面向虚拟化身的人脸表情模拟技术

姚世明[1]， 李维浩[1]， 李蔚清[2]， 苏智勇[1]

(1. 南京理工大学自动化学院，江苏 南京 210094；
2. 南京理工大学计算机科学与工程学院，江苏 南京 210094)

摘 要：为了实现基于增强现实的电子沙盘环境中的异地可视化交互功能，提出了一种面向虚拟化身的三维表情模拟技术。首先，使用 RGB 摄像头跟踪异地作业人员的表情，基于约束局部模型(CLM)提取人脸特征点数据后传输到本地；然后，采用基于径向基函数的插值算法计算虚拟化身面部网格点的坐标，驱动模型模拟出与异地作业人员相同的表情；最后，为了提高变形算法的精度和效率，提出一种基于贪心算法与人脸肌群分布的插值控制点选取和分区域插值方法。实验结果表明，该算法能够满足实际应用对实时性和真实感的需求。

关 键 词：表情模拟；径向基函数；形变模型；肌肉模型
中图分类号：TP 391 **DOI：**10.11996/JG.j.2095-302X.2019030525
文献标识码：A **文 章 编 号：**2095-302X(2019)03-0525-07

Facial Expression Simulation Technology for Virtual Avatar

YAO Shi-ming[1], LI Wei-hao[1], LI Wei-qing[2], SU Zhi-yong[1]

(1. School of Automation, Nanjing University of Science and Technology, Nanjing Jiangsu 210094, China;
2. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing Jiangsu 210094, China)

Abstract: In order to realize the offsite visualization interaction function in the electronic sand table environment based on Augmented Reality, a 3D expression simulation technology for virtual avatar is proposed. First, the expressions of remote worker is tracked using a RGB camera, the face feature point data is extracted based on constrained local models (CLM) and the data is sent to local. The interpolation algorithm based on Radial Basis Function is used to calculate the coordinates of vertices of the virtual avatar face mesh, the model is driven to simulate the same facial expression with the other offsite worker; Finally, in order to improve the accuracy and efficiency of the deformation algorithm, a selection method of interpolation control points and a sub-region interpolation method based on the distribution of the human muscle group and the greedy algorithm are proposed, The experimental results show that the algorithm can meet the real-time and realistic requirements of applications.

Keywords: expression simulation; radial basis function; deformation model; muscle mode

电子沙盘是一种虚拟化的信息显示手段，可以模拟三维、动态、可交互战场态势环境。基于增强现实(augmented reality, AR)的电子沙盘在态势展现的直观性、交互操作的便捷性以及协同研讨的高效性方面都优于传统的电子沙盘。在 AR 电子沙盘的异地协同研讨作业中引入虚拟化身技术，以虚拟人物的形式将身处异地的指挥员“投射”到同一环境，“面对面”进行沟通交流，大大提升了指挥人员交流

收稿日期：2018-11-06；**定稿日期：**2018-12-07
基金项目：“十三五”装备预研项目(315100104，41401010203)；上海航天科技创新基金项目(SAST2018009)
第一作者：姚世明(1993–)，男，河南洛阳人，硕士研究生。主要研究方向为计算机图形学、计算机视觉等。E-mail：455910419@qq.com
通信作者：苏智勇(1981–)，男，江苏泰州人，副教授，博士。主要研究方向为计算机图形学、计算机视觉等。E-mail：suzhiyong@njust.edu.cn

的充分性和高效性。人与人交流中，人脸表情是最主要和最直观的表现形式，具有逼真表情的虚拟化身，使观看者更有沉浸感。

三维人脸表情模拟技术主要包括表情动画的跟踪和模型驱动 2 部分。从上世纪 70 年代 PARKE[1] 建立第一个脸部模型到现在，三维人脸表情动画跟踪技术已经较为成熟，但实现高实时性和高真实性的人脸动画重构技术仍是目前研究的难题。对于异地交互来说，表情模拟的实时性和真实感是重要指标，本文重点研究如何提高表情重构方法的效率和精度，从而满足应用所要求的实时性和真实感。

