用于角色表情动画生成的深度学习技术

张睿思, 潘烨*

(上海交通大学电子信息与电气工程学院 上海 200240)
(whitneypanye@sjtu.edu.cn)

摘 要: 人脸动作捕捉不仅需要对人脸几何信息进行模拟, 而且需要准确传达人脸表情. 传统的人脸动作捕捉技术, 如 ARkit, 基于人脸的几何信息对面部表情进行捕捉, 但是很难让观众体验到角色表情变化. 而最近的基于情绪的动作捕捉技术, 如 ExprGen, 考虑使用人脸情绪进行面部捕捉, 但很难对角色脸部细节进行刻画. 为此, 提出将人脸几何信息和表情结合的方法, 对动画角色进行控制. 首先, 通过训练神经网络识别人脸和动画角色表情, 对人脸和动画数据集图像进行匹配. 然后, 通过训练端到端神经网络, 提取角色表情信息, 获得动画角色骨骼参数. 最后, 结合人脸几何信息对脸部关键点骨骼参数进行修正. 通过对不同人脸输入, 生成角色表情定性分析; 用4个演员视频作为输入, 带动角色运动的吸引力和强度定量分析证明了方法的准确性和实时性.

关键词: 动画制作; 深度学习; 动作捕捉
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2022.19006

Stylized Avatar Animation Based on Deep Learning

Zhang Ruisi and Pan Ye*

(*School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai* 200240)

Abstract: Animating 3D character rigs from human faces requires both geometry features and facial expression information. However, traditional animation approaches such as ARkit failed to connect character storytelling to the audience because the character expressions are hard to recognize. However, recent emotion-based motion capture techniques, such as ExprGen, consider using facial emotion for facial capture. But it is difficult to characterize the details of the character's face. A network is proposed to incorporate facial expressions into animation. Firstly, an emotion recognition neural network is used to match human and character datasets. Then, an end-to-end neural network is trained to extract character facial expressions and transfer rig parameters to characters. Finally, human face geometry is utilized to refine rig parameters. Qualitative analysis of the generated character expressions, and quantitative analysis of the attractiveness and intensity of the character expression have demonstrated the accuracy and real-time of the model.

Key words: character animation; deep learning; motion capture

在人脸面部动作捕捉中, 传统的方法(如ARkit)用相机获取人脸几何信息并映射到三维模型中. 通过学习二维视频到三维模型参数的映射关系, 获得三维模型的参数信息. 此外, 一些商业软件, 如 Faceware, 通过重构二维输入图像, 也可以获得三维模型的参数信息. 这些方法能够有效

收稿日期: 2021-03-24; 修回日期: 2021-12-22. **基金项目**: 国家自然科学基金青年科学基金(62102255); 国家重点研发计划(2019YFC1521104); 国家社科基金重大项目(I8ZD22); 上海市科委青年科技英才扬帆计划(20YF1421200). **张睿思**(2000—), 女, 在校学生, 主要研究方向为计算机图形、计算机视觉; **潘烨**(1989—), 女, 博士, 副教授, 博士生导师, 论文通信作者, 主要研究方向为虚拟现实、动画制作.

地提取人脸几何信息，但是很难让观众感受到角色表情的变化. ExprGen[1]和 DeepExpr[2]等基于表情模型尝试对这个问题进行改进，但是在对表情信息进行优化的同时，可能会丢失重要的面部几何信息，很难对动画角色面部进行细节控制. 这些问题导致在准确传达面部信息变化的同时，无法准确地表达角色的情感信息. 而在观众观看视频时，动画角色的表情和几何细节的变化对观影体验有着十分重要的影响.

本文提出了一种将人脸表情信息和几何信息相结合的方法，实现对动画角色的面部表情操纵，能够在准确地表征角色面部几何特征变化的同时，向观众传达角色表情信息. 其通过轻量级的网络结构，能够实时生成角色表情，有效地提高角色操纵效率.

