

二维人体关键点检测算法综述

史小强¹, 黄 钢^{1,2}, 苏可怡³

(1.上海理工大学健康科学与工程学院, 上海 200093;

2.上海健康医学院附属嘉定中心医院上海市分子影像学重点实验室, 上海 201318;

3.中国医药工业研究总院, 上海 201203)

✉ 858961788@qq.com; huanggang@sumhs.edu.cn; 310460791@qq.com



摘 要:随着二维人体关键点检测应用领域的扩展,对二维人体关键点检测的检测速度和精度提出了更高的要求。为了进一步分析二维人体关键点检测,首先介绍了基于传统方法和基于深度方法的二维人体关键点检测的模型,其次介绍了二维人体关键点检测的常用数据集和研究方向,最后得出了基于深度学习的二维人体关键点检测可以高精度地定位人体关键点的结论。未来,可以将二维人体关键点检测应用到更多新的领域进行更深入的研究。同时,提高模型的鲁棒性和泛化能力,探索更加高效的模型架构,以及减少模型复杂度等将成为研究热点。

关键词: 二维图像; 关键点检测; 神经网络; 深度学习

中图分类号: TP391 **文献标识码:** A

Overview of Key Point Detection Algorithms for 2D Human Body

SHI Xiaoqiang¹, HUANG Gang^{1,2}, SU Keyi³

(1. School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China;

2. Shanghai Key Laboratory of Molecular Imaging, Jiading District Central Hospital, Shanghai University of Medicine and Health Sciences, Shanghai 201318, China;

3. China State Institute of Pharmaceutical Industry, Shanghai 201203, China)

✉ 858961788@qq.com; huanggang@sumhs.edu.cn; 310460791@qq.com

Abstract: As the application field of 2D human body key point detection expands, higher requirements are put forward for the detection speed and accuracy of 2D human body key point detection. In order to further analyze 2D human body key point detection, this paper first introduces 2D human body key point detection models based on traditional methods and deep learning methods. Then, commonly used datasets and research directions for 2D human body key point detection are introduced. Finally, it is concluded that 2D human body key point detection based on deep learning can accurately locate human body key points. In future, 2D human body key point detection can be applied to more new fields for further in-depth research. At the same time, improving the robustness and generalization ability of the model, exploring more efficient model architectures, and reducing model complexity will become research focuses.

Keywords: 2D image; key point detection; neural network; deep learning

1 引言(Introduction)

随着科技的发展,含有人体信息的二维图像被应用到各个领域。人体图像包含人体骨骼关节信息,若是能有效利

用,这些信息就可以在动作识别、人工智能、疾病预测等领域发挥重要价值。目前,人体关键点检测主要用于人体姿态估计^[1]、人机交互^[2]、辅助医疗器械诊断^[3]等领域。

由于二维人体图像的复杂性较高,如受遮挡因素、环境因素、角度因素等的影响,使二维人体关键点检测受到巨大挑战。通过对二维人体关键点检测方法的研究,发现无论是单人图像还是多人图像,基于深度方法的检测精度都明显高于基于传统方法的检测精度;而且,基于传统的人体关键点检测方法主要依赖于提取人工设定的特征,不但效率低,而且会耗费大量的人力。目前,基于深度学习的方法是用于二维人体关键点检测的主流方法。

2 二维人体关键点检测(2D Human body key point detection)

二维人体关键点检测又可分为传统方法和深度方法,在深度学习方法应用于人体关键点检测之前,单人二维人体关键点检测是传统方法的研究重点。深度学习理论实现了人体关键点检测从单人二维关键点检测到多人关键点检测的推广,并且检测精度有了极大的提升,使人体关键点检测能够应用在更多的场景中。

2.1 基于传统方法的二维人体关键点检测

人体关键点检测的传统方法主要是依赖于图像的图形结构实现检测,例如边缘检测、角点检测、区域分割等技术。此外,基于图模型的方式也被提出,比如随机场模型和依赖图模型。1973年,FISCHLER等^[4]依靠图模型提出图形结构的匹配(Matching of Pictorial Structures, PSM)来检测面部结构的关键点。FELZENSZWALB等^[5]在FISCHLER的基础上对算法进行了改进和优化,使得PSM算法更加高效、准确,并且可以在更多的应用领域中使用。此后,ANDRIUKA等^[6]将PSM应用在人体关键点检测的模型中,提出了一种基于部件的模型,将人体分解为多个部件,并使用图形结构匹配算法检测每个部件的关键点,从而实现全身姿态估计。在此基础上,FELZENSZWALB等^[7]提出一种可变形部件模型(Deformable Part Model, DPM),将物体分解为多个部分并通过对局部特征进行建模捕捉物体的各种形状变化。在人体关键点检测任务中,DPM可以用于学习和识别不同人体姿态的结构模式。此后,基于图像的图形结构的人体关键点检测算法成为传统方法的主流算法。