在人脸特征点跟踪领域，国内外研究人员提出了很多有效方法。WILLIAMS[2]提出在用户面部贴有反光特性的标记点来跟踪人脸特征点运动信息。此方法精度高，但用户体验差。何钦政和王运巧[3]采用 Kinect 实现三维人脸表情参数的捕捉。但其像素较低，要求拍摄距离较近。此外，目前研究较多是基于 RGB 视频的特征点检测方法。CRISTINACCE 和 COOTES[4]提出了基于 CLM 形状模型的特征点检测法。BULAT 和 TZIMITOPOULOS[5]采用基于卷积神经网络的深度学习方法实现了基于 RGB 视频的三维人脸特征点实时跟踪，由于训练数据的限制，检测精度不高。考虑到本文应用对表情模拟的实时性和真实感要求较高，且表情采集者处在小范围移动中。以上方法中，采用高像素 RGB 摄像头跟踪特征点的方案较为合理。

在基于数据驱动的表情重构技术领域，常见方法包括基于肌肉模型、基于表情基合成和基于形变算法等来实现表情动画。基于肌肉模型的方法虽真实感强，但实时性差。而基于表情基合成的方法存在个性化特征不明显的缺点。PIGHIN 等[6]提出一种基于视频的人脸表情模拟技术，采集到特征点运动数据后，采用三维曲面插值算法驱动模型产生表情动画。目前常用的人脸模型变形的算法有拉普拉斯变形算法和径向基插值变形算法。拉普拉斯变形算法的效率高，局部细微表情的变形效果差[7]。径向基插值变形算法的平滑性好，但应用于复杂拓扑结构的人脸曲面时会出现局部失真现象，并且计算量较大[8]。SUWAJANAKORN 等[9]提出了基于语音数据的表情重构方法。GUO 等[10]使用基于卷积神经网络的深度学习方法实现了利用单张图片对人脸实时重构。但是该方法需要大量工作去建立带三维特征点数据的数据库，且对硬件性能要求较高。综上分析，实现高实时性和高逼真度的人脸动画重构仍然是一项具有挑战性的工作。

经过对以上方法的对比分析，结合应用背景，本文提出基于单目 RGB 视频驱动的人脸表情模拟技术。首先基于约束局部模型 CLM 跟踪人脸特征点的运动信息；然后采用基于径向基函数(radial-basis function，RBF)的变形算法驱动人脸网格模型来输出表情动画；并基于人脸肌肉模型和贪心算法对 RBF 插值变形算法的效率和精度进行了优化，提出了基于人脸肌群分布的分区域插值算法和插值控制点选取算法，实现了 AR 电子沙盘环境中面向虚拟化身的表情模拟功能。

1 系统框架

本系统用于 AR 电子沙盘异地可视化交互中虚拟化身的表情模拟，主要实现跟踪和重构人脸的表情动画。系统总流程如图 1 所示：第 1 部分为数据预处理。读取并记录人脸网格模型顶点的索引和坐标，选取插值控制点并与模型对应网格点绑定，计算并保存各控制点与其他网格顶点的欧氏距离；第 2 部分为特征点的运动跟踪。基于 CLM 形状模型在 RGB 摄像头采集的视频图像中搜索定位出特征点坐标；第 3 部分为插值出所有人脸网格模型顶点坐标。用检测到的插值控制点坐标训练插值函数，插值出网格模型中除控制点以外的面部顶点坐标；第 4 部分为模型的驱动。插值的结果传递给模型驱动脚本，驱动模型产生形变来输出表情动画。

