引用格式：黄一鸣，潘达，张宜春．基于人体姿态估计的京剧虚拟人物互动系统[J]．中国传媒大学学报（自然科学版），2022，29(06)：43-49.
文章编号：1673-4793(2022)06-0043-07

基于人体姿态估计的京剧虚拟人物互动系统

张宜春[1*]，黄一鸣[2]，潘达[2]
（1. 中国艺术科技研究所，北京 100007；2. 中国传媒大学媒体融合与传播国家重点实验室，北京 100024）

摘要：提出了采用基于计算机视觉的人体动作捕捉技术构建京剧虚拟人物互动系统，首先通过单目相机捕捉京剧表演者的运动画面，经由深度学习姿态估计模型预测24个人体骨骼关键点，再经由滤波处理模块对这些关键点进行平滑操作得到准确的运动信息，最后导入游戏开发引擎Unity3D中，通过将真实人体运动结点和虚拟人物关节点进行绑定，实现驱动虚拟人物仿照真人动作进行京剧表演的舞台效果。实验表明，本互动系统可以实时且准确地估计人体姿态，并能够实现与虚拟人物较好的互动体验感。

关键词：姿态估计；动作捕捉；单目相机；Unity3D；京剧
中图分类号：TP 391　**文献标识码**：A

Virtual character driving based on pose estimation technology −a case study of Beijing Opera

ZHANG Yichun[1*], HUANG Yiming[2], PAN Da[2]
(1. China institute of Arts Science and Technology, Beijing 100007, China; 2. School of Information and Communication Engineering, Communication University of China, Beijing 100024, China)

Abstract: This paper proposes to use the human motion capture technology of computer vision to build a Peking Opera virtual character interaction system. Firstly, the motion pictures of Peking opera performers are captured by monocular camera, and 24 human skeleton key points are predicted by deep learning posture estimation model. Then these key points are smoothed through the filter processing module to obtain accurate motion information. Finally, it is imported into the game development engine unity3d. By binding the real human motion nodes with the virtual character′s joint points, the stage effect of driving the virtual Peking opera characters to imitate the real action for Peking opera performance is realized. Experiments show that our proposed interactive system can estimate human posture in real time and accurately, and achieve a better sense of interactive experience with virtual characters.

Keywords: pose estimation; motion capture; monocular camera; Unity3D; Beijing Opera

1　研究背景

人体姿态估计（Human Pose Estimation, HPE）是计算机视觉领域的高级任务之一，它通过对计算机输入一幅含有人体的图像或者一段视频，从而使人们获得图像或视频中人体骨架关键点位置[1]。人体姿态估计在现实生活中的应用十分广泛，它可以应用于多个领域，例如：动作识别、姿态跟踪以及基于计算机视觉的人体动作捕捉等。

作者简介（*为通讯作者）：张宜春（1978-），中国艺术科技研究所，副研究员，主要研究领域为文化与科技融合，Email：zhangyichun@vip.sina.com；黄一鸣（1996－），男，硕士研究生，主要研究方向为计算机视觉、人体姿态估计。Email：huangyiming@cuc.edu.cn；潘达（1989－），男，讲师，博士，主要研究方向为图形图像处理、计算机视觉等。Email：pdmeng@cuc.edu.cn

人体动作捕捉技术,简称动捕(Mocap),意为记录并处理人的动作行为的技术。目前人体动作捕捉技术应用广泛,例如在体育、娱乐、医疗健康等领域动捕技术都有所涉及。根据使用设备与实现技术的不同,主流的人体动作捕捉技术可分为以下三种类型——光学动捕、惯性动捕和视觉动捕。光学动捕需要较大捕捉空间且设备昂贵、部署繁琐,但是其优点是精度极高,发展成熟。惯性动捕由惯性传感器组成,优点是使用灵活、性价比较高、便携性强;缺点是易受干扰,长时间使用易产生估计偏移量,不够精确[2]。视觉动捕是一种仅需要摄像头和计算机便可实现的动捕方法,典型实现方法是使用Kinect深度相机进行人体捕捉[3]。目前只使用普通摄像头实现动作捕捉的技术也日益发展,这使得视觉动捕成本更加低廉、部署更加便捷、前景更加广阔。

