采样卷积单元以及 5 个上采样卷积单元组成,在相同形状的单元上采用跳跃连接(Skip Connention)结构补充上采样过程中的特征丢失。f_{GD} 将 $f_{GE}(Y_T)$ 进行进一步下采样以提取特征,从而得到语言嵌入表示,再进行上采样,从而得到手势动作序列 S_F:

$$S_F = f_{GD}(f_{GE}(Y_T')) \tag{1}$$

假定所使用的手势骨骼二维关节点个数为 N,一组动作序列中手势个数为 M,则有 $S_F \in R^{2\times N\times M}$。

3.2　时序动态性判别器

仅使用 L_1 损失容易导致生成器保守地生成手势动作[28],本文引入基于时空图卷积网络的时序动态性判别器,生成更符合真实行为和真实结构的手势动作。

一个手势的所有骨架关节点集合表示为 $V_1=\{v_n \mid n=1,2,\cdots,N\}$,则一组手势动作序列的所有关节点集合表示为 $V=\{v_{m,n} \mid m=1,2,\cdots,M;n=1,2,\cdots,N\}$。构建无向图 $G=(V,E)$,表示手势动作序列上关节点之间空间结构关系和时序上的关联,G 的边集合 $E=E_1\cup E_2$,定义 A 为人体骨骼关节点之间所有固有的连接关系集合,则 $E_1=\{v_{m,i}v_{m,j} \mid m=1,2,\cdots,M;(i,j)\in A\}$ 表示一帧图像中手势动作关节点之间固有的骨骼连接关系集合,$E_2=\{v_{m,n}v_{m+1,n} \mid m=1,2,\cdots,(M-1);n=1,2,\cdots,N\}$ 表示手势动作序列中同一个关节点在不同图像帧之间的连接关系集合,因此集合 E_2 中所有与关节点 v_n 相关的边构成该点随时间变化的动作轨迹。定义关节点 $v_{m,n}$ 的邻节点集合为 $N(v_{m,n})=\{v_{m,n'} \mid l(v_{m,n},v_{m,n'})\leqslant L\}$(取邻节点范围 $L=1$),$l(v_{m,n},v_{m,n'})$ 表示从 $v_{m,n}$ 到 $v_{m,n'}$ 的最短路径长度,采样函数可表示为 $c(v_{m,n},v_{m,n'})=v_{m,n'}$。将邻节点集合中的节点分为固定的 i 个子集,每个子集都使用数字标签,映射函数为 $z_{m,n}:N(v_{m,n})\to\{0,1,\cdots,i-1\}$,权重函数可表示为 $w(v_{m,n},v_{m,n'})=w'(z_{m,n}(v_{m,n'}))$。此时,在无向图结构 $G=(V,E)$ 上的卷积运算可表示为:

$$f_{\text{out}}(v_{m,n})=\sum_{v_{m,n'}\in N(v_{m,n})}\frac{f_{\text{in}}(v_{m,n'})\cdot w'(z_{m,n}(v_{m,n'}))}{J_{m,n}(v_{m,n'})} \tag{2}$$

其中,$J_{m,n}(v_{m,n'})$ 为归一化项,平衡各子集的权重。

时序动态性判别器的网络结构由 10 个时空图卷积单元构成,前 4 个时空图卷积单元的通道数为 64,中间 3 个单元的通道数为 128,最后 3 个单元的通道数为 256。每个时空图卷积单元都经过图卷积操作、批归一化、ReLU 激活函数,接着采用残差机制,最后随机丢弃 50%的特征以避免过拟合。

在训练过程中,首先对时序动态性判别器 f_{TSD} 进行预训练,然后将生成器 f_G 生成的手势动作序列 $S_F\in R^{2\times N\times M}$ 与真实的手势动作序列 $S_G\in R^{2\times N\times M}$ 作为输入,分别提取时序动态性特征,得到生成的手势动作序列的时序动态性特征 $f_{TSD}(S_F)$ 和真实的手势动作序列的时序动态性特征 $f_{TSD}(S_G)$,将两组时序动态性特征之间的差异作为时序动态性损失函数 L_{TSD}。

$$L_{TSD}=\sum_{(Y_T,S_G)}\| f_{TSD}(S_F)-f_{TSD}(S_G)\|_1 \tag{3}$$

总损失函数 L_D 由时序动态性损失 L_{TSD} 和 L_1 损失所组成:

$$L_G=\lambda L_{TSD}+L_1 \tag{4}$$

其中,λ 为常系数。

3.3　时序动态性判别器预训练

为了训练时序动态性判别器 f_{TSD} 捕获手势动作在空间上的关系和时序上的动态性的能力,本文引入个人手势风格识别这项子任务,如图 2 所示,对时序动态性判别器进行预训练,并在数据集中随机分出了一个子集专门用于该项子任务的训练和测试。

