



南京大學

研究生毕业论文 (申请硕士学位)

论文题目 单相机动态人体三维重建

作者姓名 朱海宇

学科、专业方向 计算机视觉

指导教师 于耀 副教授

研究方向 三维重建

2010年5月

学 号 : MF1423044

论文答辩日期 : 2010 年 6 月 1 日

指导教师 : (签字)

Dynamic Human Body Modeling Using a Single RGB Camera

by

Haiyu Zhu

Directed by

Professor Yao Yu

School of Electronic Science and Engineering

May 2017

*Submitted in partial fulfilment of the requirements
for the degree of master in Computer Vision*

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目： 单相机动态人体三维重建
计算机视觉 专业 2017 级硕士生姓名： 朱海宇
指导教师（姓名、职称）： 于耀 副教授

摘要

尽管三维人体建模在虚拟现实、3D游戏等多个领域有广泛的应用，但由于人体运动的复杂性，如何低成本便捷地获取高质量人体模型一直是学者们不断探索的问题。目前获取高质量人体模型的方法主要还是通过激光雷达或者深度相机，但是这些设备价格高昂且体积较大便携性很差。于是，有学者提出用多相机系统重建人体模型，可是多相机系统环境搭建比较繁琐。不仅需要对相机的位置进行标定还需要把所有相机进行帧同步。此外，投入更多的相机也使得重建系统的成本居高不下，这也限制了多相机人体重建系统的推广与应用。随着手机等移动设备成像质量的不断提高，如何用单相机获取高质量人体模型逐渐成为学者们热衷探讨的问题。

由于人体复杂的运动、身体部位的遮挡使得单相机人体三维重建充满挑战。目前基于单相机重建的方法大多需要人为地添加一些约束，因而无法做到自动化重建出人体模型。

在运动中，人体模型局部区域内点的运动具有较高的相似度，并且随着人体的不断运动，我们有可能通过相机捕捉到之前被遮挡的信息。我们设计了一个运动学分类器，对运动中的人体根据运动信息将其划分成许多“块”，在较短的时间内“块”表面的形变可以忽略不计，因而将非刚性重建问题转化成刚性重建问题。我们把SCAPE模型和不同时刻得到的人体部位模型进行融合，并对SCAPE模型进行“刚性成分”和“非刚性成分”的分解，从而得到一个可驱动的人体三维模型。模型的“刚性成分”描述了人体的体型参数，而“非刚性成分”则描述了不同时刻下人体表面的细节信息。

实验表明：我们的单相机动态人体自动重建系统不仅具有较高的鲁棒性，而且是第一个不需要人为添加约束就可以自动重建出完整人体三维模型的系统。

关键词： SCAPE；非刚性重建；单相机；运动分类；从运动到结构

南京大学研究生毕业论文英文摘要首页用纸

THESIS: Dynamic Human Body Modeling Using a Single RGB Camera

SPECIALIZATION: Computer Vision

POSTGRADUATE: Haiyu Zhu

MENTOR: Professor Yao Yu

Abstract

Although 3D human body modeling has a wide application in virtual reality, 3D games and many other fields, it is a constant for researchers to explore how to conveniently obtain high quality huaman bodels with low cost, due to the complexity of human motion.

The state of art approach to obtaining high-quality human models is through lasers or depth cameras, which are expensive and clumsy. Thus, some researchers obtained 3D human model through multi-view reconstruction. However, it is not only to calibrate the position of the camera but also to synchronize all the cameras. In addition, more cameras contribute to high costs, which limits to promoting the multi-view reconstruction system. With the improvements of mobile phones and other mobile devices' imaging quality, how to acquire high-quality human model by single RGB camera has been a hot topic between researchers.

Due to the complexity of human motion and occlusion, it is very challenging for us to reconstructe human models using a single RGB camera. At present, most of the methods based on single-camera reconstruction need to add some constraints artificially, and thus can not reconstructe human model automatically.

During the human motion, points of human model in a small region have similar motion. And with continuous movement of the human body, it is more possible to capture missing informations through the camera. We design a kinematic classifier to vivide the moving body into many "blocks" based on the motion information, and the deformation of the "block" surface in a short period of time was negligible, thus converting the non-rigid reconstruction problem into a rigid reconstruction problem. We fuse the human body part model obtained at different times to the SCAPE

model. Finally, a drivable human model will be obtained after decomposing the "rigid component" and "non-rigid component" of SCAPE instances. The "rigid component" of the model describes the body's body parameters, while the "non-rigid component" describes the details of the human body at different times.

Experiments show that our single-camera dynamic human body reconstruction system not only has high robustness, but also the first system that can automatically reconstruct the complete human three-dimensional model without artificial constraints.

Keywords: SCAPE, non-rigid reconstruction, single RGB camera, motion classification, structure from motion

目录

目录	iv
第一章 绪论	1
1.1 三维人体建模概述	1
1.1.1 深度相机人体三维建模	1
1.1.2 多视角人体三维建模	6
1.1.3 单视角人体三维建模	9
1.2 本文主要工作	12
1.3 本文的章节组织形式	14
第二章 人体三维模型重建	16
2.1 人体模型建模	17
2.1.1 模型姿态驱动	17
2.1.2 姿态建模	18
2.1.3 模型体型驱动	20
2.1.4 体型建模	20
2.2 人体三维姿态检测	20
2.3 运动学分类	21
2.3.1 人体轮廓非刚性匹配	22
2.3.2 人体部位运动学分类	25
2.4 人体部位稠密三维重建	27
2.4.1 光束平差法	27
2.4.2 稠密三维重建	28
2.5 分解人体表面形变	30
2.5.1 模型实例获取	31
2.5.2 模型刚性、非刚性成分分解	32

第三章 结果和讨论	34
3.1 系统参数	34
3.2 定性分析	35
3.3 定量分析	38
3.4 算法复杂度分析	41
第四章 总结与展望	42
4.1 总结	42
4.2 展望	43
参考文献	44
简历与科研成果	48
致谢	49

表格

3.1 试验中重建系统相关参数的具体数值。	34
3.2 重建模型的身体参数对比。	40

插图

1.1	Kinect相机。从左往右依次是 Kinect 一代和二代。	1
1.2	KinectFusion系统流程。	2
1.3	Tong 等人 [27]的人体三维重建环境。	3
1.4	Zhang 等人 [36]人体模型重建算法。	4
1.5	Zeng 等人 [35]人体模型重建算法。	5
1.6	Newcombe 等人 [22]人体模型重建算法。	5
1.7	多相机数据采集环境。图中用蓝色方框标注的是相机。	6
1.8	Gall等人 [8] 方法示意图。	7
1.9	Liu等人 [18] 方法示意图。	8
1.10	Robertini 等人 [24] 方法示意图。	8
1.11	Guan等人 [9] 方法示意图。	10
1.12	Hasler等人 [10] 方法示意图。	11
1.13	Jain等人 [14] 方法能够对图像中的人体体型进行改变。	11
1.14	Zhou等人 [37] 模型融合方法。	12
1.15	系统概要	13
2.1	系统流程示意图	16
2.2	重建过程中使用的参数化模型。 (a): 训练得到的 SCAPE 模型; (b): 模型的不同肢体部位。	17
2.3	三角面片变形过程。	18
2.4	运动学分类步骤。 (a): 将 SCAPE 模型投影到图像上获得包含分类信息的轮廓; (b): 图像中人体的轮廓; (c): SCAPE 模型轮廓和人体轮廓配准前; (d): 利用 CPD(Coherent Point Drift) [21] 方法进行人体轮廓的匹配结果; (e): 借助 SCAPE 模型信息对图像上人体轮廓分类的结果; (f): 运动学分类结果。首先, 我们提取模型投影到图像上的轮廓以及图像上人体的轮廓; 然后, 我们通过将 (b) 中人体轮廓和 (a) 中模型的轮廓进行配准并将人体轮廓进行分类得到 (e) 中结果。最后, 我们利用全连通的条件随机场模型来获得图像上人体不同部位的分类结果 (f)。	22

2.5 利用 CPD 将图像上 SCAPE 模型轮廓和人体轮廓进行配准的结果。(a) 人体当前的姿态; (b) 驱动 SCAPE 模型到当前姿态下; (c) 人体轮廓和 SCAPE 模型轮廓配准之前; (d) 人体轮廓和 SCAPE 模型轮廓配准结果。	23
2.6 稠密光流。(a): 人上一时刻的姿态; (b): 人当前的姿态; (c): (b) 图相对于 (a) 图的光流。	26
2.7 身体部位运动学分类。左边的是通过 RGB 相机获取的原始图像; 中间的图是没有运动项参与时的分类结果; 右边的结果表明在运动项的参与下分类的结果很精确并且能够处理有遮挡的情况。	27
2.8 光束平差法示意图。	28
2.9 七个身体部位的稠密重建结果	30
2.10 当前帧下 SCAPE 模型跟上肢人体部位融合得到的实例。左边的图是非刚性融合之前 SCAPE 模型被驱动到当前帧中人体姿态下的结果, 以及 Z-buffer 算法获取到的 SCAPE 模型的前景部分。右图是通过将当前姿态下 SCAPE 模型对应部位的前景跟稠密重建模型进行非刚性融合得到的一个实例。	32
2.11 SCAPE 模型和稠密重建结果融合误差图。	33
3.1 实验环境。我们只需要一个普通的相机在合适的距离下拍摄人体的运动。	35
3.2 动态人体三维重建结果。	36
3.3 两个图像序列的对比结果。第一行是两组图片序列中的部分帧; 中间一行是我们将最终重建出来的模型驱动到第一行图片对应姿态下的结果; 最后一行是使用 Yu 的方法得到对应图片下的对比结果。	38
3.4 不同模型的重建误差。	39
3.5 重建模型和 KinectFusion 之间的平均重建误差。横坐标表示参与重建的图像数量。纵坐标表示模型的平均重建误差。随着参与重建的图像数目增加, 模型的平均重建误差逐渐降低并趋于稳定。 ..	40

第一章 绪论

1.1 三维人体建模概述

人体三维建模在游戏、计算机动画和虚拟试衣等领域都有着广泛的应用，并且在虚拟感知、计算机视觉和计算机图形学中它始终是学者们不断探讨的热点问题。在过去的几十年中，学者们探索出了许多种方法来获取人体的三维模型，例如利用三维的扫描仪获取人体三维模型的 [2, 30]，借助深度传感器获取深度图来重建人体三维模型 [7, 13, 17, 22, 23, 27, 29, 35, 36]，利用多视角重建完成人体三维重建的 [6, 8, 12, 18, 24, 38]。随着人们对模型获取成本及便捷性要求越来越高，近年来也出现了利用单个相机重建出人体三维模型的方法 [9, 10, 12, 14, 18, 34, 37]。这些方法的提出旨在更准确、更加快速的获取人体的三维模型。

1.1.1 深度相机人体三维建模

最近几年像 Kinect(如图 1.1)这类深度相机的出现，使得我们可以通过深度图或者RGB-D图像重建出物体的三维模型。Kinect 上有红外发射器、红外接收



图 1.1: Kinect 相机。从左往右依次是 Kinect 一代和二代。

器、RGB 相机三个主要传感器，红外发射器配合红外接收器可以获得场景的深度信息，而 RGB 相机主要是获取场景的颜色信息。Kinect 可以提供最大帧率 30fps 的深度图和彩色图，它的对环境的适应能力强、精度较高、价格低廉，性价比很高。Kinect 2代在一代的基础上将彩色图片的分辨率提高到了 1920×1080 ，并提高了深度信息的精度。

Newcombe 等人 [23]，Izadi 等人[13] 实现了一个名为 KinectFusion 的系统。KinectFusion 主要由四个部分组成 (1.2)：(a)、预处理阶段。将从 Kinect 获取到的深度图转换成点云，并计算出每个点的法线；(b)、三维场景表面更新。通过

跟踪新的深度图获取到的 Kinect 姿态信息将当前帧的点云融合到用截断有向距离函数(TSDF, Truncated Signed Distance Function)表示的场景中去; (c)、表面预测。通过光线投射到整个场景的 TSDF 上, 预测出场景的表面结构, 进而将当前帧的点云融合到整个场景中去, 从而实现 Kinect 姿态跟踪和场景表面更新的闭环; (d)、Kinect 姿态估计。通过多尺度的迭代最近邻算法将预测出来的场景表面跟 Kinect 得到的场景结构进行配准从而计算出 Kinect 的姿态信息。通过

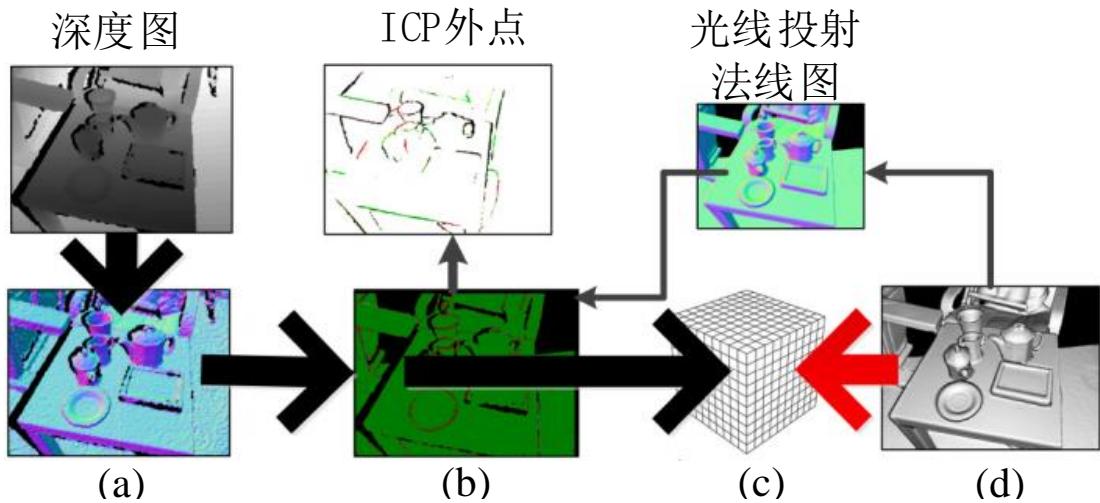


图 1.2: KinectFusion 系统流程。

GPU 加速 KinectFusion 可以实时重建出场景比较精细的三维模型, 但是它只适用于重建静态场景的三维模型。

使用 KinectFusion 重建人体模型是比较困难的, 因为 KinectFusion 只能重建静态场景的三维模型, 而在人体扫描的过程中人很难保持绝对的静止, 这就给人体模型重建带来了很大的挑战。基于 Kinect 的人体三维重建或多或少都借鉴了 KinectFusion 的思路, 他们大多利用某时刻人体局部的三维模型来恢复人体完整的三维模型。但是从局部恢复整体的问题本身就是一个欠约束的问题, 所以为了解决这样一个欠约束的问题, 目前的解决方法主要有基于模板的方法, 如 [17, 27, 29, 36]。还有一些方法是不需要人体模板就可以重建出人体三维模型的。例如, 将人体模型看作拟刚性物体(quasi-rigid object) [35], 将非刚性问题向刚性问题转换; 用稠密体变换场(Dense Volumetric Warp-field)表示模型形变 [22]; 提出新的参数化多分辨率模型 [3]; 利用非刚性光束平差法对人体模型进行全局优化 [7]。基于模型的方法能够适应人体更多复杂的姿势, 但是在和模型融合的过程中会丢失深度信息中的高频部分, 使得最终得到的模型不够精细。而其他的

重建方法可以尽可能多的保留人体表面的细节，但带来的问题就是在某些极端姿势的情况下系统可能无法获得精细的人体模型。

Weiss 等人 [29] 利用 Kinect 获取到的深度信息结合人体的轮廓、姿态等信息计算出人体的体型参数，接着用人体体型参数驱动参数化模型 SCAPE(Shape Completion and Animation for PEople) 得到目标人体参数化的三维模型。

Tong 等人 [27] 将多个视角不重叠的 Kinect 按图 1.3 中位置放置。在第一帧时用 Kinect 获取的深度信息先重建出一个比较粗糙的模板，然后将模板和当前帧下人体模型进行融合。当人体旋转一周之后就进行全局的融合来降低人体姿态变化带来的误差。最后将每一帧下形变之后的模板驱动到第一帧下并利用泊松重建得到一个完整的人体模型。

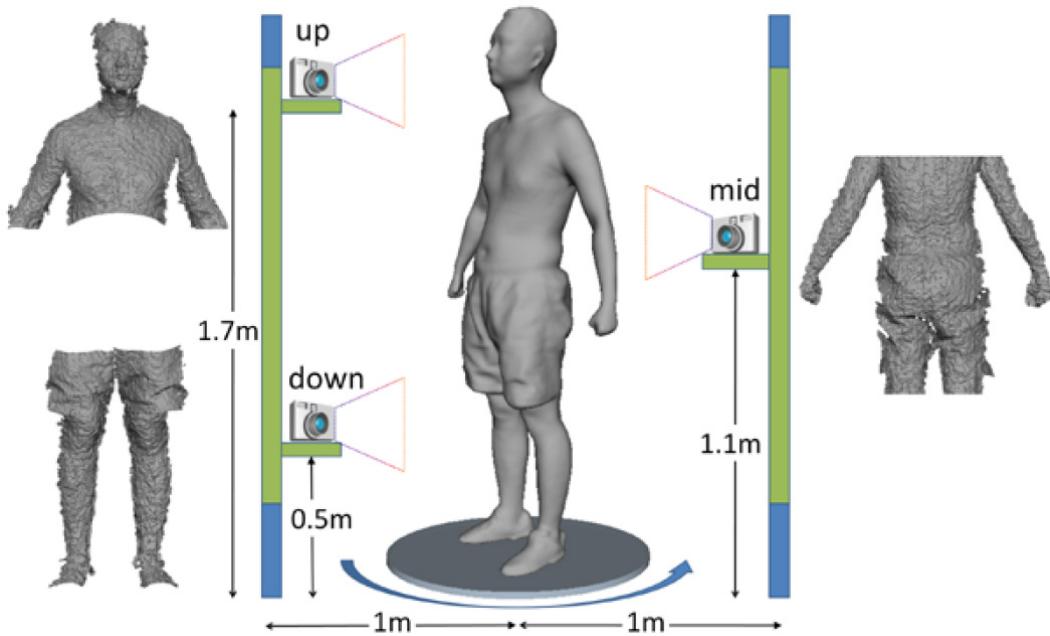


图 1.3: Tong 等人 [27] 的人体三维重建环境。

Li 等人 [17] 用一个 Kinect 从不同的角度扫描一个人近似姿态的三维模型，然后将不同角度得到的模型融合到一起得到一个完整的人体模型。每次扫描时都要讲从 Kinect 获取到的多帧深度信息进行融合来降低模型的噪声，接着将人和背景分割开，得到当前角度下人体的三维模型。由于人在 Kinect 前面旋转时姿势不可能保持一样，所以不同视角下获得的模型肯定包含了许多形变和不连续。所以他们就将某一视角下的人体模型作为模板，并将其他视角下的人体模型用刚性、非刚性融合算法将模型融合到模板中，最后利用配准过程中得到的对应关系做全局的优化得到完整的人体模型。

