考试方式：☑公开招考 □本科直博 □硕博连读

报名号：1038400806

博 士 生 科 研 计 划 书

学生姓名：李冰川

报考院系：137 软件工程系

报考专业：计算机科学与技术

研究方向：19 数字媒体技术

导　　师：(131)导师组

日　　期：2020 年 12 月 13 日

一、立题依据

□ 重点介绍本项目的科学意义、国内外现状，并附主要参考文献

人脸作为人类信息与情感表达的重要载体，是新一代人机交互方式的重要组成成分。人脸动画在工业界有着广泛的应用，但传统人脸动画制作方法费时、费力、高成本，因此，找到一个简单、方便、低成本的人脸动画制作方法具有重大意义。

人脸表情动画技术包括人脸三维重建和动画驱动两方面技术。目前获取人脸三维模型的方法主要包括三种，软件建模，仪器采集与基于图像的建模。基于人脸图像的三维重建方法非常多，常见的包括立体匹配，Structure From Motion(简称 SFM)[1]，Shape from Shading(简称 SFS)，三维可变形人脸模型(3DMM)。三维可变形人脸模型方法设备要求低，不需要生成拓扑算法简单，效果稳定，易于移动端实现实时重建，使其成为动画领域首选的方法。3DMM 方法由 Blanz[2]在 99 年首次提出，往后的改进通常是基于他们的工作。这种方法有一个关于人脸模型的“先验知识”，即可形变模型。因此不管人脸处于什么角度，都能得到较完整的人脸。现在比较常见的人脸模型有 Basel Face Model（BFM）[3]、Surrey Face Model（SFM）、FaceWarehouse、Large Scale Facial Model (LSFM)等。该方法面临的几个问题是：一是如何制作完备精准的形状基和表情基用于建模；二是如何使用这组基和图像更准确的建模；三是如何给模型贴上完整的贴图。课题的人脸重建主要基于该方法，着重解决后两个问题。

表情驱动的实现通常有实时重建和网格变形两种实现方式。前一种方法由于想要重建精准的模型需要耗费较长的时间比较难达到实时性的要求，并且这种方法最大的缺陷是没有考虑前后帧的关联性，结果不平滑，通常需要先对图像进行插值处理或者对计算得到的模型参数进行插值。目前网格变形技术主要包括空间变形、基于骨骼的蒙皮变形和表面变形三大类。空间变形技术的主要思想是把原模型嵌入到另一个局部空间中，然后通过编辑这个局部空间来改变原模型形状具有简单、高效等优点，缺少足够的细节，适合用于交互式的软件制作动画。基于骨骼的蒙皮变形技术作为空间变形技术的一种特殊类型，发展到现在技术趋于成熟在多种商业三维造型软件中已被广泛应用。其中代表性技术“线性混合蒙皮”以其简单高效的优点称为骨骼蒙皮技术中的首选。LBS 的原理是用变形矩阵来表示骨豁运动，皮肤上的点的变换是关联骨骼变换的线性组合。有两个主要缺点：一是容易出现“打结”、“塌陷”等现象；二是需要有经验的动画师手工指定骨骼关联权，通常费时费力。Pose Space Deformation 是由 Lewis[4]等人于 2000 年提出的一种几何模型变形方法。该方法在位姿空间进行差值变形，在任意的变形空间点上随意设置所需的目标变形(Target Deformations)，算法自动差值生成中间的变形结果。对于变形驱动来说，在关键姿态点定义变形目标，具有更直接的语义控制信息。基于曲面的网格变形技术较多，近年来流行的是微分域 （differential domain）网格变形，其主要思想是：使用曲面顶点的微分坐标如 Laplace 坐标，梯度坐标代替全局的笛卡尔坐标，变形过程中保持微分属性不变，可以得到保持几何细节的变形结果。比较经典的算法有 Laplace Deformation[5]、As-Rigid-As-Possible[6]等。这类算法有一个比较大的缺点是耗费时间随模型及拓扑的规模增长，并且需要足够的约束才能收敛到合理的值。课题通过网格变形驱动表情的方式希望结合这几种方法的思路提出一种有效的驱动规模较大的人脸模型变形的方法。

[1] Daugman J , Adler A , Schuckers S , et al. Structure-from-Motion[M]// Encyclopedia of Biometrics. 2009.

[2] Blanz V , Vetter T , Rockwood A . A Morphable Model for the Synthesis of 3D Faces[J]. acm siggraph, 2002:187-194.]

[3] Gerig T, Morel-Forster A, Blumer C, et al. Morphable face models-an open framework[C]//2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018: 75-82.

