

虚拟数字人手势交互设计

王晓慧, 覃京燕

(北京科技大学, 北京 100083)

摘要: **目的** 设计一个通用、动态、自学习的虚拟数字人手势生成模型, 为生成丰富情感表现力的虚拟数字人提供数据和模型基础。**方法** 首先总结隐喻手势的语义分类方法, 构建手势的量化描述语言, 为手势的语义可计算提供量化方法; 然后构建针对不同应用场景的情感语料库, 标注手势的情感、手势隐喻语义和手势的量化描述, 挖掘不同场景下虚拟数字人的行为模式, 为虚拟数字人手势生成研究提供数据基础; 最终提出基于应用场景的虚拟数字人手势生成算法, 建立抽象的交流意图与手势的物理实现之间的映射, 为生成具有环境认知能力的虚拟数字人提供模型支持。**结果** 通过两个案例验证了该方法的有效性, 解决了虚拟数字人对环境认知和互动不足的问题, 使虚拟数字人具有了行为的交互性和思想的智能性。**结论** 研究成果可推动虚拟数字人技术的发展, 为虚拟数字人在不同场景下的广泛应用提供数据和模型基础, 具有广泛的学术意义和应用前景。

关键词: 虚拟数字人; 交互设计; 手势交互; 隐喻手势; 手势生成

中图分类号: TB472 **文献标识码:** A **文章编号:** 1001-3563(2021)06-0046-07

DOI: 10.19554/j.cnki.1001-3563.2021.06.007

Gesture Interaction Design of Virtual Human

WANG Xiao-hui, QIN Jing-yan

(University of Science and Technology Beijing, Beijing 100083, China)

ABSTRACT: In order to design a general, dynamic, self-learning virtual human gesture generation model and provide data and model basis for the generation of virtual human with rich emotional expression, this paper first summarizes the semantic classification methods of metaphorical gestures, constructs a quantitative description language of gestures, and provides a quantitative method for the semantic computability of gestures. Then, the emotional corpus is built for different applications. It marks the emotion of gestures, the metaphorical semantics and quantitative description of gestures, and mines the behavior patterns of virtual human in different scenarios so as to provide a data basis for the virtual human gesture generation. Finally, a virtual human gesture generation algorithm based on application scenarios is proposed to establish the mapping between abstract communication intention and physical implementation of gestures, which provides model support for the generation of virtual human with environmental cognitive ability. As a result, the effectiveness of the method is verified by two cases, which solves the problem of virtual human's insufficient cognition and interaction to the environment. The method makes virtual humans have the interaction of behavior and the intelligence of thought. This research achievement can promote the development of virtual human technology, provide data and model basis for the wide application of virtual human in different scenarios, and has a wide range of academic significance and application prospects.

KEY WORDS: virtual human; interaction design; gesture interaction; metaphorical gestures; gesture generation

收稿日期: 2020-12-14

基金项目: 长江学者奖励项目(FRF-TP-18-010C1); 北京科技大学顺德研究生院项目(BK19AE011, BK20AF002); 佛山市促进高校科技成果服务产业发展扶持项目(2020DZXX05)

作者简介: 王晓慧(1987—), 女, 山东人, 博士, 北京科技大学副教授, 主要研究方向为计算机科学与设计学交叉。

通信作者: 覃京燕(1976—), 女, 四川人, 博士, 北京科技大学教授, 主要研究方向为人工智能与创新设计。

虚拟数字人是计算机生成的虚拟人物，具有人的几何特性和行为特性。在人工智能、虚拟现实等技术的带动下，以虚拟主播、虚拟前台、虚拟导游、虚拟客服、虚拟导师为代表的虚拟数字人已被应用在传媒、游戏、影视、金融、教育等各个领域，如腾讯发布的虚拟数字人“Siren”、央视网络春晚“小小撒”、浦发银行的“小浦”。