本文方法可以分为数据预处理、离线训练和在线生成 3 个部分. 在数据预处理部分，使用包含六度表情的多种人脸数据集，运用数据增强等技术进行人脸情绪识别；同时，对角色骨骼参数进行渲染，以获得角色对应的二维动画图像. 在离线训练部分，首先在人脸数据集和二维动画图像数据集中训练情绪识别网络，获得人脸图像和二维动画图像之间的匹配；然后，通过深度网络学习相同表情的人脸图像与角色骨骼参数之间的映射关系，获得预训练网络结构. 在在线生成部分，利用预训练网络对视频每帧的输入输出骨骼参数，再利用前后关键帧之间的联系对输出参数进行插值；然后对输入图像进行三维重构，以获得角色的运动参数；最后结合人脸图像的几何信息对骨骼参数进行细节优化. 本文方法有效地提高了角色表情识别的准确率以及方法反应速度.

本文方法与传统方法的区别如下.

(1) 传统方法仅考虑角色几何特征，本文方法将角色情绪变化引入动画三维建模中，并结合具体几何细节进行优化，提高人类动画捕捉效果.

(2) 实现动画角色实时自动控制.

(3) 传统动画制作需要用深度相机等提取脸部信息，本文方法通过对输入的二维人脸进行重构，降低了制作成本.

(4) 通过在 2 个关键帧之间进行插值，有效地平滑过渡角色面部表情，提高用户使用体验.

1 相关工作

1.1 面部表情分类与感知

表情识别与分析在人机交互、计算机图形学等领域有着广泛的应用. Ekman 等[3]提出使用表情行动编码系统表征特定表情所具有的面部肌肉状态. 通常，面部特征可以分为上面部和下面部，两者之间的关联较小. 上面部表情是指眼睛、眉毛和脸颊的表情；下面部表情是指嘴唇、鼻根部和两者之间的表情. 表情行动编码系统定义了 46 个基本的行动单元，形成的 7000 个组合能够表征大多数观察到的表情.

在动画制作中，面部表情控制单元被广泛用于动漫人物的情绪感知和操纵. 其中，FACSGen[4]通过控制行动单元控制动漫人物三维面部表情. 由于微观层面的叠加很难引起用户在整体层次上对动漫情绪进行感知，因此，HapFACS[5]对其进行改进，允许动漫制作者在控制单元和整体情绪 2 个层次上对角色表情进行控制. 但是，微观和宏观的控制也限制了角色表情的泛化性能，相同表情很难从一个角色迁移到另一个角色. 基于时空的方法[6]被用于人脸动画后期编辑，以满足动画师的特殊应用需求. 基于时空的人脸编辑技术[6]通过使用基于拉普拉斯的人脸表情合成技术，将被编辑特征点的位移在空间域传播到人脸模型的其他顶点，使用高斯函数将用户的编辑效果在时间域传播到邻近动画序列，保证人脸表情动画的平滑过渡，同时保留人脸几何细节. Song 等[7]提出使用差分子空间重构的方法自动生成角色骨骼，通过学习差分坐标的标记信息，对子空间网格进行重构，能够有效地降低变形误差，提高泛化性能. 受上述角色控制方法的启发，本文在面部控制系统考虑了角色面部情绪，使观众能够对角色情绪有更好的感知.

1.2 人脸追踪

对于人脸追踪，传统的方法是使用多层感知机或集成学习的方法对输入图像的各个窗口进行检测，找到含有人脸的部分，并进行综合. 随着深度学习的发展，深度神经网络及其各种变种的应用，使人脸检测效果大大改善. 例如，Fast R-CNN[8]和 YOLO[9]等模型能够高效、准确地检测出图像中的多张人脸. Sun 等[10]将循环卷积神经网络框架与特征级联、多尺度训练、模型预训练和关键参数的正确校准相结合，在 FDDB 数据集中获得了 83%的准确率. Wu 等[11]将分层注意力机制引入人脸识别中，通过使用高斯核模型提取人脸局部特征，并用循环神经网络对特征间关系进行建模，实现了对面部特征的层次化感知. 该模型在数据集 FDDB (face detection dataset), Wider FACE 和 UFDD(unconstrained face detection dataset)中的准确率分别为 96.42%, 94.84%

和 74.60%. 受深度学习方法的启发, 本文使用深度卷积网络对人脸进行检测与追踪.

