

硕 士 研 究 生 读 书 报 告



题目 《Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies》读书报告

作者姓名 张娇昱

作者学号 21851074

指导教师 李启雷

学科专业 软件工程

所在学院 软件学院

提交日期 二○一八年十二月

The Reading Report on “Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies”

A Dissertation Submitted to

Zhejiang University

in Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies

Major Subject: Software Engineering

Advisor: 李启雷

By

张娇昱

Zhejiang University, P.R. China

2018

摘要

本文主要是提出了一个统一的三维可变形模型，它能够用于多尺度无标记地捕捉人体运动，包括面部表情、身体运动和手势。本文中提出的第一个模型是通过拼接人体各个部分局部模型生成的“Frank”模型，这个模型通过一个单独的无缝模型实现了包括脸和手在内的部分动作的完整表达。第二个模型是对于Frank模型进行了针对日常着装人类数据的优化生成的“Adam”模型，它与初始模型具有相同的骨架层次结构，并且具有更简单的参数化。最后，本文演示了这些模型在多视图场景中对于人体总运动跟踪的应用：同时捕捉一群人的大规模身体运动和细微的面部和手部运动。

**关键词**：动作捕捉，无标记动作捕捉，三维可变形模型

Abstract

The paper present a uniﬁed deformation model for the markerless capture of human movement at multiple scales, including facial expressions, body motion, and hand gestures. The first model proposed in this paper is a model, named “Frank”, generated by splicing local models of various parts of the human body. This model realizes the complete expression of partial actions including face and hand through a single seamless model. The second model is the "Adam" model generated by the optimization of Frank's model for daily dress human data, which has the same skeleton hierarchy structure as the original model and simpler parameterization. Finally, this paper demonstrate the use of these models for total motion tracking in a multiview setup, simultaneously capturing the large-scale body movements and the subtle face and hand motion of a social group of people.

**Keywords：**motion capture, markerless motion capture, 3D deformation model

1引言

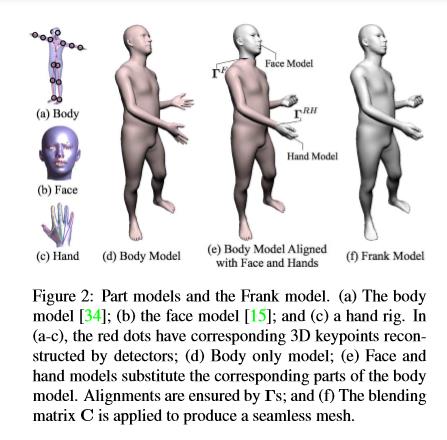
社交是人体运动[9]的重要功能之一。本文用最细微的动作交流大量的信息。在一群相互交流的人之间，一些手势，比如轻轻耸肩，快速转动头部，或者不安地将重心从一只脚移到另一只脚，都在向观察者传递关于注意力、情感和意图的关键信息。值得注意的是，这些社会信号通常是通过全身有组织的运动，如面部表情、手势和身体姿势，来传递的。这些丰富的信号在构建人类行为的活动中是比目标导向的动作更重要的，因此用机器感知和捕捉人类活动有至关重要的意义。

然而，目前还没有一种系统可以在没有标记的情况下同时追踪人体、脸部和双手。当前的无标记运动捕捉系统主要关注特定的尺度或特定的部分的动作捕捉。每个部分的捕捉系统都有自己偏爱的配置:(1)躯干和四肢的运动捕捉设计在人们可以自由移动的大空间中[19,23,47,21];(2)面部捕捉设计在以额部为主的近距离中，假设整体头部运动较小[7,26,8,11,54];(3)手指的运动捕捉设计在离手非常近的距离，在传感器测量中手区域占主导地位[37,52,45,53]。这些配置使得同时分析社会信号的全频谱变得困难。