Zhang 等人 [36] 利用 KinectFusion 获取同一个人在不同姿势下的三维模型，如图 1.4 中(a)。然后将这些模型结合人体姿态信息从而训练出这个人的参数化人体三维模型，如图 1.4 中(b)。最后只要知道某些姿态下的深度图和人体姿态就可以驱动参数模型得到当前姿态下精细的人体模型。

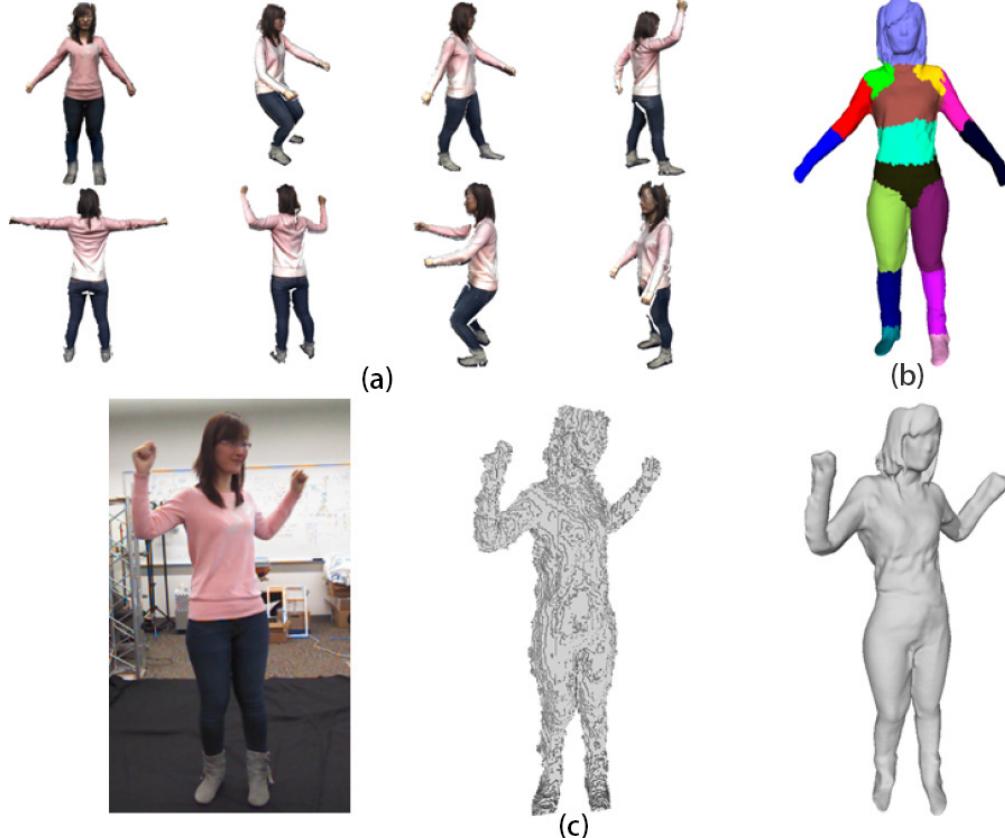


图 1.4: Zhang 等人 [36] 人体模型重建算法。

Zeng 等人 [35] 把人体模型看成很多小刚性块的组成，并且模型的非刚性形变用“图”来表示这样就降低了问题的复杂度。算法具体过程如图 1.5 所示：(1)、Kinect 获取的每帧深度信息都用模型到部分(Model-to-part Method)的方法将“图”融合到深度信息上。(2)、融合之后，对深度信息进行采样并将其融合进“图”中并更新“图”的拓扑信息。(3)、融合不同帧所有的深度信息之后就做全局的非刚性变换，将所有的空间点依据“图”变换到最后一帧中。(4)、最后通过泊松重建得到完整的人体模型。

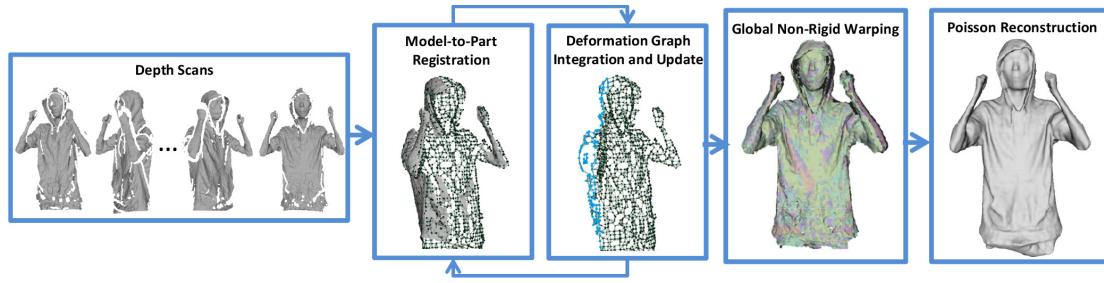


图 1.5: Zeng 等人 [35] 人体模型重建算法。

Newcombe 等人 [22] 拓展了 KinectFusion 的应用场景，他们利用稠密体变换场(Dense Volumetric Warp-field)和正则空间(Canonical Space)使得利用 Kinect 可以重建出动态人的三维模型。他们的算法的具体过程如图 1.6 所示。图中 (a)、(b) 是他们从 Kinect 得到的深度图，(d)、(e) 是重建出来的稠密模型。他们通过变换正则空间的模型到当前帧的深度图而得到稠密体变换场的参数，接着利用得到的稠密体变换场将当前深度图融合到正则空间，图 (d, f)。与此同时，变换场用一系列系数 6D 变换点来表示，并通过 K 近邻差值得到平滑的正则空间帧，图 (c)。最终，利用刚得到的正则空间帧就可以补全当前帧的模型，得到实时的动态的三维模型，图 (e)。

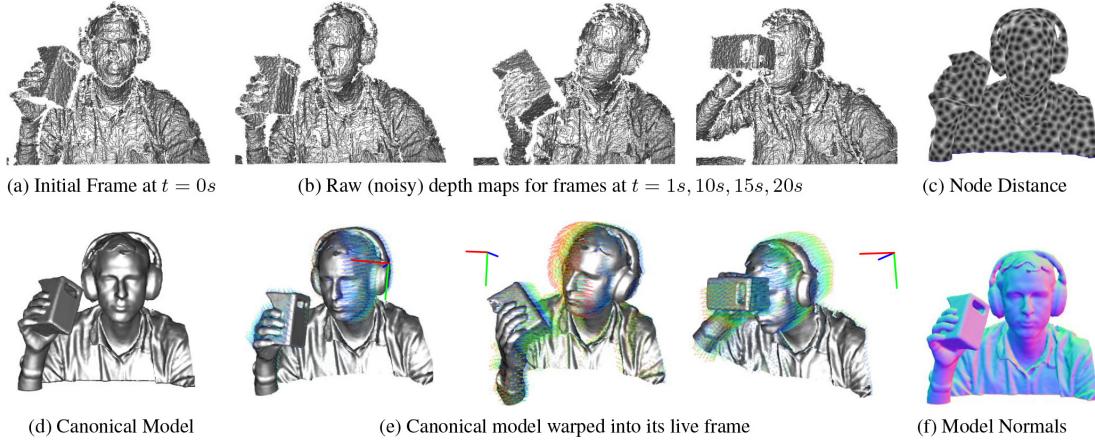


图 1.6: Newcombe 等人 [22] 人体模型重建算法。

Bogo 等人 [3] 扩展了 SCAPE 模型提出了一个名为“Delta”的多分辨率人体模型。从单目的 RGB-D 图像序列中恢复出动态人体的体型参数之后，结合高分辨率位移图获取人体表面的细节，从而得到精细的人体三维模型。Dou 等人

[7] 通过非刚性光束平差法(Non-rigid Bundle Adjustment)优化最终模型的体型和每一帧模型的非刚性参数，从而使得系统能够获取动态的并且包含大量非刚性运动的动态物体的三维模型。

1.1.2 多视角人体三维建模

多视角三维重建主要是利用多个同步且标定过的相机获取不同视角下同一时刻目标的信息来得到目标的三维模型。多视角重建环境（如图 1.7）搭建是几种重建方法中最繁琐的，不仅需要考虑相机摆放位置是否合理，还需要将相机进行帧同步，准确标定各个相机的空间位置信息。对比深度相机，多相机系统

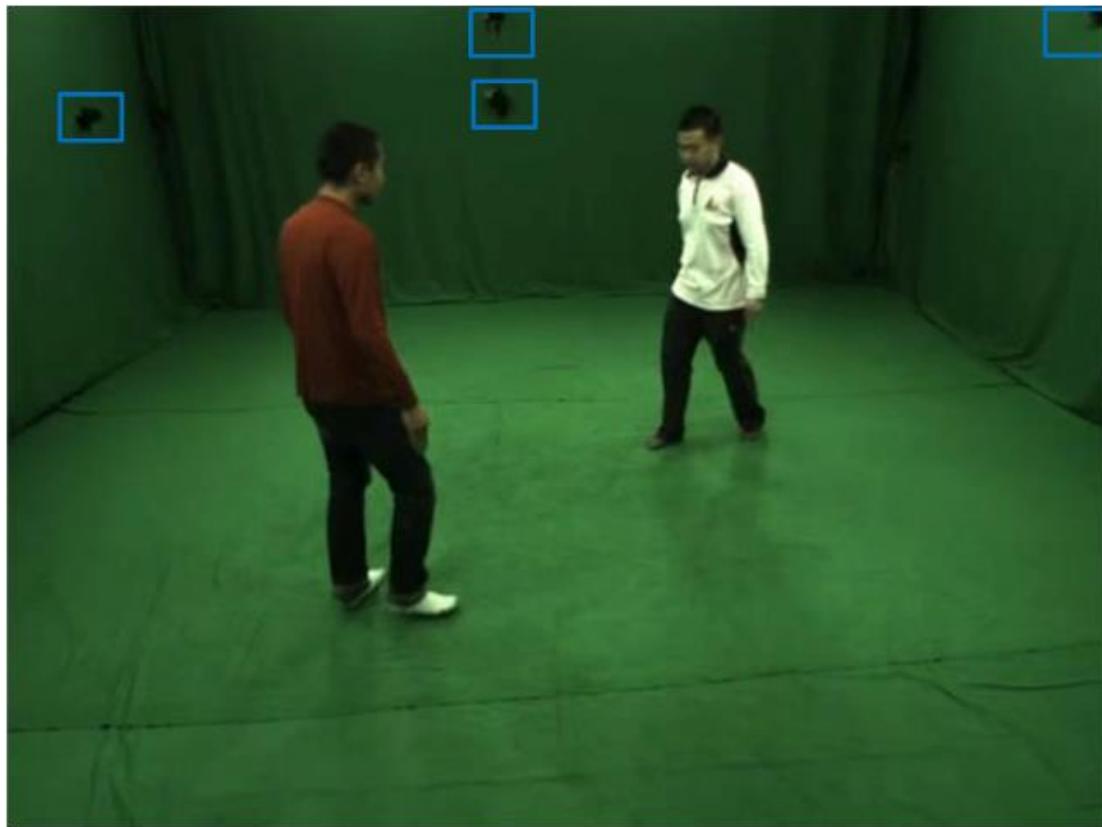


图 1.7: 多相机数据采集环境。图中用蓝色方框标注的是相机。

不光可以拿到场景的深度信息还可以获得不同角度下人体的轮廓、体型等信息。此外，多视角一定程度上也可以减少人体运动过程中的遮挡问题。

尽管多相机系统可以获取场景的深度信息，但是深度信息的质量还是无法跟深度相机相提并论，所以使用多相机进行重建的系统基本都是基于模板融合的方法。基于模板融合的方法一般会导致重建出来的模型缺失许多的细节，所以多相机重建系统的工作重点大多是尽可能恢复模型表面更多的细节信息。一

般情况下他们会选择用深度相机或者雷达扫描出人体精细的三维模型作为模板，然后将模型和人体骨架进行绑定融合得到可以用骨架驱动的三维模型，最后利用人体不同视角下的轮廓、体型信息对每帧的人体模型进行优化得到较为精细的模型 [8, 12, 18, 24]。而 Theobalt 等人 [6, 26] 抛弃了传统用骨骼驱动或将运动参数化的方法，而是用捕捉到的模型变化来驱动模型。将雷达扫描出人体精细模型转化成一个粗糙模型，对于获取的每一帧图像他们先用粗糙模型获得人的姿态信息，然后用精细模型结合人体轮廓、多视角信息得到人体模型表面的细节信息。Huang 等人 [12] 借鉴了“关键帧”的思想并将其应用到多相机重建中提高了系统的鲁棒性。Zollhofer等人 [38] 利用双目相机模拟深度相机工作获取场景的深度信息，结合体融合方法和非刚性三维重建算法重实时的重建出人体的三维模型。

Gall等人 [8] 的方法如图 1.8，他们同样用雷达先扫描出人体精细的三维模型，并将模型和人体骨骼做融合使得模型能够用人体骨架进行驱动，图 (a)。利用非监督的方法结合人体轮廓跟踪人体骨骼，图 (b,c)。将前一帧的人体模型进行变形，在拟合图像上的人体时优化化图 (c) 中的人体骨架。由于骨骼的驱动无法解决人体衣服表面的形变所以还需要将图 (d) 中的模型跟不同视角下的人体轮廓进行融合，图 (e)。最后得到实际的人体模型图 (f)

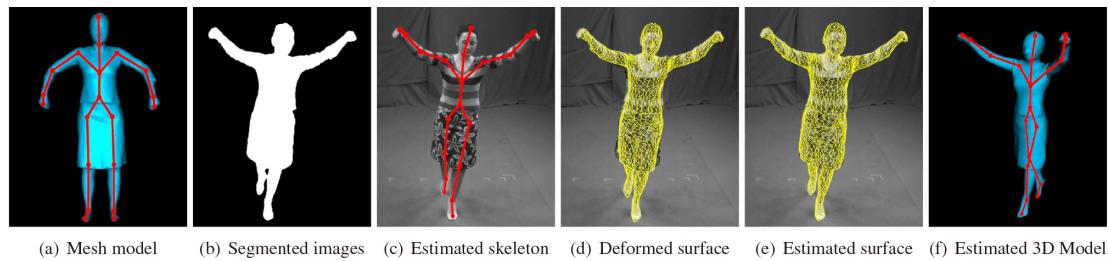


图 1.8: Gall等人 [8] 方法示意图。

Liu 等人 [18] 的方法如图 1.9 所示。他们主要的工作在于将人体运动捕捉转化为 2D 图像上的分割、3D 姿态体型估计的问题，使得系统可以获得多人的三维模型。他们也需要将人体三维模型和人体骨骼融合之后的参数模型作为初始状态，图 (a)。接着对图像 (b) 上多人的轮廓进行分割，得到不同人的身体轮廓，图 (d)。最后将当前帧的人体轮廓结合和上一帧的人体姿态、体型结合得到当前帧的人体姿态和体型，图 (e, f)。

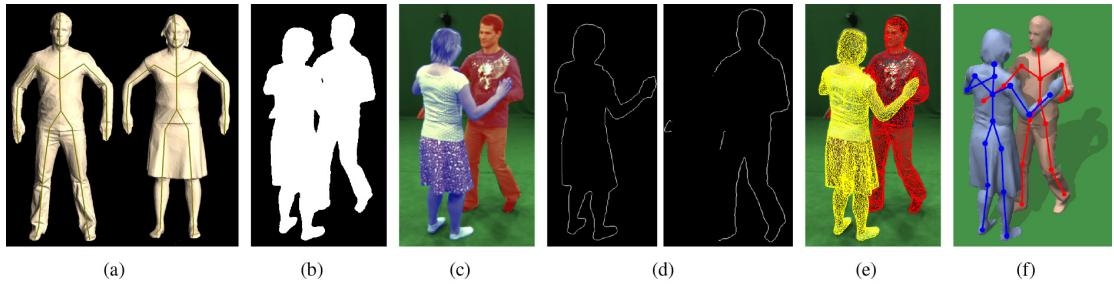


图 1.9: Liu等人 [18] 方法示意图。

Huang 等人 [12] 提出了关键帧的概念以及一种外点排除方法。非关键帧的方法主要是用同一个模型去跟当前的人体轮廓、姿态等信息进行融合以得到当前帧比较准确的人体模型和姿态，这是一种“一个匹配所有”(one-fit-all) 的策略，它在人体有遮挡或者系统噪声大时会失效。而 Huang 等人将关键帧处获取到的人体模型作为“参考模型”，重建模型时将所有的“参考模型”跟当前的人体轮廓、姿态等信息进行融合以找出一个最优解。随着“参考模型”数量的增多，系统在有遮挡的情况下就可以通过某些关键帧处的“参考模型”找出当前情况的最优匹配，从而估计出人体遮挡部分的形状，降低了系统的噪声提高了系统的鲁棒性。

Robertini 等人 [24] 在Gall等人 [8]方法的基础上提出了用三维高斯函数来表示人体体型的新方法，并将人体三维重建问题转化成连续图像序列中人体表面顶点全局优化的问题，从而能够获取模型表面更多的细节使得模型更为精细。如图 1.10 所示，他们用三维高斯函数来表示有着色的人体模型。同时将得到的多视角图像划分成很多小块并用二维高斯函数表示。最后再保证最大颜色连续性的前提下将二维高斯函数和三维高斯函数一起做优化得到模型更多的细节。

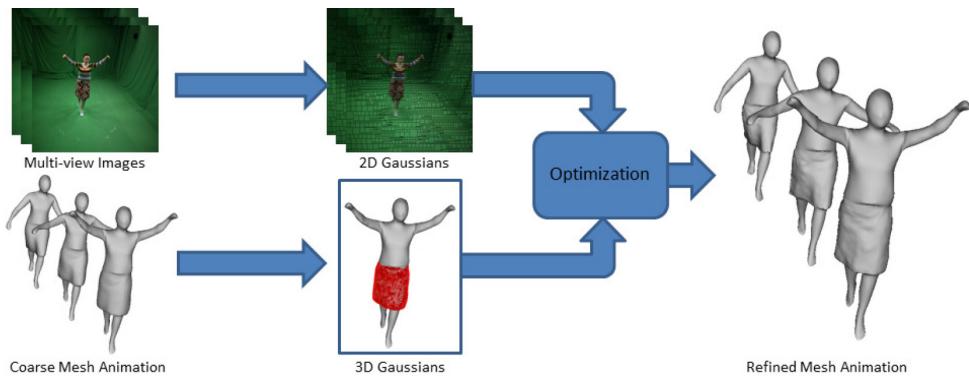


图 1.10: Robertini 等人 [24] 方法示意图。

1.1.3 单视角人体三维建模

虽然深度相机重建出来的模型精度较高，并且已经能够应用在动态的场景，但是，深度相机体积较大便携性很差，高质量深度相机价格高昂等因素限制了深度相机重建方案的普及。多相机三维重建不仅需要对多个相机的图像进行帧同步，还要对相机的位置进行标定。多个相机使用不但重建系统的成本没有降低，还使得设备环境搭建比较繁琐。上面的种种因素都制约了深度相机和多相机进行三维重建的应用场景。相比之下，单相机三维重建具备设备成本较低，系统搭建简单的优势，并且随着技术的发展，手机等移动设备使得用户可以简单快速的获取高质量的影像资料。因此，单相机三维重建有着巨大的发展潜力和广阔的应用场景。

