软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2016,27(Suppl.(2)):172−183

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 基于局部姿态先验的深度图像 3D 人体运动捕获方法*

苏 乐[1,2], 柴金祥[3], 夏时洪[1]

[1](移动计算与新型终端北京市重点实验室(中国科学院 计算技术研究所 前瞻研究实验室),北京 100190)

[2](中国科学院大学,北京 100049)

[3](Texas A&M University, Computer Science and Engineering, Texas 77843-3112, USA)

通讯作者: 夏时洪, E-mail: xsh@ict.ac.cn

**摘 要**: 提出一种基于局部姿态先验的从深度图像中实时在线捕获 3D 人体运动的方法.关键思路是根据从捕获的深度图像中自动提取具有语义信息的虚拟稀疏 3D 标记点,从事先建立的异构 3D 人体姿态数据库中快速检索 $K$ 个姿态近邻并构建局部姿态先验模型,通过迭代优化求解最大后验概率,实时地在线重建 3D 人体姿态序列.实验结果表明,该方法能够实时跟踪重建出稳定、准确的 3D 人体运动姿态序列,并且只需经过个体化人体参数自动标定过程,可跟踪身材尺寸差异较大的不同表演者;帧率约 25fps.因此,所提方法可应用于 3D 游戏/电影制作、人机交互控制等领域.

**关键词**: 运动捕获;数据驱动;深度图像;$K$ 近邻搜索;最大后验概率

中文引用格式: 苏乐,柴金祥,夏时洪.基于局部姿态先验的深度图像 3D 人体运动捕获方法.软件学报,2016,27(Suppl.(2)):172−183. http://www.jos.org.cn/1000-9825/16032.htm

英文引用格式: Su L, Chai JX, Xia SH. Local pose prior based 3D human motion capture from depth camera. Ruan Jian Xue Bao/Journal of Software, 2016,27(Suppl.(2)):172−183 (in Chinese). http://www.jos.org.cn/1000-9825/16032.htm

## Local Pose Prior Based 3D Human Motion Capture from Depth Camera

SU Le[1,2], CHAI Jin-Xiang[3], XIA Shi-Hong[1]

[1](Beijing Key Laboratory of Mobile Computing and Pervasive Devices (Advanced Computing Research Laboratory, Institute of Computing Technology, The Chinese Academy of Sciences), Beijing 100190, China)

[2](University of Chinese Academy of Sciences, Beijing 100049, China)

[3](Texas A&M University, Computer Science and Engineering, Texas 77843-3112, USA)

**Abstract**: This paper introduces a local pose prior based real-time online approach to capture 3D human animation from a single depth camera. The key idea is to learn a series of local pose prior models with K motion capture examples from a pre-established large and heterogeneous human motion database, based on automatically extracted labelled virtual sparse 3D markers from captured depth image. Then, by solving a Maximum A Posterior (MAP) problem via an iteratively optimization process, the system automatically tracks the 3D human motion sequence. The experiments show that the proposed approach robustly captures the accurate 3D human motions at 25fps. The proposed tracking system can easily applied to different actors with large different body sizes via an automatically individual body parameters calibration process. The proposed system can widely apply to 3D game/movie produce, human-machine interaction.

**Key words**: motion capture; data driven; depth image; $K$ nearest neighbor search; MAP

基于 RGBD 相机的无标记点 3D 人体运动捕获技术主要研究如何根据捕获的深度图像序列跟踪和重建准确的 3D 人体运动姿态序列.捕获的 3D 人体运动姿态序列,既可以实时、动态地生成 3D 人体骨骼动画,完成人机交互控制等任务,如游戏控制、网络社交等领域,也可以通过运动重定向和蒙皮技术,实时驱动生成自然的虚拟 3D 角色动画,如 3D 游戏开发和 3D 电影制作等领域.

* 基金项目: 中国科学院计算技术研究所创新课题(20166040)
Foundation item: Institute of Computing Technology, The Chinese Academy of Sciences (20166040)
收稿时间: 2016-05-10; 采用时间: 2016-09-07

自 2010 年微软正式推出基于红外散斑(speckle pattern)技术的 Kinect Xbox 360[1]深度相机,至今,已发展到了基于光飞行时间(time-of-fight)技术的 Kinect for Windows v2.0[2]版本.微软 Kinect 深度相机能实时在线地捕获 3D 人体运动,帧率为 30 帧/秒,是目前被学术界和工业界广泛使用的最好的基于无标记点 RGBD 图像的人体运动捕获系统.

其基本原理[3,4]是根据事先训练好的随机决策森林[5]从捕获的深度图像中自动识别 3D 人体运动姿态.尽管离线训练随机决策森林时用到了基于 Vicon 系统[6]的运动捕获数据库,但在实时在线重建 3D 人体姿态时,并未使用 3D 人体运动数据库.尽管该方法能自动识别出不同身材尺寸的表演者以及日常生活中多种类型的人体运动姿态,但受到深度数据噪声、随机决策森林泛化能力以及人体运动中肢体遮挡等因素的影响,仍不能获得合理和准确的 3D 人体运动姿态重建结果,如图 1 所示.

(a) Kinect 捕获的 RGB 图像 (b) Kinect v2.0 结果 (c) 本文方法结果

Fig.1 Full-Body pose captured by Kinect v2.0

图 1 微软 Kinect v2.0 的人体捕获结果

因此,本文在 Kinect 基于随机决策森林捕获人体运动方法的基础上,进一步提出基于数据驱动的方法,实时、在线地捕获 3D 人体运动的方法.相比非基于数据驱动的方法,通过事先建立的运动捕获数据库,主要能处理发生部分肢体遮挡时不能重建出合理和准确的人体姿态的不足.

目前,已有的基于数据驱动的方法,最典型的是采用 3D 人体模型数据库检索的方法[7,8],使用运动捕获数据驱动一个标准尺寸 3D 人体模型生成的不同姿态模型数据库或对应的多角度投影深度图像数据库,通过比较捕获的深度点云与数据库样本间的相似性,进行候选姿态检索,再通过姿态投票或非刚体注册等方法重建人体姿态.

但正如提出此类方法的原文献[7]的作者所述,如图 2 所示,该方法只能处理身材尺寸接近数据库中标准尺寸 3D 人体模型的表演者数据.但在实际应用中,表演者的身材尺寸与标准尺寸可能存在较大差异,所以已有方法一定不能获得准确的人体姿态重建结果.一条简单的解决思路是扩大数据库中的不同尺寸 3D 人体模型数量,但数据规模将呈指数级增长,大大增加了姿态检索难度,不是一种切实可行的解决方案.一种可行的解决方案是从数据库中去掉 3D 人体模型数据,只保留 3D 人体姿态数据,并针对检索姿态建模的方法,提高其数据库的表达能力.

因此,本文提出根据异构 3D 人体姿态数据检索和构建局部姿态先验模型的方法,从捕获的深度图像中实时、在线地跟踪 3D 人体运动.方法流程如图 3 所示.首先,本文从捕获的深度点云中自动提取位于人体表面的一组稀疏 3D 标记点.从事先建立的异构 3D 人体姿态数据库中,快速、自动地搜索出 $K$ 个姿态近邻,构建 3D 人体局部姿态先验模型.异构 3D 人体姿态数据库由不依赖于具体身材尺寸的关节角向量表示的 3D 人体姿态组成.