2020 年，中国人工智能产业发展联盟总体组和中关村数智人工智能产业联盟数字人工作委员会发布的 2020 年虚拟数字人发展白皮书指出，虚拟数字人制作开始从外观的数字化，逐渐深入到行为的交互化、思想的智能化。也就是说，虚拟数字人不仅要具有人的外观和行为，还要拥有人的思想，能够识别外界环境，并与人进行自然交互。因此，本文研究虚拟数字人中的手势交互，能为生成丰富情感表现力的虚拟数字人提供数据和模型基础。

随着图形界面（Graphical User Interface, GUI）、语音界面（Voice User Interface, VUI）、实体用户界面（Tangible User Interface, TUI）的发展，手势作为重要的交互方式^[1]，已被应用在汽车导航、虚拟仿真等领域^[2-3]，结合可穿戴设备可以提升手势识别的准确率^[4]。虚拟人作为新一代的交互平台，能融合其他交互技术提供多模态的交互体验^[5]。其中，手势交互大多关注于用户的手势识别^[6]。本文通过研究虚拟人手势生成，建立抽象的交流意图与手势的物理实现之间的映射，生成丰富的手势，让虚拟人的情感表现力更加丰富。

1 研究现状

1.1 虚拟数字人制作流程

虚拟数字人技术是语音识别、语音合成、自然语

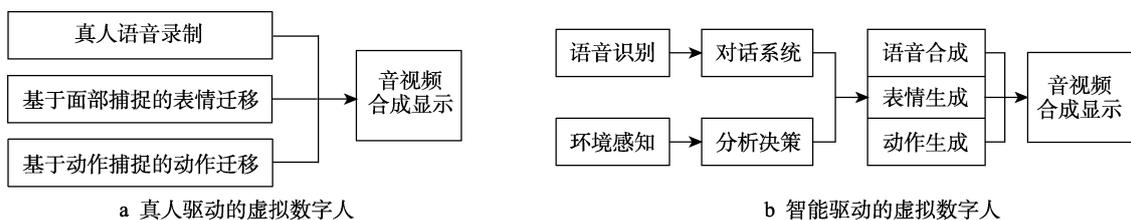


图 1 虚拟数字人制作流程
Fig.1 Generation flow of virtual human

音理解、动画生成等多种技术的融合，还处于发展阶段，尚没有通用的系统框架。根据虚拟数字人制作的两种方式总结制作流程，见图 1。真人驱动的虚拟数字人由真实人的语音和行为控制虚拟人的动作行为。随着 OptiTrack 等面部表情捕捉系统和动作捕捉系统的发展，真人驱动的虚拟数字人具有高逼真度，常用于影视制作。智能驱动的虚拟数字人由计算机自动合成其语音、表情和动作，具备用户语音识别、意图理解和环境感知的能力。

从智能驱动型虚拟数字人的技术研究角度来看，除了集成自然语言处理、语音识别、语言合成以外，大多关注于表情动作的生成^[7-8]，笔者前期也在生成丰富语义的表情和表情细节方面做了大量的工作^[9-10]。此外，大多研究关注于应用^[11-13]。目前虚拟数字人对环境的认知和互动性不强，这使得虚拟数字人难以生成个性行为，行为模拟和场景交互不尽如人意。

1.2 手势分类

语言学中将手势分为指示手势（Deictic Gestures）、隐喻手势（Metaphoric Gestures）、象似手势（Iconic Gestures）、节拍手势（Beats）^[14]。指示手势由含指示性动作的手势组成，既可引导听者注意当前感知环境中的事物，又可指向抽象的或想象的事物。象似手势是对物体的形状、运动和行为的模仿，主要通过事件或物体的形式及结构产生相似性起作用。隐喻手势所代表的概念没有物理形式，而是表征物体的抽象本质，是本文重点研究的对象。节拍手势是说说话者的手伴随着语言上下、前后有节奏地摆动。

以手势举例说明，见图 2。如图 2a：“先定一个能达到的小目标。”双手向前伸并停顿一下，表示未来的某个目标想法，表示动机。如图 2b：“比方说我先挣它一个亿。”伸出食指，指示“一”——“一个亿”，



