

基于神经网络和虚幻引擎的数字人客服系统

豆子闻, 李文书

(浙江理工大学计算机科学与技术学院, 浙江 杭州 310018)

✉ 745305676@qq.com; charlie@zstu.edu.cn



摘要:为了改善传统智能客服在基于文本交互的情境下,用户体验较差的问题。基于人脸的三维重建赋予数字人真实面孔,同时结合虚幻引擎进行实时渲染,采用 Rasa(用于构建对话机器人的开源机器学习框架)针对不同场景进行中文问答训练,经由 TTS(Text To Speech)算法转化文本为音频,再通过唇形同步算法将音频映射为数字人面部变形权重,该权重最终用于驱动面部动画,实现了用户和高保真数字人语音对话的功能。经过在不同年龄段群体进行测试和统计,用户对系统的满意率高达 96%,为提升智能客服的体验提供了优化方案。

关键词:三维重建;中文问答;数字人;唇形同步;语音合成;虚幻引擎

中图分类号:TP391 **文献标志码:**A

Digital Human Customer Service System Based on Neural Networks and Unreal Engine

DOU Ziwen, LI Wenshu

(School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

✉ 745305676@qq.com; charlie@zstu.edu.cn

Abstract: In order to improve the problem of poor user experience in traditional intelligent customer service based on text interaction, this paper proposes to endow digital humans with real faces based on 3D reconstruction and use virtual engine for real-time rendering. Rasa (an open-source machine learning framework for building dialogue robots) is used to train Chinese Q & A for different scenarios. Text is transformed into audio by TTS (Text To Speech) algorithm, and then the audio is mapped to digital human facial deformation weights by lip synchronization algorithm. This weight is ultimately used to drive facial animation and realize the function of voice conversation between users and high-fidelity digital humans. After testing and statistics in different age groups, the user satisfaction rate is as high as 96%, providing an optimization solution for improving the experience of intelligent customer service.

Key words: 3D reconstruction; Chinese Q & A; digital human; lip synchronization; speech synthesis; unreal engine

0 引言(Introduction)

在数字化时代,客户服务经历了重大变革,越来越多的消费者倾向于通过数字方式与企业沟通,而企业也在探寻更加可

靠且高效的客户服务系统^[1]。随着计算机算力的不断增强,神经网络参数量的快速提升使得智能客服更加“聪明”、完善。人工智能及时的应用彻底改变了客户服务行业,而三维重建以及

图形渲染的发展,也使得数字人更加逼真,其皮肤纹理清晰可见。数字人和客服系统相结合是数字客服这一领域最前沿的发展趋势,而在系统开发中结合使用神经网络和虚幻引擎是一种独特的方法,在以前的研究中未被广泛探索^[2]。

本文主要研究基于神经网络和虚幻引擎的数字人客服系统的开发,该系统旨在为客户创造沉浸式的互动体验,提高他们对产品或服务提供企业的满意度和忠诚度。通过模拟类人交互,数字人客服系统可以对客户的查询和投诉提供有效且个性化的响应,能显著提升客户的整体体验感。

1 系统与功能模块设计(System and function module design)

数字人客服系统架构主要包含数字人构建、中文问答、语音合成、唇形同步四大模块。用户只需要对录音设备提出自己的问题,系统的语音识别模块会自动识别出用户的语音,并转换为文字输入中文问答模块,然后中文问答模块给出回答,并经过语音合成及唇形同步模块,最终得以输出逼真的数字人面部表情和合成语音。数字人客服系统总体架构图如图1所示。