基于图像的图形结构的人体关键点检测算法主要包括用来描绘人体组成部分的部件模型及描绘部位的空间关系的空间模型。NAEGELI等^[8]将图像分割技术应用在人体关键点检测中,去除了人体图像的背景对检测精度的干扰。ACHILLES等^[9]提出了一种基于具有决策功能的部位检测器,以提高人体关键点检测的检测精度。在基于空间模型的算法中,WANG等^[10]提出了一种非树形结构,解决了人体关键点检测中自遮挡的问题。YANG等^[11]提出了混合部件模型,解决了空间约束的多变性。

基于传统方法的二维人体关键点检测需要人工设计特征,具有高效率、效果直观、方法简单的优点。但是,由于人体图像的背景复杂多变,使其检测准确度难以提升,无法达到

目前人们的需求。

2.2 基于深度方法的二维人关键点检测

随着深度学习在计算机视觉领域中的应用,目前很多人体关键点检测方法都是基于深度学习方法,并且已经取得了良好的效果。无论是单人关键点检测还是多人关键点检测,深度学习方法在其中都发挥了良好的性能。

2.2.1 基于深度方法的二维单人关键点检测

TOSHEV等^[12]第一次将卷积神经网络(Convolutional Neural Networks, CNN)应用到人体关键点检测中,并且提出了深度姿态估计DeepPose算法,首先应用DNN^[13](Deep Neural Networks)深度神经网络粗略检测出部分人体关键点的位置,其他关键点采用DNN回归预测其位置,在保证人体图像的尺寸相同且计算参数不变的情况下,级联的回归器可以获得更加精确的人体关键点的位置。DeepPose是人体关键点检测从传统方法过渡到深度方法的奠基石。

神经网络在提高精度的同时也会带来一定的弊端,随着神经网络层数的增加,其梯度会逐渐减少。对此,WEI等^[14]提出了卷积状态机(Convolutional Pose Machine, CPM),它能够直接从数据中学习图像特征和空间上下文特征。首先CPM在多阶段估计中确定人体关键点的大致位置,然后在每个阶段不断精确关键点的位置。CPM组合了多个卷积网络在多阶段的卷积网络中不断确定检测点的结果,因此解决了随着神经网络层数的增加,伴随着其梯度减少的问题。

以往的网络结构仅使用最底层的卷积特征,这可能会导致有效信息丢失,NEWELL等^[15]提出了一种能够利用多尺度特征识别人体图像的网络结构叠加的沙漏网络(Stacked Hourglass Networks, SHN),该结构使用多个沙漏网络进行级联,以提高检测效果,充分利用了人体关键点之间的相关性。

LIFSHITZ等^[16]使用一致投票的方式定位关键点位置,提出了深度共识投票(Deep Consensus Voting, DCV),在检测人体关键点位置时,不用继续依赖稀疏的关键点位置集,而是使用高密度的多目标进行投票,不仅检测效果好,还可以计算与检测图像相关的联合关键点的概率。

在深度学习层面,人体关键点检测已经取得了良好的泛化性能,但是无法在轻量级的环境中运行,ZHANG等^[17]提出了一种快速姿态估计模型FPD(Fast Pose Distillation),这是一种轻量化的人体关键点检测模型,使用了4个沙漏模型,能够以较低的计算空间执行。该网络在降低人体关键点检测的网络复杂度方面取得了较好的效果。

