无标记运动捕捉与运动映射

（申请清华大学工学博士学位论文）

培 养 单 位 ： 自动化系

学　　科 ： 控制科学与工程

研 究 生 ： 王 雁 刚

指 导 教 师 ： 戴 琼 海 教 授

二〇一四年四月

Markerless Motion Capture and Motion Retargeting

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in

Control Science and Engineering

by

Wang Yangang

Dissertation Supervisor : Professor Dai Qionghai

April, 2014

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）根据《中华人民共和国学位条例暂行实施办法》，向国家图书馆报送可以公开的学位论文。

本人保证遵守上述规定。

（保密的论文在解密后应遵守此规定）

作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘 要

无标记运动捕捉利用视觉信号数字化记录生物的运动信息，是国际上本领域前言研究的热点，在文化、军事、教育、医疗等领域有着广泛的应用价值，主要面临着两大挑战：一是捕捉高精度的运动信息，二是实现运动信息的稳定捕捉与映射。本文围绕无标记运动捕捉的两大挑战，根据运动环境的复杂程度，针对三类特定对象的运动捕捉问题，提出了三种无标记运动捕捉方法，构建了三个无标记运动捕捉的原型系统，主要工作和创新点如下：

1. 提出一种基于参数模型的单张彩色图像的深度估计方法。建立彩色图像与深度之间的非线性映射模型，针对不同类别的图像，自适应学习非线性模型的参数，与已有方法相比，提高了参数模型深度估计的精度。由该方法构建的系统可用于单相机系统的高精度无标记运动捕捉与运动映射。
2. 提出一种单相机系统的无标记运动捕捉方法。针对近距离场景无遮挡或少遮挡的实时脸部表情运动捕捉，建立自适应动态脸部表情模型，同步优化脸部基本表情及组合系数。与当前公开的方法相比，无需任何预处理，显著简化运动捕捉与运动映射流程，解决计算时间复杂度与高精度运动捕捉的矛盾。
3. 提出一种固定多相机系统的无标记运动捕捉方法。针对近距离场景复杂遮挡的手与物体交互运动捕捉，建立手、物体以及相互接触的动力学方程，提出复合运动控制器模型，采用基于接触点采样的控制器优化方法，解决复杂遮挡的交互运动捕捉难题，实现便捷和高精度交互运动的运动映射。
4. 提出一种移动多相机系统的无标记运动捕捉方法。针对大规模场景、户外不可控环境的人体运动捕捉，恢复移动相机的空间位置，重建运动对象的稠密点云，建立运动稀疏性约束与运动对象的动态纹理模型，解决大规模场景的无标记人体运动捕捉难题，提高无标记运动捕捉与运动映射的稳定性。

关键词：无标记，运动捕捉，运动映射，深度估计

Abstract

Markerless motion capture is an appealing solution for digitally recording the biologic motion information with visual signals. It is one of the hottest topics in computer vision and graphics research area and plays an important role in many fields, such as culture, military, education, medical treatment, etc. Currently, there are two major challenges existing in the markerless motion capture: one is to capture the high accuracy motion, the other is to achieve stable motion capture and retargeting results over time. Focus on the two main challenges, the dissertation proposes three types of markerless motion capture system for three specific objects motion according to the complexity of the capture environment, the main contribution of this dissertation includes:

1. Propose a depth estimation method for single color images based on the parametric model. The proposed model describes the correlation between a single color image and depth map with a kernel function in a non-linear mapping space. The model parameters are adaptively learned for different types of images by a well-defined color-depth database. Compared with the state-of-the-arts, the high resolution depth estimation result can be achieved with the learned parameters. The parametric model based depth estimation method can be used to realize the high accuracy single-view motion capture system.
2. Propose a single-view motion capture system for close-up scene, less occlusion real-time facial motion capture. An adaptive dynamic facial expression is constructed for the realtime performance and robust computations. The motion capture system jointly solves for a detailed 3D expression model of the user and the corresponding dynamic tracking parameters. Compared with the state-of-the-arts, the proposed system significantly simplifies the motion capture and motion retargeting process, while requires no user-specific training or calibration, or any other form of manual assistance. The conflict between the computation complexity and high accuracy motion capture can also be solved by the proposed system.
3. Propose a fixed multi-view motion capture system for close-up scene, severe occlusion hand manipulation motion capture. The motion capture system introduces a composite motion control to simultaneously model hand articulation, object movement, and subtle interaction between the hand and object. An optimal motion con-

trol that drives the simulation to best match the observed image data is searched based on the contact information. Convenient and high accuracy motion capture and motion retargeting results can be obtained by the proposed system.

4. Propose a moving multi-view motion capture system for large-scale scene, uncontrolled background human body motion capture. The moving camera parameters and the dense depth maps are firstly recovered from the significant amount of dynamic pixels video streams. Then, a sparse constraint and a dynamic view-dependent texture model is adopted for reducing the pose ambiguity. The proposed system can improve the stability of motion capture and motion retargeting in outdoor and large-scale environment.

Key words: markerless; motion capture; motion retargeting; depth estimation

目 录

主要符号对照表

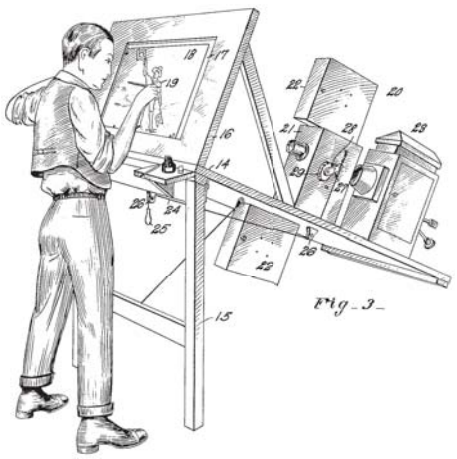
motion	运动
retargeting	映射
MoCap	运动捕捉 (Motion Capture)
prior	先验知识
blend shape	融合形状
DEM	动态表情模型 (Dynamic Expression Model)
deformation	变形
morphable	形变
ICP	迭代最近邻 (Iterated Closet Point)
VH	可视凸壳 (Visual Hull)
SGMM	超高斯混合模型 (Super-Gaussian Mixture Model)
ISA	互动模拟退火 (Interacting Simulated Annealing)
torque	扭矩
penetration	穿透
EM	期望最大 (Expectation-Maximum)
PCA	主成分分析
ICA	独立成分分析
MRF	马尔科夫随机场 (Markov Random Fields)
TOF	飞行时间 (Time of Flight)
ZNCC	零均值归一化互相关
PD	比例微分
analysis-by-synthesis	合成分析
SfM	运动恢复结构 (Structure from Motion)
SLAM	即时定位与地图重建 (Simultaneous Localization and Mapping)
CoSLAM	视觉协同即时定位与地图重建
BP	信度传播 (Belief Propagation)
Jacobian	雅克比
LCP	线性互补问题 (Linear Complementary Problem)
twist	运动旋量

第 1 章 引言

运动捕捉 (Motion Capture, MoCap) 是记录生物的运动信息并将记录下的运动信息转换成可用数学形式的一项技术[1]。近年来，随着各种新型传感器的出现及计算性能的不断提高，运动捕捉技术已经在文化（电影与游戏）、教育（虚拟现实）、医疗（生物测定与康复训练）、军事（机器人设计）等诸多应用中扮演着非常重要的角色，人们也对运动捕捉的内容与性能提出了更高的要求。虽然运动捕捉的外延在不断扩大，从各种信息源中恢复出真实有效的运动数据，并对它们进行分析、整合和映射 (retargeting)依然是运动捕捉的首要任务。值得指出的是，运动捕捉的对象可以是人、动物或其他生物，但由于人体的（包括脸、手、全身等）运动捕捉具有很强的泛化能力，并且与之相关的商业市场最为庞大，因此，以人为对象的运动捕捉在学术和商业领域占有着绝大多数的分量。

事实上，人体运动捕捉技术的发展已历经百年历史。19世纪末，静态成像技术已经非常成熟，如何显示动态的场景和对象成为当时工业界思考的难题。1878年，电影之父 Eadweard J. Muybridge 作为多相机拍摄的第一人，使用排成直线的数部相机拍摄下一匹奔跑骏马的不同图像[2]，这为在荧幕上显示动态变化的场景和人体提供了工程和理论上的保障。随着真实动态场景的记录与显示问题的解决，人们又试图解决虚拟人物的动态制作问题。1915年，漫画家 Max Fleischer 发明了转描机[3]，如图1.1(a)所示。该装置可以将拍摄的影像资料一帧一帧地投影到透明桌面上，通过临摹投影的轮廓，漫画家能够非常容易地制作出逼真的动画片，如图1.1(b)所示。转描机的出现使得制作的成本极大地降低，而在此之前，制作一部动画片通常要耗费大量的人力和财力。1937 年，迪士尼工作室即是使用数部转描机制作出了轰动世界的动画片——《白雪公主》[1]。事实上，转描机作为运动捕捉系统的前身，仅仅是一种二维的处理方法。它将需要拍摄的动态影像资料事先投影到二维平面，然后让艺术家在二维图像上进行加工和创作，以此制作出动画片。但是，人体运动的三维属性（为了便于统一，本文用x,y,z 轴构成的三维笛卡尔坐标系来表示三维空间）使得三维运动捕捉系统的出现成为必然。

20世纪70年代以来，由于受到动画制作产业的极大推动，人体运动捕捉技术获得了快速发展，一大批不同传感器类型的人体运动捕捉系统相继被发明出来。当然，随着运动捕捉技术的不断进步和成熟，现今的运动捕捉系统已经远远突破了动画制作产业。运动捕捉涉及传感器原理、动力学、信号处理、图像处理、计算机视觉、计算机图形学等众多学科，已经成为多学科交叉领域的研究热



(a) Max Fleischer发明的转描机①

(b) 转描机制作的动画帧②

图 1.1　转描机及使用转描机技术进行漫画制作

点。自20 世纪80 年代以来，其更是引起了学术领域研究人员的重点关注。美国的Biomechanics 实验室、麻省理工学院（MIT）计算机和人工智能中心（CSAIL）、卡耐基梅隆大学（CMU）机器人研究中心、纽约大学（NYU）计算机系运动中心等众多高校开展了大量的研究工作。此外，迪士尼、谷歌、微软等众多国际公司也都成立了运动捕捉的研究小组。

1.1　问题由来

目前，已知的人体运动捕捉系统可分为两大类：基于传感器或标记点的运动捕捉系统 (Sensor / Marker system)和无标记的运动捕捉系统 (Markerless system)。基于传感器或标记点的运动捕捉系统是一种“侵入”式运动捕捉。此类系统一般需要在采集对象上安装不同类型的传感器或标记点。根据所安装传感器的种类，可以将其划分为四大类：机电式、电磁式、惯性式和光学式[1]，如图1.2 所示。一般而言，基于传感器或标记点的运动捕捉系统具有准确 (accurate)、实时 (real-time)和鲁棒 (robust) 三大优势。但这些系统的缺点也是显而易见的：采集环境需要特殊定制与设计；运动对象需要穿戴带有传感器或标记点的服装，在某些应用中就很难捕捉特别自然的运动。例如，在捕捉手与物体的交互运动时，传感器或标记点便会影响手与物体之间的交互运动。

针对“侵入”式运动捕捉系统的不足，无标记的运动捕捉系统提供了一个很好的解决方案。此类系统通常只需要一部或数部相机，使用计算机视觉和图形学的相关技术手段，即可完成运动捕捉的任务[4]。由于系统不需要安装任何传感器

① 图片来源于 http://en.wikipedia.org/wiki/Rotoscoping
② 图片来源于 http://www.ufunk.net/en/insolite/disney-rotoscoping/

(a) Gypsy机电系统　(b) Motion Star电磁系统　(c) MVN惯性传感系统　(d) LED光学标记系统

图 1.2　基于传感器或标记点的商用运动捕捉系统，图片来自Google™

或标记点，当然也就无需穿着任何特殊定制的服装。无标记的运动捕捉系统对捕捉对象的干预很低，这大大改善了“侵入”式运动捕捉系统中受限的采集环境，并使目标对象能够实现更为灵活、自然的运动形态[5]；无标记运动捕捉系统可以做到结构小巧、造价低廉，因此这类系统能够胜任很多桌面级的应用环境，如捕捉脸部表情运动[6]，手与物体的交互运动[7]；此外，无标记运动捕捉系统通常需要捕捉运动对象的整体三维结构，因而具有更高的空间分辨率[8]。

与基于传感器或标记点的运动捕捉系统相比，现在的无标记运动捕捉仍然不是一项成熟的技术。虽然近年来出现的新型传感器一定程度上推动了无标记运动捕捉的发展，但由于受到运动的复杂性与不可预知性、运动过程中的遮挡等因素的影响，无标记运动捕捉系统的可靠性与稳定性还较低，并且运动捕捉的精度与质量也无法得到很好的满足[9]。当前，无标记运动捕捉技术面临着两大主要挑战：一是捕捉高精度空间分辨率的运动信息，二是实现随时间变化的、稳定的运动捕捉。

捕捉高精度空间分辨率的运动信息是无标记运动捕捉的首要目标。运动捕捉通过跟踪运动对象上随时间变化的三维“特征点”（feature points）来实现，运动的空间分辨率也就可以理解成运动对象上需跟踪的随时间变化的三维“特征点”的数量与规模。与基于传感器或标记点的运动捕捉系统只跟踪稀疏的三维“特征点”不同，无标记运动捕捉可以利用运动对象的三维模型来实现高精度空间分辨率的运动捕捉目标。本文后续章节提出的三种无标记运动捕捉系统，高精度空间分辨率运动信息的获取均依赖于高精度的运动对象三维模型。然而，由于运动对象的三维模型通常具有较高的维度，在不大幅提高无标记运动捕捉计算代价的前提下，如何合理有效地对三维模型进行降维与建模是捕捉高精度空间分辨率运动信息的首要考量因素。此外，基于传感器或标记点的运动捕捉系统可以直接地标

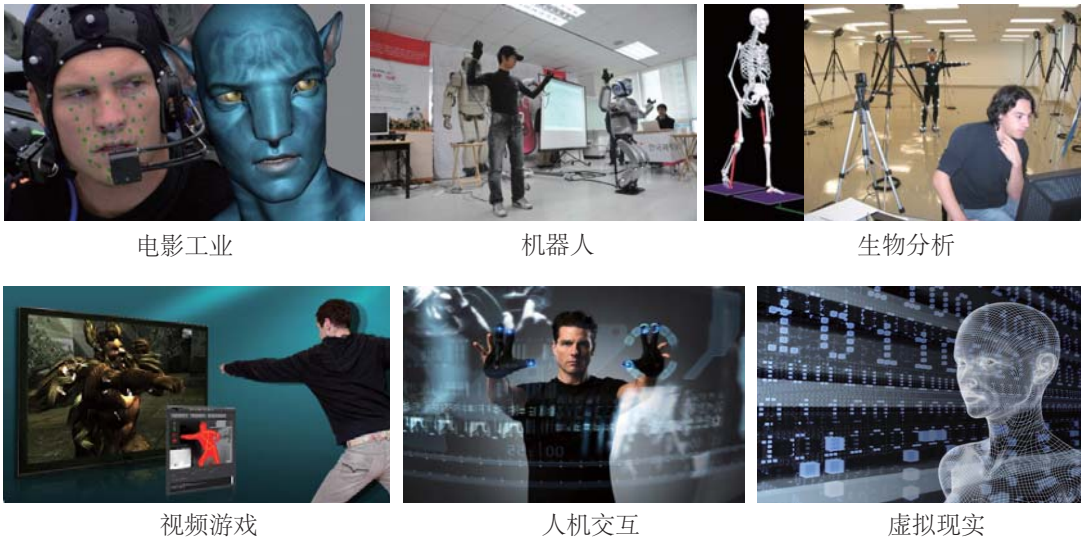


图 1.3　运动捕捉具有广泛的应用前景，图片来自Google™

记或计算出稀疏“特征点”的三维坐标，而由于没有直接的度量传感器或匹配标记点，如何设计合理的运动误差评价指标也是无标记运动捕捉技术需要关注的重点问题。

另外一方面，实现随时间变化的、稳定的运动捕捉是无标记运动捕捉的重要组成部分。由于无标记运动捕捉技术通常采用视觉信号来实现运动重构，而视觉信号很容易受到外部环境的影响，这些影响包括：多相机系统之间的同步与校准误差，运动对象自身或多运动对象之间的遮挡，环境光照的改变等，实现随时间变化的稳定运动捕捉是一件十分困难的任务[4]。除此以外，与时变视觉信号相关的数学算法的可靠性与稳定性也是需要解决的难题[9]。近些年来，随着Kinect 等可直接获取深度的相机不断普及，使用新型传感单元的无标记运动捕捉系统，为无标记运动捕捉系统提供了新的信息来源[10,11]，但在实现稳定的运动捕捉上还需要持续不断的努力。

值得一提的是，在解决上述无标记运动捕捉面临的两大挑战之外，还需要从运动映射的角度着眼考虑。运动映射是将捕捉的运动数据迁移、映射到虚拟角色上，以期望获得逼真的视觉效果，提高虚拟视频制作的生产效率。显然，运动捕捉输出的运动格式与数据会对虚拟运动映射的制作产生很大的影响。例如，以虚拟角色的脸部表情制作为例，如果采用脸部标记点的方式进行虚拟表情合成，后期会需要极多繁琐的操作；而如果采用融合形状（blend shape）的方式制作虚拟角色的脸部表情，则会极大地降低后期调整制作的工作量[12]。在随后的章节中，本文将围绕无标记运动捕捉与运动映射面临的两大挑战展开论述，根据运动过程的复杂程度，提出了三种运动捕捉系统与方法，并结合三类特定的人体运动对象（包括脸、手和全身）进行分析。本文的主要内容如图 1.5所述。

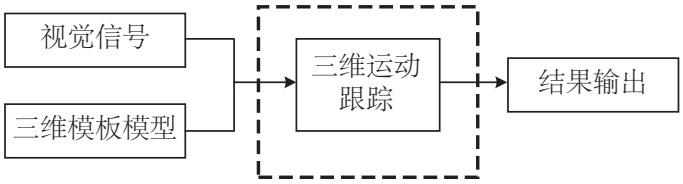


图 1.4　基于生成式方法的无标记运动捕捉的一般流程

1.2　研究现状

无标记运动捕捉经历了数十年的发展，产生了多种不同的技术方法，其中主要分为鉴别式方法和生成式方法[4,5]。鉴别式方法依赖于计算机视觉的相关算法，将运动捕捉转换成特征及物体姿态识别问题。由于受限于计算机视觉在高精度物体识别方面的限制[13]，目前基于鉴别式的无标记运动捕捉还难以获得稳定的、高精度空间分辨率的运动捕捉结果；生成式方法则将运动捕捉转换成运动跟踪问题，使用带有骨架结构的三维模型，通过最小化虚拟生成的三维模型二维投影与观测视觉信号之间的误差，实现运动捕捉的任务。一般而言，基于生成式方法的无标记运动捕捉需要事先得到运动对象的特定或一般三维模型，并对三维模型进行降维处理，得到运动对象三维模型的子空间表达。在运动捕捉过程中，根据采集到的运动对象的视觉信号，计算出运动对象在子空间表达下随时间变化的结果。图 1.4 显示了无标记运动捕捉的一般流程，其中虚线框所示为无标记运动捕捉的核心步骤，需根据运动对象及运动的不同设计特定的三维运动捕捉方法和系统。基于生成式方法的无标记运动捕捉，可以获得稳定、高精度空间分辨率的运动捕捉结果，成为无标记运动捕捉的主流做法[9]。本文的研究工作正是基于生成式方法的框架，并有益地结合了鉴别式方法的诸多特征，提出了三种无标记运动捕捉系统，实现了不同场景复杂度下、不同运动对象的无标记运动捕捉任务。

根据采集运动对象的视觉信号来源，无标记运动捕捉的输入信号源可分为多视点视频信号和单相机视频信号两类。由于多相机系统布置十分繁琐，使用单相机系统来采集视觉信号是将无标定运动捕捉实用化的一个重要方向。

传统的单相机系统只能得到二维的视觉信号，在投影成像的过程中丢失了深度维（z）信息，因此从单相机系统获得的二维视觉信号源中恢复三维运动是一个非常欠定的问题。一种思路是，首先恢复出二维视频对应的深度信息，然后利用恢复的深度信息辅助于三维运动捕捉。但从二维视频中恢复出深度信息亦是一个极其欠定的计算机视觉问题[13]。目前，关于单相机视频的深度恢复问题学术上已经做了大量的探索，主流的两种方法是半自动恢复和全自动恢复[14]。半自动恢复深度信息需要人工对一些关键帧（key frame）进行深度赋值，然后通过传播、插值等方法获取整段视频的深度信息。全自动恢复深度信息则避免了人工干预，根

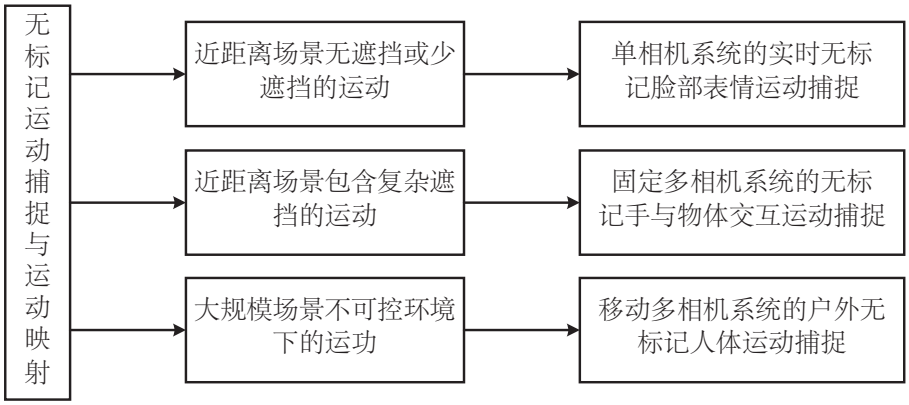


图 1.5　全文主要内容及结构示意图

据输入的二维彩色视频信号，计算机能够自动地输出深度信息。这种深度恢复的方法最终得出的深度精度是相当受限的，为了获得更好的全自动深度恢复的精度，一个有益的尝试是基于大量的数据库来指导计算机进行自动地深度恢复。

值得注意的是，近年来涌现出的新型单相机系统可以直接获得深度信息，例如微软的Kinect和华硕的Xtion。虽然这些设备获取的深度信息在精度与稳定性上还有很大的提升空间，但将这些新型单相机系统与无标记运动捕捉结合已经受到研究人员越来越多的关注。此外，这些新型单相机系统已经可以完美地达到基于传感器或标记点运动捕捉的实时性要求，因此将会具有极强的生命力。

根据采集运动的复杂程度来划分，无标记运动捕捉可分为近距离场景无遮挡或少遮挡运动捕捉，近距离场景复杂遮挡运动捕捉，以及大规模场景不可控环境运动捕捉。针对这三类运动捕捉问题，本文分别提出了单相机系统的实时脸部表情运动捕捉，固定多相机系统的手与物体交互运动捕捉，以及移动多相机系统的人体运动捕捉，如图1.5所示。

1.2.1　脸部表情运动捕捉

脸部表情能够直观地反映人类的喜怒哀乐，是人类交流的重要方式。由于诸多虚拟角色视频中都需要展现合理的脸部表情运动，因此与脸部表情相关的运动捕捉在电影工业中有着巨大的应用价值。不仅如此，脸部表情的运动捕捉能够为人类的信息交流互动提供新的形式，很容易受到普通用户的青睐，在电脑游戏、社交网络、电视、培训、客户支持乃至其他形式的在线互动应用中都会占有一席之地。

事实上，脸部表情运动捕捉一直受到学术圈和工业圈的持续关注，有关脸部表情运动捕捉的详细综述可以在文献[15] 中找到。而这项技术能否被消费级市场广泛接受的两大关键要素在于性能和易用性：性能指的是脸部表情运动捕捉系统

逼真再现特定用户表情的精度与能力；易用性的含义不仅仅包括系统的操作简单、方便与否，也包括系统采用的硬件价格是否低廉，特别地，还包括系统能否实时地再现用户的脸部表情运动。

一般来讲，高精度的无标记脸部表情运动捕捉，往往需要依赖于复杂硬件架构的运动捕捉系统，如结构光系统[16,17]，多视点相机系统[18–21]。但这些系统的计算复杂度通常较高，无法达到实时性的要求，并且以上提到的所有系统都需要专业的硬件设备以及专业人员对硬件系统进行校准。这些条件制约了脸部表情运动捕捉在消费级市场的发展潜力，也即无法同时满足：低廉的硬件设备、实时的运动捕捉以及无需任何形式的预处理的目标。

为了达到脸部表情运动捕捉的实时性要求，并且考虑到脸部表情运动通常可以在无遮挡或少遮挡的环境下进行，研究人员尝试使用单彩色相机的系统[22,23] 来实现脸部表情的运动捕捉任务。一般而言，使用单彩色相机的系统进行脸部的表情运动捕捉，需要事先建立一个较为完备的二维图像与三维脸部表情的映射数据库，然后通过特征查询的方法来捕捉脸部的表情运动。事实上，由于传统单彩色相机系统在采集过程中丢失了深度维信息，属于信息欠定问题，往往无法得到高精度的运动捕捉结果。近些年来，随着主动测量深度的新型单相机系统的不断普及，如前文介绍的Kinect 相机，使用模板三维脸[19,21]，或者建立特定用户动态表情模型（Dynamic Expression Model, DEM）[6]的实时无标记脸部表情运动捕捉技术，可大幅提高表情运动捕捉的精度与稳定性，成为实时无标记脸部表情运动捕捉技术的发展趋势。与单纯依赖彩色相机提供的颜色信息不同，Tadas 等人[24] 给出了一种结合深度和颜色信息的方法，可提高表情捕捉的精度。类似地，Weise等人[6]融合二维和三维深度信息，并将捕捉问题转化成对齐问题，同样可以获得高精度、实时的脸部表情运动捕捉结果。

基于新型可直接获取深度的单相机系统，尽管无标记脸部表情运动捕捉获得了极大地发展，但需要指出的是，已有的方法仍然存在着诸多的缺陷，最大的问题在于需要繁琐的事先训练。虽然事先扫描特定用户的三维表情模型，并对扫描的三维模型进行预处理可以获得高精度、稳定的运动捕捉结果，但这一过程通常非常耗时（一个特定用户的事先训练时间超过6 小时）。此外，除了耗时这一缺陷之外，事先训练也通常会引入额外的运动捕捉误差。原因在于，训练时通常要求用户模仿样例的脸部表情，但不同用户对样例表情的理解往往不尽相同。例如，训练张嘴的表情，嘴张开角度的大小，对每个特定的用户而言，这都是无法衡量、统一的，这就会要求用户来回往复地不断修正其三维表情模型，而这一过程显然又会耗费大量的时间。因此，降低或去除三维表情模型的获取是新型单相机系统进行无标记脸部表情运动捕捉的重要研究方向。

1.2.2　手与物体交互运动捕捉

在计算机图形学领域，制作超高真实感的手与物体交互运动的虚拟视频序列（例如抓起细长的杯把手、单手旋转魔方）是一项极具挑战性的任务，而基于数据驱动的方法为手与物体交互运动视频的制作提供了可行的技术方案[25]。顾名思义，基于数据驱动的视频制作旨在建立特定类别的交互运动数据库，然后对数据库中已有的运动数据进行插值、整合以及重映射处理，实现制作全新虚拟交互运动视频的目标。显然，建立特定类别的交互运动数据库需要捕捉不同形式的手与物体交互运动。手与物体的交互运动捕捉不仅需要恢复手的骨架运动信息、物体的运动信息，特别地，还需要捕捉两者在交互过程中的细微接触运动信息，这些细微接触运动主要表现为手的指节与物体不能出现互相穿透（penetration）、互相分离（departure）的情形。

无标记手与物体交互运动捕捉，其困难主要体现在，一方面需要处理交互运动中极高的运动复杂度，另外一方面，还需要着眼于存在复杂遮挡（包括自遮挡以及互遮挡的情形）的交互运动求解。事实上，由于手上缺少明显的特征，单纯的手势捕捉已是非常困难的问题。目前，关于手的运动捕捉，主流的做法是事先得到手的三维模型，实现手势的骨架运动跟踪[26,27]。当然，为了减少运动跟踪过程中的歧义性，利用数据库学习出合适的手势姿态分类器，然后使用学习出的分类器将观察到的图像进行分类，恢复手的三维姿态是一种常见的做法[28,29]。此外，也有大量的工作尝试降低运动跟踪中投影图与观测图之间的误差函数影响，设计了距离图、轮廓、光流等不同的特征向量[7,30,31]。值得一提的是，融合标记点的运动捕捉方法，也可以减少单纯依赖视觉运动跟踪方法的歧义性[32,33]。Zhao等人[33]提出了一个同时使用Kinect和标记点的三维手势运动捕捉系统，并阐释了如何结合这两种方法的优势。

针对手与物体的无标记交互运动捕捉，Ballan和他的同事[34]融合了传统的手势跟踪做法，通过学习手指上显著点的鉴别特征，可直接估计手与物体的运动参数。Oikonomidis 等人[35] 则将碰撞检测的思想引入到运动跟踪过程中，可同时恢复手与物体的运动参数。这些方法虽然可以有效地避免交互运动过程中的互相穿透情形，但是由于没有考虑交互运动过程中的动力学特性，无法识别以及去除手与物体互相分离的运动估计噪声。虽然，研究人员一直持续不断地致力于解决手与物体的交互运动捕捉，但目前仍然受到三方面的制约，主要表现在：（1）已有的方法与系统很容易受到遮挡以及手上缺少特征的影响，捕捉的运动往往具有歧义性；（2）目前的方法与系统主要着眼于手的关节运动，而完全忽略了手与物体的交互运动，因此也就忽略掉了手与物体之间的相互作用对求解结果的影

响；（3）到目前为止，还没有方法与系统考虑手与物体在交互运动过程中的动力学特征，由于没有捕捉运动过程中的动力学特性，捕捉的运动通常噪声很大并且不满足物理约束，更为重要的是，捕捉的手与物体的运动无法有效的用于运动映射，因而也就很难用来建立交互运动的数据库。

近些年来，随着动力学仿真在计算机图形学中的快速发展，研究人员可以十分逼真的模拟出近乎真实的交互运动[25,36]，因此，如何将动力学仿真的物理约束模型引入传统的无标记运动捕捉过程中，并且对交互运动过程中的细微接触运动进行有效地表达是一个潜在的研究方向。

1.2.3　人体运动捕捉

捕捉人体的全身运动是无标记运动捕捉的一个经典问题，其目标是恢复出人体的骨架运动信息。早先的无标记人体运动捕捉需要在一个可控的室内环境下，采用大量的固定相机实现[4,5]。为了方便提取运动对象，采集环境往往需要布置绿色的幕布背景[37]。无标记人体的运动捕捉可以使用一个简单的模板骨架[38]，通过最大化变形后的骨架与观测图像之间的一致性，求得骨架的变形参数。针对变形骨架与观测图像一致性的优化求解，使用局部优化[39]，或是全局优化[38,40]，或是两者结合的优化方法[41] 都有相关的研究工作提及。为了提高运动捕捉的精度，研究人员也尝试使用高精度的三维表面模型[42,43]。

事实上，简化采集系统已成为无标记人体运动捕捉的发展趋势。特别地，Elhayek等人[44]讨论了室内环境下无需同步多相机系统的无标记运动捕捉问题。由于该工作在进行运动恢复的过程中仅使用了轮廓信息，且只用局部优化的手段来估计运动，恢复出的三维运动精度非常有限。基于小范围内全局光照恒定的假设，Wu等人[8]在一个自然场景的室内环境下，使用手持双目相机设备实现了高精度的运动捕捉，但是恒定全局光照的假设使得他们的系统只能胜任很小范围内的无标记运动捕捉。值得一提的是，Wei等人[45]还讨论过交互式的单目视频的无标记运动捕捉问题。

随着消费级深度相机的逐渐普及，如前文提到的Kinect相机，将深度信息与传统的多视点视频信息相结合，受到学术圈越来越多的关注。Ganapathi研究组[10]和Shotton研究组[46]分别开发出一套可在室内进行实时无标记人体运动捕捉的系统，显然此类系统可以促进人机交互及体感游戏的繁荣，大大拓宽了无标记人体运动捕捉系统的应用范围。最近，也有相当一部分研究工作[11,47,48] 尝试借助于深度相机，来提高传统的无标记人体运动捕捉精度及稳定性。

值得指出的是，很多运动需要在大范围场景、不可控背景环境下才能实现，如足球运动、滑雪运动等，因此实现户外的人体运动捕捉也是无标记人体运动捕

捉的一个发展趋势。显然，与室内、可控背景环境下的无标记人体运动捕捉相比，户外、不可控环境的无标记运动捕捉难度更大，主要表现在采集环境复杂多变和捕捉对象与相机会同时发生运动。当然，针对户外场景下的人体运动捕捉问题，也有研究人员做过相应的探索。Hasler等人[49] 尝试使用手持相机在户外进行无标记运动捕捉，但他们的系统仅能在很小的场景范围内运动。为此，研究人员考虑过在运动对象的身上绑上一定数量的传感器[50,51]，以期望实现大规模场景的无标记人体运动捕捉。显然，以上提到的诸多方法均降低了运动对象的灵活度，而这与无标记运动捕捉的内在要求是相背离的。

1.3 主要研究内容

围绕无标记运动捕捉在空间分辨率和时间稳定性上的挑战，本文分别研究了三种不同运动环境的无标记运动捕捉问题，主要研究内容和研究目标分为三个部分，分别涉及：实时脸部表情运动捕捉，手与物体的交互运动捕捉以及室外人体运动捕捉，如图1.5所示。

针对上述无标记运动捕捉与运动映射研究中存在的难点问题，本文主要进行了如下研究：

- **实时脸部表情运动捕捉** 在第2 章中，针对脸部表情运动捕捉精度与易用性的突出矛盾，考虑使用单相机系统进行脸部表情的运动捕捉，由于传统相机无法获取深度的固有缺陷，建立单张彩色图像与深度图映射的非线性参数模型，提出基于该参数模型的单张彩色图像的深度估计方法，解决单张彩色图像估计深度信息求解极其欠定的理论问题。与现有的基于模型的单张图像估计深度信息的方法相比，构造图像类别关联的模型参数，通过数据库学习出类别相关的模型参数，提高了参数模型的适应性。当然，为了提高运动捕捉的精度，本文提出基于新型深度传感单相机系统的实时脸部表情运动捕捉方法。该方法相比于当前普遍需要事先进行繁琐的训练或校准的方法，核心思想是建立脸部的动态表情模型。动态表情模型的建立依赖于脸部表情的三维数据库以及变形映射技术（deformation transfer）。该方法提出变形映射的线性化结构，优化捕捉对象基本脸部表情的同时，搜索捕捉对象基本表情的组合系数，避免了任何其他形式的预处理操作，真正实现了实时脸部表情运动捕捉。实验表明，该方法获得的脸部表情运动捕捉精度与当前主流的方法相当，但显著简化了捕捉对象的工作流程。
- **手与物体交互运动捕捉** 在第3章中，针对手与物体的交互运动中细微的接触运动捕捉问题，提出了符合物理约束的复合运动控制器的模型，对手的关

节运动，物体的运动，以及它们之间相互接触的细微交互运动同时进行了捕捉，解决了交互运动捕捉中细微接触运动的捕捉问题。该方法的核心思想是利用运动控制器模拟生成大量的虚拟运动，设计了一套与输入视频匹配的运动真实性评价函数，搜索符合输入视频的最佳运动控制器参数。与以往的交互运动捕捉方法相比，该方法得到的最佳运动控制器参数可简化交互运动的运动映射，其实用性和有效性在具有不同物理属性的物体上得到进一步的验证。

- **户外人体运动捕捉** 在第4章中，针对大规模室外场景、不可控环境下的无标记人体运动捕捉问题，提出使用多个可移动手持相机来追踪运动对象拍摄的运动捕捉方法，解决了室外场景下大规模运动的无标记运动捕捉问题。该方法甚至可以在两个手机摄像头的采集环境下有效地实现人体的无标记运动捕捉。为了实现大规模运动的无标记运动捕捉，该方法首先恢复出可移动手持相机的随时间变化的空间位置关系，针对宽基线的可移动相机提出了一种新的运动对象稠密点云的计算方法，利用输入的多视点视频信号以及计算的点云恢复出骨架运动信息。为了提高骨架运动信息恢复随时间变化的稳定性，与传统的方法相比，该方法引入了一种新的稀疏性约束，并且提出了与视点相关的动态纹理模型。不同环境下、不同运动对象的大规模户外无标记运动捕捉的测试结果均体现了本文提出方法的有效性。

最后，第5章对全文做了总结，并对未来的工作进行了展望。

第 2 章　单相机系统的实时无标记脸部表情运动捕捉

本章着重讨论近距离场景无遮挡或少遮挡的运动捕捉问题，并以脸部表情运动捕捉作为具体的研究对象。一般而言，这类运动捕捉通常采用单相机系统来实现视觉信号的采集任务，其主要原因在于单相机系统具有极强的易用性以及较低的搭建成本。然而，正如引言中介绍，传统的单相机系统只能得到二维的视觉信号，在投影成像的过程中丢失了深度信息，单相机系统的运动捕捉在运动捕捉精度与易用性之间存在着固有矛盾。一种思路是，辅助使用深度信息实现高精度的三维运动捕捉。然而，从二维图像中恢复出深度信息是计算机视觉领域的经典问题。本章前半部分将重点阐述单相机系统的深度估计问题，建立了单张彩色图像与深度图映射的非线性参数模型，提出了基于该参数模型的单张彩色图像的深度估计方法。值得一提的是，随着电子工业技术的发展，将深度估计的方法用硬件集成，并实现可直接输出深度的新型单相机系统已经出现，如何高效地利用新型单相机系统输出的深度信息，降低运动捕捉的计算时间复杂度是需要解决的难题。本章后半部分将针对脸部表情的运动捕捉问题，使用可直接输出深度的新型单相机系统，提出脸部的自适应动态表情模型，降低求解空间的维度，解决了新型单相机系统计算时间复杂度与运动捕捉精度的矛盾。