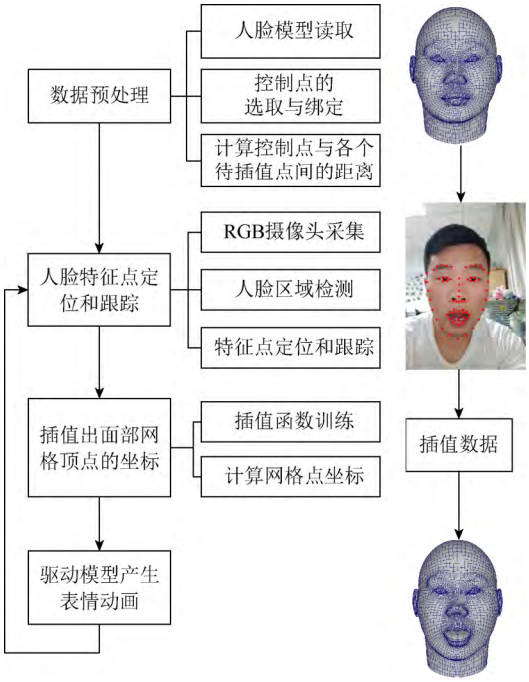


图 1 系统总体流程

2　基于 CLM 模型的人脸特征点跟踪

本文采用基于 CLM 形状模型的特征点检测方法，主要工作包括模型的训练和特征点定位，具体流程如图 2 所示。特征点检测工作一定程度上建立在 SARAGIH 等[11]工作的基础上，采用了其训练好的 CLM 形状模型。特征点定位包括人脸区域定位和特征点搜索定位 2 部分。采用经典的 VIOLA 和 JONES[12]人脸检测器检测出图像中的人脸区域，以此缩小后续的特征点搜索范围；然后基于 CLM 形状模型对人脸区域中的特征点进行拟合定位；最后采用 mean-shift 算法[13]跟踪人脸特征点的运动数据。

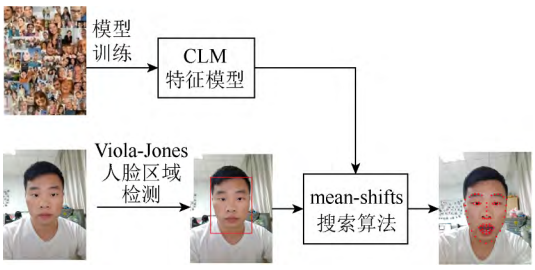


图 2　人脸特征点运动的跟踪过程

3　基于 RBF 插值的变形算法

如何利用检测到的少数特征点运动信息驱动具有大量网格点的人脸模型是本文主要解决的问题。本文选用 RBF 插值算法驱动模型形变产生表情动画，插值控制点从检测到的人脸特征点中选取，训练出人脸曲面插值函数，然后根据前一帧的模型顶点坐标插值出当前帧的顶点坐标数据，生成平滑的人脸表情动画。

3.1　RBF 插值函数介绍

RBF[14]是一种 3 层的前向神经网络，包括输入层、隐含层和输出层。RBF 的基本思想：以核函数构成隐含层空间，并对输入数据进行变换，将低维非线性数据变换到高维空间，使其在高维空间线性可分。RBF 具有结构简单、学习收敛速度快、能够逼近任意非线性函数、有效克服局部极小值问题等优点。RBF 插值函数的数学表示为

$$f(x)=\sum_{i=1}^{n}\lambda_i\varphi(\|x-c_i\|) \tag{1}$$

$$\varphi(x)=e^{\frac{\|x-c\|^2}{2\sigma^2}} \tag{2}$$

其中，$f(x)$为径向基插值函数；λ_i 为第 i 个控制点的权值；n 为插值控制点的数目；$\|x-c_i\|$为网格点 x 到插值控制点 c_i 的欧氏距离；$\varphi(x)$为隐含层核函数，隐含层到输出层采用的是简单的权值连接，将高斯径向基函数值线性加权得到训练的输出。网络结构如图 3 所示。