1.3　面部特征提取

提取人脸面部特征能够有效地获得人脸情绪特征信息, 可将其作为特征向量用于后续人脸识别和表情迁移. Zafeiriou 等[12]使用从 L_1 正则化衍生出的稀疏信号处理的方法表征面部表情, 并结合支持向量机算法对特征向量进行分类. 通过对面部网格化预处理, 该算法获得了比直接处理原始图像更好的效果, 在 CK(Cohn-Kanade)数据集中获得了 92.4%的准确率.

随着深度学习技术的发展, 深度神经网络[13]被用于对人脸图像情绪感知. 与传统特征提取任务不同的是, 在深度学习中, 人脸检测和特征提取这 2 个过程是同时进行的. 利用深度学习识别情绪可以分为 3 个步骤: 预处理、深度特征学习和深度特征提取. 其中, 预处理是指对输入图像进行人脸提取、旋转矫正和数据增强等处理. 然后, 通过端到端的学习方式完成对图像特征的提取. Yang 等[14]通过 DeRL (de-expression residue learning)方法学习人脸表情信息. 首先使用生成模型生成表情对应的中性面部图像, 虽然表情信息最终被过滤, 但是情绪信息被存储在生成模型的中间层; 然后通过学习中间层中残留的信息对面部表情进行分类. 在数据集 CK+(extended Cohn-Kanade), Oulu-CASIA, MMI 和 BU-3DFE 中, 该算法分别取得了 97.30%, 88.00%, 73.23%和 84.17%的准确率. 受采用深度学习进行人脸识别的启发, 本文使用卷积网络的输出特征向量作为面部表情的特征表示不同人脸.

1.4　动画捕捉

传统的动画捕捉方法[15-16]使用深度相机或三维扫描仪直接获取人脸面部信息, 并映射在动画角色中, 由于设备昂贵、操作平台搭建复杂, 很难获得广泛使用. 例如, Weise 等[17]将人脸的几何信息与预存储的人脸深度信息结合, 通过对概率问题进行优化, 获取 Blendshape 参数序列, 有效地提高了表情迁移的速度和稳定性. 但是其在具体使用时, 受分辨率限制, 很难捕捉到细微的人脸几何信息和动作变化, 同时也很难捕捉人脸表情的变化. Bouaziz 等[18]改进了三维参数优化算法, 通过获取低维的人脸参数表征, 有效地提高了 Blendshape 参数提取速度. 该算法首先使用 RGB-D 相机获得人脸深度图像, 将其映射到人脸三维模型中; 然后对参数进行主成分分析降维和几何变换, 将处理后的参数映射到动画模型中. 单独使用二维图像作为表情输入能够有效地降低设备需求, 使动漫表情迁移广泛使用成为可能. Cao 等[19]提出的不需要预先标定的方法, 可实现表情实时迁移, 其采用位移动态表达式(displaced dynamic expression, DDE)模型对提取的人脸二维特征点进行回归, 然后结合相机的误差优化人脸参数, 并映射在动漫头像中. DeepExp[1]模型是识别现有数据库中与给定人类图像最接近的二维表情图像的一种检索方法. 而 ExprGen[2]是通过给定人类图像生成三维风格化表情骨骼参数的方法. 受这些基于深度学习动画捕捉技术的启发, 本文使用轻量深度学习网络接受人脸关键帧输入, 输出角色骨骼参数捕捉面部表情特征, 并结合插值算法进行优化, 以提高视频前后帧之间连续性.

2　数据预处理

本文使用的数据集包括人脸和动画 2 类, 本节介绍它们的相关信息和数据预处理方法.

2.1　人脸数据集

本文使用的人脸数据集来自 CK+[20], DISFA[21], KDEF[22]和 MMI[23]. 其中, 每个数据集均选取带有六度情绪和中性 7 类标注的正面人脸图像. 对数据集中每个标签下的图像通过旋转和缩放等操作进行数据增强, 使各标签下的图像数量均等, 最后获得大约 10 000 幅图像. 本文插图和实验结果展示图像均来自上述数据集.