为了克服这一感知挑战，本文提出了一种新的生成人体变形模型，该模型能够表达人体各部分的运动。特别地，本文描述了通过无缝地将可用的部件模板模型[34,15]合并到单个骨架层次结构来构建初始主体模型“Frank”的过程。为了使模型适合于数据，本文在多个视图中利用关键点检测(例如，faces [20]， bodies [58, 16, 36]， hands[44])来获得对多个人和对象交互具有鲁棒性的3D关键点。本文利用“Frank”模型完成了70人的捕捉，并学习了一种新的变形模型“Adam”，它能够通过简化参数来捕捉头发和衣服的变化。同时本文提出了一种利用三维可变形模型获取多人全身运动的方法。最后，本文展示了本文的方法在各种社会行为序列和人-物交互中的表现，其中面部、肢体和手指动作的组合自然出现。

**2 Frank模型**

Frank模型的构建是是平衡组合现有的几个部分模型。本文中选用的模型是人体模型SMPL[34]，面部模型FaceWarehouse[15]，和本文自己设计的手部模型。需要注意的是，由于图像分辨率之间的权衡和三维扫描系统的视野问题，现在的利用三维扫描系统面部捕捉的脸部模型通常与身体的其他部位的模型构建不同。Frank模型将所有转换骨骼合并成一个单一的骨骼层次结构，但保留每个组件的原始参数，以表达身份和运动变化。作为最终的输出，Frank模型生成了捕捉人体整体运动的运动参数，并通过混合组件网格的顶点生成无缝网格。



2.1 拼接各部分模型

Frank模型构造：

其中，是Frank模型，它有动作参数，形状（身份）参数和整体变化参数。是一个表达目标物体的运动和形状的无缝网格。动作参数和形状参数是模型参数的一个集合：

, ,

B代表身体模型，F代表面部模型，LH代表左手模型，RH代表右手模型。每个部件模型分别从上述参数的子集映射到一组顶点，，， 和，其中每个网格部分的顶点数是，, 。最后得到的Frank模型，利用矩阵C来先行混合上述参数：

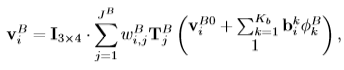
T是矩阵的转置，注意，VU的顶点比部分模型的总和要少，因为在身体模型中有多余的部分(例如，身体模型的脸和手)。特别是，本文最终的网格顶点数为：。图2 (e)为对齐后的零件模型，(f)为Frank模型在人体休息姿势中平均形状处应用混合矩阵C后的最终网格拓扑。混合矩阵C是一个非常稀疏的矩阵;大多数行将单个列设置为在其他地方为零的列，并简单地从相应的部件模型中复制顶点位置，在接缝处进行最小的插值。

在Frank模型中，所有部件都由一个骨骼层次结构刚性连接，这对于运动捕捉的输出至关重要。这种统一是通过将SMPL人体骨架的手部和面部分支替换为详细部件模型的相应骨架层次来实现的。之后再对Frank模型的所有参数进行联合优化，进行运动跟踪和特征拟合。

2.2 Body模型

在Body模型部分，本文使用了SMPL模型[34]并进行了少量修改。主体模型的定义如下:

其中，并且。这些顶点首先被描述身份或身体形状的一组混合参数替代。拿休息姿态举例，使用linear blend skinning (LBS)来获得姿态网格顶点。

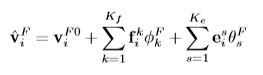


变换矩阵T编码转换为每个联合j对其余部分构成网在世界坐标系中,由遍历骨架构造层次结构从根关节构成参数 (见[34])。第j个顶点构成参数的angle-axis表示的相对旋转接头j对其父关节。忽略SMPL最后联结每只手的模型。为了简化模型，本文不使用姿态依赖的混合类型。

2.3 Face 模型

面部模型使用FaceWarehouse数据集的生成PCA模型。面部模型定义如下：

其中，并且。顶点由形状和表达式子空间组合表示：

****

最后，转换矩阵将面顶点坐标转换为世界坐标。为了保证面顶点的变换与身体的其他部分一致，本文假设面部均值与身体的平均形状一致，如图2所示，这一步在构建模型时手工完成的。利用这种方式,本文将人体模型中的头部通过联结方式全局转换为面部模型。然而,为了保持脸部与身体模型的校准, 需要一个额外的变换矩阵来补偿面部模型根位置相对于人体模型改变的误差。

