文章编号:1674 - 2869(2019)02 - 0168 - 05

基于CNN的三维人体姿态估计方法

肖澳文[1,2],刘 军[*1,2],张苏沛[1,2],杜 壮[1,2],孙思琪[1,2]
1. 智能机器人湖北省重点实验室(武汉工程大学),湖北 武汉 430205;
2. 武汉工程大学计算机科学与工程学院,湖北 武汉 430205

摘 要:针对传统三维人体姿态估计受遮挡限制的问题,提出一种基于卷积神经网络(CNN)的三维人体姿态估计方法。首先,实验模型系统采用了几段单目视频为输入源进行人体姿态识别。相对于传统的人体姿态估计方法,改进了一种顺序化的卷积神经网络用于提取人体空间信息和纹理信息。并通过对视频中人体的二维姿态估计,找出了人体头部和四肢关节点的精确位置。最后,通过投影关节点到三维空间,估计出每个人的三维姿态。实验结果表明,本文方法相比传统的姿态估计算法在人体行为上的测试平均误差从98.53 mm降低至92.88 mm,对于视频中的人体三维姿态估计有更优的精度。

关键词:三维人体姿态估计;卷积神经网络;关节点
中图分类号:TP317.4 **文献标识码**:A **doi**:10. 3969/j. issn. 1674-2869. 2019. 02. 013

Three-Dimensional Human Pose Estimation Based on Convolution Neural Network

XIAO Aowen[1,2], *LIU Jun*[*1,2], *ZHANG Supei*[1,2], *DU Zhuang*[1,2], *SUN Siqi*[1,2]
1. Hubei Key Laboratory of Intelligent Robot(Wuhan Institute of Technology), Wuhan 430205, China;
2. School of Computer Science & Engineering, Wuhan Institute of Technology, Wuhan 430205, China

Abstract: To solve the problem that the traditional three-dimensional human pose estimation performance was limited by occlusion, this paper presents a three-dimensional human pose estimation method based on convolution neural network. Firstly, some monocular videos were used as the inputs to recognize the human body postures in the experiment model. Secondly, a sequential convolution neural network was constructed to extract the spatial and texture information of human body. Thirdly, the exact position of the joint points of the head and body was found through two-dimensional human pose estimation in the video. Finally, the three-dimensional pose of each person was estimated by projecting the correlation node to the three-dimensional space. The experimental results show that the mean error reduces from 98.53 mm to 92.88 mm compared with the traditional human pose estimation algorithm, and our method has higher precision in the three-dimensional human pose estimation in the testing video.

Keywords: three-dimensional human pose estimation; convolution neural network; joint points

三维人体姿态估计是非常难的一个研究课题,通常如果不借助一些穿戴设备无法直接获取三维的人体姿态[1],而三维人体姿态的估计在人体行为的理解方面有着不可估量的研究价值[2]。人体姿态估计的研究从二维到三维的转换过程,可以看作是一个三维重建的过程,目前已有的三维

收稿日期:2018-10-23
基金项目:国家自然科学基金(61172150, 61803286);智能机器人湖北省重点实验室开放基金(HBIR 201802);武汉工程大学第十届研究生教育创新基金(CX2018197, CX2018200, CX2018212)
作者简介:肖澳文,硕士研究生。E-mail:xiaoaowen@wit.edu.cn
***通讯作者**:刘 军,博士,副教授。E-mail:liujun@wit.edu.cn
引文格式:肖澳文,刘军,张苏沛,等. 基于CNN的三维人体姿态估计方法[J]. 武汉工程大学学报,2019,41(2):168-172.

重建方法[3-5]在这一过程中可以得到好的应用。由于卷积神经网络（convolutional neural network，CNN）的优秀学习能力[6-9]，通过该网络训练模型是最好的选择之一。随着计算机动画和计算机视觉等多媒体技术的快速发展，对人体姿态进行准确的三维预测估计，在智能监控、体育训练、医疗看护以及影视制作等领域具有较大的应用价值[10]。

Bogo等[11]在2016年的ECCV会议上首次提出一种三维人体姿态估计方法，该方法首先预测二维人体关节位置，然后使用SMPL模型来创建三维人体形状网格，该网格能同时捕捉人体姿态和形状。Zhou等[12]在2016年的CVPR会议上提出一种全序列的期望-最大化算法，先训练一个深度完全卷积神经网络预测二维人体关节位置的不确定性映射，然后通过该算法实现三维人体姿态估计。2017年的CVPR会议上，Pavlakos等[13]提出一种端对端训练的方法，先输入一张彩色图像，输出人体三维姿态信息，然后采用CNN进行端对端训练，将人体姿态看作N个关节点，每个关节点有一个三维坐标(x, y, z)，根据关节点坐标估计整体的三维姿态。2016年CVPR会议上，卡内基梅隆大学Shih-En Wei团队[14]提出了卷积姿态机（Convolutional Pose Machines，CPM）方法，该方法先计算每一尺度下的部件置信度，然后累加所有尺度的置信度，最后取每个部件图中置信度最大的点做为部件位置。

