

DOI:10.3969/j.issn.1003-5060.2026.07.002

基于多模态交互的黄梅戏智能数字人设计开发

陶健林¹, 胡孟霞², 陈姣姣², 陈晓慧¹, 王广军²

(安庆职业技术学院 信息技术学院, 安徽 安庆 246003; 2. 安庆师范大学 计算机与信息学院, 安徽 安庆 246011)

摘要:在文化数字化背景下,针对非遗戏曲数字化传播交互缺失、多模态信息融合不足、黄梅戏青年群体深度认知度低的问题,文章以实现黄梅戏活态数字化传承为目标,构建基于多模态交互的黄梅戏智能数字人系统框架,并以黄梅戏经典角色冯素珍为实例开展技术验证。系统融合语音、视觉、动作三大模态:语音实现方言识别与唱腔解析,视觉融合高分辨率网络保障人脸与表情识别精度,动作依托光学-惯性捕捉及动态时间规整匹配“兰花指”等程式化动作;基于注意力机制实现多模态信息融合,支持语音、手势及沉浸式交互,形成“技术赋能-多模态融合-场景驱动”的技术路径。该研究为文化遗产数字化提供一定的技术参考,对推动多模态人机交互研究具有理论与应用价值。

关键词:智能数字人;多模态交互;黄梅戏;语音识别;人脸检测;动作捕捉

中图分类号:TP391.9

文献标志码:A

文章编号:1003-5060(2026)07-0872-10

Design and development of Huangmei Opera intelligent digital human based on multimodal interaction

TAO Jianlin¹, HU Mengxia², CHEN Jiaojiao², CHEN Xiaohui¹, WANG Guangjun²

(1. School of Information Technology, Anqing Vocational and Technical College, Anqing 246003, China; 2. School of Computer Science and Information Engineering, Anqing Normal University, Anqing 246011, China)

Abstract: Against the backdrop of cultural digitalization, aiming to address the issues of insufficient interaction in the digital communication of intangible cultural heritage(ICH) operas, inadequate fusion of multimodal information, and low level of in-depth cognition of Huangmei Opera among young groups, this study takes the living digital inheritance of Huangmei Opera as the objective, constructs a framework of Huangmei Opera intelligent digital human system based on multimodal interaction, and conducts technical verification with Feng Suzhen, a classic character from Huangmei Opera, as a case study. The system integrates three core modalities, namely speech, vision, and motion: in the speech dimension, it achieves dialect recognition and aria analysis; in the vision dimension, it integrates a high-resolution network(HRNet) to ensure the accuracy of face and facial expression recognition; in the motion dimension, it relies on optical-inertial capture and dynamic time warping(DTW) algorithm to match stylized movements such as “orchid fingers”. Furthermore, it realizes multimodal information fusion based on the attention mechanism, supporting voice, gesture, and immersive interaction, and forms a technical path characterized by technology empowerment, multimodal fusion, and scenario-driven application. This study provides technical reference for the digitalization of cultural heritage and holds theoretical and application value for advancing the research on multimodal human-computer interaction.

Key words: intelligent digital human; multimodal interaction; Huangmei Opera; speech recognition; face detection; motion capture

收稿日期:2025-09-01;修回日期:2026-03-11

基金项目:安徽省高等学校科学研究资助项目(2023AH053073);安徽省高等学校质量工程资助项目(2024cxtd275)

作者简介:陶健林(1985—),女,安徽枞阳人,安庆职业技术学院副教授;

王广军(1987—),男,安徽宣城人,博士,安庆师范大学副教授,硕士生导师,通信作者,E-mail:guangjunwang@aqnu.edu.cn.

0 引言

随着数字技术的快速发展,多模态人机交互系统正成为提升数字内容自然性、沉浸感与智能性的重要研究方向。智能数字人作为整合语音、视觉、动作等多通道信息交互载体,已逐步应用于文化传播、虚拟导览、在线教育、数字娱乐等领域^[1-2],其设计与开发涉及人工智能、计算机图形学、感知交互等多学科交叉,具有重要的理论价值与应用潜力。尽管当前数字人技术已取得显著进展,但在多模态信息的深度融合、跨场景情感化交互等方面,仍存在诸多技术挑战和优化空间。

传统戏曲作为中华优秀传统文化的重要组成部分,是国家级非物质文化遗产的核心门类,其传承与传播正迎来数字化转型的迫切需求,而黄梅戏作为极具代表性的地方戏曲剧种,在新媒体环境下逐渐陷入观众老龄化、传播形式单一、年轻群体参与度与深度认知度双低的发展^[3]。新浪热点《2024年非遗戏曲关注度趋势洞察报告》数据显示,非遗戏曲整体关注群体已呈现明显的年轻化

特征,90后群体占比最高,达到41.86%,其次是80后和00后,反映出青年群体对非遗戏曲的潜在关注需求已逐步释放;但黄梅戏作为极具代表性的地方戏曲剧种,受到传播形式单一、数字化交互体验缺失、青年向传播内容开发不足等因素的影响,尚未有效承接这一年轻化趋势,在新媒体环境下仍陷入观众老龄化、年轻群体参与度与深度认知度双低的发展困境。具体数据分布如图1所示。

综合来看,现有数字化成果多聚焦于戏曲资源的数字化存档与单向展示,缺乏能激发用户参与感的沉浸式交互设计,从而导致青年群体对戏曲文化的认知停留在表层,难以形成深度文化认同。这一现状反映出非遗戏曲在青年群体中的关注度较高但深入认知程度有限、传播仍以传统渠道为主的痛点^[4],也表明数字化、互动化传播手段亟待加强,更为智能数字人技术在非遗戏曲传播领域的应用提供了具有代表性的文化场景,同时对技术的适应性、真实性与交互性提出了更高要求。

