文章编号：2095-2163(2021)06-0139-05 中图分类号：TP 391 文献标志码：A

深度学习的二维人体姿态估计综述

张静静，宁 媛，章成学
（贵州大学 电气工程学院，贵阳 550025）

摘 要：本文针对基于深度学习的二维人体姿态估计方法进行了全面综述。首先对这些深度学习技术进行了分类、分析和比较，并介绍了二维人体姿态估计中常用的数据集和指标，最后还讨论了有待解决的问题和未来研究的挑战。

关键词：人体姿态估计；深度学习；机器视觉；图像处理；人工智能

Deep learning based 2D human pose estimation: a survey

ZHANG Jingjing, NING Yuan, ZHANG Chengxue
(College of Electrical Engineering, Guizhou University, Guiyang 550025, China)

【Abstract】 This paper provides a comprehensive overview of the two-dimensional human pose estimation methods based on deep learning. First, these deep learning technologies are classified, analyzed and compared, and the commonly used data sets and indicators in two-dimensional human pose estimation are introduced. Finally, the problems to be solved and the challenges of future research are discussed.

【Key words】 human pose estimation; deep learning; computer vision; image processing; artificial intelligence

0 引 言

人体姿态估计任务已经研究了几十年，目的是从给定的传感器输入获取人体的姿态，通常使用基于视觉的方法来获得。近年来，随着深度学习在图像分类[1]、目标检测[2]、语义分割[3]等计算机任务上的良好表现，姿态估计利用深度学习技术也取得了快速发展。主要的发展包括设计了良好且具有强大估计能力的网络，以及更丰富的数据集，用于训练网络和更实际的人体模型。虽然已有一些关于姿态估计的评论，但是国内仍然缺乏一份调查，来总结最近基于深度学习的二维人体姿态估计成果。

姿态估计作为计算机视觉基础任务之一，是一个非常重要的研究领域，可以应用于许多方面。如：动作识别、动作检测[4]、电影与动画、人体跟踪[5]、虚拟现实、人机交互、视频监控、医疗辅助、自动驾驶、运动运动分析等等。

二维人体姿态估计具有一些独具的特点和挑战。二维人体姿态估计的挑战主要集中在 3 个方面：

(1)灵活的身体结构，表明复杂的相互依赖关节和高自由度的四肢，可能导致自咬合或罕见甚至复杂的姿态。

(2)不同的身体外观，包括不同的衣服和附近人姿态的误导。

(3)复杂的环境可能导致前景遮挡、附近人遮挡，各种视角以及摄像机视图中的截断。

本次调查广泛总结了 2014 年以来发表的基于深度学习的人体姿态估计方法的里程碑式研究成果。

1 单人姿态估计

基于深度学习的单人姿态估计方法的目标是定位人体部分的关键点。典型的单人姿态估计模型框架分为 2 种：一是直接从特征中回归关键点，称之为基于直接回归的框架；二是先生成热图，并通过热图推断关键点位置，称之为基于热图的框架。

1.1 基于直接回归框架

一些研究是基于直接回归框架提出的，例如 Toshev 等人[4]提出了一种直接预测人体关键点的级联 DNN 回归器。然而，如果没有其它的过程，直接从特征图学习映射关系是很困难的。Carreira 等

基金项目：国家自然科学基金(61663005)。

作者简介：张静静(1996-)，男，硕士研究生，主要研究方向：机器视觉；宁 媛(1968-)，女，硕士，教授，硕士生导师，主要研究方向：检测技术与自动化装置、人工智能。

通讯作者：宁 媛 Email：ee.yning@gzu.edu.cn

收稿日期：2021-01-13

人[5]使用了自校正模型。通过反馈误差预测,可以逐步改进预测的关键点位置。Sun 等人[6]提出了一种称为“合成姿势回归”的结构感知方法。与其它相关工作不同的是,该方法使用骨骼而不是关节重新参数化姿势表示,骨骼之间的相互作用通过一个成分损失函数进行编码,这样的做法更原始、更稳定、并且更易于学习。Luvizon 等人[7]提出 Soft-argmax,将热图用一个完全可微的方式转换成坐标,其端到端的方式可训练网络采用基于关键点误差距离的损失函数和基于上下文的结构,使其能够获得与最先进的基于热图的框架相比较的结果。