2.1　单相机系统的深度估计

从传统单相机系统采集的二维视觉信号中恢复深度信息属于病态（ill-posed）问题。针对这一病态问题，当前国内外的研究主要有两种解决思路，一种是采用主动感知深度的传感单元，另外一种是引入与视觉信号内容相关的先验（prior）知识。在这一小节中，我们将阐述基于上述两种思路的单相机系统深度估计方法的基本原理，并且针对全自动单相机系统的深度估计，通过分析彩色图像与深度图的统计特性，提出彩色图像与深度图之间的非线性模型。进一步，本节引入数据库的知识实现非线性模型的参数估计，获得了很好的计算效率与使用灵活性。

2.1.1　相关研究

2.1.1.1　基于飞行时间的深度估计

基于飞行时间（time of flight）原理的单相机深度估计旨在让单相机系统首先激发出脉冲光波，这些脉冲光波在与三维空间物体相互作用之后，部分反射的脉

冲光波可被相机镜头感知，系统通过计算前后脉冲光波的相位差，可实现三维空间深度信息的估计任务。这是一种主动式的深度感知方法，在雷达测距中已经获得了广泛应用，基于飞行时间原理的光学测距是目前可行的最为精确的单相机系统深度估计方法。然而，由于受到反射信号的传播时间误差及脉冲光波的调制频率限制，基于飞行时间原理的单相机深度估计目前仅在较小的场景范围（测距范围在1 ~ 5*m* 左右）获得较高的深度采集精度。已知的单相机系统的高精度深度估计产品，如Mesa SwissRanger SR4000和微软的Kinect二代均使用这种原理来估计深度。

2.1.1.2　基于结构光的深度估计

结构光估计深度是将一些固定图案投影到三维空间，再使用相机拍摄投影后的图案来实现深度估计的任务。由于相机拍摄的图案受空间的三维曲面结构影响，与原始图案相比会发生不同程度的变形，因此可以利用拍摄图案与实际投影图案之间的变形量估计三维空间的曲面结构，进而获得三维空间的深度信息。一般来讲，结构光估计深度的固定图案由点、线及面阵构成，并且为了不影响传统的拍摄，这些固定图案往往被调制在红外波段。显然，结构光估计深度也是一种采用主动感知传感器的方法，其是一种便捷、高效的三维深度测量方法，与前文所述的基于飞行时间原理的深度估计方法类似，该方法也仅能在较小的场景范围内获得不错的估计结果。商业化的产品中，微软的Kinect一代是基于结构光原理实现深度估计的目标。

2.1.1.3　基于人机交互的半自动深度估计

从二维视觉信号中恢复出深度信息是一个极其欠定的问题，一种最简单有效的方法是通过人工操作，引入人类感知深度的先验知识。可以预见的是，人工操作的工作量直接决定了深度估计的最终精度。当然，实际的人机交互深度估计不可能、也并不需要对二维视觉信号的所有像素都进行人为赋值。通常的做法是让用户首先标记出特定像素或区域的深度值，然后辅以相应的传播算法来自动估计其他像素或区域的深度值，这也是人机交互的深度估计方法被称为半自动深度估计的主要原因。一般来讲，半自动深度估计出来的深度平滑，对于纹理不丰富的二维视觉信号而言，具有较好的估计精度。实际上，针对视觉信号内容的不同，半自动深度估计的算法性能也千差万别。Guttmann 等人通过考虑二维视觉信号的局部运动特性[52]，设计与图像块内容匹配的分类器来自动计算像素的深度值。Cao 提出了一种基于过分割（over-segmentation）的算法来解决相机变焦或者镜头拉伸时深度难以自动估计的问题[14]。

2.1.1.4　基于数据库的全自动深度估计

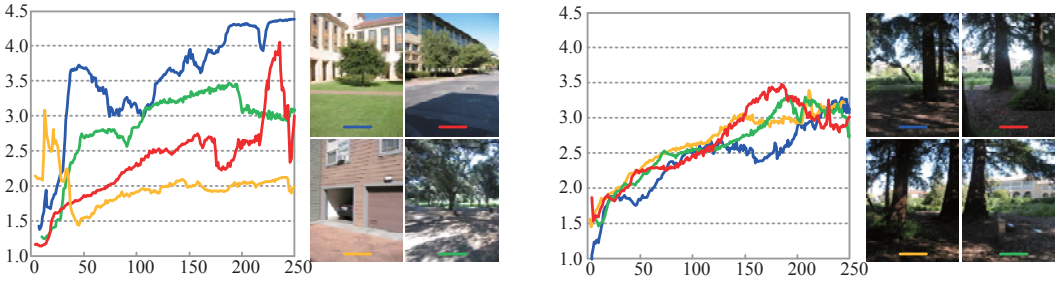
另外一类与视觉内容相关的先验知识则来源于数据库。随着可直接获取深度的相机逐渐普及，通过建立彩色图像与深度图的数据库，寻找它们之间的对应关系，建立基于数据库的全自动深度估计方法，成为近年来研究人员关注的热点。

当前，基于数据库的全自动单张彩色图像的深度估计方法可分为两类：参数化模型方法和深度迁移方法。参数化模型的方法试图通过设计合理的参数式数学模型建立彩色图像与深度图像之间的关联。Saxena和他的同事[53]在图像像素层面构建了一个复杂的多尺度马尔可夫随机场（Markov Random Fields, MRF），通过借鉴机器学习的思想，试图描述彩色图像与深度信息的一致性。他们首先从数据库中学习出与图像内容无关的模型统一参数，然后用该参数估计单张彩色图像的绝对深度值。Hoiem等人[54]则从语义层面来描述待估计的深度信息，所谓语义层面即是将深度信息量化成具有语境意义的离散区域。他们的做法是首先将彩色图像划分成诸如天空、地面或者垂直的区域，然后将垂直的区域区别对待，构建一个简单三维模型来估计区域相关的深度信息；Delage等人[55]则充分挖掘场景中的直线信息，重建出墙壁、天花板以及地板的深度信息；与以上两类做法不同，Liu等人[56]则将简单场景下的简单语义标签扩展到具有复杂结构的场景环境中。由于语义标签可以避免深度估计的遮挡及交叠问题，Saxena等人[57]于是也将他们在像素级别的数学模型，通过过分割的操作，拓展到语义层面。

近些年来，随着彩色图像与深度图的数据规模越来越大，采用直接搜索的方法成为基于数据库的全自动深度估计的一个趋势，此类深度估计方法也被称为深度迁移。其中，最具代表性的工作由Karsch在MIT做出[58]。针对待估计深度的单张彩色图像，他们首先在数据库中查找与待估计图像类似的多幅候选彩色图像，通过建立彩色图像之间的变换关系，迁移并融合数据库中候选图像的深度图，获得彩色图像的深度估计结果，显然，他们的方法得益于大规模彩色与深度图像的数据库。

2.1.2　深度估计的非线性参数模型

针对传统单相机系统的深度估计问题，引入数据库的先验知识，采用参数模型的深度估计方法可以获得很好的计算效率以及使用灵活性，受到广泛关注。目前，已有的参数模型均是预先定义复杂的特征来描述彩色图像与深度图之间的关联，通过设定相应的特征参数，建立这些复杂特征的参数化数学模型。显然，参数模型深度估计方法的重要任务之一即是寻找足够好的特征来表达颜色、深度之间的联系。虽然多尺度特征和空间一致性特征[53] 已经被证实是非常有效的描述



(a) 不同场景亮度与深度统计关系　　(b) 同一场景亮度与深度统计关系

图 2.1　图像亮度和深度值的统计关系

彩色图像与深度图关联的两类重要特征，但是否存在更好的特征是摆在参数模型深度估计方法面前的基本难题。值得指出的是，如何评价参数模型深度估计中预先定义的各类特征，以及它们对深度估计结果的影响也是一件非常困难的事情，正因如此，设定不同特征之间的权重关系也就举步维艰。本节提出了一种可自动学习最优特征的非线性参数模型，通过将彩色图像描述成图像块的集合，建立了更为灵活的图像块与深度值之间的映射关系。此外，特征的基本要素—— 滤波器在学习的过程中自动优化，以便自适应地选择合适的图像、深度特征。实验表明，本节提出的非线性参数模型可以很好地学习出彩色图像块与深度关联的特征，提高了参数模型深度估计的精度。

2.1.2.1　提出的模型

直观来讲，不同场景类别的彩色图及其对应深度图之间的统计关系是非常无序且无章可循的。尽管如此，我们能否换个角度思考：相似场景（例如，户外，森林，海滩，室内等）的图像、深度之间是否具有一致的统计规律？事实上，场景近似的图像、深度具有比较相近的统计相似性，如图2.1所示，其中，横轴为彩色图像量化后的整数亮度，其范围为[1...255]；纵轴为对数深度值（log），计算方法如下：对图像中量化的同一整数亮度，将所有出现的log深度值进行求和平均。这表明建立近似场景类型图像的彩色、深度参数模型是合理可行的，同时也指导我们参数模型的建立需对图像进行场景分类，学习出特定场景相关的参数并估计其深度信息。

给定特定类别的彩色图像I，及其对应的深度图像D，我们的目标是用一组参数集来描述它们之间的关联。与之前参数模型的方法类似[53]，我们将彩色图像看成是由图像块组成的集合，图像块的大小为15×15像素。为了避免过拟合，我们从图像I中采样出$\{x_1, x_2, ..., x_p\}$个图像块，并且在其对应的深度图上取得$\{d_1, d_2, ..., d_p\}$个深度数值，这里p是采样的样本数目。为了叙述方便，这里用列向量$\mathbf{d}$表示与所有采样图像块对应的深度值，也就是说，$\mathbf{d} = [d_1, d_2, ..., d_p]^T$。

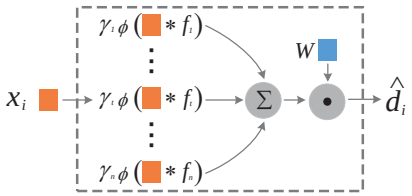


图 2.2　彩色图像块x_i的映射深度$\widehat{d_i}$计算流程

通过计算彩色图像块的映射深度值$\widehat{\mathbf{d}}$与其对应的真实深度值$\mathbf{d}$之间的平方误差和，可以建立彩色图像块与深度值之间的关联。彩色图像块的映射深度值$\widehat{\mathbf{d}}$的计算方法如下：假设矩阵$F = [f_1, f_2, ..., f_n]$，其中矩阵的每一列均是滤波器，n为滤波器的数目；对于每一个彩色图像块x_i，首先将其与n个滤波器进行卷积运算，然后利用核函数$\phi(\cdot)$将卷积后的结果映射到另外一个“空间”，针对每一个映射后的图像块使用γ_i进行加权求和得到一个统一的“计算图像块”，其中，列向量$\boldsymbol{\gamma} = [\gamma_1, \gamma_2, ..., \gamma_n]^T$表达了不同滤波器对应卷积结果的权重关系。最后，将空间权重矩阵W与之前求和得到的统一“计算图像块”进行内积运算，就可以得到彩色图像块对应的映射深度值$\widehat{d_i}$，这一计算过程如图2.2 所示。

上述的计算过程可以用下面的式子来表达

$$E = \sum_{i=1}^{p} |tr(W^T \sum_{j=1}^{n} \gamma_j \phi(x_i * f_j)) - d_i|^2. \tag{2-1}$$

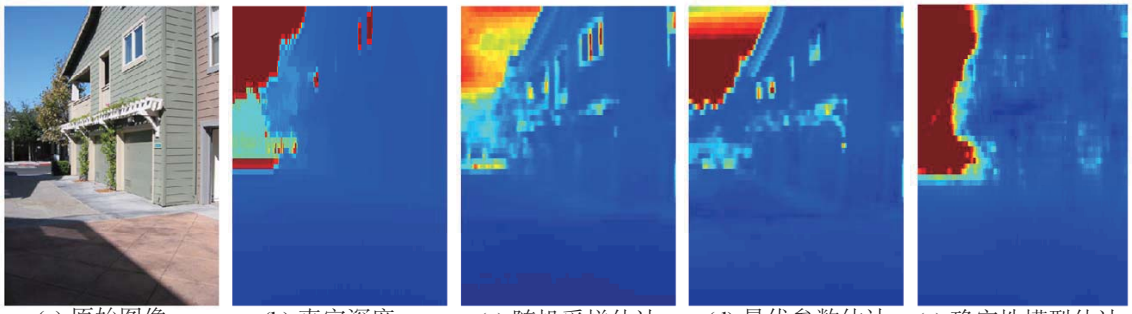
在式（2-1）中，E刻画了彩色图像的映射深度与真实深度的误差，$W, F, \boldsymbol{\gamma}$ 是所提出模型的参数，$tr(\cdot)$用来计算矩阵的迹。为了推导方便，式（2-1）也可以描述成

$$E = \sum_{i=1}^{p} |\mathbf{w}^T \phi(X_i F)\boldsymbol{\gamma} - d_i|^2, \tag{2-2}$$

其中，X_i是由x_i变形得到的矩阵表达，X_i的每一行与滤波器的大小相等；列向量$\mathbf{w}$是把权重矩阵W的所有元素进行串联重排得到的。

在下一小节中，我们将会详细介绍模型参数的估计问题。这里，我们对式（2-1）中的模型参数做进一步地分析，并且针对所提出模型的非线性映射函数$\phi(\cdot)$的具体形式，做出详细地阐述。

首先需要关注的模型参数是滤波器F，其主要用来抽取彩色图像中的纹理信息，本质上可计算出彩色图像的不同频率响应信息。与之伴随的参数$\boldsymbol{\gamma}$用来描述不同滤波器抽取频率信息的权重关系，目的在于防止单一滤波器的频率响应起决定性作用。需要说明的是，滤波器F可以通过主成分分析（PCA）或者独立成分分析（ICA）初始化得到，在这之后可以有效地学习出与类别图像相一致的滤波



(a) 原始图像　(b) 真实深度　(c) 随机采样估计　(d) 最优参数估计　(e) 确定性模型估计

图 2.3　所提模型的有效性验证

器参数。

另外一个值得关注的是滤波器F的大小。一般来讲，小的滤波器（如3×3）可以加速参数估计及深度估计的过程，但不可避免地仅考虑很小范围的邻域信息；大的滤波器（如7×7）可表达比较大范围的邻域信息，但增加了计算时间复杂度。在本文提出的模型中，由于权重矩阵W的引入已经考虑到图像块的空间邻域关系，这就使得我们可以用比较小的滤波器来实现深度估计的任务。

核函数$\phi(\cdot)$的作用是将卷积后的图像块映射到另外一个空间中，使得映射后的结果可以与深度值进行匹配。事实上，我们发现，核函数的具体形式对深度估计的结果有着很重要的影响，通过测试已有的、大量不同的核函数，我们发现一种简单并且有效的核函数形式是：$\phi(x) = \log(1 + x^2)$。相关细节会在本节后面的实验部分进行阐述。

2.1.2.2　模型参数估计

由于众多的模型参数以及非线性核函数的出现，本文所提出的模型参数估计并不是一件十分容易的事。通常的做法是，在估计其中一个模型参数时，固定其他的模型参数，然后迭代地进行优化以致最终收敛[59]。

为了方便计算式（2-2）的梯度，我们将其重新改写成如下两种矩阵形式：

$$E = \|M\phi(XF)\boldsymbol{\gamma} - \mathbf{d}\|_2^2, \tag{2-3}$$

以及

$$E = \|\Gamma\phi(F^T\widehat{X})\mathbf{w} - \mathbf{d}\|_2^2. \tag{2-4}$$

在式（2-3）中，X是将所有的X_i串接后的矩阵表达式，M的每一行都是$\mathbf{w}^T$。类似地，在式（2-4）中，$\widehat{X}$是将所有的X_i^T串接后的矩阵表达式，Γ的每一行都是$\boldsymbol{\gamma}^T$。值得注意的是，以上两式分别是关于$\boldsymbol{\gamma}$和$\mathbf{w}$的最小二乘问题，因而可以很容易地得到更新$\boldsymbol{\gamma}$和$\mathbf{w}$的闭式解。

下面，我们来着重阐述滤波器的估计问题。根据式（2-1）以及式（2-3），我们可以得到关于滤波器f_i的偏导数

$$\frac{\partial E}{\partial f_{i,t}} = \gamma_i X^T J\left(X f_{i,t}\right) M^T \left(\sum_{j=1}^{n} \gamma_j M \phi\left(X f_j\right) - \mathbf{d} \right). \tag{2-5}$$

其中，$J(Xf_{i,t})$是向量$\phi(Xf_{i,t})$雅克比（Jacobian）矩阵，该矩阵由梯度向量$\phi'(Xf_{i,t})$张成。

由于核函数$\phi(\cdot)$具有非线性特性，我们将$\phi(Xf_i)$在$f_{i,t}$处做泰勒展开（Taylor Expansion）处理，可以得到

$$\phi\left(X f_i\right) = \phi\left(X f_{i,t}\right) + J\left(X f_{i,t}\right) X\left(f_i - f_{i,t}\right). \tag{2-6}$$

这里，我们用$L_{i,t} = MJ(Xf_{i,t})X$来描述三个矩阵的乘积，将其带入到式（2-5）中并且结合式（2-6），于是我们可以得到提出的模型对滤波器$f_{i,t}$的偏导数

$$\frac{\partial E}{\partial f_{i,t}} = \gamma_i L_{i,t}^T \left[\begin{array}{l} \gamma_i L_{i,t} f_i + \sum\limits_{j=1, j\neq i}^{n} \gamma_j L_{i,t} f_j - \mathbf{d} \\ + \sum\limits_{j=1}^{n} \left(\gamma_j M \phi\left(X f_{j,t}\right) - \gamma_j L_{j,t} f_{j,t} \right) \end{array} \right]. \tag{2-7}$$

为了方便起见，我们分别定义$G_{i,t}$和$K_{i,t}$为

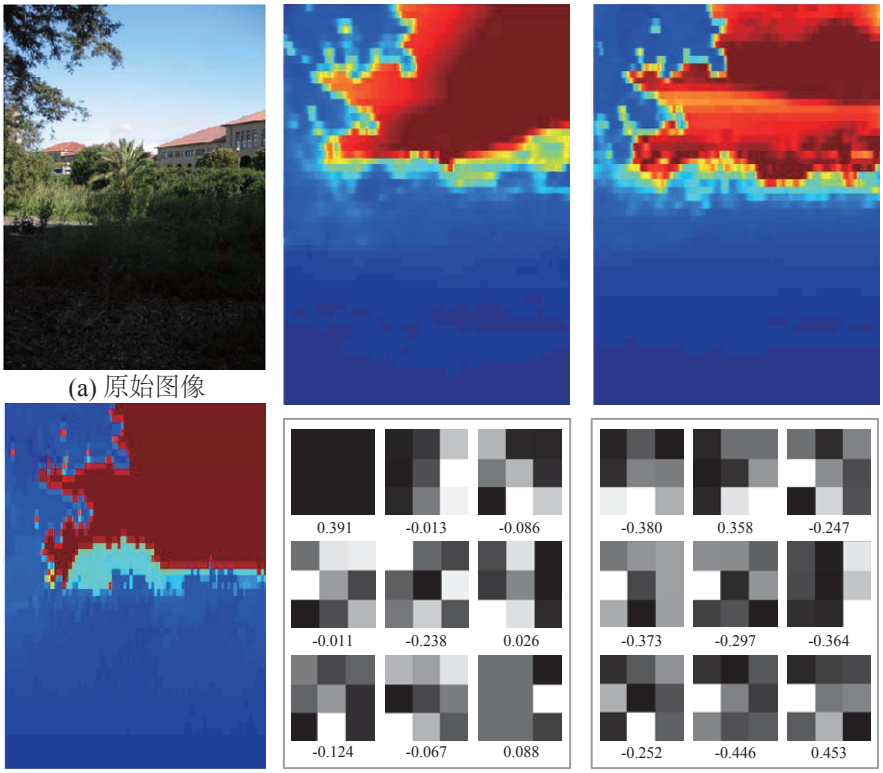
$$\begin{aligned} G_{i,t} &\triangleq \gamma_i^2 L_{i,t}^T L_{i,t}, \\ K_{i,t} &\triangleq \gamma_i L_{i,t}^T \left(\begin{array}{c} \sum\limits_{j=1, j\neq i}^{n} \gamma_j L_{i,t} f_j + \\ \sum\limits_{j=1}^{n} \gamma_j M \phi\left(X f_{j,t}\right) \\ - \sum\limits_{j=1}^{n} \gamma_j L_{j,t} f_{j,t} \end{array} \right) - \gamma_i L_{i,t}^T \mathbf{d}. \end{aligned} \tag{2-8}$$

值得一提的是，当式（2-1）达到最小时，应当保证所有滤波器的梯度为$\mathbf{0}$，也就是说所有滤波器$F = [f_1, ..., f_i, ..., f_n]$都必须满足等式$G_{i,t} f_i + K_{i,t} = \mathbf{0}$，这样我们可以很容易得到最优滤波器的闭式解。然而，实验中我们发现，直接求解线性方程组很难得到稳定的闭式解。与此同时，在求解最优滤波器的过程中，我们使用基于热启动（warm-started）的梯度下降方法来求解最优的滤波器[60]。

2.1.3　实验结果与分析

为了验证所提出模型的有效性①，我们做了大量的实验，这些实验包括：子空间优化的滤波器对深度估计精度的影响，模型参数估计时深度值的采样策略，

① 实验用的所有代码均提供在http://media.au.tsinghua.edu.cn/ygwang/spl2013depth.jsp



(a) 原始图像　(b) 真实深度　(c) 固定滤波器结果　(d) 最优滤波器结果

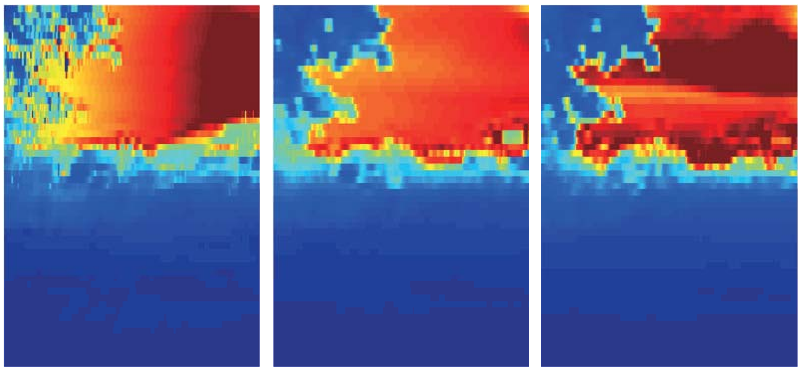
图 2.4　滤波器优化的有效性验证

以及不同核函数的选择对最终结果的影响。需要说明的是，本节实验使用的所有图片均来自文献[53]。

我们使用两种方法来进行深度估计精度的量化评价。假设估计的深度为$\widehat{d}$，真实的深度为d，第一种深度估计精度的量化评价方法是计算**相对误差（RE）**，其计算公式为$1/p * \sum_i |\widehat{d}(i) - d(i)|/d(i)$，另外一种深度估计精度的量化指标为**对数误差（LE）**，计算公式为$1/p * \sum_i \log_{10}(\widehat{d}(i)/d(i))$，上述两种计算公式中的$p$均为像素的数目。

2.1.3.1　模型有效性验证

为了验证所提出模型的有效性，我们随机采样了2000个彩色图像块以及与其对应的深度值，这个数目大约占单张图像的图像块集合的0.52‰（原始彩色图像的大小为2272×1704，并且重叠图像块的大小为15×15）。首先，我们利用原始采样图像及其深度信息来学习模型的参数，并用估计出的模型参数进行深度估计；其次，我们使用与待估计图像相近似的彩色图像及其深度图来学习模型的参数，然后用学习出的模型参数来估计深度，我们将这样一种做法称之为**参数传递**，与以前的基于模型的深度估计方法相比[53]，我们发现本文所提出的非线性参数模型可以获得更好的深度估计结果。图2.3是实验比较结果，其中，图2.3(a)是原始



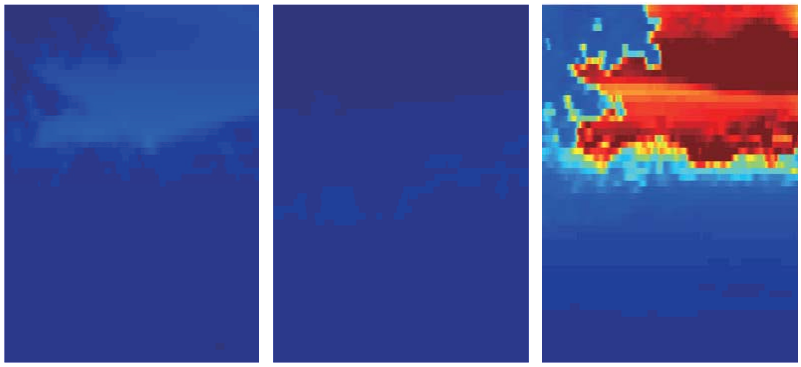
(a) 随机采样　(b) 拒绝采样　(c) 均匀采样

图 2.5　采样策略对深度估计的影响

彩色图像；图2.3(b)是真实的深度图；图2.3(c) 是使用2000 个随机采样的彩色图像块及其深度值估计模型参数，进行深度估计的结果（RE=0.3071, LE=0.1285）；图2.3(d)是使用与原始图像相近似的图像及其深度图来估计模型参数，进行深度估计的结果（RE=0.3059, LE=0.1268）；图2.3(e)是使用确定性模型参数的深度估计结果[53]（RE=0.5242, LE=0.1565）。从该图中我们可以发现：（1）使用参数重建的深度图2.3(c) 与原始深度图2.3(b)并不完全一致，这也就表明原始彩色图像块与深度之间的关系并不能完全用参数来刻画出来；（2）图2.3(c)和图2.3(d)具有相当的深度估计结果，这表明使用场景近似的模型参数完全可以进行单张彩色图像的深度估计任务。

2.1.3.2　模型参数估计的关键步骤验证

- **滤波器优化** 据我们所知，目前已知的基于参数模型的深度估计方法都采用固定的滤波器来实现特征的选取。为了验证本文提出的滤波器优化的必要性，我们比较了滤波器优化的深度估计结果。在我们的实验中，滤波器都是在子空间中进行优化的，也就是说，滤波器被描述成$F = B\widetilde{F}$，这里，B是子空间的基底，其通过对图像块的PCA分析获得；$\widetilde{F}$是滤波器在子空间下的系数，$\widetilde{F}$ 初始化为单位矩阵。图2.4分别展示了进行滤波器优化和不优化的结果，其中，（a）是原始彩色图像；（b）是真实的深度图；（c）的第一行显示了固定滤波器的深度估计结果（RE=0.4914 LE=0.1893），第二行选择显示了一些滤波器及对应的$\boldsymbol{\gamma}$；（d）的第一行显示了滤波器优化的深度估计结果（RE=0.4317 LE=0.1563），第二行选择显示了一些滤波器及对应的$\boldsymbol{\gamma}$。通过图2.4（b）和 2.4（c）的比较，我们发现滤波器优化可以获得更好的深度估计结果。此外，与固定滤波器相匹配的最优$\boldsymbol{\gamma}$ 表明，图像的亮度在估计过程中具有决定性作用。



(a) Φ_1估计结果　(b) Φ_2估计结果　(c) Φ_3估计结果

图 2.6　不同核函数对深度估计的影响

- **参数估计的采样策略** 在参数估计的过程中，我们发现选择图像块及其深度值的不同采样策略来会使参数估计的结果不同，我们测试了三种不同的采样策略：一是从整张图像上随机采样；二是考虑到采样过程中的噪声，拒绝采样那些与亮度统计规律不一致的图像块；三是将图像分成不同的网格，然后在不同的网格中进行均匀采样。图2.5显示了三种不同采样方法估计的参数进行深度估计的结果，其中，（a）是随机采样的深度估计结果（RE=0.8265, LE=0.2311）；（b）是拒绝采样的深度估计结果（RE=0.6844, LE=0.1835）；（c）是网格划分均匀采样的深度估计结果（RE=0.4317 LE=0.1563）。实验表明，为了获得更好的结果，最好使用第三种采样方法进行参数估计。
- **核函数** 针对核函数的选择问题，我们也做了相应的实验分析。我们探索了计算机视觉邻域中广泛使用的核函数形式[13]，选取了三种与统计特性一致的核函数：$\phi_1(x) = \sqrt{|x|}$，$\phi_2(x) = x^2$以及$\phi_3(x) = \log(1 + x^2)$。实验结果表明，$\phi_3$具有最好的深度估计结果。图2.6显示了不同核函数的估计结果，其中，（a）是使用ϕ_1的深度估计结果（RE=1.0843, LE=0.7673），（b）是使用ϕ_2的深度估计结果（RE=1.2658, LE=0.8711），（c）是使用ϕ_3的深度估计结果（RE=0.4317 LE=0.1563）。图2.6表明，使用ϕ_1 和ϕ_2 核函数的模型都不能获得正确的深度估计结果，其中的原因可能是ϕ_1 的梯度对于很小的x敏感，而针对大的x，ϕ_2并不能收敛。

2.2　实时无标记脸部表情运动捕捉

在前一小节中，我们详细讨论了传统单相机系统的深度估计问题，并且提出了一种传统单相机系统估计深度的非线性参数模型。本节，我们将讨论如何使用深度信息进行近距离场景无遮挡或少遮挡的运动捕捉，并以脸部表情运动捕捉作

为具体的研究对象。值得说明的是，为了尽量降低运动捕捉的系统噪声，本节运动捕捉所使用的深度信息来源于可直接获取深度的传感单元。

2.2.1　导言

实时无标记脸部表情运动捕捉，并将捕捉的运动映射到虚拟对象成为近年来研究人员广泛关注的研究点。实时脸部表情运动捕捉的优势在于，一方面可在线地提供映射反馈，便于人们对映射结果进行实时调整；另外一方面，为人类的信息交流互动提供一种新的形式。传统信息交流互动主要通过二维视频的方式进行，这样一种信息交流互动方式对替换用户的外貌有很大的局限性，而实时脸部表情的运动捕捉与映射在电脑游戏、社交网络、电视、培训、客户支持乃至其他形式的在线互动应用中，为消费级市场开启了一扇引人入胜的大门。

事实上，这项技术能否被消费级市场广泛接受的两大关键要素在于性能和易用性：性能指的是脸部表情运动捕捉系统逼真再现特定用户表情的精度与能力；易用性的含义不仅仅包括系统的操作简单、方便与否，也包括系统采用的硬件价格是否低廉。目前，基于标记点、多视点立体匹配以及三维扫描方式的脸部表情运动捕捉技术可以稳定地捕捉高精确的脸部表情运动，技术已经十分成熟，并已在工业生产中获得极大的应用。但为了获得高精度的脸部表情运动捕捉结果，这些系统都需要十分精细的校准、训练，甚至专业人员的辅助，在易用性上存在着很大的缺陷，因而也就无法被消费级市场广泛接受。

采用单相机系统的实时无标记脸部表情运动捕捉，由于在易用性方面可获得极大提高，是潜在的消费级市场产品，受到研究人员持续不断地关注。一般来说，基于二维视频的单相机无标记脸部表情运动捕捉[22,23,61]仅跟踪少量的特征点，不能得到很高的表情捕捉精度；此外，表情捕捉的性能在室内环境下也会受到光照变化的影响。近年来，随着主动测量深度的新型单相机系统的不断普及，使用模板三维脸[19,21]，或者建立特定用户动态表情模型[6]的实时无标记脸部表情运动捕捉技术可大幅提高表情运动捕捉的精度与稳定性，成为实时无标记脸部表情运动捕捉技术的发展趋势。使用动态表情模型的运动捕捉将原来的脸部表情捕捉问题转化成低维空间的非刚性对齐问题，在目前看来可以获得极好的单相机实时无标记脸部表情运动捕捉的性能。

目前，基于动态表情模型的实时无标记脸部表情运动捕捉，存在的主要问题有：（1）特定用户的动态表情模型需要事先在可控环境下采集；（2）在采集时需用户对照着样例进行照做、模仿；（3）为了获得比较满意的捕捉效果，动态表情模型往往还需要熟悉相关知识的专业人员进行处理修正，整个过程十分繁琐。以上所述的三类问题对消费级市场的普通用户而言，均是难以忍受的。

图 2.7　实时无标记脸部表情运动捕捉与运动映射结果

针对目前基于动态表情模型的实时无标记脸部表情运动捕捉存在的主要问题，本文提出了一种自适应动态表情模型，该模型由三部分构成：单位PCA模型、表情映射算子以及变形校正场。此外，本节还将会阐述在无需任何人工帮助的前提下，如何从动态模板表情自适应地优化到特定用户的动态表情? 实际上，随着用户面前的单相机系统采集到足够多的随时间变化的用户脸部数据之后，本节提出的算法就可以将动态模板表情收敛到特定用户的动态表情模型。该算法不仅可以输出特定用户的动态表情模型，而且还实现了特定用户的实时无标记脸部表情运动捕捉与映射，图2.7显示了运动捕捉及运动映射后的结果。

2.2.2　相关研究

在计算机图形学领域，脸部表情的运动捕捉技术是虚拟角色动画的基本技术，因此一直受到研究人员的关注与讨论，本小节将对相关的研究工作做一简要的回顾，有关脸部表情运动捕捉的详细综述可以在文献[15]中找到。

基于标记点的运动捕捉系统广泛地用于实时脸部表情运动捕捉[62,63]，虽然标记点可以大大简化跟踪的过程，但不可避免地，这类系统无法实现高精度的表情捕捉；最近，基于三维扫描的脸部表情运动捕捉系统（如，结构光系统[16,17]，多视点相机系统[18–21]）被开发出来，其虽然可以实现高精度的脸部表情运动捕捉，但这些系统的计算复杂度通常较高，无法达到实时性的要求。因而有研究人员结合以上两种系统的优势[64–66]，开发出基于标记点和三维扫描联合的脸部表情运动捕捉系统，在满足实时性要求的同时，不丢失脸部表情的空间分辨率。值得注意的是，以上提到的所有系统都需要专业的硬件设备以及专业人员对硬件系统进行校准，这与消费级市场的内在要求是相背离的，也即无法同时满足低廉的硬件设备、实时的运动捕捉以及无需任何形式的预处理的目标。

当然，为了达到脸部表情运动捕捉的实时性要求，研究人员也尝试了使用单彩色相机的系统[22,23,61]。由于单彩色相机系统在采集过程中丢失了深度维信息，属于信息欠定问题，脸部表情捕捉的精度往往达不到运动映射的要求。而

最近几年涌现出的可直接获取深度的新型单相机系统，如微软的Kinect和华硕的Xtion，使得实现上述的目标成为可能。与单纯依赖彩色相机提供的颜色信息不同，Tadas等人[24] 给出了一种结合深度和颜色信息的方法，可提高表情捕捉的精度。类似地，在文献[6]中，融合二维和三维深度信息，并将捕捉问题转化成对齐问题，同样可以获得高精度、实时的脸部表情运动捕捉结果。

需要指出的是，上述基于单相机系统的脸部表情运动方法，最大的问题在于需要繁琐的事先训练。虽然事先扫描特定用户的三维表情模型，并对扫描的三维模型进行预处理可以获得高精度、稳定的运动捕捉结果，但这一过程通常非常耗时（一个特定用户的事先训练时间超过6小时）。此外，除了耗时这一缺陷之外，事先训练也通常会引入额外的运动捕捉误差。原因在于，训练时通常要求用户模仿样例的脸部表情，但不同用户对样例表情的理解往往不尽相同。例如，训练张嘴的表情，嘴张开角度的大小，对每个特定的用户而言，这都是无法衡量、统一的，因此这就会要求用户来回往复地不断修正其三维表情模型，而这一过程显然又会耗费大量的时间。

本节提出的实时无标记脸部表情运动捕捉系统，无需用户进行任何预处理、预校准，为用户节约了大量的训练时间。该系统的输入为帧率30Hz、分辨率640×480 的彩色与深度图，由于脸部通常不能占满整幅画面，脸部的实际分辨率大约为160×160。我们采用了动态模板表情模型来实现特定用户的实时无标记脸部表情运动捕捉任务。与之前的方法类似[6,66]，当系统检测到用户的脸之后，系统即开始工作。区别在于，一段时间之后，系统内部的动态模板表情会收敛到特定用户的动态表情，并能很好地输出无标记运动捕捉与映射的结果。

2.2.3　自适应动态表情模型

2.2.3.1　基本概念

融合形状（blend shape）是一种三维动画制作领域非常基础的思想，其主要观点是：最终的结果可以通过所有的融合形状加权叠加得到。一般而言，脸部的融合形状由具有明显物理意义的基本表情脸构成，如中性脸、笑脸、张嘴脸等。这里，我们用$\mathbf{B}$来表示脸部的融合形状集合；用$\mathbf{b}_0$ 表示中性脸的融合形状，其是一个将三维模型的所有坐标点有序排列构成的列向量；$\mathbf{b}_i, i > 0$表示其他基本表情脸的融合形状，构成方式与中性脸一致，于是$\mathbf{B} = [\mathbf{b}_0, \ldots, \mathbf{b}_n]$。

脸部的动态表情模型正是对脸部融合形状集合表达的一个运算模型，对于一个新的脸部表情$\mathbf{F}$，其可以用融合形状集合表示为

$$\mathbf{F}(\mathbf{x}) = \mathbf{b}_0 + \Delta\mathbf{B}\mathbf{x}. \tag{2-9}$$

式（2-9）中，$\Delta\mathbf{B} = [\mathbf{b}_1 - \mathbf{b}_0 \ldots, \mathbf{b}_n - \mathbf{b}_0]$，$\mathbf{x} = [x_1, \ldots, x_n]^T$是融合形状的组合系数，范围为[0 ~ 1]。