本文主要研究基于姿态估计技术的视觉动捕技术。目前人体姿态估计技术根据输出结果的维度划分,可以分成二维人体姿态估计和三维人体姿态估计。

1.1 二维人体姿态估计

二维人体姿态估计,即给计算机输入图像信息后可以得到图像中人体关键点的二维预测坐标。二维人体姿态估计根据同一幅图中估计出的最多人数,进一步划分为单人二维姿态估计和多人二维姿态估计;二维人体姿态估计根据检测关键点的方式,又分为自顶向下(Top-Down)以及自底向上(Bottom-Up)的方法。自底向上的方法是先通过人体检测器判断出人体检测框,然后通过多次剪裁修正,最后对每个检测框进行人体关键点识别,这种方法的精度往往较高,但是识别的速度较慢。自底向上的方法则是首先根据人体关键点热力图预测人体关键点可能存在的区域,再利用诸如匈牙利算法等方法去将关键点组合成一个完整的人体骨架,这种方法的实时性较优,但是难以达到较高的检测精度,应用场景受到一定的限制。

1.2 三维人体姿态估计

由于二维人体姿态估计无法预测出人体深度信息,因此,越来越多的研究者着眼于三维人体姿态估计。三维人体姿态估计旨在通过输入的二维图像或视频,寻求三维空间上的人体关节点的位置信息。随着深度学习技术的发展以及人们对人机交互要求的提高,三维人体姿态估计越来越成为了姿态估计领域研究者热衷的课题。在三维姿态估计中,人体的表示形式一般分为两种:骨架图表示和参数化表示。

第一种表示方式是Skeleton方式,这种表示方式如图1所示,即通过人体关键点以及相邻关键点之间的连线组成相应的人体姿态[4]。

第二种表示方式是参数化的人体模型(如SMPL),这种表示方法使用身形参数、姿态参数来控制mesh网格表示人体姿态、高矮、胖瘦等信息,多用于人体三维重建。

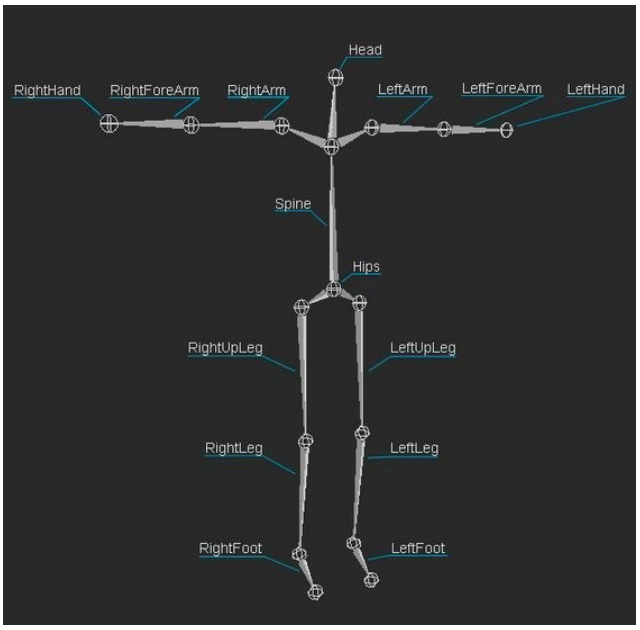

图1 姿态估计人体三维骨架图

由于三维人体姿态估计能够预测图像中人的深度信息,因此,三维人体姿态估计具有更加广泛的应用前景。

基于三维人体姿态估计技术,本文研发了基于人体姿态估计的京剧虚拟人物互动系统,使用该系统可以通过单目摄像机实现京剧艺术家表演动作的实时捕捉并将之迁移到虚拟的人物身上。

2 虚拟人物互动系统

本章主要介绍虚拟人物互动系统的四大模块,即姿态估计方法模块、虚拟人物驱动模块、卡尔曼滤波模块和视觉动捕设计模块。基于人体姿态估计的京剧虚拟人物互动系统框图见图2。其中,虚拟人物骨骼绑定以及关键点角度限制均属于虚拟人物驱动模块。