图 2 手势举例
Fig.2 Examples of gestures

用手势表示数量;并向上抬起手部,表示坚定的信念,表示精神状态。

1.3 情感语料库构建

构建用于虚拟数字人手势生成的情感语料库对于手势生成算法研究至关重要。根据情感语料库的数据来源不同,可以将情感语料库来源分为以下三类:表演型语料数据、合成型语料数据和自然型语料数据^[15-16]。

1) 表演型语料数据:根据调研分析的结果,设计语料,并邀请志愿者对于给定的语句,使用不同的情感(一般以 Ekman 六种基本情感为主,包括高兴、惊讶、悲伤、愤怒、恐惧和厌恶)进行表演。其中,表演内容分为有情景和无情景,如给志愿者六种基本情感,让他在不同情绪下,自由地表演出“就是下雨也去”的情感;对语料“我考了340分”给出两种不同的情境,一种是高兴地说,觉得自己进步很大;一种是因为失误,考出低分的心理状态。

2) 合成型语料数据:通过机器学习,学习被模仿者的语音语调、表情动作等,采集语音数据,后期输入想要表现的文本话语,另起自动生成被模仿者语音语调的语料合集。

3) 自然型语料数据:从评论、微博、新闻等收集自然的语料数据,仅通过情感标注,对数据完成分类,不对语料进行人为修改,形成情感语料库。目前以使用褒贬/正负极评价的语料库为主。

根据情感语料库的标注分类方式,可将情感语料库分为以下两类:(1)正负极(褒贬)分类方式,多见于对网络评论、新闻、微博评论的分析评价,大多基于机器学习进行分类识别^[17];(2)多种情感分类方

式,例如六种基本情感,多用于表演型语料和合成型语料的情感分析^[18]。

目前,以六种基本情感/多种基本情感为情感标注分类方式的情感语料库或情感词典,都是对于表演型语料数据或合成型语料数据的分析。而以正负极情感分类方式,对自然型语料数据进行情感标注的,主要是对网络评论、新闻热点的分析,并没有在应用虚拟数字人方面提出设计创新。

2 隐喻手势的语义分类

隐喻手势可以从交际功能和隐喻内容两个角度分类。

根据交际功能可以分为:(1)修辞(Rhetorical);(2)增强作用(Intensifier);(3)表示数量(Quantifier);(4)表示比较(Comparative);(5)表示时间(Time);(6)表示方位(Location);(7)指示作用(Deixis);(8)精神状态(Mental State)^[19]。

根据隐喻内容可以分为:(1)时间运动(Time is Motion);(2)质量大小(Quantity is Size);(3)机动式运动(Action is Self-propelled Motion);(4)变化/改变(Change is Motion);(5)内容要素(Constituents are Contents);(6)目标路径(Goal is Path);(7)保持控制(Being in Control is being above);(8)物理结构的功能(Functionability is Physical Structure);(9)意识消失(Perceptible/Awareness is out)^[20]。

以马云在 Working Families Summit 中的演讲视频“Career Ladders and Leadership”中的手势为例,展示手势的隐喻分类。隐喻手势的语义类别示例见表1。

表1 隐喻手势的语义类别示例
Tab.1 Examples of semantic categories of metaphorical gestures

手势	对应文本	隐喻类别
	早上在巴黎,下午在伦敦,晚上在布宜诺斯艾利斯	(1) 时间运动
	Comment on	(3) 机动式运动
	When women succeed, America succeeds.	(4) 变化/改变