本文提出一种面向视频的三维人体姿态估计方法，首先调用摄像头拍摄人体视频做为CNN输入，对视频中的人体进行检测定位。然后对检测到的人体单独进行二维姿态估计，获取人体二维关节点位置，将二维人体姿态估计结果与三维人体姿态重建相结合，用网格划分三维空间。通过将二维位置坐标提升为三维，投射到有效的三维姿势空间，估计出每个人的三维姿态，各自建立对应的三维人体姿态模型。相比传统人体姿态估计，本文研究结果更加立体，视觉效果明显增强。相比传统的三维人体姿态估计方法，本文方法比Ionescu等[15]和Zhou等[12]在人体行为上的测试平均误差分别提高了31.8%和5.7%，对于视频中的人体检测，二维姿态识别，三维姿态重建均有良好的效果。

1　三维人体姿态估计

1.1　姿态估计方法

CNN是一种含有深度结构的前馈神经网络模型，包含输入层、隐含层和输出层，其核心隐含层的基本结构包括卷积层、池化层和全连接层，卷积层利用卷积核提取输入数据的抽象特征，通过局部连接和权值共享来减少参数数量，输出特征图至池化层。池化层进行特征选择和信息过滤，通过下采样来进一步减小神经元个数，简化网络计算复杂度。全连接层负责连接所有的特征。卷积核的工作原理为：

$$Z^{l+1}(i,j)=\left[Z^{l}\otimes w^{l}\right](i,j)+b=\sum_{k=1}^{K_l}\sum_{x=1}^{f}\sum_{y=1}^{f}\left[Z_k^{l}(s_0 i+x,s_0 j+y)w_k^{l+1}\right]+b \tag{1}$$

$$L_{l+1}=\frac{L_l+2p-f}{s_0}+1 \tag{2}$$

其中，$(i,j)\in\{0,1,\cdots L_{l+1}\}$，$b$是偏差量，$Z^{l}$和$Z^{l+1}$表示第$l$+1层的卷积输入和输出特征图，$L_{l+1}$为$Z_{l+1}$的特征图尺寸。$Z(i,j)$对应特征图的像素，$K$为特征图的通道数，卷积层参数中，$f$是卷积核大小、$s_0$是卷积步长，$p$是填充层数。

采用一种新的多阶段卷积神经网络，通过端到端的训练，估计二维和三维人体关节点的位置。在Shih-En Wei团队[14]的CPM卷积神经网络模型基础上引入了二维融合层和三维人体姿态概率预测模型，将二维人体姿态提升为三维，并将骨骼结构的三维信息传播到二维卷积层，根据已编码的三维信息库完成三维人体姿态的预测估计。

三维人体姿态估计模型由特征提取、二维姿态预测、二维融合层、最终转换等4个不同的模块组成，如图1所示。

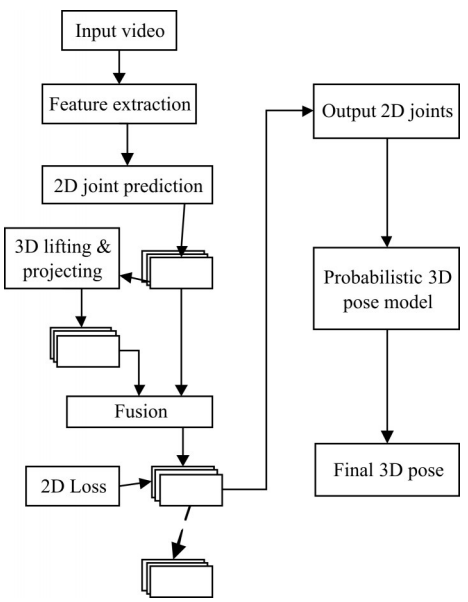

图1　三维人体姿态估计模型结构图

Fig. 1　Structure diagram of three-dimensional human pose estimation model

1)特征提取:对单帧视频的三原色(red-green blue,RGB)图像进行关节点估计获得人体的骨骼特征,将基于CNN的置信图的输出作为一个新层的输入,该层使用预训练的三维人体姿态概率预测模型将投影的二维姿态转换为三维。