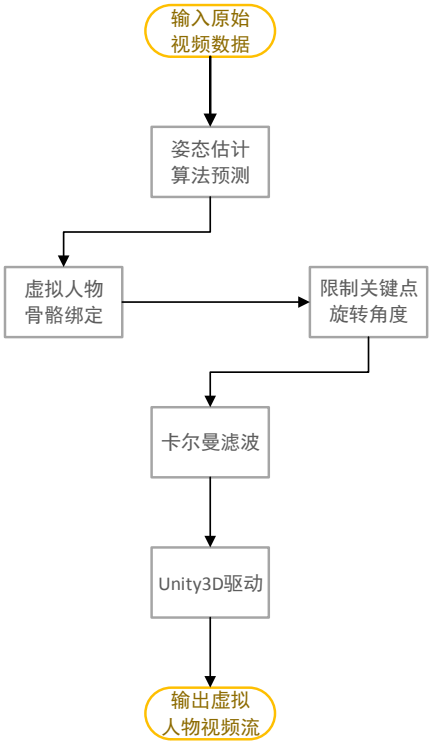

图2 虚拟人物互动系统流程图

2.1 姿态估计方法

VNect是一种针对单人且实时性较好的三维人体姿态估计方法，该算法能够通过摄像头视频流捕获人体信息，且人体关键点的预测精度较高，并实时预测出人体的三维姿态[5]。因此，本文使用VNect姿态估计算法，将之应用在人体视觉动捕和虚拟人物驱动上。此外，出于对非物质文化遗产——京剧的发扬和保护，本文使用时下流行的游戏开发引擎Unity3D，实现了对京剧艺术家动作的捕捉，并将京剧艺术家的动作迁移到了Unity3D的虚拟人物形象上，使得虚拟人物也可以做出京剧的动作，有助于增强人们对于京剧的兴趣，对传统京剧进行“新创造”。

VNect模型较好地解决了三维姿态估计中的空间歧义性问题，即解决了二维人体姿态估计结果提升到三维空间时，一个二维姿态会对应多个三维姿态的问题。VNect通过图像特征直接进行学习、提取三维隐含特征，不再分阶段提取，而是联合训练人体的二维姿态和三维姿态。

VNect模型的算法流程如图3所示，整个流程分成多阶段，依次为人体边界框提取，CNN回归，时域滤波，以及骨骼绑定，最终可以实现三维姿态估计的效果。其中最重要的就是CNN回归，用于回归人体关键点坐标和人体关键点热力图。

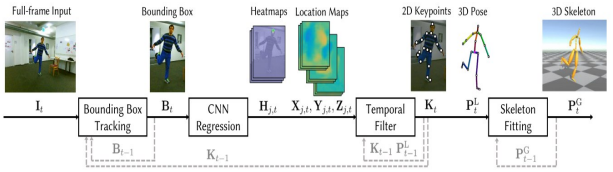

图3 VNect模型流程图

VNect算法的目的是能够实现高质量且快速的三维人体姿态估计。因此，VNect采用了联合训练的方式，计算出每个网络的热力图以及回归出每个关节点相对于根节点的三个方向的位移。卷积神经网络回归部分的结构如图4所示：

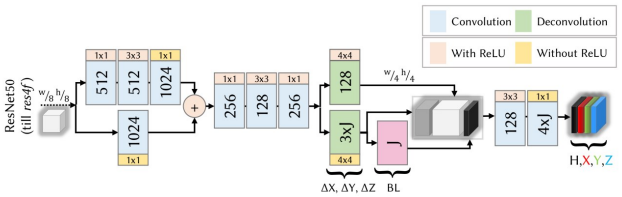

图4 VNect模型结构图

VNect的主干网络是ResNet50，仅使用到Conv4的最后一层res4f，到res4f开始进行结构修改，将连续多帧单目RGB图像（可看成视频流）作为网络输入，输出则是预测每个关键点二维坐标点热力图H和三维相对于根节点的坐标XYZ，从而以向量的形式输出。

VNect的损失函数通过关键点区域$x_j y_j z_j$和ground truth的差异进行L2范数计算：

$$Loss(x_j) = \left\| H_j^{GT} \cdot (X_j - X_j^{GT}) \right\|_2 \quad (1)$$

其中，j表示关节点的索引值，X_j表示第j个关键点在X方向的位移，X_j^{GT}是关键点的ground truth，$\cdot$代表Hadamard积。

此外，对于VNect的网络训练超参数等条件的设置同VNect保持一致。

2.2 虚拟人物驱动

虚拟人物驱动，即使用三维姿态估计技术预测人体关键点的运动信息，并将关键点信息迁移到已经绑定骨骼的三维虚拟人物形象上，读取运动数据并使用姿态估计算法回归出的运动数据驱动虚拟人物，使之进行运动。