然后,基于最大后验概率(MAP)的迭代非线性优化框架,根据已知的稀疏 3D 标记点、构建的 3D 人体局部姿态先验,以及前 2 帧重建的 3D 人体姿态,实时、在线地重建当前帧 3D 人体姿态.

最后,通过一个自动的个体化人体参数标定过程,使本文提出的3D人体运动跟踪方法能适用于具有身材尺寸差异的不同表演者.

> First experiments show that actors with different body proportions can be tracked if they are not too different from our body model. To this end, we scale the input data to roughly match the proportions the model, see Fig. 7 and the

Fig.2 Drawback of existing database method (As discribed by the author of the literature[7]).
Red line shows that the full-body pose tracking results can be affected by the differences of body sizes
图 2 已有基于数据库方法缺陷(摘自原文献[7]作者对自己工作的评价).
红色下划线部分的文字阐述了身材尺寸差异对已有工作的影响非常大

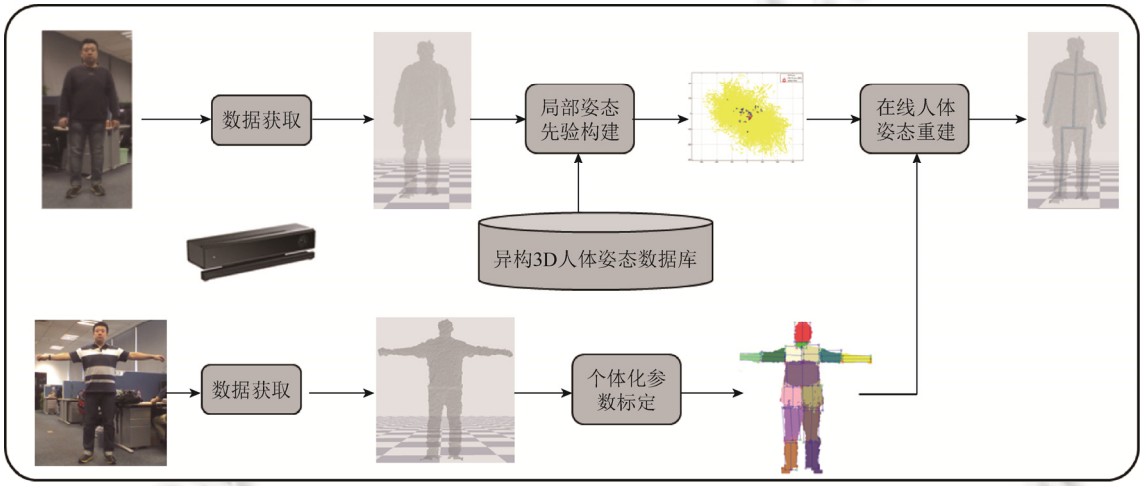



Fig.3 System overview
图 3 本文方法流程图

本文首先设计了交叉验证实验,从数据库中去掉一个运动,作为测试运动和基准数据,驱动 3D 人体模型,投影深度图像序列,用本文方法重建人体姿态,并与基准数据进行对比,如后文图 8 所示,验证了本文建立的异构3D 人体姿态数据库的姿态表达能力.其次设计了在线 3D 人体姿态重建算法各能量项的必要性验证实验,以使用 Vicon 系统同步采集并重建的人体姿态为基准数据,比较了逐一去掉某个能量项时的人体姿态重建误差,如后文图 9 所示,验证了各能量项的必要性;然后设计了与当前商用系统 Kinect 的人体姿态重建精度的对比实验,如后文图 10 所示,验证了本文方法相比 Kinect 具有更高的稳定性和准确性;最后测试了身材尺寸具有较大差异的 4 个不同表演者,如后文图 11 所示,验证了本文方法能适用于不同身材尺寸的表演者.

本文工作的主要贡献包括:

- 提出了基于局部姿态先验从捕获的深度图像中实时、在线捕获稳定和准确的 3D 人体运动方法(见第 4 节);
- 定义了从异构 3D 人体姿态数据库中基于捕获深度图像中自动提取的稀疏 3D 标记点搜索相似姿态的距离度量(见第 3 节),建立了深度点云与 3D 人体姿态间的关系,方便进行姿态检索;
- 提出了基于 3D 人体骨架数据库的自动个体化人体参数标定方法(见第 5 节).

## 1 相关工作

近年来已有很多基于 RGBD 相机的无标记点 3D 人体运动捕获技术的相关研究工作.

### 1.1 基于随机决策森林的姿态识别方法

采用直接从深度图像进行姿态识别的方法,如微软 Kinect 系统.Shotton 等人[3]和 Girshick 等人[4]分别训练基于随机森林算法的深度图像像素点分类器和回归器,再用聚类算法得到人体 3D 关节中心坐标.本文借鉴 Shotton 等人[3]的方法,根据人体骨架结构,自定义新的肢体段划分方式,以提高像素分类精度,训练了随机森林分类器.

### 1.2 基于数据库检索的方法

Baak 等人[7]提出根据深度图像特征点检测[9]结果分别进行姿态局部优化和全局数据库检索,两者中与当前捕获深度点云误差小者为最终全身人体姿态估计结果.数据库是由内嵌骨架驱动的标准尺寸 3D 人体模型构成,姿态查询度量的是检测出的深度图像特征点与数据库中模型投影深度图像特征点之间的距离差.

Ye 等人[8]提出用捕获的 3D 深度点云与数据库中标准尺寸 3D 人体模型的投影深度图像对应的 3D 深度点云进行检索匹配,找出最优匹配对应的 3D 人体模型,再将其与捕获深度点云进行非刚体注册[10−13],重建出全身人体姿态.数据库是由内嵌骨架驱动的标准 3D 人体模型投影得到的深度图像对应的 3D 深度点云构成的,查询的依据是度量当前帧 3D 深度点云与数据库中的 3D 深度点云间的距离差.但如文中作者所述,最大的缺点是所构建的数据库包含一个标准尺寸的 3D 人体模型,当用户身材与数据库中 3D 人体模型相差较大时,数据库的泛化能力有限,不能给出很好的姿态估计结果.因此,区别于上述两个基于 3D 人体模型姿态数据库检索的方法,本文建立了基于只包含异构 3D 人体姿态的数据库,通过在线构建局部姿态先验的方法,重建 3D 人体运动姿态.3D 人体运动姿态由关节角表示,理论上对于零姿态一致、但尺寸存在差异(即身材尺寸存在差异)的不同骨架,同一组关节角能得到相同的人体姿态.

本文的数据库和姿态先验模型的构建是受到已有运动捕获工作的启发,Chai 等人[14]使用运动捕获数据库构建局部姿态线性模型,从双目视频相机捕获的 3D 标记点数据重建人体运动姿态序列,Liu 等人[15]使用运动捕获数据库构建局部姿态回归模型,从惯性传感器捕获的 3D 惯性数据重建人体运动姿态.据我们所知,本文首次将根据运动捕获数据库构建局部姿态线性模型的方法应用在基于深度图像的人体运动姿态重建问题中.