[4] Lewis J P, Cordner M, Fong N. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation[C]//Proceedings of the 27th annual conference on Computer graphics and interactive techniques. 2000: 165-172.

[5] Sorkine O. Laplacian mesh processing[J]. Eurographics (STARs), 2005, 29.

[6] Sorkine O, Alexa M. As-rigid-as-possible surface modeling[C]//Symposium on Geometry processing. 2007, 4: 109-116.

二、研究内容

□ 研究目标、研究内容、创新之处和拟解决的关键问题

1. 研究目标

本课题的研究目标包括单张、多张图像的静态重建和实时重建视频帧的人脸及整个头部模型，身体部分保持不变，脖子和身体连接自然不发生严重扭曲。

2. 研究内容

本课题的人脸重建基于 3DMM 的方法，使用 BFM2019 模型数据。由于目前大部分基于 3DMM 的重建方法采用正交投影的方式，重建的结果形状和表情都不准确，所以我们采用透视投影的办法。透视投影的重建方法需要相机焦距的信息，并且使用非线性优化，时间复杂度较高，要达到实时比较困难。本课题通过分析透视投影的重建方法需要的数据和视频前后帧之间的差别，设计一种方案来达到实时透视投影重建的办法。主要研究内容包括：

1） 焦距的获取

目前大部分手机拍摄的照片都包含 exif 数据，这里面就包含了相机的焦距，数据可能异常需要自行判断。基本办法是手机的摄像头芯片宽度和 35mm 等效焦距都有一个范围，通过这两个参数和照片的分辨率可以计算相机焦距。如果照片的 exif 数据包含有焦距信息并且在正常范围内那么就采用这个焦距，否则使用一个默认值，即使不准确，重建的模型效果也比正交投影方法得到的结果好（已经试验过）。

2） 实时性

对于视频帧来说，重建的是同一个人的模型，因此理论上每一帧对应的形状参数应该一样的。假设第一帧已经得到准确的模型数据后，后面每一帧只需要根据 3D 关键点通过现行求解的方式重建。另外一种方式是根据前后帧的 3D 关键点及其他约束，对前一帧的结果进行网格变形得到后一帧的结果。

3） 动画的平滑

有些视频人物的头部动作较为剧烈，普通相机拍摄的效果不好，因此重建的时候需要插值处理，防止抖动的现象。

4） 贴图的完整性

由于贴图通常只有正脸区域的像素，因此需要对其他区域的像素通过深度学习网络进行预测。这个是用于静态重建的模型，不需要实时。

3. 拟解决的问题

a. 焦距的计算
b. 3D 关键点的计算
c. 根据 3D 关键点信息实时重建或者网格变形驱动表情
d. 模型插值处理
e. 脖子区域的变形处理

三、拟采取的研究方法、技术路线

1. 研究方法

本课题的研究方法以理论推导和程序验证为主。通过多部手机拍摄大量人脸照片进行透视投影的静态重建验证建模的准确性，实时的真人视频验证视频重建的效果。

2. 技术路线

模型数据采用 BFM2019，整个头部被切割成两部分包括脸部区域和其他部分，这两部分的边界点数和位置一样。其他部分包括后脑勺和脖子做了简化面数和点数较少，脖子和身体连接边界点数和位置一样，面部区域 4 万多个点，后脑勺和脖子总共 2 千多个点左右。对于单张图片的重建，首先推算出相机焦距，根据小孔成像的原理对每个 3D 关键点计算投影，将每一个 3D 关键点的投影和检测得到的 2D 关键点的差距作为 loss 进行非线性优化，使用 Ceres 优化库计算出投影参数、形状参数和表情参数。将整个头部旋转再进行坐标系变换（模型是倒立的），脖子通过选定 ROI 区域（后脑勺靠近脖子的整个区域）ARAP 变形，使之与身体保持自然的连接。想要得到整个头部的完整贴图，可以先通过深度学习网络的办法先得到三张照片（原始的照片，左右侧脸的照片）分别进行单张图片的重建，得到三张贴图融合成一张贴图。