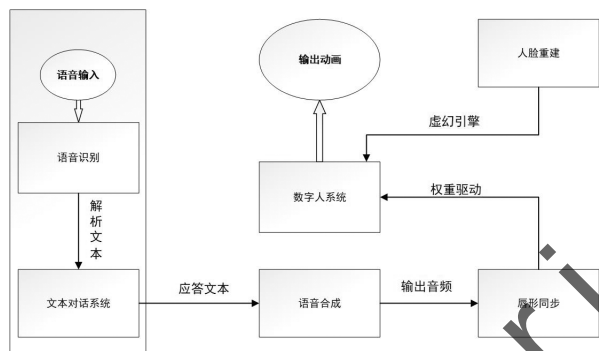


图1 数字人客服系统总体架构图

Fig. 1 Overall architecture diagram of digital human customer service system

2 基于图像的三维人脸重建(Image-based 3D face reconstruction)

在虚拟形象的构建中,本文采用基于生成器-判别器的三维可变形人脸模型(3D Morphable Models, 3DMM),主要用于从每张单独的图片中得到人脸的基础模型^[3]。在3DMM生成器阶段,使用卷积神经网络从输入图片中提取出面部姿态和光照的向量表示,以及模型中所需的参数,然后提取出面部的颜色贴图和立体构型。生成器模块的主要作用是将输入的图片转化为3D模型。

首先,使用人脸识别算法从输入图片中提取特征,这其中有两个目的:一是用于估计主体偏差并计算网络主体偏差的损失函数;二是细化纹理^[4]。其次,使用图卷积网络技术,并使用解码器、精炼器、组合器三个贴图优化模块接收生成器中的结果作为输入,并输出贴图的细化版本。人脸识别算法的输出被

传递到解码器,同时3DMM的结果传递到精炼器中以优化人脸的顶点颜色。最后,组合器综合解码器与精炼器给出的每个顶点的颜色,并将其与原始点、线、面信息结合,作为最终的输出。判别器将输出的模型再次映射到一个平面上^[5]。通过识别这张图片是原始图片还是重投影,生成器和判别器进行网络对抗训练,从而提升模型的最终生成效果。人脸三维重建流程如图2所示。

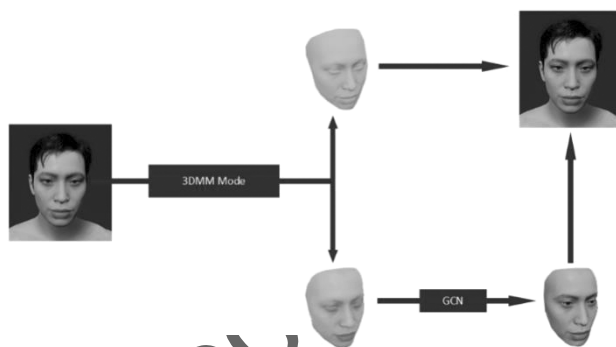


图2 人脸三维重建流程

Fig. 2 3D reconstruction process of face

3 文本对话系统(Text dialog system)

在对话系统的构建上,本文使用Rasa为基础框架,Rasa由自然语言处理库组件和核心组件两个部分组成^[6]。自然语言处理库可用于问答机器人的意图分类和实体识别,并且是一个基于有向无环图的通用型自然语言处理框架。Rasa的组件之间相互连接,形成有向无环图,并按一定的顺序运行。例如,命名实体组件必须在分词器之前运行才能正常工作。本文使用Chatito工具(Chatito是用于创建聊天机器人训练数据集的自然语言生成工具和领域特定语言)快速构建Rasa NLU训练意图识别数据集,并使用Rasa train nlu命令训练模型。

核心组件可以指定问答机器人的行为(Actions)。Rasa Core还提供了训练预测概率模型的工具,以便根据当前或历史会话预测下一步动作。首先,须要创建Domain文件,它包括意图、实体、词槽、模板和动作。其次,通过编写自定义动作(actions.py)调用API和执行相关操作。再次,构建故事(story),即模拟用户与问答机器人的实际对话。用户输入的内容会被转换为意图和实体,机器人对用户的响应则被视为动作。完成以上配置(nlu、domain、stories、config)后,即可对整个Rasa模型进行训练。Rasa的消息处理流程如图3所示。