2.2　动画数据集

本文使用的三维动画数据集来自 FERG-3D-DB 数据集[2], 包括 Mery, Bonnie, Ray 和 Malcolm 这 4 个角色. 其中每个角色包含用于控制角色表情的骨骼参数值以及对应六度情绪和中性 7 类标注, 大约有 40 000 个样例.

2.3　特征提取

对于人脸和动画数据集提取以下面部标志点作为人脸几何特征: 左右眉毛间高度(眉毛最高特征点到最低特征点高度)、左右眼睛高度(眼睛最高特征点到最低特征点高度)、鼻部宽度(鼻部最右特征点到最左特征点宽度)、左右鼻部高度(最左/右边特征点到鼻子底部高度); 嘴部宽度(最右特征点到最左特征点宽度)、嘴部高度(最高特征点到最低特征点高度)和嘴唇高度(上嘴唇到眼底部高度). 将每幅图像缩放到 256×256 像素大小, 用于后续神经网络输入.

3 网络结构设计

如图 1 所示，首先训练表情识别网络，用于动画与人脸表情识别和角色匹配；然后训练三维动画网络，用于生成人脸对应的动画骨骼参数.

3.1 表情识别网络

该过程的目的是训练神经网络，用于二维人脸图像和动漫图像的分类. 首先使用如表 1 所示的神经网络对人脸数据集进行训练. 分别对 2 个数据集进行分类，分为生气、恶心、害怕、高兴、悲伤、中性和惊讶 7 类. 在二维动漫数据集中，保持池化层前面网络结构参数不变，在动漫数据集中对训练好网络的全连接层进行微调.

使用 PyTorch 框架对网络进行端到端训练. 在训练过程中，将数据集按照 8∶1∶1 的比例划分用于训练、验证和测试. 使用随机梯度下降(stochastic gradient descent, SGD)优化器，其中动量为 0.9，权重下降为 0.000 5，初始学习率为 0.01，学习率为每 10 个周期下降为原来的 1/10. 在人脸数据集中训练 60 个周期，动画数据集中训练 50 个周期，每个训练批大小为 50.

3.2 人脸-动画角色匹配

在三维网络中，搜索与人脸数据集图像对应的动画图像作为基准值，用于三维动画网络和角色迁移网络进行训练.

首先在 Autodesk®MAYA 软件中，标注三维动画角色的特征点，按照第 2.3 节的数据预处理方法提取动画图像的面部特征向量，并对三维角色进行渲染，以获得二维图像. 然后搜索人脸数据集中的图像在对应三维动画数据集中最相似的样例. 具体的方式为：首先搜索所有动漫数据集中与表情距离最为相近的 30 幅图像；然后利用几何距离查找 30 幅图像中特征点与人脸最为相近的图像，将其作为结果输出.