1.2 基于热图框架

很多研究都采用了基于热图的框架,其中一些研究在提出的模型中利用了人类的先验信息。例如,Chen 等人[8]使用了由 DCNN 学习的具有成对关系的图形模型;Chen 等人[9]通过采用条件生成对抗网络(GANs)的训练策略来整合人体的先验知识等等。基于热图的实例如图 1 所示。

(a) 原始图像 (b) 生成的热图 (c) 检测结果

(a) Original image (b) Generated heat map (c) Detection results

图 1 基于热图的实例

Fig. 1 Example based on heatmap

网络结构设计一直是基于深度学习方法的主题。卷积式位姿机(CPM)[10]对热图进行多阶段回归,并使用中间监督来避免消失梯度。Newell 等[11]设计了一种称为“堆叠沙漏”的新型网络结构。实践证明重复自下而上、自上而下的中间监督处理,是提高人体姿势检测性能的关键。Chu 等人[12]建立了基于堆叠沙漏的基线模型,采用多上下文注意机制,使模型更加健壮和准确。此外还通过耦合沙漏残余单元来改进堆积沙漏的结构。Martinez 等人[13]提出采用深度神经网络直接使用 2D 关键点来预测 3D 关键点。实验结果表明,二维检测是导致三维人体姿态估计误差的主要原因之一。

1.3 关于两种框架的讨论

关节位置的直接回归是高度非线性的,所以不仅在映射学习中存在困难,还不能应用于多人情况(自底向上方法或一个检测框包含多个人的情况)。但是,如果采用一些特殊的技术相结合,直接回归会更可靠,因为当应用直接回归时,最终结果可以在不处理热图的情况下,以端到端的方式获得,不需要太多的更改而应用到 3D 场景。相比之下,基于热图的框架首先回归热图。热图可以可视化,可以增强人类的理解和对更加复杂的情况进行建模。基于热图的框架预测结果的精度依赖于热图的分辨率,这需要较高的内存消耗[18]。因此,对于框架选择问题没有一个绝对的结论,每种框架都有其优点和缺点。

2 二维多人姿态估计

与单人姿态估计不同,由于输入图像中没有人数的提示,多人姿态估计需要同时处理检测和定位任务。根据高层抽象还是低层像素开始计算方式的不同,人体姿态估计方法可以分为自上而下方法和自下而上方法。

自顶向下的方法通常使用人体检测器,获取输入图像中每个人的边界框,然后直接利用现有的单人姿态估计方法来预测人的姿态。预测的姿势精度很大程度上取决于对人的检测精度。整个系统的运行时间与人员数量成比例。而自下而上的方法直接预测所有人的二维关节,然后将其分组。复杂环境下关节点的正确分组是一项具有挑战性的研究课题。

2.1 自顶向下方法

自顶向下姿态估计方法的 2 个最重要的组成部分是:人体区域检测和单人姿态估计。大部分研究集中在基于现有人体检测方法上的人体部位估计。Iqbal 等人[14]使用基于卷积姿态机的姿态估计器来生成初始姿态,然后利用整数线性规划(ILP)得到最终位姿。Fang 等人[15]采用了空间转换网络(STN)、非最大抑制(NMS)和沙漏网络(Hourglass network),以便存在不精确的人体边界框时进行姿态估计。Huang 等人[16]设计了一个 CFN 网络,以 incep-v2 网络为骨干网络。该网络采用多层次监督,实现粗预测和精预测的学习。Xiao 等人[17]在 ResNet 最后一个卷积层后添加了几个逆卷积层,从低分辨率的特征中生成热图。Chen 等人[18]提出了一种级联金字塔网络(CPN),该网络利用不同层次的多尺度特征映射,从局部和全局特征中获取更多的推理,并对困难节点进行在线硬关键点挖掘损失。

基于不同 HPE 方法的相似位姿误差分布,Moon 等人[19]设计了 PoseFix 网,用来改善从任何方法估计的位姿。M. Wang 等人[20]提出了一种新颖