单视角重建相比多视角重建而言它可以获取到的信息就更少，所以大多重建方法都会先建立参数化人体模型，然后在参数化人体模型的基础上利用人体的轮廓信息作为约束求解重建问题 [9, 10, 14, 37]。但是 Yu 等人 [34] 虽然没有用参数化的模型做重建但是他们也需要先用 SFM(Structure From Motion) 的方法得到目标初始的三维模型。此外，Guan 等人 [9] 还使用从阴影到结构的方法提高重建模型的质量，消除人体遮挡带来的影响。但是，这也是的方法受限于光源的方向和强度，并且在多个光源的情况下会使得重建问题变得十分复杂。总的来说，单视角由于投影变换损失的信息过多，且现实中不可避免的遮挡问题，上述方法或多或少都需要人为的添加一些约束而无法做到自动化单视角人体三维重建。

Guan等人 [9] 的方法如图 1.11 所示。他们首先计算出单幅图片上的人体姿态、光源方向、人体体型以及前景和背景的分割，并利用公开的人体模型数据集训练出 SCAPE 模型，接着利用人体的轮廓、边缘和阴影信息，将一个静态人体模型拟合到图像上得到人体的体型信息，最后得到的三维模型可以动画等领域。由于他们充分利用了图像中的阴影、边缘信息，所以他们的方法可以应对人体有遮挡的情况。尽管如此，他们的方法在人体姿态估计时需要手动在图像上标注出人体关节点作为初值。

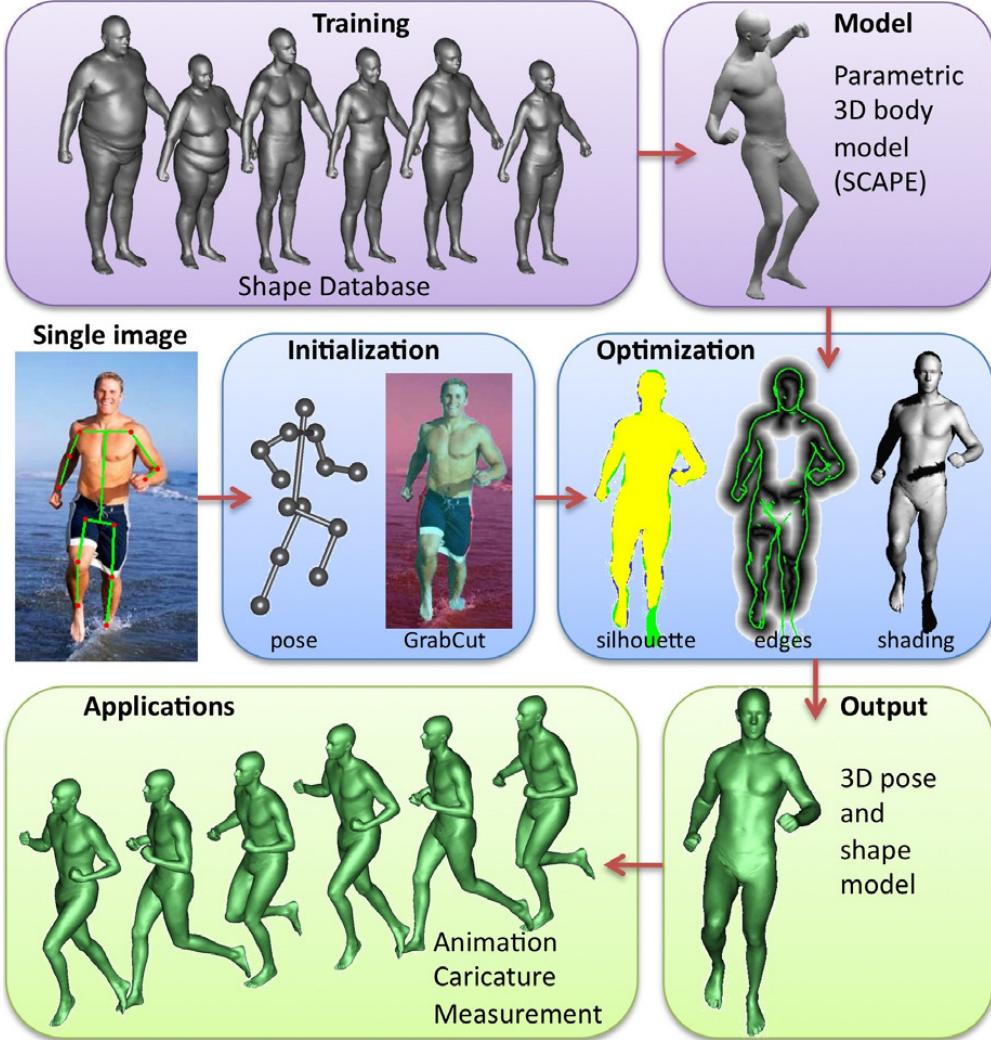


图 1.11: Guan 等人 [9] 方法示意图。

Hasler等人 [10] 没有使用比较流行的 SCAPE 模型而是利用数据集训练出一种双线性静态模型，如图 1.12(a)。这种模型的参数集可以线性地解释成人体的姿态和体型，并且模型的驱动不是依靠人体的姿态信息。然后使用拉普拉斯变换将双线性模型的体型跟图像上的人体进行融合得到图 1.12(b)。接着将人体的姿态或者体型跟图形进行融合得到图 1.12(c)。最后迭代将人体姿态、体型跟图像上人体轮廓进行融合，使得模型尽可能跟图像上人体轮廓融合 1.12(d)，最终得到人体模型 1.12(e)。虽然多张图片可以提高他们模型的精度，但是他们也需要手动标注一些特征，并且重建出来模型的表面依旧有许多的缺陷。

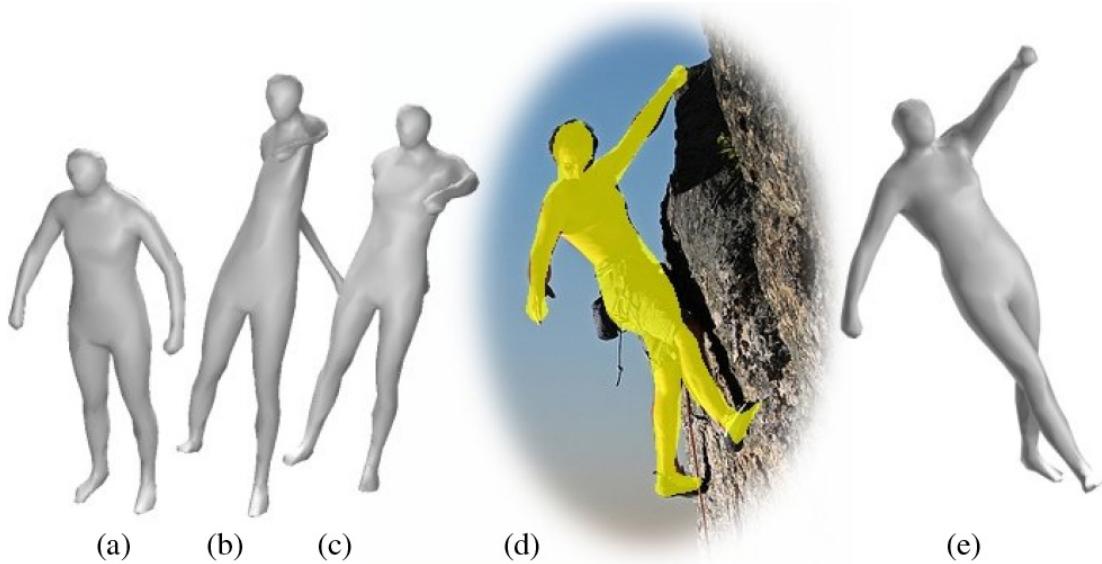


图 1.12: Hasler 等人 [10] 方法示意图。

Jain 等人 [14] 方法的特点在于当参数模型跟图像上的人体融合之后他们可以通过修改参数模型的参数（身高、腿长、肌肉量等）从而改变图像上人体的体型，如图 1.13。他们将参数化的三维模型投影到图片中的人体轮廓上，利用 KLT（Kanade-Lucas-Tomasi）算法跟踪图像上手动标定的特征点，借此同时优化出图片中人物的三维体型和姿态。他们的参数模型是有别于 SCAPE 模型的，他们没有训练 SCAPE 中每个三角面片因姿态变换而引起的变换矩阵，而是直接将模型和人体骨架进行融合绑定，因为在对图像上人体参数进行修改时是不需要这样的细节的。

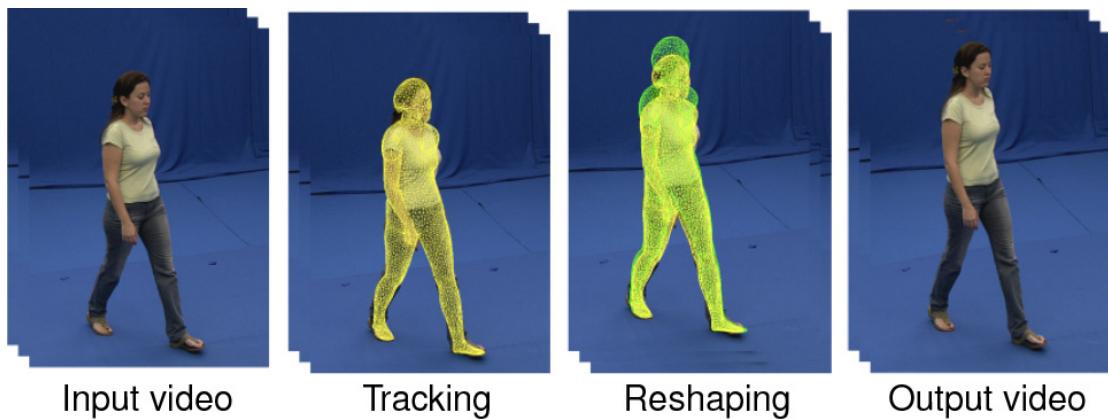


图 1.13: Jain 等人 [14] 方法能够对图像中的人体体型进行改变。

Zhou等人 [37] 的工作跟Jain等人 [14] 的工作很相似，但是他们使用的是SCAPE模型，并且在将模型和图像上人体轮廓融合时使用fit-and-adjust方法。图1.14说明了他们的融合方法。他们首先利用用户手动标注的关节点估计出人体三维姿态，图1.14(a)。接着找出模型轮廓和人体轮廓的对应点，图1.14(b)。利用人体上半身的对应点先对人体的体型、姿态进行估计并调整人体模型，图1.14(c)。再利用调整人体模型之后可以找到更多对应点（图1.14(c)人体下半身）优化人体姿态和体型，图1.14(d)。

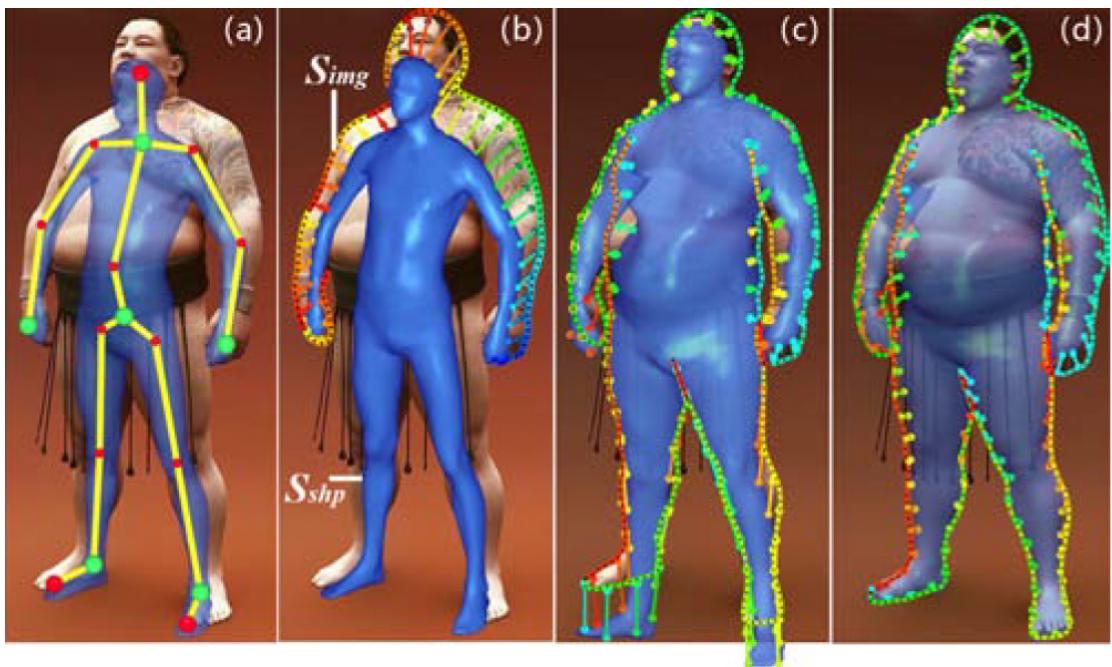


图1.14: Zhou等人 [37] 模型融合方法。

Yu等人 [34] 跟前面方法的主要区别在于他们并没有使用参数化模型，但是他们也需要用SFM的方法获取一个人脸的三维模型，然后用这个模型来做非刚性物体的实时动态重建。他们主要是用SFM重建时得到模型每个点的颜色信息，将非刚性实时重建转化成优化问题。

1.2 本文主要工作

由于人体具有较高的灵活性，单个相机获取的获取的信息十分有限，因此自动化的动态人体的单相机三维重建一直是一个充满挑战的问题。本文提出的方法能够借助一个普通的RGB相机自动的重建出一个动态的人体三维模型，如

图 1.15。图中第一行是我们通过相机获取的人体运动的序列，第二行是我们重建出来的每一帧中人体的三维模型。

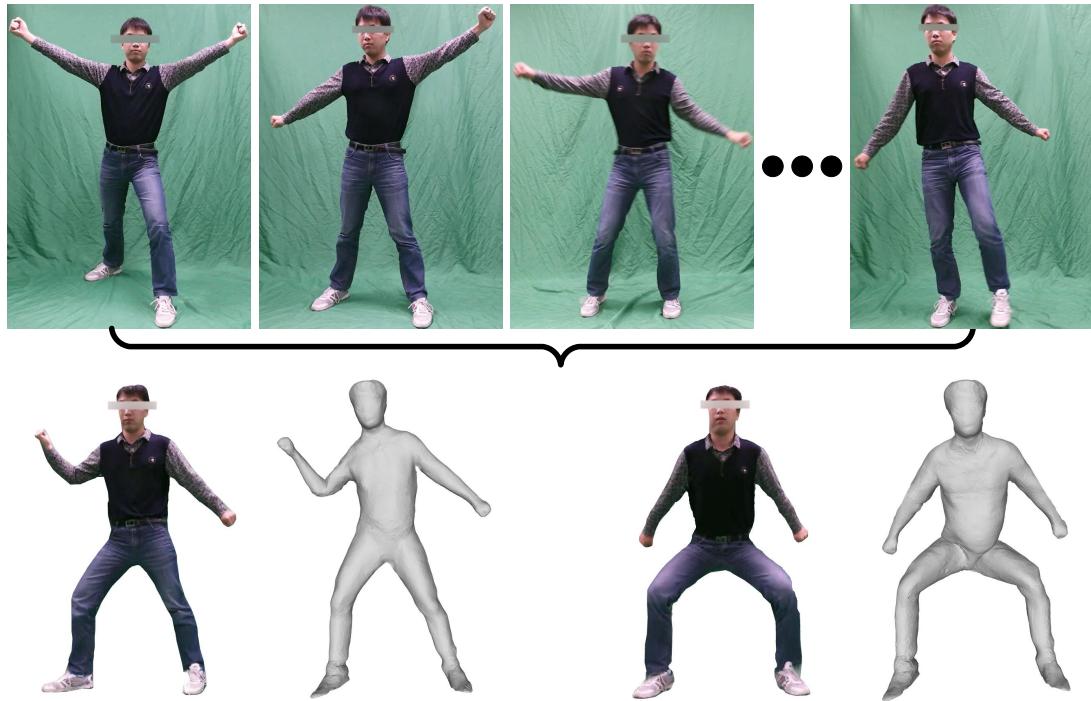


图 1.15: 系统概要

在前人的工作中，人体的运动会使得问题的约束条件不足并最终导致重建结果的质量下降。但是，我们通过对人体的运动分析发现：人体的运动可以提供一些额外的信息来帮助我们提高重建模型的质量。因为单相机图像的获取其实是空间物体做投影变换，加上物体本身的遮挡这会使得物体很大一部分信息的丢失。但如果物体能够运动起来，我们就可以在某个时间段捕获到之前丢失的一部分信息，借助这些信息我们就可以提高人体模型的重建精度。

虽然人体是一个比较复杂的非刚性体，但是由于人体骨骼的存在，人体所能产生的形变是有限的。在人体运动过程中，我们先将人体运动的序列划分成很多时间段，然后利用人体运动变化的信息我们对每段时间中的人体从运动学上将人体划分成许多的“刚性块”，这样就将人体的非刚性重建问题退化成刚性物体的三维重建。对于人体不同的“刚性快”，我们用它跟一个参数化的人体模型对应的部位进行融合变形。随着数据的不断增加，参数化模型不同的身体部分跟真实的人体的实际模型之间的差距越来越小，最终得到一个精度较高的人体三维模型。由于我们最终得到的模型本身是一个参数化的模型，因此我

们可以将其变换到人体不同姿势下如图 1.15 中第二行。因此，本文的主要工作可以总结为一下几点：

- 利用卡尔曼滤波器优化提取到的人体三维姿态；
- 率先提出将人体的运动作为分类图像上人体部位的重要依据，并在此基础上提出一个运动学分类器，用光流表示人体在图像上的运动，将图像上人体部位进行分类，具有较高的精度；
- 将 SCAPE 模型跟重建出来的人体部位模型进行融合，提取人体模型在运动过程中的刚性不变成分和非刚性成分；
- 得到人体的参数化模型，并可以将模型驱动到任意姿态下，应用场景广泛。