续表 1

手势	对应文本	隐喻类别
	Overcome those obstacles	(6) 目标路径
	That's not just	(7) 保持控制

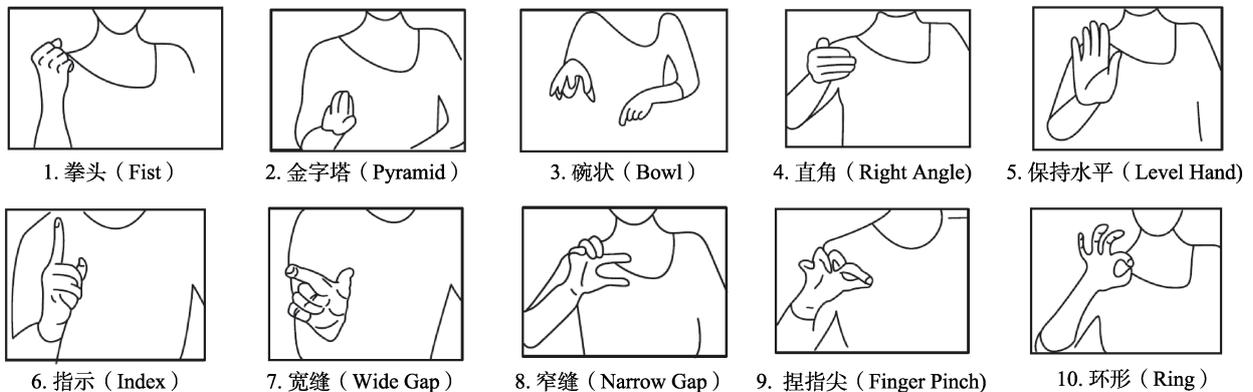


图 3 手部形状
Fig.3 Hand shape

3 手势的量化描述

提出一套手势的量化描述语言,该描述语言具有如下特点:(1)识别性强,当别人需要这套语言进行虚拟数字人的程序编写时,可以迅速理解并生成;(2)逻辑性强,为避免文字描述带来的个人主观影响,在语言的编写组成方式上,主要采用了空间坐标和阿拉伯数字进行文字归类 and 语言说明;(3)真实性强,整套语言基于对人的真实观察调研和网络资料分析,其中描述的动作和行为都具有一定的真实性。

手势的量化描述语言包括以下 5 个部分:(1)组成部分为左手、右手、两只手;(2)手部形状;(3)手心方向为向上、向下、向左、向右、面向自己、背对自己;(4)运动路径为手部起始位置、手部终止位置;(5)运动特点为快、慢、加速、突然停止。其中,手部形状包括十类^[19],见图 3。

手部位置的量化描述见图 4。在空间上,将人体模型坐标划分为 x, y, z , 不同的坐标分别对应不同的参照体。 x 轴以身体中间为坐标中心,相对地面平行且对模型实现方向垂直:左手(0)、左胳膊肘(1)、左肩(2)、头部(3)、右肩(4)、右胳膊肘(5)、右

手(6)。 y 轴以模型身体中心为出发点,视觉方向为轴方向,相对地面平行:贴近后背(0)、贴近胸(1)、身体一拳距离(2)、三拳距离(3)、五拳距离(4)、胳膊向前伸直的距离(5)。 z 轴以头顶为出发点,相对地面垂直:眉毛(0)、眼睛(1)、鼻子(2)、上嘴唇(3)、下嘴唇(4)、下颌(5)、肩部(6)、胸部(7)、腹部(8)、臀部(9)、大腿根部(10)。运动路径用手部起始位置和终止位置描述。

采用 JSON 格式描述一个手势,见图 5。5 个部分分别为:(1)组成部分 (Hand),即两只手;(2)手部形状 (Configuration),即左手和右手都保持水平;(3)手心方向,即两只手都面向自己;(4)运动路径,即两只手的起始坐标都是(2,5,1)、终止坐标(2,5,0);(5)缓慢运动。

4 用于手势生成的情感语料库构建

用六种基本情感标注自然型语料,重点研究语料选择和标注方式。情感标注采用人工标注的方式。在手势识别方面,基于本文提出的手势量化描述语言进行标注,采用半自动标注方法,在机器学习自动标注