使用融合形状的动态表情模型表示方法非常适合实时无标记脸部表情运动捕捉问题，原因在于其将捕捉问题转化成求解头部的刚体运动（全局旋转、平移）以及n 个脸部融合形状组合系数的优化问题，求解的维度大为降低。此外，更为值得一提的是，动态表情模型中的融合形状$\mathbf{B}$由特定物理意义的基本表情脸组成，这大大简化了后处理以及运动映射的任务。

考虑到已有方法与系统的缺陷，并结合融合形状思想的优势，本节提出一种自适应动态表情模型的方法，该方法无需用户进行任何预处理操作，也就是说无需事先获得特定用户的融合形状，其中最大的难题是如何对特定用户的融合形状进行建模? 本文提出用预先定义的模板融合形状作为系统的先验知识，设计相应的算法使模板融合形状收敛到特定用户的融合形状，这里用$\mathbf{B}^* = [\mathbf{b}_0^*, \ldots, \mathbf{b}_n^*]$来表示系统的模板融合形状集合矩阵。为了达到实时模板融合形状收敛到特定用户融合形状这一目标，本小节引入三个重要概念：单位PCA 模型、表情映射算子和变形校正场，如图2.8 所示，其中，特定用户的融合形状$\mathbf{B}$通过单位PCA模型、表情映射算子、以及与每个融合形状对应的变形校正场组合而成。

2.2.3.2　单位PCA模型

为了对融合形状集合$\mathbf{B}$中的中性融合形状$\mathbf{b}_0$进行建模，我们首先使用形变模型（morphable model）的方法采集不同用户中性脸的三维几何模型[67]。使用形变模型的方法采集三维脸，其最大优势在于可以得到不同用户具有相同拓扑关系的三维脸部模型。有了大量、不同用户、对齐的三维中性脸之后，我们使用主成分分析（PCA）的方法将三维中性脸进行降维处理，得到脸的单位PCA模型。这里用$\mathbf{m}$表示PCA降维得到的均值脸，$\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_l]$ 表示l 个最大特征值对应的特征向量。因此，任意一个中性融合形状$\mathbf{b}_0$可以近似表达为$\mathbf{b}_0 = \mathbf{m} + \mathbf{P}\mathbf{y}$，其中，$\mathbf{y} = [y_1, \ldots, y_l]^T$是特征向量线性组合系数。

2.2.3.3　表情映射算子

针对融合形状集合$\mathbf{B}$中的其他基本表情建模，我们使用变形映射的思想[68,69]将中性脸与其他基本表情脸建立关联。事实上，由于模板融合形状集合$\mathbf{B}^*$已知，模板中性融合形状$\mathbf{b}_0^*$到其他模板基本表情融合形状$\mathbf{b}_i^*$对应的变形映射可以事先得到。而在上一小节中，我们建立了中性脸的单位PCA模型$\mathbf{b}_0$，如果直接将模板融合形状中的变形映射关系应用到特定用户的中性融合形状$\mathbf{b}_0$，就可以得到特定用户的其他基本表情融合形状$\mathbf{b}_i$。

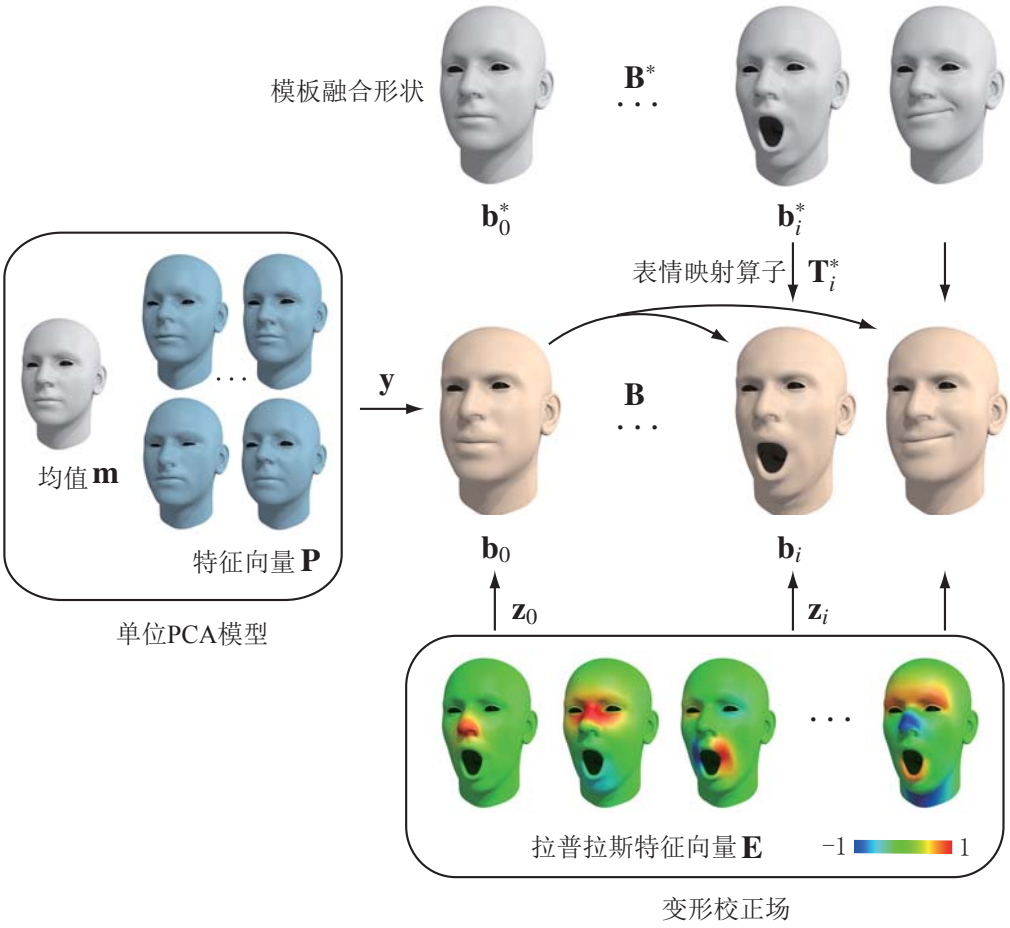


图 2.8　自适应动态表情模型

沿着这一思路，我们将其他基本表情的融合形状$\mathbf{b}_i$描述为中性融合形状$\mathbf{b}_0$的线性变换，即$\mathbf{b}_i = \mathbf{T}_i^* \mathbf{b}_0$，其中$\mathbf{T}_i^*$从模板融合形状$\mathbf{b}_0^*$与$\mathbf{b}_i^*$计算而来。值得指出的是，本节提出的$\mathbf{T}_i^*$算子与中性脸$\mathbf{b}_0^*$无关，这与已有文献给出的方法是不同的[68,70]，也正因为该算子与中性脸并不耦合，我们可以有效地对相应的融合形状进行优化计算，有关$\mathbf{T}_i^*$算子的计算推导将在全文的附录A中给出。

2.2.3.4　变形校正场

一般而言，单位PCA模型通过分解大量的三维中性脸，无法有效地表达特定用户中性脸的细节；同样地，直接从模板融合形状复制过来的表情映射算子$\mathbf{T}_i^*$也没有考虑特定用户的细节特征，与特定用户的融合形状并不完全一致。为此，我们提出一种变形校正场的方法来修正特定用户融合形状集合$\mathbf{B}$中的每个融合形状，其目的是尽量多地恢复特定用户融合形状的细节特征。我们首先计算三维模型脸的顶点邻接关系对应的拉普拉斯矩阵L，然后对该拉普拉斯矩阵进行特征值分解，得到k个最小特征值对应的特征向量$\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_k]$，根据这$k$个特征向量，我们可

以得到融合形状$\mathbf{b}_i$对应的补偿分量$\mathbf{E}\mathbf{z}_i$，其中$\mathbf{z}_i = [z_{i,1}, \ldots, z_{i,k}]^T$是特征向量的线性组合系数，这样一种做法也称为几何模型的谱分析方法，被广泛地用于三维模型的去噪、平滑处理等问题[71]。以上提出的基于拉普拉斯特征向量的变形校正场的方法有两大优势：一方面，特定用户融合形状只需优化拉普拉斯特征变量对应的k个线性组合系数，大大降低了求解的维度；另外一方面，在与高噪声的深度信号进行对齐计算时，低频拉普拉斯特征向量可以保证融合形状在优化过程中保持一定的平滑性。

2.2.3.5　自适应动态表情模型

根据前文所述的三个基本模块，本文提出的自适应动态表情模型表述如下：首先，特定用户中性脸的融合形状表示为$\mathbf{b}_0 = \mathbf{m} + \mathbf{P}\mathbf{y} + \mathbf{E}\mathbf{z}_0$，也就是说，特定用户中性脸的融合形状由单位PCA模型和变形校正场构成；其次，特定用户的每一个其他基本融合形状$\mathbf{b}_1, \ldots, \mathbf{b}_n$参数化为$\mathbf{b}_i = \mathbf{T}_i^* \mathbf{b}_0 + \mathbf{E}\mathbf{z}_i = \mathbf{T}_i^*(\mathbf{m} + \mathbf{P}\mathbf{y} + \mathbf{E}\mathbf{z}_0) + \mathbf{E}\mathbf{z}_i$，也即是由模板融合形状$\mathbf{B}^*$的表情映射算子与变形校正场构成，其中$\mathbf{m}, \mathbf{P}, \mathbf{E}, \mathbf{T}_i^*$均可以事先求得。图2.8展示了本文提出的自适应动态表情模型。

2.2.4　优化方法

本文提出的自适应动态表情模型是实时无标记脸部表情运动捕捉的核心，本小节将重点讨论参数的具体优化方法，其中包括计算精确的运动参数，以及优化特定用户对应的自适应动态表情模型参数。更为具体地讲，算法需要解决：

- 任意t时刻，与单相机系统输入深度信号对齐的脸部刚性运动，包括全局旋转$\mathbf{R}$和全局平移$\mathbf{t}$；
- 任意t时刻，式（2-9）中，脸部基本表情的融合形状系数：$\mathbf{x} = [x_1, \ldots, x_n]^T$；
- 单位PCA模型中，中性融合形状$\mathbf{b}_0$的PCA系数：$\mathbf{y} = [y_1, \ldots, y_l]^T$；
- 所有融合形状$\mathbf{b}_i, i > 0$的变形校正系数：$\mathbf{Z} = \{\mathbf{z}_0, \ldots, \mathbf{z}_n\}$，其中$\mathbf{z}_i = [z_{i,1}, \ldots z_{i,k}]^T$。

针对以上需要解决的优化问题，概括地讲，我们首先固定动态表情模型，求解最优的运动参数；然后利用求解的运动参数，重新优化模型参数：$\mathbf{y}$和$\mathbf{Z}$，通过不断迭代地对运动参数以及模型参数进行优化，实现实时无标记脸部表情运动捕捉的目标，图2.9显示了优化方法的两大步骤，针对相机系统输入的每一帧信号（包括彩色信号和深度信号），提出的方法可迭代地优化求解运动参数以及特定用户对应的动态表情模型参数。系统输出的每一帧运动参数包括：刚性运动参数（全局旋转和平移）和融合形状的组合系数。系统不仅可以输出收敛的特定用户

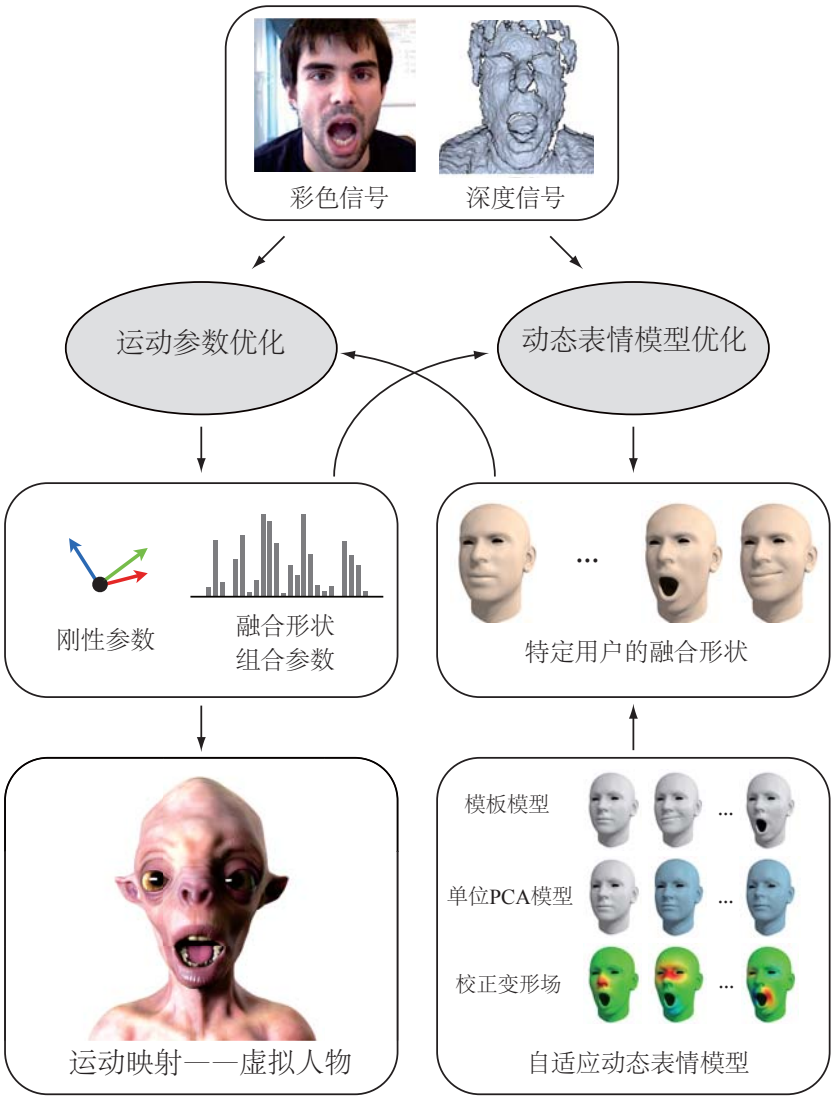


图 2.9　优化流程

动态表情模型，还能够实时输出虚拟角色的运动映射结果。本文后续的小节将从三个方面详细阐述系统的迭代优化方法。

需要说明的是，我们用上标表示时间序号，例如$\mathbf{x}^t$表示$t \in \mathbb{N}$时刻的融合形状组合系数，其中$t = 1$表示起始帧。为了描述方便，在不引起歧义的情况下，所有上标都略去不写。

2.2.4.1　算法初始化

初始化时，我们的系统需要用户以中性脸的方式进入单相机系统的视场范围，系统会实时扫描用户的中性脸[72]。通过第一次扫描，我们可以求得$\mathbf{b}_0$的初始估计，也即是求解单位PCA模型的系数$\mathbf{y}$，变形校正场的拉普拉斯向量系数$\mathbf{z}_0$，以及脸部姿态的刚性运动参数$(\mathbf{R}, \mathbf{t})$。实际上，我们采用点面约束的迭代最近邻

(a) 刚性运动参数　　(b) 融合形状组合系数

图 2.10　估计参数时所用的顶点区域，蓝色表示

（Iterated Closet Point, ICP）[73]的方法来求解这些参数，具体来讲，我们通过优化

$$\arg\min_{\mathbf{R},\mathbf{t},\mathbf{y},\mathbf{z}_0} \|\mathbf{A}_0(\mathbf{R}\mathbf{b}_0+\mathbf{t})-\mathbf{c}_0\|_2^2+\beta_1\|\mathbf{D_P}\mathbf{y}\|_2^2+\beta_2\|\mathbf{D_E}\mathbf{z}_0\|_2^2+\beta_3\|\mathbf{z}_0\|_2^2. \tag{2-10}$$

来得到中性脸$\mathbf{b}_0$的初始估计。

式（2-10）第一项中的$(\mathbf{A}_0,\mathbf{c}_0)$是由点面约束构成的矩阵，具体构造过程可以参见文献[6]；$\mathbf{D_P}\mathbf{y}$用来约束单位PCA模型的系数，$\mathbf{D_P}$是由PCA特征向量对应特征值的倒数组成的对角矩阵；$\mathbf{D_E}\mathbf{z}_0$用来约束变形校正系数，$\mathbf{D_E}$是拉普拉斯特征矩阵$\mathbf{L}$的特征向量对应特征值组成的对角矩阵[74]；最后一项用来约束变形向量的变化幅度；β_1,β_2,β_3都是正的常数。

我们使用高斯-牛顿方法优化式（2-10）[75]，在求解的初始时刻，$\mathbf{y}=\mathbf{z}_0=0$。当求解得到初始帧（$t=1$）的中性脸$\mathbf{b}_0^1$之后，其他基本融合形状可以通过中性脸的融合形状和表情映射算子得到，即：$\mathbf{b}_i^1=\mathbf{T}_i^*\mathbf{b}_0^1$，$i=1,\ldots,n$。

2.2.4.2　运动参数优化

在进行运动参数优化的阶段，脸部表情的融合形状没有做任何其他优化处理。具体来讲，我们需要求解脸部的刚性运动参数$(\mathbf{R},\mathbf{t})$，以及t 时刻的融合形状组合系数$\mathbf{x}$。事实上，通过将前一帧重建出的三维脸与当前帧观测到的深度信号进行点面ICP对齐，我们可以得到刚性运动参数的估计结果。为了使求解的刚性运动参数不至于发生大范围的跳变，点面ICP的约束仅作用于前脸的鼻子上半部分区域，如图2.10(a) 所示的蓝色区域。

给定脸部的刚性运动参数，以及脸部的融合形状$\mathbf{B}$之后，估计与当前帧深度信号最匹配的融合形状组合系数$\mathbf{x}$就可以描述成能量最小化的优化问题：

$$\arg\min_{\mathbf{x}} E_{\text{fit}}+\lambda_1 E_{\text{smooth}}+\lambda_2 E_{\text{sparse}}. \tag{2-11}$$

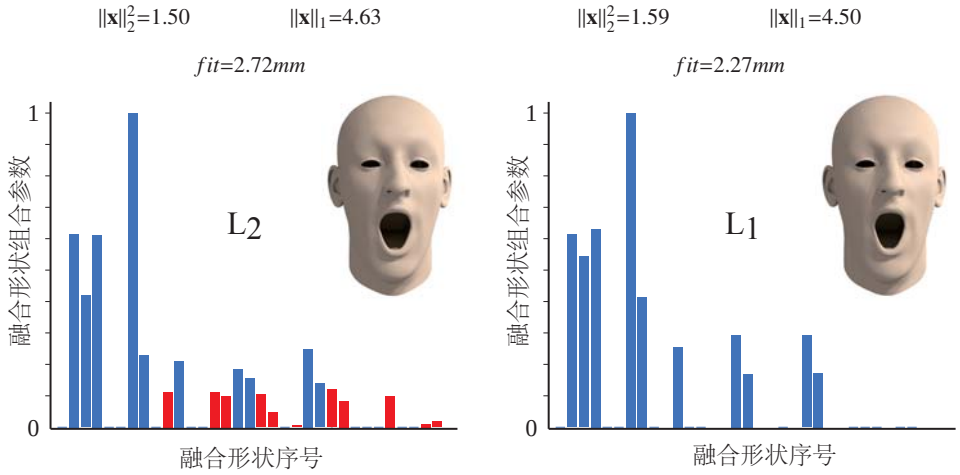


图 2.11　不同范数的正则项对式（2-11）的优化结果比较

式（2-11）的第一项为数据项，具体数学形式为：

$$E_{\text{fit}} = \|\mathbf{A}(\mathbf{b}_0 + \Delta\mathbf{B}\mathbf{x}) - \mathbf{c}\|_2^2. \tag{2-12}$$

其中，$(\mathbf{A}, \mathbf{c})$是图2.10(b)中蓝色顶点对应的对齐约束矩阵，该矩阵的详细推导可以在文献[6]找到。

式（2-11）的后两项为正则项，$E_{\text{smooth}} = \|\mathbf{x}^{t-2} - 2\mathbf{x}^{t-1} + \mathbf{x}^t\|_2^2$ 保证脸部表情运动在时间上的平滑性；而$E_{\text{sparse}} = \|\mathbf{x}\|_1$ 是L1-范数约束。我们发现L1-范数能够使运动参数的估计过程变得十分稳定，这是由于融合形状之间并不完全独立，同样的表情可以通过融合形状不同的系数组合得到，如图2.11所示。从中我们看出，L1-范数正则化可以获得更低的匹配误差（用fit 表示），但更为重要的是，大幅减少了融合形状参数的非零数目。左图红色的柱状条显示了L2-范数正则项额外使用的融合形状。。

我们使用热启动的梯度下降法来优化式（2-11）[60]，在求解优化的每一步，融合形状的参数$\mathbf{x} = [x_1, \ldots, x_n]^T$ 都被强制限定在$[0, 1]$ 之间。

2.2.4.3　动态表情模型优化

在求得脸部的刚性运动参数$(\mathbf{R}, \mathbf{t})$以及融合模型的组合参数$\mathbf{x}$之后，接下来需要对动态表情模型的PCA系数$\mathbf{y}$和拉普拉斯变形系数$\mathbf{z}_0, \ldots, \mathbf{z}_n$进行优化。

我们将式（2-12）重新改写成如下形式

$$E_{\text{fit}} = \|\mathbf{A}(\mathbf{b}_0 + \Delta\mathbf{B}\mathbf{x}) - \mathbf{c}\|_2^2 = \|\mathbf{A}[\bar{x}\mathbf{b}_0 + \sum_{i=1}^{n} x_i\mathbf{b}_i] - \mathbf{c}\|_2^2, \tag{2-13}$$

其中，$\bar{x} = 1 - \sum_{i=1}^{n} x_i$。

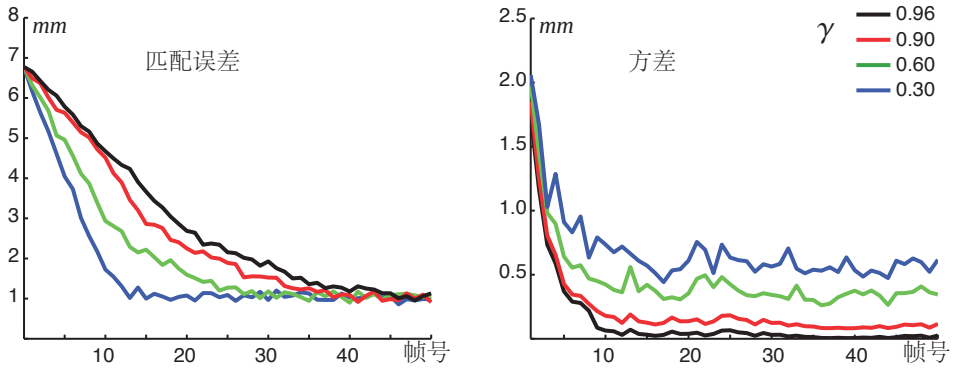


图 2.12　时间衰变因子γ的影响

由于$\mathbf{b}_0 = \mathbf{m} + \mathbf{P}\mathbf{y} + \mathbf{E}\mathbf{z}_0$，$\mathbf{b}_i = \mathbf{T}_i^* \mathbf{b}_0 + \mathbf{E}\mathbf{z}_i$，于是式（2-13）可以进一步描述成

$$E_{\text{fit}} = \|\bar{\mathbf{A}}\mathbf{u} - \bar{\mathbf{c}}\|_2^2. \tag{2-14}$$

式（2-14）中，

$$\bar{\mathbf{A}} = \mathbf{A}[(\bar{x}\mathbf{I} + \sum_{i=1}^{n} x_i \mathbf{T}_i^*)\mathbf{P}, (\bar{x}\mathbf{I} + \sum_{i=1}^{n} x_i \mathbf{T}_i^*)\mathbf{E}, x_1\mathbf{E}, \ldots, x_n\mathbf{E}],$$

且

$$\mathbf{u} = [\mathbf{y}^T, \mathbf{z}_0^T, \ldots, \mathbf{z}_n^T]^T, \quad \bar{\mathbf{c}} = \mathbf{c} - \mathbf{A}(\bar{x}\mathbf{I} + \sum_{i=1}^{n} x_i \mathbf{T}_i^*)\mathbf{m}.$$

与式（2-10）类似，我们对PCA的系数、拉普拉斯变形系数以及变形幅度进行正则化约束，于是可以得到动态表情模型优化的能量方程：

$$E_{\text{ref}} = \|\bar{\mathbf{A}}\mathbf{u} - \bar{\mathbf{c}}\|_2^2 + \beta_1\|\mathbf{D}_\mathbf{P}\mathbf{y}\|_2^2 + \sum_{i=0}^{n}(\beta_2\|\mathbf{D}_\mathbf{E}\mathbf{z}_i\|_2^2 + \beta_3\|\mathbf{z}_i\|_2^2). \tag{2-15}$$

值得注意的是，优化动态表情模型不能仅使用当前帧的信息，还必须考虑所有已经观测的数据信息。但是，无论从计算空间复杂度还是从计算时间复杂度考虑，使用所有已观测数据来优化动态表情模型是完全不可能的事情。我们采用存储空间恒定的时间聚合方法，将时间因素考虑进式（2-15）中，得到优化方程：

$$\arg\min_{\mathbf{y},\mathbf{z}_0,\ldots,\mathbf{z}_n} \sum_{j=1}^{t} \frac{\gamma^{t-j}}{\sum_{j=1}^{t}\gamma^{t-j}} E_{\text{ref}}^j, \tag{2-16}$$

式（2-16）中，t是当前帧序号；$0 \le \gamma \le 1$定义了一个随时间指数衰减的遗忘因子；E_{ref}^j是j时刻的动态表情模型优化能量方程。显然，式（2-16）有闭式解，解的形式为：

$$(\mathbf{D} + \sum_{j=1}^{t} \frac{\gamma^{t-j}}{\sum_{j=1}^{t}\gamma^{t-j}} (\bar{\mathbf{A}}^j)^T \bar{\mathbf{A}}^j)\mathbf{u} = \sum_{j=1}^{t} \frac{\gamma^{t-j}}{\sum_{j=1}^{t}\gamma^{t-j}} (\bar{\mathbf{A}}^j)^T \bar{\mathbf{c}}^j, \tag{2-17}$$

表 2.1　t时刻动态表情模型优化算法流程

序号	具体执行步骤
1	初始化：$\mathbf{M}^1 = \mathbf{0}, \mathbf{y}^1 = \mathbf{0}, s^1 = 0$
2	$s^t = \gamma s^{t-1} + 1$
3	$\mathbf{M}^t = \gamma \frac{s^{t-1}}{s^t}\mathbf{M}^{t-1} + \frac{1}{s^t}(\bar{\mathbf{A}}^t)^T \bar{\mathbf{A}}^t$
4	$\mathbf{y}^t = \gamma \frac{s^{t-1}}{s^t}\mathbf{y}^{t-1} + \frac{1}{s^t}(\bar{\mathbf{A}}^t)^T \bar{\mathbf{c}}^t$
5	输出：$\mathbf{u}^t = GaussSeidel(\mathbf{M}^t + \mathbf{D}, \mathbf{y}^t, \mathbf{u}^{t-1})$

其中，$\mathbf{D}$是包含式（2-15）正则项的对角矩阵，我们可以使用热启动的*GaussSeidel*方法来求解该式[76]。

这样，我们就可以在有限的存储资源条件下考虑所有观测到的数据信息，而不必将每一帧的信息都单独存储下来，算法流程总结在表格2.1中。图2.12显示了式（2-16）中γ参数对匹配误差以及方差的影响，其中，匹配误差通过平均每帧的ICP误差得到，方差体现三维顶点误差随时间的变化。图中表明较低的时间衰变因子可以获得更快的匹配误差下降率，但是会使得方差变大。我们发现$\gamma = 0.9$可以获得很好的优化效果。

2.2.4.4　实现细节

原则上来讲，动态表情优化可以和运动参数优化一样，一直进行下去。但是，我们可以使用一种启发式的方法来提高系统的执行效率：优化次数达到一定阈值的融合形状将不再被优化。为此，我们定义$\sigma_i = \sum_{j=1}^{t} x_i^j$来描述融合形状$\mathbf{b}_i$在$t$时候被优化的“饱和度”。当$\sigma_i > \bar{\sigma}$，融合形状$\mathbf{b}_i$在后续的帧中将认为已经收敛，不再参与优化过程。由于$\mathbf{b}_0$在自适应动态表情模型中的独特地位，我们将一直优化$\mathbf{b}_0$，直到所有的融合形状都大于给定的阈值，即$\sum_{j=1}^{t} \max(\bar{x}^j, 0) > \bar{\sigma}$。图2.13直观地展示了动态表情优化与计算时间之间的关系，其中，左图显示了当越来越多的融合形状达到收敛时，动态表情优化中融合形状的数目将不断降低。右图则显示了每一帧需要优化的融合形状数目与计算时间的关系曲线。当越来越多的融合形状达到收敛时，计算代价会逐渐降低，直到动态表情模型优化的时间几乎可以忽略不计。

所有的实验中，我们均使用了35个融合形状，即$n = 35$。单位PCA模型所用的三维脸来自于文献[67]，其中包括100名男性和100名女性，PCA模型对应的特征向量数目为50，变形校正使用的拉普拉斯特征向量个数也为50。式（2-10）和式（2-15）的参数为：$\beta_1 = 0.5$，$\beta_2 = 0.1$，$\beta_3 = 0.001$；式（2-11）的参数为$\lambda_1 = 10$，$\lambda_2 = 20$；收敛阈值$\bar{\sigma} = 10$。

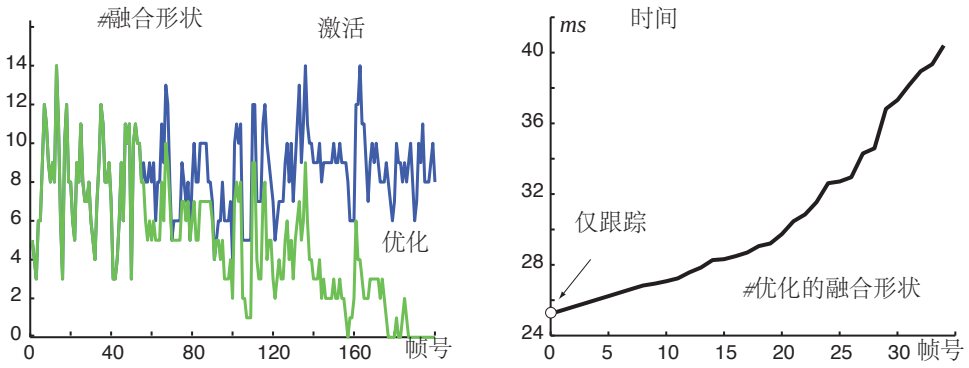


图 2.13　优化性能比较图

所有的算法均用C++完成实现，并用OpenMP[①]进行了并行化加速；线性代数的运算使用了Eigen库[②]，人脸检测以及一些图像处理使用OpenCV库[③]来完成。为了完成实时无标记人脸捕捉的系统，我们使用了基于图像的眼球跟踪算法，由于估计脸部的刚性运动参数比较准确，我们可以采用k近邻方式搜索数据库中标记、校准好的眼球运动，眼球运动的最终结果是由搜索的k近邻加权平均得到。实验表明，在主频为2.7GHz的英特尔酷睿i7处理器、显卡为英伟达GT650M、内存为16G 的计算机上，本文提出的系统可以达到25Hz的捕捉帧率。

2.2.5　实验结果与分析

我们做了许多实验来验证提出的系统与方法的性能，本小节将展示与此相关的诸多实验结果。此外，有关系统的局限性也在本章的最后做了适当地讨论。

2.2.5.1　动态表情模型实验

图2.14显示了单位PCA模型、变形校正场的特征向量数目对中性融合形状$\mathbf{b}_0$的影响。由于建立单位PCA模型的全部人脸数目十分有限（实验中仅采用100个男性和100个女性三维模型），我们发现50个特征向量已经足够描述200个人脸数据集；此外，我们发现仅用50个拉普拉斯特征向量的变形校正场就可以获得很好的运动捕捉结果。图2.15也显示了变形校正场对动态表情模型的影响，其中，第一行是单位PCA模型与表情映射的结果；第二行在融合形状（中性脸$\mathbf{b}_0$和其他表情$\mathbf{b}_i$）基础上加入校正变形场；第三行用不同颜色表达了两者之间的差别，也即是$\mathbf{E}\mathbf{z}_i$。从图中我们可以直观地发现，变形校正场对嘴部以及鼻孔附近区域的点有比较明显的影响，一般来讲，变形校正场对脸部的非对称区域可以获得更高的

① http://openmp.org/
② http://eigen.tuxfamily.org/index.php
③ http://opencv.org/

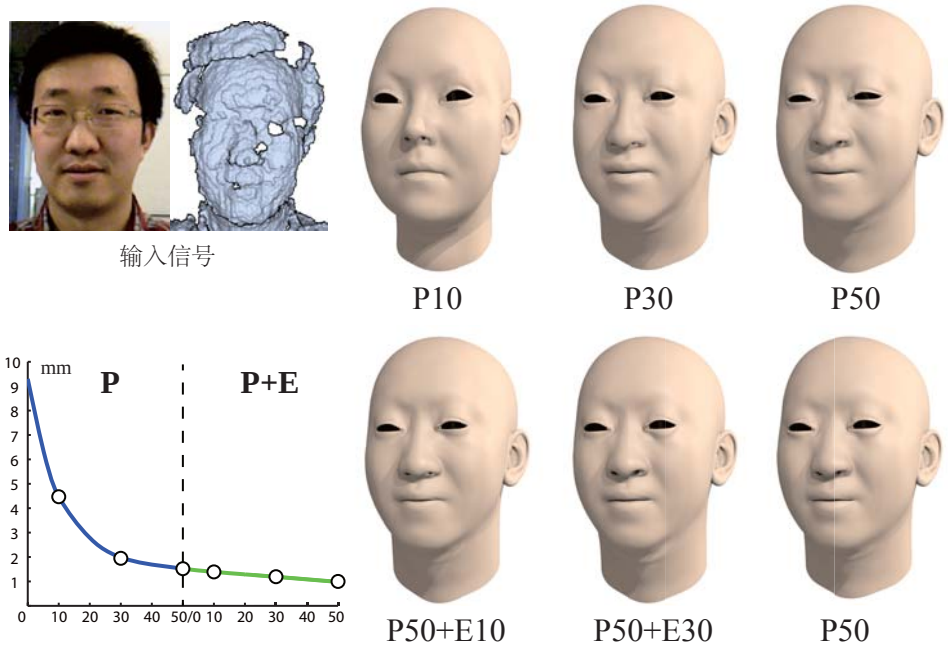


图 2.14　单位PCA模型**P**的特征向量数目、拉普拉斯矩阵**E**的特征向量数目，对中性脸$\mathbf{b}_0$的影响。左下图显示了平均ICP匹配误差与特征向量数目的关系曲线

几何建模精度。

2.2.5.2　运动参数与模型优化实验

图2.12显示了随着动态表情模型的逐步优化，平均匹配误差可以逐渐降低，图2.16显示了与此相对应的融合形状模型，其中，每一行显示了一个特定表情的融合形状，最右边的彩色小图为参考表情。此外，我们也将本文提出的系统与商业化脸部捕捉系统FaceShift[①] 做了比较，值得注意的是，FaceShift需要事先进行繁琐的训练、校准，本文提出的系统无需任何训练、校准，图2.17中表明，和只用模板表情的运动捕捉方法相比，运动动态表情优化可以明显地提高捕捉精度。并且，本文提出的方法与商业化的运动捕捉软件Faceshift（FS）相比，捕捉的运动精度相当，但是Faceshift需要进行繁琐的人工校准与训练，而本文提出的方法是全自动的。

2.2.5.3　运动映射

本文提出的表情映射算子可以保证收敛后的融合形状与模板融合形状具有类似的表情语义，因此系统输出的运动参数可以直接驱动具有同样融合形状的虚拟角色，这样无需额外处理的运动映射方式可以实现实时运动映射的任务。

