多模态人体运动同步数据集

程景铭[1)], 谢文军[2)*], 沈子祺[1)], 李琳[1)], 刘晓平[1)]

[1)] (合肥工业大学计算机与信息学院 合肥 230601)
[2)] (合肥工业大学软件学院 合肥 230601)
(wjxie@hfut.edu.cn)

摘 要: 人体运动数据集是运动数据去噪、运动编辑及运动合成等研究的重要基础. 为支撑更具通用性的多模态数据融合研究, 设计并采集一套公开的多模态人体运动数据集是亟待解决的问题. 首先设计基于传感器的动作捕捉设备采集精准的运动数据、基于体感设备采集的粗糙运动数据、基于惯性测量单元采集的局部惯性数据的采集环境; 然后基于网络时间协议实现设备间时序同步, 以及多模态数据间的空间同步; 最后分类采集了全身运动多模态数据集(HFUT multimodal motion dataset, HFUT-MMD), 包含 12 位采集者进行 6 类运动的总计 6 971 568 帧数据. 利用已有算法在 HFUT-MMD 数据集上的实验结果表明, 低精度运动数据经过模型优化能够得到与精准的运动数据相近的运动数据, 佐证了各模态数据间的一致性.

关键词: 人体运动数据; 多模态运动数据; 动作捕捉; 体感设备; 惯性测量单元
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2022.19194

Multimodal Human Motion Synchronization Dataset

Cheng Jingming[1)], Xie Wenjun[2)*], Shen Ziqi[1)], Li Lin[1)], and Liu Xiaoping[1)]

[1)] (*School of Computer Science and Information Technology, Hefei University of Technology, Hefei* 230601)
[2)] (*School of Software, Hefei University of Technology, Hefei* 230601)

Abstract: Human motion dataset is an important foundation for researches such as motion data denoising, motion editing, motion synthesis, etc. In order to support more generic studies of multimodal motion data fusion, designing and collecting a public multimodal human motion data set is an urgent problem. First, the acquisition environment is designed for precise motion data collected by sensor-based motion capture devices, rough motion data collected by body sensing devices, and local inertial data collected by inertial measurement units (IMU). Then, temporal synchronization among equipment is applied based on network time protocol (NTP) and spatial synchronization is applied among multi modal data. A full body motion dataset named HFUT-MMD is captured, which contains 6 971 568 frames in 6 types from 12 actors/actresses. The experimental results on the HFUT-MMD dataset using the existing algorithm show that the low precision motion data can be optimized to obtain the motion data similar to the accurate motion data, which corroborates the consistency between the modal data.

Key words: human motion data; multimodal motion data; motion capture; somatosensory devices; inertial measurement units

收稿日期: 2021-06-22; 修回日期: 2021-07-17. **基金项目**: 国家自然科学基金(61877016). **程景铭**(1996—), 男, 硕士研究生, CCF 学生会员, 主要研究方向为计算机图形与可视化; **谢文军**(1984—), 男, 博士, 讲师, 论文通信作者, 主要研究方向为运动数据处理, 深度学习算法; **沈子祺**(1996—), 男, 硕士研究生, 主要研究方向为计算机图形与可视化; **李琳**(1977—), 女, 博士, 副教授, 硕士生导师, 主要研究方向为计算机图形与可视化; **刘晓平**(1964—), 男, 博士, 教授, 博士生导师, CCF 理事, 主要研究方向为计算机图形学、协同计算.

以数据驱动的人体运动分析与合成可以支撑如医疗康复[1]、增强现实/虚拟现实(augmented reality/virtual reality, AR/VR)[2]等方面的应用，是热门的研究领域，同时也需要大量的运动数据支撑. 近年来，随着深度学习方法的兴起，研究人员开始使用神经网络挖掘大量运动数据背后的相关性，支撑如去噪[3]、编辑[4]和合成[5]等操作，其效果得到了显著的提升.

相应地，深度学习方法对数据的规模、可靠性和统一性也提出了更高的要求. 使用单一设备采集运动有弊端，采集得到的数据模态单一. 使用专业动作捕捉设备采集的数据精度最高，但价格昂贵，条件苛刻，需要穿戴特定服饰，导致采集运动种类有限[6]；使用相机采集运动数据会造成采集运动的环境受限[7]，采集得到的运动类型有限[8-9]；使用多相机采集运动数据可以扩充运动种类，但会增加安装采集环境的复杂性；使用惯性传感器采集运动数据，得到的数据精度较高，但数据含有的运动信息量少[10].

现有的主要采集设备有各自的优缺点，而神经网络可以在不同类型的数据之间建立连接，挖掘不同维度和结构数据间的特征关联，实现数据的扩展和去噪，因此使用多种设备采集多模态运动数据成为主流.

由多种不同类型的设备采集得到的数据形式、精度和结构不同，通过使用多模态数据间的时空同步方法可得到多模态数据. 现有的多模态数据集不能满足研究的需要，如 CMU-MMAC[11]数据集数据量少；Human3.6M[12]，Berkeley MHAD[13]等数据集采集环境安装复杂；UTD-MHAD[14]，UR[15]等数据集运动类型单一. 因此，本文设计并采集一个拥有多种运动类型的配对数据集，低精度和高精度的配对数据集能够有效地支撑去噪、编辑和合成等操作，得到高精度的运动数据. 本文的工作主要包括 3 个方面.

(1) 设计一种人体运动的多模态同步采集方案，采集包括人体运动的基准数据、骨骼关节位置以及惯性数据的全身运动数据集(HFUT multimodal motion dataset, HFUT-MMD).