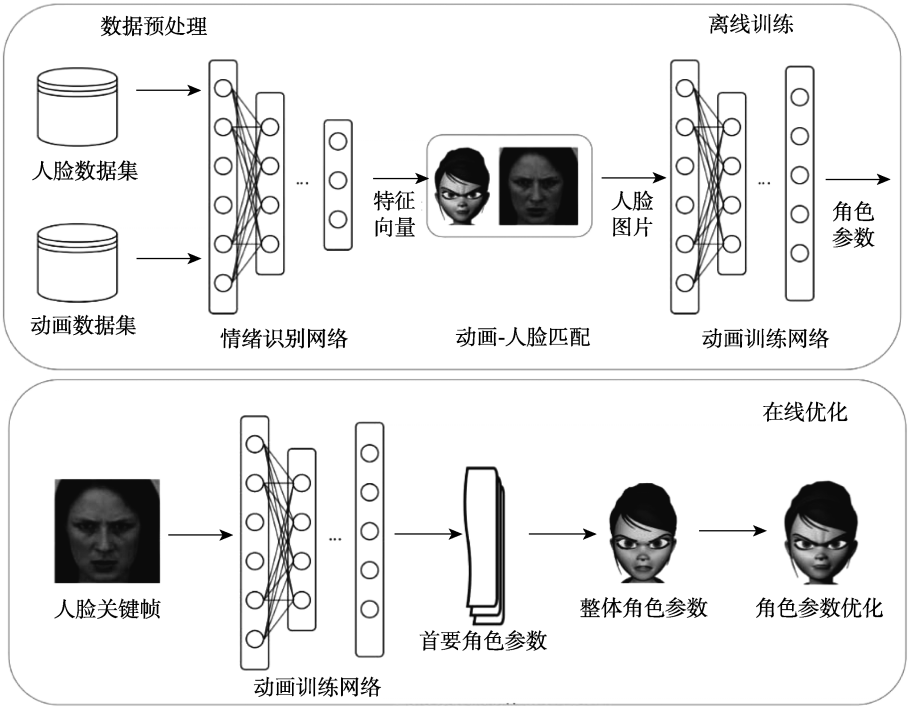


图 1 本文方法流程

3.2.1 表情距离

在表情距离中使用简森·香农散度(JS divergence, JSD)作为衡量 2 幅图像表情距离的依据. 将人脸图像和动画图像输入表 1 所示的卷积神经网络后，将 FC2 层输出的 512 维向量作为表情特征向量，并利用 JDS 计算 2 个向量之间的距离.

3.2.2 几何距离

在几何距离中，将第 2.3 节中获得的几何特征作为几何特征向量，并进行归一化. 查找归一化后与人脸几何特征向量 $\boldsymbol{h}$ 最近的动画几何特征向量 $\boldsymbol{c}$ 作为结果输出.

3.3 三维动画训练网络

该过程的目的是训练神经网络，用于生成三维动画角色参数，使用第 3.2 节中匹配的人脸-三维参数作为数据进行训练. 在表 2 所示的三维动画训练网络中，将人脸图像作为输入，将网络 FC3 层

表 1　表情识别网络参数

层	步长/块	输出
CONV1	7×7/4	56×56×96
POOL1	3×3/2	28×28×96
CONV2	5×5/1	28×28×256
POOL2	3×3/2	14×14×256
CONV3	3×3/1	14×14×384
POOL3	3×3/2	7×7×384
CONV4	3×3/1	3×3×384
POOL4	3×3/2	3×3×384
FC1		1×1×1 024
FC2		1×1×512
FC3		1×1×7

输出结果和基准值的交叉熵损失函数作为目标函数进行优化.

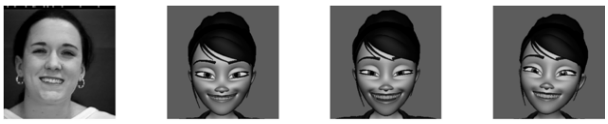
使用 PyTorch 框架对网络进行端到端训练. 在训练过程中, 使用 SGD 优化器, 其中动量为 0.9, 权重下降为 0.000 5, 初始学习率为 0.001, 学习率为每 10 个周期下降为原来的 1/10, 训练 50 个周期, 每个训练批大小为 50.

表 2　三维动画训练网络参数

层	步长/块	输出
CONV1	11×11/1	64×256×256
CONV2	1×1/2	64×128×128
CONV3	5×5/1	64×128×128
CONV4	1×1/2	64×64×64
CONV5	5×5/1	64×64×64
CONV6	1×1/2	64×32×32
CONV7	3×3/1	64×32×32
CONV8	1×1/2	64×16×16
CONV9	3×3/1	64×16×16
CONV10	1×1/2	64×8×8
CONV11	3×3/1	64×8×8
AVG POOLING	8×8/1	64×1×1
FC1		1×7
FC2		1×7

4　角色骨骼参数生成

通过神经网络的设计能够有效地捕捉动画角色的整体特征, 如角色表情等. 但是动画角色的细节参数(如眼睛睁开的幅度大小、鼻子宽度等)很难保证神经网络能够完全学习. 因此, 将角色几何特征和整体特征进行融合, 生成最终骨骼参数. 整体骨骼参数生成和骨骼参数优化后的角色图像如图 2 所示, 图 2a~图 2d 依次是输入人脸图像、角色整体骨骼参数的生成结果、角色骨骼参数的优化结果和加入运动参数的结果.



a. 输入人脸　b. 整体生成　c. 参数优化　d. 加入运动参数

图 2　不同阶段的优化结果

4.1　整体骨骼参数生成

输入人脸视频关键帧作为三维动画预训练网络的输入. 输出的骨骼参数提取了输入面部特征的情绪信息, 还需要进一步结合人脸几何信息进行优化.