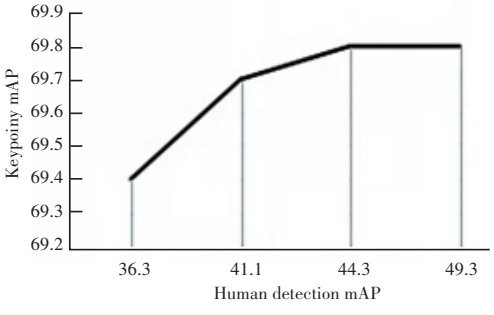
的自上而下的方法，可以解决视频中多人人体姿势估计和跟踪的问题。与现有的自上而下的方法相比，该方法不受其人员检测器性能的限制，并且可以预测未定位人员实例的姿势。

若将现有的检测网络与单一的姿态估计网络相结合，可以很容易地实现自顶向下的姿态估计方法。但是，这类方法的性能受到人体检测结果的影响，运行速度往往也不是实时的。

2.1.1　人体检测对姿态估计性能的影响

首先，自顶向下方法是进行人体检测。在人体姿态估计中，最常用的人体检测器是基于 Faster R-CNN 结构，其是一种高性能检测器。Faster R-CNN 基于不同的基础网络和扩展结构，具有许多变体，这些变体具有不同的准确性、推断时间和计算复杂度。通常，检测结果越准确，网络越复杂，此时应该考虑准确性、内存和时间之间的权衡。

大多数研究表明，用更好的人体探测器提高了人体姿态估计的精度，如图 2 所示。文献[18]的结果表明，在检测器性能较差的情况下，姿态估计器从较好的人体检测器获得了较大的增益。随着人体检测器平均精度的提高，人体姿态估计器的精度提高速度变慢。当人体检测器达更高精度时，姿态估计网络的精度则无法再提高。换句话说，人体检测器在性能一般时很重要，但在达到高性能时就不重要了。姿态估计器的增益随着更高的人体检测 AP 而非常小，尤其是当人体检测器已经足够精确时。



注：图中数据来自文献[18]。

图 2　行人检测 mAP 和关键点检测 mAP 间的关系

Fig. 2　Relationship of human detection mAP and keypoints mAP

2.1.2　NMS(非极大值抑制)

NMS 是一种常用的抑制冗余检测的方法。该技术可应用于自顶向下的人体姿态估计方法的 2 个阶段。对于人体检测，有 2 种 NMS 方法：标准 NMS 和 soft-NMS[21]。soft-NMS 在文献[12]中性能更好，同时具有与标准 NMS 相同的计算复杂度，这使得其成为一种改进人体检测的简单方法。文献[18]提出了一个基于 OKS 的 NMS，该方法考虑人类实例中关键点的相似性；文献[15]中提出的参数化姿态 NMS 是数据驱动的，这意味着所有的参数都是从数据中学来的，而不是手动设置的。该方法比文献[22]中提出的方法快很多，但比文献[18]中的 NMS 方法复杂得多。

2.2　自底向上的方法

自底向上姿态估计方法主要由人体关节检测和关节分组 2 部分组成。Deepcut[23] 使用了一种基于 Fast R-CNN 的身体部位检测器，首先检测出所有的身体部位，然后将每个部位标记为对应的部位类别，用整数线性规划，将这些部位组装成一个完整的个体。DeeperCut[24] 使用一种基于 ResNet 的更强身体部件检测器，用来探索候选关节对象之间几何外观约束的增量优化策略，从而改进了 DeepCut。Cao 等人[25] 使用 CPM 预测具有部分亲和力场(PAF)的所有身体关节候选对象。提出的 PAFs 可以编码肢体的位置和方向，将估计的关节组装成不同人的姿势。Nie 等人[26] 提出了一种姿态分割网络(PPN)，对关节分割进行联合检测和稠密回归，通过关节划分对关节构型进行局部推理。与 OpenPose 类似，Kreiss 等人[27] 设计了一个 PifPaf 网络，来预测部分强度场(PIF)和部分关联场(PAF)，来表示身体关节位置和身体关节关联。由于 PAF 的细粒度和 Laplace 损失函数的使用，该算法在低分辨率图像上运行良好。B. Cheng 等人[28] 利用高分辨率特征金字塔来学习尺度感知表示。该方法具有训练的多分辨率监控与推理的多分辨率聚合，能够更精确地解决多人姿态估计和定位关键点的尺度变化问题。