(2) 针对各个采集设备在时间上和空间上的不一致问题，基于网络时间协议(network time protocol, NTP)原理提出多模态数据时空同步的解决方案.

(3) 基于现有的算法，如卷积神经网络[16](convolutional neural networks, CNN)、双向循环自编码器[17](bidirectional recurrent autoencoder, BRA)、基于感知的双向循环自编码器[18](perceptual-based bidirectional recurrent autoencoder, BRA-P)和深度惯性姿势[19](deep intertial poser, DIP)，对数据集进行评估，验证本文创建的数据集的通用性.

1 相关工作

人体运动数据的采集方式有多种，根据采集方式的不同，大致可以将现有的数据集以数据模态分为单模态和多模态 2 种. 表 1[6,11-12,14,19-25]列出了 2 种模态数据集的主要代表，下面将分别介绍这 2 种采集方式的现有数据集.

表 1 常见运动数据集

数据集	运动片段	种类	采集者	模态
HDM05[6]	1531745	100	5	Joints
CMU-MMAC[11]	6650	23	5	RGB+3DJoints+IMU+Audio
Human3.6M[12]	3578080	6	11	RGB+D+3DJoints
UTD-MHAD[14]	861	27	8	RGB+D+3DJoints+IMU
DIP-IMU[19]	330178	5	10	IMU
KTH[20]	2391	6	25	RGB
UCF YouTube[21]	3040	11		RGB
IXMAS[22]	2340	13	10	RGB
MADS[23]	51267	5	5	RGB+D+3DJoints
UMONS-TAICHI[24]	2200	13	12	RGB+D+3DJoints
TotalCapture[25]	1892176	4	5	RGB+3DJoints+IMU

1.1 单模态数据集

早期受设备和场景的限制，只能在静态和固定的单相机视角下采集单一类型的数据，称为单模态数据，典型的数据集代表是 KTH[20]. 为进一

步探索更复杂、真实和自然的人体运动, 研究人员试图构建更逼真的数据集, 考虑了如姿势、相机视角、光照和遮挡等多种因素. 这一阶段产生许多单模态数据集, 如 UCF YouTube[21]等一系列优秀的数据集, 其中的动作样本是从好莱坞电影以及 YouTube 视频中收集得到的, 数据集的动作种类多种多样, 但样本总量偏小且数据种类单一. 由于单个相机采集位置固定且采集区域较小, 不利于部分动作的识别, 于是研究人员采用多个相机从不同的视角采集数据, 以获取更多的视觉特征, 相应的数据集有 HDM05[6]和 IXMAS[22]. HDM05 数据集使用多个相机获取运动捕捉数据, IXMAS 数据集将 5 个相机放在不同的视角, 采集 5 个不同视角的可见光图像. 相同的动作在不同的视角下有不一样的视觉特征, 利用不同视角的视觉特征可以补充单个相机采集时物体遮挡造成的数据缺失, 但多个相机的采集环境限制较多, 需要固定相机位置以及一个空旷的采集环境.

除了使用相机采集人体运动外, 还可以利用惯性测量单元(inertial measurement unit, IMU)捕获人体运动的惯性数据恢复人体姿态.

使用 IMU 采集人体运动可以避免直接视线的需要, 如 DIP-IMU[19]是穿戴 6 个 IMU 采集得到的数据集, 其优点是采集场所灵活, 不需要考虑遮挡和光照等因素. 但仅包含 IMU 数据也有一些问题, 如穿戴 IMU 节点过多会导致人体的侵入性, 穿戴稀疏的 IMU 会导致数据量太少难以实现运动恢复.

1.2　多模态数据集

随着微软公司推出了 Microsoft Kinect, 出现了如 MSR DailyActivity3D[26]等数据集. 这些数据集利用单个 Kinect 采集多种类型的数据, 从不同的角度描述人体运动, 称为多模态数据, 如可见光图像、深度图像和骨骼位置等. 使用 Kinect 采集有很多优点, 如多模态数据融合有利于提高动作识别精度; 在光线较弱的环境有可靠的深度数据等. 由于单个 Kinect 采集会产生如上述单个相机采集的一系列缺点, 因此研究人员使用多个相机采集运动数据, 出现了如 Human3.6M[12], MADS[23], UMONS-TAICHI[24]等数据集. 基于这些数据集, 能够研究多模态人体动作识别或基于多视图的人体动作识别, 并能够挖掘不同视图和模态之间的潜在互补关系.

为了采集更可靠的数据集, 研究人员使用多种设备同时采集, 最早的 CMU-MMAC[11]是同时使用光学相机、IMU 以及音频采集的数据集. CMU 数据集包含多种模态, 但使用光学动作捕捉需要固定相机和无遮挡的采集环境, 采集数据困难. 除此之外, 还有 UTD-MHAD[14]和 TotalCapture[25]等融合多种采集方式的数据集. 但大部分的数据集是为研究某种特定问题而提出的, 其规模比较小, 因此有必要建立一个大型的、可共用的多模态数据集.

本文受数据集的启发, 构建一个包含动作捕捉设备采集的基准数据、Kinect 设备采集的含较高噪声的数据, 以及惯性传感器采集的低信息量的数据的多模态数据集.

2　多模态运动数据采集设备介绍

本文采用了动捕设备、Kinect 设备和惯性传感器设备采集多模态的人体运动数据, 下面将分别对这 3 种设备详细介绍.