2)二维姿态预测:将前一层估计的三维姿态投影回图像平面,生成一组新的投影姿态置信图,上述映射封装了人体姿态间的三维依赖关系。其中,基于CNN的预测置信图沿用CPM方法提出的计算思想,使用一组卷积层和池化层,将从图像学习中提取的特征与前一阶段获得的置信图相结合,预测更新后的二维人体关节点的置信图。

3)二维融合层:在每个阶段的最后一层学习权重,将通过CNN预测的二维姿态置信图和投影后的三维姿态置信图映射融合成一个置信图传递到下一阶段的单一评估。

4)最终转换:将最终阶段输出产生的置信图投影至三维空间,并使用三维人体姿态概率预测模型将二维姿态转换为三维,从而给出最终的三维姿态估计图像。

整个人体三维姿态估计的过程包含有6个上述的结构,代表训练的不同阶段,每个阶段都会输出一组置信图来映射二维关节点的位置坐标。每一阶段的输出置信图映射及图像都做为下一阶段的输入。在三维人体姿态概率预测模型中,三维姿态层负责将二维关节点的位置坐标提升为三维,并将它们投射到有效的三维姿势空间中。然后,将通过CNN预测的二维姿态置信图和投影后的三维姿态置信图合并输出一组针对每阶段的二维关节点坐标位置。二维和三维坐标位置的准确性在各阶段都会逐步提高。每个阶段的损失只需要用二维位姿的注释来表示,而不需要三维。整个网络架构完全可逆,并可以通过反向传播实现端到端的训练。

1.2 Human3.6M数据集

Human3.6M数据集[11]的数据采集包括4个校准摄像机的高分辨率50 Hz视频,其通过高速运动捕捉系统精确的三维人体关节位置和关节角度,包含24个像素级身体部位标签,保证了准确的捕获和同步图像数据。数据集使用Human Solutions的3传感器3D扫描仪扫描所有演员,称为Vitus Smart LC3。此数据集还包含TOF数据、演员的3D激光扫描图像、准确的背景减法和人物边界框等等。应用领域为预先计算的图像描述,可视化和判别性人体姿势预测的软件,以及测试集的性能评估。数据集根据骨架给出姿势数据,使用相机参数投影三维人体关节位置并获得非常准确的二维人体姿态信息。为保证数据的一致性和使用方便性,使用相同的32个人体关节骨架。在测试集中,为减少相关的关节数量,每只手和脚只保留一个关节。获得的网格由Human Solution ScanWorks软件预处理。

模型在Human3.6M数据集上进行训练和测试,该数据集由360万个精确的三维人体姿态组成。这是一个由5名女性和6名男性受试者组成的视频和mocap数据集,从4个不同的角度拍摄,显示他们进行的典型活动(坐、走路、打招呼、吃东西等)。动作示范如图2所示。

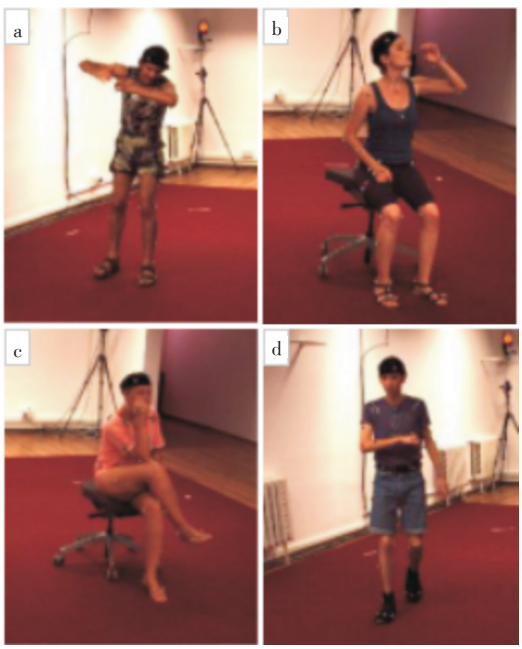

图2 Human3.6M数据集动作示范:(a)打招呼,(b)吃东西,(c)坐,(d)走路

Fig. 2 Demonstration of Human 3.6 M dataset: (a) greeting, (b) eating, (c) sitting, (d) walking

模型利用Human3.6M数据集进行训练,其中包含视频图像的训练,该估计模型训练完成后,测试也通过视频图像进行。本文获取视频图像的方法是直接调用摄像头,利用摄像头拍摄下的包含人体的视频图像,测试模型的三维人体姿态估计效果,其中包括单人和多人的视频场景。

