基于深度学习和表情 AU 参数的人脸动画方法

闫衍芙, 吕　科, 薛　健*, 王　聪, 甘　玮
(中国科学院大学工程科学学院　北京　100049)
(xuejian@ucas.ac.cn)

摘　要: 为了利用计算机方便快捷地生成表情逼真的动漫人物, 提出一种基于深度学习和表情 AU 参数的人脸动画生成方法. 该方法定义了用于描述面部表情的 24 个面部运动单元参数, 即表情 AU 参数, 并利用卷积神经网络和 FEAFA 数据集构建和训练了相应的参数回归网络模型. 在根据视频图像生成人脸动画时, 首先从单目摄像头获取视频图像, 采用有监督的梯度下降法对视频帧进行人脸检测, 进而对得到的人脸表情图像准确地回归出表情 AU 参数值, 将其视为三维人脸表情基系数, 并结合虚拟人物相对应的 24 个基础三维表情形状和中立表情形状, 在自然环境下基于表情融合变形模型驱动虚拟人物生成人脸动画. 该方法省去了传统方法中的三维重建过程, 并且考虑了运动单元参数之间的相互影响, 使得生成的人脸动画的表情更加自然、细腻. 此外, 基于人脸图像比基于特征点回归出的表情系数更加准确.

关键词: 人脸动画; 人脸运动单元; 融合变形模型; 深度学习
中图法分类号: TP391.41　　**DOI:** 10.3724/SP.J.1089.2019.17682

Facial Animation Method Based on Deep Learning and Expression AU Parameters

Yan Yanfu, Lyu Ke, Xue Jian*, Wang Cong, and Gan Wei
(*School of Engineering Science, University of Chinese Academy of Sciences, Beijing*　100049)

Abstract: To generate virtual characters with realistic expression more conveniently using computers, a method based on deep learning and expression AU parameters is proposed for generating facial animation. This method defines 24 facial action unit parameters, i.e. expression AU parameters, to describe facial expression; then, it constructs and trains corresponding parameter regression network model by using convolutional neural network and the FEAFA dataset. During generating facial animation from video images, video sequences are firstly obtained from ordinary monocular cameras, and faces are detected from video frames based on supervised descent method. Then, the expression AU parameters, regarded as expression blendshape coefficients, are regressed accurately from the detected facial images, which are combined with avatar's neutral expression blendshape and the 24 corresponding blendshapes to generate the animation of the digital avatar based on a blendshape model under real world conditions. This method does not need 3D reconstruction process in traditional methods, and by taking the relationship between different action units into consideration, the generated animation is more natural and realistic. Furthermore, the expression coefficients are more accurate based on face images rather than facial landmarks.

收稿日期: 2018-12-14; 修回日期: 2019-08-02. 基金项目: 国家重点研发计划项目(2017YFB1002203); 国家自然科学基金(61671426, 61731022, 61871258, 61471150, 61572077); 中国科学院大学优秀青年教师科研能力提升项目(Y95401YXX2); 北京市自然科学基金(4182071); 中国科学院科研装备研制项目(YZ201670). 闫衍芙(1994—), 女, 硕士研究生, 主要研究方向为数字图像处理、计算机视觉; 吕　科(1971—), 男, 博士, 教授, 博士生导师, 主要研究方向为数字图像处理、智能信息处理技术; 薛　健(1979—), 博士, 副教授, 硕士生导师, 论文通讯作者, 主要研究方向为数字图像处理、计算机图形学、科学计算可视化; 王　聪(1989—), 女, 博士研究生, 主要研究方向为智能信息处理; 甘　玮(1997—), 女, 硕士研究生, 主要研究方向为数字媒体技术.

Key words: facial animation; facial action units; blendshape model; deep learning

早在 1978 年，著名心理学家 Ekman 等[1]将面部肌肉运动和人脸表情相联系，提出了面部动作编码系统(facial action coding system, FACS)，通过对人脸肌肉运动进行详细分析，将表情动作划分为多个运动单元(action unit, AU)，并且定义了 6 种基本表情：喜悦、愤怒、悲伤、厌恶、恐惧和惊讶. 但是，面部表情在个体间的差异性也比较明显，不同人的五官特征均不相同，并且即使它们表达同样的吃惊表情，每个人也会有自己独有的特点，这也给关于人脸表情的研究带来了挑战.