2.1　动作捕捉设备

现有的动作捕捉设备主要有 2 种: 基于光学的动作捕捉设备和基于传感器的动作捕捉设备, 它们各有优缺点. 基于光学的动作捕捉设备对采集的环境要求严格, 且采集者需要穿戴贴身标记服; 基于传感器的动作捕捉设备通过绑带或贴身服装固定在身体表面, 可以外穿日常服饰, 对其他采集设备的影响小. 基于以上比较, 基于传感器的动作捕捉设备适合构建多模态运动数据集. 市面上基于传感器的动作捕捉设备具有代表性的有 Xsens Moven(简称为 MVN)和 Noitom Perception Neuron(简称为 NPN), 它们参数对比如表 2 所示.

表 2　MVN 和 NPN 的主要参数对比

参数	MVN	NPN
节点数量	最大 22 个(不含手指)	最大 32 个(包含手指)
传感器尺寸/mm^3	31.5×28.0×13.0	12.5×13.1×4.3
连接方式	无线连接	有线或无线连接
穿戴方式	紧身采集服或绑带	绑带
静态精度	无	俯仰角与翻滚角精度为±1°, 航偏角精度为±2°
最大测量范围	角速度±2000°/s, 加速度±98m/s^2	角速度 ± 2000°/s, 加速度 ± 156.8m/s^2

通过对比可见, NPN 设备传感器尺寸小, 更加轻便且包含手指运动, 能够采集更为丰富的运动信息, 虽然精度略低于 MVN, 但仍可满足应用需求. 因此, 本文选用 NPN 设备采集高精度人体运动数据.

本文选择 Biovision 层次模型(BioVision hierarchical, BVH)作为动作捕捉运动数据格式. 输出 BVH 格式的骨架由 72 个节点构成, 如图 1 所示.

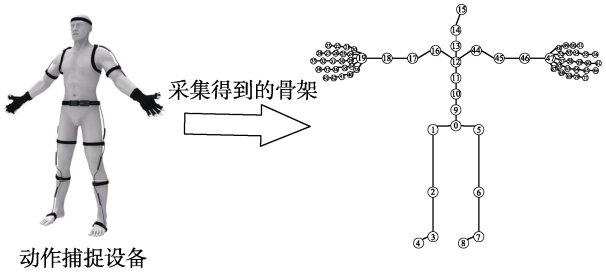


图 1 动作捕捉设备及其采集的骨架

2.2 可见光/深度相机设备

Kinect 作为当前最具代表性和普适性的体感游戏交互设备, 在各类图像设备中具有广泛的代表性, 也是相关研究和数据集的主要数据采集设备, 因此能够更有效地支持数据集的使用者验证其算法. Kinect v2 采集的骨架与动作捕捉设备采集得到的骨架相似性高, 因此本文选用 Kinect v2 设备, 如图 2 所示. Kinect v2 的一些特性如表 3 所示.

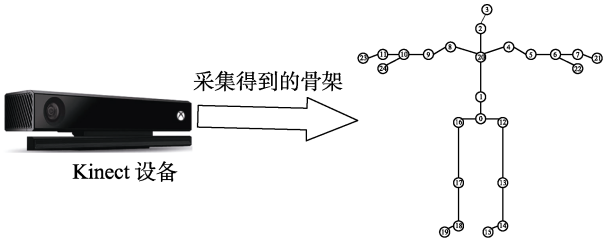


图 2 Kinect 设备及其采集的骨架

表 3 Kinect v2 特性

功能	描述
RGB 相机	1 920×1 080 像素@ 30 帧/s
深度相机	512×424 像素@ 30 帧/s
最佳采集/最大采集距离	0.8~3.5 m / 0.5~4.5 m
最佳水平/垂直采集角度	70º / 60º
人体跟踪	25 个节点最多同时允许 6 人

当前采集系统采用一台 Kinect v2 设备放置于角色正前方, 使用内置算法直接采集人体骨骼信息. Kinect 设备具有采集范围小和数据噪声大的特点, 因此, 基于官方提供的设备参数和通过对 Kinect 采集效果进行验证, 在采集场地中绘制出 2 条采集参考线, 将采集区域划分成最佳采集区、最大采集区和不可采集区 3 个区域.

2.3 IMU 设备

IMU 是测量物体三轴姿态角(或角速度)以及加速度的装置. 现实中的许多设备内置 IMU, 如智能手机、智能手表和智能手环等. 因此, 在实际场景中, 可以便捷地采集 IMU 数据. 由于本文构建的数据集需要多个设备同时采集, 使用磁场过强的 IMU 采集设备会对其他采集设备造成干扰, 影响到数据精度. 因此, 为了获取精准的运动数据, 本文采用 ATOM Matrix 作为 IMU 数据的采集设备, 内置 IMU 姿态传感器(MPU6886), 还集成了 WiFi、蓝牙等模块, 便于 IMU 数据的采集与传输.

IMU 设备及其佩戴位置如图 3 所示, 能够采集相应位置的加速度以及姿态角.

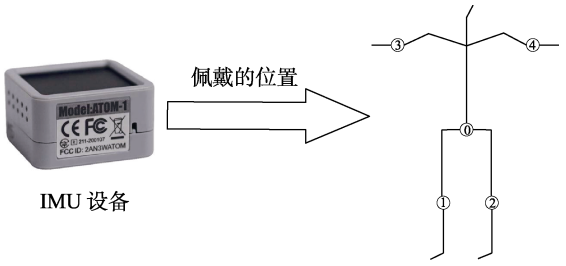


图 3 IMU 设备及其佩戴位置

MPU6886 采集得到的数据是包含加速度和陀螺仪的相关数据, 表示一种运动趋势. MPU6886 参数信息如表 4 所示.