2.2.1 blender人物模型骨骼绑定

虚拟人物实时运动需要对人物模型进行骨骼绑定，然而各个关节点的旋转限制角度不尽相同，为了使得虚拟人物的肢体动作更加真实且自然，除了使用已预测的人体关键点数据之外，还需要对虚拟人物不同关键点的旋转角度进行设置。虚拟人物关键点旋转角度设置主要依靠制作人物模型的软件进行骨骼限制，不同关键点

的旋转角度的设置详情见表1。

表1 人体主要关节点旋转角度限制

关节名	旋转角度限制
肩关节	0到180度
膝关节	0到140度
踝关节	0到45度
肘关节	0到150度
腕关节	0到80度
髋关节	0到45度

本文人物模型绑定使用建模软件blender，人物模型绑定的流程见图5。其中，人物模型绑定需要用到正向运动学(Forward Kinematics, FK)和反向运动学(Inverse Kinematics, IK)[6]。正向运动学的原理是父骨骼带动子骨骼进行运动，例如膝关节带动踝关节运动，可以实现腿部弯曲的效果。反向运动学的原理是子骨骼带动父骨骼进行运动，例如手部带动肩部进行挥手运动等。

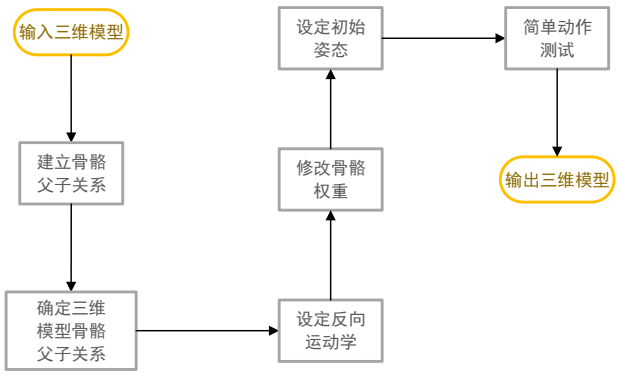

图5 骨骼绑定流程图

2.2.2 Unity3D游戏开发引擎

本文选用的虚拟人物驱动引擎是Unity3D。Unity3D是由Unity Technologies研发的3D游戏引擎，Unity3D开发功能强大，兼容性好，使之受到越来越多的开发者青睐[7]。Unity3D支持全平台，应用广泛且具有逼真的物理模拟系统，因此本文选择Unity3D作为基于人体姿态估计的京剧虚拟人物演艺系统的开发平台。

2.2.3 开放神经网络交换格式

本文使用开放神经网络交换格式(Open Neural Network Exchange, ONNX)存储深度学习模型并将之迁移到Unity3D中[8]，实现运动信息的传递。

2.2.4 骨骼数据处理与表示

基于VNect的姿态估计算法得出的人体关键点数量和Unity3D中虚拟人物驱动的关键点数量是不相同的。因此，需要对骨骼数据进行处理才可以让深度学习网络得到的人体数据应用于Unity3D中。

本文使用三维建模软件blender绑定虚拟人物的骨骼，通过增加子节点关节来提升虚拟人物运动的真实性(例如不止关注于手部位置，还关注于指关节位置)。VNect三维姿态估计算法所计算出的骨骼关键点热力图都是人体的重要关键点，算法识别到的仅为骨骼两端节点的坐标，在三维旋转角度方面没有涉及，因此卷积神经网络没有涉及到的关键点只能通过临近涉及到的关键点进行定位。神经网络预测出的二维关键点表示见图6。与关键点图相对应的关键点名称见表2。

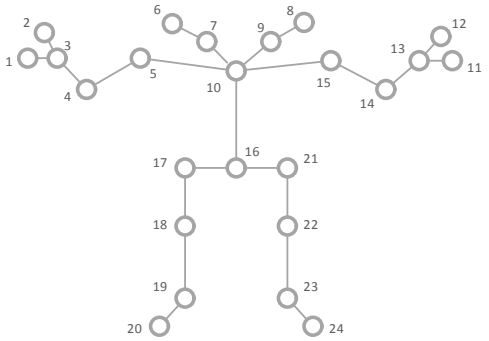

图6 神经网络预测骨架图

表2 预测关键点编号及名称

关节编号	关节名称
1	右手中指
2	右手拇指
3	右手手掌
4	右肘
5	右肩
6	右耳
7	右眼
8	左耳
9	左眼
10	鼻子
11	左手中指
12	左手拇指
13	左手手掌
14	左肘
15	左肩
16	腹部
17	右胯
18	右膝
19	右脚
20	右脚尖
21	左胯
22	左膝
23	左脚
24	左脚尖