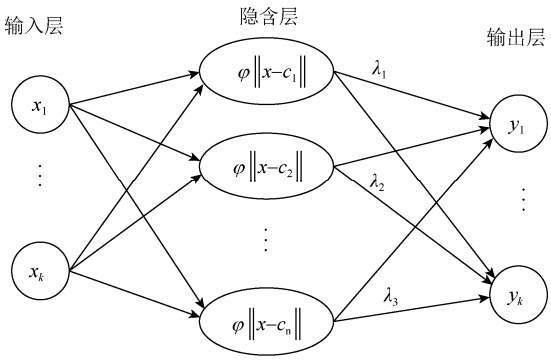


图 3　RBF 神经网络拓扑结构

式(2)为本文选用的高斯核函数。对于局部性较强的人脸，选择具有局部性的高斯核函数更有优势。在函数作用范围内，x 距离控制点 c 越近时，其函数值越大，超过一定范围后函数值接近于 0，其作用范围的大小由 σ 决定，即

$$[\lambda_1\ \ \lambda_2\ \ \cdots\ \ \lambda_n]\cdot\begin{bmatrix}\varphi\|x_i-c_1\|\\ \varphi\|x_i-c_2\|\\ \vdots\\ \varphi\|x_i-c_n\|\end{bmatrix}=f(x_i)\ \ (i=1,2,3,\cdots,n) \tag{3}$$

其中，λ 为所求的未知权值矩阵；φ 为核函数；$f(x)$为训练数据集；n 为插值控制点数目。训练插值函数就是利用特征点检测获得的数据集 $f(x)$来求出权值 λ 的大小。

3.2　基于 RBF 的人脸网格模型驱动

RBF 变形算法被广泛用于各种曲面的插值。本文将其应用于人脸模型驱动，主要包括以下工作，其中前 5 步为准备工作。

(1) 在识别的 66 个特征点中选取合适的点作为插值控制点，用于训练插值函数；

(2) 读取虚拟化身模型，记录面部网格点索引和坐标；

(3) 将选取的控制点与模型中对应位置的顶点绑定，形成映射关系；

(4) 根据肌肉模型将待插值网格点分区；

(5) 计算并存储各分区待插值点与插值控制点之间的距离，方便后续核函数的调用；

(6) 用每帧视频检测出的控制点数据训练 RBF

插值函数，计算出虚拟化身面部顶点在当前帧的坐标，传递给模型驱动脚本，从而产生对应的表情动画。

4 基于肌肉模型的RBF插值变形算法

RBF被广泛用于各种曲面的插值，在应用于人脸曲面变形时，由于人脸肌群分布有很强的局部性，并且存在眼睛和嘴巴此类孔洞模块，若直接将现有的径向基插值函数直接应用到整张人脸，实验表明将会导致表情局部失真现象。并且由于每帧都要训练插值函数，导致RBF插值变形算法的计算量较大。本文针对人脸面部的局部特殊性对RBF插值变形算法的精度和效率进行优化，从而提高表情重构的逼真度和实时性。

4.1 基于肌肉模型的分区域插值法

人脸表情的变化是通过面部肌肉的运动所产生的，有超过26块以上的肌肉对人脸的面部表情产生影响。图4为人脸的主要肌群分布示意图。高斯核函数的特性是距离控制点越近则受到的影响越大。而人脸中上下嘴唇上相邻点之间的距离很近，但实际相互影响很小，若不考虑此问题，势必会导致眼睛和嘴巴的动画失真，如图5实验结果所示，眼睛和嘴巴的张合尺度小于真实状态。为此，本文提出基于人脸肌群分布的分区域插值思想。

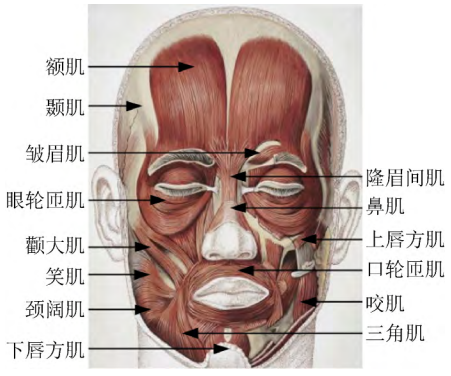
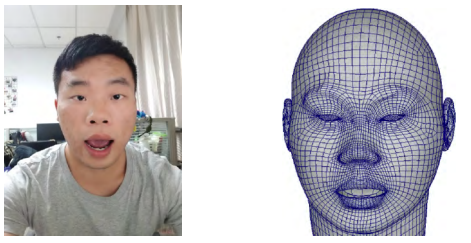