① http://www.faceshift.com/

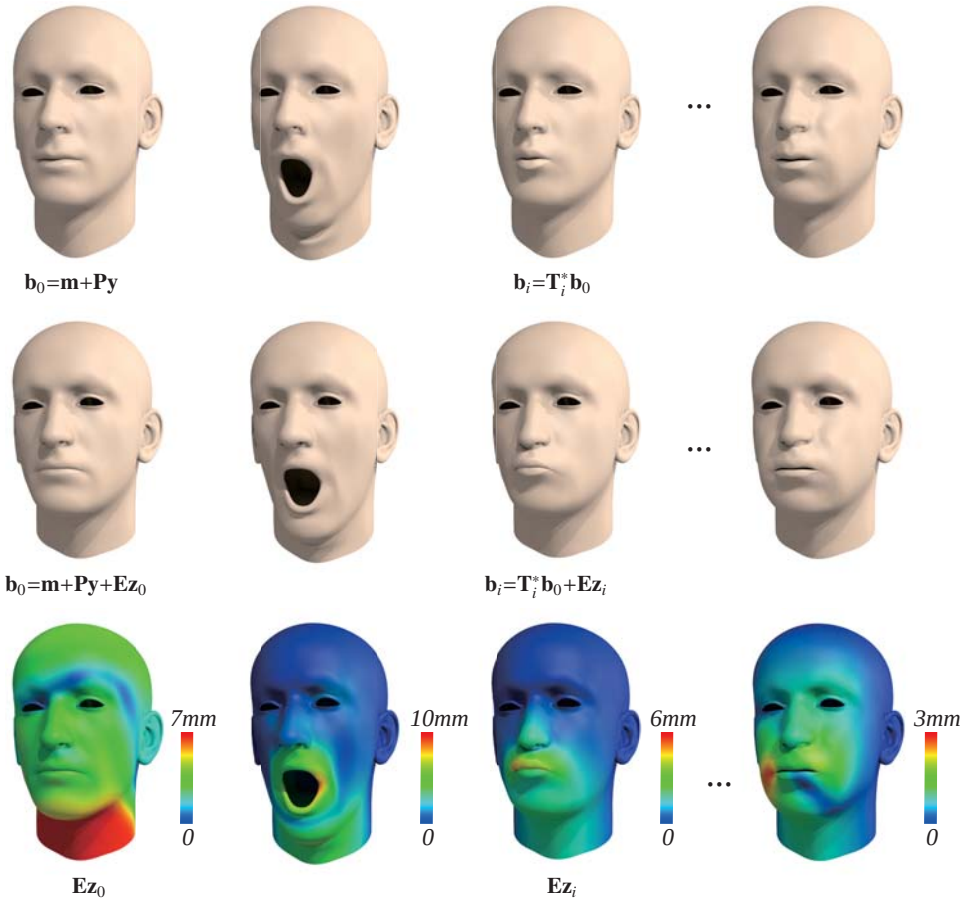


图 2.15　变形校正场的影响

图2.18显示了系统运动映射后的结果，这是使用本文提出方法的一个全新的应用。观察者直接站到屏幕之前，屏幕显示的虚拟人物可以模仿用户的表情，做同样的表情运动。

2.3　本章小结

本章着重讨论了单相机系统的实时无标记脸部表情运动捕捉问题，由于传统的单相机系统在拍摄过程中丢失了深度信息，使用单相机系统的运动捕捉属于信息欠定问题，将深度信息引入单相机系统的运动捕捉是主流的方法之一。而获取单相机系统的深度是传统的计算机视觉问题，针对这一难题，本章提出了一种非线性参数模型，通过对数据库的合理建模，学习出类别相关的彩色图像块与深度值的映射参数，提高了单相机系统深度估计的精度。另外一方面，随着可直接获取深度的新型单相机系统逐渐普及，单相机系统的深度信息获取已经十分容易，基于新型的单相机系统，本章提出了一种动态表情模型的无标记脸部表情运动捕捉方法与系统，该方法无需任何事先校准与训练即可进行特定用户的实时脸部表情运动捕捉。本章提出的方法将运动捕捉问题分解为两个子问题：运动参数的估

图 2.16　动态模型优化

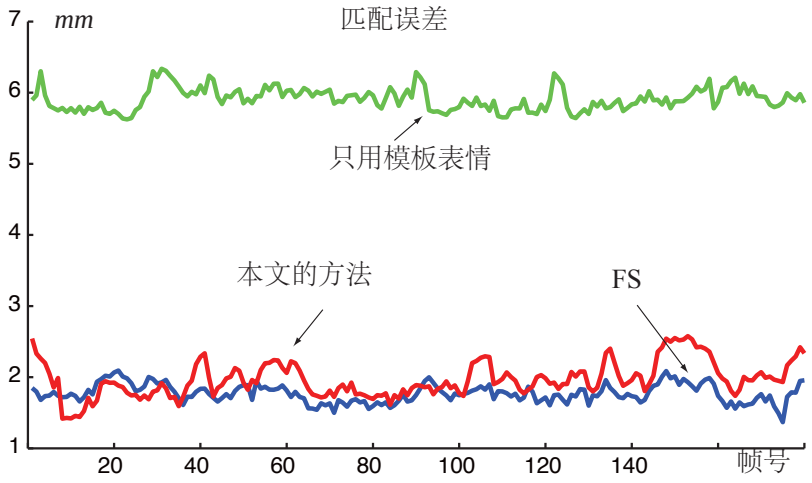


图 2.17　不同运动捕捉方法的匹配误差比较结果

图 2.18　交互式肖像

计和特定用户的表情模型参数估计。随着系统观测信息的不断增多，系统内部的模板表情可以自适应地收敛到特定用户的表情，解决了计算时间复杂与运动捕捉精度的矛盾。需要指出的是，受限于新型传感器深度获取的分辨率与精度，本章提出的方法仍有较大的改进空间。一种直接的思路是采用更好的深度传感单元提升获取深度的精度，另外一种思路是使用后续章节提出的多相机系统来实现无标记的运动捕捉。尽管如此，本章提出的运动捕捉技术，在大幅提升运动捕捉系统易用性的同时，没有降低脸部表情运动捕捉的精度，本章提出的方法与系统完全可以代替需要繁琐预处理与训练的传统运动捕捉方法。

第 3 章　固定多相机系统的无标记手与物体交互运动捕捉

在前一章中，我们讨论了近距离场景无遮挡或少遮挡的运动捕捉问题，并以实时无标记脸部表情运动捕捉作为具体的研究对象。本章将讨论另一类近距离场景的运动捕捉——包含严重遮挡的近距离场景无标记运动捕捉问题，特别地，涉及手与物体的交互运动捕捉。实际上，捕捉高度灵活的手与物体的交互过程是一个极具挑战性的难题。一方面，正如前面所述，交互运动过程中通常存在着十分严重的遮挡（包括手指之间的遮挡和手与物体之间的相互遮挡），单个视角的信息不能有效地还原真实三维运动，上一章提出的单相机系统无法实现高精度的交互运动捕捉。另外一方面，交互运动有其内在的动力学特征，如果不考虑交互运动过程中的物理特性，运动捕捉的结果通常无法直接映射到其他虚拟物体上。此外，交互运动捕捉不仅需要捕捉手与物体在运动过程中的各自姿态，还需要捕捉手与物体在交互过程中的细微接触运动，这些细微接触运动主要表现为手的指节与物体不能出现互相穿透（penetration）、互相分离（departure）的情形。针对手与物体交互这一特定的运动形式，本章提出一种固定多相机系统的无标记运动捕捉方法，建立运动过程中手的姿态、物体的姿态以及相互之间细微接触的完整动力学运动方程，提出复合运动控制器模型，改进和完善传统的手与物体交互运动捕捉方法，有效解决了包含严重遮挡的近距离场景无标记运动捕捉难题。

3.1　导言

在计算机图形学领域，制作超高真实感的手与物体交互运动的虚拟视频序列（例如抓起细长的杯把手、单手旋转魔方）是一个极具挑战性的任务。近些年来，随着各种新型传感器的出现，以及计算性能的提高，基于数据驱动的方法为手与物体交互运动视频的制作提供了一种可行的技术方案。顾名思义，基于数据驱动的视频制作旨在建立特定类别的交互运动数据库，然后对数据库中已有的运动数据进行插值、整合、修缮以及重映射处理，实现制作全新虚拟交互运动视频的目标。显然，建立特定类别的交互运动数据库需要捕捉不同形式的手与物体交互运动，然而捕捉高度灵活的手与物体的交互运动非常困难，其原因在于我们不仅要捕捉运动过程中手的姿态和物体的姿态，特别地，还需要捕捉手与物体在交互运动过程中的细微接触运动。

事实上，研究者已经探索了大量的方法与系统来捕捉手与物体的运动，目前，这些方法可以概括为三大类：基于标记点的运动捕捉、基于数据手套的运动捕捉

图 3.1　本文提出的方法和系统可以从固定多相机系统拍摄的视频中恢复出满足物理约束的真实运动数据，第一行为原始图像，第二行为捕捉的运动

和基于视频的运动捕捉。虽然研究人员针对手与物体的交互运动捕捉做了不懈的努力，但是捕捉高度灵活的手与物体的交互运动仍充满巨大的挑战。一方面，基于标记点的运动捕捉系统虽然已被广泛地使用（例如脸部表情的运动捕捉，人体全身的运动捕捉），但由于手与物体在交互过程中存在着严重的遮挡，基于标记点的运动捕捉系统很容易产生歧义的运动捕捉结果，无法有效地用于高精度的交互运动捕捉；另外一方面，虽然基于数据手套的运动捕捉系统不受运动过程中的遮挡影响，但是由于传感器的精度与布置问题，捕捉的运动信息通常伴随大量的噪声，这种类型的系统很难用于高精度的交互运动捕捉。值得一提的是，以上两种类型的系统能否捕捉手与物体交互过程中的细微接触运动还是一个未知数。

近年来，基于视频的运动捕捉系统获得了研究者极大的关注。与前面所述的两种捕捉系统不同，由于不需要标记点、手套以及任何传感器，基于视频的运动捕捉系统不会对手与物体的交互运动产生任何侵入性干扰，捕捉对象获得了极大的灵活自由度。然而，目前的基于视频的手与物体交互运动的捕捉系统却受到三方面的制约，主要表现在：（1）已有的方法与系统很容易受到遮挡以及手上缺少特征的影响，捕捉的运动往往具有歧义性；（2）目前的方法与系统主要着眼于手的关节运动，而完全忽略了手与物体的交互运动，因此也就忽略掉了手与物体之间的相互作用对求解结果的影响；（3）到目前为止，还没有方法与系统考虑手与物体在交互运动过程中的动力学特征，由于没有捕捉运动过程中的动力学特性，捕捉的运动通常噪声很大并且不满足物理约束，更为重要的是，捕捉的手与物体的运动无法有效的用于运动映射，因而也就很难用来建立交互运动的数据库。

本章提出了一种新的基于固定多相机系统的无标记手与物体交互运动捕捉系统，核心思想是提出复合运动控制器的模型，对运动过程中手的运动、物体的运动以及它们之间的细微接触运动，建立一套完整的物理动力学运动模型，利用输

入的多视点视频信号，最大化真实视频与仿真运动的一致性。本章提出的运动捕捉方法，目标是寻找最优的复合运动控制器，目的是使仿真的运动与输入的多视点视频能够最好地匹配。需要指出的是，将物理动力学运动模型与传统的基于视频的手与物体的交互运动捕捉方法结合之所以可以有效地解决前文所述的三个制约，原因有如下几点：首先，运动的求解状态（即空间结构）发生变化。传统方法仅使用多视点的视频信息，很难恢复出歧义性很大的接触点位置以及手与物体的细微接触运动，动力学模型可以消除大量的不满足物理约束的歧义状态；其次，运动的求解过程发生变化。通过将求解的姿态限定在满足物理约束的运动空间中，可以最大限度地降低运动求解过程带来的运动歧义性。此外，使用物理动力学模型的运动捕捉方法还可以为虚拟视频制作带来额外的好处：只需要简单修改控制器的相应参数（例如摩擦系数，运动速率等），就可以十分容易地进行运动映射，这就使运动捕捉的结果更方便使用。

3.2　相关研究

为了捕捉高度灵活的手与物体交互的运动数据，本文提出的方法结合基于视频的无标记运动捕捉和物理约束的动力学建模两种方法的优势。在下文中，我们将分别综述这两个方面的相关研究工作。

3.2.1　基于视频的无标记运动捕捉

基于视频的无标记运动捕捉通常采用运动跟踪的思想，特别地，针对手的无标记运动捕捉问题，主流的做法是事先得到手的三维模型，实现基于该模型骨架的运动跟踪[26,27]。一般来讲，在运动跟踪的初始时刻，方法需要手动或自动地将手的三维姿态与第一帧的视频图像进行对齐。在视频序列的后续每一帧中，虚拟生成大量的三维姿态，通过计算观测图像特征与虚拟三维姿态渲染生成的图像特征之间的误差，选择误差最小的虚拟姿态作为每一帧的跟踪结果。由于这种方法仅使用视频的信息，在使用中会有诸多的限制与歧义性，比如严重的遮挡、手上缺少明显的特征会给运动跟踪带来极大的困难。显然，仅使用视频图像的信息无法捕捉高度灵活的手与物体的交互运动。

为了减少运动跟踪过程中的歧义性，结合预先采集的运动数据，并将其作为运动跟踪的先验知识，是一种比较广泛使用的手段。总体来讲，这样一种手段有两种实际的做法：鉴别式方法和生成式方法。一般来讲，鉴别式方法[28,29]主要针对单视点的三维姿态恢复问题，它们首先从数据库中学习出合适的分类器，然后将观察到的图像用分类器进行分类，从而恢复手的三维姿态。实际上，由于手具

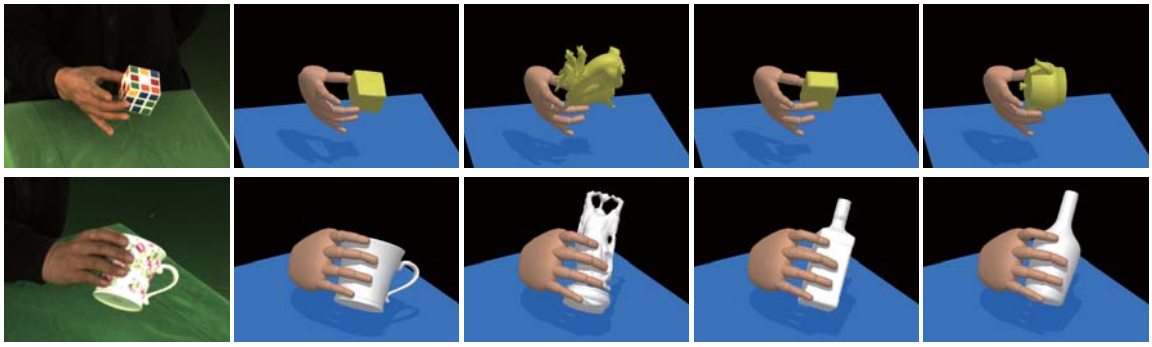


图 3.2　将捕捉的运动数据映射到具有不同外观的物体。从左到右，分别显示了：原始图像，捕捉的交互运动数据，以及将捕捉的运动数据映射到三种不同外观的物体上

有很大的灵活性，试图学习出手的整个姿态空间的分类器是几乎不可能的，因此，通过鉴别式方法进行手的姿态估计，无法有效地做到高精度的姿态恢复。值得一提的是，鉴别式的方法通常还需要比较好的初始结果。生成式方法[7,30,31] 则主要针对多视点的运动捕捉问题，一般使用带有骨架结构的三维模型，通过最小化三维模型的投影图与观测图之间的对齐误差，得到手的三维姿态估计。需要说明的是，距离图、轮廓、光流都可以用来计算投影图与观测图之间的对齐误差，并且三维骨架的运动空间可以通过数据库来预先设定合理的阈值。显然，基于数据库的先验知识仅能可靠地恢复与数据库内容一致的运动参数。此外，以上两类方法主要针对手的姿态恢复，而没有考虑手与物体的交互运动捕捉问题。

针对手与物体的交互运动捕捉问题，Ballan和他的同事[34]改进了传统的鉴别式方法，通过学习手指上显著点的鉴别特征，估计手的运动参数。而Oikonomidis等人[35] 则将碰撞检测的思想引入到生成式的运动跟踪方法中，试图同时恢复手与物体的运动参数。这些方法虽然可以有效地避免交互运动过程中的相互穿透，但是由于没有考虑交互运动过程中的动力学特性，无法识别以及去除相互分离的运动估计噪声。

值得一提的是，融合基于标记点的运动捕捉方法，也可以减少单纯依赖视频运动跟踪方法的歧义性[32,33]。Zhao等人[33]提出了一个同时使用Kinect和标记点的三维手势运动捕捉系统，并阐释了如何结合这两种方法的优势。但是，他们的系统能否捕捉手与物体的交互运动并没有得到有效地验证。

3.2.2　物理约束的动力学建模

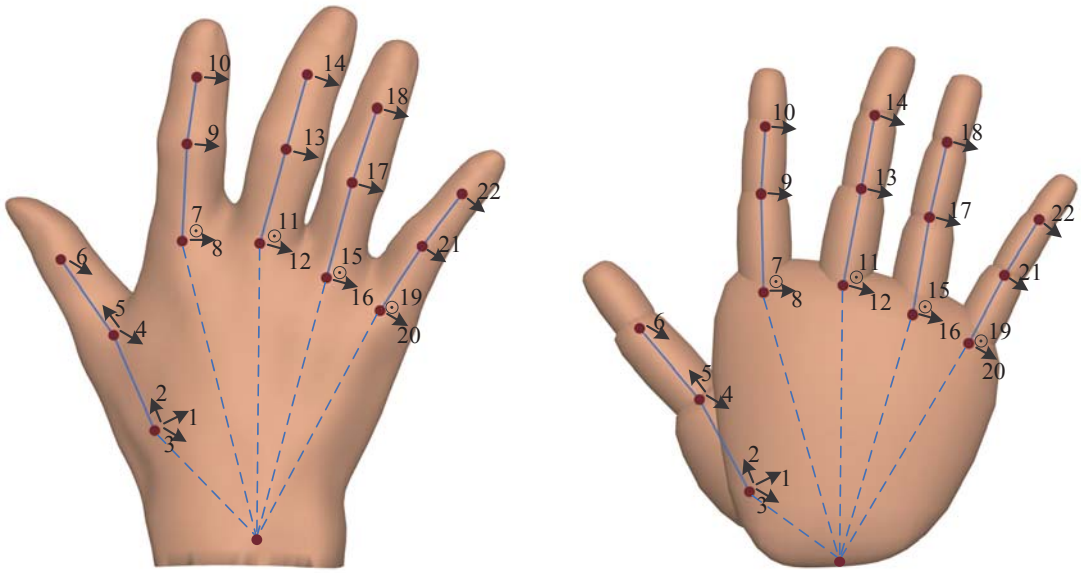
本文提出的交互运动捕捉方法与系统使用了基于物理约束的运动控制器，能够对手的运动、物体的运动以及它们之间相互的细微接触运动进行同时建模，核心思想来源于近年来发展的基于物理约束的人体全身建模[45,77–79]。与基于物理约束的人体全身建模针对的问题不同，我们主要着眼于高度灵活的手与物体交互

运动的物理建模，这其中最大的难题在于要对交互过程中的细微接触运动，以及频繁的接触点变化进行物理建模，而这在以前的研究中并没有提及。为了解决这个问题，我们提出了复合控制器的模型，其实质是融合交互过程中的比例- 微分（PD）控制器和虚拟接触力。

尽管还没有工作对交互过程中手与物体的细微接触运动进行过物理建模，但针对抓取运动，单纯地对手的姿态进行物理动力学建模却有大量的相关研究出现。Pollard[80]提出了一种从采集的抓取运动数据中，自动分析、提取满足物理约束的抓取运动控制器方法；Kry等人[81]则额外使用了接触力的采集数据库，同时分析出骨架节点的控制参数，通过调整骨架节点的控制参数，捕捉的运动数据可以很容易地映射到具有不同物理属性的物体上；Liu[25,36]则使用基于物理约束的优化方法仿真出符合动力学约束的手势运动。除此之外，也有研究人员尝试对手的肌肉、肌腱，以及它们之间的关联进行物理建模[82,83]，通过仿生学的方法来驱动手的三维模型，但这些方法能否有效地扩展用于手与物体的交互运动还是一个未知数。本文提出的无标记运动捕捉方法基于物理约束的动力学建模，无需事先采集任何运动数据，以及接触力的数据。于此同时，我们使用二维视频序列作为输入，利用复合控制器驱动带有骨架模型的手，计算虚拟生成的手势与输入的二维视频信号之间的误差，搜素最优的复合控制器。

最近，Ye等人[84]提出了一种新的虚拟视频制作方法，该方法首先分别捕捉手势运动以及物体的运动，然后将采集到的手与物体运动进行相应的处理、变换，使其满足交互运动物理约束的动力学方程，生成两者交互的虚拟运动视频。尽管他们提出的方法能够对手与物体的细微接触运动进行物理建模，但是由于没有输入的视频信号，他们方法生成的虚拟视频无法与真实的客观视觉信号进行匹配，因而也就无法进行量化评价。相反，本文提出的方法主要着眼于运动捕捉问题，不仅要求捕捉的运动数据满足物理动力学约束，还要求捕捉的运动数据与实际观测到的视觉信号相匹配。此外，本文提出方法的主要任务是寻找最优的复合控制器，而非实现满足物理约束的优化求解。

通过捕捉大量的交互运动多视点视频数据，我们验证了所提出交互运动捕捉方法与系统的有效性，特别地，我们分别测试了手与四种不同类型的物体进行交互运动的例子，这些物体包括：杯、球、魔方和棍。图3.1 显示了本文提出的方法和系统可以捕捉十分灵活的手与物体的交互运动。此外，我们还可以很容易地将捕捉的运动数据映射到具有不同属性的物体上，图3.2显示了两个捕捉的运动数据映射到三种具有不同几何外观的物体上。当然，为了验证系统的优越性能，我们也与当前最主流的标记点方法做了比较，实验表明，本文提出的固定多相机系统的无标记手与物体运动捕捉系统可以获得足够精确的结果与运动捕捉性能。



(a) 手的蒙皮模型　　(b) 手的近似刚体模型

图 3.3　手的模型拥有28个自由度，包含全局旋转和平移的6个自由度以及22个骨架节点自由度

3.3　问题归纳与概述

捕捉高精度的手与物体交互运动是非常困难的，其原因在于不仅需要对手的运动，以及物体的运动进行捕捉，还需要对它们之间的细微接触运动进行捕捉。而基于多视点视频的交互运动捕捉更为困难，不仅由于手上没有显著的特征点，还包括手与物体在交互过程中出现的严重遮挡。针对这些挑战，本节将从总体上介绍我们提出的方法与系统，以及需要解决的诸多问题。

3.3.1　状态空间

我们用一个28维的向量来描述手的三维姿态$\mathbf{q}^h \in R^{28}$，其包含了手的全局旋转和平移，以及每个骨架节点的旋转角度，图3.3 显示了每个骨架节点的自由度。骨架与三维模型表面点的关系使用线性混合蒙皮（Linear Blend Skinning, LBS）的技术来实现，有关骨架与三维模型表面点的线性化关系将在全文的附录B 中给出。物体的三维姿态用一个6 维的向量来表示 $\mathbf{q}^o \in R^6$。于是，在t 时刻，完整的状态为$\mathbf{q}_t = [\mathbf{q}_t^h, \mathbf{q}_t^o]$，其同时包含了手与物体的姿态。

3.3.2　问题建模

我们使用比例-微分（PD）的形式来建立所有骨架自由度的运动控制器，其中，每个骨架自由度的比例-微分控制器均用目标角度$\bar{\theta}$参数化。于是，手的每一

个骨架自由度所需的扭矩为

$$\tau = k_p(\bar{\theta} - \theta) + k_d\dot{\theta}. \tag{3-1}$$

式（3-1）中，θ和$\dot{\theta}$分别是当前手的骨架节点的角度和角速度；k_p和k_d是比例-微分控制器的增益和阻尼系数。需要指出的是，物体并不能主动产生任何力和扭矩，物体的运动受到手与物体之间的接触影响。原因在于物体总是被动地接受来自手的作用，而这个作用通过手与物体间的接触作用实现传递。

如果给定手与物体的初始姿态和速度$\mathbf{s}_0 = [\mathbf{q}_0, \dot{\mathbf{q}}_0]$，并且选定了合适的目标姿态$\bar{\mathbf{q}}_0, \cdots, \bar{\mathbf{q}}_{T-1}$，我们就可以利用运动控制器产生合适的骨架节点扭矩。于是，通过前向动力学的物理仿真，就可以仿真出一连串的随时间变化的手与物体姿态的运动结果$\mathbf{q}_1, \cdots, \mathbf{q}_T$。显然，这样仿真生成的运动结果符合物理动力学模型。

我们将基于固定多相机系统的无标记手与物体交互运动的捕捉建模成一个非线性优化问题，其目标是最小化仿真出的运动与输入的多视点视频之间的差别:

$$\min_{\mathbf{s}_0, \bar{\mathbf{q}}_0, \cdots, \bar{\mathbf{q}}_{T-1}} E(M(\mathbf{s}_0, \bar{\mathbf{q}}_0, \cdots, \bar{\mathbf{q}}_{T-1}), O). \tag{3-2}$$

这里，仿真的运动M不仅取决于初始状态$\mathbf{s}_0 = [\mathbf{q}_0, \dot{\mathbf{q}}_0]$，同时还取决于一系列的目标姿态$\bar{\mathbf{q}}_0, \cdots, \bar{\mathbf{q}}_{T-1}$，函数$E$用来计算仿真的运动$M$与观测到的图像$O$之间的不一致性，我们的目标即是搜索最优的运动控制器，使得仿真的运动与观测到的视频序列最匹配。为了实现这个目标，接下来，我们将介绍本文提出的方法需要解决的三个关键难题，它们分别是：基于图像的运动建模、交互运动的复合控制器和基于采样的控制器优化。在后续的小节中，我们将对这三个难题一一做详细地介绍。

基于图像的运动建模 在我们的方法与系统中，最关键的问题是如何定义一个合适的误差评价函数E来刻画仿真的运动M与观测图像O之间的不一致性? 这个问题非常具有挑战性，原因在于手与物体的交互运动通常具有十分严重的遮挡，并且手上没有十分显著的特征。此外，物体表面的单一纹理或漫反射表面也会使问题变得更为复杂。在下一小节中，我们将详细阐述我们方法与系统中基于图像的运动建模中使用的误差评价函数。

交互的复合运动控制器 为了成功实现手与物体的无标记运动捕捉，本文提出的方法与系统很大程度上取决于运动控制器的仿真能力，也就是运动控制器在多大程度上可以同时模拟手的运动、物体的运动以及它们两者之间的细微接触运动。然而，在实际操作过程中，直接搜索比例-微分控制器的目标姿态无法生成足够准确的接触力，以便同时驱动手与物体，使仿真运动结果符合观测的多视点图像数据。这是由于式（3-1）中定义的扭矩仅仅取决于手的目标姿态，而完全忽略

了物体的运动信息，这就使得仿真出的物体运动与实际图像很难匹配。为此，我们提出复合运动控制器的思想来对手与物体同时建模。为了实现复合运动控制器，我们引入了“虚拟”力的概念，并结合比例-微分控制器的优势，用于手与物体的交互运动物理仿真。“虚拟”力主要用来驱动物体，使物体运动的结果与实际观测的图像匹配。而在实际的仿真过程中，驱动物体运动的“虚拟”力被分解到手的各个关节中，这就构成了手的骨架关节的复合运动控制器，其中比例-微分控制器通过手的目标姿态运动提供，而“虚拟”力则是从物体的运动计算得到。

基于采样的控制器优化 最后一个挑战来自于怎样求解最优的控制器，使仿真的运动与实际观测图像数据吻合。事实上，最优运动控制器的求解即是对最优目标姿态的求解，也就是说，我们不仅需要搜索手与物体的初始状态$\mathbf{s}_0 = [\mathbf{q}_0, \dot{\mathbf{q}}_0]$，还需要搜索整个序列中手与物体的目标姿态$\bar{\mathbf{q}}_0, \cdots, \bar{\mathbf{q}}_{T-1}$。由于目标姿态的空间十分巨大，直接搜索求解几乎是不可能完成的事情。其次，无论手与物体之间是否存在接触，仿真函数均是时间离散的形式，因而，某个时刻目标姿态的微小变化均会产生仿真运动的巨大差异。本文提出的做法是首先利用多视点的视频数据重建出手与物体的骨架运动，然后在此运动的小范围内搜索合理的目标姿态。由于式（3-2）的梯度很难计算，并且基于采样的求解方法无需计算目标函数的梯度，我们使用了基于采样的方法来搜索最优的目标姿态。此外，我们还重点讨论了如何使用接触点的信息来加速采样的过程。

3.4　基于图像的运动建模

这一小节将将重点介绍误差评价函数，函数主要实现两个目标：（1）评价仿真运动与观测图像之间的好坏，正如式（3-2）中定义；（2）用来从多视点的视频数据中重建出手与物体的骨架运动，该骨架运动可作为后续的采样控制器参数优化的初始解。在本小节中，我们将首先介绍误差评价函数所需的先决条件，然后详细阐述评价函数的每一项，最后再介绍如何利用该误差评价函数重建出手与物体的骨架运动。

3.4.1　数据预处理

在基于固定多相机系统的无标记运动捕捉的准备阶段，我们首先用激光扫描仪分别得到手与物体的三维模型。然后，我们嵌入手的骨架，并建立手的蒙皮模型（如图3.3(a) 所示），手的各个时刻的三维姿态可以通过线性混合蒙皮技术，基于骨架变形的方法得到。此外，我们同时建立了手的近似刚体模型，也即是每个手指骨头用一圆柱体来近似表达，如图3.3(b)所示，在基于物理约束的动力学仿真

过程中（实现中使用ODE[①] 引擎），我们可以使用手的近似刚体模型，很容易地实现碰撞检测的任务。

3.4.2　误差评价函数

我们使用合成分析（analysis-by-synthesis）的方法来评价仿真的姿态$\mathbf{q}_t, t = 1, \cdots, T$与观测图像$\mathbf{I}_t^v, t = 1, \cdots, T$之间的不一致性，其中$v$是相机视角的序号。在所有的实验中，我们使用6个固定视点的相机视频作为方法与系统的输入。

给定一个假设的姿态$\mathbf{q}$，我们可以得到与之对应的变换矩阵$T_{\mathbf{q}}$，将该变换矩阵作用到手的蒙皮模型的每一个顶点，以及物体的每一个顶点上，就可以同时得到与假设姿态$\mathbf{q}$对应的手与物体的三维模型，具体计算方法可以参见附录B。由于固定多视点相机的内外参数可以事先获得[85]，我们可以使用视点依赖的纹理投影技术[86]，将三维模型投影到每个相机视角对应的二维平面，得到假设姿态生成的虚拟图像。值得一提的是，有关视点依赖纹理投影所需要的初始值将会在后面一小节中给出，事实上，为了避免跟踪过程中的误差累积问题，我们的纹理映射图在第一帧中就已经完全确定，并且在跟踪过程中不随时间变化。

评价假设姿态的虚拟图像与真实观测图像之间的不一致性有很多种做法，一种最直接的想法是衡量两者颜色之间的差别。但是，在实验过程中，我们发现仅仅使用颜色信息不足以刻画手与物体的交互运动，原因在于手、甚至物体在某些时刻会缺乏明显的颜色区分度。此外，物体的表面很多时候也不是漫反射的特性，单纯地颜色评价指标无法客观地展现物体的运动。为此，我们还考虑了轮廓和边缘信息作为它们不一致性的评价指标。

于是，假设的交互运动姿态$\mathbf{q}_t$和观测图像$\mathbf{I}_t^v, v = 1, \cdots, 6$之间的误差评价函数定义为：

$$E = w_{silh} \sum_v E_{silh}^v + w_{color} \sum_v E_{color}^v + w_{edge} \sum_v E_{edge}^v. \tag{3-3}$$

式（3-3）中，$E_{silh}, E_{color}, E_{edge}$分别用来评价轮廓、颜色和边缘的不一致性；$w_{silh}, w_{color}, w_{edge}$用来控制三项之间的权重关系。在后续所有的实验中，$w_{silh} = 1.0, w_{color} = 0.6, w_{edge} = 1.0$。

3.4.2.1　轮廓项

这一项用来保证假设姿态渲染出的虚拟图像轮廓与观测图像的轮廓相匹配。一般而言，轮廓图是一个二值图，其中背景像素为0，前景像素为1。为了使系

① http://www.ode.org/

(a) 原始图　(b) 自动抽取的轮廓图　(c) 渲染的轮廓图

图 3.4　轮廓项使用的轮廓图

统能够自动地抽取出每一帧手的轮廓图，我们将手的轮廓像素用概率模型来表示。值得说明的是，该概率模型的参数利用多视点图像第一帧的颜色直方图估计得到。我们将颜色直方图建模成超高斯混合模型（Super-Gaussian Mixture Models, SGMM），因此，图像中每个像素属于手的概率可以定义为：

$$H(x) = \sum_j \lambda_j \exp(-|x - \mu_j|^{0.8}), \lambda_j \geq 0, \sum_j \lambda_j = 1. \tag{3-4}$$

式（3-4）中，j是混合成分的序号；模型参数λ_j 和μ_j可以使用第一帧的手的像素，通过期望-最大（Expectation-Maximum, EM）算法，自动地估计出来。

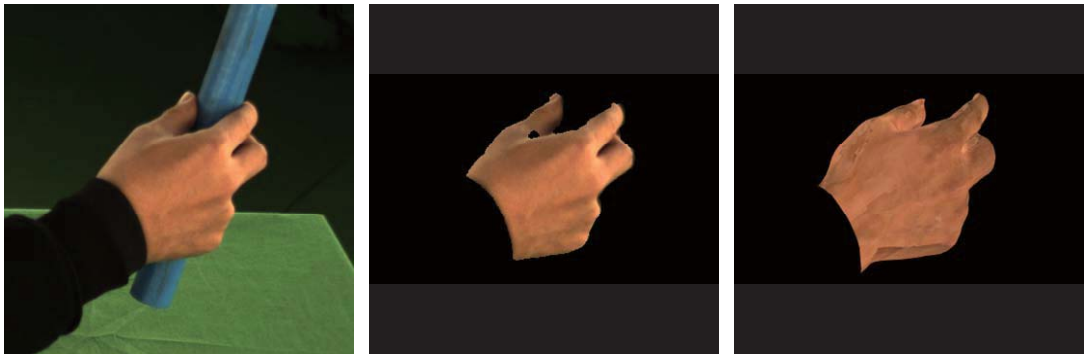
针对每一个观测的图像I_t^v，我们使用前述的高斯混合模型，可以计算出每个像素属于手的概率值。通过设定合适的阈值，我们可以自动抽取出手的轮廓图S_o。相反，针对假设姿态渲染生成的虚拟图，可以直接渲染出手的轮廓图，定义为S_r。图3.4显示了自动抽取的轮廓图与渲染的轮廓图。因此，每张输入图像的轮廓项可以定义为：

$$E_{silh}^v = \frac{\sum(S_r \cap \overline{S}_o)}{\sum S_r + \epsilon} + \frac{\sum(S_o \cap \overline{S}_r)}{\sum S_o + \epsilon} \tag{3-5}$$

上式中，$\overline{S}_o = 1 - S_o, \overline{S}_r = 1 - S_r$，它们分别是观测图轮廓$S_o$和渲染图轮廓$S_r$的补图；$\epsilon$是一个很小的常数以避免除数为0；$\sum$将所有的像素加和。直观地讲，本文定义的轮廓项用来计算渲染的虚拟轮廓与观测图轮廓之间的非重叠像素的比率，因而，最小化该项可以最大化两者之间像素的重叠率。另外值得一提的是，本文提出的轮廓项仅考虑手的轮廓图，而没有考虑物体的轮廓图，其原因在于物体的轮廓并不能像手一样，可以很有效地自动分割出来。

3.4.2.2　颜色项

我们使用视点依赖的纹理映射技术渲染出每个视点的纹理图R_t^v，图3.5分别显



(a) 原始图　(b) 手的原始颜色图　(c) 手的渲染纹理图

图 3.5　颜色项使用的纹理图

示了手的观测图像与渲染图像。这一项用来计算观测图像I_t^v与渲染图像R_t^v之间的颜色差别，一种最常用的做法是基于对应像素差别的平方和（sum of squares）来评价两张图像之间的整体差别。然而，这样一种简单地逐像素计算误差的平方和很容易失效，原因在于：（1）纹理映射图在初始时刻很难精准地对齐好；（2）存在镜面反射的物体在运动过程中的纹理与初始时刻的纹理会产生较大的差别。此外，在交互运动时，手与物体之间的遮挡阴影也会严重影响纹理图。

为此，我们提出一种鲁棒的颜色项来衡量观测图与渲染图之间的差别

$$E_{color}^{v} = \sum_{i} \min\left(\min_{j \in \mathbf{N}(i)} \left(w_{i,j} |R_i^v(i) - I_t^v(j)| \right), T \right). \tag{3-6}$$

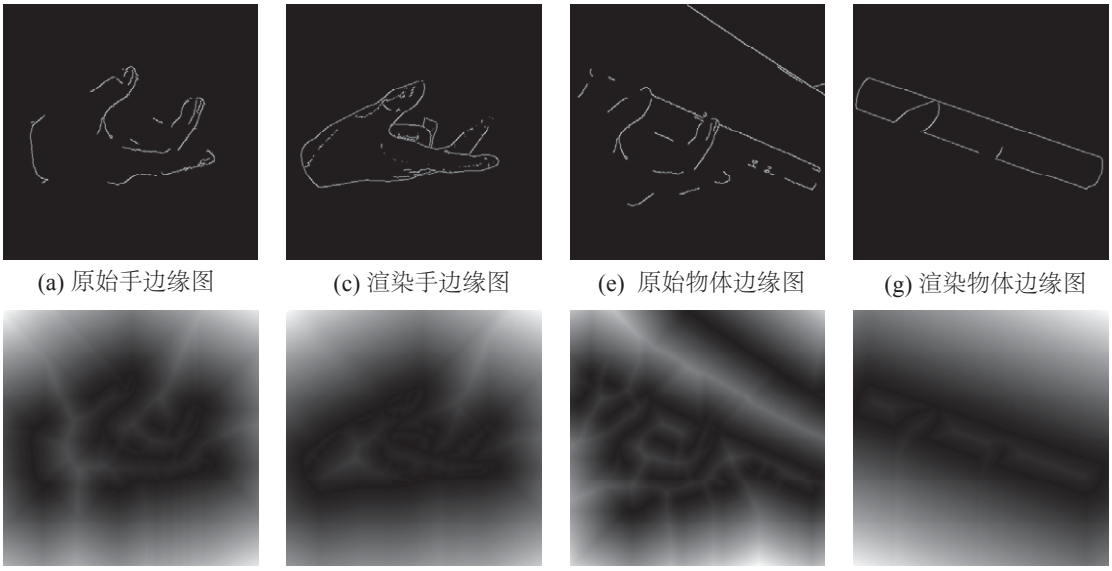
特别地，式（3-6）对渲染图像R_t^v中的每一个像素位置i，在原始观测图像中对应像素位置i的小窗口中每一个像素j，我们计算窗口中所有像素值$I_t^v(j), j \in \mathbf{N}(i)$与$R_t^v(i)$的差别，然后选择该窗口中误差最小的值作为像素$i$的颜色误差。在我们的实验中，窗口大小均选择为$5 \times 5$；我们还对窗口中的不同像素设置了一定的权重$w_{i,j}$，该权重与两像素的空间位置相关，定义为$e^{-\|i-j\|^2}$；此外，为了避免镜面反射表面带来的较大误差，我们还对颜色误差设置了一个阈值T，实验中$T = 0.25$。