表 4 MPU6886 参数

描述	参数
陀螺仪灵敏度误差	±1%
陀螺仪噪声	±4 mdps/ $\sqrt{\text{Hz}}$
加速计噪声	100 μg/ $\sqrt{\text{Hz}}$
工作温度	–40~ + 85℃

3 多模态运动数据采集方案

3.1 采集环境布局与设备关系

数据集采集场地的面积设置主要参考 Kinect 的 RGB-D 相机采集范围(采集距离为 0.5~4.5 m), 地面标记了最佳采集区和最大采集区, 用于引导采集对象约束其运动. 与此同时, 该尺寸对多数 VR/AR 设备(如 HTC Vive, Hololens)的运动空间具有良好的兼容性, 能够为虚拟现实空间中角色动

画和自然交互技术提供有效的支撑.

采集环境的空间示意图如图 4 所示.

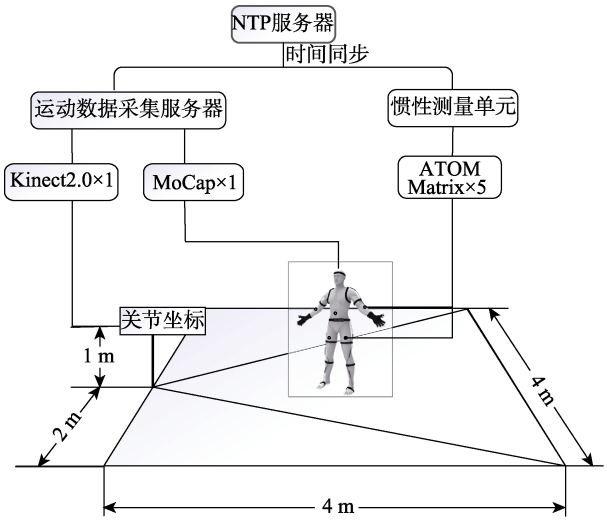


图 4　采集环境空间示意图

如图 4 所示, 采集空间布置在至少 4m×4m×2m 的室内场景中, 采集空间中除了可能存在的设备线缆和交互道具(如桌子、椅子、门窗)外, 无任何遮挡, 采集空间的背景较为单一.

当前采用 Tpose 为初始姿态, 模特应面向 Kinect 设备方向直立, 双眼平视前方, 双手水平张开, 手指并拢, 掌心向下. 真实采集环境如图 5 所示.

图 5　真实采集环境

在采集场地中, 内圈为最佳采集区, 能够完整地拍摄到模特全身; 内圈与外圈之间是最大采集区, 在这部分区域运动, 部分肢体会超出设备采集范围, 超出采集范围的关节点在骨骼位置数据中会自动预测和补全, 但 Kinect 在补全节点时会出现位置错误, 导致运动幅度很大的噪声. 在采集时, 模特被要求在采集过程中不限制运动幅度, 自然地进入该噪声区域, 生成具备该类噪声数据集.

Kinect 设备和动作捕捉设备在采集运动数据时基于各自的坐标系, 2 种设备采集的运动数据在空间表示上有区别, 需要将 2 种运动数据统一到相同坐标系下. 因此, 需要对 Kinect 运动数据变换得到在空间与动作捕捉运动数据一致的运动数据.

Kinect 骨骼运动数据的空间同步分为旋转、平移和缩放 3 步. 对于第 i 帧运动数据中第 j 个骨骼节点坐标点 (x_{ij}, y_{ij}, z_{ij}), 其空间同步坐标为

$$(x'_{ij}, y'_{ij}, z'_{ij})^{\mathrm{T}} = m \cdot (\boldsymbol{A}_{\mathrm{Rot}} \cdot (x_{ij}, y_{ij}, z_{ij})^{\mathrm{T}} - \boldsymbol{A}_{\mathrm{offset}}) \quad (1)$$

其中, m 为放大倍数; $\boldsymbol{A}_{\mathrm{Rot}}$ 为 (x_{ij}, y_{ij}, z_{ij}) 绕 y 轴旋转的旋转矩阵, 求解旋转矩阵所用到的旋转角在现有采集设备固定的情况下是定值 60°; $\boldsymbol{A}_{\mathrm{offset}}$ 为第 i 帧 Kinect 根节点数据与动作捕捉根节点数据间的偏移量, 定义为

$$\boldsymbol{A}_{\mathrm{offset}} = \boldsymbol{A}_{\mathrm{Rot}} \cdot (x_{ki0}, y_{ki0}, z_{ki0})^{\mathrm{T}} - (x_{di0}, y_{di0}, z_{di0})^{\mathrm{T}} \quad (2)$$

其中, $(x_{ki0}, y_{ki0}, z_{ki0})$ 为 Kinect 数据第 i 帧运动数据中根节点的坐标; $(x_{di0}, y_{di0}, z_{di0})$ 为动作捕捉设备第 i 帧运动数据中根节点的坐标.

经过对 Kinect 数据的旋转、平移和放大后, 可以得到一个在空间上与动作捕捉运动数据高度一致的 Kinect 骨骼运动数据, 如表 5 所示(由于 Kinect 骨骼位置数据过小, 表中空间同步之前的 Kinect 骨骼数据放大了 100 倍).

表 5　Kinect 数据空间同步前后对比

数据变换	站立	举手	跳跃
空间同步前			
空间同步后			

注. 绿色代表 Kinect, 红色代表动作捕捉.