Rasa的消息处理流程包括用户输入文本信息或语音信号,由自然语言解释器(Interpreter)将其转换成意图信息和实体信息,然后构建语义词典,包括原始文本、意图和实体,并将其传递给核心组件。Rasa的对话状态追踪器会获取当前的对

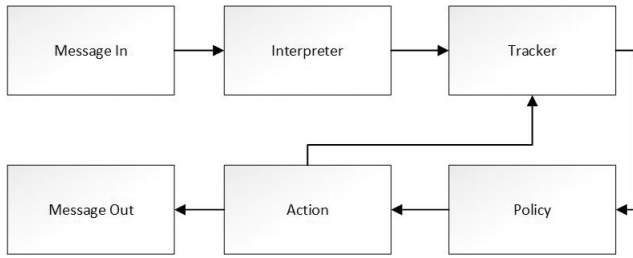


图3 Rasa的消息处理流程

Fig. 3 Message processing flow of Rasa

话状态,包括实体信息和词槽情况,并记录下来。当策略接收到当前的对话状态后,会利用特征提取组件提取对话状态的特征,并根据这些特征预测和选择下一个动作。追踪器会记录当前执行的动作,最终执行动作并将结果反馈给用户。Rasa的工程目录文件主要包括 config.yml、credentials.yml、actions.yml、domain.yml、endpoints.yml、data/nlu.yml、data/rules.yml、data/stories.yml、actions.py、data.json、data_to_neo4j.py、neo4j_Knowledge.py、models。其中,config.yml 用来存放组件配置的信息;data/nlu.yml 用来存放训练模型的数据;models 用来存放训练好的模型;domain.yml 定义了问答机器人的所有信息,包括意图、实体、词槽、动作、表单和回复等;actions.yml 用来存放动作的代码;credentials.yml 用来连接到其他服务;data/stories.yml 用于训练核心模型的故事数据集。

4 语音合成(Text-to-speech)

本文的语音合成策略是采用一种端对端的 TTS 神经网络模型,其基于典型的编码器-解码器(Encoder-Decoder)结构^[7]。Encoder 用来将文本特征转化为中间特征;Decoder 使用自回归的方式,使用中间特征与上一时刻的梅尔特征输出生成当前时刻的梅尔特征。在 Encoder 层,首先将字母转换为 512 维词向量,其次经过多层卷积操作对输入的文本序列进行上下文建模,以获得上下文特征关系,最后经过双向 LSTM 层生成编码特征。Decoder 是一个注意力机制模块,在训练时预训练网络采用真实的梅尔频谱特征作为输入进行训练,而在测试时,上一时刻线性映射的输出作为预训练网络当前时刻的输入。预训练网络为两层维度为 512 的全连接网络,作为信息瓶颈层,对学习注意力是必要的,其输入为特定一帧的梅尔频谱特征,将该时刻输出与注意力模块输出的上一时刻上下文特征进行拼接操作后,送入两层由 1 024 单元组成的 LSTM 层中,获得 LSTM 的输出。将 Encoder 输出、LSTM 输出以及累加的注意力权重变量(初始值为 0)作为注意力机制的输入,并输出这一时刻的注意力权重与这一时刻的上下文特征。其中,Encoder 的输出可以视为 Value,LSTM 的输出视为 Query,累加的注意力权重变量可以视作位置特征,获得注意力权重后,与 Encoder

的输出做加权和,得到当前时刻的上下文特征。此时,当前时刻的上下文特征再与 LSTM 的输出拼接,经过线性映射,输出目标频谱帧,最后使用 Griffin-Lim 算法将梅尔频谱重建为时域的音频信号^[8]。

5 唇形同步算法(Lip synchronization algorithm)

