软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(10):3026−3036 [doi: 10.13328/j.cnki.jos.005783]

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

一种无标记的身体与面部运动同步捕获方法*

王志勇[1,2], 王从艺[1,2], 张子豪[1,2], 袁铭择[1,2], 夏时洪[1,2]

[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)

通讯作者: 夏时洪, E-mail: xsh@ict.ac.cn

摘 要: 提供了一个无标记点的身体与面部运动同步捕获的方法.利用经过时间同步和空间标定的长焦彩色相机和 Kinect 相机来进行同步捕获.利用在环境中加入闪光来进行时间同步,使用张氏标定法进行空间标定,从而组成一组时间同步且空间对齐的混合相机(hybrid camera).然后利用 Kinect fusion 扫描用户的人体模型并嵌入骨骼.最后利用时间和空间都对齐好的两个相机来进行同步采集.首先从深度图像中得到人脸的平移参考值,然后在平移参考值的帮助下根据彩色图像的 2D 特征点重建人脸.随后,把彩色图像中得到的头部姿态传递给身体捕获结果.结果对比实验和用户调研实验均表明所提出的运动捕获的结果要好于单个的运动捕获结果.

关键词: 运动捕获;人脸人体同步运动捕获;深度图像;多线性模型方法;非刚性迭代最近点方法

中图法分类号: TP391

中文引用格式: 王志勇,王从艺,张子豪,袁铭择,夏时洪.一种无标记的身体与面部运动同步捕获方法.软件学报,2019,30(10): 3026−3036. http://www.jos.org.cn/1000-9825/5783.htm

英文引用格式: Wang ZY, Wang CY, Zhang ZH, Yuan MZ, Xia SH. Markerless motion capture method combining body capture and face capture. Ruan Jian Xue Bao/Journal of Software, 2019,30(10):3026−3036 (in Chinese). http://www.jos.org.cn/1000-9825/5783.htm

Markerless Motion Capture Method Combining Body Capture and Face Capture

WANG Zhi-Yong[1,2], WANG Cong-Yi[1,2], ZHANG Zi-Hao[1,2], YUAN Ming-Ze[1,2], XIA Shi-Hong[1]

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

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

Abstract: This paper presents a markerless synchronized motion capture method for body and face. A long-focus color camera and a standard Kinect camera are used which are synthesized and calibrated. Flush is added into the capture space to make the cameras synthesized. Then, the cameras are calibrated to get the relative transformation of the cameras by Zhang's calibration, In this way, the hybrid camera is configured. The user's body is scanned by Kinect fusion, and the skeleton is embedded for the body model. During capture, the translation reference is firstly obtained from the depth camera. After that, the facial pose and expression and identity are reconstructed by 2D feature points. Non-rigid ICP is applied to reconstruct the body pose. Finally, the head pose from the face capture camera is transformed to body capture camera to get the merged results. The comparison and user study show that the result of proposed synchronized motion capture by two cameras is better than that of single camera.

Key words: motion capture; synchronized motion capture for body and face; depth image; multilinear model; non-rigid ICP

* 基金项目: 国家自然科学基金(61772499)
Foundation item: National Natural Science Foundation of China (61772499)
本文由"自然人机交互新进展"专题特约编辑田丰、喻纯推荐.
收稿时间: 2018-08-18; 修改时间: 2018-11-01; 采用时间: 2018-12-25; jos 在线出版时间: 2019-04-29
CNKI 网络优先出版: 2019-04-30 09:19:06, http://kns.cnki.net/kcms/detail/11.2560.TP.20190430.0918.007.html

无标记虚拟人运动捕获主要研究从彩色图像或深度图像中重建人体和人脸运动的方法.Xia 等人[1]对当前的人体运动捕获与合成方法进行了整理和总结.与有标记的运动捕获系统相比,无标记运动捕获系统具有对演员干扰小、成本低、使用方便等优点,可以应用于视频直播、交互游戏以及各种其他虚拟人相关的应用中.

在演员表演或人们的日常生活者中,往往通过身体姿势和脸部表情共同表达自己的情绪或者行为.比如,人在表达喜悦或者愤怒的表情时,往往会带有相应的动作.同样地,人在进行动作时,脸上也会带有相应的表情.因此,高真实感的虚拟人动画中往往需要同时具有面部和身体运动捕获结果.然而,现有的运动捕获系统大多将面部和人体的运动捕获分开来,这就造成了需要后续再进行运动捕获结果合并.并且,后续进行运动捕获结果合并时,可能会出现面部运动中的头部动作和人体运动中的头部动作不匹配,或者两次运动不能完全同步的问题.于是产生了同时进行人体和面部运动捕获的需求.