与前人的方法相比本文所介绍的方法有如下优势：

- 与深度相机重建相比，设备成本较低，能够重建出动态人体的三维模型；
- 与多相机重建相比，数据采集环境简单，不需要进行复杂的相机标定与同步；
- 与单相机重建相比，我们不需要人为添加约束条件，高度自动化；
- 设计了一个运动学人体部位分类器，优化了人体 3D 姿态精度；
- 是首个用普通 RGB 相机自动化重建出动态人体三维模型的系统。

1.3 本文的章节组织形式

本文研究的重点研究的问题是如何利用一个普通的 RGB 相机拍摄人体运动的视频，在没有人为添加约束的前提下自动重建出高质量的人三维模型。为了验证方法的有效性和稳定性，我们建立一个数据集，将我们的方法跟别的方法进行对比验证。本文的章节组织如下：

第一章：绪论。主要介绍了人体三维重建的研究现状。我们按照重建人体模型时所用设备的不同深入分析了深度相机人体重建、多视角人体重建和单视角人体重建的研究现状。通过探讨他们方法所存在问题从而引出本文工作主要内容及其意义。最后还介绍了本文的章节组织形式。

第二章：人体三维模型重建。对我们单相机动态人体三维重建方法做了详细的介绍。通过公开的数据集我们训练出自己的 SCAPE 模型，然后借助人体姿

态检测获取每帧图像人体的三维骨架信息。结合人体运动信息，我们对图像上的人体进行运动学分类将非刚性重建问题转化成刚性重建。利用稠密光流我们获取到同一身体部位上点的对应关系，结合稀疏光束平差法我们获得不同身体部位不同时刻的三维模型和 SCAPE 模型融合得到许多实例。最后通过对实例进行刚性和非刚性运动成分的分解得到可驱动的参数化模型。

第三章：结论和讨论。我们公布了实验中所有参数的具体数值，以便于读者实现该系统。为了验证系统的有效性和鲁棒性，我们用不同的数据对系统进行验证，并跟别人的算法进行定性和定量的细致分析。最后分析了我们算法的时间复杂度，找出我们方法的效率瓶颈，找出系统可能的改进方向。

第四章：总结与展望。总结本文的工作，强调我们方法的可行性和创新性，同时指出我们方法的不足之处以及后续改进的方向。

第二章 人体三维模型重建

在这个章节中，我们将详细地介绍图 2.1 种展示的动态人体三维重建系统。系统整体上可以分为以下几个主要步骤：

1. 训练数据集，获取 SCAPE 模型。
2. 我们使用一个软件检测子来获取图像上人体 2D、3D 的关节点位置，而不需要去手动标定图像上人体的关节点位置。
3. 利用人体的运动学信息，我们优化了获取到的人体三维姿态信息，并将图像上的人体划分成许多刚性块的组合，使得动态人体三维重建从一个非刚性重建问题转换成刚性重建问题。
4. 在不同的时间段，利用身体部位的光流并结合光束平差法(Bundle Adjustment)我们获得到人体不同部位的三维形状。随着越来越多的图像参投入到模型重建，不同身体部位重建的精度不断提高。
5. 一个通用的参数化人体模型通过非刚性形变方法跟之前不同时间段获取到的人体不同部位的模型进行融合，并提取出人体运动过程中不同部位的刚性成分和非刚性成分。最终获取到人体体型模型，和每一帧下人体的三维模型。

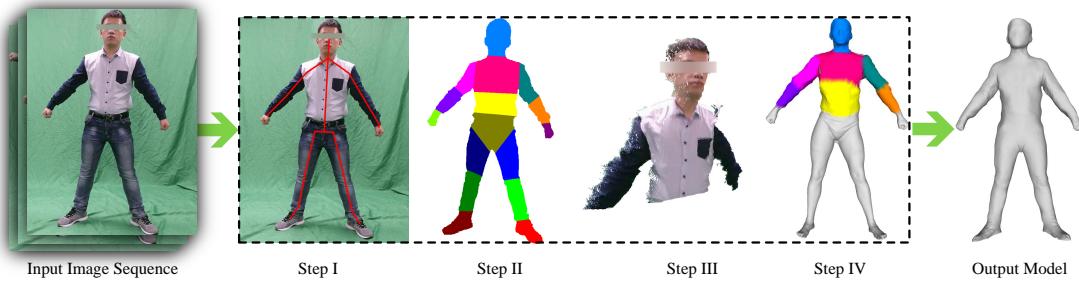


图 2.1：系统流程示意图

2.1 人体模型建模

在我们的系统中，我们利用了Hasler等人[11]提供的数据集，结合Anguelov等人[2]提供的方法训练出如图2.2所示的SCAPE模型。由于我们只需要驱动模型到不同的人体自然下而不需要对模型的体型进行变换，所以我们只需要训练出SCAPE模型中的姿态变换参数就可以了。我们使用这个模型来解决重建过程中姿态变换和模型与不同刚性块的融合问题。

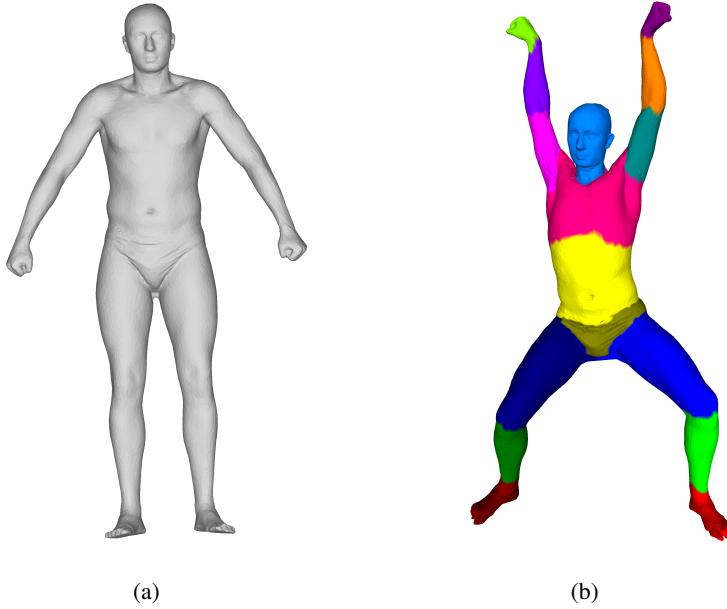


图 2.2: 重建过程中使用的参数化模型。(a): 训练得到的 SCAPE 模型; (b): 模型的不同肢体部位。

2.1.1 模型姿态驱动

我们用 Y^i 表示数据集中同一个人第 i 个不同姿态下的人体模型，我们通过对模型上的每个三角面片 p_k 进行一个两步变换从而实现用姿态驱动人体模型。每个三角面片的变换有刚性变换和非刚性变换组成。假设三角面片 p_k 由 $x_{k,1}$, $x_{k,2}$ 和 $x_{k,3}$ 三个点组成。我们先将三角面片变换到全局坐标系中，然后以三角面片 p_k 上的点 $x_{k,1}$ 建立局部坐标系并对三角面片做形变。因此对三角面片的变形就转换成对三角面片 p_k 上两条边 $\hat{v}_{k,j} = x_{k,j} - x_{k,1}, j = 2, 3$ 。

我们用一个 3×3 的线性变换矩阵 Q_k^i 来表示人体姿态 Y_i 下三角面片 p_k 的非刚性形变，用旋转矩阵 $R_{B_k}^i$ 表示与该三角面片绑定的骨骼 B_k 的刚性旋转。我们先用 Q_k^i 对三角面片 p_k 做非刚性形变，然后用 $R_{B_k}^i$ 对变换之后的三角面片做刚性旋转变换，因此可以写成：

$$v_{k,j}^i = R_{B_k}^i Q_k^i \hat{v}_{k,j}, j = 2, 3 \quad (2.1)$$

三角面片的形变过程如图 2.3 所示。这样一种形变方法的好处就是，既可以表现骨骼的刚性形变又允许模型局部的形变（例如肌肉的形变）。

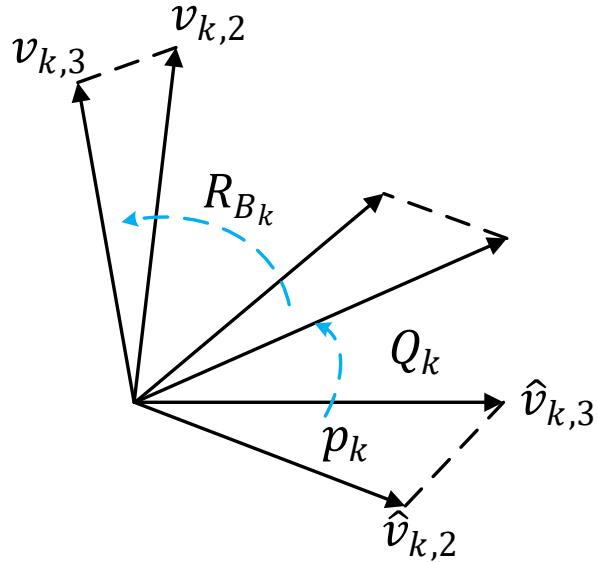


图 2.3: 三角面片变形过程。

当我们拥有某个人体姿态下的形变矩阵集 Q 和 R 时，我们就可以将模型驱动到该姿态下。首先，我们将模型上某一个特定点 y 移动到某一个位置，然后对于每一个独立的三角面片 p_k ，我们用 $R_k Q_k \hat{v}_{k,j}$ 预测出 p_k 的所有边。因为，不同三角面片预测出来的边缺乏一致性，所以为了保证变换后的模型是连续一致的，我们通过求解目标函数 (2.2) 得到模型上所有点的位置。

$$\min_{y_1, \dots, y_M} = \sum_k \sum_{j=2,3} \|R_{B_k}^i Q_k^i \hat{v}_{j,k} - (y_{j,k} - y_{1,k})\|^2 \quad (2.2)$$

2.1.2 姿态建模

上一章讲述了如何用某一姿态下的 Q 和 R 把模型驱动到该姿态下，这里就介绍如何通过数据集学到姿态变换模型中的一些参数。

因为人体的骨骼是铰链式的，所以人体的姿态可以用一个相互连接的相对旋转矩阵集来表示。如果 R_{B1} 和 R_{B2} 是两个相连骨骼的绝对旋转矩阵，那么这两个骨骼的相对关节点的相对旋转矩阵就可以简单表示为 $R_{B1}^T R_{B2}$ 。关节旋转可以很方便的用于关节相邻的两个骨骼的扭来表示。如果用 M 来表示任意的旋转矩阵，并且用 m_{ij} 表示矩阵 M 的第 i 行第 j 列的数值。那么关节旋转就可以

一个三维向量的扭 t 来表示，并且可以用公式 2.3 来表示。

$$t = \frac{\|\theta\|}{2\sin\|\theta\|} \begin{bmatrix} m_{32} - m_{23} \\ m_{13} - m_{31} \\ m_{21} - m_{12} \end{bmatrix} \quad (2.3)$$

$$\theta = \cos^{-1} \left(\frac{\text{tr}(M) - 1}{2} \right)$$

其中扭向量的方向表示了旋转轴，向量模表示旋转多少。

对于每个三角面片 p_k ，我们假设变换矩阵 Q_k^i 仅需要两临近关节点就可以计算出来，从而变换矩阵 Q_k^i 和两临近关节点 $\Delta r_{B_k}^i = (\Delta r_{B_k,1}^i, \Delta r_{B_k,2}^i)$ 组成的扭之间的回归问题的维度就大大降低了。

每个关节的旋转可以用三个参数来表示，所以整个 $\Delta r_{B_k}^i$ 只有 6 个参数，再加上一个偏移量，我们就可以用一个 7×1 的向量 $\mathbf{a}_{k,lm}$ 来表示矩阵 Q 的 9 个值，如公式 (2.4) 所示。

$$q_{k,lm}^i = \mathbf{a}_{k,lm}^T \cdot \begin{bmatrix} \Delta r_{B_k}^i \\ 1 \end{bmatrix} \quad l, m = 1, 2, 3 \quad (2.4)$$

因此，对于每个三角面片 p_k ，我们只要求解出 9×7 个值： $\mathbf{a}_k = \mathbf{a}_{k,lm} \quad l, m = 1, 2, 3$ 就可以得到 $Q_k^i = \mathcal{L}_{\mathbf{a}_k}(\Delta r_{B_k}^i)$ 。

所以现在的目标就是求解出参数 $a_{k,lm}$ 。对于每个实例 Y^i ，如果我们给出变换矩阵 Q_k^i 和每个部位的刚性旋转矩阵 R^i ，我们可以写成如公式 (2.5) 所示的最小二乘函数，其求解方式就是对于每个三角面片 k 和矩阵 $q_{k,lm}$ 分别独立求解。

$$\min_{\mathbf{a}_{k,lm}} \sum_i (([\Delta r^i \quad 1] \mathbf{a}_{k,lm} - q_{k,lm}^i))^2 \quad (2.5)$$

在实际的求解过程中，我们通过标记只有一个或两个自由度的关节点来降低模型的维度和计算量。尽管有些三个自由度的关节点在某些情况下可能导致模型过拟合，但是我们通过对关节 Δr^i 角度的观测值做 PCA 移除旋转矩阵特征值小于 0.1 的旋转轴，与此同时我们也不需要计算 Δr^i 所对应的向量 $\mathbf{a}_{k,lm}$ 。阈值 0.1 是通过观察许多特征值的有序图得到的。我们从图中发现当减少模型大约三分之一的参数数量时，模型有最小的交叉验证误差。

在数据预处理时，我们就可以计算出模型每个部位的旋转矩阵。然而，每个三角面片对于的变换矩阵 Q_k^i 任然未知。我们将用 Q_k^i 对模型进行变换得到的预测模型和数据集中实际模型进行拟合从而求解出 Q_k^i 。不幸的是，这个模型是欠约束的。因此，我们借鉴了 Sumner 等人 [25] 和 Allen 等人 [1] 的方法，并

且在此基础上引入了一个平滑项作为约束。平滑项的含义就是同一个部位相邻的三角面片应该有相似的变换。所以，对于每个模型 Y^i ，我们求解目标函数得到了每个三角面片的变换矩阵 $\{Q_1^i, \dots, Q_P^i\}$

$$\min_{\{Q_1^i, \dots, Q_P^i\}} \sum_k \sum_{j=2,3} \|R_k^i Q_k^i \hat{v}_{k,j} - v_{k,j}^i\|^2 + w_s \sum_{k_1, k_2 \text{ adj}} I(B_{k_1} = B_{k_2}) \cdot \|Q_{k_1}^i - Q_{k_2}^i\|^2 \quad (2.6)$$

其中， $w_s = 0.001\rho$ ，并且 ρ 是模型的分辨率。 $I(\cdot)$ 是一个指示函数。该函数的求解方法就是对模型每个部位和变换矩阵 Q 的每一行分别单独求解。求解到矩阵 Q 之后就可以利用公式 (2.5) 求解出每个三角面片 9×7 的回归参数 \mathbf{a}_k 最终得到的 SCAPE 模型如图 2.2。

2.1.3 模型体型驱动

模型的体型变换是跟人体的姿态没有关系的，所以我们将不同体型的人通过姿态变换将他们驱动到同一姿态下，并且为了描述人体体型变化我们引入了一个新的线性变换矩阵集 S_k^i 来表示第 i 个实例上第 k 的三角面片的形变。并且我们假设实例 i 上的三角面片 p_k 的形变是先通过姿态变换 Q_k^i ，然后再进行人体体型变换 S_k^i ，最后再进行对应身体部位的旋转变换 $R_{B_k}^i$ ，具体可用公式 2.9 表示。

$$v_{k,j}^i = R_{B_k}^i S_k^i Q_k^i \hat{v}_{k,j} \quad (2.7)$$

2.1.4 体型建模

由于人体模型是有许多的三角面片组成，结合公式 2.9 我们就可以将模型的体型参数用 $S^i = \{S_k^i : k = 1, \dots, P\}$ 。我们可以将 S^i 写成 $9 \times N$ ，并用一个线性子空间来表示：

$$S^i = \overline{U\beta^i + \mu} \quad (2.8)$$

我们可以通过 PCA(Principal Component Analysis) 求解出参数 U, μ 以及系数 β^i 。

但是这里我们还不知道每个三角面片的体型变换 S_k^i ，所以，我们利用数据集中不同体型的人计算出 S_k^i 。跟上面的思路一样，我们求解目标函数 (2.9) 就可以得到 S^i 。

$$\min_{S^i} \sum_k \sum_{j=2,3} \|R_k^i S_k^i Q_k^i \hat{v}_{k,j} - v_{k,j}^i\|^2 + w_s \sum_{k_1, k_2 \text{ adj}} \|S_{k_1}^i - S_{k_2}^i\|^2 \quad (2.9)$$

由于这里位置了只有 S_k^i ，所以我们可以直接使用最小二乘求解。

2.2 人体三维姿态检测

我们从单帧图像中恢复出人体的三维姿态来驱动 SCAPE 模型。对于每一帧图像，我们都用 Yang 等人 [33] 提供的 2D 人体姿态检测来获取图像上人体关节点

的位置，然后使用 Wang 等人 [28] 的方法将图像上人体 2D 的姿态转换成 3D 的姿态。接着，由于相邻帧图片上人体的运动是有一定关联性的，所以我们使用一个卡尔曼滤波器 [15] 来优化我们得到 2D、3D 人体姿态，从而降低了错误的 2D 关节点对 3D 关节点的影响，提高系统的鲁棒性。

因为相邻帧上的人体姿态变化是一定的关联性，所以我们把人体 3D 的姿态看做是一个在时间上离散的线性动态系统。我们假设 P_i 表示第 i 帧图片中人体的三维姿态，由于相邻帧中人体姿态具有较高的相关性，所以我们可以认为 P_i 是由第 $i-1$ 帧上人体三维姿态 P_{i-1} 变换而来。这样一种动态关系我们可以用公式表示为：

$$P_i = F_i P_{i-1} + B_i u_i + w_i \quad (2.10)$$

其中 F_i 是作用在前一时刻人体姿态 P_{i-1} 上的状态转换模型， B_i 是作用在控制向量 u_i 上的控制输入模型。 w_i 是系统中的过程噪声，并且满足常量协方差零均值的多元正态分布。在第 i 帧中，我们利用公式 (2.11) 获取人体真是三维姿态 P_i 的估计值 Z_i

$$Z_i = H_i P_i + v_i \quad (2.11)$$

其中 H_i 是一个观测模型； v_i 是系统的观测噪声，它是有常量协方差的零均值高斯白噪声模型 [15]。从图像中获取到人体的三维姿态之后，我们就使用上述的卡尔曼滤波器结合从上一帧中获取的人体三维姿态来优化当前的人体三维姿态。