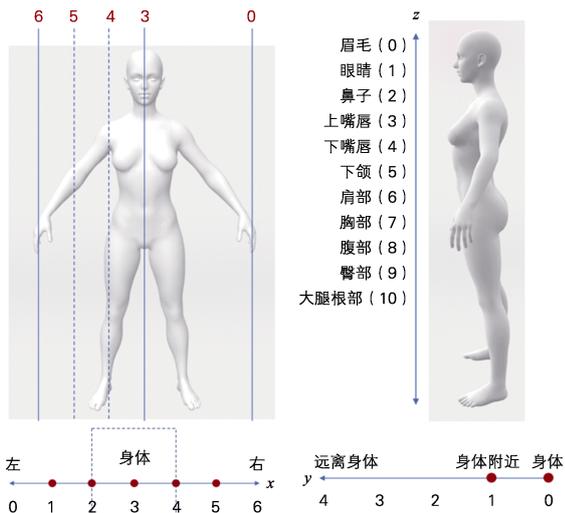


图4 手部位置的量化描述
Fig.4 Quantitative description of hand position

```

{
  "Hand": "both",
  "Configuration": { "left": "Level hand", "right": "Level hand" },
  "Orientation": { "left": "near", "right": "near" },
  "Start location": { "left": (2,5,1), "right": (2,5,1) },
  "End location": { "left": (2,5,0), "right": (2,5,0) },
  "Movement quantity": { "left": "slow", "right": "slow" }
}

```

图5 JSON格式的手势量化描述语言
Fig.5 Gesture's quantitative description language based on JSON

的基础上人工校正。

情感语料库的构建分为以下几个步骤：(1) 根据语料库应用方向选择合适的语料来源；(2) 确定情感标注内容；(3) 根据情感标注内容设计标注系统；(4) 统计数据，整合成语料库。

视频的选取标准：(1) 视频语句的发音时间基本需要限制在5s之内（通常一句话的表达，以10个字为例，在1~2s，语速、情感表达的不同会产生差异），避免因语句内容过长，影响标注者的判断；(2) 语句的情感表达直接，语句背景安静，无特殊噪音干扰；(3) 发音人普通话标准，容易分辨出情感类型。

截取视频语句后，对视频进行预处理，在视频末尾，对需要标注的语句增加2s的静止画面，方便标注者有足够的时间标注。在截取所有视频后，将全部文本整理成文档，存放入标注系统，方便标注，节约标注者判断语句的时间，可直接选择情感和情感关键词进行标注。语句时长只需要由第一遍标注的标注者进行标注。情感选择和情感关键词的单选需要标注多次。

以家庭服务型虚拟数字人的应用需要为例，构建情感语料库。从电视剧（都市现代剧）中截取片段，以其作为语料来源。视频语料来源于65集都市现代电视剧《都挺好》，根据情感表达筛选，从中截取320段视频语句，进行预处理后整理，放入标注系统，见图6。这是一个不同应用场景下的通用标注，包括选择标注视频变化、播放视频、标注六类基本情感、标

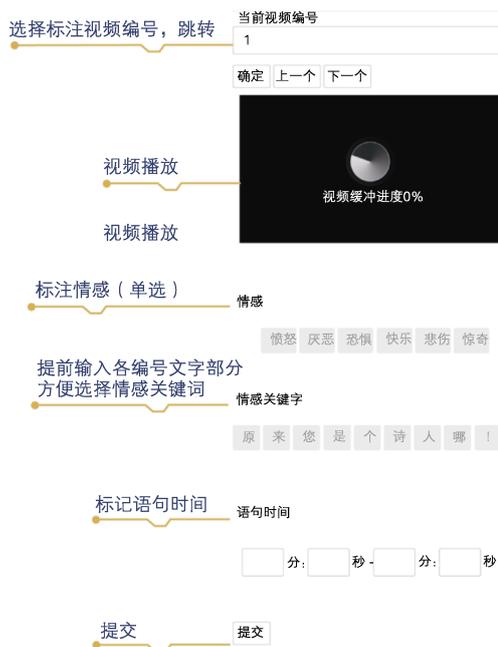


图6 标注系统界面
Fig.6 Interface of the marking system

注情感关键字、标记待标注语句的时间。其中，待标注语句是视频中的某一句话，为了准确标注情感类别，需要播放待标注语句的前后内容。待标注语句已经提前在线下转换为文本，并做好了分词。按照六种基本情感进行标注，每个语句保证至少标注两遍。最终对320个视频片段标注了情感、文本、情感关键词，其中生气语句50条、厌恶语句54条、害怕语句37条、高兴语句83条、伤心语句48条、惊喜48条。