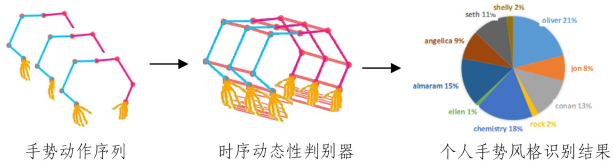


图 2　对时序动态性判别器进行预训练

Fig. 2　Pre-training of temporal dynamic discriminator

给定一组 M' 个手势动作组成的序列 $S_D\in R^{2\times N\times M'}$,输入 f_{TSD},得到动作序列在时空上的高级特征表示 $f_{TSD}(S_D)$,然后采用个人风格分类器 f_c 对相应的发言者手势动作进行风格分类,得到:

$$X_D=f_c(f_{TSD}(S_D)) \tag{5}$$

在这项任务中,本文使用 Top-1 准确率来评估方法在这项子任务上的性能,该子任务在训练集上达到了 98%的 Top-1 准确率,在测试集上达到了 82%的 Top-1 准确率,这表明时序动态性判别器能够有效地捕获手势动作在时序上的动态性,识别出不同手势风格的发言者在时序动态性上的差异。

4　实验结果与分析

为了验证本文方法的有效性,本文在 Ginosar 等所发布的公开数据集[12]上进行了实验验证,并与相关方法进行了对比分析。

4.1　数据集

Ginosar 等发布的语音手势数据集[12]共包含 10 位对象,分别是 5 位脱口秀主持人、3 位讲师和 2 位电话传教士。每个对象都有几个小时的视频,通过 OpenPose[29] 提取人体骨骼关节点的二维坐标作为手势动作的表示。本文按照 9∶1 的比例将数据集随机分为生成任务数据集和识别任务数据集,生成任务数据集用于语音驱动个人风格手势生成任务,识别任务数据集用于个人手势风格识别子任务,且这两个任务的数据均按照 8∶1∶1 的比例再随机分为训练集、验证集和测试集。

4.2　实验设置

本文方法使用 Python2.7 并基于 Tensorflow1.9 框架[30]实现,所有的实验都运行在 NVIDIA GeForce RTX 2080Ti GPU 以及 Intel Core i5-7500 CPU 上。

每秒语音生成 15 个手势动作,即生成手势的频率 $P=15$ Hz,每组完整的动作序列包含的手势个数为 $M=64$,用于表示手势动作的骨骼关节点总数 $N=49$,其中手臂部分的

骨骼关节点数 $N_a=9$，手指部分的骨骼关节点数 $N_f=40$。

在语音驱动个人风格手势生成任务的实验中，设置 batch size=16，并迭代训练 1 000 次，使用 Adam 方法[31]进行优化，学习率为 10^{-6}。在个人手势风格识别子任务的实验中，设置 batch size=64，同样使用 Adam 方法进行优化，在 10^{-5} 的学习率下迭代训练 100 次，然后学习率降低到 1×10^{-6}，继续迭代训练 500 次。

4.3 评估指标

本文使用文献[32]中的正确关键点百分比(Percentage of Correct Keypoints, PCK)作为评价指标，指标 PCK 越高，表示生成的结果越准确。如果生成的人体骨骼关节点的坐标位于实际坐标的 $\alpha\cdot(h,w)$ 个像素内，则认为这个关节点的坐标是正确的，其中 h 和 w 分别为真实标注的整个骨架坐标的最大高度差值和最大宽度差值，系数 α 设置为 0.15。

4.4 对比分析

为了评估本文提出的语音驱动个人风格手势生成方法的性能，与以下 3 种方法进行了对比分析：1)随机方法，随机选择数据集中的一个动作序列，将其作为个人风格手势生成的结果；2)speech2gesture，Ginosar 语音手势数据集上基于生成式对抗网络的基准方法[12]；3)speech2gesture-noGAN，Ginosar 语音手势数据集上去除了生成式对抗网络模块的方法，Ginosar 等认为这样会使得生成手势动作的性能指标更好。本文方法与这 3 种方法在 Ginosar 语音手势数据集中的 10 个对象上分别进行了对比分析，结果如表 1 所列。

表 1 在每个发言对象上分别评估的对比结果

Table 1 Comparative results evaluated separately on each speaker

(单位：%)

speaker	oliver	jon	conan	rock	chemistry	ellen	almaram	angelica	seth	shelly	Avg. PCK
Random	43.8	49.2	22.3	35.6	31.6	22.1	42.6	50.6	32.6	17.5	34.8
speech2gesture-noGAN	69.4	62.0	36.0	57.8	44.7	36.4	63.7	65.4	55.1	**39.3**	53.0
speech2gesture	64.8	63.2	34.5	51.4	43.4	35.9	60.0	61.7	54.5	37.7	50.7
our	**71.0**	**66.0**	**38.4**	**60.7**	**46.4**	**37.2**	**66.0**	**66.1**	**58.8**	38.5	**54.9**