Vicon cara[2]是一种可以同时捕获面部运动和身体运动的捕获系统.这个系统要求场地四周架设红外相机,用户穿好带有标记点的紧身衣,脸上贴好标记点的同时,戴上配有摄像机的头盔.这些额外的要求不但大大提高了运动捕获成本,而且难免会对演员的表演造成干扰.此外,这类方法往往还需要手工制定标记点的对应关系.虽然已有研究者[3]对自动标记点标注的方法进行了研究,但目前提供的方法需要消耗时间,并且会影响系统的稳定性.

为此,我们设计了一种在无标记点的情况下进行面部捕获和身体运动捕获的方法.利用视场角小但分辨率较高的彩色相机捕获面部运动,同时使用视场角大但分辨率较低的 Kinect 深度相机捕获人体运动,以实现面部和身体运动的协同捕获.

在开始捕获之前,首先对两个相机进行时间同步和空间标定,然后开始数据捕获.捕获完成之后,首先把两部分运动捕获结果分别进行运动重建,然后根据相机标定结果把两部分运动结合起来,并利用两个相机的信息差异来帮助改善捕获结果:针对身体捕获的 Kinect[4]中含有彩色图像所不具有的真实的头部平移参数,而针对人脸捕获的彩色相机中具有 Kinect 所不具有的高分辨率人脸形状.通过将两者的信息相结合,可以得到更好的捕获结果.

实验结果表明,我们的方法能够同时捕获高质量的面部和人体运动,从而得到头部动作一致并且时间上对齐好的全身运动.同时,两个捕获目标互相引导的结果要明显好于两部分运动捕获分别进行的结果.通过结果展示和用户调研验证了这一点.

本文工作的主要贡献包括以下几点.

(1) 首次提出了一种同步捕获身体与面部运动及其动画的方法,并且动画结果经过了 user study 验证;

(2) 提出了一种通过赋予环境光照特征来对不同采集设备进行时间同步的方法;

(3) 提出了结合面部动画捕获设备和身体捕获设备各自的特点,帮助改善整体精度的方法.

1 相关工作

人体运动捕获和面部运动捕获都是虚拟人运动领域重要的研究分支,目前已有很多使用不同数据源、不同建模方式的运动捕获方法.这里,重点描述无标记点的运动捕获方法.

Shotton 等人[5]和 Girshick 等人[6]将人体姿态重建问题建模成一个回归问题,使用随机森林来进行回归.这类方法可以得到很好的身体运动重建结果.但是,为了同时捕获全身运动,Kinect 深度相机的距离必须比较远.这就导致了在 Kinect 采集到的图像中头部所占的像素数量很少,无论是深度图还是彩色图都很难分辨出头部的姿态和表情,造成捕获结果中头部基本是不动的.

Ye 等人[7]和 Su 等人[8,9]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.

Wang 等人[14]通过从若干帧深度图像中使用非刚体 ICP 进行对齐来得到人体模型,进而用来捕获人体运动.

Cao 等人[15,16]和 Wang 等人[17]利用彩色图像来重建得到的人脸运动.由于他们只使用彩色图像的信息,通过 2D 到 3D 的匹配算法来恢复 3D 信息,虽然能够得到视觉效果很好的结果,但是真实的深度信息是很难从彩色图像中完全恢复的,因此对头部的前后移动不可能完全真实地恢复出来.

Wang 等人[18]通过单帧图像得到了精细的人脸模型.但是,这样的系统在计算上比较复杂.同样地,由于只使用了彩色相机,这样的系统不可能获得很好的头部前后位置信息.

以上这些工作虽然在面部表情捕获以及身体运动捕获中取得了不错的效果,但是并没有考虑到利用两个相机各自的特点进行人脸和人体运动的同步捕获,如果应用在人体动画和人脸动画同时捕获系统中,会遗漏另一相机所带来的信息.据我们所知,本文是首次使用无标记点的方法将不同尺度的运动捕获结果结合起来,得到一个头部旋转角度一致、同步的身体加面部运动捕获结果.我们的算法流程图如图 1 所示.