5 虚拟数字人手势生成算法

建立抽象的交流意图与手势的物理实现之间的映射。并不是每句话都要伴有手势，并且同一句话在不同场景下的手势也会有所不同，例如同一句话在讲笑话和播放新闻时的语调、表情和手势都会不同。因此，对虚拟数字人的应用场景进行深层次的设计，研究基于应用场景的虚拟数字人手势自动生成算法，通过环境的变化让虚拟数字人产生不同的动作。

这里抽象的交流意图是基于自然语言处理，将文本映射到手势语义的分类中。基于上述构建的语料库构建分类模型，首先给出判断是否需要手势。因为手势是辅助和增强语义表达的，因此对于不确定是否需要手势的情况，倾向于给出不需要手势的结果。对所构建的不同场景的语料库进行统计，分析库中所有动作的模式及其与语义的对应关系。人类有丰富情感表现力的手势是一个有限集，整理语料库中的有语义的动作，进行情感、手势隐喻语义、原始文本与手势量化描述语言的映射关系，并构建分类模型。对于任意的文本，通过自然语言理解，构建该文本到情感、手势隐喻语义的分类，与原始文本的关联关系，最终生

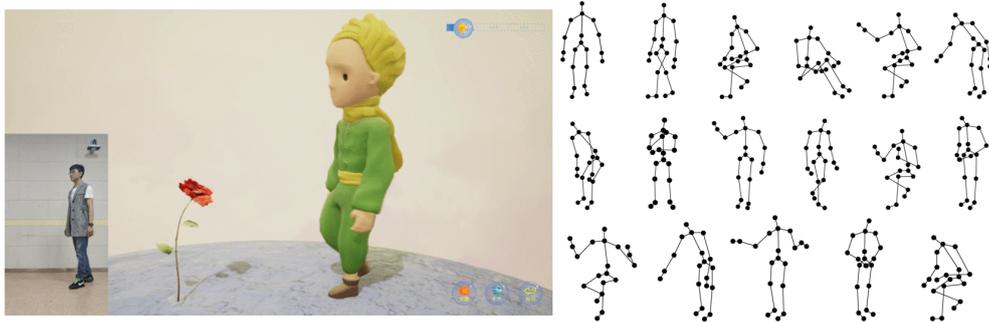


图 7 “小世界”：虚拟数字人陪伴儿童戏剧表演

Fig.7 “Little World”: virtual humans accompany children on dramatic performance

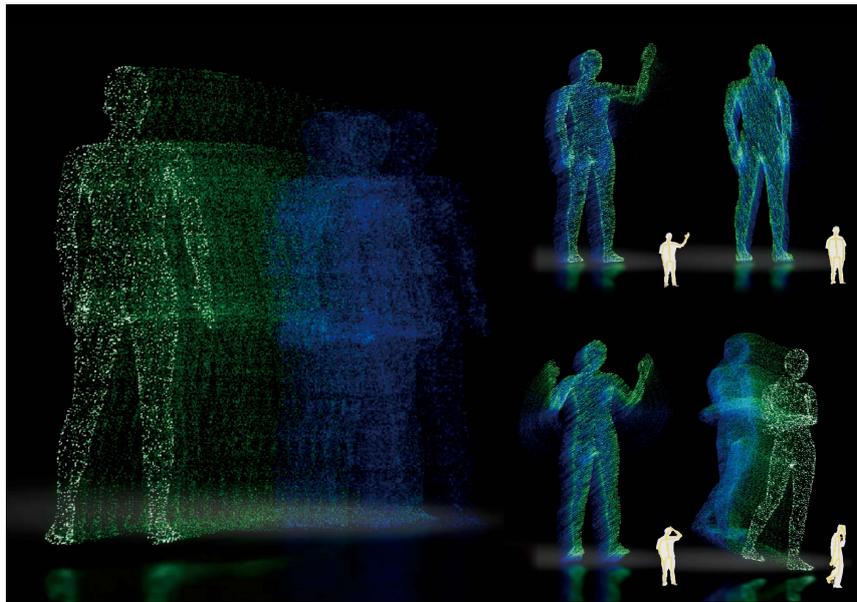


图 8 “AI 之镜”：可视化 AI 自我认知过程

Fig.8 “AI Mirror”: visualize AI's self-cognition process

成虚拟数字人手势。

从语义到手势量化描述的映射过程是一对多的分类问题。针对不同应用场景，使用不同语料库进行训练。对于同一个隐喻手势语义细分类别下的多个手势，采用自然语言处理技术，对库中手势文本与输入文本进行语义匹配，最终选择最匹配的手势。该算法非常依赖语料库，因此语料库的构建至关重要。因为人类有丰富情感表现力的手势是一个有限集，相比于手势的量化描述中所定义的维度组合，有语义的手势只占很少的一部分。因此，本文没有采用直接预测手势量化描述中每一个维度的数值，而是保守地从库中匹配手势，以保证虚拟数字人生成的每一个手势都准确无误，对于不确定的手势，虚拟数字人宁愿不做。

6 手势交互案例设计