从表 1 所列数据可以看出，本文方法在数据集中所有对象上的实验结果都要优于 speech2gesture 方法，平均 PCK 提升了 4.2%，这表明所提方法能够捕获手势动作在时序上的动态性，并有效用于语音驱动个人风格手势生成任务中。此外，speech2gesture-noGAN 方法去除了生成式对抗网络模块，没有对生成的手势动作进行判别或平滑处理，虽然使其在性能指标 PCK 上优于 speech2gesture 方法，但其平均 PCK 提升了 2.3%，但其模型容易导致生成的手势趋向于所有手势结果的均值。而本文方法不仅在大多数对象上的实验结果都要优于 speech2gesture-noGAN 方法，平均 PCK 也提升了 1.9%，并且所引入的时序动态性判别器能够防止生成的手势趋向于所有手势结果的均值。

图 3 给出了本文方法与 speech2gesture 和 speech2gesture-noGAN 的可视化对比结果，第一列为语音对应的真实手势动作，后三列分别为本文方法与 speech2gesture 和 speech2gesture-noGAN 根据发言者音频所生成的手势动作。

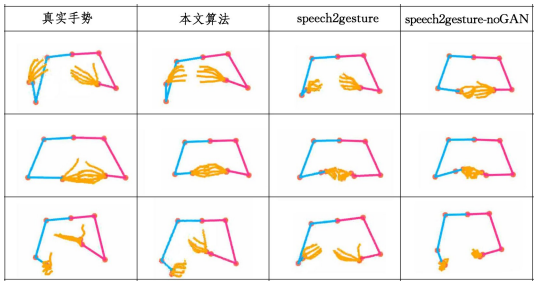


图 3 可视化对比结果

Fig. 3 Visualization of comparison results

从图 3 中可以看出，speech2gesture-noGAN 方法生成的手势动作总是较为保守的，且手指部分的骨骼关节生成效果不自然，这是由于 speech2gesture-noGAN 仅仅使用了 L_1 损失函数；speech2gesture 比 speech2gesture-noGAN 方法生成的手势动作更加自然且贴近真实的结果，但仍然有些保守，因为 speech2gesture 方法忽略了手势序列在时间和空间结构上的关系；而本文方法生成的动作学习了手势序列在时序上的动态性特征，利用了手势动作序列在时间和空间结构上的关系，从而使所生成的动作能够较自然地还原真实手势的效果，且手指部分较为自然。

结束语 本文提出了一种基于时空图卷积网络的语音驱动个人风格手势生成方法，通过在个人手势风格识别子任务上对时序动态性判别器进行预训练，能够有效地捕获手势动作在时序上的动态性，提升语音驱动个人风格手势生成任务的性能。在公开数据集上进行了实验验证，所提方法在性能指标(PCK)上比相关方法有明显的提升，所生成的手势动作更自然。

参 考 文 献

[1] YAGHOUBZADEH R, KRAMER M, PITSCH K, et al. Virtual agents as daily assistants for elderly or cognitively impaired people[C]// International Workshop on Intelligent Virtual Agents. Berlin, Springer: 2013: 79-91.

[2] LI J, KIZILCEC R, BAILENSONJ, et al. Social robots and virtual agents as lecturers for video instruction[J]. Computers in Human Behavior, 2016, 55: 1222-1230.

[3] PACELLA D, LÓPEZ-PÉREZ B. Assessing children's interpersonal emotion regulation with virtual agents: The serious game Emodiscovery[J]. Computers & Education, 2018, 123: 1-12.

[4] TAN S M, LIEW T W. Designing embodied virtual agents as product specialists in a multi-product category E-commerce: The roles of source credibility and social presence[J]. International Journal of Human-Computer Interaction, 2020, 36(12): 1136-1149.

[5] YOON Y, KO W R, JANG M, et al. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots[C]// 2019 International Conference on Robotics and Automation(ICRA). IEEE, 2019: 4303-4309.

[6] VAN VUUREN S, CHERNEY L R. A virtual therapist for speech and language therapy[C]// International Conference on

Intelligent Virtual Agents. Cham:Springer,2014:438-448.

[7] KANG S H,FENG A W,SEYMOURM,et al. Smart Mobile Virtual Characters: Video Characters vs. Animated Characters [C]// Proceedings of the Fourth International Conference on Human Agent Interaction. 2016:371-374.