3.4.2.3　边缘项

这一项用来计算渲染图像边缘图与观测图像边缘图之间的差别。我们使用二值图来表达边缘图，其中边缘的像素设置为1，非边缘的像素设置为0。由于边缘不仅与颜色纹理有关，还与它们的三维几何以及光照有关，因此引入边缘项可以提高跟踪结果的稳定性，尤其是对那些缺少纹理细节的物体而言。

特别地，我们的边缘项定义为

$$E_{edge}^{v} = \frac{\sum e_r^{hand} \cdot d_o^{hand}}{\sum e_r^{hand} + \epsilon} + \frac{\sum e_o^{hand} \cdot d_r^{hand}}{\sum e_o^{hand} + \epsilon} + 2\frac{\sum e_r^{object} \cdot d_o^{object}}{\sum e_r^{object} + \epsilon}. \tag{3-7}$$



(a) 原始手边缘图　(c) 渲染手边缘图　(e) 原始物体边缘图　(g) 渲染物体边缘图

(b) 原始手距离变换图　(d) 渲染手距离变换图　(f) 原始物体距离变换图　(h) 渲染物体距离变换图

图 3.6　边缘图以及距离变换图

其中，e_r^{hand} 和 e_o^{hand} 分别是手的渲染图像边缘图和观测图像边缘图，实际上，我们使用了canny算子来抽取他们的边缘图；d_r^{hand} 和 d_o^{hand} 则是边缘图的距离变换图，在距离变换图中的每一个像素值的大小取决于最近边缘的距离，也就是说距离变化图可以很鲁棒的对边缘图的差别进行比较。图3.6分别显示了手与物体的边缘图与距离变换图。

直观地来讲，式（3-7）的第一项和第二项双向描述了手的渲染图像边缘图与观测图像边缘图的差别。相反，第三项中仅仅单向计算了物体的渲染图像边缘图与观测图像边缘图之间的差别，这是由于观测图像中物体的轮廓图并不容易提取出来，因此我们仅仅考虑了单向的物体边缘图之间的差别。与轮廓项一致，ϵ也是一个很小的常数，主要是为了避免除数为0的情况出现。

3.4.3　骨架运动重建

现在，我们将讨论如何使用固定多相机系统的多视点视频数据，计算式（3-3）的姿态误差值，进行手与物体的骨架运动重建。这一步对本文提出的系统与方法相当关键，重建的结果将作为后续物理仿真的目标姿态的初值。

3.4.3.1　姿态跟踪

我们的骨架运动重建方法将顺序地进行手与物体的三维姿态估计。给定手与物体的初始结果，通过最小化式（3-3），我们可以估计出每一帧的手与物体的三维姿态。为了保证跟踪结果的平滑性，我们引入一个平滑项来保证运动跟踪结果

不会出现较大的跳变，该平滑项定义为$\|\mathbf{q}_t - 2\mathbf{q}_{t-1} + \mathbf{q}_{t-2}\|^2$。我们使用互动模拟退火（Interacting Simulated Annealing, ISA）算法来搜索每一帧的最优值，与模拟退火算法一致，基于互动粒子系统的互动模拟退火算法可以收敛到全局最优解。简单来讲，我们首先根据前一帧的跟踪结果，随机生成一些姿态，这些姿态统称为粒子，通过式（3-3）评价每一个粒子对应的代价之后，可以给每个粒子赋予一定的权重，根据粒子之间的权重关系，我们可以重新采样得到新的粒子，以上过程不停地重复直至收敛。于是，最终的姿态也就是所有姿态粒子的加权求和。在我们的实验中，迭代的次数为25，每一步迭代采用300 个姿态粒子。

3.4.3.2　姿态初始化

由于顺序进行姿态跟踪需要姿态初值，这一小节将讨论如何得到第一帧的姿态。实际上，初始值通过将三维模型与多视点的图像进行对齐得到。首先，手与物体的全局姿态通过分别将手与物体的可视凸壳（Visual Hull, VH）与三维模型进行对齐，使用迭代最近邻的方法得到粗略的结果。然后，我们将三维模型与图像的精确对齐描述为一个优化问题，由于还没有对齐的纹理图，优化方程仅包含式（3-3）中的轮廓项和边缘项，值得一提的是，初始时刻手与物体相互分离，分别得到手与物体的轮廓非常容易。与姿态跟踪采用的方法一致，我们采用互动模拟退火的算法来求解该对齐优化问题，分别得到手与物体的初始姿态。

3.5　交互的复合运动控制器

这一小节主要讨论手与物体的仿真运动，以及它们发生交互运动时，细微接触的动力学建模问题。与以往的研究工作不同，为了考虑物体对手的反作用影响，我们提出了复合运动控制器的模型。

3.5.1　比例-微分控制器

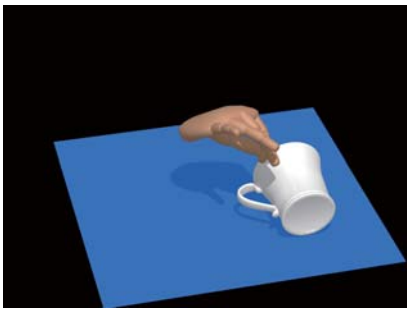
为了实现物理仿真的任务，我们对手上的任一骨架节点均设置了一个比例-微分控制器。于是，在任意时刻，手上每个骨架节点内在的扭矩可以用下式计算

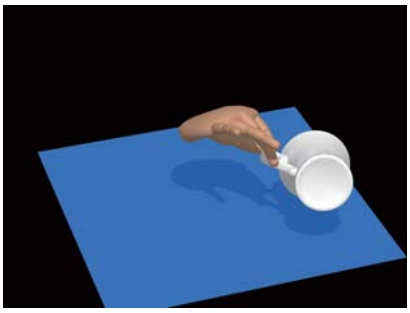
$$\tau_{pd} = k_p\left(\overline{\mathbf{q}}^h - \mathbf{q}^h\right) + k_d\dot{\mathbf{q}}^h, \tag{3-8}$$

式（3-8）中，$\mathbf{q}^h$和$\dot{\mathbf{q}}^h$分别是当前时刻手的骨架节点的角度和角速度，$\overline{\mathbf{q}}^h$是手的目标姿态，比例- 微分控制器的任务是驱动手从当前姿态$\mathbf{q}^h$往目标姿态$\overline{\mathbf{q}}^h$运动。

理论上来讲，直接搜索比例-微分控制中手的目标姿态$\overline{\mathbf{q}}^h$可以产生合适的接触点扭矩τ_{pd}，然后仿真出手与物体的交互运动。但在实际操作过程中，即使给出了

(a) 原始图像





(b) 比例-微分控制器仿真结果　(c) 复合运动控制器仿真结果

图 3.7　比例-微分控制器与复合运动控制器的仿真结果比较

手的真实运动路径，搜索交互运动中合适的比例-微分控制器的目标姿态仍然是一件非常困难的事情。以抓取桌上的杯子为例，在手的参考运动附近搜索手的目标姿态可能对“裸手”，或者轻质的慢速物体运动有效，但是对于正常的物体运动，通常需要较大的骨架节点力和扭矩来抵消物体运动的影响。换句话讲，抓取物体所需要的目标姿态与单一手的目标参考姿态有很大的不同。因此，试图只在手的参考姿态附近，直接搜索交互运动的目标姿态是非常容易失效的，如图3.7所示。

这就启发我们使用一个“虚拟”力和扭矩来抵消物体的惯性运动力以及重力作用。需要指出的是，“虚拟”力并不能直接作用于被动的物体上，相反，“虚拟”力需要通过手与物体的接触作用传递。因而，我们需要将等效的“虚拟”力分解到手指的各个骨架关节中。在我们的方法中，我们使用“虚拟”力对应的手指关节扭矩，驱动物体按照目标运动轨迹运动，图3.8显示了物体运动到预设位置所需的“虚拟”力和扭矩，产生与“虚拟”力和扭矩同等效果的增广接触力和扭矩，以及分解出来的手指关节所需的额外力和扭矩。

3.5.2　“虚拟”力和增广接触力

为了驱动物体沿着预设的路径运动（这个预设的路径不仅包括参考位置$\overline{\mathbf{q}}^{o,p}$，还包括参考旋转$\overline{\mathbf{q}}^{o,r}$），我们引入了“虚拟”力的思想。实际仿真中，由于“虚拟”力是通过接触点力实现的，本小节将重点讨论如何计算合适的接触点力和扭

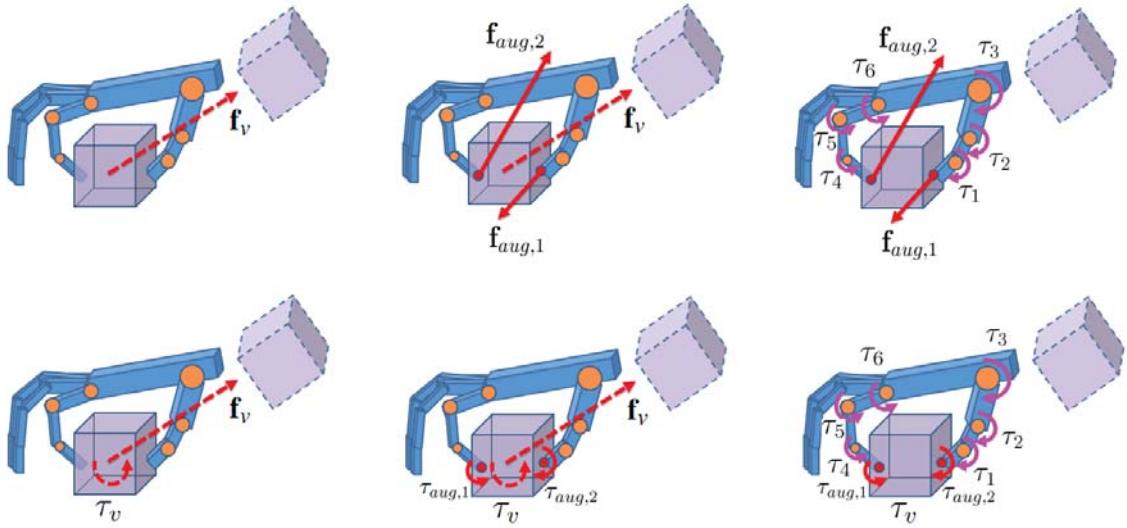


图 3.8　使物体运动到目标姿态所需的“虚拟”力和扭矩、增广接触力和扭矩以及手指关节力和扭矩的图示。（左图）“虚拟”力$\mathbf{f}_v$和扭矩τ_v；（中图）为了实现“虚拟”力和扭矩同等效果的增广接触力$\mathbf{f}_{aug,i}$和增广接触扭矩$\tau_{aug,i}$；（右图）将增广接触力和扭矩分解到各个手指关节的力和扭矩

矩来实现等效的“虚拟”力和扭矩。

首先，“虚拟”力$\mathbf{f}_v$的作用是使物体沿着预设的路径运动。因此，我们可以定义一个虚拟的比例-微分控制器，来实现“虚拟”力的作用（如图3.8 中第一列第一行）

$$\mathbf{f}_v = k_p\left(\overline{\mathbf{q}}^{o,p} - \mathbf{q}^{o,p}\right) + k_d\dot{\mathbf{q}}^{o,p} - m_o\mathbf{g}. \tag{3-9}$$

式（3-9）中，$\mathbf{q}^{o,p}$和$\dot{\mathbf{q}}^{o,p}$分别是物体当前的位置和线速度；$\overline{\mathbf{q}}^{o,p}$是物体的目标位置。“虚拟”力的任务是驱动物体从当前位置$\mathbf{q}^{o,p}$运动到目标位置$\overline{\mathbf{q}}^{o,p}$。此外，在该式中，我们引入了物体的重力，其目的是去除物体质量对仿真运动的影响。

与手的仿真不同，我们无法直接利用式（3-9）中定义的虚拟比例-微分控制器对物体进行仿真，原因在于物体的运动是被动的，并且它的运动完全取决于手与物体之间接触点的力和扭矩。因此，根据上式定义的“虚拟”力，我们必须计算出等效的接触点力，我们将其称为增广接触力$\mathbf{f}_{aug}$，如图3.8中第二列第一行所示。计算等效的增广接触力面临着两个挑战：（1）当手与物体存在多个接触点时，解并不唯一。例如，当手用四个指头旋转魔方时，会有很多同样的接触点力使得物体具有同样的运动；（2）增广接触力必须要在摩擦椎中，并且需要保持和手与物体之间的接触点的摩擦方向一致。

为了解决上述提到的两点挑战，我们首先使用ODE引擎中的线性互补问题（Linear Complementarity Problem，LCP）求解工具包，计算出当前时刻手与物体

的初始接触点力。然后用该接触点力来正则化需要求解的增广接触力，也就是说，最终求得的增广接触力不能和初始接触点力相差太远。更为具体地讲，在任一时刻，我们首先用ODE 引擎检测出手与物体的所有接触点，然后使用线性互补问题求解工具包计算出每个接触点的接触点力$\mathbf{f}_{c,i}, i = 1, \cdots, K$，其中$K$是接触点的数目。通过惩罚与初始接触点力$\mathbf{f}_{c,i}, i = 1, \cdots, K$的变化程度，我们可以正则化增广接触力$\mathbf{f}_{aug,i}, i = 1, \cdots, K$的解空间。此外，为了保证增广接触力$\mathbf{f}_{aug,i}, i = 1, \cdots, K$在接触点的摩擦椎里面，我们进一步限制增广接触力与初始接触点力的方向相同，也即是$\mathbf{f}_{aug,i} = w_i \mathbf{f}_{c,i}, w_i \geq 0, i = 1, \cdots, K$。

在我们方法的实现中，我们将增广接触力的求解，形式化成一个优化问题。通过求解一个动态规划问题，我们可以得到最优的增广接触力：

$$\begin{aligned} \underset{w_1, \cdots, w_K}{\arg\min} \quad & \|\mathbf{f}_v - \sum_{i=1}^{K} w_i \mathbf{f}_{c,i}\|^2 + \lambda_1 \sum_{i=1}^{K} (w_i - 1)^2 \\ s.t. \quad & w_i \geq 0, i = 1, \cdots, K. \end{aligned} \tag{3-10}$$

式（3-10）中的第一项保证增广接触力$\mathbf{f}_{aug,i} = w_i \mathbf{f}_{c,i}$的合力与虚拟力$\mathbf{f}_v$一致；第二项为正则项，其作用是惩罚增广接触力与LCP求解工具包的初始接触点力的差别；不等式约束用来保证增广接触力与初始接触点力具有相同的方向；权重$\lambda_1 = 1$。

到目前为止，我们仅仅考虑了使物体产生平动的“虚拟”力，下面我们将进一步考虑使物体发生转动的“虚拟”力矩τ_v。类似于“虚拟”力的思想，“虚拟”力矩的任务是使物体按照预定的旋转运动，也就是说物体可以从当前的旋转姿态$\mathbf{q}^{o,r}$运动到目标旋转姿态$\overline{\mathbf{q}}^{o,r}$。同样地，我们需要估计出每个接触点的增广接触力矩$\tau_{aug,i}, i = 1, \cdots, K$，如图3.8第二列第二行所示。此外，增广接触力矩需要补偿每个接触点处增广接触力产生的力矩。

我们将每个接触点的增广接触力矩建模成扭转扭矩（torsional torque）：$\tau_{aug,i} = l_i \vec{n}_i$，其中，$l_i$是第$i$个增广接触扭矩的大小，$\vec{n}_i$是第$i$个接触点对应的三维曲面的法向，模为1。在机器人和人机交互领域[87]，扭转扭矩被广泛地用于同一部位不同接触力的扭矩建模。从数学角度考虑，作用于同一表面的多个接触点力的效果可以看作是作用于压力中心（Center of Pressure，COP）的接触点力和扭转扭矩的组合。值得指出的是，压力中心通常需要保证垂直合力产生的扭矩为0，即垂直合力需要穿过压力中心，这样，压力中心的多个接触点力产生的扭矩也即是多个摩擦力产生的合成扭转扭矩。此外，使用扭转扭矩的另外一个好处是我们可以对挠性问题的接触扭矩进行建模，这样手与物体的细微接触运动也就可以进行合理、有效的表达。

同样地，我们将增广接触扭矩也描述为一个动态规划问题，也即：

$$\underset{l_1,\cdots,l_K}{\arg\min} \quad \| \sum_{i=1}^{K} \left(\mathbf{r}_i \times \mathbf{f}_{aug,i} + l_i \vec{n}_i \right) - \tau_v \|^2 + \lambda_2 \sum_{i=1}^{K} l_i^2, \tag{3-11}$$

其中，$\mathbf{r}_i$是物体的质心到第i个接触点的向量。第一项刻画了“虚拟”扭矩与增广接触力和扭矩的组合扭矩之间的差别，第二项为正则项；权重$\lambda_2 = 1$。

根据式（3-10）和式（3-11），接触点力和扭矩可以同时优化，也可以依次优化。在我们的实验中，我们首先优化得到接触点力的大小，然后再优化扭矩。

3.5.3　复合运动控制器

一旦我们计算出了增广接触力和扭矩，我们可以使用雅克比（Jacobian）变换矩阵，得到手的每个骨架关节对应的，驱动物体从当前姿态到目标姿态$\overline{\mathbf{q}}^o$ 所需的增广骨架节点扭矩，如图3.8第三列所示：

$$\tau_{hand,v} = \sum_{i=1}^{K} J^T \mathbf{f}_{aug,i} + A^T \tau_{aug,i}. \tag{3-12}$$

式（3-12）中，$\tau_{hand,v}$是驱动物体沿着目标轨迹运动所需的手指关节的扭矩；矩阵J和A分别是当前姿态下，将手指关节角速度映射为第i个接触点线速度和角速度的雅克比矩阵。

这样，我们的手与物体交互运动的控制器包含两部分：（1）将当前手势驱动到目标手势的比例-微分控制器τ_{pd}；（2）将物体从当前姿态驱动到目标姿态对应的增广骨架节点扭矩$\tau_{hand,v}$，也即是

$$\tau_{joint} = \tau_{pd} + \tau_{hand,v}. \tag{3-13}$$

于是，我们建立了最终的扭矩τ_{joint}与当前时刻的手与物体姿态$\overline{\mathbf{q}}_t = [\overline{\mathbf{q}}^h, \overline{\mathbf{q}}^o]$ 的关联，这样一个复合运动控制器的表达不仅考虑了手的运动，物体的运动，也考虑了两者之间的细微接触运动，其中包括每个接触点的位置、大小、摩擦力乃至物体的重力影响。因此，它可以用来同时跟踪手与物体的交互运动。

3.6　基于采样的控制器优化

在这一小节中，我们将重点阐述如何搜索最优的运动控制器，使得仿真的手与物体交互运动结果与真实的图像匹配，也就是说，我们不仅需要搜索出初始状态$\mathbf{s}_0 = [\mathbf{q}_0, \dot{\mathbf{q}}_0]$，还需要搜索出所有帧的目标姿态$\overline{\mathbf{q}}_0, \cdots, \overline{\mathbf{q}}_{T-1}$。为了降低搜索空间的维度，一种思想是在重建的骨架运动附近搜索最优的目标姿态，同样地，有关

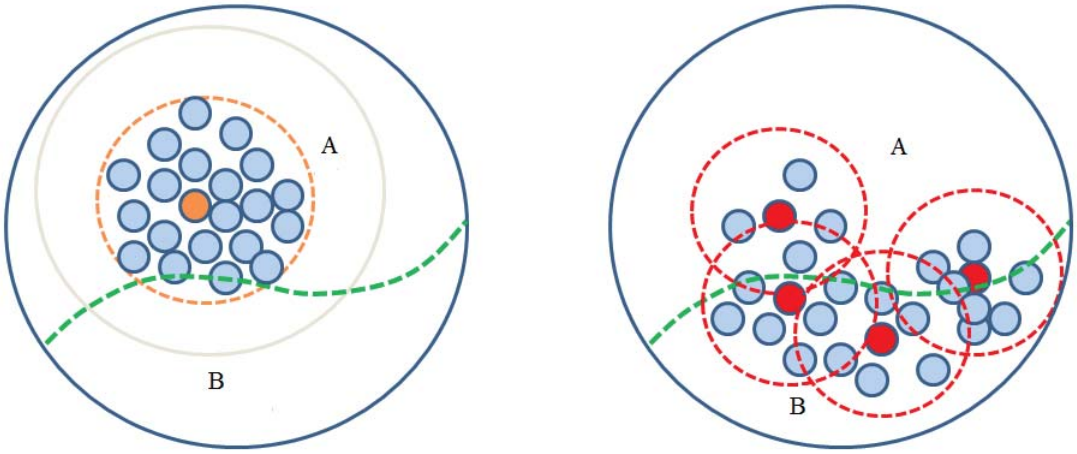


图 3.9　基于接触点的采样，图中的A和B分别代表物体和可能的手指骨头接触的两种不同接触状态区域

初始状态的搜素也是在参考骨架运动的附近进行。很显然，基于采样的搜索方法非常适合本问题的求解，原因在于手与物体的交互运动中存在着大量未定义的接触点，而有关这些未定义的接触点的梯度求取是很难实现的。

这种思想简单来讲即是，在任一t时刻，使用重建的骨架运动作为目标姿态的初始值，然后在该初始值附近进行采样，得到一些新的目标姿态，利用这些新的目标姿态进行物理仿真，于是可以得到一系列仿真运动的状态，通过式（3-3）定义的目标函数，选择最优的状态并作为$t + \delta t$时刻的初始值，重复上述过程直至遍历所有的帧，最终可以得到与图像最匹配的仿真运动。在我们的实验中，前后两帧的时间步长设为0.05s，而仿真时间步长设为0.0005s。

由于直接从多视点图像中重建的骨架运动噪声很大，并且极有可能重建出错误的手与物体的接触运动。以抓取杯子把手为例，多视点视频重建出的骨架参考运动几乎是完全错误的，如图3.13所示。上述简单的随机采样方法没有考虑手与物体的接触点作用，因此直接在参考姿态的附近进行采样会产生大量具有错误接触点信息的样本，很有可能无法实现交互运动的捕捉任务。为此，我们需要提出其他的方法来解决该问题。

3.6.1　基于接触点的采样

本小节将讨论如何在采样过程中考虑接触点的信息，这里面临的最大挑战在于精确的接触点检测几乎是不可能完成的任务，其原因在于重建的骨架运动噪声很大。为了解决这个问题，我们采取的做法是将给定误差范围之类的所有点都认为是潜在的接触点，然后在采样过程中，枚举所有的接触可能性，然后针对不同

的接触类型进行采样。

给定从重建的骨架运动中估计的初始目标姿态，我们首先计算物体与手的每一个骨头之间的最近距离，以此来确定物体与每一个手的骨头之间是否存在接触点。如果这个最近距离大于一定的阈值，我们就认为物体与手指骨头之间没有接触点，否则，则存在一个潜在的接触点。对于任何一个潜在的接触点，其是接触点与不是接触点的概率均相等。这就允许我们在一定的阈值之类对噪声很大的骨架运动，枚举出所有可能的接触状态。也就是说，对于k个潜在的接触点，一共有2^k个接触状态，每一个接触状态都有相应的概率值。基于接触点的采样方法首先根据接触状态的概率值，采样出一定数量的接触状态，然后针对这些接触状态，我们使用逆向工程技术，修正初始目标姿态，使得修正后的新姿态与采样的接触状态一致接着，这些新姿态将作为生成目标姿态的“种子”姿态，也即是说，我们在这些新姿态周围进行采样，得到目标姿态，而不是在原来的重建出的骨架运动姿态附近采样。图3.9 中的左图表明，在不考虑接触点状态的情况下，直接在高噪声的参考姿态附近进行随机采样，采样得到的结果可能会产生错误接触状态的结果。“橙色”和“蓝色”的圆圈分别代表从骨架跟踪得到的参考姿态，以及在其附近采样出的目标姿态；而图3.9 中的右图表明，基于接触点的采样方法首先基于手与物体潜在的接触状态，生成少量的“种子”姿态，然后在每个“种子”姿态附近采样得到目标姿态。“红色”的圆圈代表基于接触点采样方法得到的“种子”姿态。显然，这样一种基于接触点的采样方法考虑了接触点的状态属性，可以加速采样收敛的过程。

3.6.2　样本选择

接下来的问题是如何选择采样的样本。事实上，每一个样本都对应于式（3-2）的目标函数值。一种贪心的策略是在采样的每一步均保留函数值最小的样本，然而这样一种选择策略很容易陷入局部极值，这是因为每一帧的函数最小值对应的样本，并不能保证在连续帧上都具有同时的最小值。为此，我们使用文献[88]提到的样本选择策略。首先，我们将样本的能量从大到小排列，然后舍弃50%能量较大的样本，我们用E_{low}和E_{high}来指代剩余的样本中，能量最低样本和能量最高的样本，并且用m表示我们需要保留的样本数目。采样出的m个样本$s_0, \cdots, s_{m-1}$需要满足两个准则：（1）采样的最终样本应能覆盖$[E_{low}, E_{high}]$范围；（2）应尽可能地采样到能量靠近E_{low}的样本。为此，我们使用多项式的采样策略，即：$s(x) = (E_{high} - E_{low})x^6 + E_{low}$，其中，$x = i/m, i = 0, \cdots, m-1$。我们选择那些能量最接近$s(x)$ 的样本。经过试验发现，每一步使用1500个样本并且保留$m = 150$个样本，可以产生足够好的采样结果。

表 3.1　物理仿真参数

节点名	k_p	k_d	节点名	k_p	k_d
全局平移	400.0	50.0	全局旋转	600.0	50.0
拇指$_1$	10.0	0.3	拇指$_2$	8.0	0.2
拇指$_3$	4.0	0.4	食指$_1$	4.0	0.2
食指$_2$	3.0	0.1	食指$_3$	2.0	0.1
中指$_1$	4.0	0.4	中指$_2$	4.0	0.1
中指$_3$	2.0	0.1	无名指$_1$	4.0	0.4
无名指$_2$	3.0	0.1	无名指$_3$	2.0	0.1
小指$_1$	4.0	0.4	小指$_2$	3.0	0.1
小指$_3$	2.0	0.1			

3.7　实验结果与分析

通过捕捉多个手与物体交互运动的例子，我们验证了本文所提方法和系统的有效性。在所有的实验中，我们使用帧率为20fps，分辨率为1024×768 的6 个固定视角的同步相机来拍摄多视点的视频序列，仿真运动使用ODE引擎来实现。图3.10 显示了一些帧的运动结果。表3.1 显示了运动控制中使用的所有仿真参数，其中，手指头名称的下标按照指根到指尖的顺序。例如：拇指$_1$是拇指根部的一个节点；拇指$_3$则是最接近拇指尖的一个节点。所有的仿真结果均使用同样的控制器参数，即k_p 和k_d。

3.7.1　实际数据测试

我们将本文提出的固定多相机系统的无标记手与物体的交互运动捕捉系统用于4种不同物体的交互运动捕捉，它们分别是：球、杯子、魔方和木棍。我们假设物体具有均匀的质量分布，通过离散化物体的实心三维模型得到物体的体素结构，我们将离散的体素进行积分处理可以计算得到每一个物体的转动惯量。图3.10显示我们的系统可以很好地捕捉手与物体的交互运动，图3.10的第四行也表明我们的系统能够捕捉手与物体之间的细微接触运动。除此之外，我们还对纹理特征丰富以及无纹理特征的物理进行了捕捉，其中，球和木棍具有几乎一致的纹理颜色，而魔方和杯子则具有复杂的纹理结构，实验结果表明物体的纹理特征对系统的稳定性影响不大。系统也表明物体表面的反射特性对捕捉结果影响也不是很大，例如杯子、魔方和木棍的表面具有比较明显的非漫反射属性，而球的表面则具有明显的漫反射属性。此外，我们的系统还可以捕捉复杂背景下的快速运动，如图3.10第五行所示，其中，从左到右依次是：原始图；附在原始图上的运动捕捉

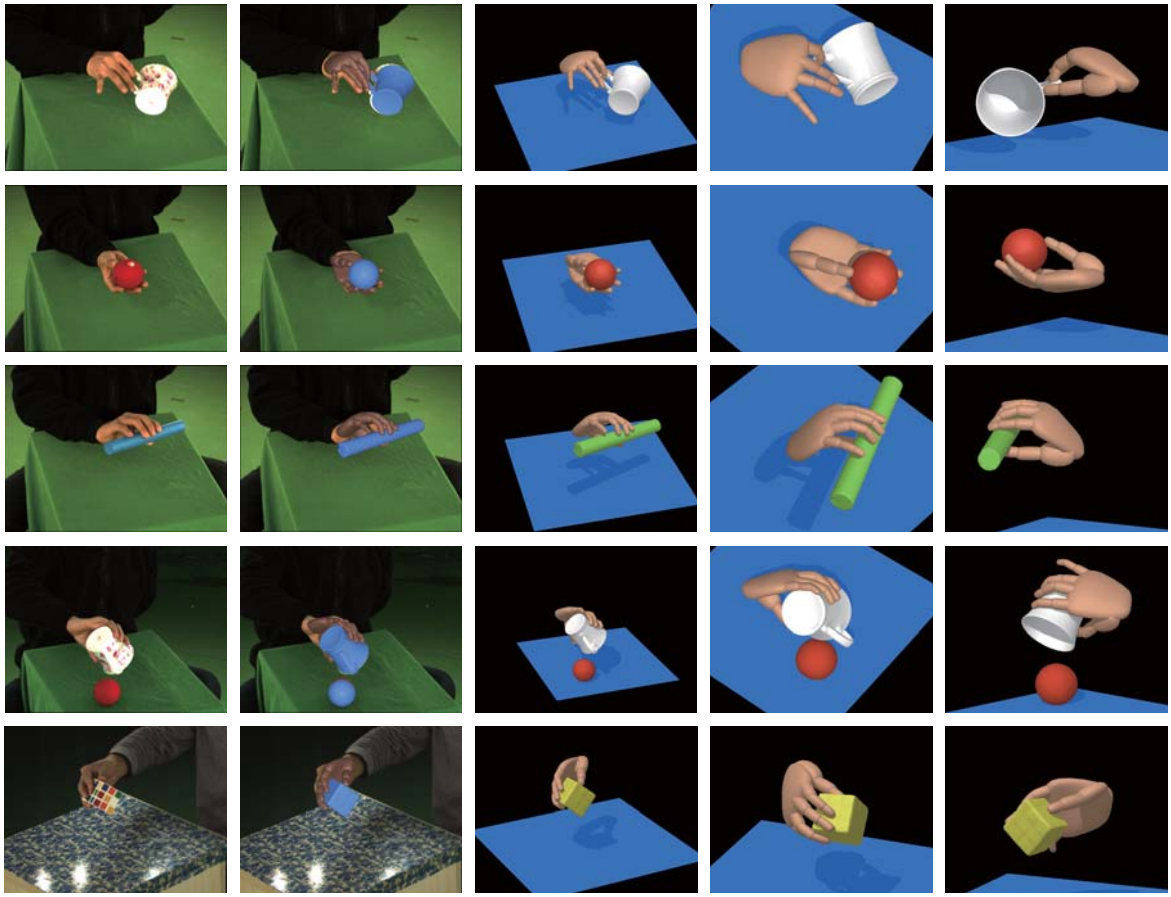


图 3.10　固定多相机系统的无标记手与物体交互运动捕捉结果

结果；相同视点的运动捕捉结果；以及其他两个视角的运动捕捉结果。

3.7.2　运动捕捉结果泛化

从多视点视频中恢复出基于物理约束的运动控制器的一大好处是可以直接将重建的运动控制器作用于具有不同物理属性的物体上。在我们的实验中，我们将捕捉的运动映射到具有完全不同几何结构的其他物体上，获得很好的运动映射结果，如图3.2所示。此外，我们还可以改变运动过程中其他物理参数，例如摩擦系数，获得新的虚拟运动。

3.7.2.1　运动映射到新的物体

交互运动的映射特别具有吸引力的地方在于我们可以只捕捉少量物体的交互运动数据，然后直接将交互运动的结果映射到其他虚拟物体上，显然虚拟物体有可能客观并不存在，因此运动映射将会大大简化虚拟交互运动视频制作的流程。给定从多视点视频中恢复的运动控制器，我们首先将目标物体与原始物体对齐，这一过程包括将两者的质心以及两者质心坐标系同时对齐，然后，我们直接将恢

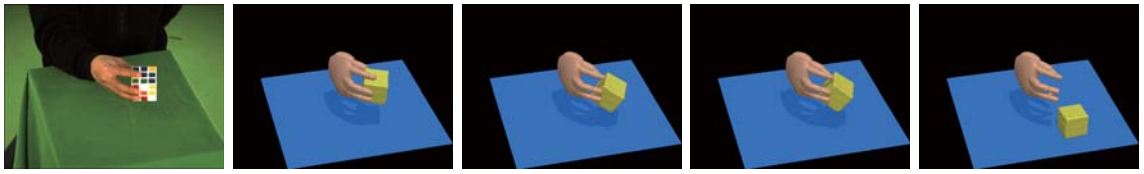


图 3.11　摩擦系数对结果的影响

复的运动控制器作用到目标物体上，即可以得到运动映射的结果。图3.2显示了一些运动映射后的结果。

3.7.2.2　改变物理参数

此外，我们还可以改变恢复的运动控制器中的物理参数，我们可以修改物体的摩擦系数。图3.11显示了一个抓取魔方的例子，其中，从左往右，依次是：原始图，摩擦系数从0.75 减小到0.3，0.2甚至0.01的结果。实验表明，随着摩擦系数的逐渐降低，手上的物体变得极度光滑，以至于从手上掉了下来。

3.7.3　与基于标记点系统的比较

在这一小节，我们首先介绍了与其他近似系统的比较结果，其中包括基于标记点的运动捕捉系统（Vicon① 系统），以及将标记点系统与Kinect系统结合的运动捕捉系统，实验结果表明我们的系统可以获得更好的实验结果。

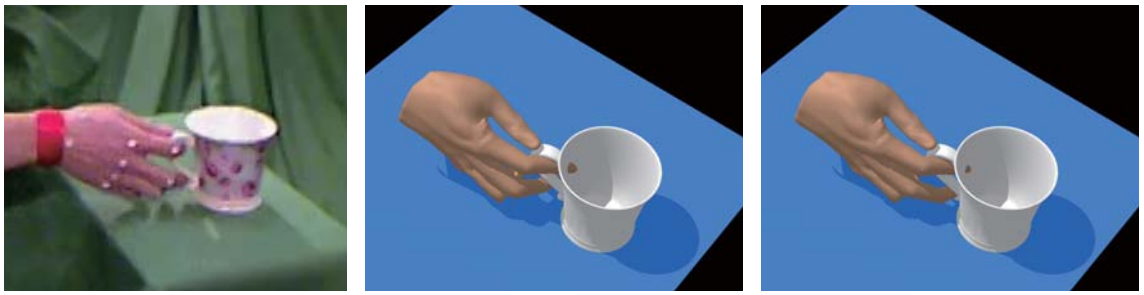


图 3.12　Vicon系统及Zhao的方法[33]结果

3.7.3.1　与Vicon系统的比较

由于我们无法将固定多相机系统的无标记运动捕捉系统与基于标记点的运动捕捉系统放在同一个屋子里，我们使用同样的物体做同样的交互运动来进行比较实验。我们采用的基于标记点的运动捕捉系统使用的是Vicon光学标记点系统，特别地，我们将21个标记点贴在手上，6个贴在物体上，并且使用了12个相机来捕捉手与物体的交互运动，图3.12的第一列显示的即是贴上标记点的手与物体。对捕

① http://www.vicon.com/

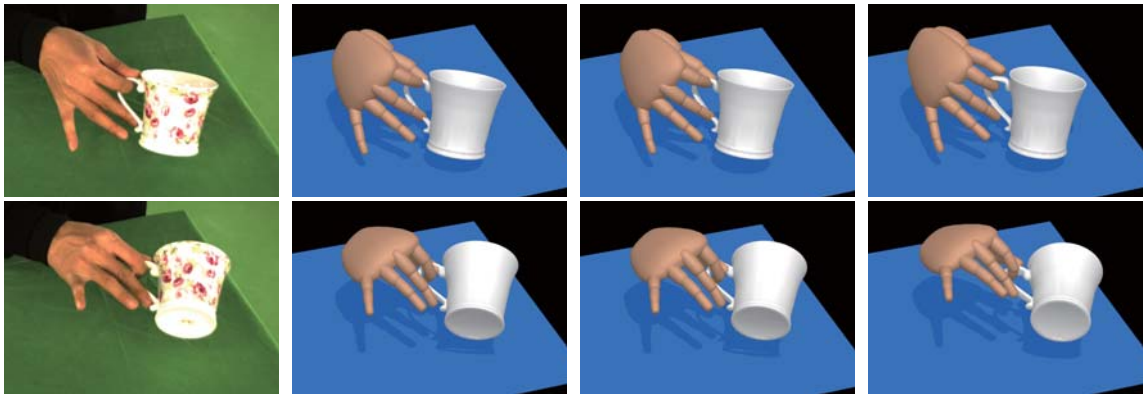


图 3.13　骨架运动重建与基于物理的运动捕捉结果比较

捉下来的每一帧，我们手工标记出所有的标记点，然后利用这些标记点进行手与物体的姿态重建。图3.12的第二列显示的是使用Vicon系统的捕捉结果。实验结果表明基于光学标记点的系统不能很好的重建出手与物体之间的接触点位置，此外，手与物体之间的接触运动也不能很好的重建出来。