在游戏开发引擎Unity3D中，骨骼转动有多种表示方法，例如通过方向余弦、欧拉角、四元数表示等。其中，四元数是使用频率最高的一种表示方法，四元数计算高效，可以表示状态和旋转动作，且能通过欧拉角转换。因此，本文选用Unity3D游戏开发中最常用的四元数来表示。

在完成人物模型的骨骼绑定、开发平台的选择、深度学习模型的迁移以及三维骨骼数据的四元数表示后，基本可以实现虚拟人物的运动，但是此时由于可能会有一些帧没有捕捉到所有的人体关键点、背景环境的干扰以及遮挡难题的存在，会对连续帧的姿态估计造成干扰，因而并不会令每一帧图像都能预测出平滑的运动表示。如果没有人为处理，神经网络预测出的骨骼运动就会发生闪跳现象，因此需要后续的进一步处理。

2.3 卡尔曼滤波

卡尔曼滤波器（Kalman filter）是一种常用的连续信号跟踪滤波器，且卡尔曼滤波器是一种纯粹的时域滤波器，无需在频域设计后再在时域实现，因此，该滤波器占用的内存很小，适用于视觉动作捕捉的平滑处理[9]。

2.3.1 卡尔曼滤波器原理

卡尔曼滤波器分为非线性方程直接法和线性方程间接法。人体姿态可视为域内线性运动，因此，选用线性卡尔曼滤波器作为运动平滑的优化器。

线性卡尔曼滤波器公式为：

$$\hat{x}_k = \hat{x}_{k-1} + K_k(\hat{z}_k - \hat{x}_{k-1}) \tag{2}$$

其中，$\hat{x}_k$为当前的估计值，$\hat{x}_{k-1}$为上一状态测量值，$\hat{z}_k$为测量值，K_k为卡尔曼增益，卡尔曼增益的表达式如下：

$$K_k = \frac{e_{estimate_{k-1}}}{e_{estimate_{k-1}} + e_{estimatemt_{k-1}}} \tag{3}$$

其中，$e_{estimate_{k-1}}$为估计误差，$e_{estimatemt_{k-1}}$为测量误差。当估计误差大于测量误差时卡尔曼增益趋近1，估计值$\hat{x}_k$同测量值$\hat{z}_k$近似；当测量误差大于估计误差时卡尔曼增益趋近0，估计值$\hat{x}_k$同上一时刻测量值$\hat{z}_{k-1}$近似。

2.3.2 卡尔曼滤波器在动捕方面的应用

在虚拟人物驱动过程中，神经网络预测出的人体关键点跳闪现象通过卡尔曼滤波得到缓解，针对人体运动数据在各个坐标轴上的规律，通过卡尔曼滤波进行处理，为视觉动捕系统开发的算法性能稳定性提供了保障。

2.4 视觉动捕系统设计

基于人体姿态估计的京剧虚拟人物互动系统除了由姿态估计算法模块、虚拟人物驱动模块、卡尔曼滤波模块组成外，还有一个重要组成模块即为视觉动捕系统模块。以下将主要介绍本文研发的视觉动捕系统的实现功能、系统界面UI设计等方面。

2.4.1 动捕系统实现的功能

本文基于人体姿态估计技术研发出的视觉动捕系统同传统京剧的结合可以实现京剧表演动作的“迁移”，即将京剧艺术家展示的京剧动作进行捕捉，并“迁移”给虚拟人物，从而实现姿态估计和京剧表演的有机结合。Unity3D游戏引擎支持计算机摄像头调用[10]。本文采用的姿态估计算法模型可以实现使用计算机摄像头对于三维人体关键点的捕获。通过摄像头进行的虚拟人物驱动见图7。