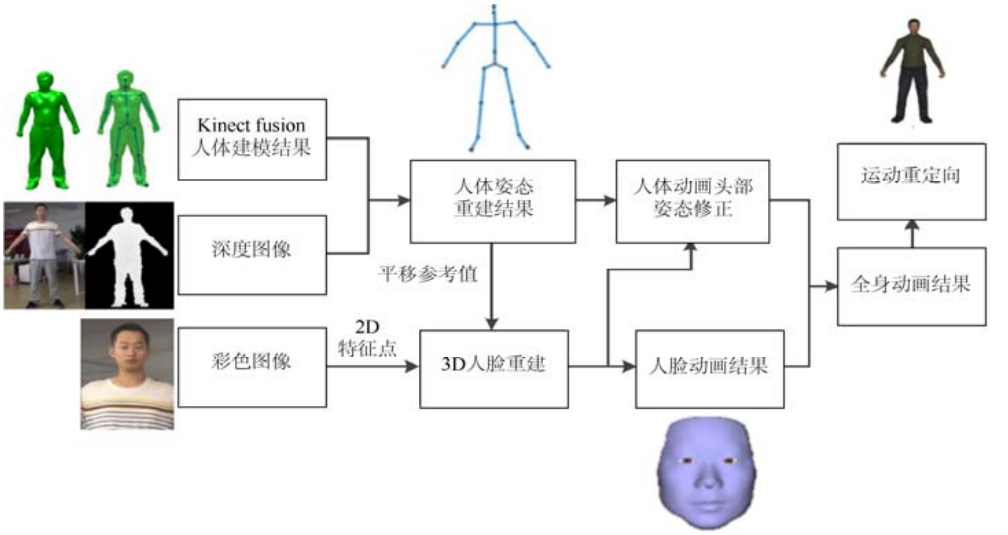

Fig.1 Pipline of our algorithm

图 1 算法流程图

对于人脸捕获部分,我们从 2D 图像中得到特征点,然后结合 Kinect 相机中得到的平移参考值解算头部姿态.对于人体捕获部分,首先使用 Kinect Fusion 扫描一个人的模型,接着手动对其嵌入骨骼,之后使用非刚体 ICP 进行肢体姿态恢复,然后把人脸捕获得到的头部姿态通过两个相机的相对矩阵传递给身体捕获空间,最后把求得的身体姿态和表情基权重重定向到另一个角色上渲染并加以展示.

2 相机的同步与标定

如前所述,为了能够同时采集人体和面部运动,我们使用两个相机分别捕获人脸和人体运动.采集设备如图 2 所示.

Fig.2 The capture equapments. Including a Kinect depth camera and a color camera of a smart phone

图 2 采集设备,包括一个 Kinect 深度相机和一个手机的彩色相机

2.1 彩色相机和Kinect深度相机同步

为了能够将捕获到的运动合在一起,我们要把两个相机的时间和空间信息对齐在一起.这一节主要描述相机的时间同步和空间标定.

为了使捕获到的面部和人体运动数据时间同步,必须要将两个相机拍摄的视频进行时间同步,否则,将使后续的过程中使用相同的人脸姿态参数对不同时刻的人脸表情进行求解,产生误差.这里,我们通过产生特定外部光照的方式来实现这一目标.在开始数据采集之后,首先在场景中进行几次闪光,每次闪光持续大约 0.1s,在两个相机的视频录制中会同时留下闪光的记录.我们通过在得到的视频中统计高亮像素数量来找到闪光出现的时间,从而在两个视频中分别找到对应的零时刻.由于我们可以预先分别得到两个相机的帧率,因此可以通过零时刻的对齐将两个视频进行时间对齐.实践中,两个相机的帧率都设置为 30fps.这样,通过首帧图片的对齐,就实现了整个视频序列的时间同步.在实验中,对齐结果的时间误差小于 1 帧.

2.2 彩色相机和Kinect深度相机标定

由于我们的 Kinect 深度相机和彩色相机的位置、角度、焦距都不相同,因此需要在使用中进行相机标定来得到两个相机的相对外参矩阵以完成后面的采集数据的融合.我们使用 Zhang 等人[19]的方法,用两个相机同时对棋盘格进行拍照,从而标定出两个相机各自的相机参数.将彩色相机和 Kinect 相机的外参标定结果分别记为 M_{color} 和 M_{depth},则 Kinect 深度相机与彩色相机的相对矩阵为

$$M_{color}^{depth} = M_{depth} M_{color}^{-1} \tag{1}$$

这个矩阵的意义是把采集面部运动的彩色相机空间中的运动数据变换到采集身体运动的 Kinect 深度相机空间的变换矩阵.利用这个相对变换矩阵,可以在分别捕获到人体运动和面部运动之后,把两部分数据整合在一起,从而得到完整的人体运动.

3 面部动画捕获

为了从彩色视频中得到高质量的面部动画捕获结果,我们使用基于多线性模型[15,20]的综合分析方法.首先从彩色图像中提取人脸特征点,然后利用多线性人脸模型进行人脸重建,最后求解人脸的姿态和表情基权重,得到人脸动画姿态和表情基权重.

3.1 2D特征点获取

我们使用类似于 Ren[21]的级联随机森林+线性回归的方式来得到脸部特征点位置.在级联的每一级中,我们从人脸图像中提取出灰度差特征训练随机森林,然后把随机森林的分类结果作为 0-1 特征,建立从 0-1 特征到形状变化的回归模型.实践中使用的是线性回归模型,数学描述为

$$\Delta S^t = w^t \Phi^t (I^t, S^{t-1}) \tag{2}$$

其中,w^t 是回归得到的线性模型的权重,Φ^t 表示随机森林图提取出的特征的函数.I 是长焦相机采集到的 2D 图像,S^{t-1} 是前一步的形状.整个过程的初始形状选择为训练数据库中所有形状的平均形状.通过这样的逐步迭代,可以得到最终的 2D 点形状 S^t.