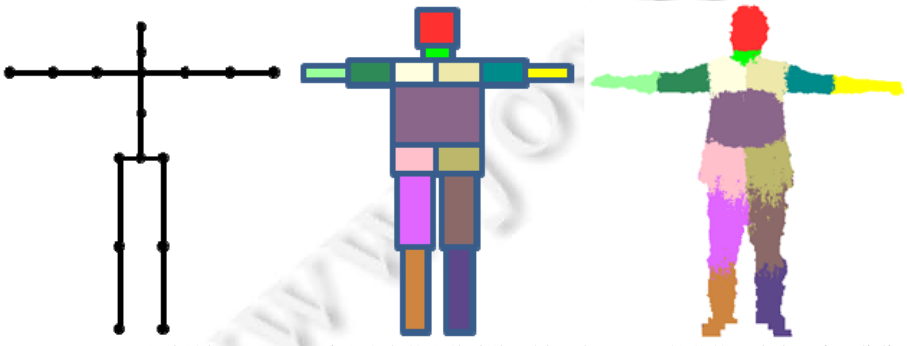
## 2 数据获取与姿态数据库

### 2.1 数据获取

本文使用的是微软 Kinect V2.0 深度相机,以 30 帧/秒的帧率,实时获取分辨率为 640×480 的深度图像序列.深度图像像素点表示为 $\boldsymbol{x}$,对应深度值和 3D 点分别表示为 $d(\boldsymbol{x})$和 $\boldsymbol{p}$.

### 2.2 3D人体姿态表示

如图 4(a)所示.



(a) 3D 人体骨架　(b) 本文定义的人体肢体段划分结果　(c) 捕获的 T 姿态深度图像像素点所属肢体段类别的识别结果

Fig.4 Defined full-body skeleton and body-parts

图 4 人体骨架及肢体段定义

本文定义3D人体姿态为关节自由度(degree of freedom,简称Dof)构成的向量$\boldsymbol{q}\in\boldsymbol{R}^{36}$,具体包括root (6 Dof)、upperback(3 Dof)、r/lclavicle (2 Dof)、r/lhumerus (3 Dof)、r/lradius (1 Dof)、neck (2 Dof)、head (1 Dof)、r/lfemur (3 Dof)、r/ltibia (1 Dof)和r/lfoot (2 Dof).

### 2.3 异构3D人体姿态数据库

本文在从深度图像自动提取稀疏3D标记点与在线自动构建姿态先验时均需使用事先建立好的3D人体姿态数据库.从美国CMU运动捕获数据库[16]中挑选了82个总时间接近1.5小时的运动序列,运动类型包括:走路、跑步、打拳、踢腿、跳跃、跳舞、挥手、健身和打高尔夫等.使用运动重定向技术[17]进行了骨架归一化.

## 3 在线局部姿态先验构建

### 3.1 稀疏3D标记点提取

从Kinect深度相机捕获得到的原始深度图像,像素点对应3D空间中的人体表面点,是稠密的3D点云,本文提出从捕获的深度图像中自动、在线地提取位于人体表面且与人体肢体段具有对应关系的稀疏3D标记点集,作为在线构建局部姿态先验和在线3D人体姿态跟踪过程中3D空间约束,重建3D人体姿态的方法.

首先,本文将人体划分为15个肢体段,分别为head,neck,r/l shoulder,r/l upper arm,r/l lower arm,torso,r/l waist,r/l upper leg和r/l lower leg,如图4(b)所示.区别于Shotton等人的工作[3],本文使用的肢体段分类方式能获得更鲁棒的像素分类结果.训练了一个随机决策森林分类器[5]对捕获的深度图像像素点完成分类识别,确定深度图像像素点与人体肢体段类别的概率分布.

然后,本文使用MeanShift迭代聚类算法[18],自动、快速地确定出该概率密度的前15个极大值点,这些极大值点分别对应15个人体肢体段,本文将这些3D聚类中心点看成是具有语义信息的稀疏3D标记点,如图5所示.

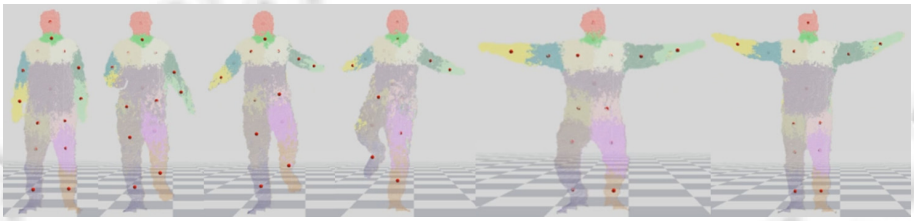


Fig.5 Sparse 3D virtual markers detection results. Captured 3D depth point clouds with body-parts (different color) and sparse 3D virtual markers (black dots) detection results

图5 稀疏3D标记点提取结果.图中显示的是捕获的深度图像分别对应的3D点云,不同颜色表示不同的人体肢体段,黑色3D点集是提取的虚拟稀疏3D标记点

### 3.2 在线$K$近邻搜索

已知当前帧提取的稀疏3D标记点$\tilde{\boldsymbol{a}}$、前2帧重建姿态$[\tilde{\boldsymbol{q}}_{-2},\tilde{\boldsymbol{q}}_{-1}]$以及事先建立的3D人体姿态数据库$\boldsymbol{Q}=\{\boldsymbol{q}_n\,|\,n=1,\ldots,N\}$,对于数据库$\boldsymbol{Q}$中的每个姿态$\boldsymbol{q}_n$,本文定义搜索距离$d_{\text{query}}$为

$$d_{\text{query}}(\tilde{\boldsymbol{a}},\boldsymbol{q}_n)=\alpha\rho\left(\boldsymbol{a}_n-R(\boldsymbol{a}_n,\tilde{\boldsymbol{a}})\tilde{\boldsymbol{a}}\right)+\beta\left\|\boldsymbol{q}_n-2\tilde{\boldsymbol{q}}_{-1}+\tilde{\boldsymbol{q}}_{-2}\right\|_L^2,$$

其中,因为稀疏3D标记点存在误差,所以本文使用Lorentzian鲁棒距离度量$\rho(\mathbf{e})=\log\left(1+\mathbf{e}^2/\left(2\sigma^2\right)\right)$代替欧式距离度量,$\sigma$是用于鲁棒估计的标量,实验中设为0.05. $R(\boldsymbol{a}_n,\tilde{\boldsymbol{a}})$是数据库姿态$\boldsymbol{q}_n$对应3D标记点$\boldsymbol{a}_n$与当前帧提取的3D标记点$\tilde{\boldsymbol{a}}$之间的相对旋转矩阵.上式中第1项度量的是当前帧提取的3D标记点与数据库姿态对应3D标记点之间的距离;第2项度量的是数据库姿态与前2帧重建姿态之间连续变化程度的大小.实验中,权值$\alpha$,$\beta$分别为0.75和0.25.部分搜索结果如图6所示.