IMU 设备有自己独立的坐标系, 需要将其坐标系统一. 文献[19]提出了一种 IMU 数据的坐标系解决方案, 本文也使用此方案将四肢的 IMU 数据归一化到以腰节点为根节点的坐标系中, 实现 IMU 数据坐标系的统一.

3.2　多模态数据的时间同步

多模态数据的时间同步是以动作捕捉数据作

为基准数据进行数据重采样. 本文的数据集系统采集的有动作捕捉设备采集的 BVH 格式的数据、Kinect 采集的节点位置、ATOM Matrix 设备采集到的 IMU 数据. 本文构建的数据集使用多种设备采集, 设备之间的时钟不一致, 采集得到的时间戳不同. 因此, 首先需要解决的问题是统一各个设备之间的时钟, 以获得基于同一个时钟下的时间戳. 本文的设计方案是将一台本地计算机搭建成时间同步服务器, 各个设备在采集时使用时间同步服务器的时间戳, 这样保证了多个设备使用同一个时钟.

动作捕捉设备的采集软件安装在本地计算机上, 动作捕捉设备通过直连的方式将数据传输到本地计算机, 因此使用本地计算机的时间戳. 动作捕捉软件在采集数据时没有保存采集的起始时间, 因此需要通过额外记录鼠标语言的方式记录采集起始时间. 采集软件会记录运动数据的固定步长 s 和姿态数据 f 帧, 因此其总时间长度应为 $s\cdot f$. 通过捕获点击采集按钮和结束按钮的时间戳信息, 能够获取开始时间 t_s 和结束时间 t_e, 鉴于交互到响应存在延迟, 经过多次实验测得交互到响应的延迟是稳定的, 因此数据集采用人工标记的 λ 作为同步基准.

Kinect 设备与动作捕捉设备相同, 也是直连本地计算机. 本文通过修改微软官方提供的人体跟踪软件开发工具包(Software Development Kit, SDK)的代码, 在实时追踪人体姿态的同时保存每帧运动的时间戳, 此时采集得到的时间戳是基于本地计算机的时间戳, 与动作捕捉设备的时间戳在同一个时钟下.

IMU 设备是一种微型的开发板, 其内部无独立的时钟, 无法获得完整的时间戳. 由于该设备内部集成了 WiFi 模块, 因此本文在开发板上实现了 NTP, 获取时间同步服务器的时间. NTP 原理如图 6 所示.

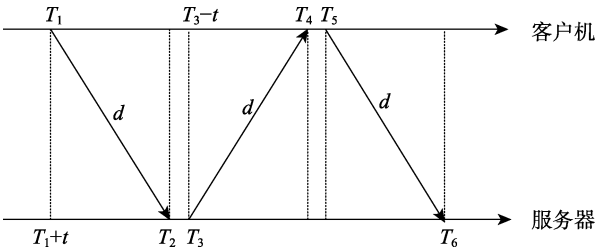


图 6 NTP 原理($T_1\sim T_4$ 为标准 NTP, T_5 和 T_6 为本文扩展)

其中, $T_1\sim T_6$ 均为时间戳; d 为传输时延. T_1 为客户机发送 NTP 请求时间; T_2 为服务器收到 NTP 请求时间戳; T_3 为服务器回复 NTP 请求时间戳; T_4 为客户机收到 NTP 回复时间. 由于本文使用的 IMU 设备内部没有时钟, 无法获得 T_1 和 T_4 的时间戳, 仅有 T_2 和 T_3 的时间戳. 因此, 在标准 NTP 的基础上, 当客户机收到 NTP 回复包时, 在 T_5 向服务器发送一个包含 T_2 和 T_3 的 NTP 包, 服务器在 T_6 接收到 NTP 回复包, 服务器可以利用 T_6, T_2 和 T_3 计算出传输时延 d. IMU 设备把 (T_2-d) 作为接收到的数据帧的时间戳. 经过多次重复实验, 计算出传输时延 $d=(33\pm13\ \text{ms})$.

通过以上方法, 本文将各设备的时间戳统一到时间同步服务器, 实现了各设备的时钟同步. 而设备间采集的起止时间各不相同, 采样频率也不同. 因此, 需要对各设备采集的数据重采样. 动作捕捉软件的采样频率为 60 帧/s, Kinect 的采样频率为 30 帧/s, IMU 设备的采样频率约为 15 帧/s. 本文以动作捕捉采集软件采集到数据作为基准数据, 其他设备采集的数据对齐动作捕捉数据. 重采样是利用记录的时间戳对各个设备采集的数据在时间上裁剪对齐, 并使用样条插值实现 Kinect 和 IMU 数据与动作捕捉数据的同步. 文献[27]证明了 3 次自然样条插值后数值的稳定性, 本文使用样条插值前后数据对比如图 7 所示.

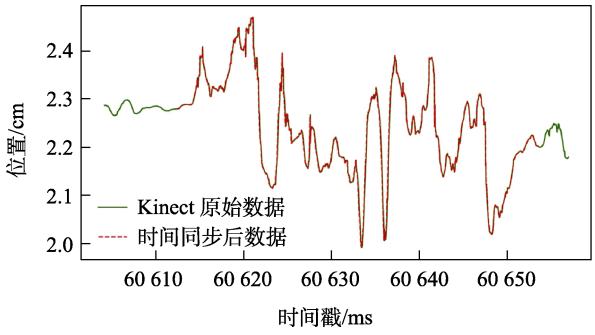


图 7 Kinect 数据时间同步前后对比

3.3 数据集方案的索引与标定

HFUT-MMD 数据集由 12 人(8 男 4 女, 编号为 00~11)采集, 包含 6 个动作类别(分别是广播体操、八段锦、体育运动、五禽戏、日常活动和舞蹈, 编号为 S_1~S_6), 共 34 种动作, 总计 6971568 帧. 表 6 和表 7 分别列出了数据集采集模特和每类动作的采集帧数等信息, 其中采集时有动作捕捉设备和 Kinect 同时采集的配对数据, 有动作捕捉设备和 IMU 同时采集的配对数据, 也有动作捕捉设备、Kinect 和 IMU 三者同时采集的配对数据.