音频信号最终需要转化为面部动画,本文使用唇形同步算法 Audio2Face^[9]。Audio2Face 网络由三部分组成,其中第一部分包含 1 个自相关分析层和 5 个卷积层,在数据分析阶段,使用线性预测编码提取音频特征,再通过卷积神经网络对特征进行压缩,提取短时特征,该特征反映了人发音时面部动画附带的特定语气、音素等。在发音阶段,压缩后的短时特征再经过 5 个卷积层,提取相邻序列帧的相关性,最终输出整体特征图。除此之外,发声时的情绪状态也至关重要,该参数由神经网络自动学习生成,并逐层拼接到发音网络。在输出网络,数据需要经过两个全连接层,该网络将特征数据映射到面部模型的 61 维变形权重(Blend Shape),其中第一个全连接层将特征映射到表情系数,第二个全连接层将表情系数映射到 BlendShape^[10]。唇形同步算法流程如图 4 所示。

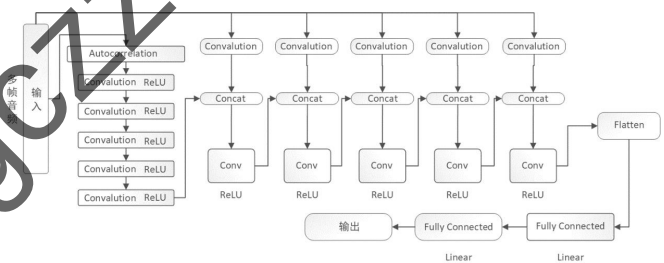


图4 唇形同步算法

Fig. 4 Lip synchronization algorithm

最终输出的权重用于控制面部动画,如图 5 所示为权重对面部动画的影响,例如当权重 MouthRight 的值变大时,面部网格的嘴部表现出右撇的姿态。

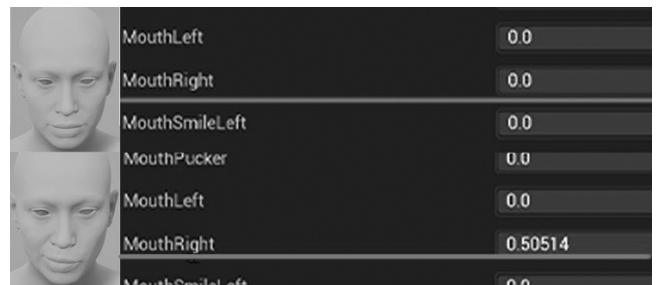


图5 权重 MouthRight 为 0.50514 时对面部动画的影响

Fig. 5 Influence on facial animation when MouthRight is 0.50514

6 系统实现(System implementation)

在虚幻引擎中,将三维重建的人脸模型重新映射为 Metahuman 模型,这是一种带有身体骨骼绑定以及毛发的虚

拟人制作系统,面部动画的表达由变形权重的曲线经由姿势映射形成,本文使用 Livelink 数据流作为 Python 端和虚幻引擎的通信方式,总共将 61 维权重传输给虚幻引擎,权重名分别如下:

EyeBlinkLeft, EyeLookInLeft, EyeLookUpLeft, EyeSquintLeft, EyeWideLeft, EyeBlinkRight, EyeLookDownRight, EyeLookInRight, EyeLookOutRight, EyeLookUpRight, EyeSquintRight, EyeLookDownLeft, EyeWideRight, JawForward, JawLeft, JawRight, JawOpen, MouthClose, MouthFunnel, MouthPucker, MouthLeft, MouthRight, EyeLookOutLeft, MouthSmileLeft, MouthSmileRight, MouthFrownLeft, MouthFrownRight, MouthDimpleLeft, MouthDimpleRight, MouthStretchLeft, MouthStretchRight, MouthRollLower, MouthRollUpper, MouthShrugLower, MouthShrugUpper, MouthPressLeft, MouthPressRight, MouthLowerDownLeft, MouthLowerDownRight, MouthUpperUpLeft, MouthUpperUpRight, BrowDownLeft, BrowDownRight, BrowInnerUp, BrowOuterUpLeft, BrowOuterUpRight, CheekPuff, CheekSquintLeft, CheekSquintRight, NoseSneerLeft, NoseSneerRight, TongueOut, HeadYaw, HeadPitch, HeadRoll, LeftEyeYaw, LeftEyePitch, LeftEyeRoll, RightEyeYaw, RightEyePitch, RightEyeRoll.