4.2　骨骼参数优化

利用 OpenCV 提取的二维图像深度等信息进行三维重构, 获得头部转动关键点的坐标值, 结合神经网络生成角色骨骼参数进行细节优化, 将优化后的骨骼参数作为结果进行输出. 具体方法是: 利用 OpenCV 提取读入二维人脸图像的面部特征点(眼睛最高、最低、最左、最右特征点, 鼻子两侧特征点, 嘴巴眼睛最高、最低、最左和最右特征点, 眼珠中心特征点)坐标和相机内参, 通过求解 N 点透视位姿问题获得面部旋转矩阵; 将该面部旋转矩阵转换为对应的坐标值, 获得在三维空间中人脸原始的特征点坐标; 通过计算人脸特征点坐标并转换为角色骨骼参数, 与第 4.1 节中使用神经网络生成的骨骼参数叠加, 选取其中对应骨骼参数最大值对角色面部细节进行优化. 在优化过程中, 综合对角色眼睛、鼻子和嘴部进行优化, 而非单独对角色嘴部进行优化.

4.3　骨骼参数参数插值

面部表情一般很难发生较大变化, 因此可以利用线性插值算法在角色表情变化的 2 个关键帧之间进行插值, 使表情过渡更加平滑. 根据前后 2 个关键帧骨骼参数计算中间 5 帧的骨骼参数间隔值, 将其作为插值结果, 如图 3 所示. 可见 2 个关键帧插值后角色表情过渡更加平滑、自然.

图 3　2 个关键帧之间的插值结果

5 案例分析

5.1 角色情绪识别准确率

在训练人脸表情识别网络和角色情绪识别网络时，将其中 80%的图像用于训练集进行训练，10%的图像用于验证集进行验证，10%的图像用于测试集进行测试. 不同角色的情绪识别准确率如表 3 所示. 测试结果证明，设计的神经网络能够有效地识别人脸和动漫角色 Mery 的表情. 可以看出，训练的神经网络能够较有效地对人脸和角色表情进行分类，因此在人脸-动画角色匹配时，前一层全连接层输出能够较好地作为角色情绪特征.

表 3 情绪识别准确率

角色	准确率/%
人脸	78.96
Mery	70.89

5.2 迁移结果展示

本文方法能够有效地将人脸表情迁移到不同的动画角色中，同时，针对不同的人脸也能迁移到相同的角色，保证方法的鲁棒性. 图 4 展示人脸情绪变化的图像序列迁移到 Mery 的结果，图 5 展示不同人脸情绪变化的图像迁移到 Mery 的结果. 可以看出，针对单张人脸，本文方法能够保证关键帧前后表情变换的准确性；针对不同人脸，本文方法能够准确地识别出不同表情，保证不同输入的鲁棒性. 其中，基于情感的方法是指 ExprGen[2]根据人脸表情利用深度学习方法生成角色骨骼参数的算法；基于几何的方法①是指利用 Faceware 软件生成角色骨骼参数.

a. 相同人脸不同表情输入

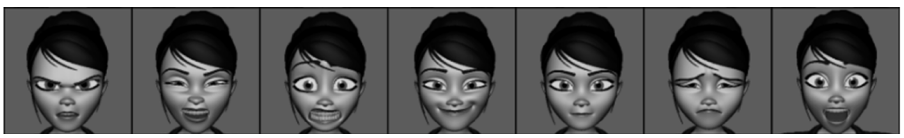
b. 本文

c. 基于几何的方法①

图 4 不同方法对相同人脸的迁移结果

a. 不同人脸不同表情输入

b. 本文



c. 基于情感的方法[2]

图 5 不同方法对不同人脸的迁移结果

① https://facewaretech.com

5.3　与其他方法迁移效果对比

本文方法与仅基于情绪进行角色迁移和仅基于几何信息进行角色迁移的动画效果进行对比，最终结果如图 5 所示. 可以看出，与基于情感的角色迁移方法相比，本文方法能够对嘴部和眼睛进行更为精细的控制，提高观众对于角色情绪变化感知程度.