表 6　模特数据信息分类及采集帧数

模特编号	性别	身高	动作捕捉帧数	Kinect 帧数	IMU 帧数
09	女	155	132 948	132 948	132 948
10	女	160	43 034	13 258	43 034
05	女	165	179 902	179 902	0
07	女	170	33 680	33 680	0
04	男	170	171 231	171 231	171 231
00, 03, 06, 11	男	175	1 860 469	1 860 469	335 640
01, 02, 08	男	180	540 317	540 317	395 329

表 7　动作数据分类及采集帧数

动作编号	动作类别	动作捕捉帧数	Kinect 帧数	IMU 帧数
S_1	广播体操	621 208	621 208	129 944
S_2	八段锦	659 291	629 515	583 859
S_3	体育运动	764 570	764 570	272 335
S_4	五禽戏	179 902	179 902	0
S_5	日常活动	702 930	702 930	92 044
S_6	舞蹈	33 680	33 680	0
总计		2 961 581	2 931 805	1 078 182

4　多模态同步运动数据应用

4.1　Kinect 骨骼运动数据去噪

Kinect 相机在体感游戏领域应用广泛, 游戏者在体验虚拟游戏时, 关注的是虚拟体验感. 提高虚拟体验最重要的是提高运动的识别精度, 但是 Kinect 相机采集得到的骨骼位置数据有噪声, 采集者仅在正对相机时可以获得较好的采集效果. 而实际运动中, 采集者会不可避免地出现侧身和弯腰等动作, 对于这类动作, Kinect 采集出的骨骼位置数据噪声很大, 不能恢复正常的运动, 因此需要对 Kinect 数据去噪, 得到更加平滑自然的运动. Holden 等[16]提出了一个基于深度学习的角色运动合成框架, 使用卷积自编码器生成运动的流形, 该流形与人体运动更加契合. 为了更好地生成流形, Holden 等还在训练好的自编码器上叠加一个前馈神经网络, 可以根据高层参数形成更加逼真的运动序列. 该运动序列可以在运动流形空间中优化, 产生高质量、平滑的运动序列, 从而达到数据去噪的效果. 在 Holden 等提出的 CNN 的基础上, 基于双向长短期记忆网络(long short-term memory, LSTM)的自编码器 BRA[17]和 BRA-P[18]对生成的人物添加骨骼长度和平滑度的约束, 以生成更平滑、自然的人体运动. 本文使用这 3 种方法评估本文创建的数据集, 将 S_1~S_6 中的“男 175”数据随机划分, 一部分作为训练数据训练网络, 另一部分作为测试数据测试网络性能. 本文在训练 CNN, BRA, BRA-P 的基础上, 观察到网络生成的运动中手部运动数据去噪效果不好, 故本文在原有约束的基础上对手部加了更高的权重. 通过测试, 在测试运动序列上截取了模特站立、挥手、弯腰、右转和左转的结果, 如表 8 所示.

表 8　CNN, BRA 和 BRA-P 运动数据测试结果

数据	站立	挥手	弯腰	右转	左转
Kinect 采集数据					
CNN[16]输出					
BRA[17]输出					
BRA-P[18]输出					
精确数据					

通过实验结果可以明显地看出, 本系统采集的 Kinect 骨骼位置数据在经过 3 种方法的优化之后, 可以得到一个接近精确数据的结果. 特别是在一些 Kinect 采集效果较差的方向, 如弯腰和侧身等运动, Kinect 采集到的节点会有错位现象. 通过实验结果可以看出, CNN 可以在 Kinect 采集效果较差时, 恢复整个骨架且运动相近, 即网络生成的全身运动数据与真实运动相近. 但网络的特性平滑了许多动作, 造成运动幅度达不到精确数据的结果. 后续的改进网络 BRA 和 BRA-P 使用双向 LSTM 可以最大化保留运动的前后联系, 因此运动的幅度能够达到精确数据的效果, 网络生成的运动与精确数据近乎相同.

文献[17-18]提出了位置误差、骨骼长度误差以及平滑性误差用于比较网络的性能, 本文在实验时也用这 3 个指标评价 3 种方法的优劣, CNN, BRA 和 BRA-P 的各种误差比较如表 9 所示.

通过实验结果可以看出, BRA 和 BRA-P 在位置的精确度上优于 CNN. BRA-P 与 BRA 相比虽然在数值上没有很大的改进, 但在视觉效果上更自然、流畅, 即网络生成的运动更加趋近真实运动. 本文创建的数据集中运动的视觉效果在这些方法上表现良好, 由此得出本文数据集具有较好的适用性.