本文实现了基于CUDA GPU的并行在线$K$近邻姿态搜索.首先创建$N$个一维GPU线程,每个GPU线程针

对数据库中姿态 $\boldsymbol{q}_n$ 分别计算搜索距离 $d_{\text{query}}(\tilde{\boldsymbol{a}},\boldsymbol{q}_n)$;然后使用堆排序算法求出搜索距离最小的前 $K$ 个姿态近邻 $\boldsymbol{Q}_K=\{\boldsymbol{q}_1,\ldots,\boldsymbol{q}_K\}$.

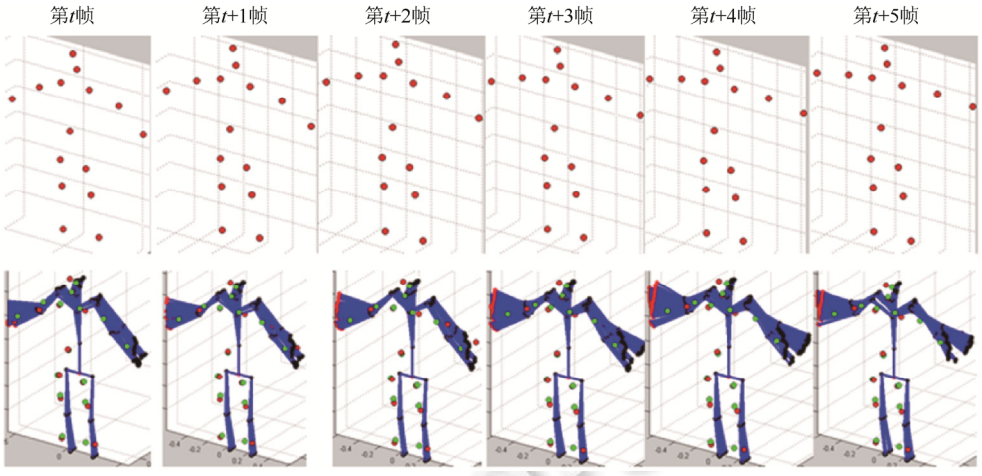



Fig.6 Online *K*NN poses searching results. Example seven consistent frames, *K* most similar poses (2nd line) are selected according to sparse 3D virtual markers (1st line) for each frame

图 6 在线 $K$ 姿态近邻搜索结果.图中显示了连续 7 帧,每帧根据各自捕获的稀疏 3D 标记点(第 1 行),从数据库中搜索出的前 $K$ 个最相似姿态(第 2 行)

### 3.3 局部线性模型构建

已知当前帧搜索出的 $K$ 个姿态近邻 $\boldsymbol{Q}_K$,本文使用主成分分析技术建立当前帧待求 3D 人体姿态 $\tilde{\boldsymbol{q}}$ 的局部线性模型表示为

$$\tilde{\boldsymbol{q}}=\bar{\boldsymbol{q}}+\boldsymbol{P}_B\cdot\boldsymbol{\varphi},$$

其中,$\bar{\boldsymbol{q}}$ 是 $K$ 个姿态近邻 $\boldsymbol{Q}_K$ 的均值向量,$\boldsymbol{P}_B$ 是 $K$ 个姿态近邻 $\boldsymbol{Q}_K$ 的前 $B$ 维主成分向量构成的矩阵,$\boldsymbol{\varphi}$ 是当前帧 3D 人体姿态的低维表示向量.

姿态近邻搜索的 $K$ 值和局部线性模型的低维向量维度 $B$ 是通过交叉验证自动确定的.构建每帧的待求 3D 人体姿态的局部线性模型时,首先构建一组包含不同 $K$ 值和 $B$ 值的局部线性模型集合,然后通过留一(leave-one-out)交叉验证过程自动选出 3D 人体姿态重建结果最好的那个局部线性模型,从而确定 $K$ 值和 $B$ 值.实验中,通常 $K$ 值不超过 50,$B$ 值不超过 7 维.

## 4 在线 3D 人体姿态重建

本节重点介绍如何根据在线提取的稀疏 3D 标记点(见第 3 节)和构建的局部姿态先验(见第 4 节),以及前 2 帧重建的 3D 人体姿态,跟踪重建出当前帧 3D 人体姿态.

本文将在线 3D 人体姿态跟踪问题形式化为一个最大后验概率(maximum a posteriori,简称 MAP)问题进行迭代优化求解.具体地说,求最满足当前帧稀疏 3D 标记点 $\tilde{\boldsymbol{a}}$、$K$ 个姿态近邻 $\boldsymbol{Q}_K=\{\boldsymbol{q}_1,\ldots,\boldsymbol{q}_K\}$,以及前 2 帧重建的 3D 人体姿态 $[\tilde{\boldsymbol{q}}_{-2},\tilde{\boldsymbol{q}}_{-1}]$ 的当前帧 3D 人体姿态 $\tilde{\boldsymbol{q}}$:

$$\tilde{\boldsymbol{q}}=\arg\max_{\boldsymbol{q}}\ Pr(\boldsymbol{q}\mid\tilde{\boldsymbol{a}};\boldsymbol{q}_1,\ldots,\boldsymbol{q}_K;\tilde{\boldsymbol{q}}_{-2},\tilde{\boldsymbol{q}}_{-1})\propto\arg\max_{\boldsymbol{q}}\ Pr(\tilde{\boldsymbol{a}}\mid\boldsymbol{q})Pr(\boldsymbol{q}\mid\boldsymbol{q}_1,\ldots,\boldsymbol{q}_K)Pr(\boldsymbol{q}\mid\tilde{\boldsymbol{q}}_{-2},\tilde{\boldsymbol{q}}_{-1}).$$

实验中,通常最小化负对数函数,具体形式化为

$$\tilde{\boldsymbol{q}}=\arg\min_{\boldsymbol{q}}\ \underbrace{-\ln Pr(\tilde{\boldsymbol{a}}\mid\boldsymbol{q})}_{E_{marker}}\underbrace{-\ln Pr(\boldsymbol{q}\mid\boldsymbol{q}_1,\ldots,\boldsymbol{q}_K)}_{E_{prior}}\underbrace{-\ln Pr(\boldsymbol{q}\mid\tilde{\boldsymbol{q}}_{-2},\tilde{\boldsymbol{q}}_{-1})}_{E_{smooth}},$$

其中,第 1 项 $E_{marker}$ 是稀疏 3D 标记点约束项,第 2 项 $E_{proior}$ 是局部姿态先验约束项,第 3 项 $E_{smooth}$ 是姿态变化平

滑项.