2.3 运动学分类

虽然上面获取到的人体姿态已经提供了图像中人体部位分类的粗略信息，但是由于相机的透视投影、人体运动存在遮挡等因素，直接用前面获取到的人体姿态做人体部位分类是十分粗糙不准确的。因为人体是靠骨骼支撑起来的，运动中骨骼的形变程度非常小，所以在较短的时间内我们可以忽略肢体部位表面的形变，认为该部位的运动是一个刚性体。尽管人体上临近肢体间是靠关节连接的，但是属于不同肢体部位的运动是不同的。在这里我们提出了一个运动学的分类器，借助人体不同部位的运动信息来比较准确的划分出人体的不同部位，从而进一步优化人体的二维、三维姿态，为后面重建出高质量的人体模型打下坚实的基础。在物理学中，刚体是一个在运动中不发生任何形变的物体，换而言之，我们可以认为刚体在运动过程中都有着相似的位移。根据之前获取的人体三维姿态并结合 SCAPE 模型，我们将人体划分成 16 个刚性块。图 2.4 说明了对图像上人体部位进行运动学分类的全过程。

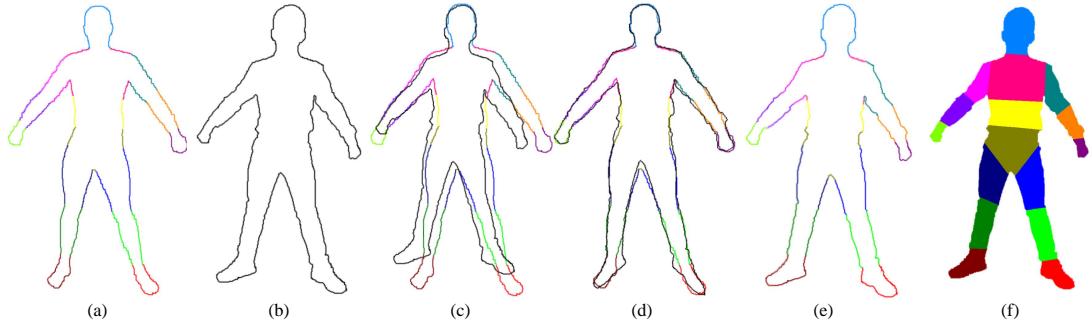


图 2.4: 运动学分类步骤。 (a): 将 SCAPE 模型投影到图像上获得包含分类信息的轮廓; (b): 图像中人体的轮廓; (c): SCAPE 模型轮廓和人体轮廓配准前; (d): 利用 CPD(Coherent Point Drift) [21] 方法进行人体轮廓的匹配结果; (e): 借助 SCAPE 模型信息对图像上人体轮廓分类的结果; (f): 运动学分类结果。首先, 我们提取模型投影到图像上的轮廓以及图像上人体的轮廓; 然后, 我们通过将 (b) 中人体轮廓和 (a) 中模型的轮廓进行配准并将人体轮廓进行分类得到 (e) 中结果。最后, 我们利用全连通的条件随机场模型来获得图像上人体不同部位的分类结果 (f)。

2.3.1 人体轮廓非刚性匹配

SCAPE 模型本身包含的人体肢体分类信息是我们对图像上人体进行分类的重要依据之一。因为 SCAPE 模型跟图像上人体没有对应关系, 所以 SCAPE 模型上的分类信息是无法直接使用的, 因此, 我们借助一致性点漂移算法(CPD, Coherent Point Drift Algorithm) [21] 将 SCAPE 模型上的初始分类信息转换到当前图像上的人体上将其作为后续运动学人体分类的一个重要依据。

由于我们使用的 SCAPE 模型本身跟目标人体模型之间的相似度较低, 再加上从图像中提取的三维人体姿态包含误差, 所以将 SCAPE 模型驱动到当前姿态下并将其投影到二维空间得到的 SCAPE 模型轮廓和人体轮廓重叠度很低, 如 2.5 中第三列。为了尽可能地从 SCAPE 模型上获取比较准确的分类信息, 我们将 SCAPE 模型驱动到当前姿态下的二维轮廓和图像上的人体轮廓进行非刚性匹配, 从而得到 SCAPE 模型二维轮廓跟图像上人体轮廓的映射关系, 如图 2.5。

CPD 算法将两个点集的非刚性配准问题转换成高斯混合模型(GMM, Gaussian Mixture Model)概率密度函数的参数估计。CPD 算法将其中一个点集的每个点看作高斯混合模型中每个分量的质心, 而将另一个点集的点当做高斯混合模型生成的数据。通过最大化高斯混合模型的后验概率将两个点集进行拟合并

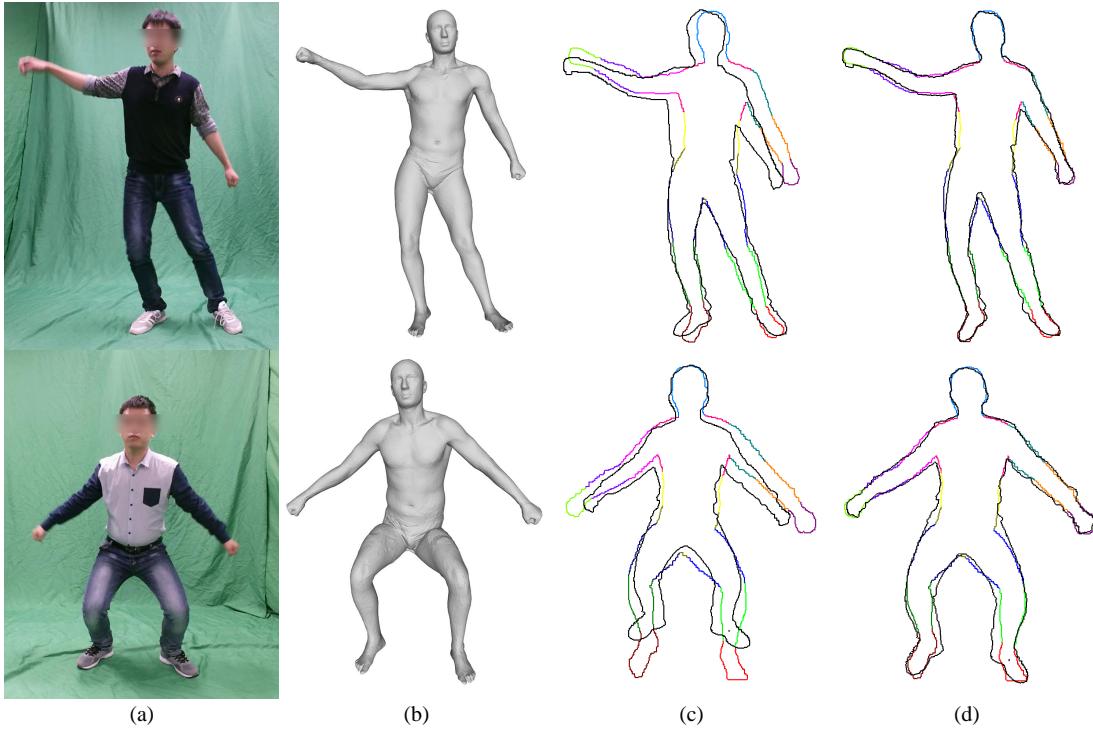


图 2.5: 利用 CPD 将图像上 SCAPE 模型轮廓和人体轮廓进行配准的结果。(a) 人体当前的姿态; (b) 驱动 SCAPE 模型到当前姿态下; (c) 人体轮廓和 SCAPE 模型轮廓配准之前; (d) 人体轮廓和 SCAPE 模型轮廓配准结果。

可以获取两个点集之间的对应关系。算法的核心是将模板点集高斯混合模型每个分量的质心看作一个整体，在保证点集内部拓扑关系的情况下向目标点集漂移。

假设 $\mathbf{X}_{N \times D} = [\mathbf{x}_1, \dots, \mathbf{x}_N]$ 为目标点集， $\mathbf{Y}_{M \times D} = [\mathbf{y}_1, \dots, \mathbf{y}_M]$ 为模板点集，其中 D 为点集的维度， N, M 分别为模板点集和数据点集的数量。我们将点集 \mathbf{Y} 中的点看作是混合高斯模型的质心，点集 \mathbf{X} 中的点为该混合高斯模型生成的数据点。混合高斯模型的概率密度函数定义为：

$$p(\mathbf{x}) = \sum_{m=1}^{M+1} P(m) p(\mathbf{x}|m) \quad (2.12)$$

其中 $p(\mathbf{x}|m) = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}_m\|^2}{2\sigma^2}\right)$ 。为了处理模型中的噪声和外点，我们向混合高斯模型中添加了一个额外的联合分布 $p(\mathbf{x}|M+1) = \frac{1}{N}$ 。我们在所有的混合高斯模型中都使用了相等的协方差 σ^2 和相等的成员概率 $P(m) = \frac{1}{M}, m =$

$1, \dots, M$ 。联合分布概率的权重为 $w, 0 \leq w \leq 1$ ，从而混合模型的形式为：

$$p(\mathbf{x}) = w \frac{1}{N} + (1 - w) \sum_{m=1}^M \frac{1}{M} p(\mathbf{x}|m) \quad (2.13)$$

我们用参数集 θ 来重新表示混合高斯模型的质心，并且通过最大化似然概率求解出混合高斯模型的质心。此外，我们还可以最小化负对数似然概率函数 (2.14) 求解混合高斯模型的质心，这跟最大化似然概率求解是等价的。

$$E(\theta, \delta^2) = - \sum_{n=1}^N \log \sum_{m=1}^{M+1} P(m) p(\mathbf{x}_n|m) \quad (2.14)$$

我们还定义了两点 \mathbf{y}_m 和 \mathbf{x}_n 是对应点的可能性为给定点集上混合高斯模型质心的后验概率： $P(m|\mathbf{x}_n) = P(m) p(\mathbf{x}_n|m) / p(\mathbf{x}_n)$ 。

我们使用最大期望(EM, Expectation Maximization)算法来求解出 θ 和 σ^2 。EM 算法的思路是先给参数猜测出一个初值(“old”参数值)，然后用贝叶斯利用计算出混合成分的后验概率分布 $P^{old}(m|\mathbf{x}_n)$ ，这就是 EM 算法中的“E-step”。然后，通过最小化负对数似然概率 (2.15) 求解出新的(“new”)参数值，这是 EM 算法中的“M-step”。

$$Q = - \sum_{n=1}^N \sum_{m=1}^{M+1} P^{old}(m|\mathbf{x}_n) \log(P^{new} p^{new}(\mathbf{x}_n|m)) \quad (2.15)$$

EM 算法通过反复迭代求解“E-step”和“M-step”直到函数收敛于某一极值。我们用 θ 和 σ^2 将公式 (2.15) 改写为：

$$Q(\theta, \sigma^2) = \frac{1}{2\sigma^2} \sum_{n=1}^N \sum_{m=1}^M P^{old}(m|\mathbf{x}_n) \|\mathbf{x}_n - \tau(\mathbf{y}_m, \theta)\|^2 + \frac{N_p D}{2} \log \sigma^2 \quad (2.16)$$

其中 $N_p = \sum_{n=1}^N \sum_{m=1}^M P^{old}(m|\mathbf{x}_n) \leq N$ ，当且仅当 $w = 0$ 时 $N = N_p$ 。 P^{old} 表示从已知参数计算到的混合高斯模型的后验概率：

$$P^{old}(m|\mathbf{x}_n) = \frac{\exp\left(-\frac{1}{2} \left\| \frac{\mathbf{x}_n - \tau(\mathbf{y}_m, \theta^{old})}{\theta^{old}} \right\|^2\right)}{\sum_{k=1}^M \exp\left(-\frac{1}{2} \left\| \frac{\mathbf{x}_n - \tau(\mathbf{y}_m, \theta^{old})}{\theta^{old}} \right\|^2\right) + c} \quad (2.17)$$

其中有 $c = (2\pi\sigma^2)^{D/2} \frac{w}{1-w} \frac{M}{N}$ 。当目标函数 (2.14) 还未收敛于极值时，最小化函数 (2.16) 也就最小化了 (2.14)。 τ 表示点集的变换。由于，将 SCAPE 模型的二维轮廓配准到图像上人体轮廓上，所以整个拟合过程是非刚性的。因此， τ 就可以用 Tikhonov [5, 31] 正则框架来求解。

2.3.2 人体部位运动学分类

我们把每一帧图像看成是定义在变量 $\{I_1, \dots, I_N\}$ 上的随机场 \mathbf{I} , 其中 I_i 是图像上像素 i 所处位置 p_i 处的像素值, N 是图像上像素点的个数。此外, 随机场 \mathbf{X} 是定义在变量集 $\{X_1, \dots, X_N\}$ 上用来给每个像素 i 确定其分类标签 X_i 的。我们用 \mathbf{d}_i 来表示像素 i 在相邻帧上的位移, 并且通过由 Brox 和 Malik [4] 提出的大位移光流算法来计算出位移的具体数值。如公式 (2.18) 所示, 通过最小化条件随机场的吉布斯能量来获得每个像素的最优分类结果。

$$E(\mathbf{I}, \mathbf{X}) = \sum_i \psi_u(I_i, X_i) + \sum_{i < j} \psi_p(X_i, X_j, I_i, I_j) \quad (2.18)$$

其中, 变量 i 和 j 的取值范围是 $[1, N]$ 。单点势能(unary potential) ψ_u 由公式 (2.19) 推导而出, 它表示了标签 X 和输入图像 I 的匹配程度。

$$C(I_i, X_i) = \underbrace{\omega_s u_s(p_i, \mathbf{S}(X_i))}_{silhouette} + \underbrace{\omega_j u_j(p_i, \mathbf{B}(X_i))}_{joints} + \underbrace{\omega_m u_m(\mathbf{d}_i, \mathbf{D}(X_i))}_{motion} \quad (2.19)$$

s.t.

$$\omega_s + \omega_j + \omega_m = 1$$

其中 ω_s , ω_j 和 ω_m 是权重因子, 并且 $\mathbf{S}(X_i)$ 表示每个分类标签 X_i 所对应的图像上人体轮廓点。因为 SCAPE 模型上有不同肢体部位所对应点集的信息, 所以不同标签 X_i 所对应的人体轮廓点可以通过将 SCAPE 模型投影到图像上, 利用 CPD (Coherent Point Drift) [21] 算法将模型的轮廓和人体的轮廓进行非刚性配准得到。轮廓项能量 u_s 是由计算图像上 p_i 处的点和人体轮廓上点的最小二乘距离表示的。

$$u_s(p_i, \mathbf{S}(X_i)) = \exp\left(-\frac{\min(|p_i - \mathbf{S}(X_i)|^2)}{2\theta_s^2}\right) \quad (2.20)$$

$\mathbf{B}(X_i)$ 是图像上被标记为 X_i 的骨骼的重心。因此, 关节项能量 u_j 通过公式 (2.21) 计算得到。

$$u_j(p_i, \mathbf{B}(X_i)) = \exp\left(-\frac{|p_i - \mathbf{B}(X_i)|^2}{2\theta_j^2}\right) \quad (2.21)$$

其中参数 θ_s 和 θ_j 控制图像上像素点间的连接强度和范围。

此外, 我们提出了运动项能量 u_m 来表示像素 I_i 的运动跟刚性块 X_i 的运动的匹配程度。 $\mathbf{D}(X_i)$ 表示标签为 X_i 的骨骼的重心位移, 这个是通过相邻帧上

人体姿态的变化计算得到。像素 I_i 的运动是通过 LDOF (Large Displace Optical Flow) [4] 光流算法获取到的，我们将得到的光流映射到颜色空间，用不同的颜色来表示光流的方向，用不同的亮度表示光流的大小，结果如图 2.6 中(c)所示。从图中可以看出，在某一时刻同一部位的运动是极其相似的，从而说明不同部位的运动特点确实是划分不同身体部位的可靠依据。

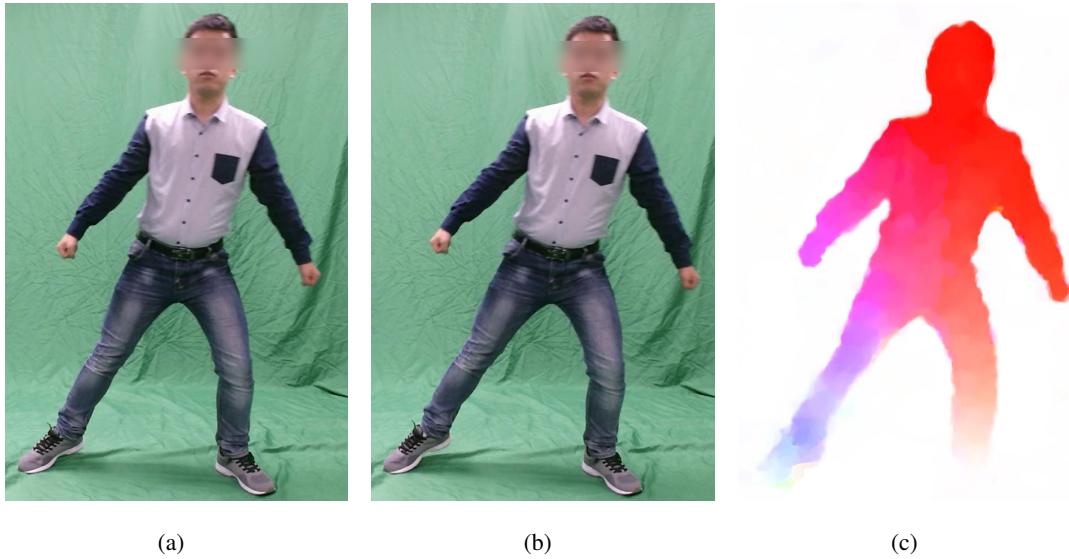


图 2.6: 稠密光流。(a): 人上一时刻的姿态; (b): 人当前的姿态; (c): (b) 图相对于(a)图的光流。

$$u_m(\mathbf{d}_i, \mathbf{D}(X_i)) = \rho_{\mathbf{d}} \left(1 + \frac{\mathbf{d}_i \cdot \mathbf{D}(X_i)}{|\mathbf{d}_i| |\mathbf{D}(X_i)|} \right) \quad (2.22)$$

运动项能量通过评估像素 I_i 和标签为 X_i 的骨骼重心的位移匹配程度提高了图像上人体部位分类的精度。

公式 (2.18) 中成对点势能(pairwise potentials) ψ_p 是由公式 (2.23) 计算得到。

$$\psi_p(X_i, X_j, I_i, I_j) = \mu(x_i, x_j) \exp \left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|X_i - X_j|^2}{2\theta_\beta^2} \right) + \mu(x_i, x_j) \exp(u_m(\mathbf{d}_i, \mathbf{d}_j) - 1) \quad (2.23)$$