表 9 CNN, BRA 和 BRA-P 的误差比较

方法	位置误差	骨骼长度误差	平滑性误差
CNN[16]	0.054587	0.362243	0.564507
BRA[17]	0.003319	0.365130	0.472645
BRA-P[18]	0.003183	0.432121	0.449990

4.2 IMU 人体运动重建

在医疗康复中，通常需要检测人体运动评判病人恢复程度，受采集环境的限制，不用相机采集相关数据. 稀疏的 IMU 刚好能够满足这个需求，没有环境的限制，对人体侵入性小. 但使用稀疏 IMU 设备的弊端是携带的运动信息量少，采集得到的数据仅包含方向和加速度，很难直接根据现有的运动数据重构人体姿态. Huang 等[19]提出了一种基于深度学习的人体姿态估计(deep inertial poser, DIP)方法，只需要 6 个 IMU 输入就能重构人体姿态，且实时运行. 该方法将 6 组 IMU 数据输入双向循环 LSTM 网络，生成可以驱动 SMPL[28]模型的数据，输入网络的数据需要经过校准方向、补偿传感器漂移和归一化等操作. DIP 论证了对 IMU 数据每帧标准化和使用旋转矩阵代替旋转角的恢复效果更好，因此，本文也将采集得到的 IMU 数据预处理达到 DIP 中叙述的最佳训练效果. 训练所需的基准数据由动作捕捉设备采集的数据转化得到. 使用 S_2, S_3, S_5 的数据做训练，其余部分做测试，测试结果如图 8 所示.

a. 基准运动

b. DIP 生成运动

图 8 S_2 全身运动测试结果

实验结果证明，本文创建的数据库在 DIP 中表现良好，能够恢复绝大部分运动，但由于本文采集的数据库中仅包含 5 个 IMU 节点，不包含头节点，故涉及头部的精细运动恢复不出来，头部运动只能靠全身运动推测，如图 9 所示.

综上所述，本文创建的数据库可以应用到 IMU 人体运动重建中，给相关的研究人员提供了可靠的 IMU 数据.

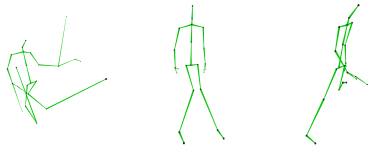
a. 基准运动

b. DIP 生成运动

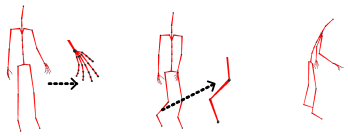
图 9 S_2 头部运动测试结果

4.3 融合 Kinect 和 IMU 数据重构人体运动

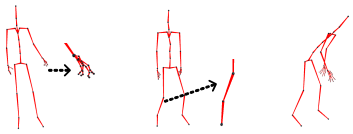
Kinect 设备是光学相机，容易受到遮挡或超出采集范围而产生噪声，IMU 设备是惯性传感器，传感器数量少且包含的运动信息少，但不受环境限制. 文献[14]提出融合 Kinect 和 IMU 数据，可以提高数据精度，本文使用融合 Kinect 和 IMU 数据可以改善 Kinect 运动数据在肢体残缺或出现位置错误时造成的噪声，并使用双向 LSTM 网络重构人体运动；实验对比使用单一 Kinect 数据、单一 IMU 数据和 Kinect 融合 IMU 数据作为输入，测试结果如图 10 所示.



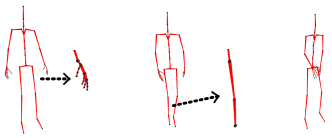
a. Kinect 原始噪声运动



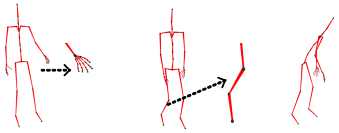
b. 基准运动



c. 单一 Kinect 输入测试结果



d. 单一 IMU 输入测试结果



e. Kinect 融合 IMU 测试结果

图 10 Kinect 融合 IMU 与单一 Kinect、单一 IMU 对比测试结果

实验表明, Kinect 数据可以恢复更多的运动细节以及运动幅度; IMU 数据仅有运动方向, 故恢复的运动在骨架上维持较高的一致性. 在 Kinect 数据异常时, 融合 IMU 数据可以给 Kinect 数据一定的补充, 保证运动的正确性, 得到更优的运动重构结果.

5 结　语

本文创建了 HFUT-MMD, 旨在为人体姿态估计方面发展提供助力. HFUT-MMD 包括 6 类动作, 包含基准数据、Kinect 骨骼位置数据和 IMU 数据, 它们可以从不同的角度描述人体姿态. 动捕设备采集的基准数据一般比较精确, 但价格昂贵, 不能普及; 一些价格便宜的 Kinect 和 IMU 设备更容易普及, 但精度较低, 恢复运动较难. 所以, 本文引用了几种方法分别在 Kinect 骨骼数据去噪、IMU 数据重构人体姿态、融合 Kinect 和 IMU 数据重构人体运动等方面进行了实验, 验证了低精度设备恢复人体运动的可行性. 本文介绍了单一数据的恢复运动、Kinect 数据和 IMU 数据融合的初步尝试, 多模态数据的融合也有相关研究人员在探索, 并且取得了不错的成果. 本文希望 HFUT-MMD 数据集可以提供给相关的研究人员更有效的多模态同步数据集.