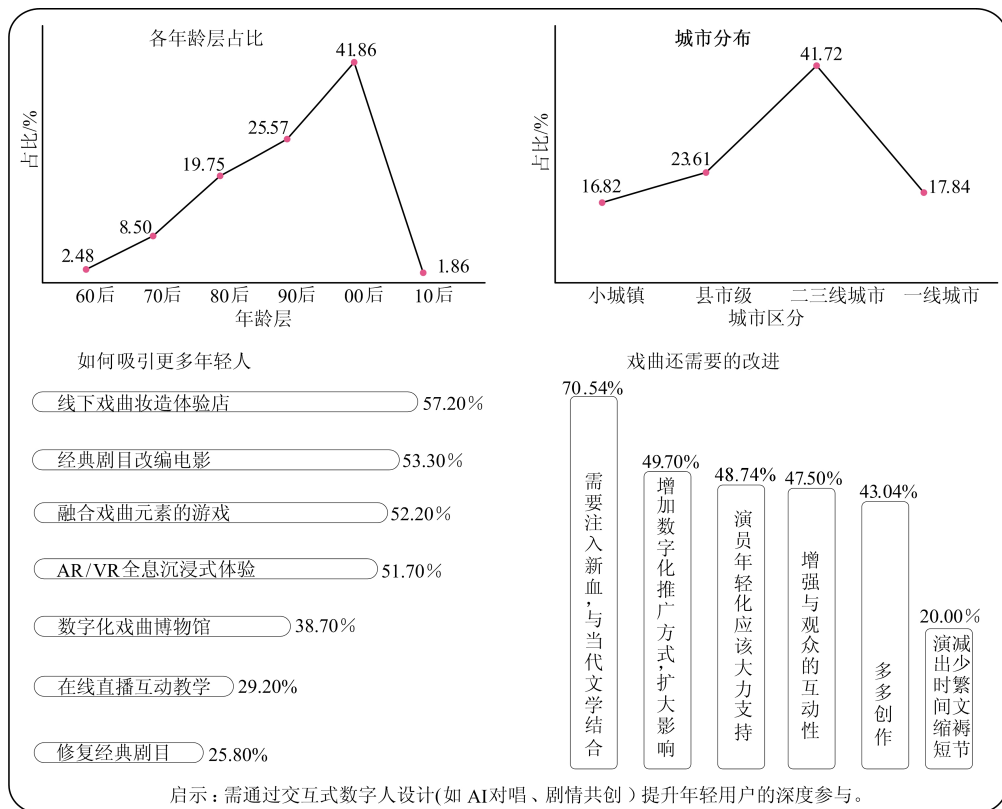


图1 非遗戏曲关注度趋势洞察报告数据

本研究以黄梅戏为研究对象,针对非遗戏曲数字化传播交互缺失、青年群体深度认知度不足

等问题,以实现黄梅戏活态数字化传承、破解青年传播困境为目标,构建基于多模态交互的黄梅戏

智能数字人系统框架,并以经典角色冯素珍为实例开展技术验证与应用研究。研究重点探索语音、视觉与动作三大模态的协同机制与深度融合方法,通过虚实融合的交互场景验证系统效能,并设计问卷调查法量化评估系统在提升青年群体黄梅戏深度认知度、参与积极性方面的效果。本研究不仅为黄梅戏等非遗戏曲的数字化传播提供可参考的技术路径,更致力于形成一套具有可迁移性与扩展性的智能数字人设计框架^[5],推动多模态交互理论在文化遗产活化、数字内容生成及跨媒体传播等领域的深入应用,为非遗戏曲的青年化传播与创新发展提供技术支撑与实践参考。

1 多模态交互设计的理论基础

1.1 虚实共生视角下的多模态认知神经机制

在虚实共生视角下,多模态交互机制主要涉及大脑对来自虚拟与现实不同感官通道的信息的处理和整合过程。大脑通过感知、认知和运动等多个模块,对视觉、听觉、触觉等多种模态的信息进行接收、处理和响应,从而实现与虚实融合环境的有效交互^[6]。感知模块负责接收来自虚实共生环境中的不同感官通道的信息;认知模块对感知到的虚实融合信息进行进一步的分析和理解,大

脑通过认知处理器,对不同模态的虚实信息进行融合和关联,形成对虚实共生环境的综合认知^[7];运动模块根据认知模块的决策,生成相应的运动指令,通过运动程序执行器控制身体的动作,实现与虚实共生环境的交互。例如,在人与黄梅戏智能数字人的交互中,数字人可通过基于哈尔类(Haar-like)特征的面部检测技术捕捉用户表情(视觉信息),通过基于梅尔频率倒谱系数(Mel-frequency cepstral coefficients, MFCC)的语音识别技术接收用户指令(听觉信息),经大脑皮层处理与融合后,生成精准的交互反馈。

1.2 多模态交互的核心维度

在人工智能技术快速迭代的当下,多模态交互系统通过语音、动作与视觉三大核心维度的深度协同,正重塑人机交互的范式。基于自然语言处理、计算机视觉和动作捕捉技术的融合创新,数字人作为多模态交互的典型载体,实现了从基础感知到智能创造的跨越式发展。区别于传统的单模态交互系统,这种三维度的协同机制通过高维信号重构,不仅模拟人类的外在形态,更赋予数字人理解意图、感知情绪和动态反馈的拟人化交互能力^[8]。