其中 θ_α 和 θ_β 参数表示了图像上像素点间的相似程度和临近程度。标签兼容函数 $\mu(x_i, x_j)$ 通过波特模型得到， $\mu(x_i, x_j) = [x_i \neq x_j]$ 。 ψ_p 的作用是处于某个区域中有相似运动的像素点更有可能属于同一个肢体部位。我们使用由 Krahenbuhl 等人 [16] 实现的高效平均场求解方法来优化目标函数 (2.18)。

为了检验公式 (2.19) 中我们提出的运动项的有效性，我们让 $\omega_m = 0$ ，然后用同样的求解方法来求解上述的稠密条件随机场。对比结果如图 2.7 所示，包

含运动项所得到的分类结果(图中右边的结果)比没有利用运动项获得的结果(图中中间的结果)要精确很多。有运动项存在的情况下人体的右腿和腹部之间的划分更加准确，并且在手臂遮挡住胸口的情况下也能准确区分手臂和胸部。

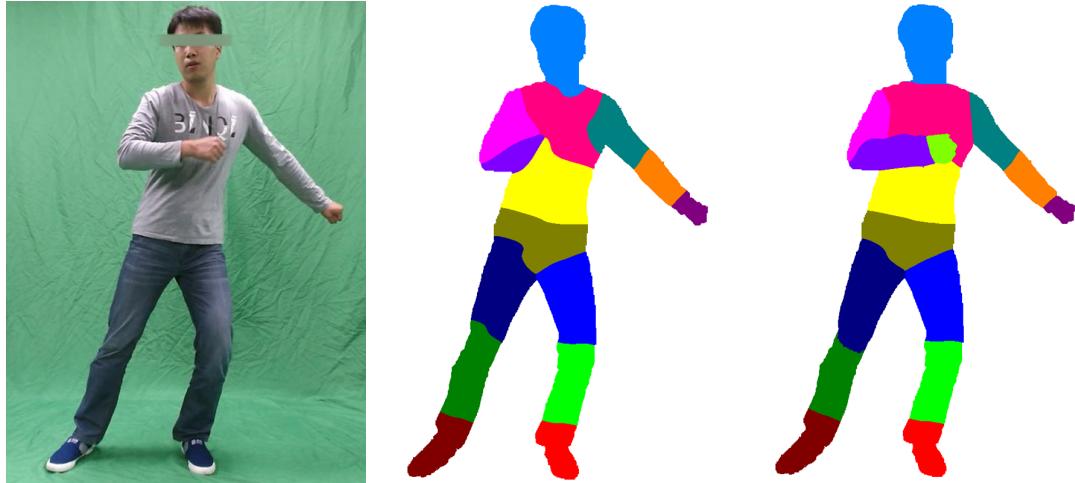


图 2.7: 身体部位运动学分类。左边的是通过 RGB 相机获取的原始图像；中间的图是没有运动项参与时的分类结果；右边的结果表明在运动项的参与下分类的结果很精确并且能够处理有遮挡的情况。

2.4 人体部位稠密三维重建

在对图像上的人体进行运动学分类之后，再结合亚像素级的稠密光流我们就可以找到图像上每个身体部位像素点之间的对应关系，结合人体的三维姿态利用稀疏光束平差法我们就可以重建出人体各部位的稠密三维模型。

2.4.1 光束平差法

图 2.8 解释了光束平差法的原理。一般情况下光束平差法是用在基于特征的三维重建算法中的最后一步。光束平差法主要是通过最小化特征点的重投影误差的平法和，对整个三维场景的结构、相机的姿态做全局的优化。由于目标函数参数众多且是非线性的，所以一般使用非线性最小二乘算法进行求解。光束平差法使用的最广泛的非线性优化算法是列文伯格-马夸尔特(LM,Levenberg Marquardt)算法。由于列文伯格-马夸尔特 [20] 阻尼策略比较高效且实现起来比较容易，所以列文伯格-马夸尔特算法能够在较多未知参数的情况下快速收敛，从而比较适合求解光束平差法的目标函数。在光束平差法中，列文伯格-马夸尔特的雅克比矩阵是一个较大的稀疏矩阵，所以在求解光束平差法时要对求解过程进行改进使用稀疏化 [19] 的方法进行求解。

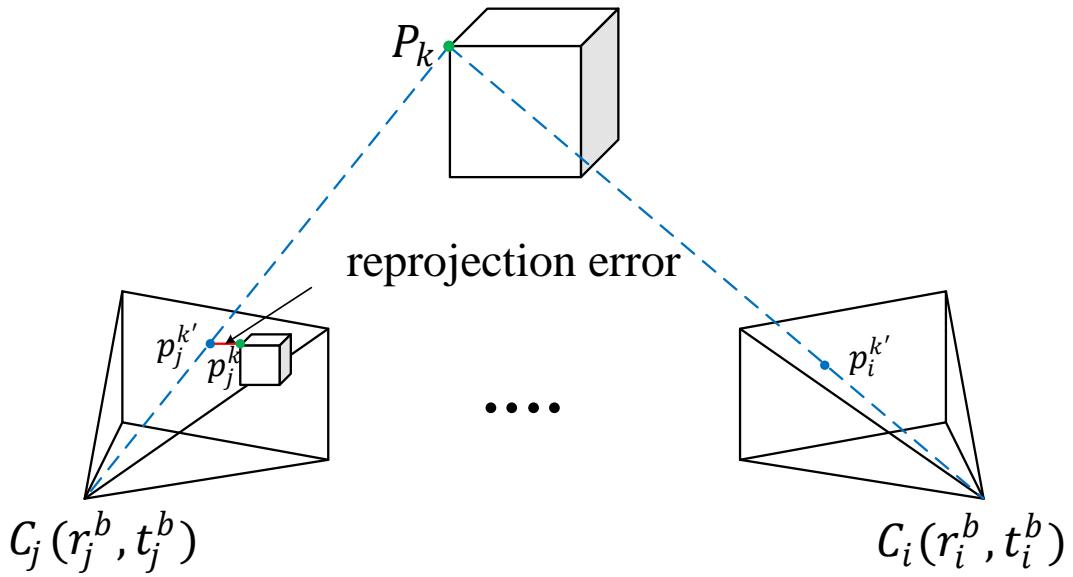


图 2.8: 光束平差法示意图。

我们用 $\pi: \mathbb{R}^3 \rightarrow \mathbb{R}^2$ 表示投影函数，因此光束平差法的目标函数就是：

$$E_{BA} = \sum_{j=f_b}^i \sum_{k \in B_b} \|p_j^k - \pi(r_j^b P_k + t_j^b)\|^2 \quad (2.24)$$

其中帧的索引是从 f_b 变化到 f_i 。身体部位 B_b 上第 k 个点的三维空间位置我们用 P_k 来表示，点 P_k 在第 j 帧上的投影是 p_j^k 。 $T_j^b = [r_j^b | t_j^b]$ 表示相机在第 j 帧时拍摄身体部位 B_b 时相对的相机位置。这里相机的相对位置 T_j^b 是区别于相机的绝对位置 T_j^* 的，相机相对位置不仅包含了身体部位 B_b 的运动还包含了相机本身的运动。以身体部位 B_b 的运动为例，我们可以将参考帧 f_b 对应的人体三维骨架驱动到当前帧 j 帧下就可以获得相机对于身体部位 B_b 的相对位置变换矩阵 $\Delta T_{f_b \rightarrow j}^b$ 。因此，相机相对位置 $T_j^b = \Delta T_{f_b \rightarrow j}^b T_{f_b}^*$ 。

我们使用稀疏光束平差法 [19] 来最小化代价函数 (2.24)，我们使用的稀疏光束平差法在最小化代价函数 (2.24) 时使用稀疏矩阵来存储雅克比矩阵，从而在保证计算效率的同时也不会占用过多的内存。

2.4.2 稠密三维重建

通过对每一帧上的人体进行运动学的分类，利用相邻帧之间的稠密光流我们得到不同身体部位在相邻帧上的像素点之间的对应关系。我们之所以选择用光流来获取点的对应关系，是因为在实际的情况下因身体部位的形变、肢体的运动会导致常见的特征检测只能得到十分稀疏的特征点对，特征点数量过少导

致重建问题欠约束而无法求解。换而言之，目前存在的特征检测方法只适用于静态场景，无法对包含非刚性形变的场景进行特征匹配。我们把获取到的图像序列划分成许多较短的时间段，在这些时间段中对于某个特定的身体部位表面的形变就可以忽略不计，这样我们就把人体的非刚性重建转换成某个时间段中各个身体部位的刚性重建，也就可以使用从运动到结构的经典方法来获取人体不同部位的不同时间段的三维模型。我们对不同身体部位进行三维重建的步骤通过算法 1 进行了详细的说明。

Algorithm 1 身体部位稠密三维重建

```

1:  $i$ : the frame number of the input image
2:  $l_i^b$ : the 2D barycenter of  $B_b$  in the frame  $f_i$ 
3:  $f_b$ : the number of start frame of part  $B_b$ , denoted as ref for simplification
4: for  $b \leftarrow 1$  to 16 do
5:   if  $\|l_i^b - l_{ref}^b\| > 30$  then
6:      $T_{ref}^b \leftarrow T_{ref}'$ 
7:     for  $j \leftarrow f_b + 1$  to  $i$  do
8:        $T_j^b \leftarrow \Delta T_{ref \rightarrow j}^b T_{ref}^b$ 
9:       derive dense optical flow from  $f_{ref}$  to  $f_j$ 
10:    end for
11:     $P^b \leftarrow \min E_{BA}$ 
12:     $f_b \leftarrow f_i$ 
13:  end if
14: end for

```

算法 1 中 f_i 表示 RGB 相机获取到的图像序列中的第 i 帧图像， B_i 表示人体上第 i 个身体部位。

对于每个身体部位，因为在较短的时间内该部位表面的形变可以忽略不计，所以我们认为它在较短时间内运动是刚性运动。我们使用 LDOF(Large Displacement Optical Flow) [4] 算法计算不同帧之间亚像素级的稠密光流，并用稠密光流来跟踪身体部位在这较短时间内运动，从而也就获得图像上身体部位点的对应关系。当任意身体部位在图像上的连续运动距离超过阈值(20个像素)时，

例如，身体部位 B_b 的运动距离超过阈值时，我们选取 f_b 帧(身体部位 B_b 在图像上开始运动)到 f_i (身体部位 B_b 在图像上的运动距离达到阈值)帧之间的所有帧，利用 f_b 帧到 f_i 帧之间相邻帧间的稠密光流获取到像素点之间的对应关系，我们利用光束平差法(Bundle Adjustment)就可以获取身体部位 B_b 的三维模型。

图 2.9展示了我们利用亚像素级稠密点对应关系重建出的不同身体部位的稠密模型。我们先将 (2.24) 中 P_k 看作未知参数，并将其初始深度值设定为2米到4米之间的随机值，然后用稀疏光束平差法计算出三维模型，接着将 (2.24) 中 r_j^b, t_j^b 看作未知数，从而优化人体的姿态参数。尽管亚像素级的稠密光流比较准确，但是系统中还是包含很多的噪声（身体部位分类错误、稠密光流噪声点等），但是通过系数的光束平差法我们对所有的帧进行优化得到比较精细的三维模型并优化人体的三维姿态。



图 2.9: 七个身体部位的稠密重建结果

通过最小化公式 (2.24) 中的重投影误差来确定身体部位在图像上运动范围的阈值为 20 个像素。当设置的阈值过小，进行重建时的短基线就会过短，从而导致算出点的深度值不精确，在公式 (2.24) 中体现为重投影误差较大。如果设置的阈值过大，所选取的帧数就会较多，身体部位表面的非刚性形变较大而不能忽略，那么模型的重投影误差也会较大。所以为了获得比较理想的阈值，我们先选一个比较小的阈值，然后将阈值逐渐增大。选取整个过程中重投影误差最小时的阈值作为我们最终的结果。

2.5 分解人体表面形变

在这章中，我们将人体表面的形变分解成刚性成分和非刚性的成分。由于人体在运动过程中人的体型是不会发生改变的，所以人体表面的刚性成分主要

是由人体体型参数和衣服的刚性运动部分组成。为了得到人体运动的刚性部分和非刚性部分，我们对上面重建出来的人体各部位三维模型进行进一步处理。将已知的 SCAPE 模型和重建出来的人体部位模型结合起来，利用非刚性形变算法将 SCAPE 模型和不同时间段重建出来的人体部位模型进行融合得到模型的许多实例，然后对这些实例的运动进行分解，得到运动中人体模型表面的刚性形变和非刚性形变成分。

2.5.1 模型实例获取

在人体运动过程中，人体模型上的刚性成分是指人体模型运动过程一直保持不变的成分，它跟人的姿态变化没有任何关系。换而言之，即便是在不同的时间不同的姿态下人体模型上的同一部位的刚性部分都是相同的。由于上述重建算法在某个时间区间中重建的模型可能只是几个肢体部位(如图 2.9)而不是一个完整的人体模型，并且这些模型只是包含了人体模型上的几何信息并没有详细的语义信息。所以，为了获取到当前人体模型的刚性部分，我们需要将之前在运动学分类 2.3 中使用的 SCAPE 模型和不同时刻肢体部位模型的进行非刚性融合，从而将人体模型上的刚性、非刚性部分分解转化为 SCAPE 模型上的刚性、非刚性分解。虽然从图 2.2 中可以看出我们使用的 SCAPE 模型和我们目标人体模型之间的差异是很大，但是，从运动学分类优化人体姿态 2.3 一节以及本章节中可以看出 SCAPE 模型和目标模型之间的巨大差异对我们的人体部位分类和人体模型成分分解并没有多大影响，这也表明了我们系统的鲁棒性很高，泛化能力很强。

首先，我们利用之前从图片中获得的人体三维姿态将 SCAPE 模型驱动到当前图片中人体的姿态下，然后用 Z-buffer 算法获取该姿态下相机所能获取到的 SCAPE 模型的前景部分，再将前景部分不同的身体部位跟它对应的稠密三维重建结果借助 CPD 算法进行非刚性融合，融合之后我们还获得了 SCAPE 模型和稠密重建模型上的点对应关系。融合之后的 SCAPE 模型就是当前帧下面人体模型的一个实例，部分结果可以参考图 2.10。

为了检验 SCAPE 模型和稠密模型进行非刚性融合的结果，我们计算了融合之后的 SCAPE 实例上所有点跟稠密模型上的对应点之间的欧氏距离并将其作为融合误差，通过将融合误差映射到颜色空间来直观表示身体不同部位的融合误差结果，融合误差见图 2.11。如图所示，SCAPE 模型进行非刚性融合的结果是可以接受的。对应点之间最大的误差只有 0.5 cm，并且只是左手臂上的一个小部分。SCAPE 上前景部分参与融合的点数是 2689，因此得到的平均融合误差仅为 0.8 mm。尽管融合结果包含一些噪声点，但是这些噪声是可容忍的并且对后面的结果几乎不会产生什么影响。

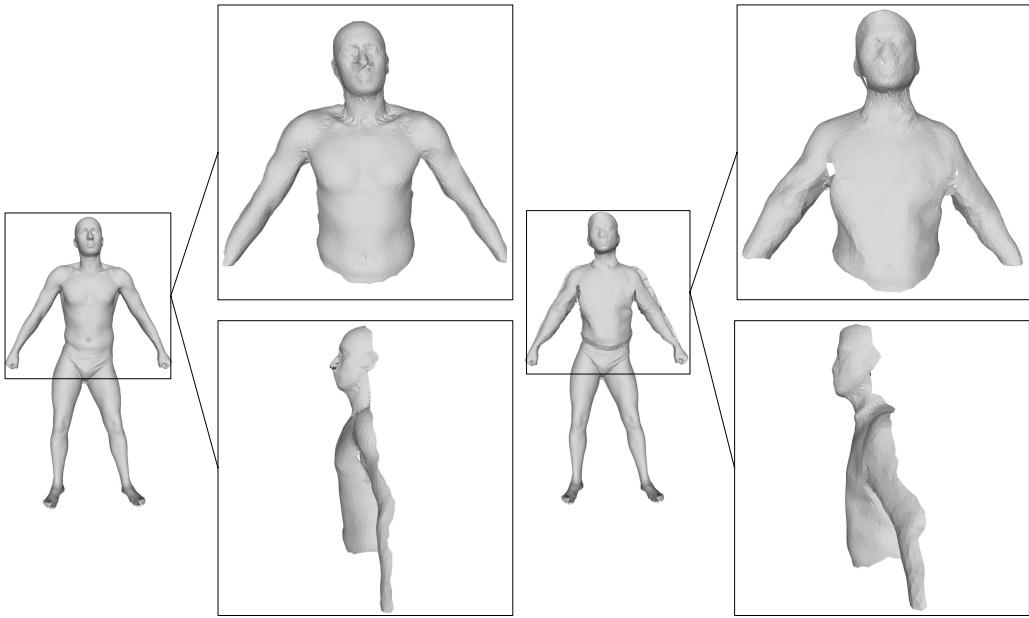


图 2.10: 当前帧下 SCAPE 模型跟上肢人体部位融合得到的实例。左边的图是非刚性融合之前 SCAPE 模型被驱动到当前帧中人体姿态下的结果, 以及 Z-buffer 算法获取到的 SCAPE 模型的前景部分。右图是通过将当前姿态下 SCAPE 模型对应部位的前景跟稠密重建模型进行非刚性融合得到的一个实例。

2.5.2 模型刚性、非刚性成分分解

每次 SCAPE 模型和重建出来的肢体部位稠密模型融合都会得到一个 SCAPE 模型的实例, 我们将得到的所有 SCAPE 实例结合起来推导出实例间的变换矩阵, 通过对变换矩阵进行分解而得到人体模型运动过程中的刚性成分和非刚性成分。我们模型的形变计算是使用类似于训练 SCAPE 时所用的方法, 我们这里的 SCAPE 实例就类似于 [2] 中训练集。我们用一个 3×3 矩阵 D_k^j 来表示 SCAPE 模型上第 k 个三角面片从初始状态下表换到第 j 处实例时的形变矩阵, D_k^j 是通过最小化目标函数 (2.25) 得到。

$$\min_{D^j, R^j} \sum_k \sum_{l=2,3} \rho \|R_k^j D_k^j \hat{u}_{k,l} - u_{k,l}^j\|^2 + \rho \omega_d \sum_{k_1, k_2} \|D_{k_1}^j - D_{k_2}^j\|^2 \quad (2.25)$$

其中, R_k^j 代表刚性旋转矩阵, 并且 $\hat{u}_{k,l} = v_{k,l} - u_{k,1}$, $l = 2, 3$ 表示我们 SCAPE 模型上三角面片 k 的两条边。同样, $u_{k,l}^j$ 代表实例 j 上三角面片 k 的两条边 $u_{k,l}^j$ 。 ρ 是一个指示函数, 在对实例 j 上三角面片 k 进行形变之后, $\rho = 1$; 否则, $\rho = 0$ 。 $\omega_d = 1e^{-3}$ 不仅可以用来避免相邻三角面片间发生较大的形变, 还减少