3.2 3D人脸姿态、形状和表情重建

我们使用多线性模型来进行人脸形状的建模和重建,使用 facewarehouse[11]人脸数据库来进行人脸建模.我们使用多线性模型将人脸数据库组织成一个 3 阶张量,3 个维度分别对应于每个模型的顶点坐标(vertex)、人脸形状(identity)以及人脸表情(expression).通过张量分解[22],可以将数据进行压缩,同时,基本上不损失数据库的信息,形式为

$$T \approx C_r \times_2 U_{id} \times_3 U_{\exp} \tag{3}$$

这里,T 是原始的数据集,C_r 是经过压缩的核心张量,U_{id} 和 $U_{\exp}$ 分别是在后两个维度上的变换矩阵.$\times_2$ 和$\times_3$ 分别表示张量在第 2 个和第 3 个维度上与矩阵相乘.具体可以参考文献[22]等研究工作.实践中,T 的原始维度是

34530×150×47,C_r 的保留维度为 34530×50×25.

然后,对于任意人脸,我们可以将其表达为数据库中的人脸的线性插值,数学形式为

$$face = C_r \times_2 w_{id} \times_3 w_{\exp} \tag{4}$$

其中,w_{id} 和 $w_{\exp}$ 分别是人脸形状和人脸表情两个维度上的权重.得到的结果 $face$ 即为当前权重下重建得到的人脸的顶点坐标值.

与文献[11,12]等工作处理的从彩色图像重建人脸的问题不同,由于我们使用了 Kinect 深度相机来捕获人体运动,可以在深度数据的帮助下得到一个全局平移的参考值 T_g.这个全局平移参考值的获取方式如下:

$$T_g = (M_{color}^{depth})^{-1}\overline{P_{depth}} \tag{5}$$

其中,$\overline{P_{depth}}$ 是 Kinect 相机坐标系下深度图采集到的头部区域的中心位置.

我们使用特征点匹配的方法来求解人脸的姿态、形状和表情参数,将问题描述为一个优化问题,所使用的损失函数为

$$\underset{R,T,w_{id},w_{\exp}}{\arg\min}\ E = \sum_{i=1}^{n} \| fea_{2d} - \Pi(R(C_r \times_2 w_{id} \times_3 w_{\exp}) + T) \|^2 + \lambda_{trans} \| T - T_g \|^2 + \lambda_{id} w_{id}^T \sum_{id}^{-1} w_{id} + \lambda_{\exp} w_{\exp}^T \sum_{\exp}^{-1} w_{\exp} \tag{6}$$

其中,第 1 项是拟合的 2D 点与采集到的 2D 点的误差项,第 2 项意味着求解得到的平移项要与从深度图像得到的平移参考值尽可能地接近,第 3 项和第 4 项分别是形状和表情参数的先验项,fea_{2d} 是图像中的 2D 特征点,Π 是相机从 3D 空间到 2D 空间的投影矩阵.我们使用不考虑偏斜和畸变的针孔相机模型,则 Π 的形式为

$$\Pi = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{7}$$

R 和 T 是当前帧的旋转和平移参数,w_{id} 和 $w_{\exp}$ 是形状和表情的参数.Σ是数据库中参数的协方差矩阵.λ_{trans}、λ_{id}、$\lambda_{\exp}$ 分别是平移参考值、人脸形状以及人脸表情的正则项权重,实验中分别取为 1、0.001、0.001.

这个问题是一个标准的最小二乘问题.在求解中,我们把所有的参数分为 3 组:R 和 T 一起分为一组,W_{id} 和 $w_{\exp}$ 分别作为一组,使用迭代优化的方法来进行求解.在每一次迭代过程中,每次固定 3 组参数中的两组来优化剩余的一组.通过不断地迭代,可以逐渐逼近整个问题的最优解.实践中,我们发现,优化过程整体迭代 3 次就可以得到收敛的结果.

对于每一步的问题,我们都使用 levenberg marquardt 优化方法[23]来进行求解.

注意,我们只在第 1 帧优化求解人脸形状参数 w_{id}.在之后的运动捕获中,这一项参数始终保持不变.在视频追踪中,第 1 帧的 R 使用零值进行初始化,T 使用深度图像中得到的参考值来初始化,w_{id} 和 $w_{\exp}$ 使用先验的平均值进行初始化.后续的每一帧使用前一帧的优化结果作为初始值.

4 人体运动捕获

这一节给出人体运动获取方法.我们将从深度图像中重建人体运动的问题建模为一个反向运动学问题.为此,首先利用 Kinect fusion[24]来获得标准姿态的人体模型,然后为人体模型嵌入骨骼,最后用模型匹配采集到的深度图像来得到正确的姿态.