5.4　用户测试

选择 4 名专业演员[24] (2 男 2 女)的 7 种不同表情(中性、生气、悲伤、害怕、恶心、高兴、惊讶)的视频，利用本文方法、基于几何和基于情感方法，分别生成 28 个视频的骨骼参数，并将每帧的骨骼参数进行渲染，合成视频，用于用户测试.

共选择 32 名同学(16 男 16 女，平均年龄 20 岁)作为用户，每人观看一套原始演员或动画角色的视频(一名演员/演员生成角色骨骼参数的 7 个表情组成的 7 个视频为一套)，每个视频被重复观看 2 遍. 要求每位用户回答以下 3 个问题.

问题 1. 角色的表情是以下 7 个中的哪一个? 用户在中性、生气、悲伤、害怕、恶心、高兴和惊讶 7 个选项中选择一个作为最终结果.

问题 2. 角色的整体吸引力如何? 用户使用鼠标在屏幕上移动滑块，以 1~7 的等级评定吸引力. 1 分表示“完全没有吸引力”，7 分表示“非常有吸引力”.

问题 3. 角色所表现出的情感有多强烈? 用户使用鼠标单击在屏幕上移动滑块，以 1~7 的等级对强度进行评级. 1 分表示“完全不强烈”，7 分表示“非常强烈”.

对于用户识别情绪的准确率进行分析，结果如表 4 所示. 实验对象对原始演员、基于几何、基于情感和本文方法视频的角色情感强度和角色吸引力强度感受如表 5 所示. 本文方法与基于几何的方法相比，能够更好地表征用户情绪；与基于情感的方法相比，拥有更好的角色情绪吸引力.

表 4　不同方法在用户测试中视频情绪识别准确率

方法	准确率/%
原始演员	41.71
基于几何①	18.89
基于情感[2]	27.33
本文	28.94

① https://facewaretech.com

表 5　不同方法在用户测试中视频情感强度和吸引力(平均值)

方法	强度	强度 p 值	吸引力	吸引力 p 值
原始演员	4.79	0.005	5.02	0.017
基于几何①	4.02	0.036	3.98	0.045
基于情感[2]	4.63	0.016	4.25	0.010
本文	4.58		4.62	

5.5　结果分析

对本文方法、基于几何的方法和基于情感的方法的 p 值进行分析. 在强度方面，与基于情感的方法 p=0.016 相比，本文方法没有明显的提高；但是与基于几何的方法 p=0.036 相比，有显著的提高. 在吸引力方面，与基于情感的方法 p=0.010 相比，本文方法没有显著的提高，但与基于几何的方法 p=0.045 相比，有显著的提高.

6　结　语

本文提出一种结合人脸表情和几何特征的实时动画生成方法，通过对人脸表情迁移，有效地提高了观众对于动画人物的角色感知度；同时，通过几何特征控制使对角色细节操纵成为可能. 通过插值优化算法，实现了角色实时控制和自动生成. 由于角色表情和几何特征变化在前后帧之间可能存在不连贯的问题，如何将两者结合，实现前后平滑过渡将是之后研究的重点.