在虚幻引擎端,该 61 维权重实时修改变画蓝图中的动画曲线值,如图 6 所示。

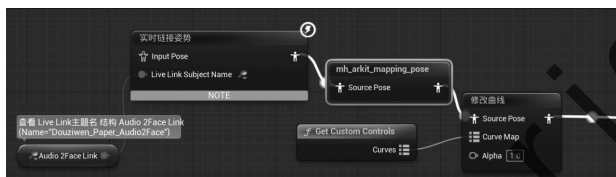


图 6 虚幻引擎中动画蓝图权重应用

Fig. 6 Application of animation blueprint weight in unreal engine

数字人客服系统虚幻引擎端如图 7 所示,本文研究抽取不同年龄段的实验者共 50 人,其中本科生 30 人,硕士研究生 15 人,教师 5 人,实验者针对不同的问题向数字人客服提问,其中 48 人表示对数字人客服的服务高度满意,整体满意度为 96%。



图 7 数字人虚幻引擎端展示

Fig. 7 Unreal engine end display of digital human

7 结论 (Conclusion)

本文设计并实现的基于神经网络和虚幻引擎的数字人客服系统,创造性地将客服系统与数字人相结合,利用神经网络和虚幻引擎创建了一个高度逼真和交互性强的虚拟对象,使客户能够以自然和直观的方式与数字人进行交互。调查和测试结果显示,客户对这种新的服务方式反应积极,多数实验者对数字人客服系统的服务表示高度满意。统计结果也表明,数字人客服系统的使用对客户服务运营的效率和有效性产生重大影响,同时能帮助企业降低成本,提高客户满意度,以及增加收入。

参考文献 (References)

- [1] 宋倩茜,马双. 电商平台智能客服与人工客服的顾客感知价值对比研究[J]. 商展经济,2022(22):38-40.
- [2] 王泓贤. 基于虚幻引擎的 3D 动画创作研究与应用[J]. 文化产业,2020(30):10-11.
- [3] 黄炎辉. 基于图像的三维人脸建模及其应用研究[D]. 西安:西北工业大学,2019.
- [4] TAVOOSI J. Designing a new recurrent convolutional neural network for face detection and recognition in a color image[J]. Iran Journal of Computer Science,2021,4(3):185-194.
- [5] 徐志良. 基于 GAN 的人脸属性解耦编辑研究[D]. 武汉:华中科技大学,2021.
- [6] 刘宇杰,陈家豪,宋晖,等. 基于 Rasa 的任务型对话系统设计与实现[J]. 现代计算机,2022,28(3):108-112.
- [7] CHEN L J, REN J, CHEN P F, et al. Limited text speech synthesis with electroglottograph based on Bi-LSTM and modified Tacotron-2 [J]. Applied Intelligence, 2022, 52(13):15193-15209.
- [8] MASUYAMA Y, YATABE K, KOIZUMI Y, et al. Deep Griffin-Lim iteration: trainable iterative phase reconstruction using neural network[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 15(1):37-50.
- [9] 蔡国鑫. Audio2Face 基于音频文件智能生成虚拟角色面部动画[J]. 现代电影技术,2021(9):60-61.
- [10] PARK J H, KIM J, SONG J, et al. Simultaneous simplification of multiple triangle meshes for blend shape[J]. Journal of the Korea Computer Graphics Society, 2019, 25(3):75-83.

作者简介:

豆子闻(1994-),男,硕士生。研究领域:三维重建,虚拟现实。
李文书(1975-),男,博士,教授。研究领域:图像处理,认知建模,虚拟现实,物联网集成开发。