截止到目前,基于深度方法的二维单人关键点检测研究非常丰富且全面,现有的检测研究多集中在提高算法的检测精度和检测准确率方面。

2.2.2 基于深度方法的二维多人关键点检测

在单人关键点检测中,人数是已知的,但是对于多人关键点检测来说,人数和位置都是未知的,因此在人体关键点

检测任务中,多人关键点检测的难度更大。基于深度方法的多人关键点检测主要分为2类:一类是自上而下的方法,另一类是自下而上的方法,具体分类如图1所示。

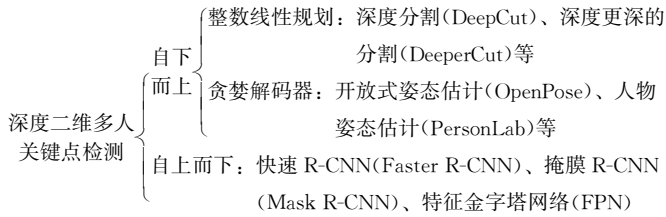


图1 深度二维人体关键点检测方法

Fig. 1 Deep 2D human body key point detection method

2.2.2.1 自下而上

自下而上的原理是首先估计图像中所有人的人体关键点,然后按照一定的方法把所得到的人体关键点的位置分配给图像中的每个人,从而得到图像中每个人体关键点的位置。

PISHCHULIN等^[18]提出了DeepCut,先使用CNN选出人体部位的区域,将所选出的区域作为一个节点,把所有的区域节点合成一张图像,然后计算节点之间的关联性权重,将其看成ILP(Integer Linear Program)线性编程问题,这种方法通过图论节点的聚类,有效地进行非极大值的抑制。此外,此算法在识别被遮挡的身体部位方面取得了良好的效果。在DeepCut的基础上,INSAFUTDINOV等^[19]为了降低其网络复杂度,提出了性能更好的残差网络DeeperCut。DeepCut和DeeperCut的方法都属于整数线性规划,在解决多变量最优决策的问题上有着良好的效果,但是该方法对数据的准确性要求很高且计算量大。

CAO等^[20]提出了一种高效检测多人关键点检测的方法OpenPose。通过网络训练将身体部位跟图像相互联系,利用全局上下文进行编码。该架构通过同一顺序预测过程的两个分支联合学习局部位置及其关联性,实现了高检测精度的同时具备实时性的多人关键点检测。基于OpenPose算法,OSOKIN等^[21]提出了轻量化的OpenPose,在保证精度的同时,还提高了人体关键点检测的检测速度。苏波等^[22]通过改进OpenPose,优化了缩减预测阶段的重复分支,提高了检测精度。

基于单人关键点检测,NEWELL等^[23]提出了一种通用检测分组方法,用标签将每个检测与同一组中的其他检测相关联。将关联嵌入方法与堆叠Hourglass网络集成在一起,为每个身体关节生成和标记热图,然后将具有相似标签的身体关节分组到个人身上,从而实现多人关键点检测。

2.2.2.2 自上而下

自上而下的原理主要包括目标检测和人体关键点检测两个部分,首先利用人体检测器检测出图像中的所有人体,然后对选出的每个人使用单人关键点检测。因此,自上而下的方法主要侧重于人体检测器的研究,常用的人体部位检测器有Faster R-CNN(Faster Region-Convolutional Neural Networks)^[24]、Mask R-CNN(Mask Region-Convolutional Neural Networks)^[25]、

FPN(Feature Pyramid Networks)^[26]等。

Mask R-CNN的原理是先将图像中人体框选出来,然后分割出特征图用于关键点检测。与其他自上而下的人体关键点检测方法相比,基于Mask R-CNN的方法在精度上表现更好,同时速度也较快,这也是它被广泛使用的原因之一。比如,宋玲等^[27]提出的SN Mask R-CNN是在Mask R-CNN的基础上加入通道重组,以提升其网络性能,在保留精度的同时提高了其检测速度。

PAPANDREOU等^[28]提出了一种自上向下的多人关键点检测方法,首先利用Faster R-CNN估计出目标人体在图像中的尺寸和位置,然后根据检测出目标人体的关键点使用ResNet估计每个关键点的偏移量和密度热图,最后将偏移量作用在热力图上完成多人关键点检测。

CHEN等^[29]提出一种级联金字塔网络(Cascaded Pyramid Network, CPN),首先粗略地检测人体关键点,然后利用RefineNet网络检测被遮挡的以及比较难检测的关键点,这个模型能够同时兼顾人体关键点的局部信息和全局信息。

3 常用数据集(Common datasets)

在人体关键点检测的发展过程中有很多公开的数据集,比如利兹体育姿势数据集(LSP)^[30],这是一种包含多种人体体育姿势的数据集,其中分为竞技、羽毛球、体操、跑酷等运动姿态,一共约有2 000张图像且带有姿势注解;宾夕法尼亚大学提供的好莱坞电影中的图像数据集(FLIC)^[31],包含约5 000张图像,其中每张图像都被标注了10个上半身关节点;来自马克斯·普朗克计算机科学研究所的数据集(MPII)^[32],涵盖了410种人体动作的照片,整个数据集包含25 000张图像,标注了其中40 000名人体的关节点;最常用的是由微软团队提供的微软通用上下文中的物体数据集(MSCOCO)^[33],其数据场景非常丰富,MSCOCO数据集是目前主流的人体关键点检测数据集,它的数据量主要包括训练集、验证集及测试集,一共有330 000张图片,标记的图像超过200 000张。除此之外,还有比赛用的数据集,如智能算法挑战赛AI Challenger,其中包含700 000个人体关键点标注,300 000张图片场景标注和语义描述数据,是国内迄今公开的规模最大的科研数据集。表1列出了6种常用的人体关键点检测的数据集,包括数据集的检测类别、关键点数量及数据集包含的图片数量。