6.1 “小世界”：虚拟数字人陪伴儿童戏剧表演

“小世界”交互艺术装置是虚拟数字人陪伴儿童戏剧表演系统，发布在 2020 年计算机顶级国际会议

ACM Multimedia 首届交互艺术展中，全球总共入围 7 个项目^[21]。每个孩子都是他独特世界中的首席演员，为了帮助他们提高表演水平，“小世界”互动艺术让虚拟数字人在戏剧表演中陪伴孩子。“小世界”为角色营造戏剧场景和虚拟数字人物，设计了 17 种典型语义的动作交互，见图 7，为儿童戏剧表演提供了一个新形式、新平台，实现儿童从观看者到表演者的转变，由被动接受变成主动参与，激发了儿童的创新意识和学习兴趣，同时使儿童充分发挥他们的想象力。

6.2 “AI 之镜”：可视化 AI 自我认知过程

“AI 之镜”交互艺术装置是可视化 AI 自我认知过程，发布在 2020 年计算机顶级国际会议 ACM Multimedia 首届交互艺术展中^[22]。如变色龙一般模仿的 AI 逐渐产生意识，AI 对其历史会有怎样的看法，人类对 AI 会有怎样的感受？“AI 之镜”从强 AI 视角可视化其自我认知机制，并引起人们对人工智能的反思。通过虚拟数字人和真人的手势动作交互，见图 8。在无意识模仿的第一阶段，视觉神经元感知环境

信息,而镜像神经元模仿人类行为,如图8中第一行的两个动作。然后,从长期的模仿中产生语言和意识,呈现为诗句和情感空间中的坐标。在最后的自主意识行为阶段,如图8中第二行的两个动作,镜像神经元将与用户进行更和谐的行为或者产生自主运动,让用户像照镜子一样发觉可能未曾注意的行为特征并引发反思。

6.3 方法的局限性讨论

本文总结的隐喻手势语义分类方法和构建的手势量化描述语言,提供了通用的手势语义可计算方法。提出的语料库构建方法,可实现针对不同应用场景的动态构建。虚拟数字人手势生成模型是基于所构建数据库的自学习过程。因此,设计了一个通用、动态、自学习的虚拟数字人手势生成模型,为生成丰富情感表现力的虚拟数字人提供数据和模型基础。该方法非常依赖数据库,因此工作量较大。为了保证虚拟数字人的动作要么不做,做就要做得准确,便采用了保守的数据匹配方法,使生成结果是一个有限集,由此数据库中样本较少,可能会导致多样性不足。在未来研究中,收集“小世界”和“AI之镜”等应用中用户产生的手势动作,与积累的大量的不同应用场景的语料库整合,构建基于深度学习的虚拟数字人手势生成模型,生成具有丰富语义表现力的动作。