多模态数字人功能分类如图 2 所示。

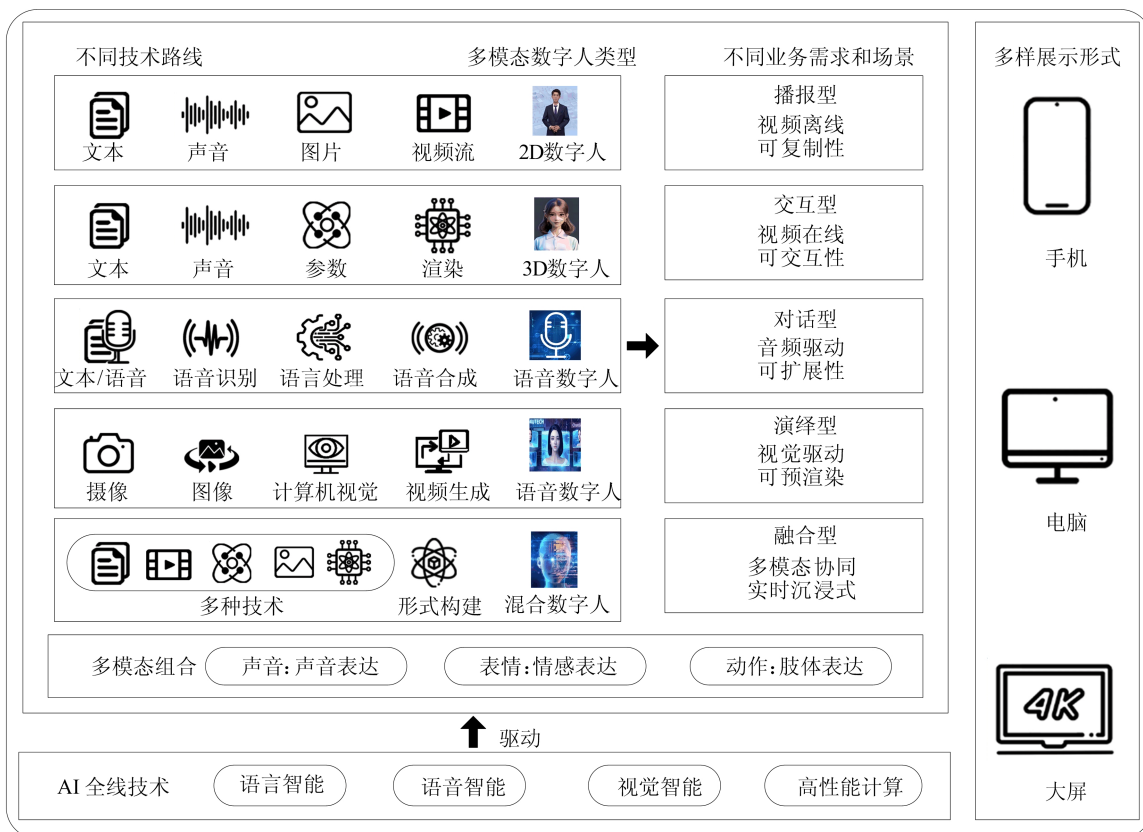


图 2 多模态数字人功能分类

总之,多模态交互的核心是语音、动作和视觉,这 3 个维度的协同工作能够实现更自然、高效的交互^[9],实现“拟人化、场景化、文化化”的效果。

1.2.1 语音识别理论模型

语音交互是多模态数字人实现自然沟通的核心,在声学建模层面,技术经历了从传统高斯混合模型-隐马尔可夫模型(Gaussian mixture model-hidden Markov model, GMM-HMM)到卷积增强型融合架构(convolution-augmented transformer, Conformer)的演进。传统 GMM-HMM 模型通过隐马尔可夫模型描述语音信号的时序状态转移,结合高斯混合模型对各状态的观测值概率进行建模,但存在忽视协同发音影响的缺陷;Conformer 则融合卷积神经网络的局部特征提取能力与转换器的全局依赖建模能力,既能精准捕捉特定场景下的韵律细节(如特定频段集中的语音信号),又能有效解析语义逻辑。

特征提取将语音信号转换为机器学习模型可处理的特征向量,其中 MFCC 是关键技术,转换公式为:

$$f_{\text{mel}}(f) = 2595 \times \lg\left(1 + \frac{f}{700}\right) \quad (1)$$

其中: f 为语音信号实际频率; f_{mel} 为转换后的梅尔频率。式(1)的核心在于模拟人耳听觉特性,通过梅尔频率与实际频率的非线性映射来优化特征表达。例如在黄梅戏智能数字人中,“平词”唱腔主要分布在 200~800 Hz 频段,而“花腔”分布在 800~1 500 Hz 频段。MFCC 通过对数变换强化了这两个关键频段的区分度,既保留了“平词”唱腔的中低频细节,又突出了“花腔”特有的高频装饰音,使提取的声学特征更符合人类聆听黄梅戏时的感知习惯。为后续戏曲场景下的语音交互提供理论与特征提取基础。

1.2.2 动作交互理论模型

动作交互的核心技术路径为“捕捉—建模—响应”,其理论基础是将特定场景下的动作转化为可量化的运动学参数。通过 8 摄像头 OptiTrack 光学动作捕捉系统,在动作执行体关节处粘贴反光标记点,记录程式化动作的三维坐标轨迹,获取关节角、运动速度等原始数据;同时辅以惯性动作捕捉传感器,弥补光学捕捉在遮挡场景下的数据缺失,确保复杂动作的数据完整性。

在动作建模阶段,需建立场景化动作向量库,可用公式表示为:

$$A = \{a_1, \dots, a_i, \dots, a_k\} \quad (2)$$

其中: A 为黄梅戏程式化动作库; a_i 为动作库中第 i 个动作单元,包含动作名称(如“单指”“双指”“翻身甩袖”等)、关节运动轨迹及持续时长等属性; k 为动作库中动作单元的总数量。

针对黄梅戏程式化动作的非标准化执行特点(不同演员的动作幅度、节奏存在差异),动态时间规整(dynamic time warping, DTW)算法可实现动作序列的非线性对齐,有效解决用户模仿与标准模板之间的时间轴偏差问题,为戏曲动作的精准匹配与交互响应提供核心算法支撑。

1.2.3 视觉交互理论模型

视觉交互的核心是通过面部特征捕捉与表情识别,实现数字人与用户的情感共鸣,其技术体系涵盖人脸检测-特征定位-表情分类这 3 个阶段。人脸检测阶段采用自适应提升算法 AdaBoost 优化的 Haar-like 特征,通过“边缘、线性、中心环绕”3 类矩形特征,快速筛选图像中的人脸区域,检测准确率需达 98% 以上,确保复杂背景下(如戏曲舞台灯光)的有效识别。