人脸动画技术旨在从源主体中获取其脸部表情信息和头部姿态信息，驱动目标模型生成表情动画. 该技术在数字娱乐领域、虚拟现实领域都有极高的应用价值. 对于动漫产业，可利用计算机自动合成人脸表情模型，从而节省了大量的人工设计工作. 在虚拟现实领域中，虚拟世界的角色通过人脸动画技术可以模拟人类的各种面部表情，从而实现更为友好的人机交互.

现阶段的人脸动画方法受限于高昂的面部表情捕捉设备、不准确的表情系数计算和复杂的三维人脸表情模型重建. 针对这些问题，基于深度学习的人脸动画方法利用普通单目摄像头获取视频图像，基于深度学习技术对整幅人脸图像的表情参数进行提取，进而基于 blendshape 模型生成人脸表情动画.

1 相关工作

关于人脸动画的研究越来越受到国内外科研人员的重视. 其中，Cao 等[2]于 2013 年提出了针对特定用户实时的人脸动画系统，该系统不需要借助于特殊的外部设备，通过对预先采集一系列规定表情和姿势的图像，并对每幅图像上的人脸特征点位置进行半自动标定，以训练出针对该用户的三维人脸形状回归器；根据二维特征点和三维人脸形状上顶点的对应关系，计算出相应的人脸姿势参数和表情系数. 之后，Weng 等[3]提出了运行在移动设备上的基于表演的人脸动画系统，它基于二维图像直接获取人脸的动作参数和表情系数，在满足了跟踪准确性的前提下，减少了所需回归的目标的维度，省去了三维特征点回归的步骤，直接回归姿态参数和表情参数；与此同时，也提高了目标模型表情变换的逼真度. 以上 2 种方法虽然在实时性和准确性上都有所保证，但都是对特定用户进行表情数据获取，以驱动目标模型；当要获取任意新用户的表情数据时，都需要耗费大量的时间进行烦琐的数据采集、标定及回归器的训练工作. 针对现有技术的不足，Cao 等[4]通过公开的三维人脸表情数据集直接学习得到回归器，并基于带偏移量的动态表情模型(displaced dynamic expression, DDE)，根据二维图像回归出精确的人脸特征点位置和三维人脸形状. 回归出的人脸特征点的二维信息也将用于调整摄像机内参矩阵和用户身份，以更好地匹配当前用户的表情信息，回归阶段和调整阶段交替进行，即可快速精确地进行人脸表情获取并驱动目标模型进行表情动画. 之后，Thies 等[5]继续对人脸表情生成技术进行研究，利用非刚性模型的捆绑技术来处理面部身份的恢复问题；并基于稠密光照一致性原理跟踪源视频序列与目标视频序列的表情变化，通过对目标模型快速有效的变形来实现面部表情重现. 本节将从人脸表情数据获取和表情动画生成来具体介绍人脸动画技术的研究现状.

1.1 人脸表情数据获取

人脸表情数据获取可以借助于特殊设备直接获取精确的三维人脸形状. Li 等[6]利用由 8 个红外摄像头组成的设备，在人脸上贴上标记点，以跟踪定位得到人脸上的关键点；并将其映射至人脸上的稠密点集，从而得到人脸的三维模型. Weise 等[7]利用三维数字化扫描仪获取到投影结构光图谱，从而得到稠密且质量较高的表情数据. Weise 等[8]于 2011 年结合深度相机 Kinect 直接获取人脸的 RGB-D 信息，基于动态表情模型实时捕获人脸的刚性和非刚性参数，从而估计出人脸表情数据.

上述方法虽然获取的表情数据较为准确，但设备昂贵、对于表演者具有的侵入性、应用场景受限等也是其不可忽视的缺点. 故而基于普通网络摄像头的表情动作捕捉对于普通用户更为适合，可以通过人脸检测、人脸关键点定位和人脸对齐技术进行人脸表情动作捕捉. 传统的人脸关键点定位技术主要有 ASM, AAM 和 CLM 等方法. 基于回归分析的跟踪方法，尤其是级联形态回归(cascade pose regression, CPR)方法[9]也被广泛用于人脸特征点定位. 该方法给定形态初始猜测值，利用不同