对于视频的实时重建，目前的方案是第一帧先采用单张图片的重建方法。后面每一帧根据 3D 关键点的坐标和形状参数可计算得到所有信息。考虑到连续的两帧，3D 点变化较小，根据上一帧得到的 3D 关键点 p 、2D 关键点 I_p 和当前帧的 2D 关键点 I_q 计算当前帧的 3D 关键点 q 可采用以下公式计算：

$$p=\begin{pmatrix} x_{p1} & \cdots & x_{pk} \\ y_{p1} & \cdots & y_{pk} \\ z_{p1} & \cdots & z_{pk} \end{pmatrix}, q=\begin{pmatrix} x_{q1} & \cdots & x_{qk} \\ y_{q1} & \cdots & y_{qk} \\ z_{q1} & \cdots & z_{qk} \end{pmatrix}$$

$$p=\begin{pmatrix} p_1^t \\ p_2^t \\ p_3^t \end{pmatrix}, q=\begin{pmatrix} q_1^t \\ q_2^t \\ q_3^t \end{pmatrix}$$

$$I_p=\begin{pmatrix} u_{p1} & \cdots & u_{pk} \\ v_{p1} & \cdots & v_{p1} \end{pmatrix},\ I_q=\begin{pmatrix} u_{q1} & \cdots & u_{qk} \\ v_{q1} & \cdots & v_{q1} \end{pmatrix}$$

$$\begin{cases} f\left(q_1^t/q_3^t-p_1^t/p_3^t\right)=\left(I_q\right)_1^t-\left(I_p\right)_1^t \\ f\left(q_2^t/q_3^t-p_2^t/p_3^t\right)=\left(I_q\right)_2^t-\left(I_p\right)_2^t \end{cases}$$

$$\Rightarrow$$

$$\begin{cases} f\left(q_1^t-p_1^t\right)/p_3^t\approx\left(I_q\right)_1^t-\left(I_p\right)_1^t \\ f\left(q_2^t-p_2^t\right)/p_3^t\approx\left(I_q\right)_2^t-\left(I_p\right)_2^t \end{cases}$$

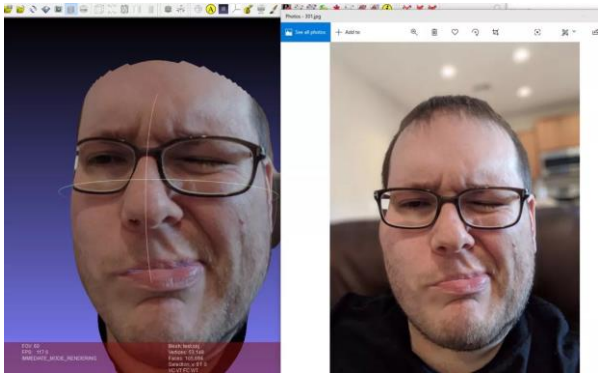
得到 3D 关键点之后，通过前后帧（可近似认为是刚性变化）的 3D 关键点计算相对旋转和平移，累加前一帧的投影参数得到当前帧的投影参数。另一种方式是以计算得到的当前帧的 3D 关键点为目标及其他约束（例如人脸轮廓边界）对前一帧进行网格变形得到当前帧的结果，这种方式要求对脸部区域进行简化，减少点数和面数，以达到实时性的要求。

对于脖子区域的变形处理采用拉普拉斯变形，ROI 区域选择后脑勺靠近脖子部分和整个脖子区域进行计算，时间复杂度为 $O\left(n^2\right)$，n 是 ROI 区域的点数较少（ROI 区域几百个点）可以达到实时性的要求。在前后两帧之间插入一帧模型和图片。插入的这一帧形状参数和所有帧一样，表情参数根据前后帧的表情参数线性加权，旋转通过前后帧的四元数进行球面 Slerp 插值，平移通过前后帧的平移进行线性加权。

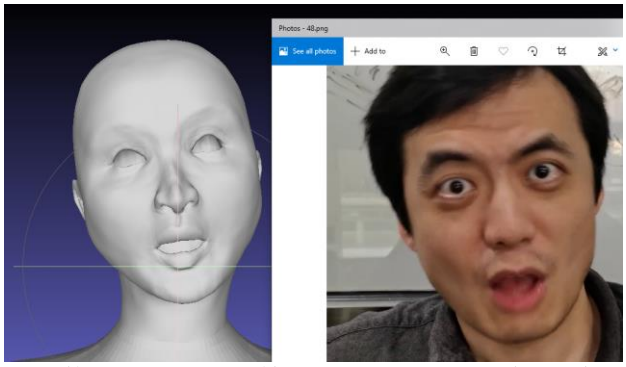
四、考核指标

□ 预期成果

1. 单张、多张图片静态建模形状和表情准确，多张图片静态建模贴图完整。



2. 视频能够实时建模，表情准确，没有抖动现象，脖子和身体保持自然连接。



3. 目前使用开源的 dlib 检测的 2D 关键点只有 68 个，表情效果不够逼真，需要更多的关键点信息。