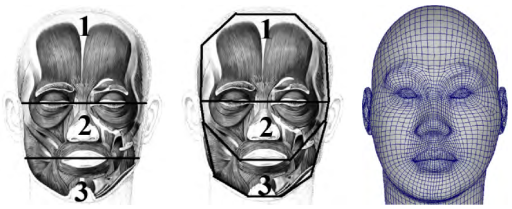
图4 人脸肌群分布图



(a) 输入表情 (b) 表情失真现象

图5 人脸局部特性导致的失真

人脸表情是由人脸肌肉群所驱动的。现有的分区域插值方法[8]仅利用对称性从眼睛和嘴巴上下的水平分割线将模型分为多区域，如图6(a)所示。本文根据人脸的肌群分布、人脸对称性以及孔洞区域的特殊性将待插值的人脸网格模型按照图6(b)分为3个区域分别进行插值。图6(c)为分区的人脸网格模型。



(a) 传统分区方法 (b) 本文分区方法 (c) 待分区模型

图6 网格模型分区示意图

对于各分区的临界处，相邻分区对临界的点都会产生一个插值位移数据，且大小和方向均不一致。从图6(b)可以看出，各分区的相邻边界线基本都处在各肌群块的临界处，这些位置的点在表情动作中发生的位移变化本身是很小的。此外，由于边界距离大部分特征点较远，得到的插值数据较小，实验也表明边界处点的位移变化值是很小的。所以本文选择较为简单的求和取均值的方法来解决边界过渡问题。

4.2 基于贪心算法和肌肉模型的控制点选取算法

电子沙盘环境中的异地可视化交互对实时性的要求较高，RBF插值算法的计算量与控制点的数目和待插值点的数目成正比。为了提高算法效率，应选择尽可能少的插值控制点来满足应用所需的表情逼真度。RENDALL和ALLEN[15]采用贪心算法来完成曲面插值控制点的选取工作。余重基和李际军[16]提出根据面部肌群分布来选取插值控制点。

本文结合贪心算法与人脸肌群分布提出一种新的插值控制点的选取方式。将人脸网格模型按照图6(b)分为3个区域分别进行插值，每个区域都要从本区域中选取部分特征点作为插值控制点，选取的插值控制点数据用来训练插值函数，图7为人脸的特征点检测结果。

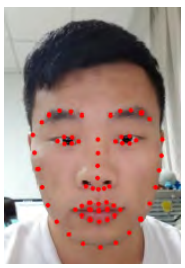


图7 人脸特征点分布示意图

插值控制点选取的基本思想过程：以图 6(b)区域 1 的控制点选取为例，从图 7 中看出该区域有 20 个人脸特征点(处在 2 个区域分割线上的特征点为 2 个区域共用)。首先要根据肌肉模型从该区域 20 个特征点中选取初始控制点。区域 1 中眉毛主要受皱眉肌和额肌的驱动，两者分别控制眉毛做横向和纵向运动，故选取眉毛两端及中间共 2 个特征点作为控制点，同理，上眼睑选取两眼角和上端共 3 个特征点作为控制点。于是，得到 12 个初始控制点集合 $C^0=(c_1,c_2,c_3,\cdots,c_{12})$。设检测到的特征点坐标值为 $g(x)$，将初始控制点的坐标数据带入式(3)，得到

$$[\lambda_1 \quad \lambda_2 \quad \cdots \quad \lambda_{12}]\cdot\begin{bmatrix}\varphi\|c_i-c_1\|\\ \varphi\|c_i-c_2\|\\ \vdots\\ \varphi\|c_i-c_{12}\|\end{bmatrix}=f(c_i)\ (i=1,2,\cdots,12) \tag{4}$$