### 4.1 稀疏3D标记点约束项

惩罚的是根据重建的 3D 人体姿态 $\boldsymbol{q}$ 计算出的稀疏 3D 标记点与从捕获的深度图像中提取的对应稀疏 3D 标记点 $\tilde{\boldsymbol{a}}$ 间的距离.

$$Pr(\tilde{\boldsymbol{a}}\,|\,\boldsymbol{q}) \propto \exp\left(-\frac{\left\|\boldsymbol{f}(\boldsymbol{q};\tilde{\boldsymbol{s}},\tilde{\boldsymbol{v}})-\tilde{\boldsymbol{a}}\right\|^2}{2\sigma_{marker}^2}\right) \Rightarrow E_{marker}=-\ln Pr(\tilde{\boldsymbol{a}}\,|\,\boldsymbol{q}) \propto \left\|\boldsymbol{f}(\boldsymbol{q};\tilde{\boldsymbol{s}},\tilde{\boldsymbol{v}})-\tilde{\boldsymbol{a}}\right\|^2,$$

其中, $\boldsymbol{f}(\boldsymbol{q};\tilde{\boldsymbol{s}},\tilde{\boldsymbol{v}})$ 是前向运动学方程,计算已知 3D 人体姿态 $\boldsymbol{q}$、3D 人体骨架 $\tilde{\boldsymbol{s}}$ 和稀疏 3D 标记点相对父关节偏移量 $\tilde{\boldsymbol{v}}$ 时的稀疏 3D 标记点坐标, $\tilde{\boldsymbol{s}}$ 和 $\tilde{\boldsymbol{v}}$ 是通过本文提出的自动 3D 人体参数标定方法计算获得的,详见第 5 节.仅使用此约束项通过逆向运动学方法逐帧也可以求解 3D 人体姿态[20],但从捕获的深度图像中提取的稀疏 3D 标记点存在噪声而不准确.因此,本文使用第 3.2 节中的 Lorentzian 鲁棒距离度量代替欧式距离度量,其中 $\boldsymbol{e}=\boldsymbol{f}(\boldsymbol{q};\tilde{\boldsymbol{s}},\tilde{\boldsymbol{v}})-\tilde{\boldsymbol{a}}$ , $\sigma$ 是用于鲁棒估计的标量,实验中设为 0.05.

因为从捕获的深度图像中提取的稀疏 3D 标记点与人体肢体段存在一一对应的关系,所以一旦某些肢体段对应的稀疏3D标记点在人体运动过程中由肢体自遮挡造成了丢失,本文算法能自动进行判断,并将丢失的标记点从此约束项中剔除.

### 4.2 局部姿态先验约束项

惩罚的是重建的 3D 人体姿态 $\boldsymbol{q}$ 与在线搜索出的 $K$ 个姿态近邻 $\boldsymbol{Q}_K=\{\boldsymbol{q}_1,\ldots,\boldsymbol{q}_K\}$ 构成的局部空间 3D 人体姿态概率分布的满足程度.假设局部空间中的 $K$ 个姿态近邻构成多维高斯分布,则局部姿态先验约束项最大化.

$$Pr(\boldsymbol{q}\,|\,\boldsymbol{q}_1,\ldots,\boldsymbol{q}_K) \propto \exp\left(-\frac{\left\|\boldsymbol{P}_B^T\left(\boldsymbol{P}_B(\boldsymbol{q}-\bar{\boldsymbol{q}})\right)+\bar{\boldsymbol{q}}-\boldsymbol{q}\right\|^2}{2\sigma_{prior}^2}\right)$$

$$\Rightarrow E_{prior}=-\ln Pr(\boldsymbol{q}\,|\,\boldsymbol{q}_1,\ldots,\boldsymbol{q}_K) \propto \left\|\boldsymbol{P}_B^T\left(\boldsymbol{P}_B(\boldsymbol{q}-\bar{\boldsymbol{q}})\right)+\bar{\boldsymbol{q}}-\boldsymbol{q}\right\|^2,$$

其中,向量 $\bar{\boldsymbol{q}}$ 和矩阵 $\boldsymbol{P}_B$ 分别是局部空间中 $K$ 个姿态近邻 $\boldsymbol{Q}_K=\{\boldsymbol{q}_1,\ldots,\boldsymbol{q}_K\}$ 的姿态均值向量和协方差矩阵前 $B$ 维主成分向量构成的矩阵.

### 4.3 姿态变化平滑项

惩罚的是重建的 3D 人体姿态 $\boldsymbol{q}$ 与前 2 帧重建姿态 $[\tilde{\boldsymbol{q}}_{-2},\tilde{\boldsymbol{q}}_{-1}]$ 间的变化速度的平滑度.

$$Pr(\boldsymbol{q}\,|\,\tilde{\boldsymbol{q}}_{-2},\tilde{\boldsymbol{q}}_{-1}) \propto \exp\left(-\frac{\left\|\boldsymbol{q}-2\tilde{\boldsymbol{q}}_{-1}+\tilde{\boldsymbol{q}}_{-2}\right\|^2}{2\sigma_{smooth}^2}\right) \Rightarrow E_{smooth}=-\ln Pr(\boldsymbol{q}\,|\,\tilde{\boldsymbol{q}}_{-2},\tilde{\boldsymbol{q}}_{-1}) \propto \left\|\boldsymbol{q}-2\tilde{\boldsymbol{q}}_{-1}+\tilde{\boldsymbol{q}}_{-2}\right\|^2,$$

其中, $\tilde{\boldsymbol{q}}_{-2}$ 和 $\tilde{\boldsymbol{q}}_{-1}$ 分别是第 $t$–2 帧和第 $t$–1 帧重建的 3D 人体姿态.本文假设人体运动姿态在连续 3 帧约 10ms 的时间间隔内是近似匀速变化的,当前帧 3D 人体姿态只与前 2 帧重建姿态有关,因此平滑项度量的是相邻 2 帧 3D 人体姿态的速度变化.

综合上述 3 个约束项,得到最终的能量函数比表示为

$$\tilde{\boldsymbol{q}}=\arg\max_{\boldsymbol{q}}\ \lambda_1\left\|\boldsymbol{P}_B^T\left(\boldsymbol{P}_B(\boldsymbol{q}-\bar{\boldsymbol{q}})\right)+\bar{\boldsymbol{q}}-\boldsymbol{q}\right\|^2+\lambda_2\rho\left(\boldsymbol{f}(\boldsymbol{q};\tilde{\boldsymbol{s}},\tilde{\boldsymbol{v}})-\tilde{\boldsymbol{a}}\right)+\lambda_3\left\|\boldsymbol{q}-2\tilde{\boldsymbol{q}}_{-1}+\tilde{\boldsymbol{q}}_{-2}\right\|^2,$$

其中, $\lambda_1,\lambda_2,\lambda_3$ 是各约束项权值,实验中分别设为 0.2,5,1.本文优化过程中首先根据一阶 Taylor 展开将各非线性约束项线性化,并推导了解析的 Jacobian 矩阵;然后将 Kinect 给出的第 1 帧 3D 人体姿态作为初解,由 Levenberg-Marquardt 迭代非线性优化算法求解当前帧 3D 人体姿态 $\tilde{\boldsymbol{q}}$ .实验中,迭代优化过程平均每帧迭代次数不超过 5 次.

## 5 自动 3D 人体参数和标记点标定

为使本文提出的在线 3D 人体姿态跟踪方法能适用于身材尺寸差异较大的不同表演者,本文提出一个基于 3D 人体骨架数据库,自动根据捕获的深度图像标定个体化 3D 人体骨架参数 $\tilde{\boldsymbol{s}}$(骨骼段长度)和稀疏 3D 标记点相对人体骨架父关节相对偏移参数 $\tilde{\boldsymbol{v}}$ 的方法.同一个表演者只需要进行一次标定过程.