4.1 人体模型的获取

我们使用 Kinect fusion 来获得人体模型.通过将 Kinect 相机绕人体一圈,捕获到一个粗糙的人体扫描模型,然后手动地为得到的人体模型嵌入骨骼,结果如图 3 所示.得到骨骼模型和人体模型之后,使用线性骨骼蒙皮模型[25]建立起骨骼对蒙皮的驱动关系:

$$p' = \sum_{j=1}^{m} w_j(p) T_j p \tag{8}$$

其中,$w_j(p)$是第 j 个骨骼对顶点的控制权重,T_j 是第 j 个骨骼自身的局部矩阵.

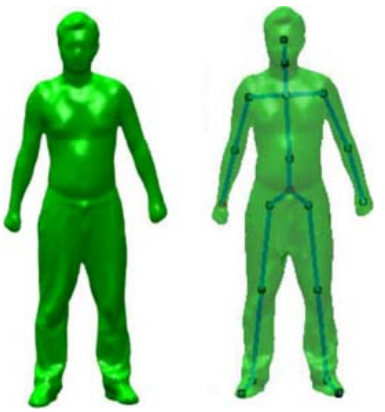

Fig.3 The human body model from Kinect fusion and embedded skeleton

图 3 Kinect fusion 得到的人体模型和嵌入的骨骼

4.2 基于Robust ICP对齐的肢体姿态估计

本节描述得到骨骼和蒙皮之后,根据深度图像求解当前姿态的方法.根据人体的网格模型和深度图像,我们把求解姿态问题建模为一个反向运动学问题,并用优化的方法来求解,目标函数为

$$\min_{q} = w_{icp}E_{icp} + w_{smooth}E_{smooth} + w_{sil}E_{sil} + w_{bound}E_{bound} + w_{prior}E_{prior} \tag{9}$$

其中,优化自变量 q 是运动的关节角和根节点位置参数.目标函数中第 1 项是运动数据的 icp 项,表示重建得到的蒙皮要尽可能地与采集到的深度数据一致.数学形式为

$$E_{icp} = \sum_{i=1}^{n} \| \vec{n} \cdot (p_i(q) - P_i) \|^2 \tag{10}$$

其中,$p_i(q)$是在姿态 q 的条件下,身体的第 i 个顶点根据前向运动学和线性混合蒙皮(LBS)得到的 3D 位置.P_i 则是与身体上的第 i 个顶点距离最近的深度点.$\vec{n}$ 是相应的网格上的点的法线方向.

第 2 项是运动平滑项,用来防止求解得到的姿态发生突变.

$$E_{smooth} = \| q_{i-2} - 2q_{i-1} + q \|^2 \tag{11}$$

其中,q_{i-2} 和 q_{i-1} 分别代表前两帧和前一帧的关节角度.

第 3 项是轮廓项,表示得到的蒙皮的轮廓要尽可能地与采集到的深度图像的轮廓一致.

$$E_{sil} = \sum_{i=1}^{m} \| p_{2d}(i) - p_{sil}(i) \|^2 \tag{12}$$

其中,$p_{2d}(i)$是模型上的第 i 个轮廓点在相机平面的投影,$p_{sil}(i)$是深度图中距离其最近的轮廓点 2D 坐标.这一项用来进行深度图和人体模型的轮廓对齐,保证重建得到的姿态和采集到的深度图像之间的轮廓对应关系.

第 4 项是关节角边界项,对应于求解得到的关节角要处于合理范围内.

$$E_{bound} = \begin{cases} 0, & q_{\min} \leqslant q \leqslant q_{\max} \\ 1, & \text{else} \end{cases} \tag{13}$$

最后一项是运动姿态先验项,主要用来保证没有深度点云约束的关节部分满足运动数据的统计规律,从而避免出现不合理的运动姿态.对这一项,我们首先把 CMU 数据库中关节角的分布建模为高斯分布,并统计整个分布的均值 q_{mean} 和方差Σ,然后对当前姿态 q,计算 q 在这个高斯分布中的概率,将概率的负对数作为运动先验项.这一项的数学形式可以表示为 q 与 q_{mean} 的马氏距离(Mahalanobis distance).

$$E_{prior} = (q - q_{mean})^T \sum^{-1} (q - q_{mean}) \tag{14}$$

目标函数中的 w 是各项目标函数的权重.实验中依次取为 1、0.01、3、20、1.

目标函数的每一项都是最小二乘形式,我们可以使用高斯牛顿法来求解.在求解中,每次迭代都需要重新为

ICP 项和轮廓项找到最近点.这样,通过迭代最终得到一个匹配结果比较好的姿态 q.

4.3 头部旋转与全身运动的结合

由于深度图像分辨率比较低,只能用来求解大尺度的人体姿态.而头部的姿态在深度图中的特征不够明显,只使用深度图来进行姿态重建无法得到好的结果.如图 6 中第 2 列所示,重建得到的身体运动在头部几乎是不动的,需要使用人脸运动捕获的结果来帮助修正头部姿态.