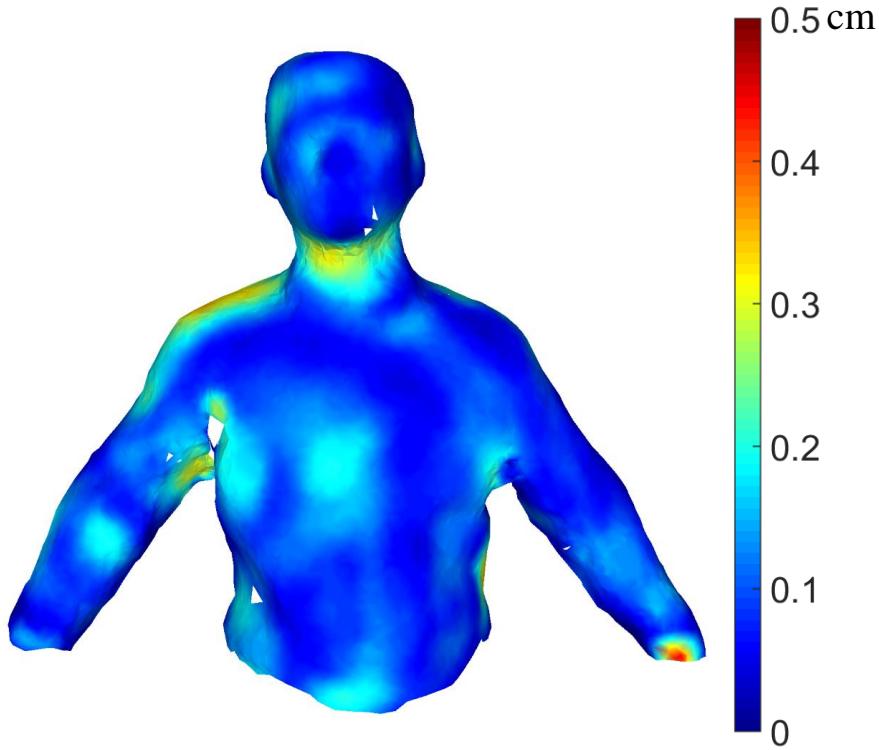


图 2.11: SCAPE 模型和稠密重建结果融合误差图。

稠密模型上噪声点对模型运动成分分解的负面影响。因为形变矩阵 D_k^j 和旋转矩阵 R_k^j 都是未知的，所以目标函数 (2.25) 完全是非线性的。对于优化这种目标函数，我们通过给旋转矩阵 R_k^j 设定一个初值，然后求解 D_k^j ，然后利用 D_k^j 来更新旋转矩阵的值，然后再求解 D_k^j ，如此反复迭代的分别求解 D_k^j , R_k^j 直到目标函数收敛于某处极值。求解的具体过程是：在求解出 D_k^j 之后，旋转矩阵 R_k^j 可以通过扭 ω , $R_k^{j \text{ new}} \leftarrow (\mathbf{I} + [\omega]_{\times}) R_k^{j \text{ old}}$ 来更新它的值，其中 $[\cdot]_{\times}$ 表示叉乘矩阵。所以，当计算出一个 D_k^j 值之后，我们最小化目标函数 (2.26) 得到扭 ω 。

$$\min_{\omega^j} \sum_k \sum_{l=2,3} \rho \| (\mathbf{I} + [\omega]_{\times}) R_k^j D_k^j \hat{u}_{k,l} - u_{k,l}^j \|^2 + \omega_t \sum_{b1,b2} \|\omega_{b1} - \omega_{b2}\|^2 \quad (2.26)$$

其中， b_1 和 b_2 分别表示相邻的两个肢体部位。在反复迭代计算 D_k^j 和 R_k^j 的值直到收敛于某处极值，就得到实例跟初始 SCAPE 模型之间的形变矩阵 D^j 。

因为我们的初始 SCAPE 模型包含了 K 个三角面片，因此每个实例相对于初始模型的形变矩阵 D^j 可以转换成一个维度为 $9 \times K$ 的向量，而这个 $9 \times K$ 的向量可以由一个线性子空间生成： $D^j = U\beta^j + \mu$ 。这里 μ 是就是模型运动中的刚性成分，这个可以通过计算 D^j 的均值得到。当我们获得的实例数量足够多时，我们可以通过 PCA(Principal Component Analysis)很容易的推导出模型运动过程中的非刚性成分 β^j 。

第三章 结果和讨论

为了检验三维人体动态重建系统的鲁棒性和有效性，我们在一个包含 30 个人进行不同运动的数据集上测试了我们的重建系统。图 3.1 展示了我们实验环境，软硬件配置为：

- 处理器：Intel Core i7, 3.4 GHz
- 内存：24 GB RAM
- 软件平台：Windows 7, MATLAB 2015
- 相机：SONY CX260, 分辨率 1920×1080 , 帧率 30 fps

数据集中所有的视频都是使用同一个分辨率为 1920×1080 的 SONY CX260 在图 3.1 中的环境下拍摄的。每个视频大约包含 450 帧图像。一般情况下，我们每 10 帧可以获取一个 SCAPE 模型实例，并且所有的数据都是离线的情况下在上述台式机上处理的。为了评估系统重建出来的精度，我们分别使用我们的方法、KinectFusion 和 Yu 等人 [34] 提出的单目视觉非刚性重建算法进行了定性和定量的对比。此外，我们还将我们的方法跟 Xu 等人 [32] 提出的在人体大幅度运动下使用 Kinect 测量人体体型参数的方法进行的对比。

3.1 系统参数

文章前面的章节中没有详细说明公式中一些参数的具体数值，在我们的实验中这些参数都被设置为固定的值，具体参考表 3.1。

公式	参数值		
公式 (2.19)	$\omega_s = 5/16$	$\omega_j = 1/16$	$\omega_m = 5/8$
公式 (2.20) (2.21) (2.22)	$\theta_s = 60$	$\theta_j = 60$	$\rho_d = 0.5$
公式 (2.23)	$\theta_\alpha = 60$	$\theta_\beta = 25$	

表 3.1: 试验中重建系统相关参数的具体数值。

对于表 3.1 中前两行我们有 6 个参数需要确定数值，但是 ρ_d 被设置为 0.5 来归一化公式 2.22 中的运动项能量。此外，因为 (2.19) 中 $\omega_s + \omega_j + \omega_m = 1$ ，所

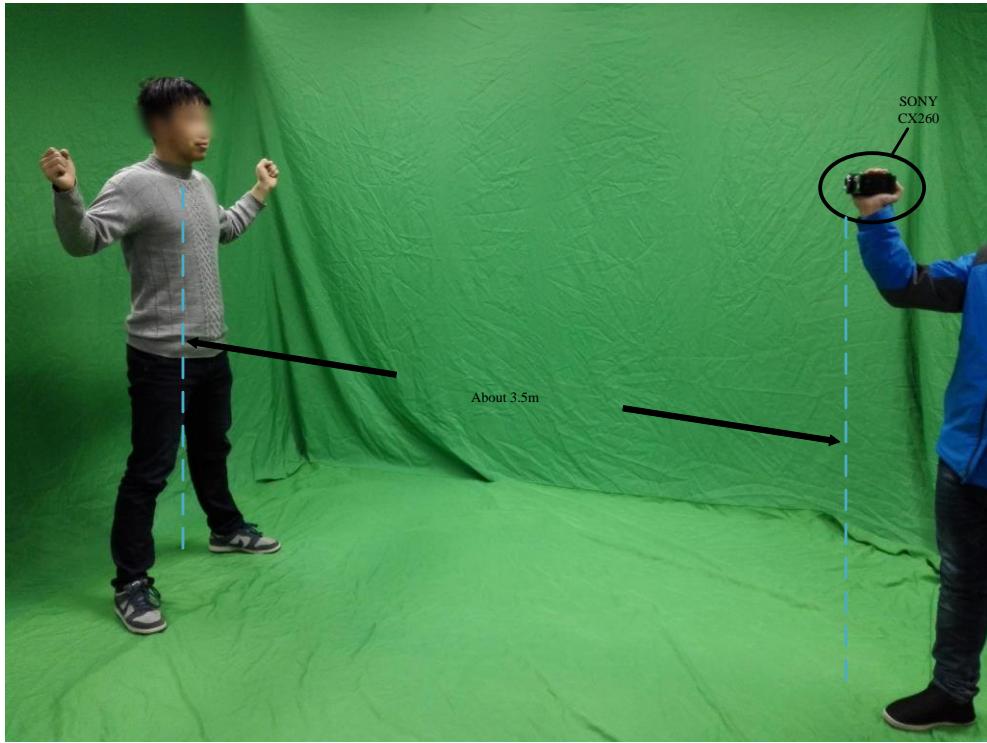
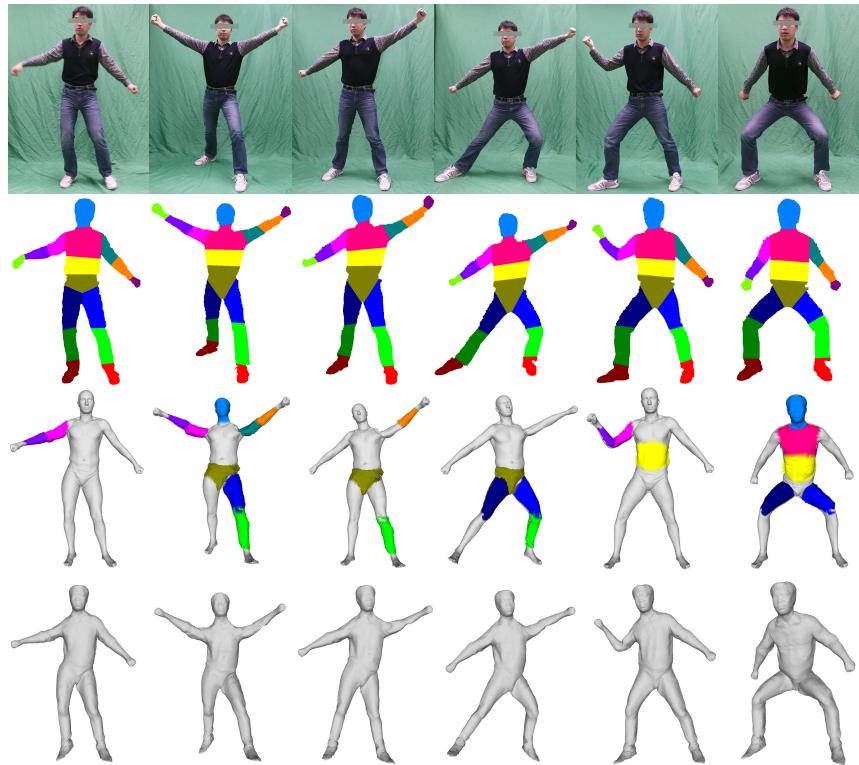


图 3.1: 实验环境。我们只需要一个普通的相机在合适的距离下拍摄人体的运动。

以，我们只需要确定 ω_s , ω_j , θ_s 和 θ_j 的数值。受 Krähenbühl 和 Koltun 在文章 [16] 中确定核函数参数数值方法的启发，我们使用跟他们类似的方法来确定表 3.1 中所提到的参数值。具体过程就是：首先，我们手动标定同一个人不同姿态下 20 张图片的身体部位作为对照数据；然后在改变 ω_s , ω_j , θ_s 和 θ_j 值的同时求解公式 (2.19)，得到 20 张图片上身体部位的分类结果，并计算身体部位分类的误差值；最终选取误差最小时的 ω_s , ω_j , θ_s , θ_j 值。最后再用同样的方法来确定 θ_α 和 θ_β 的数值。

3.2 定性分析

在图 3.2 中我们展示了系统重建出两组数据中不同人在不同姿态时的三维模型。尽管我们没有使用优化算法和局部搜索算法来进一步优化得到的人体三维姿态，但是我们的系统还是从图像序列中恢复出相对准确的人体姿态。我们的系统并不需要运动对象的先验知识就可以重建出带有细节的人体三维模型，系统的泛化能力比较强。



(a)



(b)

图 3.2: 动态人体三维重建结果。

图中每个子图的第一行是图像序列中的部分帧；第二行是二维的人体运动学分类结果；第三行是将 SCAPE 模型和当前帧重建出来的身体部位模型进行融合得到的 SCAPE 实例，不同的身体部位我们用不同的颜色进行了标注；最后一行是将最终重建出的人体模型驱动到当前帧的姿态下并加上当前帧的非刚性成分得到的模型。

图 3.2 中第二行的运动学分类结果说明了我们运动学分类器是比较准确和鲁棒的。正如图中第三行所展示的，我们的 SCAPE 模型和目标人体的三维模型的相似度是十分低的，也就是说我们的系统在不需要对人体提前进行测量的情况下就可以进行重建人体模型，侧面体现了我们的系统的潜在应用场景很广泛。对于不同时间段获取的身体部位模型跟 SCAPE 模型融合的结果，我们分别用不同的颜色在对应的 SCAPE 实例中标注出来了。我们通过不断处理更多的图片来更新重建出来的人体三维模型，随着处理的图片数量的增加模型的精度也逐渐提高。最后我们将模型驱动到不同帧中人体对应的姿态下，在加入当前帧中人体运动的非刚性成分之后就得到拥有细节的人体三维模型，结果如图中最后一行所示。对比图中的 (a)、(b) 子图，尽管我们使用的是同一个 SCAPE 模型但是最终得到的都是特定人的精细模型。

目前最新的非刚性动态重建方法是由 Yu 等人 [34] 提出的，他们先让重建目标保持固定姿态使用从运动到结构的经典算法获取目标稠密的三维模型作为模板，然后重建目标开始运动并重建出每一帧上目标的三维模型。跟他们相比我们的系统有如下的优势：

1. 应用场景广泛。他们的算法需要先让重建目标静止以获得目标的初始三维模型，并且他们在文章中所演示的事例都是从比较近的视角去重建一个比较小的物体，例如手、布娃娃。而我们使用的 SCAPE 模型是一个比较通用的模型，并且对于重建人体模型时不要求去刻意配合，能够提高获取人体模型时的用户体验。
2. 鲁棒性较高。我们用数据集对他们的算法进行了测试比较，结果如图 3.3 所示。虽然我们通过稠密重建算法获取到了两个高质量的人体模型（图 3.3 中最后一行第一个和第四个图），但是他们的算法依旧不能重建出较好的人体模型（图 3.3 中最后一行），并且随着人体运动幅度的增大他们重建出来的模型上的噪声点越来越多。相比之下，我们却可以在使用同一个 SCAPE 模型的前提下，充分利用人体运动所带来的信息不断优化模型的精度，最终得到高质量的人体模型（如图 3.3 中第二行）。
3. 模型质量更高。Yu 的方法重建出来的模型只是人体被拍到部分的三维模

型，是一个不完整的三维模型。而我们的 SCAPE 模型可以补全每一帧中不可见部位的形状



图 3.3: 两个图像序列的对比结果。第一行是两组图片序列中的部分帧；中间一行是我们将最终重建出来的模型驱动到第一行图片对应姿态下的结果；最后一行是使用 Yu 的方法得到对应图片下的对比结果。

我们仔细分析了他们的算法，认为他们的系统无法重建出高质量人体模型的原因是：他们每一帧的模型重建都是通过对最初的模板进行形变而得到，由于他们的模型驱动并不是通过骨骼来驱动的，所以随着人体运动幅度的增大不能将初始模板变换到当前人体的姿态下面，最终导致模型上的噪声点越来越多，模型的质量越来越低。毫无疑问，我们的系统是目前唯一能够通过单个普通相机高效重建动态人体三维模型的系统。

3.3 定量分析

根据我们的调研，目前还没有系统能用单个 RGB 相机重建出一个动态人体的三维模型，因此我们将得到的模型和 KinectFusion 的结果进行比较，并分析我们系统重建模型的误差，结果如图 3.2。人体的真实模型是通过 KinectFusion

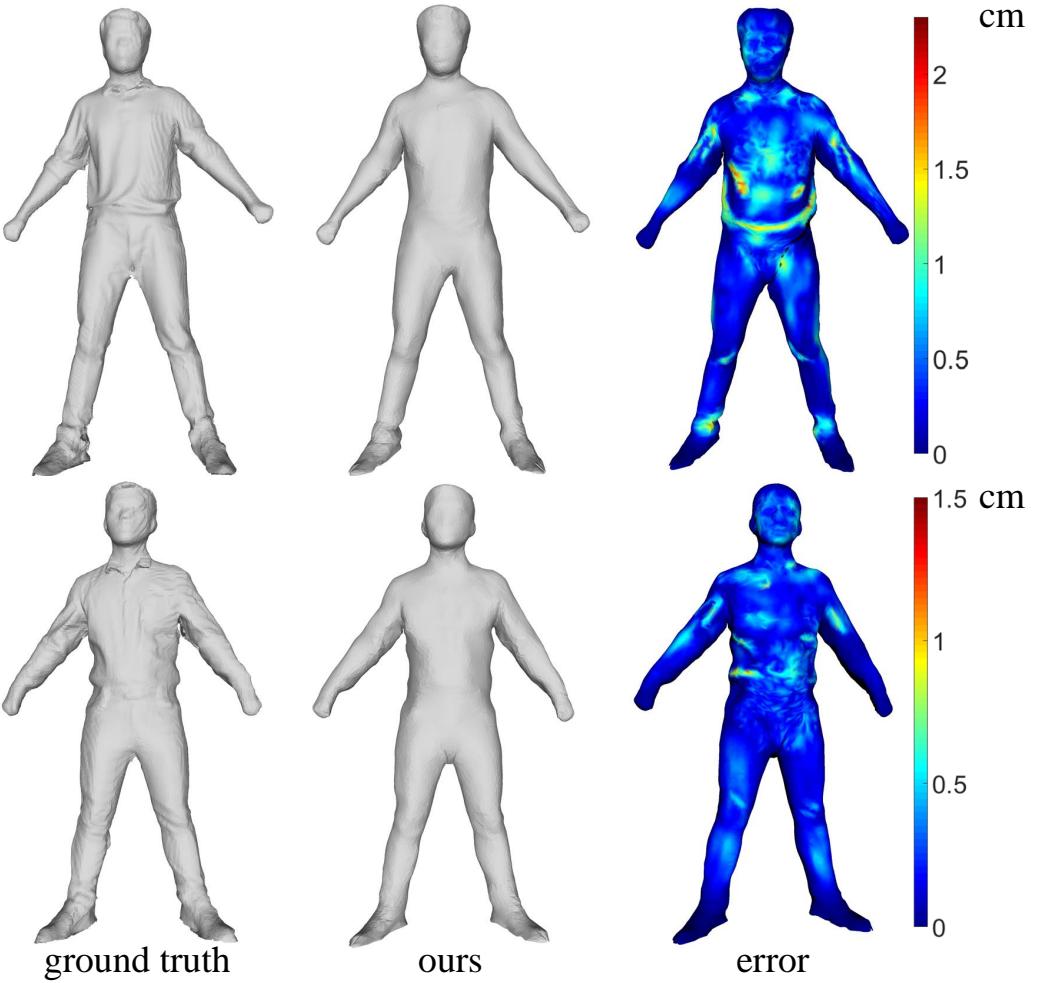


图 3.4: 不同模型的重建误差。

扫描静止时候的人得到的。我们使用的模板是由 12500 个点和 25000 个面组成。为了得到比较准确的重建误差，我们先将模型驱动到 KinectFusion 结果的相同姿势下，接着用刚性 ICP(Iterative Closest Point) 将模型的正面跟 KinectFusion 的结果进行配准，然后将对应点之间的欧式距离作为模型的重建误差并将其映射到色彩空间以直观的展现各部位的重建误差。