2 结果与讨论

实验运行环境:服务器配置为:CPU[Interl(R) Core(TM)i7-8700 CPU @3.20 GHz],显卡(NVIDA GeForce GTX 1080Ti),系统:64位Ubuntu 16.04 LTS,内存:16 GB,磁盘:(3 TB),固态硬盘(256 GB)。

实验平台为：Tensorflow 1.4.0，OpenCV 3.0，Python 3.5。

实验数据集：Human3.6M数据集。

实验采用摄像头拍摄获得视频作为输入，并获得二维关节点和三维姿态估计结果，实验结果如图3所示。由图3(a)和图3(b)可以看出本文结果在复杂场景下首先能准确识别出目标人体，并获得人体二维姿态的关键关节点及骨骼线条，不同的部位用不同颜色的线条表示，展示出了头部和四肢的关节点位置。

三维人体姿态估计结果如图3(c)和图3(d)所示，结果能准确反映人体三维关节姿态及三维网格空间中人体各部位所在位置的三维空间坐标，三维人体估计模型一共有16个关键关节点，均在图3(c)和3(d)中标出，不同部位用不同颜色的线条表示，方便区分。

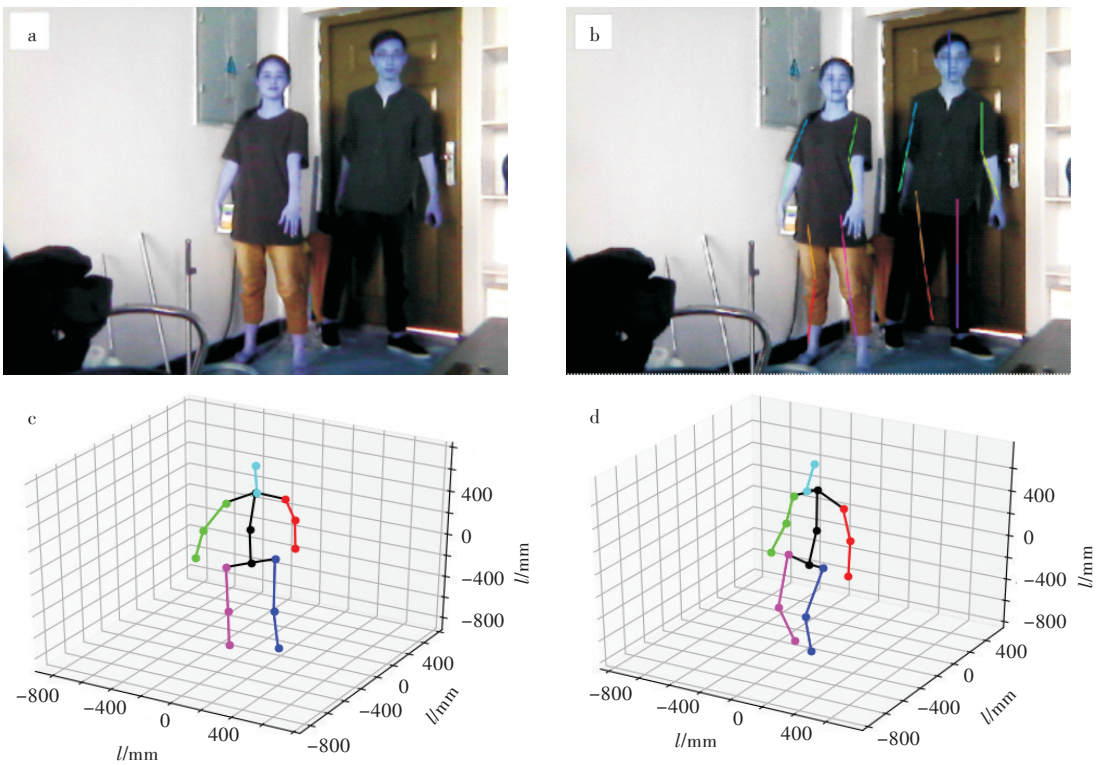

图3　视频的三维人体姿态估计：(a)输入视频，(b)关节点演示图，(c)左边人体三维估计，(d)右边人体三维估计

Fig. 3　Three-dimensional human pose estimation of videos: (a) input videos, (b) demonstration of joint points, (c) three-dimensional pose estimation of left person, (d) three-dimensional pose estimation of right person