表1 人体关键点检测数据集

Tab.1 Datasets of human body key points

数据集	类别	关键点数量/个	图片数量/张
LSP	单人	14	2 000
MPII	单人、多人	16	25 000
MSCOCO	多人	17	330 000
FLIC	单人	10	5 000
HKD	多人	14	300 000
AI Challenger	多人	14	380 000

4 研究展望(Research prospect)

(1)数据集存在的问题: 目前使用的数据集中大多是人体部位的常规动作, 随着二维人体关键点检测的发展, 这些数据集已经不能满足实际应用的要求, 比如跌倒检测、医学康复动作分析等; 而且, 目前的研究大多使用有监督的训练模式, 所使用的数据集都是经过人力标注的, 这不但浪费人力资源, 还会使模型的关键点检测精度降低。因此, 今后的研究中应当尽可能地增加数据集的种类, 使人体关键点检测可以应用到更多的场景中。

(2)研究中出现的问题: 目前, 人体关键点检测算法大多比较复杂, 网络运算量大, 因此出现了不少关于轻量化二维人体关键点检测网络, 比如 VGGNet、ResNet、GoogleNet 等。已经有很多研究者为了降低网络参数的数量, 将轻量化的网络结构应用到人体关键点检测中。

(3)三维人体关键点检测: 三维人体关键点需要基于二维的基础, 就是在二维人体关键点结果的基础上加上深度信息。深度信息的引入, 使三维的人体关键点检测比二维人体关键点检测更加精准, 拥有更高的研究价值。二维人体关键点的发展为实现更好的三维人体关键点检测奠定了基础。

5 结论(Conclusion)

二维人体关键点检测作为计算机视觉的基本任务, 在各个领域已经发挥了重要的作用。未来, 二维人体关键点检测在网络复杂度及应用领域还有很大的发展空间。本文对二维人体关键点检测的发展从传统方法到深度方法的应用方面进行了阐述, 综合分析了提升检测精度和检测性能的相关模型, 并对二维人体关键点检测的发展前景进行了展望。

参考文献(References)

- [1] BARTOL K, BOJANIĆ D, PETKOVIĆ T, et al. A review of body measurement using 3D scanning[J]. IEEE Access, 2021, 9:67281-67301.
- [2] DOCEKAL J, ROZLIVEK J, MATAS J, et al. Human keypoint detection for close proximity human-robot interaction[C]// IEEE. 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids). New York, USA: IEEE, 2022:450-457.
- [3] AHMED I, JEON G, PICCIALI F. A deep-learning-based smart healthcare system for patient's discomfort detection at the edge of Internet of things[J]. IEEE Internet of Things Journal, 2021, 8(13): 10318-10326.
- [4] FISCHLER M A, ELSCHLAGER R A. The representation and matching of pictorial structures[J]. IEEE Transactions on Computers, 1973, 22(1):67-92.
- [5] FELZENSZWALB P F, HUTTENLOCHER D P. Pictorial structures for object recognition[J]. International Journal of Computer Vision, 2005, 61(1):55-79.
- [6] ANDRILUKA M, ROTH S, SCHIELE B. Pictorial structures revisited: people detection and articulated pose estimation[C]// STEFAN R. 2009 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2009:1014-1021.
- [7] FELZENSZWALB P, MCALLESTER D, RAMANAN D. A discriminatively trained, multiscale, deformable part model[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2008 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE Computer Society, 2008:1-8.
- [8] NÄGELI T, OBERHOLZER S, PLÜSS S, et al. Flycon: real-time environment-independent multi-view human pose estimation with aerial vehicles[J]. ACM Transactions on Graphics, 2018, 37(6):182.1-182.14.
- [9] ACHILLES F, ICHIM A E, COSKUN H, et al. Patient mocap: human pose estimation under blanket occlusion for hospital monitoring applications[C]// OURSELIN S, JOSKOWICZ L, SABUNCU M R, et al. Medical Image Computing and Computer Assisted Intervention. Cham, Switzerland: Springer, 2016:491-499.
- [10] WANG J B, QIU K, PENG H W, et al. AI Coach: deep human pose estimation and analysis for personalized athletic training assistance[C]// PHILLIPS W, RILOFF E. Proceeding of the 27th ACM International Conference on Multimedia. New York, USA: ACM, 2019:2228-2230.
- [11] YANG Y, RAMANAN D. Articulated pose estimation with flexible mixtures of parts[C]// Institute of Electrical and Electronic Engineers. 2011 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE Curran Associates, 2011: 1385-1392.
- [12] TOSHEV A, SZEGEDY C. DeepPose: human pose estimation via deep neural networks[C]// Institute of Electrical and Electronic Engineers. 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, US: IEEE, 2014:1653-1660.
- [13] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6):84-90.
- [14] WEI S E, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE, 2016:4724-4732.