最后得到的每个面部顶点位置如下：

其中，变换矩阵直接由决定，用于校准面部模型和人体模型

2.4 Hand模型

本文中的手部模型为自己设计的网格模型。手部模型有个关节，网格依然使用linear blend skinning (LBS)。手模型有一个固定的形状，但本文为每跟骨头引入缩放参数，以适应不同的手指大小。其中第j个联合变换是对于其父关节的欧拉角旋转的参数化,每个维度都添加一个额外的各向异性比例因子。特别地，骨骼定位框架的第j个关节的线性变换变为了，将欧拉角表示为一个3\*3旋转矩阵，是对角线为X，Y，Z缩放因子的对角矩阵。在世界坐标中，手的顶点由具有权值的LBS给出:

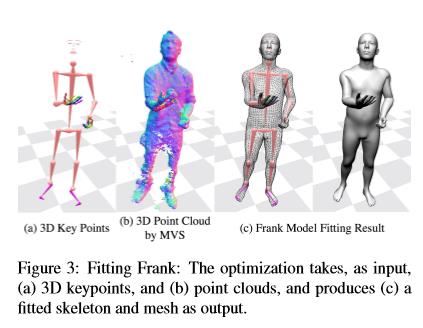
其中，表示第i个顶面的平均形状，是每根骨骼的混合变换，是针对人体模型中相应的手关节进行变换，是校准人体模型与手部模型位置。

**3 Frank模型动作捕捉**

本文的运动捕捉方法在很大程度上依赖于三维关键点对应的拟合网格，这些三维关键点是通过对多个摄像机视图的二维关键点检测三角剖分得到的。为了获取形状信息，本文还使用多视图立体重建生成的点云。通过优化框架进行模型拟合，使对应的模型节点与表面点和三维关键点检测之间的距离最小，迭代最接近点(ICP)与三维点云之间的距离最小。

3.1 三维检测

本文在框架中合并了两种类型的测量，如图3所示:(1)对应的三维关键点，映射到网格模型上已知的关节或表面点(见图2);(2)多视图立体重建中不对应的三维点，本文使用ICP进行匹配。



**3D Body, Face, and Hand Keypoints**:在每个可用视图中使用OpenPose检测器[27]，使用Cao等人[16]方法在躯干上生成2D 关键点，使用Simon等人[44]方法生成头部和手部的关键点。利用已知的摄像机标定参数进行重建，利用[29]方法从二维检测中得到三维人体骨架。通过三角剖分二维手部姿态检测得到三维手部关键点，采用[44]方法，对面部关键点进行三角剖分。注意，在具有相互遮挡或运动模糊的具有挑战性的场景中，如果没有足够的2D检测来进行三角剖分，那么三维关键点的子集可能会完全丢失。

**3D Feet Keypoints**: 使用OpenPose检测器会丢失关键线索——脚上的关键点。对于动作捕捉，这是一个准确确定脚的方向的必要特征。因此，本文训练了一个关键点检测器，用于检测大脚趾尖、小脚趾尖和脚掌。本文在COCO数据集中的大约5000个person实例中的每只脚注释这3个关键点，并使用Wei等人的架构[58]，脚周围有一个由3D body检测决定的包围框。本文还在Panoptic Studio中使用Simon et al.[44]的多视角自展（bootstrapping）方法来提高质量。

**3D Point Clouds**:本文使用商业软件RealityCapture[1]从多视图图像中获取三维点云，并给出相应的点法线。

3.2 目标函数

本文首先独立地匹配序列中的每一帧。为了清晰起见，本文从表示法中删除时间索引，并描述单个框架的流程，优化了以下成本函数:

多阶段的Ceres Solver库[3]的Levenberg–Marquardt算法被用来避免局部最小化。

**Anatomical Keypoint Cost：**对应与网格模型直接对应的三维关键点。这一项包括身体和手的关节(或末端效应器)，也包括与网格表面相对应的点(如面部关键点和手指和脚趾尖)。这两种类型的对应的顶点表示为组合通过回归矩阵, C表示对应的数量和是模型中顶点的数量。设D为某一帧中可用检测集。Cost为:

其中，代表对应回归矩阵中的一行并且用少量的顶点代表一个插入的位置，表示三维检测点。 是这一项的相对权重。

**ICP Cost:** 三维点云的测量与模型网格的匹配不是先验的。因此，本文在每个求解器迭代过程中使用迭代最近点(ICP)来建立它们与网格的对应关系。找到点云中离每个网格顶点最近的3D点，并计算点对面的残差，即法向距离，

其中，是距离顶点j最近的三维点，表示点的正则化，是这一项的相对权重。

**Seam Constraints:**组成Frank模型的部分模型由骨架层次结构严格链接。然而，每个部件模型的独立表面参数化可能会在部件之间的边界引入不连续性(例如，一个肥壮的手臂和一个纤细的手腕)。为了避免这种伪影，本文通过惩罚每个部分接缝周围顶点的最后两个环之间的差异，以及以质心坐标表示的身体模型中相应的最近点，来鼓励接缝部分周围的顶点接近。

**Prior Cost:** 根据特定帧中的可用的测量值的数量，参数集可能不是唯一确定的(例如，手指的宽度)。更重要的是，由于头发和衣服的原因，三维点云存在噪声，模型无法很好地解释，SMPL和FaceWarehouse网格无法捕捉到这些信息，会导致ICP过程中的对应错误。此外，模型的节点位置不一定与用于训练二维探测器的标注标准一致。因此，本文被迫对模型参数设置先验，以避免模型对这些噪声源过拟合，。每个部分的先验由对应的形状和位姿先验定义，对于每个参数本文使用零均值标准正态先验，比例因子除外，比例因子被鼓励接近1。

**4 Adam模型**

Adam模型，相比Frank模型，是一个简化参数的整体人体动作捕捉模型。特别是，这个新模型具有单个关节层次结构和所有形状自由度的通用参数，将脸、手和身体形状捆绑在一起，避免了单独的部件参数化或seam约束的需要。为了建立模型，需要将重建的网格与不同受试者的所有身体部位(脸、身体和手)对齐，这样模型就可以了解到这些部位的变化。为了做到这一点，本文利用了Frank模型，并将其应用于一个由70名受试者组成的数据集，在这个数据集中，每个受试者在一个多视图摄像机系统中执行一个短距离的运动。本文为每个人选择5个不同姿势的框架，得到350个网格，然后用Frank模型重建它们，生成带有关节位置的对齐网格来构建Adam。因为本文从穿衣服的人身上得到模型，所以混合形状在一个粗糙的层面上解释了衣服的变化。

4.1 适应服装和头发

Frank模型捕捉人体和面部的形状变化，但不考虑服装或头发，因为它保留了部分模型的原始模型空间([34]和[15])。为了学习一组新的线性混合形状，更好地捕捉穿着衣服的人的粗略几何形状，也粗略地建模头发，本文需要网格更准确地匹配源数据的几何形状。为此，本文沿着每个顶点的法线方向，在形状空间之外变形网格。Frank模型中的每个顶点,变形网格顶点表示为:

其中，是一个补偿Frank模型顶点与三维点云之间差异的沿着每个顶点的法线方向标量。本文把建模问题视为一个线性系统：

4.2 检测目标回归

在LBS模型的关节位置(即，骨骼变形的三维旋转中心)和关键点检测的位置(这来自于对二维图像中解剖关节位置的人工标注猜测)。如图4所示。这种差异会导致模型在实现低关键点代价(，尤其是肩膀和臀部)的同时，也会导致模型不适合。本文解决这个问题通过计算一个新的回归函数，，这个回归函数使得人体模型顶点与预期的三维关键点的位置相对应。然而，为了能够学习这些回归函数，本文需要模型顶点的实例以及三维关键点的检测。