3.7.3.2　与Zhao方法[33]的比较

最近，Zhao和他的同事将Kinect系统和基于标记点的光学系统结合在一起，获得了高精度的手势运动捕捉结果，本文也将我们的系统与之进行了比较。由于他们系统的主要目标是捕捉手的自由运动，因此我们将他们的系统进行了一定程度的拓展，使其能够捕捉手与物体的交互运动。与他们原来的方法一致，我们将交互运动的捕捉描述成一个优化问题，也即是最大化重建的姿态与观测数据之间的一致性，其中，观测数据包括标记点的位置，以及Kinect获得的深度信息，求解同样采用粒子滤波的优化方法来搜索最优的手与物体的姿态。尽管结合Kinect的深度信息可以提高手与物体姿态的跟踪精度，然而他们的方法依然被手与物体之间的严重遮挡问题所困扰，无法正确地捕捉手与物体的细微接触运动，图3.12的第三列显示的是他们方法的运动捕捉结果。

3.7.4　与骨架跟踪结果的比较

我们将本文提出的基于复合运动控制器的运动捕捉方法与传统的骨架跟踪方法进行了比较，实验结果表明，骨架跟踪方法获得的运动信息噪声很大并且跟踪结果容易出现运动跳变，当然也就无法捕捉手与物体之间的细微接触运动。图3.13显示了一些帧的结果，其中，从左往右分别显示了原始图像、骨架跟踪结果、骨架跟踪平滑结果、引入碰撞检测的骨架跟踪结果，以及基于物理的运动捕捉结果。值得说明的是，直接平滑处理骨架跟踪的结果并不能消除跟踪过程中手

与物体之间的错误接触点，换句话说，单纯使用骨架跟踪的方法并不能确定手的关节与物体之间是否存在接触点。其主要其原因在于，手与物体之间存在着十分严重的遮挡，并且图像评价函数也具有较大的误差。

此外，我们还在骨架跟踪过程中引入碰撞检测，并做了实验分析。特别地，我们在式（3-3）中加入碰撞检测项来惩罚手与物体互相穿透的情形，与之前求解类似，采用了同样的参数。尽管在骨架跟踪过程中引入碰撞检测项可以有效地避免手与物体出现互相穿透的情形，但是无法保证跟踪的结果具有准确的接触点，并且手与物体之间的细微接触运动也不能保证满足物理动力学方程。相反，为了避免手与物体出现互相穿透的情形，骨架跟踪的结果有时变得更为糟糕，如图3.13倒数第二列所示。因此，在骨架跟踪过程中引入碰撞检测，并不能有效地捕捉手与物体之间的接触运动。

得益于物理约束的运动控制器，我们的系统不仅可以产生符合观测图像数据的交互运动，还可以使捕捉的运动符合物理动力学模型。更为重要的是，恢复出基于物理约束的复合运动控制器，使我们可以更容易地进行交互运动的运动映射，而这样一种优势在以往的骨架跟踪系统中都是无法想象的。

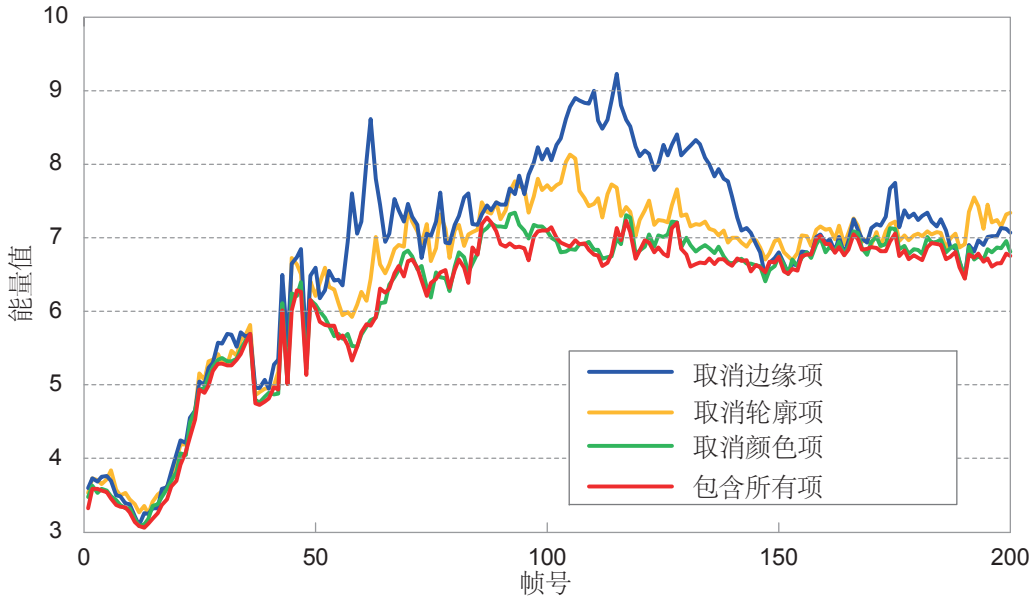
3.7.5　更多评测与比较

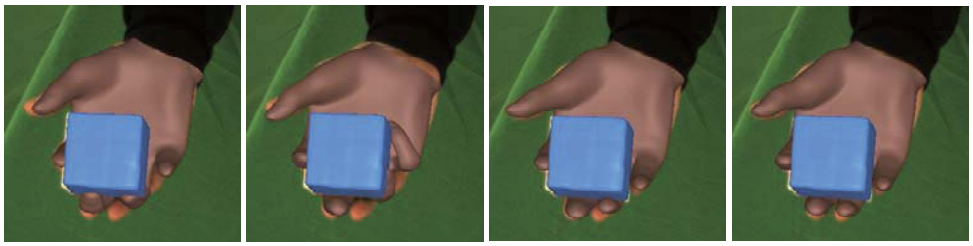
3.7.5.1　“虚拟”力

为了验证本文所提出的系统中，复合运动控制器“虚拟”力的作用，我们分别比较了在复合运动控制器中加入和取消“虚拟”力的实验。特别地，我们比较了使用式（3-8）的比例- 微分控制器和使用式（3-12）的复合控制器的差别，我们使用同样的采样策略和参数进行最优控制器的搜素。实验结果表明，没有“虚拟”力，系统无法正确地实现手与物体的交互运动捕捉，如图3.8所示。这是由于没有“虚拟”力的比例- 微分控制器没有考虑物体的运动，因此也就无法捕捉物体的运动。

3.7.5.2　基于接触点采样

这一实验主要是验证基于接触点采样方法的有效性。为了便于比较，我们使用同样的1500个采样样本进行随机采样和基于接触点的采样。实验结果表明，如果在采样过程中不考虑接触点的信息，随机采样并不能有效地捕捉手与物体之间的交互运动。相反，本文提出的方法能够成功实现手与物体的交互运动捕捉。





(a) 取消边缘项　(b) 取消轮廓项　(c) 取消颜色项　(d) 包含所有项

图 3.14　骨架跟踪的图像项评价

3.7.5.3　减少相机数目

我们的系统使用固定多相机实现无标记的手与物体交互运动捕捉。为了分析相机数目对运动捕捉的影响，我们测试了不同相机数目的运动捕捉性能。从实验中，我们发现相机数目从6 个减少到5个时，捕捉的运动质量并没有很大程度的降低。然而，当相机数目少于5个时，捕捉的运动质量开始大幅度的降低。这可能由于比较糟糕的骨架跟踪结果对运动控制器优化的影响。

3.7.5.4　图像项

骨架跟踪的目标函数包含三项，分别是：轮廓项、颜色项和边缘项，为了分析这三项之间的重要性关系，我们做了相关的实验来分析三者对骨架跟踪结果的影响。特别地，我们分别去掉式（3-3）中的一项，然后进行多视点视频的骨架跟踪。为了便于比较，骨架跟踪过程中使用的所有参数均是相同的。

由于没有真实的运动数据，我们使用下面两种方法对重建的骨架进行比较

- 第一种方法是直接比较式（3-3）的能量函数值，图3.14中第一行显示了不同跟踪方法得到的跟踪结果对应的能量，能量曲线反应了不同跟踪方法的跟

踪误差。

- 第二种方法是将跟踪出来的结果投影到新的相机视角上。特别地，我们将其中的一个相机视角排除在骨架跟踪算法之外，然后将所有方法得到的跟踪结果往该相机上进行投影，图3.14中（a-d）显示了不同跟踪方法投影得到的结果图。

两种比较方法同时表明，结合三种图像项的方法可以获得更好的骨架跟踪结果。

3.7.5.5　时间性能

在我们所有的实验中，基于六个固定视点的骨架跟踪，每一帧大约需要2－2.5分钟，而每一帧基于采样的控制器优化需要差不多1－1.5分钟。所有的实验均在主频为2.7GHz的英特尔酷睿i7处理器、显卡为英伟达GT650M、内存为16G 的计算机上进行测试。使用没有经过优化的代码，我们的运动捕捉方法每一帧大约需要3－4分钟。

3.8　本章小结

本章着重讨论了近距离场景包含严重遮挡的无标记运动捕捉问题，特别地，以手与物体的交互运动作为具体的研究对象。我们提出了一个基于固定多相机的无标记运动捕捉系统，该系统可以捕捉灵活的手与物体的交互运动。由于该系统不需要标记点、手套以及传感器，因此对手与物体的无标记运动捕捉而言，具有很大的吸引力。此外，本章提出的系统可以很自然地对手的关节、物体的运动，以及它们之间相互接触的细微运动进行建模，得到与观测图像一致的、符合物理约束的逼真运动捕捉结果。

通过捕捉四个物体的诸多不同的交互运动情形，我们验证了本章所提方法与系统的性能。实验结果表明，本章提出的系统对物体外观的变化以及背景环境的复杂程度都具有良好的鲁棒性。此外，将捕捉的运动控制器映射到具有不同物理特性的新物体上，证明了所提方法和系统具有很强的泛化能力。

本章提出的系统结合了基于视频的运动捕捉和基于物理约束的运动建模两种技术的优势。数学上来讲，基于物理约束的运动建模是一个病态问题，因为有很多种调整策略使得运动满足物理动力学方程，但是仅有少部分的运动与真实的图像数据吻合。通过同时考虑物理约束以及观测的图像数据，我们可以生成与真实图像数据匹配的物理运动结果。当然，基于多视点视频的运动捕捉也可以利用物理约束来减少模型的不确定性，从而确保重建运动在物理上是合理、可行的。

本章提出的系统与方法最大的缺陷是计算代价高昂，因此，使用图形处理器（GPU）或者并行计算来对本章提出的系统进行加速将是一个值得探讨的问题。事实上，本章提出的无标记运动捕捉系统的绝大部分模块都可以使用图形处理器进行加速，一方面，由于涉及图像处理的计算，式（3-3）可以用图形处理器进行加速；另外一方面，两个搜索求解过程（互动模拟退火和随机采样）也可以很容易地实现基于图形处理器的加速。此外，修改和重用捕捉的交互运动数据（包括运动迁移、插值和合成）也是潜在的研究方向。当然，测试更多不同物体、不同交互运动形式的运动捕捉也是值得完善的工作，这也必将会为基于数据驱动的交互运动视频制作，建立一个全面而详细的数据库。

第 4 章　移动多相机系统的户外无标记人体运动捕捉

在前两章中，我们讨论了近距离场景的无标记运动捕捉问题，并且针对运动的复杂程度提出了两种不同的运动捕捉系统。但是，前述章节中提出的运动捕捉系统都只能胜任实验室环境下的无标记运动捕捉，而涉及大规模场景的诸多户外运动，例如滑雪运动、滑板运动、篮球运动、足球运动等等，固定单相机或多相机系统均无法满足此类运动捕捉的需求。为此，针对大规模室外场景、不可控环境下的无标记运动捕捉问题，本章提出移动多相机系统的运动捕捉系统，并以户外无标记人体运动捕捉作为具体的研究对象，解决了室外场景下大规模运动的无标记运动捕捉问题，移动多相机系统甚至可以在两个手机摄像头的采集环境下有效地实现人体的无标记运动捕捉。为了实现大规模运动的无标记运动捕捉，本章提出的方法首先恢复出可移动手持相机的随时间变化的空间位置关系，针对宽基线的可移动相机系统，提出一种新的运动对象稠密点云的计算方法，利用输入的多视点视频信号以及计算的点云信息，恢复出人体的骨架运动。为了提高骨架运动信息恢复随时间变化的稳定性，与传统的方法相比，本章提出的方法引入了一种新的稀疏性约束，并且提出了与视点相关的动态纹理模型。不同环境下、不同运动对象的大规模户外无标记运动捕捉的测试结果均体现了本章提出方法的有效性。

4.1　导言

无标记运动捕捉技术能够捕捉运动对象的骨架运动信息，而无需运动对象穿戴任何传感器或带标记点的服装，对运动对象的干预很低，在近些年获得了快速的发展[41,43,89]。但是，受限于小范围场景，以及相机位置需要事先校准确定，甚至需要干净的背景环境（如绿幕布背景），目前广泛采用的无标记运动捕捉技术，能够捕捉的运动类型十分有限[42,90]。最近，Wu等人[8]设计了一种手持式的双目相机系统，可以在自然背景环境下实现无标记的运动捕捉任务，大大简化了无标记人体运动捕捉系统搭建与使用的复杂程度。然而，由于使用了特定的全局光照模型，他们提出的系统仅能在一个小范围场景内移动，因此也就无法大幅增加可捕捉的运动类型。

本章着重讨论了大规模场景的无标记人体运动捕捉问题，主要原因在于很多运动只有在大范围场景下才能得以施展，比如滑雪运动、滑板运动、篮球运动、足球运动等等。事实上，针对大规模场景下的人体运动捕捉，曾有研究人员讨论

(a) 手机拍摄帧　(b) 重建的点云　(c) 恢复的骨架　(d) 运动映射结果

图 4.1　移动多相机系统的无标记人体运动捕捉与运动映射

过[50,51]，但是这些方法需要在运动对象身上绑上一定的传感器或者相机。显然，这些类型的处理手段降低了运动对象的灵活度，不符合无标记运动捕捉的内在要求。为此，针对大规模场景的无标记人体全身运动捕捉问题，本章提出了一种移动多相机的无标记运动捕捉系统和方法，该方法首先从移动的多相机系统中重建出三维点云，依赖于重建的点云以及彩色图像，恢复出运动对象的骨架运动信息。为了实现大规模场景的拍摄任务，该方法允许拍摄者跟着运动对象进行拍摄，实验表明，跟踪距离可持续数百米之远。此外，本章提出的方法仅需要数个手持相机，而这些相机可以由几个不同的拍摄者跟着运动对象进行拍摄。在最简单的情况下，只使用两个移动手机即可完成对运动对象的拍摄数据采集。与Wu 等人[8] 的方法相比，本章提出的方法进一步简化了捕捉系统搭建与使用的复杂度。图4.1 中，（a）显示了系统输入视频的一些帧，（b－c）分别显示了重建的点云和骨架；（d）则显示了使用本章提出的方法，可以很容易地将运动捕捉结果映射到虚拟对象上。

值得指出的是，由于高精度的三维点云可以用来很好地估计运动对象的骨架信息[46]，本章提出的方法首先从移动的多相机视频数据中恢复出前景对象的三维点云。事实上，从手持的移动相机视频数据中，直接重建出运动对象的三维点云，也有一些研究工作出现[49,91]。但是，以前的点云重建工作依赖于运动恢复结构（Structure from Motion，SfM）[92]技术。需要说明的是，这项技术通过静态的像素点估计相机的外参，而运动变化的前景像素在运动恢复结构中被认为是野值（outliers），也即噪声。这一问题会极大地限制运动恢复结构技术在移动相机系统无标记运动捕捉中的应用：一方面，估计移动的相机参数需要大范围的静态背景像素，这就要求运动的前景所占像素不能太大；另外一方面，运动的前景对象具有较大的像素占比，才能实现高精度的无标记运动捕捉，以即是说，基于运动恢复结构的技术无法保证运动的前景对象具有较高的图像分辨率（不能超过10% 的

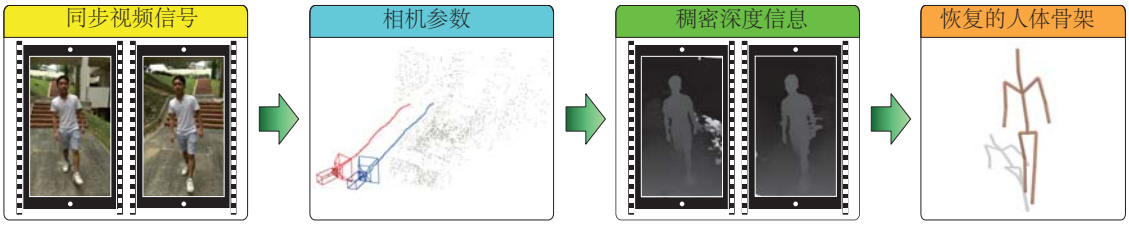


图 4.2　移动多相机无标记运动捕捉系统的流程图

像素占比），这就使得基于运动恢复结构技术的运动捕捉，主要依赖前景运动对象的稀疏视点的轮廓信息，因而运动捕捉结果质量非常有限。

幸运的是，最近一项叫做“视觉协同即时定位与地图重建”①（CoSLAM）的技术[93]为我们提供了处理该问题的另外一个手段。该方法可以同时恢复相机的运动、静态像素点、以及动态像素点的三维信息。在实际拍摄过程中，我们使用了多个移动相机，跟随运动对象近距离拍摄，其中，运动对象的像素占比达到50%仍然可以重建出高精度的三维点云。尽管如此，“视觉协同即时定位与地图重建”恢复出的三维点云是非常稀疏的，很难直接用于大规模场景的无标记运动捕捉。因此，我们需要解决宽基线移动相机的立体匹配问题，也即是重建出运动对象稠密的三维点云。据我们所知，只有最近的两个工作[91,94]试图求解过该问题，其中Taneja等人[91] 的工作重建出来的点云对于运动捕捉而言过于粗糙；Jiang的方法[94]，每一帧大约需要15 分钟的处理时间，计算复杂度太高。为此，本文提出了一种高效的算法，使每帧处理时间降低到大约1 分钟，获得与Jiang的方法近似相当的三维点云重建精度，并且可以满足运动捕捉的要求。

重建出的稠密三维点云紧接着会用于无标记的运动捕捉中，需要指出的是，为了减少点云噪声、少量相机视点以及光照变化对运动捕捉系统的影响，本章提出了一种新的无标记运动捕捉方法，结合了线性优化与非线性优化的优点，探索了深度信息、轮廓信息以及颜色信息对捕捉精度的影响。在运动捕捉的每一帧，我们先用线性优化的方法得到一个初始的人体骨架姿态，然后再使用全局优化的方法来进一步完善求解的人体骨架姿态。为了提高运动捕捉系统的稳定性，在线性骨架姿态求解阶段，本章引入了一个骨架姿态的稀疏约束。此外，我们还使用了一种新的纹理更新策略，以避免大规模场景下光照变化对运动捕捉的影响。

4.2　相关研究

本文提出的移动多相机系统的无标记人体运动捕捉系统以多个移动相机的拍

① 该技术的代码可以在以下网址找到：https://github.com/danping/CoSLAM

摄视频作为输入，使用“视觉协同即时定位与地图重建”的技术估计出相机的运动以及稀疏的点云信息，然后根据估计的相机参数以及拍摄的视频，本文提出了一种新的三维点云重建算法可以获得稠密的运动对象三维点云，最后运动捕捉系统结合运动对象的多种信息恢复出人体的三维骨架运动，系统流程如图4.2所示。需要说明的是，在拍摄之前，多视点相机利用闪光灯手动估算出同步的整数帧差，当然也可以采用文献[49]提到的自动同步算法。因此，在后续的小节中，我们都假设多视点视频的信号是已经同步好的。此外，值得注意的是，“视觉协同即时定位与地图重建”技术重建出的误差通常要高于2 个像素，这对于运动捕捉而言，具有一定的挑战。为此，本文提出一种新的三维点云重建算法得到每一帧稠密的深度图，这些计算出的深度图与其他图像信息一起，被用来估计人体的三维骨架运动。

4.2.1　视觉协同即时定位与地图重建

视觉协同即时定位与地图重建的目标是从多个相机视频序列中同时恢复出相机的位置以及场景的三维结构。当然，基于视觉的即时定位与地图重建技术不仅仅局限于多个相机，特别地，研究人员已详细讨论过基于单目相机的视觉即时定位与地图重建技术，目前来看，主要有两种方法：一种是基于运动恢复结构的技术，或依次恢复场景的三维结构和相机位置[95]，或利用捆绑约束优化的方法对场景三维结构和相机位置同时进行求解[96]；另外一种方法是将该问题描述成一个贝叶斯推理问题，通过卡尔曼滤波[97]、粒子滤波[98] 等方法实现问题的优化求解。Strasdata等人[99]仔细比较了两种方法的优劣后指出：基于运动恢复结构的技术可以获得更为精确的结果，而基于滤波方法的处理手段可以更为快速地得到结果。

然而，基于单目相机的视觉即时定位与地图重建技术很难有效地处理动态场景的情形，尤其是当前景运动对象具有比较大的像素占比[93]，因此基于多相机的视觉即时定位与地图重建逐渐成为主流。为此，Nister等人[100]提出一种使用双目立体相机的窄基线系统实现相机姿态与场景三维结构同时恢复的任务。但正如他们指出的一样，窄基线的双目立体相机会严重影响三维结构恢复的质量；Paz等人[101]则将双目立体相机分离开，可有效估计相机的旋转；Zou等人[93]在此基础之上使用多个不同步移动的相机，得到了很好的相机姿态估计与场景三维结构恢复的结果。

需要指出的是，当前视觉协同即时定位与地图重建技术只能恢复较为稀疏的三维点云，并且与场景类别还有比较大的相关性，特别地，需要场景中存在较为明显的角点特征。这些问题也是视觉协同即时定位与地图重建需要解决的难题与挑战。

4.2.2　动态场景的立体重建

目前为止，有关立体重建的话题已经有很多研究工作涉及，文献[102]也对立体重建的相关研究工作做了一个非常详尽的综述。值得一提的是，绝大多数立体重建工作都局限于静态场景，只有少量的研究工作关注动态场景的立体重建问题，并且在仅有的动态场景立体重建方法中，大都使用固定的多相机系统[103–107]。

为了增加系统采集的灵活性，我们使用移动多相机系统来进行数据采集，显然，移动多相机系统使得立体重建变得极为困难，屈指可数的研究工作曾讨论过移动相机的动态场景立体重建问题[91,94,108]。Ballan等人[108]首先重建出静态背景的三维结构，而前景的三维结构仅用一个平面标识出来；Taneja等人[91]则用体素的结构来表示前景移动的物体，显然这种表示方法得到的前景点云非常粗糙，难以胜任运动捕捉的任务；Jiang等人[94]提出的动态场景三维重建的结果似乎可以用来作为运动捕捉的输入，但是他们的方法每处理一帧都大约需要15 分钟的时间。为此，我们提出了一种新的立体重建方法，其可以大幅提高处理时间而不降低三维重建的精度。

4.2.3　无标记运动捕捉

关于无标记运动捕捉，目前已有大量的方法与系统，这里我们仅讨论与本文研究工作相关的部分无标记运动捕捉研究。早先的无标记运动捕捉技术需要在一个可控的室内环境下，采用大量的固定相机实现[4,5]。这些系统一般使用用生成模型的方法，也即是利用一个简单的模板骨架[38]，通过最大化变形后的骨架与观测图像之间的一致性，求得骨架的变形参数。针对变形骨架与观测图像一致性的优化求解，使用局部优化[39]，或是全局优化[38,40]，或是两者结合的优化方法[41] 都有相关的研究工作提及。为了提高运动捕捉的精度，研究人员也尝试使用了高精度的三维表面模型[42,43]。

事实上，简化采集系统已成为无标记运动捕捉的发展趋势。特别地，Elhayek等人[44]讨论了室内环境下无需同步多相机系统的无标记运动捕捉问题；Hasler等人[49]则尝试过使用手持相机在户外进行无标记运动捕捉。由于以上两个工作在进行运动恢复的过程中仅使用了轮廓信息，且只用局部优化的手段来估计运动，恢复出的三维运动精度非常有限。基于小范围内全局光照恒定的假设，Wu等人[8]在一个自然场景的环境下，使用手持双目相机设备实现高精度的运动捕捉，但是恒定全局光照的假设使得他们的系统只能胜任很小范围内的无标记运动捕捉。值得一提的是，Wei等人[45]还讨论过交互式的单目视频的无标记运动捕捉问题。

随着消费级深度相机的逐渐普及（例如微软的Kinect相机），Shotton研究组[46]和Ganapathi研究组[10]开发出一个可在室内进行实时无标记运动捕捉的系统，显然此类系统可以大大促进人机交互及一些体感游戏的繁荣。最近，也有相当一部分研究工作[11,47,48]尝试提高此类系统的稳定性以及运动捕捉的精度。

本文主要着眼于大范围场景下的无标记运动捕捉问题。考虑到生成模型的诸多优势，例如精度高、可靠性强，本文提出一个基于移动多相机系统的无标记人体全是运动捕捉系统，该系统结合在求解人体姿态的过程中结合了局部优化和全局优化的方法优势。在运动捕捉方法，本文的主要贡献是提出了一个新的三维人体姿态求解算法，其中包括：（1）一个全新的骨架运动稀疏约束；（2）一个新颖的自适应纹理更新策略。

值得指出的是，许多传感器，诸如惯性传感单元（IMUs）、加速度计，也可以用来作为大规模场景的无标记运动捕捉。例如，商业化的惯性运动捕捉系统（如XSens① 的MVN系统）通过一个紧致的衣服来对户外的人体运动进行捕捉。但是，这些类似的系统仅能记录下骨架节点的旋转信息，而无法记录下统一坐标系下人体运动的全局三维信息，这对于某些应用场合是无法接受的。

4.3　三维点云重建

正如前面小节所述，“视觉协同即时定位与地图重建”可以获得每一帧的移动相机参数。而在相机参数已知的情况下，根据针孔相机的数学模型[92]，三维点云与视角深度图只需进行简单的线性变换。因此，本节讨论的三维点云重建，也可以理解成深度图的估计问题。事实上，假设相机的内参为f_x, f_y, α, u, v，其中，f_x, f_y分别为相机x轴和y轴的焦距；u, v为主点坐标；α为畸变系数。于是，空间三维点$[x, y, z]^T$ 与投影的二维坐标$[X, Y]^T$之间的关系可以表示成

$$\begin{bmatrix} f_x & \alpha & u \\ 0 & f_y & v \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = S \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}. \tag{4-1}$$

其中，S为比例因子。如果深度图已知，也即$[X, Y]^T$坐标下的深度值z已知，那么二维坐标$[X, Y]^T$对应的三维点$[x, y, z]^T$可以很容易得到

$$z = z, y = \frac{z(Y - v)}{f_y}, x = \frac{z(X - u) - \alpha y}{f_x}. \tag{4-2}$$

为了估计每个相机视角的深度图，我们首先使用基于信度传播（Belief

① http://www.xsens.com/

Propagation, BP）的立体匹配算法[109,110]，引入可见性判断和区域分割的平面拟合机制，重建出所有相机第一帧的深度图。对于每个视角视频的其他帧，我们使用深度传播的思想以期望得到所有帧的深度图。此外，为了提高深度传播的稳定性，针对图像中不同的像素点，我们采取了不同的深度传播策略。实验表明，约有超过70% 的图像像素点的深度值可以直接利用前一帧的深度传播得到，而剩余图像像素的深度需要使用深度填充算法完成深度估计的任务。为了便于描述，我们用$I_m = \{I_m^t | 1 \le t \le N\}$表示第$m$个相机的视频帧，其中，$1 \le m \le M$，$M$是所使用相机的数目，在本文的实验中，$M = 2$。我们的目标是恢复出每个相机的深度图$\{Z_m^t | 1 \le m \le M; 1 \le t \le N\}$。我们使用$\mathbf{x}_m^t$ 表示第m 个相机第t 帧$\mathbf{x}$像素的深度值，在不引起歧义的情况下，将其简写为$\mathbf{x}$。

4.3.1　深度初始化

这一小节将简要介绍深度的初始化，详细内容可以参看文献[110]。我们将待求解的深度区间$[d_{min}, d_{max}]$离散化成300层，通过信度传播的算法优化如下方程

$$E = \sum_{\mathbf{x}} E_d(\mathbf{x}, Z_m^1(\mathbf{x})) + \sum_{(\mathbf{x},\mathbf{y})\in\mathbb{N}} E_s(\mathbf{x}, \mathbf{y}, Z_m^1(\mathbf{x}), Z_m^1(\mathbf{y})). \tag{4-3}$$

其中，E_d, E_s分别是数据项和平滑项。这样，我们就可以得到图像中每个像素的深度值。式(4-3)中数据项的计算方法如下，假设像素$\mathbf{x}$ 的深度值为$Z_m^1(\mathbf{x})$，于是我们可以将该像素对应的三维点投影到各个相机视角，通过比较各个相机视角下，投影像素之间DAISY 描述子[111]的一致性误差，得到数据项E_d。需要指出的是，如果某个像素点在其他相机下都不可见，我们则通过同一相机前后帧之间DAISY 描述子的一致性误差作为该像素点的数据项。式(4-3)中平滑项用来保证邻域像素具有比较一致的深度值。此外，为了保证得到更为精确的深度估计，我们使用均值平移(mean shift)[112]的图像分割手段，分区域进行深度的初始化。

4.3.2　深度传播

假设所有相机视角第t帧的深度图已经计算得到，本小节将给出一个有效的计算第$t + 1$帧深度的传播算法。我们的主要想法是，针对不同类型的像素采用不同的传播策略。我们首先根据像素在其他相机视角下的可见性，将像素做一个分类，其中，计算像素$\mathbf{x}_m^t$在第n个相机视角可见性的公式如下：

$$V_{m\to n}^t(\mathbf{x}) = \begin{cases} 1 & |Z_{m\to n}^t(\mathbf{x}) - Z_n^t(\mathbf{y})| < \eta, \\ 0 & \text{其他.} \end{cases} \tag{4-4}$$

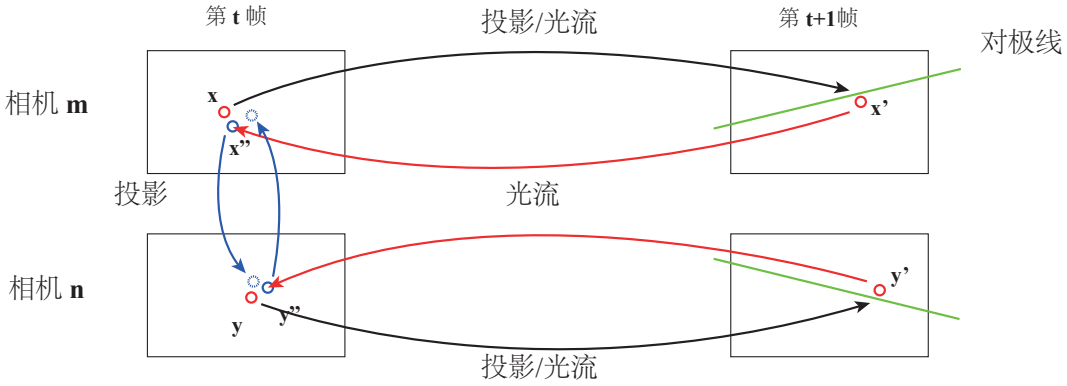


图 4.3　可见点的深度传播

式(4-4)中，$\eta = 0.02(d_{max} - d_{min})$。由于$\mathbf{x}_m^t$像素在第$m$个相机下的深度值为$Z_m^t(\mathbf{x})$，根据第$m$ 个相机和第n个相机之间的外参关系，我们得到$\mathbf{x}_m^t$像素在第n个相机下的匹配像素$\mathbf{y}_n^t$。我们用$Z_{m\to n}^t(\mathbf{x})$表示$\mathbf{x}_m^t$像素在第n个相机下深度值，实际上它等于$Z_m^t(\mathbf{x})$，而匹配像素$\mathbf{y}_n^t$ 在第n个相机下的深度值用$Z_n^t(\mathbf{y})$ 表示，如果这两个值之间的误差小于一定的阈值，我们就认为$\mathbf{x}_m^t$像素在第n个相机下可见。我们用$\mathcal{V}_{m\to n}^t$ 表示第m个相机在第n 个相机中所有可见的像素点，也即是：$\mathcal{V}_{m\to n}^t = \{\mathbf{x}|V_{m\to n}^t(\mathbf{x}) = 1\}$。于是，第$m$ 个相机的可见性图是所有相机视角下可见性图的集合

$$\mathbb{V}_m^t = \cup_{n\neq m}\mathcal{V}_{m\to n}^t. \tag{4-5}$$

为了简化深度传播，$\mathbb{V}_m^t$中的每一个像素有且仅有一个可见的参考相机视角。也就是说，如果某个像素点在多于一个相机视点下可见，我们选择式（4-4）中，深度误差最小的视角作为该像素的唯一可见参考相机视角。

4.3.2.1　静态点传播

我们首先考虑如何对所有可见的静态点进行传播处理。事实上，由于事先并不知道$\mathbb{V}_m^t$中哪些像素点是静态点，哪些像素是动态点，我们只能假设所有的像素点都是静态点，也即是将$\mathbb{V}_m^t$中的所有像素点进行传播，如图4.3所示。对于第m相机下的每个像素$\mathbf{x}_m^t$，假设其在第n 个相机下是可见的，并且其对应的像素为$\mathbf{y}_n^t$，由于第$t+1$帧的相机参数已知，直接将这些像素点对应的三维坐标往后一帧上投影，就可以分别得到其在后一帧上对应的像素点$\mathbf{x}'^{t+1}_m$及$\mathbf{y}'^{t+1}_n$。这样，像素点$\mathbf{x}'^{t+1}_m$的深度将可以直接通过投影得到，如图4.3 中的黑线所示。

显然，上述的假设不是完全都成立的，为此我们需要检查并去除那些不合理的假设像素点。由于$\mathbf{x}'^{t+1}_m$和$\mathbf{y}'^{t+1}_n$ 的深度均由投影而来，因此这两个像素点必须满足对极几何约束，在我们的方法中，我们还同时考虑了光流约束。正如

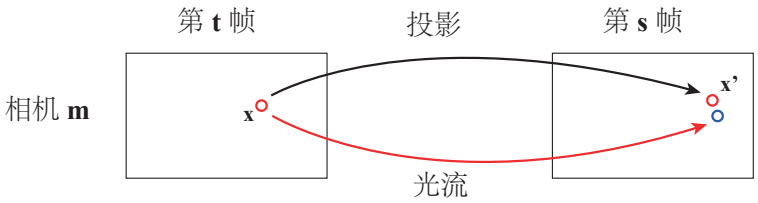


图 4.4　遮挡点的深度传播

图4.3中的红线所示，我们使用光流跟踪的方法，分别得到$\mathbf{x}'^{t+1}_m$和$\mathbf{y}'^{t+1}_n$在前一帧的对应$\mathbf{x}''^{t}_m$和$\mathbf{y}''^{t}_n$。在理想情况下，应当满足$\mathbf{x}''^{t}_m = \mathbf{x}^{t}_m$及$\mathbf{y}''^{t}_n = \mathbf{y}^{t}_m$。实际上，我们允许光流跟踪和对极线存在一定的误差，因此只判断了$\mathbf{x}''^{t}_m$与$\mathbf{y}''^{t}_n$之间的一致性。特别地，根据像素$\mathbf{x}''^{t}_m$的深度，我们将其投影到第n个相机视角下（如图4.3中的蓝线所示），要求投影后的像素与$\mathbf{y}''^{t}_n$在T_1个像素误差范围内（在我们的实验中$T_1 = 1$），否则，我们认为该像素就不是一个合理的静态点。当然，为了对称性的考虑，我们也将像素$\mathbf{y}''^{t}_n$往像素$\mathbf{x}''^{t}_m$对应的相机视角上投影，要求满足同样的约束条件。

4.3.2.2　动态点传播

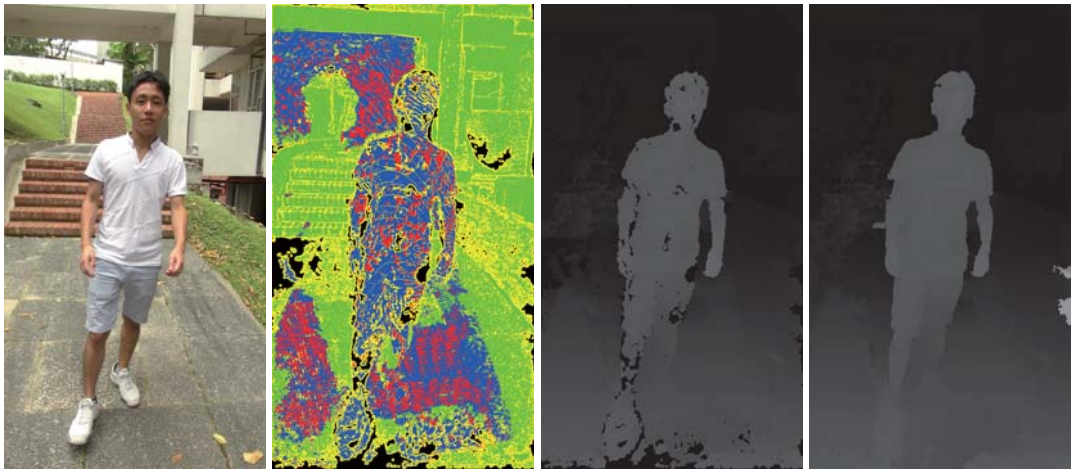
对于那些在第一阶段没能通过的像素点，我们均认为是动态点，并且采用了不同的传播方案，同样也在图4.3中展现出来。对于像素$\mathbf{x}^{t}_m$及其对应像素点$\mathbf{y}^{t}_n$，我们使用光流跟踪的方法，得到它们各自对应的像素点$\mathbf{x}'^{t+1}_m$和$\mathbf{y}'^{t+1}_n$，如图4.3中的黑线所示。$\mathbf{x}'^{t+1}_m$的深度值可以通过已知的相机参数，及$\mathbf{x}'^{t+1}_m$、$\mathbf{y}'^{t+1}_n$之间的三角关系得到。