特征定位阶段需兼顾精度与实时性。传统主动形状模型(active shape model, ASM)通过构建平均面部轮廓模板,迭代调整 16 个特征点(眼角、嘴角等)与图像边缘的贴合度,适用于静态表情捕捉;而主动外观模型(active appearance model, AAM)进一步融合形状与纹理信息,可定位 68 个精细特征点,但计算量较大;因此采用高分辨率网络(high-resolution network, HRNet)作为核心方案,通过多分辨率特征分支保留眉毛、瞳孔等小区域细节,端到端回归特征点坐标。动态捕捉则通过光流法计算帧间位移,公式为:

$$\Delta(x, y, t) = \frac{I(x + \Delta x, y + \Delta y, t + 1) - I(x, y, t)}{\sqrt{\Delta x^2 + \Delta y^2}} \quad (3)$$

其中: x, y 为图像像素在二维平面的横、纵坐标(即像素在画面中的水平与垂直位置); t 为视频帧的时间序列; I 为对应坐标与时间帧下的像素灰度值; $\Delta x, \Delta y$ 分别为围殴相邻帧间像素在横、纵方向的位移量。式(3)通过计算像素灰度值随坐标和时间的变化,实时捕捉用户面部因听唱段共情产生的动态表情变化(如嘴角微动、眼角收缩),为后续表情分析提供连续的帧间运动数据。

而表情分类环节需适配黄梅戏表演中的复合情感,优先通过几何特征量化戏曲特有的表情状态,核心采用嘴角-鼻翼距离计算公式,即

$$d = \sqrt{(u_j - u_i)^2 + (v_j - v_i)^2} \quad (4)$$

其中: u_i 、 v_i 为嘴角特征点坐标; u_j 、 v_j 为鼻翼特征点坐标; d 为两点间欧氏距离。通过该公式可量化“羞涩”(嘴角上扬,使距离减小)、“悲愤”(嘴角下撇,使距离增大)等表情特征,结合 HRNet 定位的特征点数据,为数字人提供表情反馈的基础决策依据,最终实现视觉交互的情感化升级。

1.3 文化数字化中的多模态设计

在数字博物馆中,多模态设计整合了多种符号资源,如文字、图像、音频、视频等,能够更全面、生动地呈现文化内容,通过文字介绍、高清文物图片、三维(3D)模型展示、语音讲解以及相关历史场景的视频重现等多种模态的结合,让观众可以全方位地了解文物背后的文化内涵^[10]。数字动漫通过角色形象的视觉设计、故事情节文字脚本、音效与配乐的配合等多种模态元素的协同作用,打造出具有独特魅力和文化内涵的动漫作品,不仅在国内受到欢迎,还能更好地走向国际市场,传播中国文化^[11]。在一些历史文化遗址的数字重现中,在虚拟现实(virtual reality, VR)或增强现实(augmented reality, AR)技术构建的数字文化空间中,用户可以通过视觉、听觉、触觉等多种感官与文化内容进行交互,可以身临其境地感受古代建筑的宏伟、聆听历史故事的讲解,甚至通过手势操作与虚拟场景中的元素进行互动,从而获得全新的文化体验。

文化数字化中的多模态设计强调通过整合多

种模态元素(如文字、图像、音频、视频、交互等)来更全面、生动地呈现文化内容,提升文化体验的丰富性和感染力,增强文化的传播力和影响力^[12]。多模态设计能够满足不同用户的需求和偏好,使文化数字化产品更具吸引力和互动性。在实际应用中,多模态设计原则被广泛应用于数字博物馆、数字文化创意产品、数字文化空间构建等多个领域,通过多种模态的协同作用,为用户带来全新的文化体验,推动文化的传承与发展。

2 多模态交互设计与虚拟交互实现

2.1 黄梅戏数字人的多模态交互设计

基于认知神经学理论与多模态交互技术,本研究构建以黄梅戏经典角色冯素珍为核心载体的多模态交互设计框架。该框架依托大模型技术底座,深度融合黄梅戏专业知识图谱,实现自由主题对话、剧情讲述、才艺展示等多样化功能,整体架构如图 3 所示。

交互框架设计理念遵循“文化真实性-技术可行性-用户参与性”三位一体原则:文化真实性体现在数字人的戏服装扮、程式化动作、方言唱腔等要素严格遵循黄梅戏艺术规范;技术可行性通过融合面部捕捉、语音识别、3D 实时渲染等技术保障交互流畅稳定;用户参与性借助即兴对唱、动作模仿、剧情共创等互动设计,激发用户对传统文化的兴趣。



图3 黄梅戏智能数字人多模态交互设计框架

该交互框架通过语音、动作、视觉多模态信息的协同处理与反馈,构建自然流畅的交互体验。系统既能准确理解用户的语音指令与动作意图,又能通过精准的面部表情和肢体动作传递黄梅戏特有的艺术感染力,实现技术赋能与文化遗产的有机统一,体现了多模态技术在非遗数字化保护

与创新中的应用价值。

2.1.1 多模态输入设计

多模态输入设计深度融合了语音、动作、视觉及触觉多通道交互方式,其技术实现建立在严谨的理论模型之上。

语音输入模块支持异构语言(方言-普通话)

混合识别,结合场景化语料库,直接采用1.2.1节所述的MFCC特征提取技术完成声学特征提取,生成的高区分度特征向量为连接主义时间分类(connectionist temporal classification, CTC)端到端声学建模提供输入,实现对黄梅戏唱腔词汇、点播指令、文化咨询等用户意图的精准捕捉,同时保留安庆方言与黄梅戏唱腔的频段特征。

视觉输入模块通过摄像头捕捉用户手势与体态,基于Haar-like特征与AdaBoost算法的级联结构实现鲁棒的人脸检测,其中弱分类器权重计算与迭代更新样本权重的机制确保了在复杂戏曲舞台灯光下仍能维持高检测率。该模块进一步结合ASM完成面部关键点定位,并同步识别用户表情与服饰颜色,以动态调整数字人的互动风格。

