# 基于深度学习的数字人驱动技术研究与软件实现

段　超[1]　张杭凯[2*]
1.杭州硅基聚欢科技有限公司　浙江杭州　310000
2.杭州伏特机器人科技有限公司　浙江杭州　310000

**摘　要：**随着人工智能和深度学习技术的快速发展，数字人作为一种融合计算机图形学、语音合成、自然语言处理与行为驱动等多学科技术的智能体，正广泛应用于虚拟客服、虚拟主播、在线教育等领域。本文围绕基于深度学习的数字人驱动技术展开系统研究，深入分析了当前主流的面部表情生成、语音驱动唇形同步、动作捕捉与行为建模等关键技术，并基于深度神经网络构建了完整的数字人驱动系统软件架构。本文首先阐述了数字人的定义与分类，分析其在不同场景下的应用需求，介绍了开发一套可实时交互的数字人软件系统的框架，为推动虚拟人技术的实用化和智能化发展提供了理论基础与技术支撑。

**关键词：**深度学习；数字人驱动；实时交互；神经网络

## 序言

随着元宇宙概念的提出与虚拟现实技术的持续演进，数字人作为一种具备人类外貌、语音和行为能力的智能虚拟形象，日益受到科技界与产业界的高度关注。传统的数字人驱动技术多依赖于规则系统和模板库，存在自然度不高、响应滞后等问题。近年来，深度学习的崛起为数字人驱动带来了革命性的突破，使其在图像识别、动作生成和多模态交互等方面实现了质的飞跃。尤其是生成对抗网络（GAN）、卷积神经网络（CNN）、循环神经网络（RNN）和变换器模型（Transformer）等的应用，为数字人从“能动”向“智能”发展提供了强有力的支持。

## 一、数字人驱动技术概述

### （一）数字人定义与分类

数字人是以数字化技术构建的具有人类形象与行为特征的虚拟角色，广泛应用于社交媒体、虚拟演艺、在线教育、数字营销等领域。按其实现方式与智能程度的不同，数字人可分为三类，第一类是静态数字人，主要用于图像渲染和视觉呈现，具备高度写实的外观但缺乏交互能力[1]。第二类是驱动型数字人，具备一定的语言、动作响应能力，依赖脚本或行为模型进行驱动[2]。第三类是智能型数字人，基于人工智能与深度学习技术，能实现多模态感知、语义理解与自然交互[3]。从形象呈现上也可分为写实型数字人（如数字明星、虚拟偶像）和卡通型数字人（如教育动画角色）[4]。在实际开发中，数字人构建需综合考虑技术可行性、成本控制与用户接受度，因此对其进行合理分类与定位，有助于提升数字人系统的实用性与交互性，也为后续技术实现提供明确的开发路径[5]。

### （二）数字人驱动技术

数字人驱动技术是指实现虚拟人面部表情、语音同步、身体动作与行为逻辑等功能的核心技术体系，主要包括语音驱动技术、表情与动作捕捉技术、自然语言理解与行为控制技术。当前，深度学习技术的引入极大提升了驱动系统的智能性与实时性。例如，在唇形同步方面，基于语音到图像的深度模型（如Wav2Lip）能够实现语音驱动的高拟合度口型生成；在动作生成上，采用LSTM、Transformer等神经网络可以根据语义指令预测肢体动作序列，实现自然流畅的动作反应；在人脸表情方面，GAN与3D Morphable Model的结合使得表情生成更具真实感与可控性。此外，情感识别、知识图谱驱动的语义交互与强化学习机制的引入，也使数字人在对话理解、情绪表达和个性塑造方面更加智能。