基于概率主成分分析(probabilistic principal component analysis, PPCA)方法[16]，对该三维人体姿态估计模型做出了在Human3.6M数据集上部分人体动作行为的姿态估计测试，跟传统的方法估计的误差对比结果如表1所示。Ionescu等[15]提出Human3.6数据集，并对该数据集进行了傅里叶核近似测试。Zhou等[12]用二维姿势标注来训练CNN联合回归量和单独的3D mocap数据集，从而学习三维稀疏基础来建立三维估计模型。本文模型的姿态估计测试在各个人体动作行为上的误差均为最低值，在吃东西行为上误差为79.12 mm，在坐行为上误差为118.96 mm，在走路行为上误差为78.28 mm，在打招呼行为上为95.17 mm，性能相比前两种方法均有不同程度的提升，平均误差达到92.88 mm，相比Ionescu等[15]的傅里叶核近似法提高了31.8%，相比Zhou等[12]的三维稀疏回归法提高了5.7%，估计效果明显更加优越。3种方法平均误差对比如表1所示。

表1　不同行为下姿态估计测试误差对比

Tab. 1　Comparison of test errors of human pose estimation in different behavior videos　　mm

方法比较	三维人体姿态估计测试误差				
	吃东西	坐	走路	打招呼	平均误差
傅里叶核近似法[15]	132.37	151.57	96.60	164.39	136.23
三维稀疏回归法[12]	87.05	124.52	79.39	103.16	98.53
本文方法	79.12	118.96	78.28	95.17	92.88

3 结 语

本文提出了一种基于CNN的人体姿态识别方法,从视频中估计三维的人体姿势,在估计误差上优于传统的解决方案。该方法能有效地将视频图像中的人体姿态从二维升级到三维空间,可以在单人和多人场景下识别出每一个人的三维人体姿态,输出三维人体模型图。后续在三维人体姿态估计领域的研究中,会重点关注对输入图像的预处理,减弱图像阴影对图像识别的影响,使得识别效果更加准确。同时,做到实时的视频人体姿态识别也将是今后研究的方向之一。

参考文献

[1] 辛义忠,邢志飞. 基于Kinect的人体动作识别方法[J]. 计算机工程与设计,2016,37(4):1056-1061.

[2] 赵海峰,费婷婷,王文中,等. 结合个性化建模和深度数据的三维人体姿态估计[J]. 计算机系统应用,2016,25(11):118-125.

[3] 陈起凤,刘军,李威,等. 三维重建中线段匹配方法的研究[J]. 武汉工程大学学报,2018,40(4):446-450.

[4] 郭盛威,章秀华,范艳,等. 三维重建表面几何特征的提取与参数测量计算[J]. 武汉工程大学学报, 2016,38(2):185-188.

[5] 刘军,李娜,刘鹏. 双目视觉立体标定方法的改进[J]. 武汉工程大学学报,2013,35(10):68-73.

[6] 周志华,陈世福. 神经网络集成[J]. 计算机学报,2002,25(1):1-8.

[7] 张苏沛,刘军,肖澳文,等. 基于卷积神经网络的验证码识别[J]. 武汉工程大学学报,2019,41(1):89-92.

[8] 王钰清,陆文凯,刘金林,等. 基于数据增广和CNN的地震随机噪声压制[J]. 地球物理学报,2019,62(1):421-433.

[9] 吴和保,李晓微,龙玉阳,等. 人工神经网络快速预测蠕墨铸铁的性能[J]. 武汉工程大学学报,2013,35(10):63-67.

[10] 李天峰. 基于多媒体技术的三维人物图像动态重构[J]. 现代电子技术,2018,41(9):68-71.

[11] BOGO F, KANAZA W A, LASSNER C, et al. Keep it SMPL: automatic estimation of 3D human pose and shape from a single image [J]. Springer International Publishing,2016,10(6): 561-578.

[12] ZHOU X, ZHU M, LEONARDOS S, et al. Sparseness meets deepness: 3D human pose estimation from monocular video [J]. IEEE Conference on Computer Vision and Pattern Recognition, 2016, 537(1): 4966-4975.

[13] PAVLAKOS G, ZHOU X, DERPANIS K G, et al. Coarse-to-fine volumetric prediction for single-image 3D human pose [J]. IEEE Conference on Computer Vision and Pattern Recognition, 2016, 139(1): 1263-1272.

[14] WEI S E, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines [J]. IEEE Conference on Computer Vision and Pattern Recognition, 2016, 511(1): 4724-4732.

[15] 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 & Machine Intelligence, 2014, 36(7): 1325-1339.

[16] 杨博雄,杨雨绮. 利用PCA进行深度学习图像特征提取后的降维研究[J]. 计算机系统应用,2019,28(1):279-283.

本文编辑:陈小平