动作输入模块利用穿戴式传感器捕捉用户手

势与体态,与1.2.2节所述的黄梅戏场景化动作向量库进行匹配,通过DTW算法完成用户动作与程式化模板的非线性对齐,快速识别用户的动作交互意图,为多模态融合提供动作维度的特征输入。触觉输入模块结合VR手柄或触觉手套,模拟戏曲道具的交互反馈,通过多通道感知融合增强沉浸感。

各模态信息通过注意力机制融合策略实现深度整合,形成统一的交互理解。

2.1.2 多模态处理与融合

多模态处理与融合是实现智能数字人自然交互与情感共鸣的核心环节,其目标是通过整合语音、情绪、视觉信息以提升交互效率与用户体验,契合黄梅戏文化传播需求^[13]。黄梅戏智能数字人情感化设计模型如图4所示。

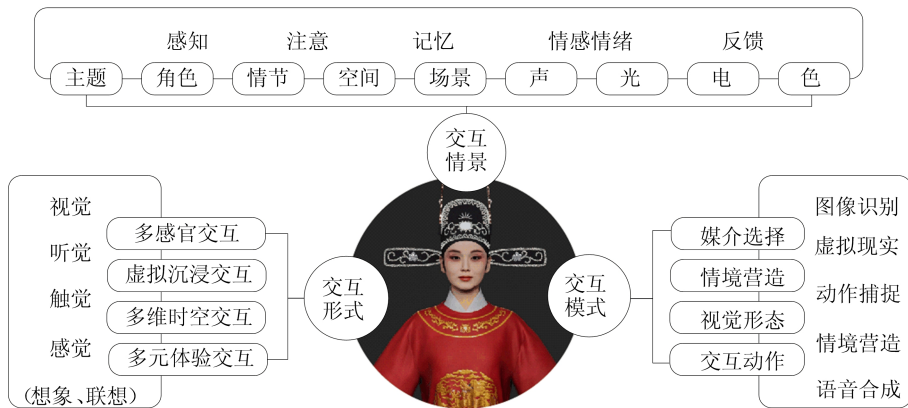


图4 黄梅戏智能数字人情感化设计模型

本系统采用基于注意力机制的多模态融合策略,为来自不同模态的特征向量分配合适的权重。其核心融合公式定义如下:

$$z = \sum_{i=1}^M \alpha_i f_i \quad (5)$$

其中: f_i 为来自第*i*个模态(如语音、视觉、动作)的特征向量; α_i 为由注意力网络计算得到的该模态的动态权重,权重之和为1; z 为融合后的统一语义表征向量; M 为模态总数。

本研究采用交叉注意力机制实现模态权重的动态计算,以语音模态为基础查询,视觉、动作模态特征为键与值,通过注意力网络学习不同交互场景下的模态贡献度。

注意力权重的计算过程为:先对各模态特征向量进行层归一化处理,通过线性层将其映射至同一特征空间,再通过softmax函数归一化得到权重值。交叉注意力机制可实现模态间的特征关

联挖掘,例如:在戏曲表演交互场景中,自动提升动作模态的权重占比;在语音指令交互场景中,强化语音模态的特征贡献。

在交互过程中,系统并行处理来自不同通道的信息。基于自然语言处理技术,数字人解析语音与文本输入的语义,提取用户意图与情感倾向,结合场景化知识图谱准确理解戏曲相关词汇与典故,确保回应专业性^[14]。与此同时,视觉通道捕捉的用户表情与手势以及动作通道识别的程式化动作意图被一同送入融合中心。该系统利用式(5)所描述的融合机制,将不同模态的特征向量在统一的语义空间中进行加权整合。如当用户同时发出“唱《女驸马》选段”语音指令和“做出兰花指”手势指令时,多模态信息的协同作用能使系统更精准地锁定“戏曲表演”这一交互场景,从而触发数字人同步启动唱腔输出与对应的身段动作,实现从单向指令接收到双向情境化交互的跨越。

2.1.3 多模态输出设计

多模态输出设计旨在通过语音、动作、视觉的协同呈现,验证多模态输出的协同性与场景的适配性。

语音输出采用安庆方言与普通话的自适应切换机制,唱段严格遵循“平词”的叙事性韵律与“花腔”的轻快性节奏,念白语气则紧密结合角色情绪,并同步输出唱词字幕以适配不同用户的认知需求。动作输出以黄梅戏程式化规范为核心,“云手”“台步”“兰花指”等经典身段通过精细的骨骼绑定与蒙皮技术得以流畅再现,其动作幅度与节奏严格对标戏曲规范。视觉输出则聚焦于文化符号的高保真还原,数字人的戏服严格参照《女驸马》经典造型,通过高精度建模呈现细节;其面部表情基于实时交互反馈动态生成,通过参数化控制实现如式(4)所量化的“羞涩”“悲愤”等符合剧情的复合情绪,从而极大地增强了演绎的情感共鸣与用户沉浸感。

此外,系统支持“剧情共创”的个性化输出,允许用户通过语音或文本修改剧情片段,数字人可据此实时调整台词、动作与表情,实现“参与式”戏

曲体验。

2.2 基于冯素珍原型的数字人开发实践

2.2.1 黄梅戏经典元素的数字化提取

在冯素珍原型数字人的开发中,对黄梅戏经典元素进行系统性数字化提取是保障文化真实性的首要步骤。此项工作围绕服饰、唱腔、身段、道具四大核心元素展开。

服饰数字化采用高分辨率图像采集与三维扫描技术,以高精度捕捉戏服的纹理、形状及动态褶皱形态,并对蟒袍“团龙”刺绣、官帽“顶珠”等细节进行数据标注。唱腔提取则聚焦于“平词”“花腔”两大板式,由国家一级演员录制标准唱段,并依据旋律类型、节奏模式及音高范围进行结构化分类标注,基于 MFCC 特征提取结果构建可查询、可分析的黄梅戏专属唱腔数据库,为数字人语音输出的唱腔还原提供标准化数字资产。