的回归算子不断修正，迭代多次至最后一级，从而得到最终输出. Cao 等[10]基于 CPR 算法，利用随机蕨(random fern)进行双层级联回归，通过最小化回归误差直接回归出人脸三维特征点. Xiong 等[11]提出有监督梯度下降法，其学习训练数据 SIFT 特征的梯度下降方向，相比于之前的方法极大地提高了定位效率. 近几年，由于硬件设备的发展、GPU 的广泛应用，基于深度学习的人脸关键点定位技术取得了巨大的成功. Luo 等[12]利用深度信念网络(deep belief network, DBN)进行分层解析定位人脸关键点. Wu 等[13]基于受限玻尔兹曼机实现人脸在不同姿态、不同表情下的特征点定位. Zhang 等[14]提出基于多任务的深度学习检测方法，可以在准确定位人脸特征点的同时对性别等其他任务进行学习. 在此基础上，Wu 等[15]使用彩色图像对全连接层的特征图像进行高斯混合模型分类，然后利用具有相似特征的图像训练其对应的回归器. 此外，Zhang 等[16]基于多任务级联卷积网络(multi-task cascaded convolutional networks, MTCNN)同时完成人脸检测和人脸对齐这 2 个任务，先通过 PNet 提取 region proposal 并获得概率图，从而得到人脸检测矩形框；接着通过 RNet 对矩形框进行修正移除重叠的矩形框；最后，利用 ONet 重新判断矩形框中的人脸图像是否合格并回归出人脸的 5 个特征点. 与传统的基于可变形模板的方法和基于回归分析的方法相比，基于深度学习的人脸特征点检测更加精确，但对硬件要求较高，同时在训练模型时非常耗时且需要很多训练技巧.

1.2　表情动画生成

表情动画生成可以基于融合变形模型(blendshape 模型)来实现，Lewis 等[17]通过采集人脸在不同姿态和不同表情下的表情模型数据，新的人脸表情模型则由已有的模型插值得到. Cao 等[18]于 2014 年发布了三维人脸表情模型数据库 FaceWarehouse，其利用深度相机对每个用户采集一个中立表情和 19 个特殊表情，并基于双线性模型生成该用户的 47 个三维表情形状. 此后大多数研究方法均使用该数据库，并基于融合变形模型进行人脸表情生成.

根据肌肉模型生成表情动画是基于人脸的面部解剖学原理，利用计算机对人脸肌肉进行模拟，从而生成各种特定的表情运动. FACS 将人脸分为多个 AU，每一个 AU 都控制面部的一部分肌肉，从而控制面部各种运动行为. Waters[19]将人脸分为骨骼层、肌肉层和皮肤层；骨骼控制肌肉的运动，肌肉带动皮肤，进而形成多种多样的表情. Lee 等[20]进一步对该模型进行改进，提出了简化的 3 层模型分别模拟生物组织、肌肉组织层和颅骨结构，并增加了牙齿等器官；该模型可以逼真准确地模拟人脸表情变化，但结构过于复杂，计算量过大且需要进行人为调节.

参数化模型旨在基于人脸相似的拓扑结构，根据设置好的参数值和参数规则将其进行组合，进而生成丰富的人脸表情动作；它可以克服插值法的一些限制，更加直观地控制人脸外形. Pandzic 等[21]提出了以 MPEG-4 形式的标准人脸动画参数模型，在 MPEG-4 标准中，明确了人脸定义参数(facial definition parameter, FDP) 和人脸动画参数(facial animation parameter, FAP). 其中，FDP 利用 84 个特征点对中性表情人脸进行了重新定义，包括人脸的几何坐标、位置信息、纹理信息、场景信息及特征点需满足的数量关系；FAP 参数在中性表情人脸上实现变形，以描述人脸的动态表情变化.

基于肌肉模型进行表情生成时，对骨骼、肌肉、皮肤运动关系的配置都需要大量的工作. 骨骼蒙皮模型则将面部结构简化为 2 层，分别为骨骼层和皮肤层. 现有的蒙皮技术主要有刚性绑定和柔性绑定，其中刚性绑定算法中一个皮肤顶点只受一个骨骼控制，而柔性绑定算法中皮肤上某个顶点的最终位置由多条骨骼控制. 商业软件 Maya, 3D Max 中已集成的柔性绑定算法可以非常方便地布置骨骼、绑定皮肤. 此外，Kavan 等[22]也实现了骨骼蒙皮模型的自动制作. Li 等[23]基于脸部表情肌建立人脸表情骨骼，用条形骨骼模拟线性肌、环形骨骼模拟眼轮匝肌，同时用 2 个眼部控制器来专门控制眼球转动动作.