- [15] NEWELL A, YANG K Y, DENG J, et al. Stacked hourglass networks for human pose estimation[C]// BASTIANL, NICU S, MAX W, et al. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, Berlin: Springer, 2016:483-499.
- [16] LIFSHITZ I, FETAYA E, ULLMAN S. Human pose estimation using deep consensus voting[C]// European Conference on Computer Vision Amsterdam. European Conference on Computer Vision. Berlin, Germany: ECCV, 2016:246-260.
- [17] ZHANG F, ZHU X T, YE M. Fast human pose estimation[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA: IEEE COMPUTER SOC, 2019:3517-3526.
- [18] PISHCHULIN L, INSAFUTDINOV E, TANG S Y, et al. Deep-cut: Joint subset partition and labeling for multi person pose estimation[C]// Institute of Electrical and Electronics Engineers. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2016:4929-4937.
- [19] INSAFUTDINOV E, PISHCHULIN L, ANDRES B et al. Deep-ercut: a deeper, stronger, and faster multi-person pose estimation model[C]// LEI B, MATAS J, SEBE N, et al. European Conference on Computer Vision. Berlin, Germany: ECCV, 2016:34-50.
- [20] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]// Institute of Electrical and Electronics Engineers. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE Computer Society, 2020:172-186.
- [21] OSOKIN D. Real-time 2D multi-person pose estimation on CPU: lightweight OpenPose[C]// MARIA D M, GABRIELLA S D B, ANA F. Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods. Cham, Switzerland: Springer, 2019:744-748.
- [22] 苏波,柴自强,王莉.基于改进 OpenPose 的视频动作评分方法研究[J].软件工程,2022,25(9):22-26.
- [23] NEWELL A, HUANG Z, DENG J, et al. Associative embedding: end-to-end learning for joint detection and grouping[C]// GUTON I, LUXBURG U V, BENGIO S, et al. Advances in Neural Information Processing Systems. USA: NIPS, 2017:2277-2287.
- [24] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(6): 1137-1149.
- [25] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 386-397.
- [26] LIN T Y, DOLLÁR P, GIRSHICK R B, et al. Feature pyramid networks for object detection[C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2017: 936-944.
- [27] 宋玲,夏智敏.人体关键点检测的 Mask R-CNN 网络模型改进研究[J].计算机工程与应用,2021,57(1):150-160.
- [28] PAPANDREOU G, ZHU T, CHEN L C, et al. PersonLab: person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model[C]// FERRARI V, HEBERT M, WEISS Y, et al. Proceedings of the 15th European Conference on Computer Vision. Berlin, Germany: Springer, 2018:282-299.
- [29] CHEN Y L, WANG Z C, PENG Y X, et al. Cascaded pyramid network for multi-person pose estimation[C]// Institute of Electrical and Electronics Engineers. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2018: 7103-7112.
- [30] JOHNSON S, EVERINGHAM M. Clustered pose and nonlinear appearance models for human pose estimation[C]// British Machine Vision Association. Proceedings of the British Machine Vision Conference 2010. British: British Machine Vision Association, 2010:1-11.
- [31] SAPP B, TASKAR B. Modex: multimodal decomposable models for human pose estimation [C]// Institute of Electrical and Electronics Engineers. Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2013:3674-3681.
- [32] ANDRILUKA M, PISHCHULIN L, GEHLER P, et al. 2D human pose estimation: new benchmark and state of the art analysis[C]// Proceedings of the 2014 IEEE Conference on computer Vision and Pattern Recognition. New York, USA: IEEE, 2014:3686-3693.
- [33] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: common objects in context[C]// FLEET D, PAJDLA T, SCHIELE B, et al. Computer Vision-ECCV 2014. Berlin, Germany: IEEE, 2014: 740-755.

作者简介:

史小强(1999-),男,硕士生.研究领域:医学图像处理.

黄钢(1961-),男,博士,教授.研究领域:核医学分子影像.本文通信作者.

苏可怡(1997-),女,硕士生.研究领域:细胞图像处理.