本文使用了美国 CMU 运动捕获数据库中的不同尺寸的人体骨架(acclaim skeleton file,简称 ASF)集合,基于主成分分析技术建立 3D 人体骨架先验模型.

$$\boldsymbol{s}=\boldsymbol{H}\boldsymbol{\tau}+\bar{\boldsymbol{s}},$$

其中,$\bar{\boldsymbol{s}}$ 和 $\boldsymbol{H}$ 分别表示骨架均值向量和前 $h$ 维主成分向量构成的矩阵,$\boldsymbol{\tau}$ 是骨架的低维向量.同时,将人体躯干近似看成椭圆柱体,将其余肢体段看成圆柱体,如图 4(b)所示,并且假设人体左右肢体段尺寸是对称的.

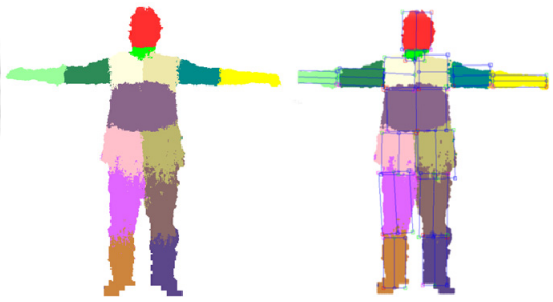
首先,让表演者身体摆成“T”姿态,用深度相机连续捕获 10 帧深度图像,为提高肢体段识别精度,本文使用专门针对“T”姿态训练的随机决策森林(如图 7(a)所示)来提取每帧的稀疏 3D 标记点坐标,表示为 $\left\{\tilde{\boldsymbol{a}}_t^* \mid t=1,\ldots,10\right\}$.

然后,基于随机一次性采样(RANdom SAmple consensus,简称 RANSAC)算法[19],根据第 1 帧捕获深度图像估算出个体化人体参数 $\boldsymbol{s}_0$.具体做法是基于 RANSAC 算法,针对每类像素点拟合一个圆柱体模型[13].

最后,在已知捕获的 10 帧深度图像中稀疏 3D 标记点集合和 3D 人体骨架先验模型,将从第 1 帧估计出的个体化人体参数作为初始解,通过迭代优化求解下面的能量方程,得到最终的个体化 3D 人体参数.

$$\arg\min_{\boldsymbol{s},\boldsymbol{v}} \sum_{t=1}^{10}\left\|\boldsymbol{f}\left(\boldsymbol{s},\boldsymbol{v};\boldsymbol{q}^*\right)-\tilde{\boldsymbol{a}}_t^*\right\|^2+\gamma\left\|\boldsymbol{H}_h^T\left(\boldsymbol{H}_h\left(\boldsymbol{s}-\bar{\boldsymbol{s}}\right)\right)+\bar{\boldsymbol{s}}-\boldsymbol{s}\right\|^2,$$

其中,$\gamma$是权值,设为 0.5.标定过程是自动完成的,标定一次约耗时 1min.标定结果如图 7(b)所示.



(a) T 姿态时深度图像像素分类结果 (b) 各肢体段标定结果

Fig.7 3D actor calibration results

图 7 自动 3D 人体参数标定

## 6 实 验

本文实验主要针对不同表演者和不同运动类型,基于异构 3D 人体姿态数据库,构建局部姿态先验模型,进行 3D 人体运动跟踪的方法进行有效性、准确性和通用性测试.

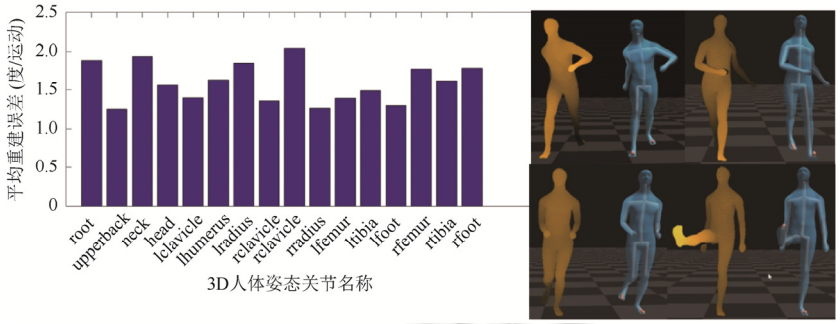
本文的实验平台是 Intel(R) Xeon(R) CPU E3-1240 V2 @ 3.40GHz 3.40GHz,内存 16GB,显卡 NVIDIA GeForce GTX 780 Ti,操作系统 Windows 7,CUDA 版本 5.5.

实验时,本文的 3D 人体姿态跟踪方法在实现代码未经优化的情况下,稀疏 3D 标记点提取过程、局部姿态先验构建过程和非线性迭代优化过程分别耗时为 5ms、20ms 和 15ms,因此能达到平均约 25 帧/秒.

### 6.1 交叉验证实验

本文设计了针对异构 3D 人体姿态数据库的姿态表达能力进行测试的交叉验证实验.实验中,从异构的 3D 人体姿态数据库中去掉一个运动序列,将该运动序列作为测试运动和基准运动.测试时,先用该运动驱动虚拟 3D 人

体模型,投影为深度图像,根据本文提出的 3D 人体运动跟踪方法重建运动姿态序列,再比较重建出的关节角度值与原运动的基准关节角度值之间的误差,平均重建误差是 1.5 度/帧,如图 8 所示.其中,柱状图(左)是从数据库中选取了 6 个运动序列(包括走路、跑步、踢腿 3 种类型)作为测试数据的各关节角度平局重建误差,最大平均误差不超过 2.5 度/关节;(右)是测试数据的个别帧重建结果截图,黄色点云是输入的投影深度点云,蓝色人体模型是重建的 3D 人体姿态模型.验证了异构 3D 人体姿态数据库的姿态表达能力.该验证实验是在合成数据上进行的,投影的深度图像相比 Kinect 捕获的深度图像,数据噪声较低.