2　本文系统框架

本文提出一种基于深度学习回归表情参数并生成人脸动画的方法，其总体框架如图 1 所示. 首先从单目摄像头获取到视频帧图像，基于有监督的梯度下降法检测人脸的二维特征点，并对图像进行裁剪，以缩小提取人脸特征时的搜索空间. 然后，裁剪过后的图像将通过卷积神经网络 VGG-Face 得到人脸的特征向量，该特征向量会再通过一个简单的 3 层神经网络，并基于欧氏距离损失回归出 AU 参数. 最后，基于 Blendshape 模型，

利用 AU 参数和与之对应的三维人脸表情模型生成逼真的人脸动画.

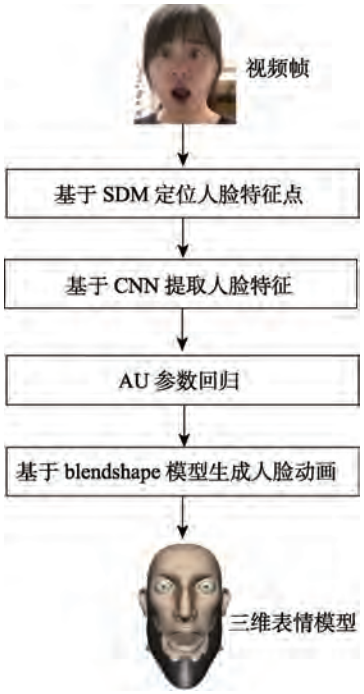

图 1 本文方法的总体框架

3 表情 AU 参数获取

3.1 数据集

为获得足够的人脸表情训练数据，本文采集并建立了人脸表情数据集 FEAFA(a well-annotated dataset for facial expression analysis and 3D facial animation). FEAFA 数据集基于 FACS 重新定义了不包括中立表情在内的 24 个 AU，分别是左眼闭合、右眼闭合、左眼睑提升、右眼睑提升、左眉毛向下、右眉毛向下、左眉毛上扬、右眉毛上扬、下巴向下、下巴左移、下巴右移、左嘴角上扬、右嘴角上扬、左嘴角外展、右嘴角外展、抿上嘴唇、抿下嘴唇、下巴向外、上嘴唇向上、下嘴唇向下、嘴角向下、嘟嘴、脸颊鼓起及鼻子皱起.

为了使标注过程更加简便，同时使定义的每一个 AU 对应于一个三维人脸表情形状，本文首先对 FACS 中定义的 AU 进行细分，如将 FACS 的眼睛闭合 AU 视为左眼闭合和右眼闭合 2 个不同的 AU；同样的细分方式适用于 FACS 中的上眼睑提升、眉毛下压、外眉毛上扬、嘴唇伸展和下巴滑动. 另外，将抿嘴细分为抿上嘴唇和抿下嘴唇 2 个不同的 AU，本文用中立表情和 24 个 AU 来刻画在不同个体上普遍的人脸表情.

本文对 122 个参与者在自然环境下采集出 123 段视频. 122 个参与者均来自亚洲，其中，戴眼镜的有 40 人，有刘海的有 38 人，年龄在 0~15 岁之间的有 4 人，15~60 岁之间的有 110 人，61~80 岁之间有 8 人. 此外，每段视频持续 7~112 s；每段视频包含 4~29 个人脸表情，视频均在非实验室环境(自然环境)下录制；因此每个视频有不同的光照条件和不同的帧率.

视频帧收集完成之后，经过训练的标注人员对每一个视频帧所涉及的 AU 进行标注. 每一个 AU 参数用 0~1 的浮点数刻画特定表情下各个 AU 相对于其中立情况的变形程度. 偏离程度越低，该 AU 参数值越小即接近于 0；偏离程度越高，该 AU 值越高即接近于 1. 本文最终建立的 FEAFA 数据集包含 99 356 幅标注完成的人脸表情图像.