计算权值系数矩阵 $\boldsymbol{\lambda}=[\lambda_1 \quad \lambda_2 \quad \cdots \quad \lambda_{12}]$，从而得到

$$f(x)=\sum_{i=1}^{12}\lambda_i\varphi(\|x\text{-}c_i\|) \tag{5}$$

然后把其他 8 个特征点前一帧的坐标带入式(5)，插值出当前帧的坐标 $f(x_i)$。并按照式(6)计算插值误差 $\Delta^0=(\Delta_1,\Delta_2,\Delta_3,\cdots,\Delta_8)$，即

$$\Delta=|f(x_i)-g(x_i)| \tag{6}$$

按照贪心算法的思想，将插值误差最大的特征点加入控制点集合 C 得到 C_1。重新以 C_1 作为控制点集合进行插值并比较误差，重复迭代，直至该区域的所有特征点插值误差小于设定的误差上限时，即完成该区域的控制点选取工作。插值控制点选取算法流程如图 8 所示。

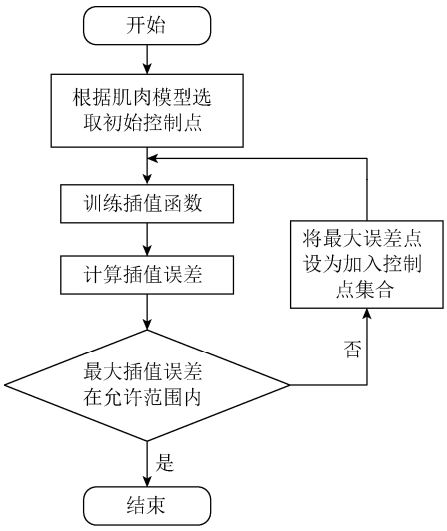


图 8　插值控制点选取算法流程

5　实验和分析

采用第三方软件 FaceGen 软件生成三维人脸模型作为表情动画的载体，人脸模型的网格顶点为 1 781 个。实验中，系统的输入为 RGB 视频流，采用本文的人脸特征点检测和分区域 RBF 插值算法获得人脸模型网格顶点的实时坐标数据，然后通过 socket 通信将数据发送给 Unity 3D 脚本，驱动模型产生动画。