## 二、核心算法与模型设计

### （一）输入端处理模块

输入端处理模块是数字人驱动系统的首要环节，其

**作者简介：**
段超，男，汉，浙江省杭州市，1980年10月30日，硕士，软件开发+数字人；
张杭凯，男，汉，浙江省绍兴市，1989年10月22日，本科，机电类软件开发相关。

主要任务是从多种数据源中获取并预处理驱动信息，为后续的生成模型提供准确、稳定的输入信号。常见的输入形式包括语音、文本、图像以及视频信号，其中语音和文本通常用于驱动口型和语义动作生成，图像和视频则用于捕捉用户的面部表情、头部姿态及肢体动作。为了提高输入的准确性与鲁棒性，该模块首先需要通过深度学习模型进行信号清洗与特征提取。例如，语音信号经过声纹识别和情感识别模型处理后，提取出说话人特征与情绪状态；视频信号则通过姿态估计和面部关键点检测网络（如OpenPose或MediaPipe）实现对人脸表情与动作的精准捕捉。此外，为了适应不同设备与输入场景，输入模块还需具备一定的容错性与实时处理能力，常采用轻量化模型结构并结合边缘计算部署策略。最终，处理后的数据被统一编码为高维特征向量，输入至动作或表情生成模型中，实现端到端的驱动效果。

### （二）动作/表情生成模型

动作与表情生成模型是数字人驱动系统的核心，其目标是将输入端提取的语音、文本或图像特征映射为自然、协调的人体动作和面部表情。该模型通常采用基于时序建模的深度学习框架，如循环神经网络（RNN）、长短期记忆网络（LSTM）或更先进的Transformer结构，以捕捉驱动信号中的时间动态特征。在表情生成方面，模型需对口型、眼神、眉毛等细微动作进行精准控制，因此通常结合三维人脸重建与面部肌肉建模方法，通过显式地建模面部动作单元（Action Units）提升表情的真实度。动作生成部分则侧重于肢体的连贯性与物理合理性，一般采用骨骼动画参数作为输出，通过人体运动模型（如SMPL）进行约束和还原。为增强生成内容的自然性与多样性，常引入生成对抗网络（GAN）或变分自编码器（VAE）结构，对生成样本进行质量控制。通过训练大规模人机交互数据集，模型逐步学习到语义与动作之间的映射关系，从而实现数字人在语境下的自然表现。

### （三）多模态融合机制

多模态融合机制在数字人驱动技术中扮演着桥梁作用，主要负责将来自不同输入通道（如语音、文本、图像等）的特征信息进行统一建模和协调，以生成具有一致性和表现力的动作与表情。由于各模态间存在异构性和不同的时空分布，融合机制的设计需解决模态对齐、权重分配与语义统一等问题。当前主流的融合策略可分为早期融合、晚期融合和中间融合三类，其中中间融合方式因其在保持模态独立性与实现深层语义整合之间的良好平衡，应用较为广泛。该机制一般采用多模态注意力网络（Multi-modal Attention Network）或跨模态Transformer架构，实现对不同模态信息的加权整合。在实际应用中，例如在数字人口播场景中，系统需要同时理解语音语调中的情绪变化、文本中的语义内容以及图像中的面部动作，因此融合模块需具备对情绪、语义和视觉动作之间复杂关系的建模能力。融合后的统一特征被输入至生成模型中，指导数字人做出与语境一致的动态响应，从而提升系统的智能交互能力和人机交流自然性。

## 三、数字人驱动软件系统设计与实现

### （一）系统总体架构设计

数字人驱动软件系统总体架构基于模块化与层次化设计理念，主要包括数据输入层、深度学习模型层、驱动渲染层与前后端交互层。数据输入层负责采集用户输入的信息，包括音频、文本、图像等，通过预处理模块进行规范化与特征提取，为后续模型计算提供高质量输入。深度学习模型层是系统的核心，集成了多模态感知模型（如语音识别、情感识别、人脸关键点检测）与生成模型（如语音合成、面部动画生成），利用深度神经网络完成对数字人行为的智能驱动。驱动渲染层则负责将模型输出的指令转化为数字人形象的动态表现，包括面部表情、口型变化、姿态动作等，通过三维图形引擎实时渲染呈现。前后端交互层确保系统各模块之间数据传输的高效性与稳定性，采用WebSocket等实时通信机制，提升系统响应速度和用户体验。整个架构设计强调模块解耦、灵活扩展和系统性能优化，为数字人系统的实际部署与多场景应用提供了技术支撑。