2.2.2 黄梅戏数字人模型的制作流程

黄梅戏数字人模型的制作遵循“原画—粗模—高模—低模—贴图—烘焙—制作UV—骨骼—绑定蒙皮—应用”的标准化流程,在精准还原文化特征的同时兼顾多终端平台的渲染效能,完整制作流程如图 5 所示。

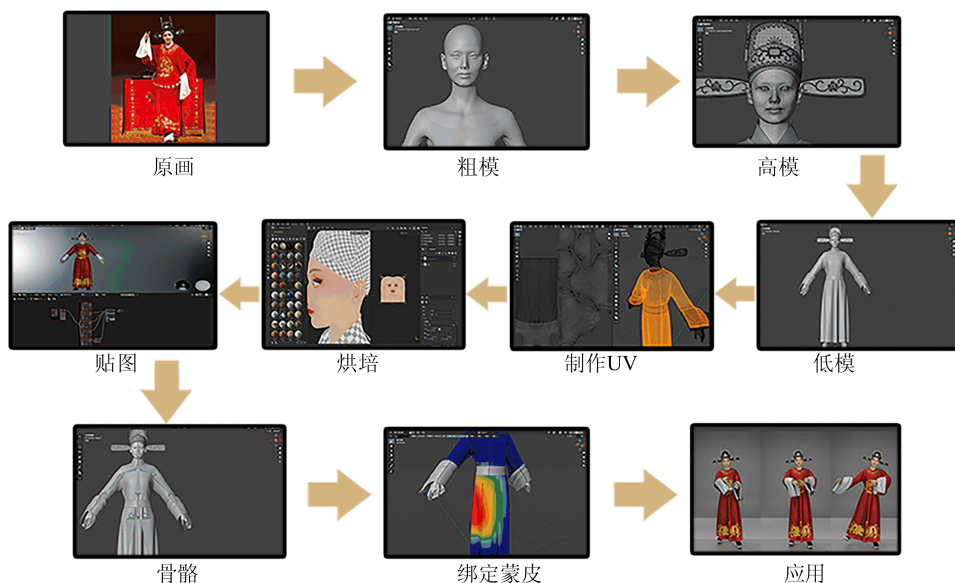


图 5 黄梅戏数字人模型的制作流程

制作伊始,团队基于《女驸马》的舞台造型完成冯素珍的原画设计,精准定义其面部特征以及戏服与发饰细节。随后通过粗模搭建确立基础框架,并利用细分曲面技术构建高精度模型,精细雕刻出戏服刺绣的针脚与面部肌理。为适配移动设备与 VR 等平台的实时渲染要求,通过烘焙法线贴图与位移贴图等技术,将高模的视觉细节完整

映射至游戏级低模上。在贴图阶段,采用基于物理的渲染材质系统来表现戏服丝绸的光泽感与皮肤的次表面散射效果,增强了模型的视觉真实性。骨骼绑定阶段设置了大量控制节点,重点优化颈部与手部关节,以确保“转头甩袖”的流畅性与“兰花指”的灵活变换。最后,通过手工权重调整与一系列典型动作的测试校准,有效避免模型在动态

表演中变形失真。最终交付了一个兼具视觉真实性与交互适配性的高保真黄梅戏数字人模型。

2.2.3 虚实融合的交互场景设计

在开发基于冯素珍原型的黄梅戏数字人时,本研究设计了一系列虚实融合的交互场景,旨在通过创新的交互体验,让用户深入感受黄梅戏的艺术魅力和文化内涵。该研究利用VR技术,创建了一个沉浸式的虚拟舞台环境。用户可以通过VR设备,观看冯素珍数字人表演《女驸马》等经典黄梅戏剧目。通过高保真的视觉和音频效果,用户仿佛置身于真实的戏曲舞台,近距离感受黄梅戏的艺术魅力。

多模态交互是虚实融合交互场景的核心。系统整合了语音、手势、表情等多种交互模态,为用户提供自然流畅的交互体验。例如:用户可以通过语音指令,让冯素珍数字人表演特定的唱段;通过手势动作,控制数字人的动作;通过表情识别,与数字人进行情感交流。这种多模态交互设计,不仅提高了交互的自然性和效率,也大大增强了用户的沉浸感和参与感。

黄梅戏数字人交互视频效果如图6所示。



图6 黄梅戏数字人交互视频效果

2.2.4 动态表情与动作捕捉系统的精准适配

为确保数字人的表演能够真实还原黄梅戏的艺术风格,采用了高精度的动态表情与动作捕捉系统,并对数据进行了针对性的优化与适配。在动作捕捉方面,采用CMAvatar软件协同光学与惯性动作捕捉系统。光学系统用于高精度记录“水袖挥舞”等大范围运动轨迹;惯性系统则通过穿戴式传感器弥补光学系统在遮挡场景下的数据缺失。两者融合为数字人动作库提供了高精度的数据源。动作捕捉画面如图7所示。

表情捕捉则采用高分辨率摄像头结合HR-Net定位面部关键点,精准记录演员表情的细微肌肉运动数据,并将这些数据与三维模型的

Blendshape驱动系统进行适配。黄梅戏演员驱动画面如图8所示。

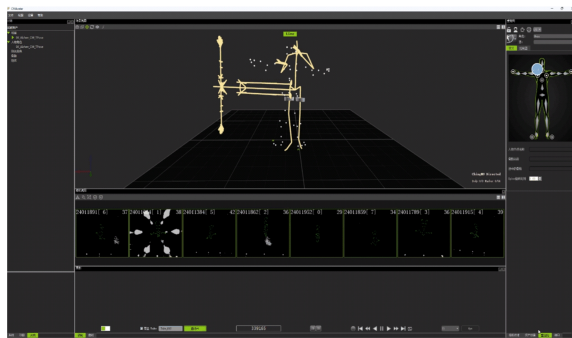


图7 动作捕捉动作画面



图8 黄梅戏演员驱动画面

在此基础上,该系统依据黄梅戏的身段与表情特点对捕捉数据进行了艺术化调整与优化,例如依据专业规范,利用公式评估并微调“兰花指”的手指弯曲角度与“台步”的步幅,使其误差控制在预设阈值内。同时,通过专业演员的测试反馈,严格校准了动作、表情与唱词情感的匹配度。