3.2 人脸检测

在获取人脸表情数据时，首先对视频序列中取得的每一帧图像进行人脸检测. 本文采用 Xiong 等[11]提出的有监督的梯度下降法(supervised descent method, SDM)进行人脸特征点检测. 给定人脸的初始形状，SDM 算法不断迭代使得初始人脸形状逐渐回归至真实人脸形状，求解公式为

$$f(x_0+\Delta x)=\left\|h(d(x_0+\Delta x))-\phi_*\right\|_2^2 \tag{1}$$

其中，$h(d(x_0+\Delta x))$ 表示从该人脸图像上的特征点提取出的 SIFT 特征；ϕ_* 表示人脸真实特征点的 SIFT 特征；给定初始形状 x_0，将 x_0 回归至人脸真实形状 x_*，需求解出使得 $f(x_0+\Delta x)$ 最小的 Δx. 由于 SIFT 算子不可导，梯度下降方向和牛顿方向均不可用，故该方法将目标函数的导数，也就是增量 Δx 描述为简单的线性结构，不需要计算牛顿法中的 Hessian 矩阵和 Jacobian 矩阵，可直接利用最小二乘法进行求解. SDM 算法通过改变对增量的计算方法，在保证人脸特征点检测的鲁棒性的同时，大大降低了计算量. 表情图像在进行人脸检测之后，对其裁剪、缩放以满足后续卷积神经网络的输入要求.

3.3 基于 VGG-Face 回归 AU 参数

VGGNet[24]通过堆叠小尺寸的卷积核建立更深层次的网络结构，它不仅在图像分类任务中表现优异，并且适合于多种迁移学习任务，是从图像中提取 CNN 特征的首选算法. 本文将裁剪过后的表情图像通过 VGG-Face 得到人脸的 2 622 维特征向量，再通过一个简单的 2 层全连接神经网络结构以学习 24 维 AU 参数向量. 每一个全连接层都添

加 ReLU 层以实现非线性性，并且，不同于手动降维，添加 Dropout 层使神经网络自动决定对于 AU 参数回归任务最适合的特征维度．网络的具体结构如图 2 所示．

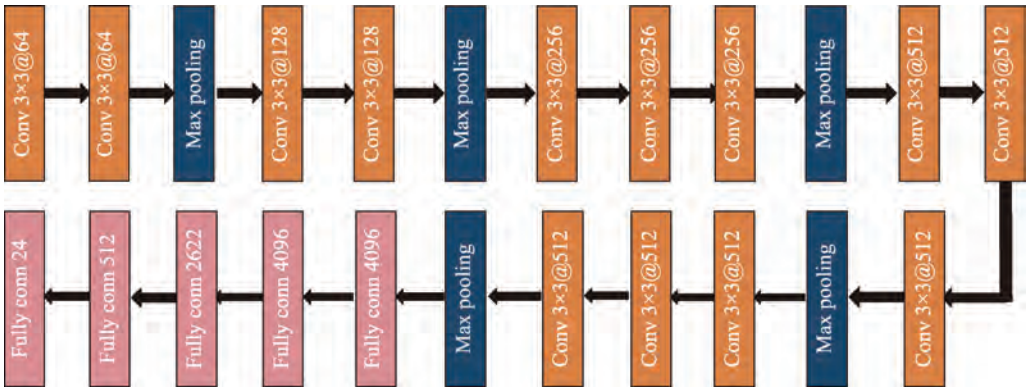

图 2　AU 参数回归网络结构

不同于 AU 密度估计算法单独计算每一个 AU 所对应的损失，本文使用欧氏距离损失函数衡量 24 个 AU 参数的全部损失，总结为

$$L_{\text{train}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i\right\|_2^2 \tag{2}$$

其中，L_{train} 代表所有训练集样本的平均损失；$\hat{\boldsymbol{y}}_i$ 和 $\boldsymbol{y}_i$ 均为 24 维向量；$\boldsymbol{y}_i$ 为 AU 参数的真值，即标注值，$\hat{\boldsymbol{y}}_i$ 为 AU 参数的估计值，即回归结果；N 代表训练样本个数．相比于基于人脸的二维或三维特征点进行表情参数提取，本文将整幅人脸表情图像通过卷积神经网络回归出更为准确的表情参数．此外，由于各个 AU 之间的相互影响，相比于已有的 AU 检测算法独立地识别每个 AU 是否存在，本文对 24 个 AU 参数进行联合回归，让神经网络更好地学习各个 AU 之间的依赖关系．