[8] HOLLER J,LEVINSONS C. Multimodal language processing in human communication[J]. Trends in Cognitive Sciences,2019,23(8):639-652.

[9] BAVELAS J,GERWING J,SUTTON C,et al. Gesturing on the telephone: Independent effects of dialogue and visibility[J]. Journal of Memory and Language,2008,58(2):495-520.

[10] POUW W,HARRISON S J,DIXON J A. Gesture-speech physics: The biomechanical basis for the emergence of gesture-speech synchrony[J]. Journal of Experimental Psychology:General,2020,149(2):391.

[11] BUTTERWORTH B,HADARU. Gesture,speech,and computational stages: A reply to McNeill[J]. Psychological Review,1989,96(1):168-174.

[12] GINOSAR S,BAR A,KOHAVI G,et al. Learning individual styles of conversational gesture[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:3497-3506.

[13] WANG X,MENG H H,JIANG X T,et al. Survey on Character Motion Synthesis Based on Neural Network[J]. Computer Science,2019,46(9):22-27.

[14] XIN Q Q,CHEN Z X,FENG X X,et al. Movement Drive and Control Constraints of Virtual Hand Based on Multi-curve Spectrum[J]. Computer Science,2014,41(1):126-129,151.

[15] MARSELLA S,XU Y,LHOMMET M,et al. Virtual character performance from speech[C]// Proceed-ings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation. 2013:25-35.

[16] THIEBAUX M,MARSELLA S,MARSHALLA N,et al. Smartbody: Behavior realization for embodied conversational agents[C]// Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1. 2008:151-158.

[17] NEFF M,KIPP M,ALBRECHT I,et al. Gesture modeling and animation based on a probabilistic recreation of speaker style [J]. ACM Transactions on Graphics(TOG),2008,27(1):1-24.

[18] SADOUGHI N,BUSSO C. Speech-driven animation with meaningful behaviors[J]. Speech Communication, 2019, 110: 90-100.

[19] ALEXANDERSON S,HENTER G E,KUCHERENKO T, et al. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows[C]// Computer Graphics Forum. 2020, 39(2):487-496.

[20] GUO D,TANG S G,HONG R C,et al. Review of Sign Language Recognition, Translation and Generation[J]. Computer Science,2021,48(3):60-70.

[21] KUCHERENKO T,NAGY R,JONELL P,et al. Speech Properties Gestures:Gesture-Property Prediction as a Tool for Generating Representational Gestures from Speech[J]. arXiv: 2106.14736,2021.

[22] HASEGAWA D,KANEKO N,SHIRAKAWA S,et al. Evaluation of speech-to-gesture generation using bi-directional LSTM network[C]// Proceedings of the 18th International Conference on Intelligent Virtual Agents. 2018:79-86.

[23] KUCHERENKO T,HASEGAWA D,HENTER G E,et al. Analyzing input and output representations for speech-driven gesture generation[C]// Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. 2019:97-104.

[24] YUNUS F,CLAVEL C,PELACHAUD C. Sequence-to-Sequence Predictive Model: From Prosody To Communicative Gestures[C]// International Conference on Human-Computer Interaction. Springer,Cham,2021:355-374.

[25] REBOL M,GÜTI C,PIETROSZEK K. Passing a Non-verbal Turing Test: Evaluatina Gesture Animations Generated from Speech[C]// 2021 IEEE Virtual Reality and 3D User Interfaces (VR). IEEE,2021:573-581.

[26] HABIBIE I,XU W,MEHTA D,et al. Learning Speech-driven 3D Conversational Gestures from Video[J]. arXiv:2102.06837, 2021.

[27] YAN S,XIONG Y,LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]// Thirty-second AAAI Conference on Artificial Intelligence. 2018.

[28] REN X,LI H,HUANG Z,et al. Music-oriented dance video synthesis with pose perceptual loss[J]. arXiv: 1912.06606, 2019.

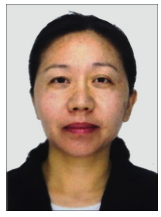
[29] CAO Z,HIDALGO G,SIMON T,et al. OpenPose:realtime multi-person 2D pose estimation using Part Affinity Fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019,43(1):172-186.

[30] ABADI M. TensorFlow:learning functions at scale[C]// Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming. 2016.

[31] KINGMA D P,BA J. Adam:A method for stochastic optimization[J]. arXiv:1412.6980,2014.

[32] YANG Y,RAMANAN D. Articulated human detection with flexible mixtures of parts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2012,35(12):2878-2890.

ZHANG Bin, born in 1997, postgraduate. His main research interests include cross-modal generation and computer vison.



LIU Chang-hong, born in 1977, Ph.D, associate professor, is a member of China Computer Federation. Her main research interests include computer vison, cross-modal retrieval and hyperspectral image processing.