由于我们已经提前标定了相机相对旋转矩阵 R_{color}^{depth},利用人脸运动捕获结果重建得到的头部姿态为

$$R_{depth}^{t} = R_{color}^{depth} R_{color}^{t} \tag{15}$$

这样,我们就将人脸与人体的运动结合到了一起.这里,虽然头部运动求解可以同时得到一个平移参数,但是人的骨骼模型中头部没有平移自由度,强行把平移参数加入会造成模型穿刺甚至崩溃,因此不使用面部运动追踪得到的平移参数.

得到运动捕获结果之后,需要把运动数据和表情参数重定向到虚拟人的身体和脸上.由于用来进行运动采集的模型的骨架结构和表情基的组成方法都是与虚拟人相同的,因此只需把追踪得到的关节角和表情基权重直接赋予虚拟人就可以得到运动重定向的结果,之后即可用来驱动虚拟人来得到动画结果.

5 实验结果

本文使用低成本的深度和彩色相机来捕获全身运动.为了验证全身运动跟踪算法的有效性,我们在实际数据上对算法进行了测试.Kinect 深度图像和彩色图像的分辨率分别设置为 320×240 以及 640×480.经过时间同步的 Kinect 数据和彩色相机如图 4 所示.

(a) 两个相机的捕获结果对比 (b) Kinect 捕获到的彩色图像头部放大的结果

Fig.4 The captured images

图 4 相机捕获结果示例

可以看到,Kinect 采集到的图片头部区域像素数量较少,很难辨识头部姿态和表情.虽然可以看到 Kinect 得到的图像能够捕获到整个身体,但是面部非常模糊,无法识别出人脸的姿态和表情,需要长焦的彩色相机帮助捕获人脸.

同时,为了说明全身运动捕获能够得到更加自然、真实的虚拟人动画,我们在视觉上对比了全身动画和只跟踪人脸或者肢体单个部分产生的动画.实验结果表明,本文的方法能够得到更加真实的虚拟人动画.

采用本文方法,人脸 2D 特征点加上 3D 重建时间为每帧 20ms,使用的处理器是 Intel i7-6700.目前,我们的方法在身体重建部分没有进行速度优化,在 matlab 中速度为每秒 3 帧.

5.1 两个相机的运动捕获结果

为了验证本文算法的有效性,我们首先用 Kinect fusion 扫描一个人的模型,然后使用两个经过同步和标定的相机采集了表演者大约 2 分钟的运动,之后按照本文的方法先分别进行运动合成,最后再将两部分结合在一起.表演者主要在模拟和别人谈话时候的一些姿态和人脸表情.我们从 3 个角度展示了合成动画的渲染结果,如图 5 所示.

可以看到,合成的动画同时捕获到了身体与面部的动作,包括肢体动作、头部姿态、面部表情,由此得到了更高真实感的人体动画捕获结果.由图 5 可以看到,我们的方法利用彩色相机和 Kinect 深度相机,真实地重建出

了人脸和人体运动.

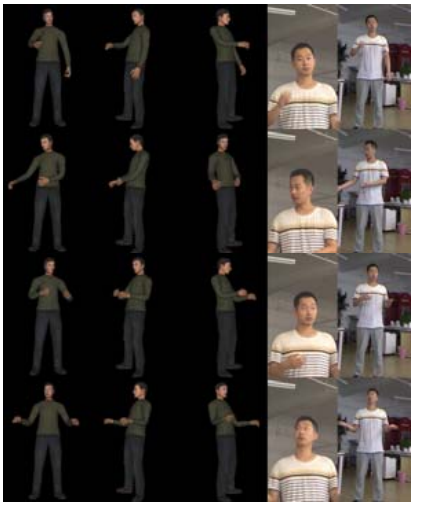

Fig.5 The rendered motion and the video input of the Kinect and color camera

图 5 从 3 个不同角度渲染的运动结果与两个相机的视频输入的对比

5.2 虚拟人动画真实感验证

为了验证身体和人脸捕获结合的方法能够得到更佳真实感的动画结果,本节将采用我们的方法后全身动画驱动结果与单独的人脸动画以及身体动画的结果进行对比.

从图 6 结果可以看出,同时捕获身体运动与面部运动的结果是真实、自然的,而仅仅捕获其中任意一个的结果都会感觉不够自然.此外,只做身体运动捕获的结果显示,头部基本是不动的,通过面部运动捕获补充得到了人的头部姿态.每一行的第 1 格是人脸、人体同步捕获的结果,第 2 格是单独捕获肢体动作的结果,第 3 格是单独捕获人脸动作的结果,第 4 格和第 5 格分别是长焦彩色相机与 Kinect 的彩色相机的视频输入.