图7 摄像头实时的虚拟人物驱动

2.4.2 动捕系统的界面设计

在视觉动捕系统的UI设计中，本工程采用了模块化的设计思想，动捕系统主界面见图8所示。

图8 摄像头实时的虚拟人物驱动

主要分为三部分——舞台、虚拟人物、视频窗口以及选项框。其中，舞台选用了传统的京剧舞台图片，人物模型采用的是NicoNico网站中的人物形象，选项框分为两大部分，依次是视频源和人物模型，视频源可以选取网上或电脑本地的京剧动作视频，或者通过同计算机连接的摄像头捕获的视频数据，对之进行读取并驱动。

3 实验与分析

以下为对本文提出的基于姿态估计算法的视觉动作捕捉系统的性能进行实验分析。

3.1 虚拟人物互动系统实验开发环境

基于人体姿态估计的京剧虚拟人物互动系统开发所用相关工具、使用的编程语言等条件见表3。

表3 动捕系统开发环境表

开发项	开发工具
操作系统	Windows10
开发平台	Unity3D、Pycharm
开发语言	C#、Python
开发建模	Blender、3DMax
CPU型号	I7-7800X
显卡配置	RTX 3060

3.2 视觉动捕系统精度评价实验

精度评价实验是视觉动捕系统中最重要的实验,它主要的考虑因素是姿态估计算法的精确性,姿态估计算法的精确性越高,则视觉动捕系统的可靠性就越高。

3.2.1 评价指标

姿态估计领域的精度实验最常见的姿态估计算法性能评价指标就是平均关节位置误差(Mean Per Joint Position Error, MPJPE)[11]。平均关节位置误差是预测关键点同ground truth之间的平均欧氏距离,单位是毫米。MPJPE的公式如下:

$$err_j = \frac{\sum_i (\| p_{ij} - p_{ij}^{gt} \|)}{N} \tag{4}$$

其中,N代表骨骼图中关键点的数量,p_{ij}为预测出的关键点位置,p_{ij}^{gt}表示标签中关键点的实际位置。

3.2.2 数据集

本文精度实验选用的是Human 3.6M数据集。Human3.6M数据集中含有360万三维人体姿势及对应图像。该数据集由11个实验者的动作组成,依次存放于S1-S11文件夹。本文选用S1、S5、S6、S7、S8作为训练集,S9和S11作为测试集,且主要选用测试集中的Walk、Smoke、Wait作为测试姿态。三维姿态估计算法精度实验的结果见表4。

表4 三维姿态估计算法精度比较表

方法	Walk	Smoke	Wait
DconvMP	77.6	89.5	86.5
StructNet	97.4	100.2	99.4
VNect	**56.0**	**78.9**	**74.2**

表4所示实验选用的比较指标为平均关节位置误差,单位为毫米。表中最后一行加粗的结果所对应的方法为本文所使用的姿态估计方法,可以看到在三个精度实验的结果中,VNect所得到的平均关节位置误差是最小的,证明VNect算法具有更高的姿态估计精度。

3.3 视觉动捕系统实时性评价实验

实时性实验主要是以算法达到的帧率(Frame rate)来进行衡量。帧率是指位图图像连续出现在显示器上的频率,单位是帧每秒。本文选用另外两个主流的姿态估计算法(DconvMP和StructNet)训练的模型,将之嵌入到视觉动捕系统中进行性能对比。实时性实验的结果见表5。

表5 三维姿态估计算法实时性比较表

方法	帧率
DconvMP	2
StructNet	5
VNect	**30**

实时性实验比较表中最后一行加粗的数值最大,表示VNect姿态估计算法的实时性最好,30FPS已经可应用于视觉动作捕捉。

3.4 卡尔曼滤波效果评价实验

本实验主要检验卡尔曼滤波是否对驱动的虚拟人物运动具有平滑作用。对于京剧演示视频中同一帧的动作,有无卡尔曼滤波的效果对比图见图9。根据三幅图图中红色框标明的脚部细节可知,未经卡尔曼滤波处理的脚部会因为噪声干扰而翘起,经过卡尔曼滤波处理的脚部姿态会更加贴近视频图中的原始姿态。由此可知,使用卡尔曼滤波算法的虚拟人物驱动可以避免单帧的预测错误,提升虚拟人物动作的正确性。