在黄梅戏身段动作优化阶段,该数字人通过计算原始捕捉数据与黄梅戏艺术规范模板之间的误差,进行针对性修正,其调整量量化评估公式为:

$$E_{\text{RMS}} = \sqrt{\frac{1}{N} \sum_{j=1}^N (p_j^{\text{captured}} - p_j^{\text{standard}})^2} \quad (6)$$

其中: E_{RMS} 为均方根误差,用于衡量捕捉数据与标准模板的总体偏差; p_j^{captured} 和 p_j^{standard} 分别为第 j 个关节在捕捉数据和标准模板中的位置坐标; N 为关节总数。

在根据计算结果进行调整后,可以再次计算均方根误差。若其值显著减小,则证明优化是有效的,从而形成了一个“捕捉—计算均方根误差—优化—再计算均方根误差验证”的闭环质量控制流程。最终提升数字人的表现力与艺术感染力。黄梅戏演员手部驱动测试画面如图9所示。

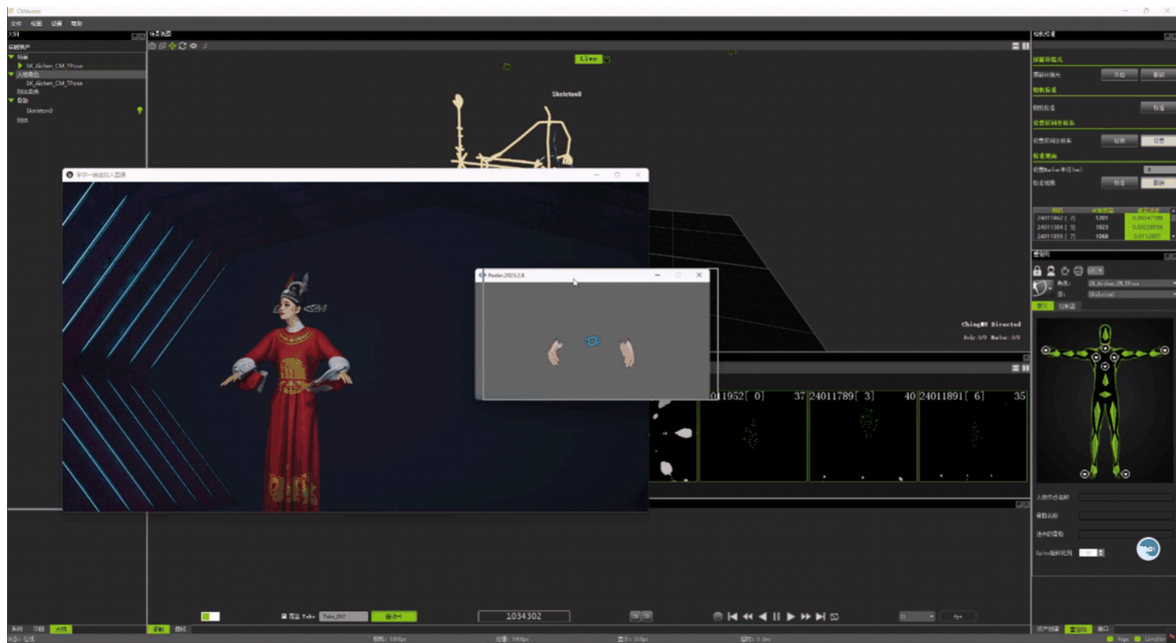


图 9 黄梅戏演员手部驱动测试画面

3 系统用户测试与效果分析

为客观验证本系统对解决非遗戏曲数字化传播交互缺失、青年群体深度认知不足等问题的实际效能,呼应“实现黄梅戏活态数字化传承”的研究目标,本研究开展规模性用户测试,结合问卷调查评估系统的交互体验、文化传播效果及用户接受度,以此验证所提方案的有效性 with 优越性。

本次用户测试共回收有效问卷 107 份,样本人口统计学特征见表 1 所列。样本以 18~25 岁青年群体为主(占比 66.4%),与研究“非遗戏曲青年化传播”的核心目标高度契合;女性用户占比 60.7%,男性用户占比 39.3%,整体样本结构具有良好的代表性,能够有效反映目标用户群体的真实体验与反馈。

表 1 样本人口统计学特征

统计项目	分类	占比/%
性别	男	39.3
	女	60.7
年龄/岁	18~25	66.4
	26~55	29.9
	≥56	3.7

本研究采用李克特 7 级量表(1 表示非常不同意,7 表示非常同意)设计用户体验问卷,围绕本文核心研究内容,选取多模态控制、认知参与、数字人形象 3 个关键维度,共 6 道题项,重点验证系统多模态交互技术的有效性、黄梅戏文化传播

的实际效果及数字人设计的体验感。测试基于系统搭建的基础交互场景开展,确保用户能够全面体验系统功能并给出真实反馈。

各维度及题项的均值±标准差统计结果见表 2 所列。由表 2 可知,整体来看,所有维度均值均高于 7 分量表中间值(4 分),表明用户对系统的交互体验、文化传播效果及数字人设计均持积极认可态度。

表 2 黄梅戏智能数字人系统用户测试核心维度统计结果

评价维度	题项	均值±标准差	整体维度均值±标准差
多模态控制	能否语音交互	4.83±0.96	4.84±0.88
	能否手势交互	4.85±0.93	
认知参与	能否提升知识	5.25±1.21	5.31±1.06
	能否获取新内容	5.36±1.05	
数字人形象	是否行为自然	4.90±0.99	4.92±0.91
	是否外貌舒适	4.83±0.96	