参考文献(References):

[1] Peng Xiaolan, Chen Hui, Wang Lan, *et al.* Rehabilitation of aphasia after stroke with physical and virtual integrated system[J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(2): 256-265(in Chinese)
(彭晓兰, 陈辉, 王岚, 等. 虚实融合交互系统辅助脑卒中后失语症康复训练[J]. 计算机辅助设计与图形学学报, 2019, 31(2): 256-265)

[2] Qin Pu, Yang Chenglei, Li Huiyu, *et al*. Virtual reality shooting recognition device and system using MEMS sensor[J]. Journal of Computer-Aided Design & Computer Graphics, 2017, 29 (11): 2083-2090(in Chinese)
(秦溥, 杨承磊, 李慧宇, 等. 采用MEMS传感器感知的虚拟现实射击识别设备与系统[J]. 计算机辅助设计与图形学学报, 2017, 29(11): 2083-2090)

[3] Holden D, Saito J, Komura T, *et al*. Learning motion manifolds with convolutional autoencoders[C] //Proceedings of SIGGRAPH Asia 2015 Technical Briefs. New York: ACM Press, 2015: Article No.18

[4] Holden D, Kanoun O, Perepichka M, *et al*. Learned motion matching[J]. ACM Transactions on Graphics, 2020, 39(4): Article No.53

[5] von Marcard T, Rosenhahn B, Black M J, *et al*. Sparse inertial poser: automatic 3D human pose estimation from sparse IMUs[J]. Computer Graphics Forum, 2017, 36(2): 349-360

[6] Müller M, Röder T, Clausen M, *et al*. Documentation mocap database HDM05[R]. Bonn: Computer Graphics Technical Reports, Universität Bonn, 2007

[7] Blank M, Gorelick L, Shechtman E, *et al*. Actions as space-time shapes[C] //Proceedings of the 10th IEEE International Conference on Computer Vision, Volume 1. Los Alamitos: IEEE Computer Society Press, 2005: 1395-1402

[8] Laptev I, Marszalek M, Schmid C, *et al*. Learning realistic human actions from movies[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2008: 1-8

[9] Niebles J C, Chen C W, Li F F. Modeling temporal structure of decomposable motion segments for activity classification[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2010: 392-405

[10] Zhang M, Sawchuk A A. USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors[C] //Proceedings of the ACM Conference on Ubiquitous Computing. New York: ACM Press, 2012: 1036-1043

[11] de la Torre F, Hodgins J, Bargteil A, *et al.* Guide to the Carnegie Mellon University multimodal activity (CMU-MMAC) database[R]. Pittsburgh: Carnegie Mellon University, 2009

[12] Ionescu C, Papava D, Olaru V, *et al*. Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligen1723ce, 2014, 36(7): 1325-1339

[13] Ofli F, Chaudhry R, Kurillo G, *et al*. Berkeley MHAD: a comprehensive multimodal human action database[C] //Proceedings of the IEEE Workshop on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2013: 53-60

[14] Chen C, Jafari R, Kehtarnavaz N. UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor[C] //Proceedings of the IEEE International Conference on Image Processing. Los Alamitos: IEEE Computer Society Press, 2015: 168-172

[15] Kepski M, Kwolek B. Fall detection using ceiling-mounted 3D depth camera[C] //Proceedings of the International Conference on Computer Vision Theory and Applications. Los Alamitos: IEEE Computer Society Press, 2014: 640-647

[16] Holden D, Saito J, Komura T. A deep learning framework for character motion synthesis and editing[J]. ACM Transactions on Graphics, 2016, 35(4): Article No.138

[17] Li S J, Zhou Y, Zhu H S, *et al*. Bidirectional recurrent autoencoder for 3D skeleton motion data refinement[J]. Computers & Graphics, 2019, 81: 92-103

[18] Li S J, Zhu H S, Zheng L P, *et al*. A perceptual-based noise-agnostic 3D skeleton motion data refinement network[J]. IEEE Access, 2020, 8: 52927-52940

[19] Huang Y H, Kaufmann M, Aksan E, *et al*. Deep inertial poser: learning to reconstruct human pose from sparse inertial measurements in real time[J]. ACM Transactions on Graphics, 2018, 37(6): Article No.185

[20] Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach[C] //Proceedings of the 17th International Conference on Pattern Recognition. Los Alamitos: IEEE Com-

puter Society Press, 2004: 32-36

[21] Liu J G, Luo J B, Shah M. Recognizing realistic actions from videos “in the wild”[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2009: 1996-2003

[22] Weinland D, Boyer E, Ronfard R. Action recognition from arbitrary views using 3D exemplars[C] //Proceedings of the IEEE 11th International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2007: 1-7

[23] Zhang W C, Liu Z G, Zhou L Y, *et al*. Martial arts, dancing and sports dataset: a challenging stereo and multi-view dataset for 3D human pose estimation[J]. Image and Vision Computing, 2017, 61: 22-39

[24] Tits M, Laraba S, Caulier E, *et al*. UMONS-TAICHI: a multimodal motion capture dataset of expertise in Taijiquan gestures[J]. Data in Brief, 2018, 19: 1214-1221

[25] Trumble M, Gilbert A, Malleson C, *et al*. Total capture: 3D human pose estimation fusing video and inertial sensors[C] //Proceedings of the British Machine Vision Conference 2017. Swansea: BMVC Press, 2017: 1-13

[26] Wang J, Liu Z C, Wu Y, *et al*. Mining actionlet ensemble for action recognition with depth cameras[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2012: 1290-1297

[27] Zhang Sanyuan, Sun Shouqian, Pan Yunhe. Cubic algebraic curves interpolation with geometric constraints[J]. Chinese Journal of Computers, 2001, 24(5): 509-515(in Chinese)
(张三元, 孙守迁, 潘云鹤. 基于几何约束的三次代数曲线插值[J]. 计算机学报, 2001, 24(5): 509-515)

[28] Loper M, Mahmood N, Romero J, *et al*. SMPL: a skinned multi-person linear model[J]. ACM Transactions on Graphics, 2015, 34(6): Article No.248