视频图 无卡尔曼滤波处理 有卡尔曼滤波处理

图9 卡尔曼滤波效果对比图

3.5 视觉动捕系统性能主观评价实验

本实验主要用于测试使用者对于此虚拟人物互动系统的主观交互体验感。实验邀请了5位受试者，依次在摄像头前做站立、挥手、行走、侧身动作，受试者通过观察视觉动捕系统中的虚拟人物的运动效果以及体验到的虚拟人物驱动系统的交互感受，根据虚拟人物运动的精确度、流畅度等主观感受进行评分。受试者编号为1-5，分数区间设置为0-100分，实验结果见表6。

表6 视觉动捕系统性能主观实验评价表

受试者	站立	挥手	行走	侧身
1	90	90	89	91
2	95	94	88	92
3	97	96	90	95
4	96	95	92	96
5	96	94	91	92

五位受试者对于站立动作、挥手动作和侧身动作的打分均在90分以上，证明站立、挥手和侧身动作的表现较好；行走动作分数略低于90分，效果不如前两个动作效果，但是分数依然较高，证明本文视觉动捕系统带给使用者较好的体验。主观评价受试者1的站立动作、挥手动作、行走动作、侧身动作实验效果图见图10。

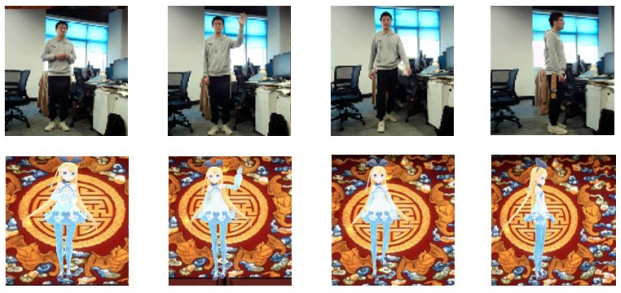

图10 主观评价实验效果图

综合以上四个实验的结果，可以发现基于VNect姿态估计算法的虚拟人物互动系统的精度、速度、平滑性、受试者评价都较优，基于人体姿态估计的京剧虚拟人物互动系统的性能较好。

5 小结

本文研发了一种基于人体姿态估计的京剧虚拟人物互动系统，仅通过向计算机输入京剧教学视频或单目普通相机捕捉的视频流，即可将京剧艺术家的动作迁移到虚拟人物身上，并实时驱动虚拟人物进行京剧表演。本动作捕捉系统在可靠性和实时性方面均达到了较好的效果。在未来的开发中，可以考虑将之迁移到移动设备中，在保证精度和速度的前提下，让基于此姿态估计算法的应用更加普及。

参考文献（References）：

[1] 邓益侬, 罗健欣, 金凤林. 基于深度学习的人体姿态估计方法综述[J]. 计算机工程与应用, 2019, 55(19):22-42.

[2] 张鋆豪, 何百岳, 杨旭升, 张文安. 基于可穿戴式惯性传感器的人体运动跟踪方法综述[J]. 自动化学报, 2019, 45(08):1439-1454.

[3] 杨扬. 基于深度相机的三维重建与运动捕捉[D]. 南京:南京邮电大学, 2020.

[4] ChenY, Tian Y, HeM. Monocular human pose estimation: A survey of deep learning-based methods[J]. Computer Vision and Image Understanding, 2020, 192.

[5] MehtaD, Sridhar S, Sotnychenko O, et al. VNect: real-time 3D human pose estimation with a single RGB camera[J]. ACM Transactions on Graphics, 2017, 36(4):1-14.

[6] 杨彬, 李和平, 曾慧. 基于视频的三维人体姿态估计[J]. 北京航空航天大学学报, 2019, 45(12):2463-2469.

[7] 朱杰. 基于Unity3D游戏人工智能的研究与应用[D]. 广州:广东工业大学, 2020.

[8] Nishida A, Soga M. Development of a learning environment for sketching human body with pose change using Motion Capture [J]. Procedia Computer Science, 2021, 192: 3696-3703.

[9] 陈伟. 基于四元数和卡尔曼滤波的姿态角估计算法研究与应用[D]. 秦皇岛:燕山大学, 2015.

[10] 吴亚峰，杜化美，张月霞. Unity3D开发实战详解[M]. 北京:人民邮电出版社, 2013.

[11] AhnB, Choi DG, ParkJ. Real-time head pose estimation using multi-task deep neural network[J]. Robotics and Autonomous Systems, 2018, 103:1-12.

编辑：王谦