为了去除不稳定的传播点，我们检查了对应点之间的对极线关系，也即是要求$\mathbf{x}'^{t+1}_m$和$\mathbf{y}'^{t+1}_n$必须处在误差为T_1的对极线上。此外，我们还要求它们之间的DAISY描述子误差必须在T_2范围内，实际实验中发现，$T_2 = 0.15$可以获得不错的传播效果。

4.3.2.3　遮挡点传播

所有在像素点集$\mathbb{V}^{t}_m$之外的像素点仅在第m个相机下可见，因此我们使用同一相机的相邻帧进行传播处理。假设像素点$\mathbf{x}^{t}_m$是静态点，并且将他的三维坐标投影到第s帧，得到对应像素点$\mathbf{x}'^{s}_m$，如图4.4中的黑线所示。于是，像素$\mathbf{x}'^{s}_m$的深度可以由投影决定。

为了更好的传播结果，我们使用颜色一致性的度量准则而没有采用DAISY描述子，其原因在于相邻相机的基线通常很小。如果RGB颜色误差大于阈值T_3，我们就认为传播的深度是无效的。在我们的实验中，RGB的颜色范围为$[0, 255]$，阈值$T_3 = 2$。接着，我们检查光流与深度投影之间的一致性。我们使用光流跟踪的



(a) $t+1$帧原始图　(b) 不同区域点　(c) 深度传播结果　(d) 最终深度图

图 4.5　深度传播

方法得到像素$\mathbf{x}_m^t$在第s帧上的对应像素点，如图4.4中的红线所示。我们要求光流跟踪得到的像素点的深度与$\mathbf{x'}_m^s$的深度在T_1范围内。我们使用了前后20 帧的信息来进行深度传播检查，如果超过一半的传播满足前面所述的检查，就认为当前帧的深度传播是有效的。

针对上述三种传播都失效的像素点，我们通过检测其邻域范围的像素点是否已经有深度值，如果有，则将该像素点在5×5窗口内所有深度的均值赋给当前像素点。图4.5显示了深度传播过程的中间结果。(a)图为原始彩色图；(b)图用不同颜色显示了不同传播策略的像素点，其中，红色像素点为静态点，蓝色像素点为动态点，绿色像素点为遮挡点，黄色像素点的深度通过空间平滑计算得到；(c)图显示了所有深度传播之后的结果，图中显示有一些空洞；(d)图显示了补洞之后的深度结果。

4.3.2.4　深度补洞

经过了前面所述的传播之后，仍然会有没有赋值的像素点，如图4.5(c)所示。为此，我们提出了两步深度补洞的策略。首先，我们将已经有深度传播值的像素点用平面拟合的方法进行修正。具体来讲，针对离散化的300层深度，我们仅考虑每一个有深度传播值的像素点中，与传播深度值最接近的5层。为了估计这些像素点在每一层的深度值，我们需要知道像素点在各个相机之间的可见性。可见性判断的具体做法是，根据像素的深度值，将该像素投影到其他相机上，如果投影像素在其他相机的帧中，我们就认为该像素在另外相机上可见的。如果像素点可被多个相机看到，我们用各个相机间DAISY 描述子最小值作为数据项；如果像素点在所有相机上均不可见，我们

用相邻帧间的颜色和光流一致性来计算数据项。特别地，我们计算相邻20 帧的$C_c(\mathbf{x}_m^t, \mathbf{x'}_m^s)C_o(\mathbf{x}_m^t, \mathbf{x'}_m^s)$（也就是说，$|s-t| \le 10$），数值最小的10 个数的均值来处理遮挡的情形。其中，$C_c(\mathbf{x}_m^t, \mathbf{x'}_m^s) = 1 - \sigma/(\sigma + |I_m^t(\mathbf{x}) - I_m^s(\mathbf{x'})|)$ 用来评价颜色的一致性；$C_o(\mathbf{x}_m^t, \mathbf{x'}_m^s) = 1 - \sigma/(\sigma + |\mathbf{x}_m^t + o_{t\to s}(\mathbf{x}) - \mathbf{x'}_m^s|)$用来评价光流的一致性，实验中$\sigma = 14$。$o_{t\to s}(\mathbf{x})$用来表示像素$\mathbf{x}$从第$t$帧到第$s$帧的光流向量。于是，对于每个像素点最可能的5 个深度层，我们都可以使用均值移动的分割方法拟合出一个平面，然后将到拟合平面的距离加到5个可能的深度里，最终的深度值取数据项最小的那个。

在经过前面的平面拟合后，可能还会存在部分的空洞像素点，我们再次对式（4-3）进行求解，而对于那些前面拟合优化后的像素点，它们的深度值固定不变。由于这些像素点的可见性是未知的，因此我们对于300层的深度值计算两次，第一次是假设像素点可见，另外一次是假设像素点不可见。于是，每个像素点就会有600种选择可能性，接着使用信度传播的方法得到最终的深度值。

4.4　运动捕捉优化求解

本节将重点讨论如何利用已经计算出的深度，以及输入的多视点视频，实现运动捕捉的目标。为此，我们需要建立运动对象的模板三维模型$\mathcal{M}$。建立运动对象模板三维模型的具体方式如下：我们首先得到运动对象在固定姿态下的多视点图像，然后使用VisualSFM技术① 和多视点三维重建的方法[37]重建出一个近似的三维模型，然后重新三角化该近似模型，保证最终的模板三维模型$\mathcal{M}$拥有差不多5000个顶点。

我们将一个拥有36个自由度的骨架手动嵌入到模型$\mathcal{M}$中，然后自动地计算出模型表面点与骨架的变形权重关系[113]。为了便于描述，人体的骨架用$\chi = (\xi, \Theta)$表示，其中ξ是6个自由度的全局旋转、平移，$\Theta = \{\theta_1, ..., \theta_n\}$是骨架角度的向量。于是，骨架姿态$\chi$下的变形模型$\mathcal{M}(\chi)$表面的每一个三维点$\mathbf{v}(\chi)$可以使用线性混合蒙皮方法计算得到[41]。在后面的小节中，我们使用$\mathcal{M}$表示变形后的三维表面模型。

我们的目标是通过调整骨架姿态χ，使变形模型$\mathcal{M}$与图像以及重建的点云匹配。我们首先针对第一帧得到粗对齐结果，然后逐帧优化骨架的运动。其中，每一帧的骨架求解都分为两步处理：（1）对E_{loc}函数进行线性求解，得到骨架姿态的初始结果；（2）利用线性优化求解的结果，通过对E_{glob}函数非线性优化，得到更为精细的骨架姿态。线性优化可以大幅减少骨架姿态的搜索空间，而非线性优化则可以提高骨架运动恢复的稳定性与精度。值得一提的是，本文提出的运动捕

① http://ccwu.me/vsfm/

捉方法需要事先得到前景运动对象的轮廓图，我们使用文献[114]提出的半自动方法进行前景分割，当然，文献[115]提出的全自动分割方法也可用来实现前景分割的目的。

4.4.1　线性优化求解

我们的线性优化求解方程如下：

$$E_{loc}(\chi) = \lambda_{cont} \sum_{m} E_m^{cont} + \lambda_{dep} E^{dep} + \lambda_{reg} E^{reg}, \tag{4-6}$$

其中，m是相机序号，前两项为数据项，最后一项为正则项，$\lambda_{cont} = 4.5, \lambda_{dep} = 1.0, \lambda_{reg} = 10.0$，$E_m^{cont}$刻画了模型$\mathcal{M}$和图像的轮廓之间的一致性，$E^{dep}$则描述了模型$\mathcal{M}$和点云之间的匹配关系。针对每帧的线性优化，我们通过10次迭代找到匹配关系，求解最优的骨架姿态。

线性优化方法采用基于热启动的梯度下降法[60]，并且，为了得到更好的计算效率，我们对$\mathbf{v}(\chi)$进行线性化处理，也即是

$$\mathbf{v}(\chi + \nabla\chi) = \mathbf{v}(\chi) + J\nabla\chi. \tag{4-7}$$

其中，J是雅克比矩阵[116]，附录B中给出了线性化的具体推导过程。这样，我们就可以从连续变化的χ中求得$\mathcal{M}(\chi)$。

4.4.1.1　轮廓线项

我们将模型$\mathcal{M}(\chi)$上的每一个三维顶点往第m个相机上投影，找到与轮廓线上的像素点$\mathbf{x}$最接近的三维顶点$\mathbf{v}_i$。像素$\mathbf{x}$与相机中心可以定义一条射线，称作Plücker线，表示为$l_i = (a_i, b_i)$[41]。由于$\mathbf{v}_i$必须在这条线上，因此轮廓线项为

$$E_m^{cont} = \sum_i \left\| \mathbf{v}_i(\chi) \times a_i - b_i \right\|_2 . \tag{4-8}$$

4.4.1.2　深度项

对于重建点云上的三维点$\mathbf{p}_i$，我们将所有满足$\mathbf{n}(\mathbf{v}_i) \cdot \mathbf{n}(\mathbf{p}_i) > 0.1$约束的模型表面点中，距离最近的模型表面点$\mathbf{v}_i$作为$\mathbf{p}_i$的匹配点。我们使用点面距(point-plane distance)来评价深度的不一致性，其计算公式如下

$$E^{dep} = \sum_i \left\| \mathbf{n}(\mathbf{p}_i) \cdot (\mathbf{v}_i(\chi) - \mathbf{p}_i) \right\|_2 . \tag{4-9}$$

需要指出的是，点面距与点点距(point-point distance)相比，有着更好的收敛性能[6]。

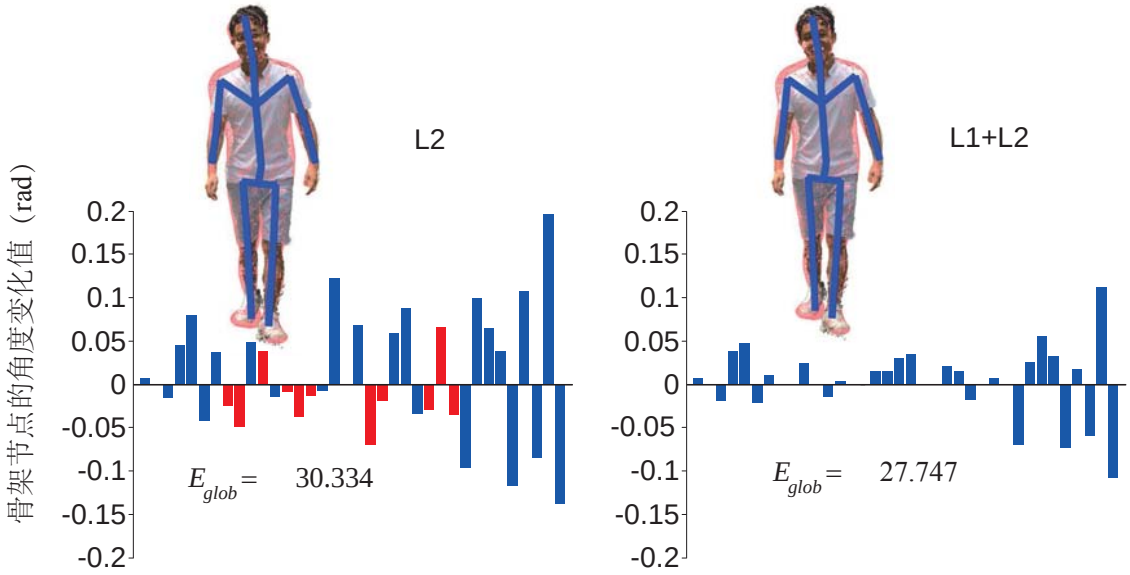


图 4.6　是否引入L1范数正则项的线性骨架求解比较结果

4.4.1.3　正则项

我们的线性系统使用的正则项包括两部分，也即

$$E^{reg} = \|\nabla\chi\|_1 + \left\|\nabla_p\chi\right\|_2. \tag{4-10}$$

其中，第二项是χ^{t+1}与三阶自回归估计姿态的L2范数。

式（4-10）中第一项是$\nabla\chi = \chi^{t+1} - \chi^t$的L1范数。这个新提出的范数能够保证线性优化成功实现运动捕捉。实际上，它要求骨架运动在时间上具有稀疏性，也就是说，应保证$\nabla\chi$具有尽量少的非零元个数。这个约束对于绝大多数的运动，甚至是快速运动都是合理的。这个基于稀疏约束的正则项可以显著提高骨架运动跟踪的鲁棒性，因为运动捕捉的输入是噪声很大的点云，并且仅有少量视角的轮廓线信息。显然，这些不稳定的信息输入使得不同的骨架节点组合有可能产生相同的E_{contc}和E_{depc}的拟合误差，而L1范数可以保证骨架节点运动与前一帧相比，出现尽量少的变动，这样一种稀疏性约束可以大幅降低求解过程中的不确定性。

图4.6显示了本文提出的L1正则项的优势，我们将不含$\|\nabla\chi\|_1$的正则项与包含$\|\nabla\chi\|_1$正则项估计出的$\nabla\chi$分别显示在图4.6的左右两边，其中横轴是骨架节点的序号，纵轴是估计的$\nabla\chi$角度，以弧度为单位。通过降低非零骨架节点的角度，包含L1正则项的线性骨架优化可以获得更低的E_{glob}。左图中的红色柱状条显示了只使用L2范数的骨架姿态求解后，额外需要运动的骨架节点。

4.4.2　非线性优化求解

前面一小节提出的线性骨架优化只是提供了骨架姿态的一个近似解，接下来，我们使用该近似解作为非线性优化的初始值，通过分析综合的方法来求解一个更为复杂的非线性目标函数，我们同时考虑了颜色、深度以及轮廓的信息：

$$E_{glob}(\chi) = \lambda_{dep}E^{dep} + \lambda_{sil}\sum_{m} E_m^{sil} + \lambda_{color}\sum_{m} E_m^{color}. \tag{4-11}$$

其中，组合系数分别为：$\lambda_{dep} = 2.5, \lambda_{sil} = 1.5, \lambda_{color} = 2.0$；同样地，$m$为相机的序号。我们使用与前一章相同的非线性采样方法：互动模拟退火来优化式（4-11）求取最优的χ^{t+1}。特别地，我们迭代地采样300个不同的骨架姿态，变形得到三维模型$\mathcal{M}$，然后计算匹配误差E_{glob}，这一采样选择过程一共持续20次。

4.4.2.1　深度项

这一项与式（4-9）相同。由于线性骨架姿态求解可以获得比较好的初始结果，并且从计算效率的角度考虑，在非线性的步骤中，我们简化了深度的匹配搜索过程。具体来讲，我们不再针对三维点云的每一个点，在模型上搜索最近的模型表面点作为三维点云的匹配点。相反，我们将三维模型的表面点$\mathbf{v}_i$投影到深度图像上，深度图上的像素点对应的三维点，即认为是与三维模型表面点$\mathbf{v}_i$匹配的三维坐标点。

4.4.2.2　轮廓项

这一项用来计算观测图像的前景轮廓S_m与三维模型$\mathcal{M}$向第m个相机上投影后得到的轮廓S'_m之间的不一致误差，由于两个轮廓均是针对第$t+1$帧而言，这里我们将上标省略掉了，我们的轮廓项定义为：

$$E_m^{sil} = \frac{\|S_m \cap \bar{S}'_m\|}{\|S_m\|} + \frac{\|S'_m \cap \bar{S}_m\|}{\|S'_m\|}. \tag{4-12}$$

其中，轮廓S的范数定义为轮廓中的像素个数，$\bar{S}$为轮廓S的补图。

4.4.2.3　颜色项

输入的彩色图像与三维模型渲染生成的纹理图像的差别度量是一个非常重要的一致性评价指标。然而，颜色项很少被以前的方法采用，原因在于对三维模型$\mathcal{M}$的外观进行建模是非常困难的一件事：一方面，使用固定的纹理无法处理由于光照变化带来的外观变化，而这一情形在大范围场景的户外很容易出现；另外

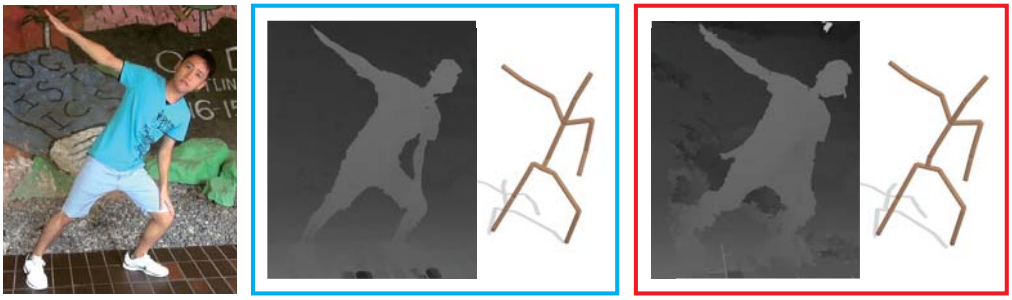


图 4.7　本文提出的立体重建方法与Jiang方法[94]的比较结果

一方面，简单的随时间变化的纹理更新又很容易受模型姿态的求解失效，反而使得跟踪的结果更为糟糕。

对于模型$\mathcal{M}$上的每一个点，我们对每个视角，都保存一个与视点相关的纹理图像块，并且该图像块不断地进行更新，之所以这样做的原因是不同相机在随时间变化的过程中颜色都是有差别的。我们用$c_m^{t+1}(i)$表示三维模型上的i顶点在第m个相机$t+1$帧上的纹理图像块，我们利用三维模型的第i个顶点在第m个相机上，可见的最近K帧加权平均求得$c_m^{t+1}(i)$，特别地，

$$c_m^{t+1}(i) = \sum_{s\in C} \exp\left(-E_{glob}(\chi^s)\right) \cdot I_m^s(i) \Big/ \sum_{s\in C} \exp\left(-E_{glob}(\chi^s)\right) \tag{4-13}$$

其中，$I_m^s(i)$是第i个顶点在m相机上第s帧的投影图像，大小为11×11的矩阵；$E_{glob}(\chi^s)$是式（4-11）在第s帧的结果；C是K帧的集合。上式中赋予较低E_{glob}值更大的权重，原因是我们认为较低的E_{glob}意味着更好的骨架姿态求取，也即是更好的几何纹理对齐结果。

有了每个像素对应的纹理图像块之后，我们使用零均值互相关（ZNCC）来比较图像块之间的差别，我们的颜色项定义为

$$E_m^{color} = \sum_i ZNCC\,(i, m, t+1). \tag{4-14}$$

这里，$ZNCC\,(i, m, t+1)$是$c_m^{t+1}(i)$和输入视频帧I_m^{t+1}在第i个顶点在m相机下投影的11×11图像块的零均值互相关值。

4.5　实验结果与分析

我们使用2－3个手持相机拍摄了大量的测试数据，拍摄相机使用的是SONY HDR-CX700，该相机的帧率为50帧/每秒。此外，我们尝试使用了两个iPhone手机拍摄了一组实验数据。所有拍摄的视频数据的分辨率为1920×1080，为了计算效率的考量，视频降采样为960×540。所有数据均在处理器主频为2.7GHz的英特

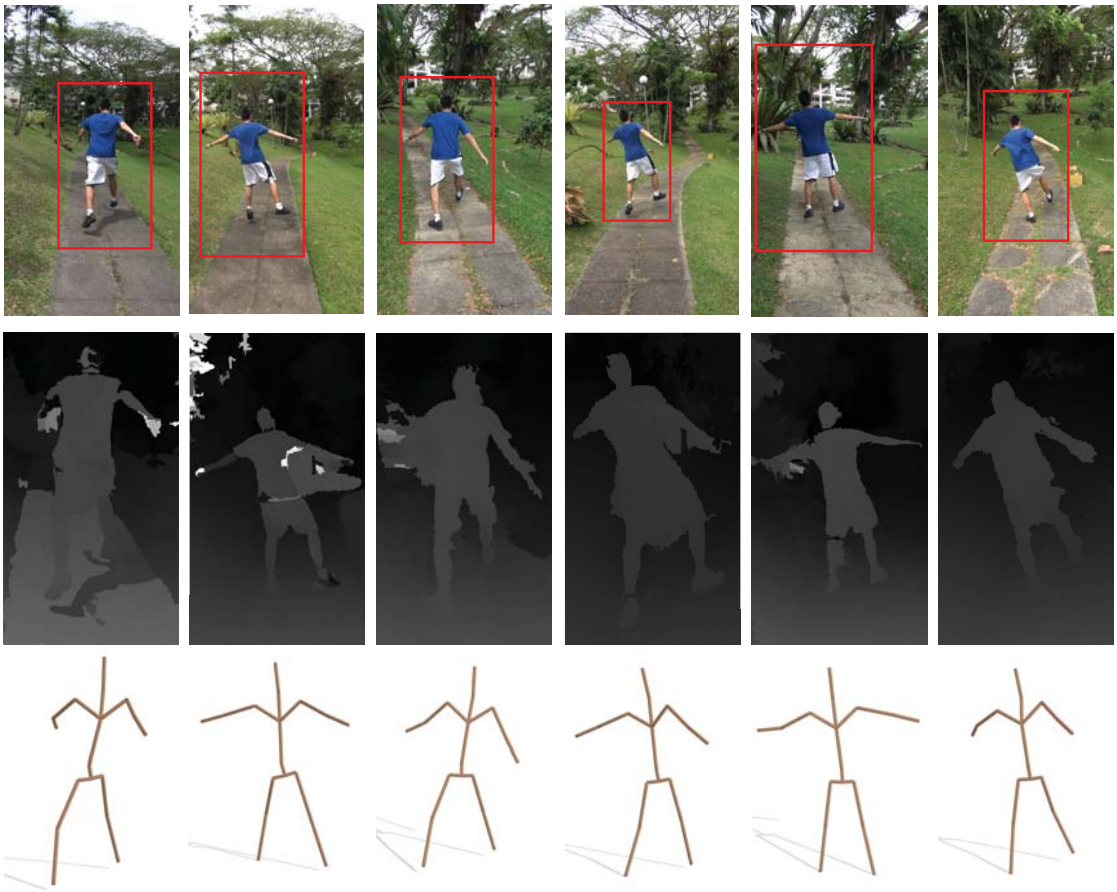


图 4.8　快速运动超过50m的一个例子

尔酷睿i7处理器、显卡为英伟达Quadro FX1800、内存为16G 的计算机上进行了测试。其中，三维点云重建耗时大约为每帧每相机1分钟；运动捕捉大约为每帧每相机15s。当然，所有测试用的代码并没有做任何的优化加速。

大量的测试数据表明本文提出的系统具有比较高的可靠性。其中，图4.1显示了一组由两个手机拍摄的例子，需要注意的是运动对象大约有50%的像素占比。与传统的运动恢复结构不同，协同视觉即时定位与地图重建仍然可以较好地恢复出相机的运动参数。如图4.1(b)所示，我们的立体重建方法可以恢复出高精度的三维点云。运动捕捉重建的骨架在图4.1(c)中显示出来，并且我们可以将捕捉的运动映射到虚拟物体上，如图4.1(d)所示。如此简单的装置系统证明我们的方法可以拓宽无标记运动捕捉的应用范围，具有广泛地应用前景。

图4.8显示了一个大范围场景的快速运动例子，其中运动对象在前面运动，后面跟着两个相机拍摄。该运动对象移动速度很快，并且超过50m。本文提出的方法仍能较好地恢复出运动对象的骨架运动信息。第二行显示了估计的深度图，值得注意的是，在某些遮挡区域存在比较严重的深度估计错误，但我们的运动捕捉方法仍能鲁棒地恢复出骨架姿态。

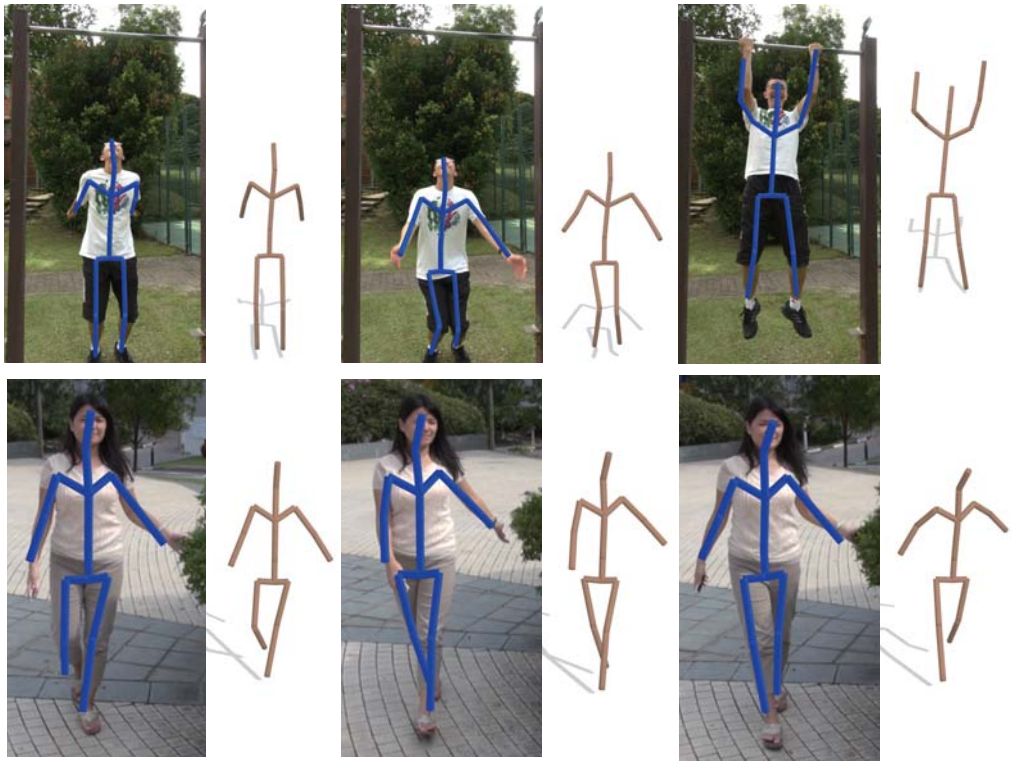


图 4.9　另外两个例子

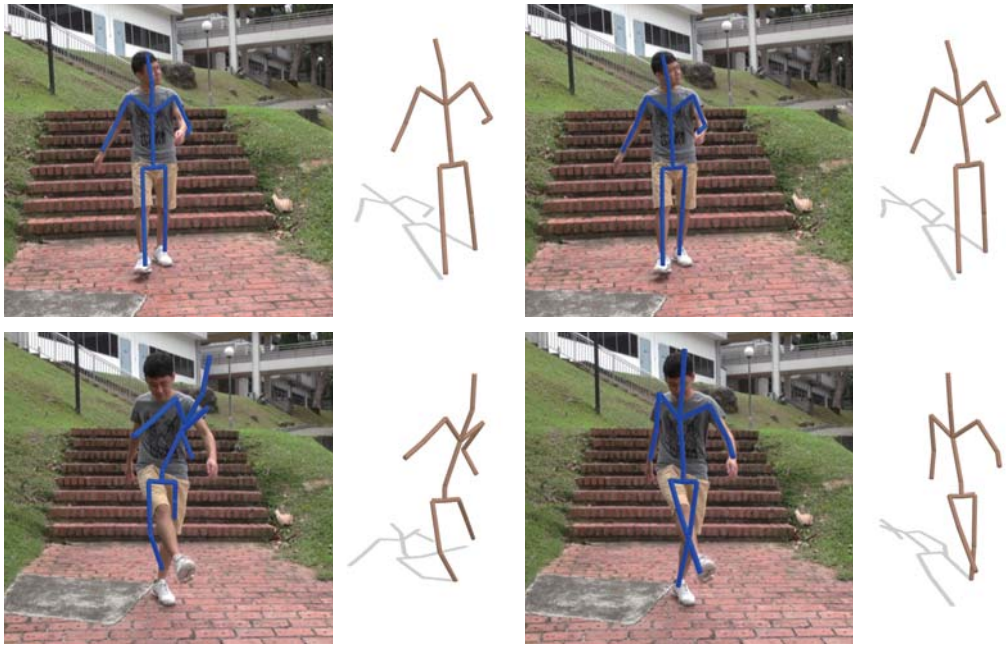


图 4.10　与Gall方法[41]的比较结果，左图和右图分别显示了Gall的方法和我们的方法

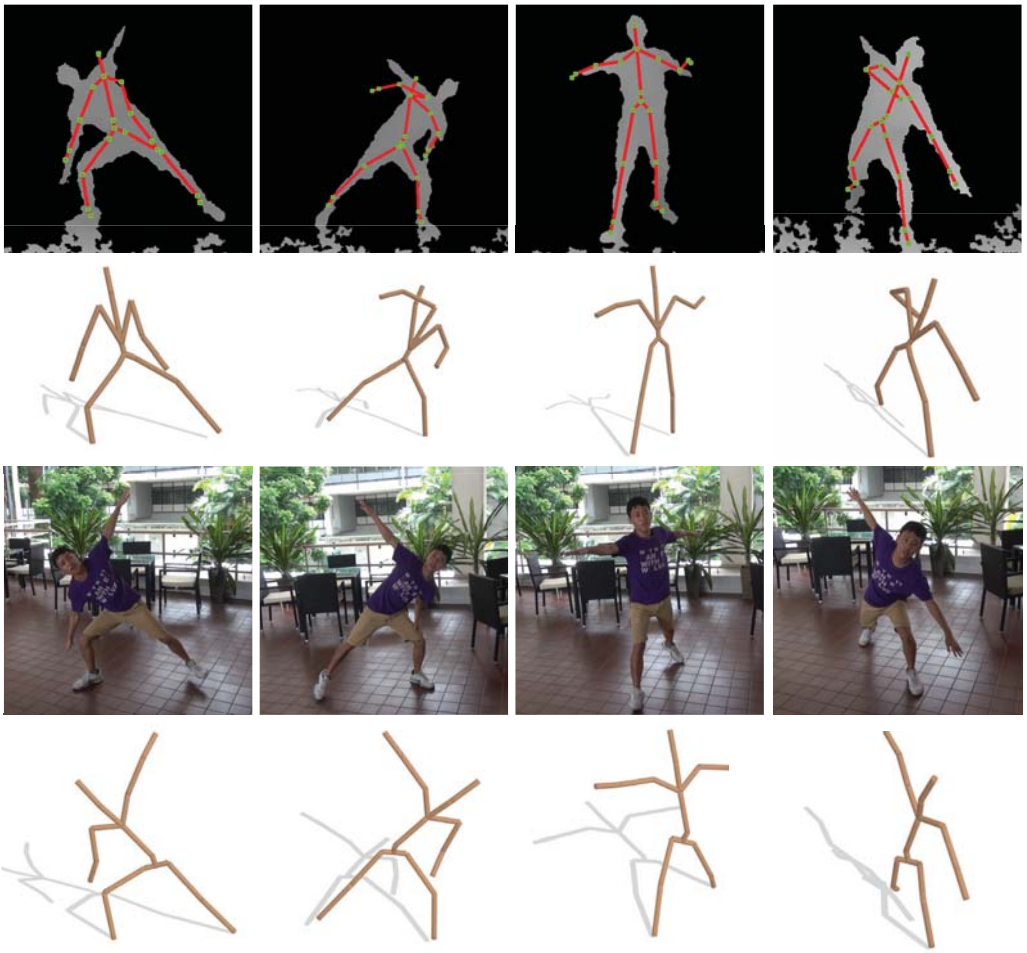


图 4.11　与Kinect方法[46]的比较结果

图4.9显示了另外两个例子。第一行显示了引体向上的运动捕捉；第二行则显示了一个女孩沿着花坛边在走动的例子。为了更好地评价运动捕捉的精度，我们将恢复出来的骨架附着在彩色图像上。

- **与Jiang的方法[94]比较** 我们将本文提出的立体重建方法与文献[94]的方法进行了比较，然后使用两种不同方法得到的深度图来做运动捕捉。图4.7中，蓝色和橙色的帧分别显示了Jiang 的方法与我们方法得到的不同深度图。从图4.7中可以看出，Jiang的方法可以获得比本文方法更好的深度图。但值得一提的是，本文提出的方法比他们的方法快10 倍。尽管如此，两种方法获得骨架姿态是十分接近的。这也就间接证明了我们的运动捕捉方法对于噪声较高的深度具有比较优良的鲁棒性。
- **与Gall的方法[41]比较** 我们也比较了文献[41]提出的运动捕捉方法，其也同时运用了线性和非线性的优化方法，但是他们没有考虑点云的信息。图4.10中左图显示的是Gall的方法，右图则显示了我们的方法得出的运动捕捉结果。从图中也可以看出，我们的方法可以得到更好的骨架姿态估计效果，原因在

于我们不仅考虑了深度信息，还针对少量视角设计了稀疏约束及纹理更新的策略。

- **与Kinect的结果[46]比较** 此外，我们还与微软Kinect的骨架姿态估计结果进行了比较。图4.11 显示了一个半开放的室外场景。第一行和第二行分别显示了Kinect获得的深度图以及从它的SDK中估计的骨架；最后两行则显示了我们的深度图与骨架的骨架。显然，Kinect的骨架估计结果很容易受到户外糟糕的深度传感器的影响，而我们的方法可以获得更好的运动捕捉效果。

4.6　本章小结

本章提出了一种针对大规模场景的人体运动捕捉方法与系统，该方法允许拍摄者手持相机，对运动对象进行跟踪拍摄。实际上，为了实现大规模场景的无标记运动捕捉任务，运动对象需要具有很高的图像像素占比。而针对这一情形，传统的运动恢复结构的技术很难有效地估计出相机参数。为此，本章使用了视觉协同即时定位与地图重建的技术，获得了移动相机的运动参数之后，提出了一种有效的立体重建方法，可以重建出稠密的运动对象三维点云信息。此外，为了提高运动捕捉方法对移动相机，以及少量视角的鲁棒性，我们引入了一个新的稀疏运动先验知识和动态纹理跟新方法。实验表明，我们提出的方法能够在简单装置下获得较好的运动捕捉结果，例如两个手持手机相机镜头，我们相信这样一种装置可以大大拓宽无标记运动捕捉的应用范围。

目前，本章提出的系统还存在着诸多问题，主要表现为系统的稳定性还有待提高。一方面，由于户外场景环境不可控，运动跟踪的结果很容易受到背景噪声的影响；另外一方面，在跟踪拍摄的时候，很难做到全角度的拍摄，因此总会存在遮挡等信息丢失的问题，而这进一步加剧了运动捕捉的难度。展望本章的研究工作，一种可能性是引入人体运动的数据库信息，减少遮挡带来的歧义性。此外，沿用前一章提出的基于物理约束的动力学仿真模型，在求解过程中可以规避大量不合理的姿态，是值得探索的可行方案。

第 5 章　总结与展望

5.1　本文工作总结

本文针对无标记运动捕捉面临的两大技术挑战，也即捕捉高精度的运动信息、实现运动信息的稳定捕捉，根据运动环境的复杂程度，提出了三种无标记运动捕捉系统，分别涉及单相机系统、固定多相机系统和移动多相机系统，特别地，实现了实时脸部表情运动捕捉、手与物体交互运动捕捉、以及户外大规模场景的人体运动捕捉目标。

单相机系统的实时无标记脸部表情运动捕捉系统，使用新型可直接输出深度的单相机系统，提出脸部的自适应动态表情模型，降低了求解空间的维度，解决了新型单相机系统计算时间复杂度与运动捕捉精度的矛盾。由于脸部表情运动属于近距离场景下的少遮挡运动，具有极强易用性以及较低搭建成本的单相机系统，足以满足视觉信号的采集任务。但正如引言中介绍，传统的单相机系统只能得到二维的视觉信号，在投影成像的过程中丢失了深度信息，单相机运动捕捉在系统的易用性与捕捉的运动精度之间存在着固有矛盾。一种思路是，使用深度信息辅助实现高精度的三维运动捕捉。然而，从二维图像中恢复出深度信息是计算视觉的经典问题，为此，本文首先讨论了单相机系统的深度估计问题，建立了单张图像与深度图映射的非线性参数模型，提出了基于该参数模型的单张彩色图像的深度估计方法。值得一提的是，随着电子工业技术的发展，单相机系统的深度估计可直接由电子芯片完成，同步输出场景的深度信号，拓宽了单相机系统的信号采集能力。本文紧接着针对脸部表情的运动捕捉问题，高效利用了新型传感单元输出的深度信息，实现了实时无标记脸部表情运动捕捉与运动映射目标。

固定多相机系统的无标记手与物体交互运动捕捉系统，可以捕捉灵活的手与物体的交互运动，由于该系统不需要标记点、手套以及传感器，因此对手与物体的无标记运动捕捉而言具有很大的吸引力。本文提出的固定多相机系统结合了基于视频的运动捕捉和基于物理约束的运动建模两种技术的优势。从数学上来讲，基于物理的运动建模是一个病态问题，因为有很多种调整策略使得运动满足物理动力学方程，但是仅有少部分的运动与真实的图像数据吻合。通过同时考虑物理约束以及观测的图像数据，我们可以仿真出与真实图像数据匹配的物理运动结果。当然，基于多视点视频的运动捕捉也可以利用物理约束来减少模型的不确定性，确保重建的运动满足合理的物理约束。简而言之，本文提出的系统可以很自

然地对手的关节运动、物体的运动，以及它们之间相互接触的细微运动进行建模，通过建立运动过程中手的姿态、物体的姿态以及相互之间细微接触的完整动力学运动方程，提出复合运动控制器模型，改进和完善传统的手与物体交互运动捕捉方法，得到与观测图像一致的符合物理约束的逼真运动捕捉结果，有效解决了包含严重遮挡的近距离场景无标记运动捕捉难题。

移动多相机系统的户外无标记人体运动捕捉系统，针对大规模室外场景、不可控环境下的无标记运动捕捉问题，首先恢复出可移动手持相机的随时间变化的空间位置关系，针对宽基线的可移动相机系统，提出一种新的运动对象稠密点云的计算方法，利用输入的多视点视频信号以及计算的点云信息，恢复出人体全身的骨架运动。为了提高骨架运动信息恢复随时间变化的稳定性，与传统的方法相比，本文提出的系统在求解过程中引入一种新的稀疏性约束，并且提出与视点相关的动态纹理模型，解决了室外场景下大规模运动的无标记运动捕捉问题。特别地，移动多相机系统甚至可以在两个手机摄像头的采集环境下有效地实现人体的无标记运动捕捉，可极大地拓宽无标记运动捕捉的应用范围。

5.2　未来工作展望

单相机系统的无标记脸部表情运动捕捉系统无需用户进行任何预处理、预校准，可实现特定用户的实时无标记脸部表情运动捕捉，为用户节约了大量的训练时间，提高了运动捕捉的性能和易用性，可广泛应用于消费级市场。但由于当前捕捉深度信息的单相机系统分辨率很低，且获得的深度信号噪声较大，基于新型直接获取深度传感器的单相机方法在稳定性方面还有待提高。考虑到新型传感器技术的不断发展，这一缺陷并构成本文提出的单相机系统的主要问题。实际上，如何获得更高精度的融合形状才是本文提出方法最大的考量因素，其原因在于该方法使用了单位PCA模型建立中性脸，这一假设在大尺度范围能够有效的工作，但是当需要获得更高精度的中性脸，例如额头皱纹等脸部细节时，基于PCA模型的方法无法体现出来。此外，本文使用的动态模板表情也存在着相应的局限，一般来讲，儿童和成年人具有不同的动态表情模型，因此该系统可能需要针对不同年龄段的人群做一区别对待。而由于本文使用的单位PCA模型仅包含成年人，因此有关这方面的问题还需要进一步分析与讨论。

固定多相机系统的无标记手与物体交互运动捕捉系统考虑交互运动过程中的内在动力学物理属性，建立运动过程中手的姿态、物体的姿态以及相互之间细微接触的完整动力学运动方程，提出复合运动控制器模型，改进和完善传统的手与物体交互运动捕捉方法，有效解决了包含严重遮挡的近距离场景无标记运动捕捉

难题。但是交互运动计算的时间复杂度还很高，离真正的实用化还有较大的提升空间。如何使用并行计算或图形处理器的方式来加速当前的运动捕捉模块是值得关注的一个研究方向，尤其是基于多视点图像的骨架运动跟踪，以及基于复合控制器的物理仿真。此外，在未来的研究工作中，我们还可以考虑如何利用已经捕捉的运动数据，通过修缮、重用数据库中的运动数据实现完全虚拟运动的合成与制作，提高当前制作交互运动视频的效率。

移动多相机系统的无标记人体运动捕捉系统使用两个甚至多个可手持移动的相机，通过恢复移动相机的相对空间位置以及运动对象的稠密点云，引入运动的稀疏性约束以及与视点相关的动态纹理更新框架，解决了室外不可控环境、大规模场景下的无标记运动捕捉问题。尽管我们已经验证了移动计算设备，例如手机，可以满足运动捕捉系统数据采集的任务，但整个系统的计算代价与复杂度仍然没有得到有效地解决，离正在的移动平台还有相当大的一段距离。此外，系统的稳定性与精度也有很大的提升空间。目前，系统还无法达到任意场景、任意运动的精准捕捉，主要原因在于不可控环境与大规模场景的复杂性，以及对象运动的不可预知性。随着近年来移动端视频数据的规模逐年增加，有关机器学习、数据挖掘方法的不断发展，运动的不可预知性在一定程度上可以通过某些先验知识得以限定，进而可以提高移动多相机系统的稳定与可靠性。此外，在未来的研究中，我们也期望众多的新型传感器能够介入移动相机平台中，通过捕捉深度、光谱等其他信息也可以使运动捕捉的精度与鲁棒性得以更好地解决。

最后，总结全文，我们关注的问题是无标记运动捕捉与运动映射，核心方法是基于无接触的视觉信号，关键任务有两个，一是在空间分辨率上，实现高精度的运动重构，二是在时间分辨率上，实现鲁棒的运动跟踪。为此，根据运动场景的复杂程度，我们分别研究了近距离场景下无遮挡或少遮挡的运动捕捉，近距离场景下存在严重遮挡的运动捕捉，以及大规模场景的运动捕捉。并且针对这些场景，分别选择了脸部表情运动、手与物体的交互运动、人体运动作为具体的研究对象。事实上，自然界的运动千变万化，而运动是物质的固有属性，我们不可能将所有的运动都纳入无标记运动捕捉的范畴。本文讨论的无标记运动捕捉方法与系统力求在实现具体运动对象的捕捉任务前提下，解决当前无标记运动捕捉技术面临的挑战，发展和完善当前无标记运动捕捉的理论与技术，实现无标记运动捕捉方法与系统从工业生产应用向民用消费级市场的转变。

参考文献

[1] Menache A. Understanding motion capture for computer animation and video games. Burlington, Massachusetts: Morgan Kaufmann, 2000.