参考文献(References):

[1] Aneja D, Colburn A, Faigin G, *et al*. Modeling stylized character expressions via deep learning[C] //Proceedings of Asian Conference on Computer Vision. Heidelberg: Springer, 2016: 136-153

[2] Aneja D, Chaudhuri B, Colburn A, *et al*. Learning to generate 3D stylized character expressions from humans[C] //Proceedings of the IEEE Winter Conference on Applications Press, 2018: 160-169

[3] Ekman P, Friesen W. Facial action coding system: a technique for the measurement of facial movement[M]. Palo Alto: Consulting Psycholo- Gists Press, 1978

[4] Roesch E B, Tamarit L, Reveret L, *et al*. FACSGen: a tool to synthesize emotional facial expressions through systematic manipulation of facial action units[J]. Journal of Nonverbal Behavior, 2011, 35(1): 1-16

[5] Amini R, Lisetti C. HapFACS: an open source API/software to generate FACs-based expressions for ECAs animation and for corpus generation[C] //Proceedings of Humaine Association Conference on Affective Computing and Intelligent Interaction. Los Alamitos: IEEE Computer Society Press, 2013: 270-275

[6] Wan Xianmei, Jin Xiaogang. Spacetime facial animation editing[J]. Journal of Computer-Aided Design & Computer Graphics, 2013, 25(8): 1183-1189(in Chinese)
(万贤美, 金小刚. 时空人脸表情动画编辑[J]. 计算机辅助设计与图形学学报, 2013, 25(8): 1183-1189)

[7] Song S L, Shi W Q, Reed M. Accurate face rig approximation with deep differential subspace reconstruction[J]. ACM Transactions on Graphics, 2020, 39(4): Article No.34

[8] Girshick R. Fast R-CNN[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 1440-1448

[9] Redmon J, Farhadi A. YOLOv3: an incremental improvement[OL]. [2021-03-24]. https://arxiv.org/abs/1804.02767

[10] Sun X, Wu P C Hoi S C H. Face detection using deep learning: an improved faster RCNN approach[J]. Neurocomputing, 2018, 299: 42-50

[11] Wu S Z, Kan M N, Shan S G, *et al*. Hierarchical attention for part-aware face detection[J]. International Journal of Computer Vision, 2019, 127(6/7): 560-578

[12] Zafeiriou S, Petrou M. Sparse representations for facial expressions recognition via11 optimization[C] //Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2010: 32-39

[13] Liu X F, Kumar B V K V, You J, *et al*. Adaptive deep metric learning for identity-aware facial expression recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2017: 522-531

[14] Yang H Y, Ciftci U, Yin L J. Facial expression recognition by de-expression residue learning[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 2168-2177

[15] Gonzalez-Franco M, Ofek E, Pan Y, *et al*. The Rocketbox library and the utility of freely available rigged avatars[J]. Frontiers in Virtual Reality, 2020, 1: 561558

[16] Pan Y, Mitchell K. PoseMMR: a collaborative mixed reality authoring tool for character animation[C] //Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops. Los Alamitos: IEEE Computer Society Press, 2020: 758-759

[17] Weise T, Bouaziz S, Li H, *et al*. Realtime performance-based facial animation[J]. ACM Transactions on Graphics, 2011, 30(4): Article No.77

[18] Bouaziz S, Wang Y Y, Pauly M. Online modeling for realtime facial animation[J]. ACM Transactions on Graphics, 2013, 32(4): Article No.40

[19] Cao C, Hou Q M, Zhou K. Displaced dynamic expression regression for real-time facial tracking and animation[J]. ACM Transactions on Graphics, 2014, 33(4): Article No.43

[20] Lucey P, Cohn J F, Kanade T, *et al*. The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression[C] //Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. Los Alamitos: IEEE Computer Society Press, 2010: 94-101

[21] Mavadati S M, Mahoor M H, Bartlett K, *et al*. DISFA: a spontaneous facial action intensity database[J]. IEEE Transactions on Affective Computing, 2013, 4(2): 151-160

[22] Goeleven E, de Raedt R, Leyman L, *et al*. The Karolinska directed emotional faces: a validation study[J]. Cognition and Emotion, 2008, 22(6): 1094-1118

[23] Pantic M, Valstar M, Rademaker R, *et al*. Web-based database for facial expression analysis[C] //Proceedings of the IEEE International Conference on Multimedia and Expo. Los Alamitos: IEEE Computer Society Press, 2005: 5

[24] Livingstone S R, Russo F A. The Ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in north American English[J]. PLoS One, 2018, 13(5): e0196391