在认知参与维度下,“能否获取新内容”(5.36±1.05)与“能否提升知识”(5.25±1.21)得分均处于较高水平,直接验证了系统对青年群体黄梅戏深度认知的提升效果,有效回应了“非遗戏曲青年群体关注度高但深入认知有限”的核心问题。多模态控制维度表现良好(整体均值 4.84±0.88),语音交互与手势交互得分接近且标准差较小,用户体验一致性高,验证了本研究多模态技术方案的有效性。数字人形象维度表现达标(整体均值 4.92±0.91),行为自然度与外貌舒适度得

分均衡,说明数字人在黄梅戏角色形象设计、动作还原方面基本满足用户期望,为文化传播提供了良好的视觉与交互载体。

4 结 论

本研究利用多模态人机交互技术,设计并开发了一套智能数字人系统。该系统通过深度融合语音、视觉与动作三大模态,实现了从多通道感知、上下文理解到情感化输出的全链路交互。在语音维度,系统整合 MFCC 特征提取与 CTC 端到端建模技术,有效支持了复杂声学环境下的连续语音识别与情感韵律分析;在视觉维度,采用基于 HRNet 的高分辨率特征保持网络,显著提升了人脸检测、微表情识别与实时追踪的精度和鲁棒性;在动作维度上,融合光学与惯性动作捕捉系统,并引入 DTW 算法进行动作序列的弹性匹配,实现了高保真、可复用的动作库构建与响应机制。研究进一步提出了一种基于注意力权重的多模态信息融合策略,使系统能够依据交互场景动态调整不同模态的贡献度,从而在黄梅戏角色互动等应用实例中展现出良好的自然性与情境适应性。该技术框架为构建具有深度感知与拟人反馈能力的智能数字人提供了系统性方法论。

研究表明,本文所提出的多模态交互架构与数字人设计范式,不仅能够服务于黄梅戏等特定文化遗产的数字化保存与创新传播,其模块化、可扩展的特性也使其具备向更广泛领域迁移的潜力,例如,可应用于虚拟教学、数字娱乐、智能导览、远程协作等多个场景,推动人机交互向更智能、更沉浸的方向演进。

值得关注的是,随着 AIGC 技术的快速发展,高仿真虚拟数字人的生成效率与逼真度将得到进一步提升^[15],为本框架的进一步优化与落地提供了新的技术支撑。未来工作将着力于跨模态预训练模型的集成、轻量化实时渲染方案的探索,以及开放域对话与个性化交互能力的增强;同时结合非遗数字化“从呈现到体验”的最新发展趋势,拓展移动端、元宇宙^[16]等多平台适配场景,降低沉浸式交互的设备门槛,以期推动智能数字人技术在黄梅戏等非遗传播、文旅体验等更多元场景中的普适应用与持续创新^[17-18]。

[参 考 文 献]

- [1] 翟雪松,吴庭辉,李翠欣,等. 数字人教育应用的演进、趋势与挑战[J]. 现代远程教育研究,2023,35(6):41-50.
- [2] 谢景伟,施薇. 智能数字人在对外汉语教学中的应用效果探索:个性化学习与跨文化交互实践[J]. 新课程教学(电子版),2025(17):162-164.
- [3] 朱沁怡. 数字技术赋能我国传统戏曲的现状、困囿与破局之路[J]. 戏剧之家,2025(13):21-23.
- [4] 叶建波. 新时代青少年戏曲传承的意义、困境与路径研究[J]. 潇湘文化研究,2025(2):58-63.
- [5] 常宏. 虚拟数字人在非遗传承发展中的应用[J]. 人民论坛,2024(2):103-105.
- [6] 王国庆,裴云强,杨阳,等. 多模可信交互:从多模态信息融合到人-机器人-数字人三位一体式交互模型[J]. 中国科学:信息科学,2024,54(4):872-892.
- [7] 王党校,郑一磊,李腾,等. 面向人类智能增强的多模态人机交互[J]. 中国科学:信息科学,2018,48(4):449-465.
- [8] 田媛,聂昕晓,刘海诺,等. 混合课堂学生认知投入的多模态量化机制与干预[J]. 心理科学进展,2024,32(10):1578-1592.
- [9] 陶建华,巫英才,喻纯,等. 多模态人机交互综述[J]. 中国图象图形学报,2022,27(6):1956-1987.
- [10] 盛榕,杨健,姚连军. 多模态视域下红色文化数字传播路径研究:以金寨为例[J]. 安徽工业大学学报(社会科学版),2024,41(4):39-42.
- [11] 黄允怡. 多模态隐喻理论下的广府文化创意产品设计研究[D]. 杭州:中国美术学院,2021:63-65.
- [12] 刘佳. 数字化赋能非遗对外传播多模态语料库平台构建研究[J]. 上海翻译,2023(6):34-39.
- [13] 许海振,梁成福,王谦. 基于多模态对话技术下的电商导购数字人研发及应用[C]//中国新闻技术工作者联合会. 中国新闻技术工作者联合会2024年学术年会论文集. 济南:山东舜网传媒股份有限公司,2024:159-164.
- [14] 王妍. 面向数字人交互的视觉跨模态学习研究[D]. 合肥:安徽大学,2023:1-4.
- [15] 黄慎泽,吕欣. AIGC 技术驱动下的高仿真虚拟数字人设计研究[J]. 艺术设计研究,2024(6):95-101.
- [16] 朱永琼,宋章通,方浩. “文旅元宇宙”中虚拟数字人的应用[J]. 传媒,2023(3):55-57.
- [17] 林乐怡,朱津涛,钟凯茜. AI 虚拟数字人技术下的非遗文化创新传播研究:以宋韵文化“谷小雨数字人”为例[J]. 中原文化与旅游,2024(19):40-42.
- [18] 史玉平. 融媒体视域下曲艺类非遗文化的数字化传播路径探析:以凤阳花鼓为例[J]. 社会科学前沿,2023,12(9):4979-4984.

(责任编辑 胡亚敏)

[1] 翟雪松,吴庭辉,李翠欣,等. 数字人教育应用的演进、趋势