4　人脸动画生成

在得到准确的人脸表情数据之后，本文将基于该表情数据驱动三维虚拟人物生成表情动画．Cao 等[18]于 2014 年发布基于 FACS 的三维人脸表情数据库 FaceWarehouse，利用深度相机 Kinect 采集年龄分布在 7~80 岁的用户在不同表情下的 RGB-D 数据，包括一个中立表情和 19 个其他表情，同时对面部特征点进行标注．之后对初始人脸三维形状模板变形，使其分别符合该用户不同表情下的深度信息，并建立二维图像上人脸特征点与三维面部网格上顶点的对应关系．在此基础上，FaceWarehouse 借鉴 Li 等[25]的方法，针对每个用户生成其中立表情形状和 46 个其他三维表情形状；这些三维形状网格具有相同的拓扑结构，故可将其组成三阶张量 $\boldsymbol{C}_r$ (11k 形状顶点×150 个体×47 个表情)．当张量中的用户身份系数确定时，即可计算出该用户的表情融合模型．基于该模型和三维表情模型系数即可生成该用户的任一特殊表情形状，可总结为

$$\boldsymbol{F} = \boldsymbol{C}_r \times_2 \boldsymbol{w}_{\text{id}}^{\text{T}} \times_3 \boldsymbol{w}_{\text{exp}}^{\text{T}} \tag{3}$$

其中，$\boldsymbol{w}_{\text{id}}$ 是用户的个体系数；$\boldsymbol{w}_{\text{exp}}$ 是表情系数．基于 FaceWarehouse 数据集，本文将选取中立表情形状和其中可表达人脸复杂表情的 24 个三维表情形状，并建立其与自定义 AU 参数的对应关系，使得每一个 AU 对应于一个三维表情形状，部分对应关系如图 3 所示．

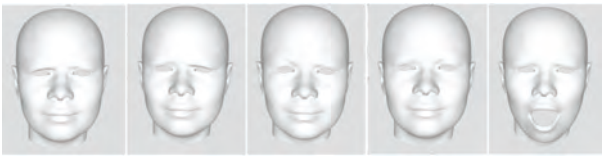

图 3　部分 AU 与三维人脸表情形状对应关系

事实上，任何具有中立表情形状和 24 个对应表情形状的虚拟人物该方法都可以驱动生成特定表情下的人脸动画，可以基于已有的三维模型数据库也可以自己构建表情基模型．利用卷积神经网络得到的源用户在特定表情下的 AU 参数，根据 AU 参数来对目标对象所对应的 25 个三维表情形状进行形状融合即可得到特定表情下的三维人脸模型．并且，基于 EPnP 算法[26]可以计算出头部运动的刚性变换，所以某一用户在特定姿态和特定表情下的三维人脸模型 $\boldsymbol{F}$ 可总结为

$$\boldsymbol{F} = \boldsymbol{RH} + \boldsymbol{t} \tag{4}$$

其中，$\boldsymbol{H} = \boldsymbol{B}_0 + \sum_{i=1}^{24} e_i(\boldsymbol{B}_i - \boldsymbol{B}_0)$ 表示具有特定表情的三维人脸形状，$\boldsymbol{B}_0$ 表示中立表情的三维人脸形状，$\boldsymbol{B}_i$ 为第 i 个 AU 对应的三维人脸表情形状；e_i 为回归得到的第 i 个 AU 参数值. 结合人脸姿态估计出的头部旋转矩阵 $\boldsymbol{R}$ 和平移向量 $\boldsymbol{t}$，并对三维人脸网格进行渲染即可生成逼真的人脸表情动画.

5 实验与分析

本文基于 Jia 等[27]开发的深度学习框架 Caffe (convolutional architecture for fast feature embedding)进行卷积神经网络的搭建，训练均在具有 12 GB 显存的 NVIDIA Titan X GPU 上进行. 在验证阶段，所有的网络参数(权重、偏置)和超参数(学习率、迭代次数、权重衰减)均被调整至最优以最小化损失函数. FEAFA 数据集在进行人脸特征点定位操作后，得到 94 649 幅表情图像；考虑同一个参与者同时在训练集和验证集中出现会影响回归结果，本文依据不同的参与者划分训练集和验证集. 由于数据集规模不大，本文只划分了训练集和验证集，并未留出测试集. 其中，训练集包含 106 位参与者和 78 343 幅表情图像；验证集包含 16 位参与者和 16 306 幅表情图像.