5.1　常见表情的模拟效果

图 9 为本文方法对常见几种表情的模拟效果。系统的表情模拟速度在 24 fps 左右基本满足应用要求的实时性。

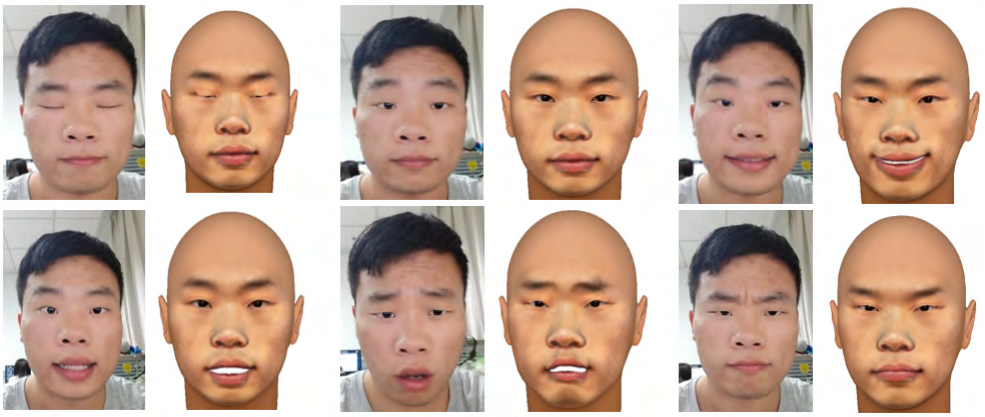


图 9　常见表情的模拟效果

5.2　实验对比分析

图 10 展示了本文的基于人脸肌肉模型的分区域 RBF 插值变形算法的优化效果。采用传统的方法[16]直接对整张人脸网格插值并驱动模型产生

的表情在眼睛和嘴巴处的失真较为明显，主要表现为张合尺度小于实际输入。实验结果表明，本文方法在嘴巴和眼睛处有较好的优化效果，输出的表情更接近于真实的表情。

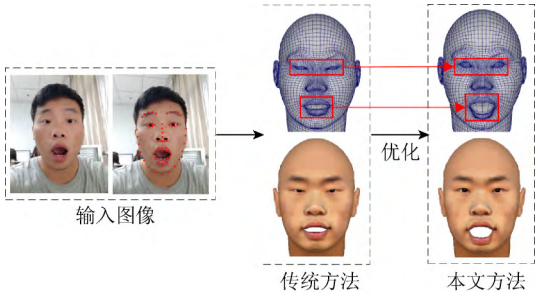


图 10 分区域 RBF 插值变形算法的优化效果

图 11 为表情模拟逼真度的实验对比。图中左侧是文献[7]基于拉普拉斯变形算法的实验结果，右侧为本文基于 RBF 插值变形算法的表情模拟效果。两者均采用数据驱动人脸网格模型形变来产生表情动画。实验结果表明，在整体上 2 种方法均实现了表情的还原，但本文方法在局部细微表情的处理上表现更好。特别是在眼睛和眉毛等局部区域的表情模拟效果更接近原表情，这得益于高斯核函数的局部特性。

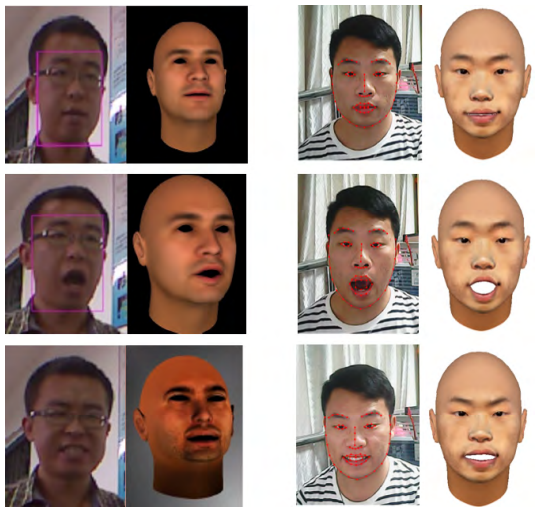


图 11 表情模拟效果的实验对比

5.3 AR 环境中的表情模拟效果测试

本文的研究目的在于实现面向虚拟化身的表情模拟功能，并将其应用于 AR 电子沙盘环境中的异地可视化交互功能。实验中，将表情模拟功能集成到 unity3d 应用中，并部署到微软的 AR 设备 HoloLens 上。实验效果如图 12 和图 13 所示，两图均为 HoloLens 设备的截图。

图 12 中左端为 PC 机和 RGB 摄像头对人脸特征点的实时检测画面，右端为佩戴 HoloLens 设备的观察者所看到添加在现实环境中的虚拟表情模拟效果。

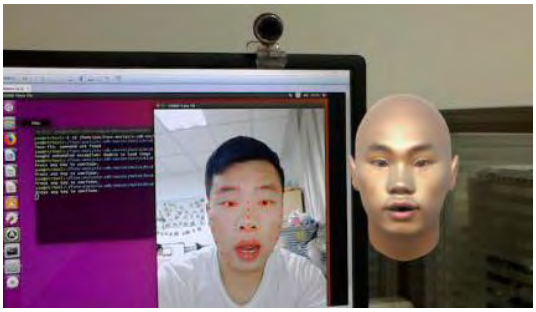


图 12 基于增强现实的表情模拟效果

图 13 为 AR 电子沙盘异地可视化交互场景中虚拟化身的表情模拟效果。图中虚拟化身代表异地的作业人员，右侧展示了走近虚拟化身时近距离观察面部表情的效果。

图 13 AR 沙盘环境中虚拟化身的表情模拟效果

近几年基于深度学习进行三维人脸建模的研究工作有很多[5,9-10]，其中文献[10]基于卷积神经网络方法不仅实现了人脸表情的实时重构，而且表情细节还原度很高，每帧重构时间在 20 ms 左右。但是基于深度学习的方法需要大量的前期工作来建立带人脸三维数据的数据库，并且对硬件的性能要求较高，而可穿戴设备 HoloLens 的计算能力较低。本文的特征点检测和插值工作在普通 PC 机中完成，驱动模型产生表情动画的工作在 HoloLens 中实现，此方法工作量较小，对硬件性能要求低，而且表情重构的实时性和逼真度也满足了应用的需求。