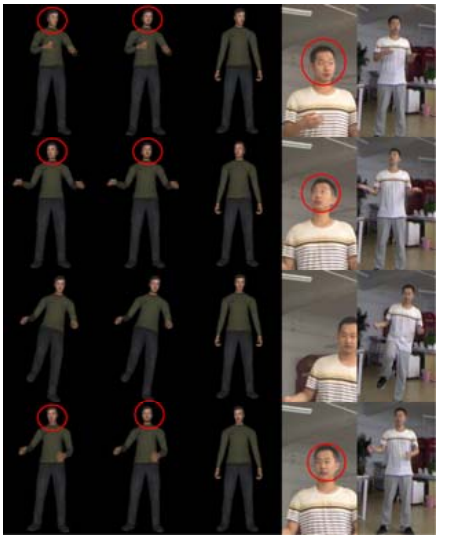

Fig.6 Comparison of separated body and face capture and our synchronized motion capture

图 6 人体、人脸分开进行捕获与同步捕获的对比

5.3 运动质量的用户调研评价

为了更好地评价所捕获到的运动质量,我们通过用户调研得到定量的运动质量评分.我们把合成的运动渲染并录制成视频,然后让 18 个用户对于全身动画、仅肢体动画以及仅人脸动画进行真实感打分(1~5 分制),1 分代表最不真实,5 分代表最真实.其中,4 个用户对动画领域比较熟悉.

图 7 显示了分数的均值和方差.从给出的数据可以看到,身体和人脸动画同步捕获的结果要明显好于仅进行人脸动画或者仅进行肢体动画捕获的结果.这也说明,我们的人体人脸同步捕获技术能够得到更真实的运动.图 7 中显示了 3 种不同运动捕获结果的用户评价,5 分对应最真实,1 分对应最不真实.

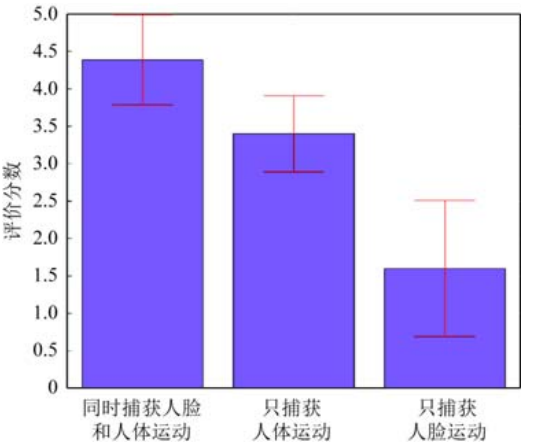

Fig.7 Result of user study

图 7 用户调研的结果

同时,我们对用户调研结果用 t 检验(t-Test)进行了显著性分析[26],见表 1.

Table 1 Result of user study

表 1 用户调研结果

名称	结果(a)	结果(b)	结果(c)
分数	4.39±0.60	3.40±0.51	1.60±0.91
名称	(a)与(b)的 P 值		(a)与(c)的 P 值
数值	0.000 116		<0.0001

结果显示,$P<0.05$,说明同时捕获两种运动的结果与分别进行运动捕获的结果在用户打分上存在统计学差异.也就是说,使用两个相机进行运动捕获的整体结果要好于分开的结果.

6 结 论

本文提出了一种一个 Kinect 深度相机和一个长焦彩色相机同步捕获身体与面部运动的方法.该方法具有成本低、易于架设等特点.该方法通过相机的同步和标定把两个相机捕获到的动画结合起来,得到完整的全身运动.此外,两个相机采集到的信息还可以互相补充,改善运动重建结果.通过捕获表演者的运动的方法验证了这一算法的有效性,并与单独进行身体运动以及人脸运动捕获的结果进行了对比.分别通过图片对比和用户调研验证了我们的同步捕获方法所得到的动画结果更加真实、自然.

现有方法和系统的缺点之一是长焦彩色相机的视野较小,这限制了表演者可以使用的空间大小,不太适合需要大范围运动的场景.一种解决方法是使用多个彩色相机,当人脸位置发生变化时,使用相应相机捕获不同空间位置的人脸.

另一个缺点是当前的方法只具有人脸和人体的运动捕获,距离完整的虚拟人捕获系统还缺少手部的捕获.

这也是未来需要进一步提高效果的方向之一.一种可能的方式是加入几个不同角度的 Kinect 相机来进行手部运动的捕获.

此外,如果使用一个类似于 SMPL[27]的人体参数化模型,可以根据人体图像直接得到一个人的模型,例如文献[28],这样得到的人体模型与现在的方法相比,既可以省去 Kinect fusion 和手动嵌入骨骼的工作,又可以得到更精细的人体模型.

在未来的工作中,我们可以把长焦彩色相机的结果与带 marker 的运动捕获系统(如 Vicon Cara)进行对比.此外,还可以使用人机交互中的方法(如显著性分析),更加详细地对结果进行评价.