事实上，网络参数的初始化策略对最终的回归结果和泛化能力都有着重要的影响，目前广泛运用的初始化策略是基于高斯分布或均匀分布的，但是这 2 个分布所能提供的信息少之又少. 本文采用面向人脸图像分类 VGG-Face 训练模型 VGG_FACE.caffemode 作为预训练模型，以提供更丰富的人脸信息；该模型由牛津大学的 VGG(visual geometry group)小组提供. 然后，对网络进行 fine-tune，以得到 AU 参数回归的训练模型，并将在 5 个测试集上的结果进行平均，得到最终的回归结果.

为比较不同网络结构对 AU 参数回归的影响，本文对比了基于 VGG_Face 与基于 AlexNet，GoogLeNet 和 VGG_16 的网络模型的损失大小. 从 AlexNet，GoogLeNet 和 VGG_16 的最后一个全连接层得到人脸的 1 000 维特征向量，并在其之后加上维度为 512 的中间层和维度为 24 的输出层. 对这 3 个网络分别采用公开发布在 Caffe Model Zoo 中的 bvlc_alexnet 模型、bvlc_googlenet 模型和 VGG_ILSVRC_16_layers 模型作为 AU 参数回归的预训练模型. 表 1 展示了利用这 4 种神经网络提取人脸特征所得到 AU 参数回归的欧氏距离损失；其中，训练集包含 106 位参与者的 78 343 幅表情图像，验证集包含 16 位参与者的 16 306 幅表情图像. 实验结果表明，基于 VGG-Face 的网络模型回归出来的 AU 参数更为准确.

表 1 基于各个神经网络回归出 AU 参数的损失大小

特征提取器	欧氏距离损失
AlexNet	0.352
GoogLeNet	0.311
VGG-16	0.280
VGG-Face	0.268

在得到回归出的 AU 参数之后，结合虚拟人物的中立表情形状和与 AU 参数相对应的 24 个三维表情形状，以及计算出的头部姿态信息，即可驱动其生成人脸动画. 图 4 展示了驱动一个简单的虚拟人物生成人脸动画的结果，初始模型来源于文献[28-29].

於俊等[30]提出一种基于单目摄像机由视频驱动的人脸动画方法，在不包含渲染三维人脸模型的情况下，其合成一帧动画平均所需时间为 1.81 s；而本文基于深度学习网络 AlexNet 合成一帧人脸动画仅需要 0.05 s. 此外，文献[2,4-5]的人脸动画系统也与本文类似，但是这些方法都是基于特征点来重建三维人脸以计算表情系数，而基于特征点的方法只能获取人脸的部分表情信息；本文则基于整副人脸表情图像回归表情参数，可以获得更加完整的人脸表情信息，使获得的表情参数更为准确，从而生成更为逼真细腻的人脸动画.

图 4 人脸动画生成效果

6　总结与讨论

本文提出了一种基于表情 AU 参数和深度学习的人脸动画生成方法，它不需要对相机手工标定，也不需要对用户进行预先训练，基于任一视频帧即可驱动虚拟人物实现人脸动画．首先从普通单目摄像头获取视频图像，并基于有监督的梯度下降法对视频帧进行人脸检测，利用卷积神经网络对人脸图像进行特征提取；基于此构建出对 24 个 AU 参数进行回归的网络模型，AU 参数即为三维人脸表情基系数．对源视频序列的每一帧表情图像准确地回归出 AU 参数值，并结合虚拟人物相对应的 24 个基础三维表情形状和中立表情形状，基于表情融合变形模型驱动虚拟人物进行人脸动画生成．该方法考虑了 AU 参数之间的相互影响，使得生成的人脸动画的表情更加自然、细腻；并且，它基于人脸图像回归表情系数，比基于特征点回归出的表情系数更加准确，同时可以更好地适应快速运动和剧烈的光照变化．

参考文献(References):

[1] Ekman P, Friesen V W. Facial action coding system: manual[M]. Palo Alto: Consulting Psychologists Press, 1978

[2] Cao C, Weng Y L, Lin S, *et al*. 3D shape regression for real-time facial animation[J]. ACM Transactions on Graphics, 2013, 32(4): Article No.41

[3] Weng Y L, Cao C, Hou Q M, *et al*. Real-time facial animation on mobile devices[J]. Graphical Models, 2014, 76(3): 172-179

[4] 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

[5] Thies J, Zollhöfer M, Stamminger M, *et al*. Face2Face: real-time face capture and reenactment of RGB videos[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2387-2395