6　结论及展望

本文采用基于 CLM 模型的特征点检测方法实现了人脸表情的跟踪，对如何高效且真实的模拟人脸动画表情进行了研究。由于人脸肌群分布复杂，导致人脸有很强的局部性。对此，本文提出了基于肌肉模型的分区域插值方法，提高了表情模拟的真实感；同时结合肌群分布和贪心算法，设计了插值控制点的选取方式，减少了不必要的训练数据，从而提高了表情模拟的效率。实验表明，本文提出的表情模拟方案在保证真实感的同时满足应用要求的实时性。

本文系统还存在需要改进的地方。在模型驱动时，考虑到表情运动时人脸的深度数据变化很小，模型的网格点的深度值固定采用初始中性表情的初值，这样可能会导致局部失真，后续打算采用深度学习的方法直接获取特征点三维数据来解决这一问题。

参 考 文 献

[1] PARKE F I. Computer generated animation of faces [C]// AVM′72 Proceedings of the ACM Annual Conference. New York: ACM Press, 1972: 451-457.

[2] WILLIAMS L. Performance-driven facial animation [J]. ACM SIGGRAPH Computer Graphics, 1990, 24(4): 235-242.

[3] 何钦政, 王运巧. 基于 Kinect 的人脸表情捕捉及动画模拟系统研究[J]. 图学学报, 2016, 37(3): 290-295.

[4] CRISTINACCE D, COOTES T. Automatic feature localisation with constrained local models [J]. Pattern Recognition, 2008, 41(10): 3054-3067.

[5] BULAT A, TZIMIROPOULOS G. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230, 000 3D facial landmarks) [C]//2017 IEEE International Conference on Computer Vision (ICCV). New York: IEEE Press, 2017: 1021-1030.

[6] PIGHIN F, AUSLANDER F, LISCHINSKI D, et al. Realistic facial animation using image-based 3D morphing [EB/OL]. [2018-09-01]. https://www.researchgate.net/publication/2326717_Realistic_Facial_Animation_Using_Image-Based_3D_Morphing.

[7] 孙世全. 表情驱动的实时 Laplacian 网格变换面部表情模拟[D]. 秦皇岛: 燕山大学, 2015.

[8] 张满囤, 霍江雷, 单新媛, 等. 基于 Kinect 与网格几何变形的人脸表情动画[J]. 计算机工程与应用, 2017, 53(14): 172-177.

[9] SUWAJANAKORN S, SEITZ S M, KEMELMACHER-SHLIZERMAN I. Synthesizing Obama [J]. ACM Transactions on Graphics, 2017, 36(4): 1-13.

[10] GUO Y D, ZHANG J Y, CAI J F, et al. CNN-based real-time dense face reconstruction with inverse-rendered photo-realistic face images [EB/OL]. [2018-10-15]. https://arxiv.org/abs/1708.00980.

[11] SARAGIH J M, LUCEY S, COHN J F. Real-time avatar animation from a single image [C]//Face and Gesture 2011. New York: IEEE Press, 2011: 117-124.

[12] VIOLA P, JONES M J. Robust real-time face detection [J]. International Journal of Computer Vision, 2004, 57(2): 137-154.

[13] SARAGIH J M, LUCEY S, COHN J F. Deformable model fitting by regularized landmark mean-shift [J]. International Journal of Computer Vision, 2011, 91(2): 200-215.

[14] 缪报通, 陈发来. 径向基函数神经网络在散乱数据插值中的应用[J]. 中国科学技术大学学报, 2001, 31(2): 135-142.

[15] RENDALL T C S, ALLEN C B. Reduced surface point selection options for efficient mesh deformation using radial basis functions [J]. Journal of Computational Physics, 2010, 229(8): 2810-2820.

[16] 余重基, 李际军. 一种基于 RBF 网络生成人脸表情的算法[J]. 计算机应用, 2005, 25(7): 1611-1615.