因此，本文首先以原始关节位置作为检测目标，拟合Frank模型(附加形状变化)，得到所有受试者的对齐网格。基于这些输出，本文可以用三维关键点测量的位置作为目标，而不是Frank模型的关节位置来构建回归矩阵。与SMPL[34]中的联合回归类似，本文首先在每个检测目标附近选择一个顶点子集，并估计这些顶点的固定稀疏线性组合，该组合近似于所有拟合网格中3D关键点的位置。这个优化是一个具有非负约束的l1 -正则化最小二乘问题，本文附加地把顶点权和加起来，得到一个插值。

注意，这个新的回归机只用于Eq.(12)中的优化，而来自SMPL [34]的的原始联合回归机用于LBS。然而，本文也在关节回归矩阵中加入了行来解释额外的手指关节，本文用同样的方法求解。由此产生的矩阵是,是adam模型的顶点的数量(同Frank模型一样)，是adam模型的关节数量模型包括21个身体关节和每个手20个手指关节(包括5指尖)。

4.3 构造形变空间

与∆位移进行模型拟合后, 本文运用LBS变换的逆变换将每一帧都转换到休息姿态。当拟合的表面扭曲到这个标准的姿势，本文做PCA分析建立关节线性形状空间，捕捉形状变化在整个身体。在3.3节中，由于本文的MVS点云分辨率太低，无法匹配面部表情，所以本文将面部表情的表达从面部模型中分离出来并且从FaceWarehouse模型获得。

Adam模型的参数定义：

其中，并且。

**4.4 Adam模型动补**

使用Adam获取全身运动的代价函数类似于Eqn. 11，没有seam项:

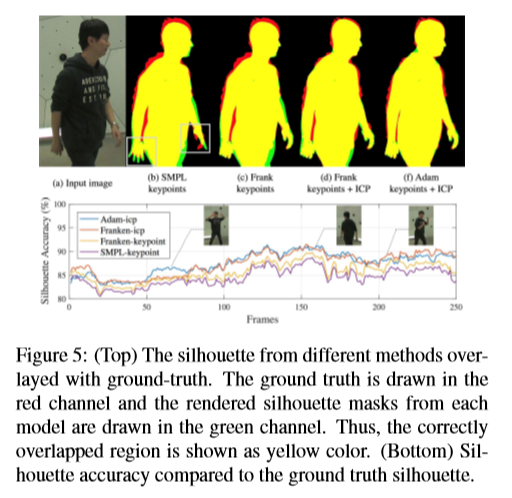
然而，Adam比Frank更易于优化:它为所有部件提供了一组统一的形状和姿态参数，并且不需要不同模型之间的seam约束。

**5 结果**

本文使用两个模型，Frank和Adam，在各种具有挑战性的序列上执行全动作捕捉。对于实验，使用CMU panoptical Studio捕获的数据集[28,29]。使用140个VGA摄像机重建三维人体关键点，480个VGA摄像机重建足部关键点，31个高清摄像机重建面部和手部关键点，以及三维点云。本文将模型与简化的SMPL模型[34]进行比较。

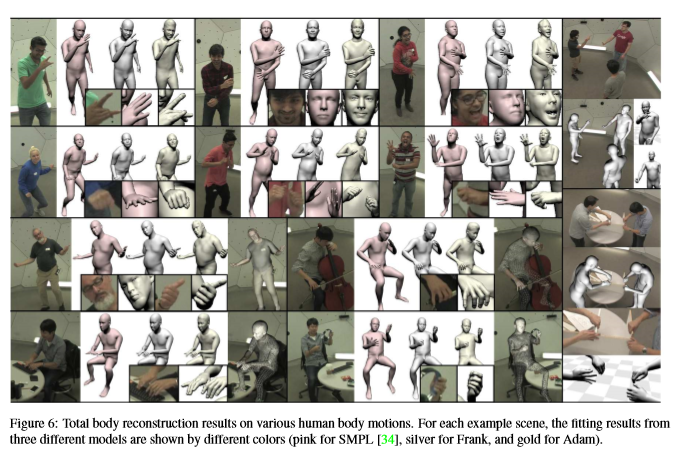
5.1 定量评价

本文通过测量5个不同视点的地面真实轮廓的重叠度来评估每个模型在10秒钟的序列中与移动的人匹配的程度。为了得到地真轮廓，本文使用高斯模型对每个像素的颜色进行背景减影算法，通过形态学变换进行后处理去除噪声。以下是比较结果：