如图 3.4 所示，图中第一行是第一个图像序列的重建结果，它的平均重建误差只有 0.3 厘米；图中第二行是第二个图像序列的重建结果，其平均重建误差只有 0.2 厘米。并且我们系统的最大重建误差只有 1.5 厘米，并且只出现在模型上一个很小区域，对模型的应用影响不大。对于第一个图像序列来说，重建误差较大的部分主要出现在模型的腰部，从图 3.2 中左边部分图片可以看出，

因为人物身穿的衣服在腰部最为宽松，所以对应模型腰部的非刚性成分较为丰富，重建误差相对较大。虽然第二个图像序列中人的肢体运动跟第一个不一样，如图 3.2，但是由于第二个人的衣着比较贴身，所以第二个图像序列的重建误差相对较小。

我们得到的模型可以应用到许多的领域，尤其是虚拟试衣。在我们的实验中，我们重建了 30 个不同的人体模型，并且用 Xu 等人 [32] 使用的方法测量了模型的身体参数。我们计算了 30 个人体模型的身体参数误差并将结果和 Xu 的方法进行了比较细致的对比，如表 3.2 所示。尽管，Xu 的方法使用了深度相机来改善人体参数测量的误差，但是我们的系统跟他们相比有较高的精度和易用性：在仅仅使用一个 RGB 相机的情况下，我们系统在人体胸围、上半身长度、大腿周长的测量上的平均误差都比 Xu 的方法小。

Errors	Arm Length	Chest Girth	Neck to Hip Distance	Hip Girth	Thigh Girth
Error of Xu's (cm)	1.2	3.2	4.5	3.1	2.1
Error of Ours (cm)	1.4	2.3	3.7	3.3	1.9

表 3.2: 重建模型的身体参数对比。

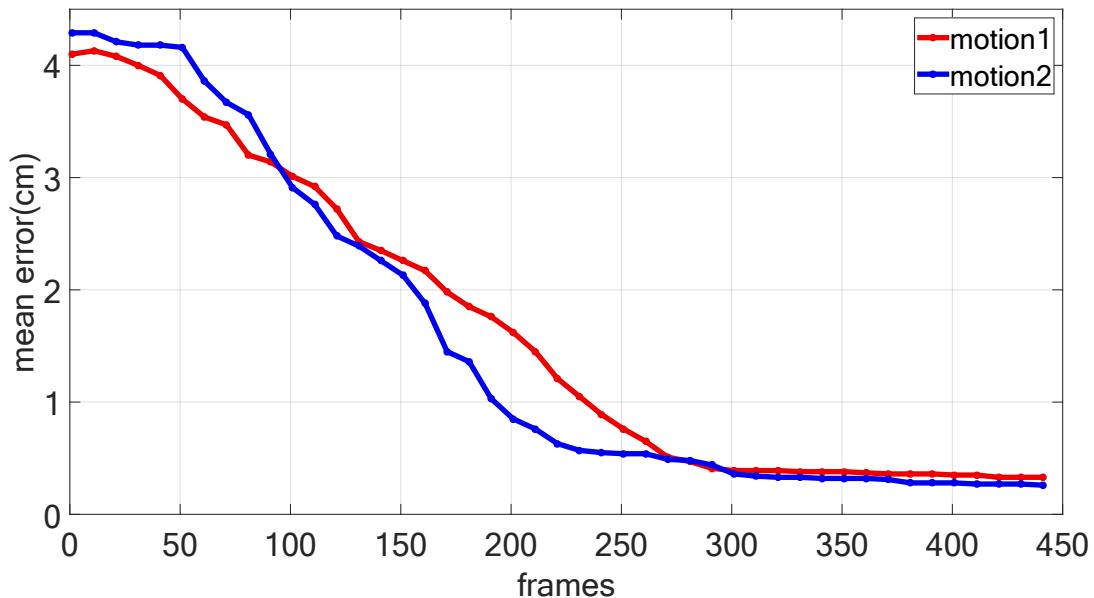


图 3.5: 重建模型和 KinectFusion 之间的平均重建误差。横坐标表示参与重建的图像数量。纵坐标表示模型的平均重建误差。随着参与重建的图像数目增加，模型的平均重建误差逐渐降低并趋于稳定。

此外，我们绘制了模型的平均重建误差和视频帧数之间的关系图来研究输入视频的帧数对重建模型精度的影响，如图 3.5。从图 3.5 可以看出，400 帧之前随着视频帧数的增加重建模型的平均误差不断减小；但是 400 帧以后，随着帧数的增加模型的平均误差下降不明显逐渐趋于稳定。因此，为了保证模型的精度，我们在应用中获取的视频帧数最好大于 400 帧。

3.4 算法复杂度分析

由于 3D 姿态提取、稠密光流获取、稠密条件随机场、光束平差的计算量都很大，因此我们的系统还不能做到实时，处理数据只能进行离线处理。当前情况下每处理一帧图像需要大约 8 分钟时间，其中运动学分类耗费大约 5 分钟是每帧图像处理过程中耗时最多的一部分。在我们的实验中，我们发现最耗时的步骤是人体姿态提取和运动学分类，这两者的时间开销加起来占据了整体计算时间的 90%。根据文献 [33] 的分析，姿态提取的时间复杂度是 $O(N)$ ，其中 N 是图像的尺寸。文献 [16] 详细阐述了稠密条件随机场的具体实现，求解稠密条件随机场问题主要由消息传递（Message Passing）、一致性转换（Compatibility Transform）和局部更新（Local Update）组成。其中计算瓶颈主要是消息传递，是用一个时间复杂度为 $O(N)$ 的高位滤波算法实现的，其中 N 是图像的尺寸。综上所述，我们系统的整体时间复杂度为 $O(Nn)$ ，其中 n 是系统处理的图像帧数。

尽管目前我们的系统还是比较耗时的，但是它的确是第一个通过普通 RGB 相机自动重建出动态人体三维模型的系统。此外，由于图像序列上不同时间段之间是没有影响的，因此我们的系统是可以进行并行化的，我们下一步的打算就是将系统进行并行化并用 GPU 进行加速，提高系统整体的运算速度。

第四章 总结与展望

4.1 总结

本文主要实现了利用单个 RGB 相机自动重建出运动人体的三维模型。由于相机成像相当于把空间中的物体投影到二维空间，损失了很多的信息。所以传统的重建方法要么选择不同的设备（例如深度相机或者多个 RGB 相机）来获取目标物体更多的信息事先物体三维模型重建；要么就是保证目标物体是静止的，通过单相机的不同视角来重建出目标物体模型。尽管 Yu 等人 [34] 通过事先获取目标物体的几何信息事先了单相机动态物体的三维重建，但是他们方法的缺陷也是显而易见的：需要更多的用户交互，是一种非自动化的动态模型重建方法，他们需要目标先保持静止以便获取物体初始状态下的三维模型；鲁棒性不高，由于他们的模型变换不是通过骨骼驱动的，所以如图 3.3 中所示，目标物体运动幅度较大时模型的质量较低；此外，重建出来的模型只是物体的一个表面，并不是完整的模型。相比之下，我们的方法优势较为明显：

- 是一种完全自动化的动态人体模型重建系统。我们不需要知道目标人体模型的先验信息，不需要用户的交互，只利用参数化的 SCAPE 模型可以自动重建出动态人体的三维模型；
- 模型获取成本低。相比深度相机以及多相机采集环境，我们系统的只需要普通的 RGB 相机就可以完成数据获取，成本很低；
- 分离出了人体运动中的刚性成分和非刚性成分，应用场景广阔。分析人体模型中的刚性成分我们可以得到人体模型的体型参数，可以用于虚拟试衣等应用场景，人体模型的非刚性成分可以用于获得人体不同姿态下的精细模型，可以用于计算机动画等应用场景；
- 可获得完整的人体模型。对于相机无法拍摄到的地方，我们可以用 SCAPE 模型进行补全；
- 重建出来模型的精度较高。如图 3.4 中所示，我们系统的重建误差较低，模型质量较高；
- 系统鲁棒性较高。在使用同一个 SCAPE 模型的情况下，我们分别用不同的数据集对系统的鲁棒性进行了验证，都得到质量较高的人体模型。

4.2 展望

由于图片的分辨率不能太低，在我们的实验中因为图像的分辨率为 1920×1080 ，处理每一帧图像都很耗时，所以目前情况下只能离线重建出人体的三维模型而不能实时地获取人体的模型。但是，不同时间段进行人体不同部位的三维模型重建是相互独立的，因此我们的系统是可以并行化并用 GPU 进行加速的。除此之外，虽然我们得到的模型可以应用于虚拟试衣、计算机动画等领域，但是我们的模型还不够精细，所以下一阶段我们的工作重点就是尽可能恢复模型上更多的细节并将算法并行化。

参考文献

- [1] Allen, B., Curless, B., Popovic, Z., 2003. The space of human body shapes: reconstruction and parameterization from range scans. international conference on computer graphics and interactive techniques 22, 587–594.
- [2] Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J., 2005. Scape: shape completion and animation of people, in: ToG, ACM. pp. 408–416.
- [3] Bogo, F., Black, M.J., Loper, M., Romero, J., . Detailed full-body reconstructions of moving people from monocular rgbd sequences .
- [4] Brox, T., Malik, J., 2011. Large displacement optical flow: descriptor matching in variational motion estimation. PAMI 33, 500–513.
- [5] Chen, Z., Haykin, S., 2002. On different facets of regularization theory. Neural Computation 14, 2791–2846.
- [6] De Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H., Thrun, S., 2008. Performance capture from sparse multi-view video. international conference on computer graphics and interactive techniques 27, 98.
- [7] Dou, M., Taylor, J., Fuchs, H., Fitzgibbon, A., Izadi, S., 2015. 3d scanning deformable objects with a single rgbd sensor, in: PAMI, pp. 493–501.
- [8] Gall, J., Stoll, C., De Aguiar, E., Theobalt, C., Rosenhahn, B., Seidel, H.P., 2009. Motion capture using joint skeleton tracking and surface estimation, in: CVPR, IEEE. pp. 1746–1753.
- [9] Guan, P., Weiss, A., Bălan, A.O., Black, M.J., 2009. Estimating human shape and pose from a single image, in: ICCV, IEEE. pp. 1381–1388.
- [10] Hasler, N., Ackermann, H., Rosenhahn, B., Thormahlen, T., Seidel, H.P., 2010. Multilinear pose and body shape estimation of dressed subjects from image sets, in: CVPR, IEEE. pp. 1823–1830.

- [11] Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., Seidel, H.P., 2009. A statistical model of human pose and body shape, in: Computer Graphics Forum, Wiley Online Library. pp. 337–346.
- [12] Huang, C.H., Boyer, E., Navab, N., Ilic, S., 2014. Human shape and pose tracking using keyframes, in: CVPR, IEEE. pp. 3446–3453.
- [13] Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., 2011. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. uist. Proc Uist , 559–568.
- [14] Jain, A., Thormählen, T., Seidel, H.P., Theobalt, C., 2010. Moviereshape: Tracking and reshaping of humans in videos, in: ToG, ACM. p. 148.
- [15] Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. Journal of Fluids Engineering 82, 35–45.
- [16] Krähenbühl, P., Koltun, V., 2012. Efficient inference in fully connected crfs with gaussian edge potentials. arXiv preprint arXiv:1210.5644 .
- [17] Li, H., Vouga, E., Gudym, A., Luo, L., Barron, J.T., Gusev, G., 2013. 3d self-portraits. ToG 32, 187.
- [18] Liu, Y., Gall, J., Stoll, C., Dai, Q., Seidel, H.P., Theobalt, C., 2013. Markerless motion capture of multiple characters using multiview image segmentation. PAMI 35, 2720–2735.
- [19] Lourakis, M.I., Argyros, A.A., 2009. Sba: A software package for generic sparse bundle adjustment. TOMS 36, 2.
- [20] Mor茅, J.J., 1977. The levenberg-marquardt algorithm: Implementation and theory. Lecture Notes in Mathematics 630, 105–116.
- [21] Myronenko, A., Song, X., 2010. Point set registration: Coherent point drift. PAMI 32, 2262–2275.
- [22] Newcombe, R.A., Fox, D., Seitz, S.M., 2015. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time, in: CVPR, pp. 343–352.

- [23] Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A., 2011. Kinectfusion: Real-time dense surface mapping and tracking, in: ISMAR, IEEE. pp. 127–136.
- [24] Robertini, N., De Aguiar, E., Helten, T., Theobalt, C., 2014. Efficient multi-view performance capture of fine-scale surface detail, in: 3DV, IEEE. pp. 5–12.
- [25] Sumner, R.W., Popović, J., 2004. Deformation transfer for triangle meshes. ToG 23, 399–405.
- [26] Theobalt, C., Aguiar, E.D., Stoll, C., Seidel, H.P., Thrun, S., 2010. Performance Capture from Multi-View Video. Springer Berlin Heidelberg.
- [27] Tong, J., Zhou, J., Liu, L., Pan, Z., Yan, H., 2012. Scanning 3d full human bodies using kinects. Visualization and Computer Graphics, IEEE Transactions on 18, 643–650.
- [28] Wang, C., Wang, Y., Lin, Z., Yuille, A.L., Gao, W., 2014. Robust estimation of 3d human poses from a single image, in: CVPR, IEEE. pp. 2369–2376.
- [29] Weiss, A., Hirshberg, D., Black, M.J., 2011. Home 3d body scans from noisy image and range data, in: ICCV, IEEE. pp. 1951–1958.
- [30] Werghi, N., 2007. Segmentation and modeling of full human body shape from 3-d scan data: A survey. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 37, 1122–1136.
- [31] Willoughby, R.A., 2012. Solutions of ill-posed problems (a. n. tikhonov and v. y. arsenin). Siam Review 21, 266–267.
- [32] Xu, H., Yu, Y., Zhou, Y., Li, Y., Du, S., 2013. Measuring accurate body parameters of dressed humans with large-scale motion using a kinect sensor. Sensors 13, 11362–11384.
- [33] Yang, Y., Ramanan, D., 2011. Articulated pose estimation with flexible mixtures-of-parts, in: CVPR, IEEE. pp. 1385–1392.
- [34] Yu, R., Russell, C., Campbell, N., Agapito, L., 2015. Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video, in: ICCV, University of Bath.

- [35] Zeng, M., Zheng, J., Cheng, X., Liu, X., 2013. Templateless quasi-rigid shape modeling with implicit loop-closure, in: CVPR, IEEE. pp. 145–152.
- [36] Zhang, Q., Fu, B., Ye, M., Yang, R., 2014. Quality dynamic human body modeling using a single low-cost depth camera, in: CVPR, IEEE. pp. 676–683.
- [37] Zhou, S., Fu, H., Liu, L., Cohen-Or, D., Han, X., 2010. Parametric reshaping of human bodies in images, in: ToG, ACM. p. 126.
- [38] Zollhöfer, M., Nießner, M., Izadi, S., Rehmann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., Loop, C., Theobalt, C., 等, 2014. Real-time non-rigid reconstruction using an rgb-d camera. ToG 33, 156.

简历与科研成果

基本情况

朱海宇，男，汉族，1991年8月出生，江苏省泰州人。

教育背景

2014.9~2017.6	南京大学电子科学与工程学院图像工程实验室	硕士
2010.9~2014.6	南京大学电子科学与工程学院	本科

攻读硕士学位期间完成的学术成果

- [1] SCI 2 区论文：Zhu H, Yu Y, Zhou Y, Du S. Dynamic Human Body Modeling Using a Single RGB Camera. Sensors. 2016;16(3):402.
- [2] 发明专利：一种自动人体三维重建系统，除导师外第一完成人。

攻读硕士学位期间参与的科研课题

1. TODO

致谢

首先感谢我的指导老师于耀老师，在我研究生阶段他不仅给予了我悉心的指导，教会了我进行科学研究的一般方法，让我对计算机视觉领域三维重建问题有了较深的理解，还在具体的科研项目中启发式指导我解决所遇到的问题，让我在实践中形成了自己的方法论。在跟于老师讨论 SCAPE 模型的计算过程中，感受到了于老师严谨细致的治学态度，学习了于老师高质量的实现代码，收获颇丰。虽然计算机视觉领域的知识一直在快速的迭代更新，但是于老师教会的方法却足以让我受益终身。与此同时，我还要感谢都思丹教授、周余副教授，他们不仅帮助我解决学习、科研上遇到的难题，还关心我生活上所遇到的困难。他们在我的研究生阶段的科研过程中也给予了大量的指导和帮助，使得我在进入实验室不久就进入了状态，找到了自己感兴趣的领域，并收获了一定的成果。跟三位老师一起学习计算机视觉的基础知识和研读人体建模的前沿文章的时光确实是令人怀念和难忘的。

接下来，我也要感谢图像工程实验室的所有同学，尤其是高之泉同学、唐炳晓同学、李云同学、陈希同学，在和他们的交流学习中，他们总是乐于分享自己的知识、观点和理解，让我从中学到了许多知识和方法并发现了自身的不足。在和高之泉同学一起采集人体运动数据的日子虽然很辛苦，但他还是仔细认真地帮我采集数据，减少了我许多的工作量。每一个科研成果的背后，都有一支互相帮助、积极交流、主动分享、团结互助的优秀团队。

虽然本科阶段的微电子知识跟研究生阶段的研究内容不相干，但是那些知识拓宽了我的视野，使我的思路更加开阔。在这里要感谢本科所有的老师，他们各具特色、认真严谨的教学让我终身难忘。

最后，感谢南京大学电子科学与工程学院 14 级全体同学，这是一个充满爱、团结互助的集体。在短暂的研究生生活中，每次遇到困难总会有人给予我无私的帮助。在此，我想对大家说一声谢谢，跟你们在一起的日子很快乐，最后衷心地祝愿大家前程似锦。