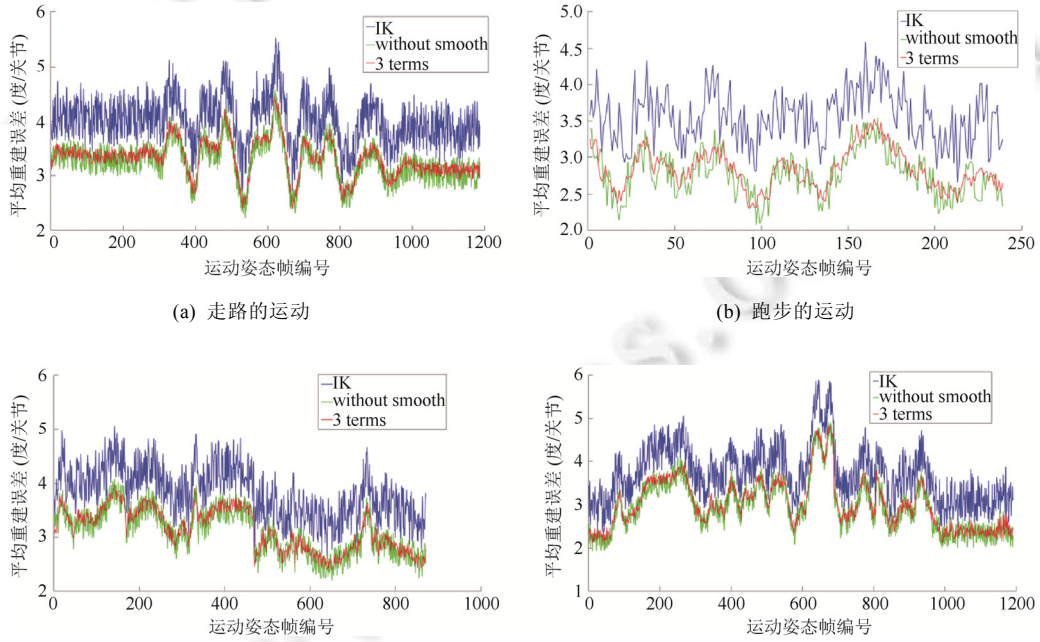

(a) T 姿态时深度图像像素分类结果 (b) 各肢体段标定结果

Fig.8 Cross-Validation

图 8 交叉验证

### 6.2 在线3D人体姿态重建算法各能量项的必要性验证实验

本文设计了针对在线 3D 人体姿态重建算法 3 个能量项的必要性进行测试的验证实验,如图 9 所示.其中,图 9(a)~图 9(d)分别是走路、跑步、打拳和跳舞 4 种运动的关节角平均误差.红色曲线表示同时使用 3 个能量项,绿色曲线表示未使用姿态平滑项,蓝色曲线表示仅使用稀疏 3D 标记点约束项通过 IK[20]的姿态重建结果.




(a) 走路的运动 (b) 跑步的运动

(c) 打拳的运动 (d) 跳舞的运动

Fig.9 Evaluations for three energy terms

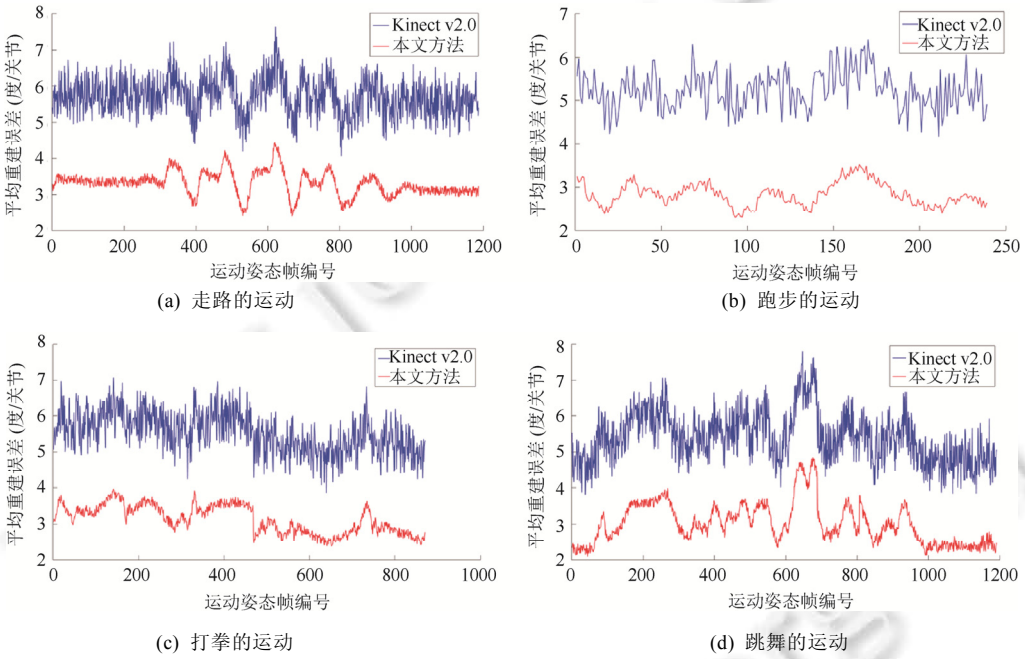
图 9 算法有效性实验

本文使用的测试数据是由 Kinect 采集的 4 种类型的真实人体运动数据,分别为走路(1 187 帧)、跑步(239 帧)、打拳(871 帧)和跳舞(1 191 帧);使用的基准数据是由 Vicon 系统同步采集的标记点数据重建的 3D 人体运动数据.实验中,比较了仅使用稀疏 3D 标记点能量项、稀疏 3D 标记点能量项加局部姿态先验项,以及同时使用 3 个能量项的 4 种类型运动姿态的重建误差.从测试结果上看,对于走路、跑步、打拳和跳舞 4 种类型的运动,仅使用稀疏 3D 标记点重建出的 3D 人体姿态的关节角平均误差均大于加入了局部姿态先验后的关节角平均误差;并且,再加入姿态平滑项的重建结果最好.验证了在线 3D 人体姿态重建算法各能量项是缺一不可的.

因此,通过在稀疏 3D 标记点基础上加入局部 3D 人体姿态先验,相比单纯使用稀疏 3D 标记点跟踪出的 3D 人体姿态,其稳定性和准确性更高.

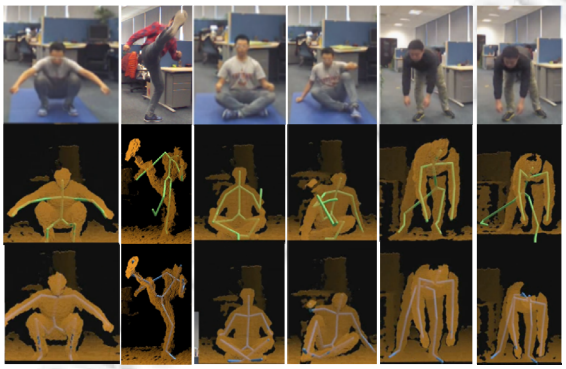
### 6.3 与Kinect重建姿态的准确性对比实验

本文设计了与目前商用系统 Kinect 的 3D 人体姿态重建结果的对比实验.实验中,测试数据和基准数据与第 6.2 节一致,3D 人体姿态重建误差对比结果如图 10 所示.




(a) 走路的运动 (b) 跑步的运动

(c) 打拳的运动 (d) 跳舞的运动



(e) 截取了几帧本文方法(蓝色骨架)与 Kinect(绿色骨架)的姿态重建结果

Fig.10 Comparation with Kinect v2.0

图 10 算法对比实验

图 10(a)~图 10(d)分别是走路、跑步、打拳和跳舞 4 种运动的关节角平均误差,蓝色和红色曲线分别表示 Kinect v2.0 和本文方法的关节角平均误差.使用本文算法跟踪出的 3D 人体姿态关节角平均误差小于 Kinect v2.0 的重建结果,姿态曲线也更平滑,验证了本文的 3D 人体运动跟踪方法具有更高的稳定性和准确性

### 6.4 针对不同身材尺寸表演者的有效性和通用性验证实验

本文设计了 3D 人体运动跟踪方法针对不同身材尺寸表演者的有效性和通用性实验.实验中,用 Kinect 分别捕获了 4 位身材尺寸(见表 1)具有较大差异表演者的运动数据,重建出 3D 人体姿态序列,部分姿态重建结果如图 11 所示.其中,图 11(a)~图 11(d)分别是 4 个不同的表演者的同步采集的 RGB 图像序列(各自第 1 行)和对应的 3D 人体姿态跟踪结果(各自第 2 行),验证了本文方法针对不同身材尺寸表演者也具备有效性和通用性.