5.2 定性结果

本文在面部和手部动作自然发生的序列上运行本文的方法。这些序列包括用于构建Adam的70人的短距离动作，多人的社交互动，灵巧的手部动作的家具构建序列，音乐表演(大提琴和吉他)，以及常见的日常动作如打字。在这里，本文还定性地比较了本文的模型(Frank为银色，Adam为金色)和SMPL(没有pose-blendshapes，粉色)[34]。注意，基于本文模型的全身运动捕捉通过捕捉手和脸的细微细节来产生更多的真实感。



**5 小结**

本文中一共提出了两种模型，Frank模型和Adam模型。Frank模型是通过拼接已有的人体模型、面部模型和自己构造的手部模型来获得整个人体的完整模型。Adam模型则是在Frank模型的基础上训练出的统一模型，简化了参数，但是在捕捉完整的人体数据时表现也非常出色。

本文得到了两个观点：

1. 通常被认为与身体和面部捕捉相比更具有挑战性的无标记的手部动作捕捉，在本文的结果中显示了更好的定位质量。人体关节位于人体内部，对穿着衣服的受试者来说很难定位，一旦面部不面对任何摄像头，面部重建的准确性就会大大降低。然而，手通常是裸露的，而手关键点检测器[44]甚至在自遮挡情况下也能提供高可信度的猜测测量，可以在多个视图中进行融合。
2. 无标记运动捕捉技术有可能最终超越基于标记的技术。基于标记的方法存在严重的闭塞问题，使得很难同时捕获身体和双手，而本文的方法仍然可以利用基于学习的关键点探测器来测量闭塞部分。

参考文献

[1] Realitycapture software. www.capturingreality. com/.

[2] Vicon motion systems. www.vicon.com.

[3] S. Agarwal, K. Mierle, and Others. Ceres solver. http: //ceres-solver.org.

[4] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. In ToG, 2005.

[5] A. Baak, M. M¨uller, G. Bharaj, H.-P. Seidel, and C. Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In Consumer Depth Cameras for Computer Vision. Springer, 2013.

[6] L. Ballan, A. Taneja, J. Gall, L. Van Gool, and M. Pollefeys. Motion capture of hands in action using discriminative salient points. In ECCV, 2012.

[7] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. High-quality single-shot capture of facial geometry. In TOG, 2010.

[8] T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. Sumner, and M. Gross. High-quality passive facial performance capture using anchor frames. In TOG, 2011.

[9] R. Birdwhistell. Kinesics and context: Essays on body motion communication. In University of Pennsylvania Press, Philadelphia., 1970.

[10] F. Bogo, A. Kanazawa, C. Lassner, P. V. Gehler, J. Romero, and M. J. Black. Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. In CoRR, 2016.

[11] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer. High resolution passive facial performance capture. In TOG, 2010.

[12] C. Bregler, J. Malik, and K. Pullen. Twist based acquisition and tracking of animal and human kinematics. In IJCV, 2004.

[13] T. Brox, B. Rosenhahn, J. Gall, and D. Cremers. Combined region and motion-based 3D tracking of rigid and articulated objects. In TPAMI, 2010.

[14] C. Cao, D. Bradley, K. Zhou, and T. Beeler. Real-time highﬁdelity facial performance capture. In TOG, 2015.

[15] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. In TVCG, 2014.

[16] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multiperson 2d pose estimation using part afﬁnity ﬁelds. In CVPR, 2017.

[17] K. M. Cheung, S. Baker, and T. Kanade. Shape-fromsilhouette across time part i: Theory and algorithms. In IJCV, 2005.

[18] S. Corazza, L. M¨undermann, E. Gambaretto, G. Ferrigno, and T. P. Andriacchi. Markerless Motion Capture through Visual Hull, Articulated ICP and Subject Speciﬁc Model Generation. In IJCV, 2010.

[19] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. In SIGGRAPH, 2008.

[20] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. F. Cohn. Intraface.