References:

[1] Xia SH, Gao L, Lai YK, Yuan MZ, Chai JX. A survey on human performance capture and animation. Journal of Computer Science and Technology, 2017,32(3):536–554.

[2] Zelezny M, Krnoul Z, Jedlicka P, *et al*. Analysis of facial motion capture data for visual speech synthesis. In: Proc. of the Int'l Conf. on Speech and Computer. 2015. 81–88.

[3] Xia SH, Su L, Fei XY, Wang H. Toward accurate realtime marker labeling for live optical motion capture. Visual Computer, 2017,33(6-8):993–1003.

[4] Zhang Z. Microsoft Kinect sensor and its effect. IEEE Multimedia, 2012,19(2):4–10.

[5] 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.

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

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

[8] Su L, Chai JX, Xia SH. Local pose prior based 3D human motion capture from depth local pose prior based 3D human motion capture from depth. Ruan Jian Xue Bao/Journal of Software, 2016,27(S2):172−183 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16032.htm

[9] Xia SH, Zhang ZH, Su L. Cascaded 3D full-body pose regression from single depth image at 100 fps. In: Proc. of the IEEE VR. 2018.

[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 Auto-nomous 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 Graphics, 2012,31(6):188:1–188:12.

[14] Wang KK, Zhang GF, Xia SH. Templateless non-rigid reconstruction and motion tracking with a single RGB-D camera. IEEE Trans. on Image Processing, 2017,26(12):5966–5979.

[15] Cao C, Weng Y, Zhou S, Tong Y, Zhou K. Facewarehouse: A 3d facial expression database for visual computing. IEEE Trans. on Visualization and Computer Graphics, 2013,20(3):413–425.

[16] Cao C, Hou Q, Zhou K. Displaced dynamic expression regression for realtime facial tracking and animation. ACM Trans. on Graphics, 2014,33(4).

[17] Wang CY, Shi FH, Xia SH, Chai JX. Realtime 3D eye gaze animation using a single RGB camera. ACM Trans. on Graphics (TOG)—Proc. of the ACM SIGGRAPH, 2016,35(4):118:1–118:14.

[18] Wang H, Xia SH. Construction facial shape with details single image. Journal of Computer-Aided Design & Computer Graphics, 2016,29(7):1256–1266 (in Chinese with English abstract).

[19] Zhang Z. A flexible new technique for camera calibration. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000, 22(11):1330–1334.

[20] Vlasic D, Brand M, Pfister H. Face transfer with multilinear models. ACM Trans. on Graphics, 2005,24(3):426–433.

[21] Ren S, Cao X, Wei Y. Face alignment at 3000 fps via regressing local binary features. In: Proc. of the Computer Vision and Pattern Recognition. 2014.

[22] Kolda TG, Bader BW. Tensor decompositions and applications. SIAM Review, 2009,51(3):455–500.

[23] Lourakis MIA. levmar: Leven-berg-Marquardt nonlinear least squares al-gorithms in C/C++. 2004. http://www.ics.forth.gr/~lourakis/levmar/

[24] Newcombe RA, Izadi S, Hilliges O, *et al*. KinectFusion: Real-time dense surface mapping and tracking. In: Proc. of the 10th IEEE Int'l Symp. on Mixed and Augmented Reality (ISMAR). IEEE, 2011. 127–136.

[25] Lewis JP, Cordner M, Fong N. Pose space deformations: A uni-fied approach to shape interpolation and skeleton-driven deformation. In: Proc. of the ACM SIGGRAPH 2000, Annual Conf. Series, ACM SIGGRAPH. 2000.

[26] Box JF. Guinness, Gosset, Fisher, and Small Samples. Statistical Science, 1987.

[27] Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ. SMPL: A skinned multi-person linear model. ACM Trans. on Graphics (TOG)—Proc. of the ACM SIGGRAPH Asia, 2015,34(6):248:1–248:16.

[28] Bogo F, Kanazawa A, Lassner C, Gehler P, Romero J, Black MJ. Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. arXiv:1607.08128, 2016. https://arxiv.org/abs/1607.08128 [doi: 10.1007/978-3-319-46454-1_34]

附中文参考文献:

[8] 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(S2):172–183. http://www.jos.org.cn/1000-9825/16032.htm

[18] 王涵,夏时洪.单张图片自动重建带几何细节的人脸形状.计算机辅助设计与图形学学报,2016,29(7):1256–1266.

王志勇(1989－),男,天津人,学士,主要研究领域为计算机图形学,虚拟人动画.

袁铭泽(1988－),男,硕士,主要研究领域为计算机图形学,虚拟人动画,人工智能.

王从艺(1990－),男,博士,主要研究领域为计算机图形学,虚拟人动画,人工智能.

夏时洪(1990－),男,博士,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,三维深度学习,数字几何处理,物理建模,动力学控制,运动分析与理解.

张子豪(1993－),男,学士,主要研究领域为计算机图形学,虚拟人动画,人工智能.