**Table 1** Four actors with large different body sizes

**表 1** 身材尺寸不同的 4 个表演者

| 编号 | 性别 | 身高(cm) | 体重(kg) |
|---|---|---|---|
| (a) | 男 | 180m | 90 |
| (b) | 男 | 175 | 60 |
| (c) | 男 | 182 | 55 |
| (d) | 女 | 162 | 50 |

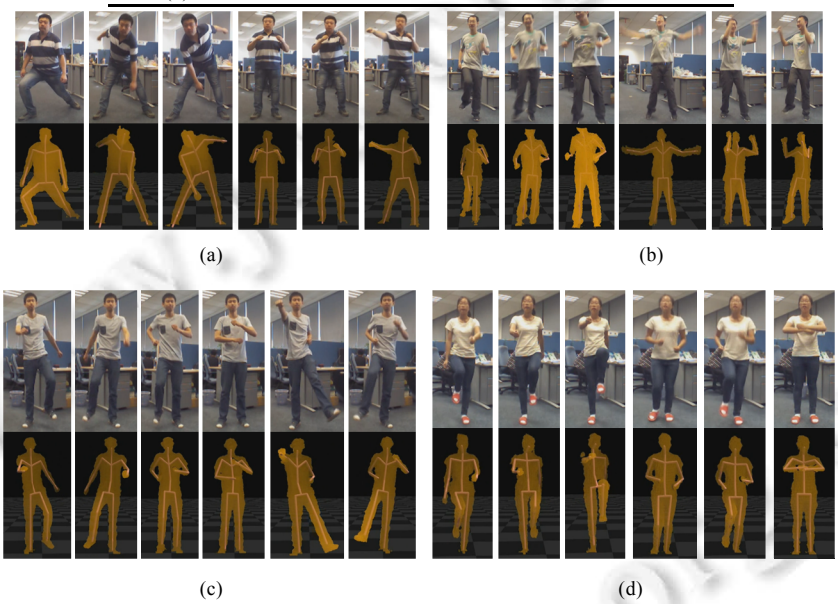



Fig.11 Evaluation for algorithm generalization on actors with different body sizes

图 11 针对不同身材尺寸表演者的算法通用性实验结果

## 7 总结与讨论

本文提出了一种基于局部姿态先验模型的从深度图像中实时在线跟踪 3D 人体运动的方法.该方法无需人工初始化 3D 人体姿态,也无需事先构建个体化 3D 人体模型,即能自动根据捕获的深度图像实时在线地跟踪重建出稳定和准确的 3D 人体运动姿态动画序列.同时,通过一个自动根据捕获的标准姿态下的人体深度图像自动估计个体化人体参数的过程,进一步提高了 3D 人体姿态跟踪算法的泛化能力与实用性,具有很高的应用价值.例如,该方法可以应用在 3D 游戏开发、3D 电影制作、人机交互控制、虚拟网络社交、体育训练与运动康复等多个领域中.

本文算法需要使用从捕获的深度图像提取的稀疏 3D 标记点作为空间约束,但这些标记点的准确性受预先训练的随机决策森林分类器的泛化能力和捕获的深度图像深度值精度等因素的影响,并且发生肢体自遮挡时

会发生丢失再出现情况.本文主要通过使用 3D 姿态数据库构建姿态先验的方法降低稀疏 3D 标记点噪声和少量丢失再出现所产生的影响.如何处理更严重的肢体自遮挡人体姿态,如前滚翻或被周围环境物体遮挡等运动时,会存在稀疏 3D 标记点同时大量丢失再出现的情况,是本文的下一步研究工作.

**References**:


[1] KINECT. Microsoft Kinect for Xbox 360. 2010.

[2] KINECT. Microsoft Kinect for Windows. 2015.

[3] Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A. Real-Time human pose recognition in parts from a single depth image. In: Proc. of the CVPR. IEEE, 2011.

[4] Girshick R, Shotton J, Kohli P, Criminisi A, Fitzgibbon A. Efficient regression of general-activity human poses from depth images. In: Proc. of the IEEE 13th Int'l Conf. on Computer Vision. 2011. 415–422.

[5] Breiman L. Random forests. Mach. Learning, 2001,45(1):5–32.

[6] http://www.vicon.com/

[7] Baak A, Muller M, Bharaj G, Seidel HP, Theobalt C. A data-driven approach for real-time full body pose reconstruction from a depth camera. In: Proc. of the IEEE 13th Int'l Conf. on Computer Vision (ICCV). 2011. 1092–1099.

[8] Ye M, Wang X, Yang R, Ren L, Pollefeys M. Accurate 3D pose estimation from a single depth image. In: Proc. of the IEEE 13th Int'l Conf. on Computer Vision. 2011. 731–738.

[9] Plagemann C, Ganapathi V, Koller D, Thrun S. Real-Time identification and localization of body parts from depth images. In: Proc. of the ICRA. 2010.

[10] Grest D, Woetzel J, Koch R. Nonlinear body pose estimation from depth images. In: Proc. of the DAGM. Vienna, 2005.

[11] Grest D, Kruger V, Koch R. Single view motion tracking by depth and silhouette information. In: Proc. of the 15th Scandinavian Conf. on Image Analysis (SCIA). 2007. 719–729.

[12] Knoop S, Vacek S, Dillmann R. Fusion of 2D and 3D sensor data for articulated body tracking. Robotics and Autonomous Systems, 2009,57(3):321–329.

[13] Wei X, Zhang P, Chai J. Accurate realtime full-body motion capture using a single depth camera. ACM Trans. on Graph., 2012, 31(6):1–12.

[14] Chai J, Hodgins J. Performance animation from low-dimensional control signals. ACM Trans. on Graphics, 2005,24(3):686–696.

[15] Liu H, Wei X, Chai J, HA I, Rhee T. Real-Time human motion control with a small number of inertial sensors. In: Proc. of the Symp. on Interactive 3D Graphics and Games. ACM, 2011. 133–140.

[16] CMU Mocap Database. http://mocap.cs.cmu.edu/

[17] Gleicher M. Retargeting motion to new characters. In: Cohen M, ed. Proc. of the SIGGRAPH'98 Annual Conf. Series. Addison Wesley: ACM, 1998. 33–42.

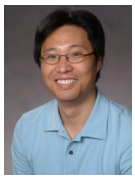
[18] Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis. IEEE Trans. on PAMI, 2002,24(5).

[19] Fischler MA, Bolles RC. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Comm. of the ACM, 1980,24(6):381–395.

[20] Zhao J, Badler N. Inverse kinematics positioning using nonlinear programming for highly articulated figures. ACM Trans. on Graphics (TOG), 1994,13(4):313–336.



**苏乐**(1984－),男,天津人,博士,主要研究领域为计算机图形学,计算机视觉,虚拟现实,人工智能.



**柴金祥**(1975－),男,博士,副教授,博士生导师,主要研究领域为计算机图形学,计算机视觉,虚拟现实,人工智能.

**夏时洪**(1974－),男,博士,研究员,博士生导师,CCF 高级会员,主要研究领域为计算机图形学,虚拟现实,人工智能.