7 结语

虚拟数字人手势交互能丰富其情感表现力。首先,总结了隐喻手势的语义分类方法,构建了手势的量化描述语言,为手势的语义可计算提供了量化方法;提出了针对不同应用场景的情感语料库构建方法,以自然型语料为例,构建了丰富情感的语料库,为虚拟人手势生成研究提供了数据基础;最终构建了基于应用场景的虚拟人手势生成方法。通过两个交互艺术装置展示了虚拟数字人手势交互过程,挖掘了手势交互的广泛应用场景。

参考文献:

- [1] WILLIAMS A S, ORTEGA F R. Understanding Gesture and Speech Multimodal Interactions for Manipulation Tasks in Augmented Reality Using Unconstrained Elicitation[J]. Proceedings of the ACM on Human-Computer Interaction, 2020, 26(12): 3479-3489.
- [2] ZHAO Dan, LIU Yue, WANG Yong-tian, et al. Analyzing the Usability of Gesture Interaction in Virtual Driving System[C]. Osaka: IEEE Conference on Virtual Reality and 3D User Interfaces, 2019.
- [3] 邹序焱, 何汉武, 吴悦明, 等. 基于手势交互的虚实融合实验仿真技术[EB/OL]. (2020-11-02)[2020-12-14]. <http://kns.cnki.net/kcms/detail/11.3092.v.20201030.1432.005.html>.

- [4] 刘璇恒, 邓宝松, 裴育, 等. 穿戴式手势交互系统与识别算法研究[J]. 小型微型计算机系统, 2020, 41(11): 2241-2248.
- LIU Xuan-heng, DENG Bao-song, PEI Yu, et al. Wearable Gesture Interaction System and Recognition Algorithm[J]. Journal of Chinese Computer Systems, 2020, 41(11): 2241-2248.
- [5] KIM K. Improving Social Presence with a Virtual Human via Multimodal Physical-Virtual Interactivity in AR[C]. Montreal: CHI Conference on Human Factors in Computing Systems, 2018.
- [6] XIAO Yang, YUAN Jun-song, THALMANN D. Human-virtual Human Interaction by upper Body Gesture Understanding[C]. Singapore: Proceedings of the 19th ACM Symposium on Virtual Reality Software and Technology, 2013.
- [7] TISSERAND Y, AYLETT R, MORTILLARO M, et al. Real-time Simulation of Virtual Humans' Emotional Facial Expressions, Harnessing Autonomic Physiological and Musculoskeletal Control[C]. Scotland: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, 2020.
- [8] 熊伟. 基于三维动画的虚拟人物表情动作系统设计[J]. 现代电子技术, 2020, 43(20): 97-101.
- XIONG Wei. Design of 3D Animation Based Expression and Action System for Virtual Character[J]. Modern Electronics Technique, 2020, 43(20): 97-101.
- [9] 张申, 贾珈, 王晓慧, 等. 基于语义维度的人脸表情生成[J]. 清华大学学报(自然科学版), 2011, 51(1): 80-84.
- ZHANG Shen, JIA Jia, WANG Xiao-hui, et al. Facial Expression Synthesis Based on Semantic Dimensions [J]. Journal of Tsinghua University, 2011, 51(1): 80-84.
- [10] 王晓慧, 贾珈, 蔡莲红. 基于小波图像融合的表情细节合成[J]. 计算机研究与发展, 2013, 50(2): 387-393.
- WANG Xiao-hui, JIA Jia, CAI Lian-hong. Expression Detail Synthesis Based on Wavelet-Based Image Fusion[J]. Journal of Computer Research and Development, 2013, 50(2): 387-393.
- [11] SCASSELLATI B, BRAWER J, TSUI K, et al. Teaching Language to Deaf Infants with a Robot and a Virtual Human[C]. Montreal: 2018 CHI Conference on Human Factors in Computing Systems, 2018.
- [12] YAO H, SIQUEIRA A G D, FOSTER A, et al. Toward Automated Evaluation of Empathetic Responses in Virtual Human Interaction Systems for Mental Health Scenarios[C]. Scotland: IVA 20: ACM International Conference on Intelligent Virtual Agents. ACM, 2020.

(下转第76页)