[2] Williams A L. Republic of Images: A History of French Filmmaking. Cambridge, Massachusetts: Harvard University Press, 1992.

[3] FLEISCHER M. Method of producing moving-picture cartoons, October 9, 1917. US Patent 1,242,674.

[4] Moeslund T B, Hilton A, Krüger V. A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding, 2006, 104(2):90–126.

[5] Poppe R. Vision-based human motion analysis: An overview. Computer Vision and Image Understanding, 2007, 108(1):4–18.

[6] Weise T, Bouaziz S, Li H, et al. Realtime Performance-Based Facial Animation. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2011, 30(4):77.

[7] Wang R Y, Popović J. Real-time hand-tracking with a color glove. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2009, 28(3):63.

[8] Wu C, Stoll C, Valgaerts L, et al. On-set performance capture of multiple actors with a stereo camera. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2013, 32(6):161.

[9] Gall J, Rosenhahn B, Seidel H P. An introduction to interacting simulated annealing. Proceedings of Human Motion: Understanding, Modelling, Capture, and Animation. Houten, Netherlands: Springer Netherlands, 2008: 319–345.

[10] Ganapathi V, Plagemann C, Koller D, et al. Real time motion capture using a single time-of-flight camera. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2010.

[11] Wei X, Zhang P, Chai J. Accurate realtime full-body motion capture using a single depth camera. ACM Trans. Graph., 2012, 31(6):188.

[12] Morph target animation. http://en.wikipedia.org/wiki/Morph_target_animation.

[13] Szeliski R. Computer vision: algorithms and applications. London, United Kingdom: Springer London, 2010.

[14] Cao X, Li Z, Dai Q. Semi-automatic 2d-to-3d conversion using disparity propagation. IEEE Transactions on Broadcasting, 2011, 57(2):491–499.

[15] Pighin F, Lewis J P. Performance-Driven Facial Animation. Proceedings of ACM SIGGRAPH Courses, 2006.

[16] Zhang L, Snavely N, Curless B, et al. Spacetime faces: high resolution capture for modeling and animation. ACM Trans. Graph., 2004, 23(3):548–558.

[17] Weise T, Li H, Gool L V, et al. Face/Off: Live Facial Puppetry. Proceedings of ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA), 2009.

[18] Furukawa Y, Ponce J. Dense 3D motion capture for human faces. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2009.

[19] Bradley D, Heidrich W, Popa T, et al. High Resolution Passive Facial Performance Capture. ACM Trans. Graph., 2010, 29(3):41.

[20] Beeler T, Hahn F, Bradley D, et al. High-quality passive facial performance capture using anchor frames. ACM Trans. Graph., 2011, 30(4):75.

[21] Valgaerts L, Wu C, Bruhn A, et al. Lightweight binocular facial performance capture under uncontrolled lighting. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2012, 31(6):187.

[22] Chai J, Xiao J, Hodgins J. Vision-based control of 3D facial animation. Proceedings of ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA), 2003.

[23] Saragih J M, Lucey S, Cohn J F. Real-time avatar animation from a single image. Proceedings of Automatic Face and Gesture Recognition (FG), 2011.

[24] Baltrušaitis T, Robinson P, Morency L P. 3D Constrained Local Model for Rigid and Non-Rigid Facial Tracking. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2012.

[25] Liu C K. Dextrous manipulation from a grasping pose. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2009, 28(3):59.

[26] Rehg J M, Kanade T. Model-based tracking of self-occluding articulated objects. Proceedings of International Conference on Computer Vision (ICCV), 1995.

[27] La Gorce M, Fleet D J, Paragios N. Model-based 3d hand pose estimation from monocular video. IEEE Trans. Pattern Analysis and Machine Intelligence, 2011, 33(9):1793–1805.

[28] Wu Y, Lin J Y, Huang T S. Capturing natural hand articulation. Proceedings of International Conference on Computer Vision (ICCV), 2001.

[29] Zhou H, Huang T S. Tracking articulated hand motion with eigen dynamics analysis. Proceedings of International Conference on Computer Vision (ICCV), 2003.

[30] Athitsos V, Sclaroff S. Estimating 3D hand pose from a cluttered image. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2003.

[31] Romero J, Kjellstrom H, Kragic D. Hands in action: real-time 3D reconstruction of hands in interaction with objects. Proceedings of International Conference on Robotics and Automation (ICRA), 2010.

[32] Hoyet L, Ryall K, McDonnell R, et al. Sleight of hand: perception of finger motion from reduced marker sets. Proceedings of ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games (I3D), 2012.

[33] Zhao W, Chai J, Xu Y Q. Combining marker-based mocap and rgb-d camera for acquiring high-fidelity hand motion data. Proceedings of ACM/Eurographcs Symposium on Computer Animation (SCA), 2012.

[34] Ballan L, Taneja A, Gall J, et al. Motion capture of hands in action using discriminative salient points. Proceedings of European Conference on Computer Vision (ECCV), 2012.

[35] Oikonomidis I, Kyriazis N, Argyros A A. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. Proceedings of International Conference on Computer Vision (ICCV), 2011.

[36] Liu C K. Synthesis of interactive hand manipulation. Proceedings of ACM/Eurographcs Symposium on Computer Animation (SCA), 2008.

[37] Liu Y, Dai Q, Xu W. A point-cloud-based multiview stereo algorithm for free-viewpoint video. IEEE Trans. on Visualization and Computer Graphics, 2010, 16(3):407–418.

[38] Deutscher J, Blake A, Reid I. Articulated body motion capture by annealed particle filtering. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2000.

[39] Bregler C, Malik J, Pullen K. Twist based acquisition and tracking of animal and human kinematics. International Journal of Computer Vision, 2004, 56(3):179–194.

[40] Gall J, Rosenhahn B, Brox T, et al. Optimization and filtering for human motion capture. International Journal of Computer Vision, 2010, 87(1-2):75–92.

[41] Gall J, Stoll C, De Aguiar E, et al. Motion capture using joint skeleton tracking and surface estimation. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2009.

[42] Vlasic D, Baran I, Matusik W, et al. Articulated mesh animation from multi-view silhouettes. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2008, 27(3):97.

[43] Wu C, Varanasi K, Theobalt C. Full body performance capture under uncontrolled and varying illumination: A shading-based approach. Proceedings of European Conference on Computer Vision (ECCV), 2012.

[44] Elhayek A, Stoll C, Hasler N, et al. Spatio-temporal motion tracking with unsynchronized cameras. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2012.

[45] Wei X, Chai J. Videomocap: modeling physically realistic human motion from monocular video sequences. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2010, 29(4):42.

[46] Shotton J, Fitzgibbon A, Cook M, et al. Real-time Human Pose Recognition in Parts from Single Depth Images. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2011.

[47] Baak A, Müller M, Bharaj G, et al. A data-driven approach for real-time full body pose reconstruction from a depth camera. Proceedings of Consumer Depth Cameras for Computer Vision, 2013.

[48] Taylor J, Shotton J, Sharp T, et al. The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2012.

[49] Hasler N, Rosenhahn B, Thormahlen T, et al. Markerless Motion Capture with Unsynchronized Moving Cameras. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2009.

[50] Vlasic D, Adelsberger R, Vannucci G, et al. Practical Motion Capture in Everyday Surroundings. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2007, 26(3):35.

[51] Shiratori T, Park H S, Sigal L, et al. Motion Capture from Body-mounted Cameras. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2011, 30(4).

[52] Baker S, Szeliski R, Anandan P. A layered approach to stereo reconstruction. Proceedings of Computer Vision and Pattern Recognition (CVPR), 1998.

[53] Saxena A, Chung S, Ng A. 3-d depth reconstruction from a single still image. International Journal of Computer Vision, 2008, 76(1):53–69.

[54] Hoiem D, Efros A A, Hebert M. Automatic photo pop-up. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2005, 24(3):577–584.

[55] Delage E, Lee H, Ng A Y. A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2006.

[56] Liu B, Gould S, Koller D. Single image depth estimation from predicted semantic labels. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2010.

[57] Saxena A, Sun M, Ng A. Make3d: Learning 3d scene structure from a single still image. IEEE Trans. Pattern Analysis and Machine Intelligence, 2009, 31(5):824–840.

[58] Karsch K, Liu C, Kang S B. Depth Extraction from Video Using Non-parametric Sampling. Proceedings of European Conference on Computer Vision (ECCV), 2012.

[59] Mairal J, Bach F, Ponce J, et al. Online dictionary learning for sparse coding. Proceedings of International Conference on Machine Learning (ICML), 2009.

[60] Fu W J. Penalized Regressions: The Bridge versus the Lasso. Journal of Computational and Graphical Statistics, 1998, 7(3):397–416.

[61] Amberg B, Blake A, Vetter T. On compositional Image Alignment, with an application to Active Appearance Models. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2009.

[62] Lin I C, Ouhyoung M. Mirror MoCap: Automatic and efficient capture of dense 3D facial motion parameters from video. The Visual Computer, 2005, 21(6):355–372.

[63] Deng Z, Chiang P Y, Fox P, et al. Animating blendshape faces by cross-mapping motion capture data. Proceedings of ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games (I3D), 2006.

[64] Ma W C, Jones A, Chiang J Y, et al. Facial performance synthesis using deformation-driven polynomial displacement maps. ACM Trans. Graph., 2008, 27(3):121.

[65] Bickel B, Lang M, Botsch M, et al. Pose-Space Animation and Transfer of Facial Details. Proceedings of ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA), 2008.

[66] Huang H, Chai J, Tong X, et al. Leveraging motion capture and 3D scanning for high-fidelity facial performance acquisition. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2011, 30(4):74.

[67] Blanz V, Vetter T. A morphable model for the synthesis of 3D faces. ACM Trans. Graph. (Proceedings of SIGGRAPH), 1999, 32(3):187–194.

[68] Sumner R W, Popović J. Deformation transfer for triangle meshes. ACM Trans. Graph., 2004, 23(3):399–405.

[69] Li H, Weise T, Pauly M. Example-Based Facial Rigging. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2010, 29(3):32.

[70] Botsch M, Sumner R, Pauly M, et al. Deformation Transfer for Detail-Preserving Surface Editing. Proceedings of Vision, Modeling & Visualization, 2006.

[71] Levy B, Zhang R H. Spectral Geometry Processing. Proceedings of ACM SIGGRAPH Courses, 2010.

[72] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2001.

[73] Rusinkiewicz S, Levoy M. Efficient Variants of the ICP Algorithm. Proceedings of International Conference on 3D Digital Imaging and Modeling (DIM), 2001.

[74] Botsch M, Kobbelt L, Pauly M, et al. Polygon Mesh Processing. Natick, Massachusetts: A K Peters, 2010.

[75] Madsen K, Nielsen H B, Tingleff O. Methods for Non-Linear Least Squares Problems (2nd ed.). Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby: Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2004.

[76] Barrett R, Berry M, Chan T F, et al. Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, 2nd Edition. Philadelphia, PA: SIAM, 1994.

[77] Brubaker M A, Fleet D J. The kneed walker for human pose tracking. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2008.

[78] Vondrak M, Sigal L, Jenkins O C. Physical simulation for probabilistic motion tracking. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2008.

[79] Vondrak M, Sigal L, Hodgins J, et al. Video-based 3d motion capture through biped control. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2012, 31(4):27.

[80] Pollard N S, Zordan V B. Physically based grasping control from example. Proceedings of ACM/Eurographcs Symposium on Computer Animation (SCA), 2005.

[81] Kry P G, Pai D K. Interaction capture and synthesis. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2006, 25(3):872–880.

[82] Tsang W, Singh K, Fiume E. Helping hand: an anatomically accurate inverse dynamics solution for unconstrained hand motion. Proceedings of ACM/Eurographcs Symposium on Computer Animation (SCA), 2005.

[83] Sueda S, Kaufman A, Pai D K. Musculotendon simulation for hand animation. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2008, 27(3):83.

[84] Ye Y, Liu C K. Synthesis of detailed hand manipulations using contact sampling. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2012, 31(4):41.

[85] Zhang Z. Flexible camera calibration by viewing a plane from unknown orientations. Proceedings of International Conference on Computer Vision (ICCV), 1999.

[86] Debevec P, Yu Y, Borshukov G. Efficient view-dependent image-based rendering with projective texture-mapping. Vienna, Austria: Springer Vienna, 1998.

[87] Lee S H, Goswami A. Ground reaction force control at each foot: A momentum-based humanoid balance controller for non-level and non-stationary ground. Proceedings of International Conference on Intelligent Robots and Systems (IROS), 2010.

[88] Liu L, Yin K, Panne M, et al. Sampling-based contact-rich motion control. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2010, 29(4):128.

[89] Li G, Wu C, Stoll C, et al. Capturing Relightable Human Performances under General Uncontrolled Illumination. Comput. Graph. Forum, 2013, 32(2):275–284.

[90] De Aguiar E, Stoll C, Theobalt C, et al. Performance capture from sparse multi-view video. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2008, 27(3):98.

[91] Taneja A, Ballan L, Pollefeys M. Modeling Dynamic Scenes Recorded with Freely Moving Cameras. Proceedings of Asian Conference of Computer Vision (ACCV), 2010.

[92] Hartley R, Zisserman A. Multiple View Geometry in Computer Vision. 2 ed., New York, NY, USA: Cambridge University Press, 2003.

[93] Zou D, Tan P. CoSLAM: Collaborative Visual SLAM in Dynamic Environments. IEEE Trans. Pattern Analysis and Machine Intelligence, 2013, 35(2):354–366.

[94] Jiang H, Liu H, Tan P, et al. 3D Reconstruction of Dynamic Scenes with Multiple Handheld Cameras. Proceedings of European Conference on Computer Vision (ECCV), 2012.

[95] Royer E, Lhuillier M, Dhome M, et al. Localization in urban environments: monocular vision compared to a differential GPS sensor. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2005.

[96] Mouragnon E, Lhuillier M, Dhome M, et al. Real time localization and 3d reconstruction. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2006.

[97] Davison A J. Real-time simultaneous localisation and mapping with a single camera. Proceedings of International Conference on Computer Vision (ICCV), 2003.

[98] Eade E, Drummond T. Scalable monocular SLAM. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2006.

[99] Strasdat H, Montiel J M, Davison A J. Visual slam: Why filter? Image and Vision Computing, 2012, 30(2):65–77.

[100] Nistér D, Naroditsky O, Bergen J. Visual odometry. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2004.

[101] Paz L M, Piniés P, Tardós J D, et al. Large-scale 6-DOF SLAM with stereo-in-hand. IEEE Trans. on Robotics, 2008, 24(5):946–957.

[102] Seitz S M, Curless B, Diebel J, et al. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2006.

[103] Zhang Y, Kambhamettu C. Integrated 3D scene flow and structure recovery from multiview image sequences. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2000.

[104] Tao H, Sawhney H, Kumar R. Dynamic depth recovery from multiple synchronized video streams. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2001.

[105] Zitnick C L, Kang S B, Uyttendaele M, et al. High-quality Video View Interpolation Using a Layered Representation. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2004, 23(3):600–608.

[106] Larsen E S, Mordohai P, Polleyfeys M, et al. Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. Proceedings of International Conference on Computer Vision (ICCV), 2007.

[107] Yang M, Cao X, Dai Q. Multiview video depth estimation with spatial-temporal consistency. Proceedings of British Machine Vision Conference (BMVC), 2010.

[108] Ballan L, Brostow G J, Puwein J, et al. Unstructured Video-based Rendering: Interactive Exploration of Casually Captured Videos. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2010, 29(4):87.

[109] Sun J, Zheng N N, Shum H Y. Stereo Matching Using Belief Propagation. IEEE Trans. Pattern Analysis and Machine Intelligence, 2003, 25(7):787–800.

[110] Zhang G, Jia J, Wong T T, et al. Consistent Depth Maps Recovery from a Video Sequence. IEEE Trans. Pattern Analysis and Machine Intelligence, 2009, 31(6):974–988.

[111] Tola E, Lepetit V, Fua P. DAISY: An Efficient Dense Descriptor Applied to Wide-Baseline Stereo. IEEE Trans. Pattern Analysis and Machine Intelligence, 2010, 32(5):815–830.

[112] Comaniciu D, Meer P. Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE Trans. Pattern Analysis and Machine Intelligence, 2002, 24(5):603–619.

[113] Baran I, Popovic J. Automatic rigging and animation of 3D characters. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2007, 26(3):72.

[114] Bai X, Wang J, Simons D, et al. Video SnapCut: Robust Video Object Cutout Using Localized Classifiers. ACM Trans. Graph. (Proceedings of SIGGRAPH), 2009, 28(3):70.

[115] Xiao J, Wang J, Tan P, et al. Joint affinity propagation for multiple view segmentation. Proceedings of International Conference on Computer Vision (ICCV), 2007.

[116] Murray R, Li Z, Sastry S. A Mathematical Introduction to Robotic Manipulation. Boca Raton, Florida: CRC Press, 1994.

致　谢

衷心感谢导师戴琼海教授对本人的帮助与指导。3D作为蓬勃发展的朝阳产业，受到工业界和学术界持续不断的关注，是戴老师将我引入到与此相关的学术研究中。他治学严谨，对待工作一丝不苟，教导我研究当持之以恒，贵在勤奋。戴老师身体力行，对待科研的满腔热血一直鼓舞我刻苦努力，他以身作则的态度将使我终生受益。

感谢BBNC的所有老师和同学，你们的热情帮助与支持贯穿了我的整个博士阶段！感谢瑞士洛桑联邦理工大学的Pauly Mark教授，美国德克萨斯农机大学的柴金祥教授，新加坡国立大学的谭平教授和微软亚洲研究院的童欣研究员，与他们的讨论让我获益良多。

感谢我的家人多年来的支持，特别感谢我的夫人——彭聪女士的理解、包容以及鼓励，家庭与亲情永远是我奋斗的源动力！

本课题承蒙国家自然科学基金、清华大学博士生短期出国访学基金、清华大学国际会议基金资助，特此致谢！

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

附录 A　线性脸部表情映射算子

本附录将介绍脸部表情映射算子的推导过程。也即是，如何将模板中性融合形状$\mathbf{b}_0^*$与其他表情融合形状$\mathbf{b}_i^*$之间的变形关系，作用于特定对象的中性融合形状$\mathbf{b}_0$，得到特定对象的其他表情融合形状$\mathbf{b}_i$。为此，我们首先根据模板融合形状$\mathbf{b}_0^*$ 与$\mathbf{b}_i^*$ 中的p组对应三角面片，求出这p组对应三角面片的变形仿射变换矩阵，构成仿射变换集合$\{\mathbf{S}_1^*,\ldots,\mathbf{S}_p^*\}$。需要说明的是，由于仿射变换矩阵无法唯一确定每个对应三角形的变形关系，我们需要引入一个垂直于三角形平面的额外顶点，将三角形构成四面体来计算仿射变换矩阵。

假设模板中性融合形状$\mathbf{b}_0^*$的三角面片$\mathbf{v}_0^*$构成的四面体为：$\{\mathbf{v}_{01}^*,\mathbf{v}_{02}^*,\mathbf{v}_{03}^*,\mathbf{v}_{04}^*\}$，

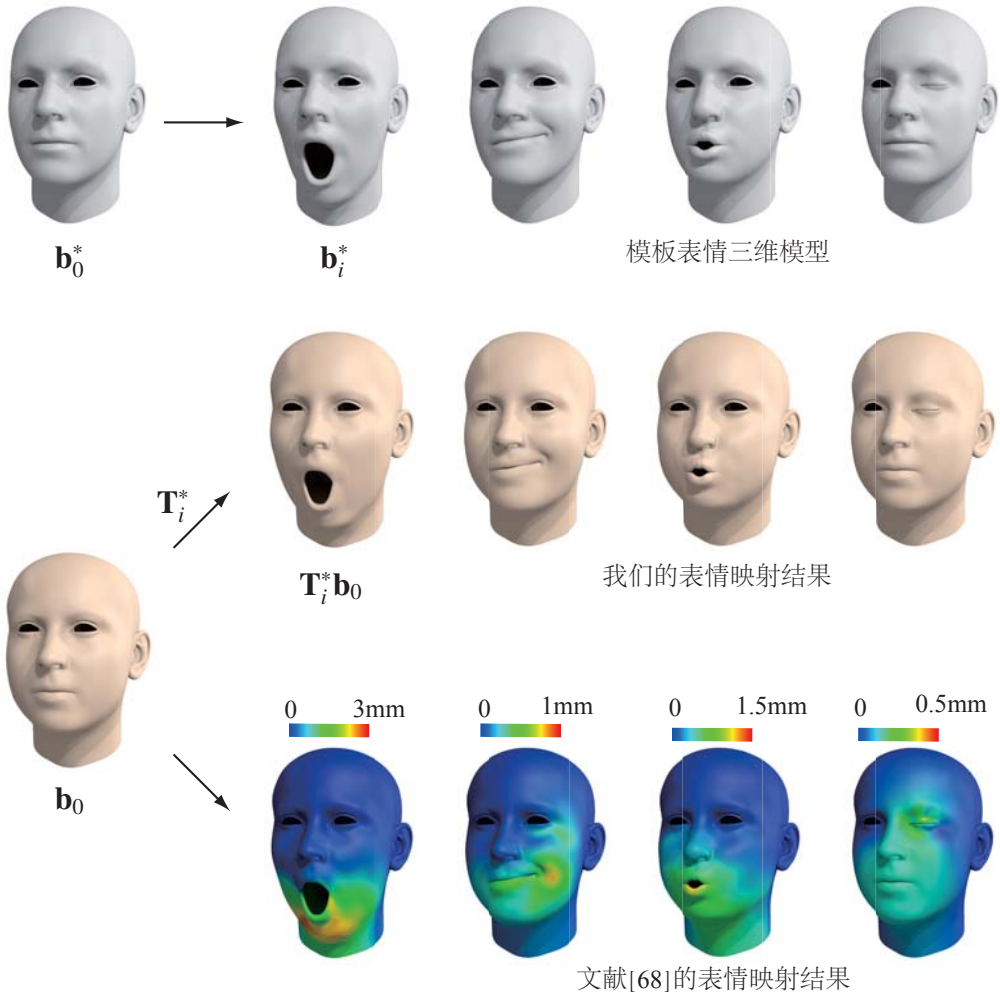


图 A.1　将模板模型（上）传递到特定用户的表情模型（中），与文献[68] 提出的方法相比，可以获得几乎同样的传递效果

其中，$\mathbf{v}_{04}^*$为新加入的顶点。在模板其他表情的融合形状$\mathbf{b}_i^*$中，与$\mathbf{v}_0^*$对应的三角面片假设为$\mathbf{v}_i^*$，其构成的四面体表示为：$\{\mathbf{v}_{i1}^*, \mathbf{v}_{i2}^*, \mathbf{v}_{i3}^*, \mathbf{v}_{i4}^*\}$。

于是，两个三角形对应的仿射变换$\mathbf{S}^*$矩阵可以计算为

$$\mathbf{S}^* = \mathbf{S}_i^* \mathbf{S}_0^{*(-1)}. \tag{A-1}$$

其中，$\mathbf{S}_i^* = [\mathbf{v}_{i2}^* - \mathbf{v}_{i1}^*, \mathbf{v}_{i3}^* - \mathbf{v}_{i1}^*, \mathbf{v}_{i4}^* - \mathbf{v}_{i1}^*]$，$\mathbf{S}_0^* = [\mathbf{v}_{02}^* - \mathbf{v}_{01}^*, \mathbf{v}_{03}^* - \mathbf{v}_{01}^*, \mathbf{v}_{04}^* - \mathbf{v}_{01}^*]$。

我们将变形映射问题描述为优化问题

$$\arg\min_{\mathbf{b}_i} \sum_{j=1}^{p} \|\mathbf{S}_j^* \mathbf{t}_{0j} - \mathbf{t}_{ij}\|_2^2 + \|\mathbf{F}(\mathbf{b}_i - \mathbf{b}_0)\|_2^2. \tag{A-2}$$

其中，$\mathbf{t}_{ij} = [\mathbf{v}_{i2} - \mathbf{v}_{i1}, \mathbf{v}_{i3} - \mathbf{v}_{i1}]_j$是特定对象的融合形状$\mathbf{b}_i$中第$j$个三角面片的两条边；$\mathbf{F}$刻画了$\mathbf{b}_0$变形成$\mathbf{b}_i$需要固定的三角形顶点，其是一个对角矩阵。

进一步地，式（A-2）可以改写成

$$\arg\min_{\mathbf{b}_i} \|\mathbf{H}_i^* \mathbf{G}\mathbf{b}_0 - \mathbf{G}\mathbf{b}_i\|_2^2 + \|\mathbf{F}(\mathbf{b}_i - \mathbf{b}_0)\|_2^2. \tag{A-3}$$

其中，$\mathbf{G}$矩阵是面片顶点与边的转换算子；$\mathbf{H}_i^*$是将模板中性融合形状$\mathbf{b}_0^*$变形成模板融合形状$\mathbf{b}_i^*$的仿射映射集合构成的矩阵。

显然，式（A-3）有闭式最优解

$$\mathbf{b}_i = \mathbf{T}_i^* \mathbf{b}_0. \tag{A-4}$$

其中，

$$\mathbf{T}_i^* = (\mathbf{G}^T\mathbf{G} + \mathbf{F})^{-1}(\mathbf{G}^T\mathbf{H}_i^*\mathbf{G} + \mathbf{F}). \tag{A-5}$$

这样，我们就推导出了使用模板融合形状的表情映射算子，很显然，$\mathbf{T}_i^*$是线性映射算子。值得一提的是，为了避免$\mathbf{T}_i^*$受到$\mathbf{b}_0$的影响，我们使用图的拉普拉斯矩阵（Graph Laplacian），而没有采用反正切拉普拉斯矩阵（Cotan Laplacian）[68,70]，实验结果表明，这样处理不仅对最终的变形结果影响不大（如图A.1 所示），反而可以简化动态表情模型的优化过程。

附录 B　基于骨架的三维模型表面点线性化

针对手与物体的交互运动捕捉，以及人体运动捕捉，我们需要对三维模型进行变形处理。由于骨架模型符合运动对象的客观运动形态，基于骨架的三维模型变形来实现运动捕捉任务是研究人员广泛采用的技术方案。这其中的一项关键技术是：线性混合蒙皮（Linear Blend Skinning，LBS）[113]，其主要思想是，三维模型表面的每一个顶点坐标均由骨架节点的变换加权而来，形式化为

$$v'_p = \left(\sum_{i=1}^{N} w_i T_i \right) v_p. \tag{B-1}$$

其中，v_p和v'_p分别是变换前和变换后的三维模型表面顶点的坐标；N为骨架节点的数目；T_i为第i个骨架节点的变换矩阵；w_i是模型表面顶点相对于第i个骨架节点的权重。模型表面点的坐标变换计算本质上即是求解所有骨架节点的变换矩阵。

一般而言，骨架模型被描述成一个树状结构。以第i个骨架节点为例，由于所有i节点的父节点均会影响i节点的运动，因而，第i个骨架节点的变换矩阵T_i 需要考虑所有i节点父节点的变换。另外值得一提的是，骨架节点的旋转自由度通常不止一个，例如手腕骨架关节即有2个自由度，在计算每个骨架节点的变换矩阵时，要考虑所有与该节点相关的骨架自由度。这里，我们用M表示骨架节点自由度的总数，显然，M与N不一定相等。

本附录将讨论如何将式（B-1）用M个骨架节点自由度来线性化处理。为此，我们引入旋量理论（screw theory）来描述刚体的运动[116]，其基本思想是认为任何一个三维刚性运动可以表示为绕一个三维轴的旋转，以及在该轴上的平移。这样一种螺旋运动（screw motion）形式，可以用运动旋量（twist）来描述。运动旋量有两种表示方法，一种是用6 维向量表示：

$$\xi = [v_1, v_2, v_3, \omega_x, \omega_y, \omega_z]^T, \tag{B-2}$$

另外一种则是用4×4的矩阵形式表示：

$$\hat{\xi} = \begin{bmatrix} 0 & -\omega_z & \omega_y & v_1 \\ \omega_z & 0 & -\omega_x & v_2 \\ -\omega_y & \omega_x & 0 & v_3 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{B-3}$$

式（B-2）和（B-3）中，$\boldsymbol{\omega} = [\omega_x, \omega_y, \omega_z]^T$为单位旋转轴；$\mathbf{v} = [v_1, v_2, v_3]^T$的计算公式为，旋转中心$\mathbf{p}$与旋转轴$\omega$的叉积，即$\mathbf{v} = \mathbf{p} \times \boldsymbol{\omega}$。以第$i$ 个骨架节点为例，

p为i骨架节点的三维坐标；ω是该骨架节点的旋转自由度。如果旋转角度为标量θ，则$\theta\hat{\xi}$可以用来表示三维刚性运动。

根据运动旋量的级联定理[116]，变换T_i可以表示成

$$T_i = e^{\theta_0\hat{\xi}_0} \cdot \prod_{j\in Parent(i), j\neq 0} e^{\theta_j\hat{\xi}_j}. \tag{B-4}$$

式（B-4）中，$Parent(i)$是第i个骨架节点的所有父节点的所有自由度集合；θ_j是第j个自由度的旋转角度；$\hat{\xi}_j$是第j个自由度对应的旋量矩阵。需要指出的是，$\theta_0\hat{\xi}_0$表示全局运动，为了便于讨论，我们将其分离出来。式（B-4）可以进行泰勒展开，线性化为

$$T_i = I + \theta_0\hat{\xi}_0 + \sum_{j\in Parent(i), j\neq 0} \theta_j\hat{\xi}_j. \tag{B-5}$$

其中，I为单位矩阵。

将式（B-5）带入式（B-1）中，并且考虑到$\sum_{i=1}^{N} w_i = 1$，可以得到

$$v'_p = \left(I + \theta_0\hat{\xi}_0 + \sum_{m=1}^{M}\left(\sum_{j\in Children(m)} w_j\right)\theta_m\hat{\xi}_m\right)v_p. \tag{B-6}$$

其中，$Children\,(m)$是第m个自由度对应骨架节点的所有子节点的骨架节点集合。

我们用$\bar{w}_m = \sum_{j\in Children(m)} w_j$，并整理式（B-6）后可以得到

$$\left(\theta_0\hat{\xi}_0 + \sum_{m=1}^{M}\bar{w}_m\theta_m\hat{\xi}_m\right)v_p = v'_p - v_p. \tag{B-7}$$

假设顶点$v_p = [x, y, z, 1]^T, v'_p = [x', y', z', 1]^T$，式（B-7）可以形式化为

$$A\mathbf{x} = b. \tag{B-8}$$

其中，

$$\mathbf{x} = [v_1, v_2, v_3, \omega_x, \omega_y, \omega_z, \theta_1, \cdots, \theta_M]^T,$$

$$b = [x' - x, y' - y, z' - z]^T,$$

且

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & z & -y & \left(v_{1,1} - \omega_{z,1}y + \omega_{y,1}z\right)\bar{w}_1 & \cdots & \left(v_{1,m} - \omega_{z,m}y + \omega_{y,m}z\right)\bar{w}_m \\ 0 & 1 & 0 & -z & 0 & x & \left(v_{2,1} + \omega_{z,1}x - \omega_{x,1}z\right)\bar{w}_1 & \cdots & \left(v_{2,m} + \omega_{z,m}x - \omega_{x,m}z\right)\bar{w}_m \\ 0 & 0 & 1 & y & -x & 0 & \left(v_{3,1} - \omega_{y,1}x + \omega_{x,1}y\right)\bar{w}_1 & \cdots & \left(v_{3,m} - \omega_{y,m}x + \omega_{x,m}y\right)\bar{w}_m \end{bmatrix}.$$

于是，三维模型表面点可以用M个骨架节点自由度线性化表示。

个人简历、在学期间发表的学术论文与研究成果

个人简历

1987 年 11 月 10 日出生于江苏省大丰市。

2005 年 9 月考入东南大学仪器科学与工程学院测控技术与仪器专业，2009 年 7 月本科毕业并获得工学学士学位。

2009 年 9 月免试进入清华大学自动化系攻读博士学位至今。

发表的学术论文

[1] Wang Y G, Min J Y, Zhang J J, et al. Video-based hand manipulation capture through composite motion control. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2013), 2013, 32(4):43. (SCI 收录，检索号：183TM，影响因子：3.361)

[2] Wang Y G, Wang R P, Dai Q H. A parametric model for describing the correlation between single color images and depth maps. IEEE Signal Processing Letters, 2013. (SCI 源刊，已录用，影响因子：1.674)

[3] Bouaziz S, Wang Y G, Pauly M. Online modeling for realtime facial animation. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2013), 2013, 32(4):40. (SCI 收录，检索号：183TM，影响因子：3.361)

研究成果

[1] 戴琼海, 王雁刚. 用于图像分割的方法及系统: 中国, CN101814183B. (中国专利授权号).

[2] 戴琼海, 王雁刚. 基于可变阶马尔科夫随机场的图像表示方法: 中国, CN101964106B. (中国专利授权号).

[3] 戴琼海, 王雁刚. 基于深度相机的测量三维场景深度的装置: 中国, CN102073050B. (中国专利授权号).

[4] 刘烨斌, 王雁刚, 戴琼海. 一种三维手势运动重建方法和系统: 中国, CN102262783B. (中国专利授权号).