[6] Li B B, Zhang Q, Zhou D S, *et al*. Facial animation based on feature points[J]. Telkomnika Indonesian Journal of Electrical Engineering, 2013, 11(3): 1697-1706

[7] Weise T, Li H, van Gool L, *et al*. Face/Off: live facial puppetry[C] //Proceedings of the ACM SIGGRAPH /Eurographics Symposium on Computer Animation. New York: ACM Press, 2009: 7-16

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

[9] Dollár P, Welinder P, Perona P. Cascaded pose regression[C] // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2010: 1078-1085

[10] Cao X D, Wei Y C, Wen F, *et al*. Face alignment by explicit shape regression[J]. International Journal of Computer Vision, 2014, 107(2): 177-190

[11] Xiong X H, de la Torre F. Supervised descent method and its applications to face alignment[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2013: 532-539

[12] Luo P, Wang X G, Tang X O. Hierarchical face parsing via deep learning[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2012: 2480-2487

[13] Wu Y, Wang Z G, Ji Q. Facial feature tracking under varying facial expressions and face poses based on restricted Boltzmann machines[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2013: 3452-3459

[14] Zhang Z P, Luo P, Chen C L, *et al*. Facial landmark detection by deep multi-task learning[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2014: 94-108

[15] Wu Y, Hassner T, Kim K, *et al*. Facial landmark detection with tweaked convolutional neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 3067-3074

[16] Zhang K P, Zhang Z P, Li Z F, *et al*. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503

[17] Lewis J P, Mooser J, Deng Z G, *et al*. Reducing blendshape interference by selected motion attenuation[C] //Proceedings of the Symposium on Interactive 3D Graphics and Games. New York: ACM Press, 2005: 25-29

[18] Cao C, Weng Y L, Zhou S, *et al*. FaceWarehouse: a 3D facial expression database for visual computing[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(3): 413-425

[19] Waters K. A muscle model for animation three-dimensional facial expression[J]. ACM SIGGRAPH Computer Graphics, 1987, 21(4): 17-24

[20] Lee Y, Terzopoulos D, Waters K. Realistic modeling for facial animation[C] //Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM Press, 1995: 55-62

[21] Pandzic I S, Forchheimer R. MPEG-4 facial animation: the standard, implementation and applications[M]. New York: John Wiley & Sons, Inc, 2002

[22] Kavan L, Sloan P P, O'Sullivan C. Fast and efficient skinning of animated meshes[J]. Computer Graphics Forum, 2010, 29(2): 327-336

[23] Li D X, Sun C, Hu F Q, *et al*. Real-time performance-driven facial animation with 3ds Max and Kinect[C] //Proceedings of the 3rd International Conference on Consumer Electronics, Communications and Networks. Los Alamitos: IEEE Computer Society Press, 2014: 473-476

[24] Simonyan K, Zisserman A. Very deep convolutional networks

for large-scale image recognition[OL]. [2018-12-14]. https://arxiv.org/abs/1409.1556

[25] Li H, Weise T, Pauly M. Example-based facial rigging[J]. ACM Transactions on Graphics, 2010, 29(4): Article No.32

[26] Lepetit V, Moreno-Noguer F, Fua P. EP*n*P: an accurate *O*(*n*) solution to the P*n*P problem[J]. International Journal of Computer Vision, 2009, 81(2): 155-166

[27] Jia Y Q, Shelhamer E, Donahue J, *et al*. Caffe: convolutional architecture for fast feature embedding[C] //Proceedings of the 22nd ACM International conference on Multimedia. New York: ACM Press, 2014: 675-678

[28] Free3D. Centaur base mesh 3D model[OL]. [2018-12-14] https://free3d.com/3d-model/free-base-mesh-centaur--67384.html

[29] CG Model. Disney Infinite Vanellope: Model No.71840[OL]. [2018-12-14]. http://www.cgmodel.com/model-71840.html(in Chinese)
(CG 模型网. 迪士尼无限云妮露: 作品编号 71840[OL]. [2018-12-14]. http://www.cgmodel.com/model-71840.html)

[30] Yu Jun, Wang Zengfu. 2D-3D facial video coding/decoding at ultra-low bit-rate[J]. Acta Electronica Sinica, 2013, 41(1): 185-192(in Chinese)
(於 俊, 汪增福. 极低码率下的 2D-3D 人脸视频编解码[J]. 电子学报, 2013, 41(1): 185-192)