**表1　虚拟数字人基础技术架构**

| 多模态感知模型 | 2D数字人 | 3D数字人 |
|---|---|---|
| 人物生成 | 无 | 人物建模等 |
| 人物表达 | | 语音生成、动画生成（驱动、渲染）等 |
| 合成显示 | | 终端显示技术 |
| 识别感知 | | 语音语义识别、人脸识别、动作识别等 |
| 分析决策 | | 知识库、对话管理等公众号·时空创意界 |

### （二）软件关键模块实现

在数字人驱动软件系统中，关键模块的实现直接关系到系统性能与智能程度。首先，语音识别与文本生成模块采用基于Transformer结构的预训练模型（如Wav2Vec2.0和GPT），实现自然语言的高精度识别与生成处理，支持用户语音指令的实时转化与语义理解。其次，情感识别模块利用多模态输入数据（音频、图像）构建

融合网络模型，准确识别用户语调、面部表情等情感信息，为驱动数字人做出相应的表情和语气变化提供依据。第三，动作与表情生成模块基于时序生成模型（如LSTM或Diffusion Model），通过输入文本或语音的情感标注驱动三维面部表情点和骨骼动作参数，实现与语义同步的拟人化表演。此外，三维模型绑定与渲染模块则依托Unity或Unreal Engine进行数字人骨骼绑定、皮肤蒙皮和高效渲染，确保视觉表现的真实性与流畅度。上述关键模块以高内聚、低耦合方式集成于系统中，通过统一接口与通信协议实现数据交互与功能协调，形成了完整的智能数字人驱动能力。

**（三）前后端交互设计**

为了保障数字人驱动系统的流畅运行与良好用户体验，前后端交互设计采用了分布式架构与实时通信机制。前端基于Vue.js构建，主要负责用户界面展示与交互逻辑处理，包括音视频采集、文本输入、驱动结果展示等功能，具备高度的响应性与可视化效果。后端采用Python语言与FastAPI框架开发，集成深度学习模型服务、数据处理逻辑及接口控制模块，支持高并发请求与异步处理能力。前后端之间的数据传输采用RESTful API与WebSocket结合的方式：静态资源和常规请求通过HTTP传输，而需要实时响应的语音识别、动作渲染控制等环节则依赖WebSocket进行双向通信，有效降低延迟，提高系统响应速度。此外，接口设计遵循标准化、模块化和可扩展性原则，确保系统易于后期迭代和功能拓展。为增强用户操作的沉浸感与互动性，系统还引入了状态同步机制与缓存机制，避免前后端状态不一致的问题。整体前后端交互设计有效支撑了数字人从输入到响应的完整链路，实现了“所说即所动”的智能驱动体验。图2为某单位数字人驱动软件系统。


图2　某单位数字人驱动软件系统

## 结语

综上所述，本论文围绕基于深度学习的数字人驱动技术展开研究，并实现了相关的软件系统，旨在探索虚拟人驱动技术的关键路径与实际应用。通过对面部表情识别、动作捕捉与语音驱动等核心技术的分析与实现，验证了深度学习算法在提升数字人自然度、交互性和实时性方面的优势。实践结果表明，结合卷积神经网络（CNN）、循环神经网络（RNN）以及Transformer等模型，可以有效提升数字人动作生成与语音表情同步的精度与流畅性。尽管在实时性能、资源消耗和多模态融合方面仍存在一定挑战，但本研究为数字人技术的进一步发展提供了技术基础与可行路径。

## 参考文献

[1]祝智庭，胡姣.教育数字化转型的本质探析与研究展望[J].中国电化教育，2022（4）：1-8，25.

[2]刘月霞，郭华.深度学习：走向核心素养（理论普及读本）[M].北京：教育科学出版社，2018.

[3]夏峰平.数字化赋能深度学习：数据驱动与个性化路径的融合[J].中学教学参考，2024（6）：28-31.

[4]齐勇，门泽木，解思源，成润泽.基于数字孪生和深度学习的风力与光伏发电预测方法研究[J].软件工程，2024，28（3）：57-63.

[5]润治，王瑞琪，刘继彦，等.基于CNN-Bi-LSTM功率预测的海岛综合能源系统优化调度[J].全球能源互联网，2023，6（1）：88-100.