近年来，已有一些方法可以实现一次性预测。Newell 等人[29] 引入了一种单级深度网络架构，可以同时进行检测和分组。该网络可以生成每个关节的检测热图，以及包含每个关节的分组标签的关联嵌入图。Papandreou 等人[30] 提出了一种用于姿态估计和实例分割的无检测框多任务网络。该网络可以同步预测每个人所有关键点的关节热图和其之间的相对距离，并按照一种基于树结构运动图的贪婪解码过程进行分组。Kocabas 等人[31] 提出，结合多任务模型和一种新的分配方法来处理人体关键点估计，完成检测和语义分割任务。其主干网是 ResNet 和 FPN 的结合，具有关键点和个人检测子网的共享特性。A. Varamesh 等人[32] 设计了一个使用混合密度网络进行空间回归的框架，实现了对象检测和人体姿势估计的框架。

目前自底向上方法的处理速度非常快,有些方法可以实时运行。然而,复杂的背景和人体遮挡会对性能产生很大的影响。自顶向下的方法在几乎所有标准数据集上都取得了最先进的性能,但其处理速度也受到检测人数的限制。

2.2.1 热图生成方法

目前,有 3 种方法可用来生成热图:一是在每个关键点位置,通过二维高斯激活设置热图;二是将圆心为关键点,半径为 R(超参数)的圆心内所有位置的像素值设为 1,其它位置设为 0,当采用这种热图时,通过预测位置偏置图来更准确地定位关键点;三是生成一个二进制掩模。

2.2.2 关键点连接方法

在自下而上的方法中,关键点连接是一个重要的步骤。Deepcut[23] 使用 CNN 只是学习外观特征,使用其它手工定义的几何特征拟合 logistic 模型进行配对概率估计。然而,Deepercut[24] 将人工计算的特征改为由深度神经网络生成的学习特征,大大提高了 AP。2 种方法都对几何特征采用 logistic 模型来模拟成对关节的亲和力。PAFs[25] 和关联嵌入[22] 以深度学习的方式与热图同时学习。当涉及到将关节分组到人体实例时,其更加直接。这 2 种方法的性能比文献[23]提到的更好。这是因为深度神经网络的容量更大,并且直接从数据中学习,既可以捕捉局部特征,也可以捕捉全局背景。

3 数据集与评价指标

3.1 数据集

早期的数据集中包含的图片背景相对简单,图像数量太少,无法进行训练,并不适合基于深度学习的方法。基于深度学习方法中常用的数据集包括 MSCOCO、MPII、LSP、FLIC、PoseTrack 和 AI Challenger 等。其中,LSP 数据集中的图像来自体育活动场景,FLIC 数据集是从好莱坞电影中收集得到的。LSP 和 FLIC 数据集相对较小,只包含特定类型的活动。最新的数据集,如 MSCOCO 和 AI Challenger,在类别数量上则更丰富。

3.2 评价指标

不同的数据集具有不同的特征(例如,不同范围的人体尺寸、上身/全身)和不同的任务要求(单/多姿态估计),因此用于 2D 人体姿态估计的评估指标也有所不同。

(1)部位正确估计百分比(Percentage of Correct Parts,PCP):为早期姿态估计的评估指标,用于评估肢体的定位精度,若肢体的 2 个端点在相应真值端点的阈值内,则该肢体被正确定位;

(2)关节点正确定位百分比(Percentage of Correct Keypoints,PCK):评估人体关节点定位的准确率,若候选关节点落在真实关节点的阈值像素内,则该候选关节点是正确的;

(3)关节点平均精度(Average Precision of Keypoints,APK):通过 PCK 评估将预测的姿态分配给真值姿态后,由 APK 得出每个关节点定位准确的平均精度;

(4)对象关节点相似度(Object Keypoint Similarity,OKS):多人姿态估计评价指标,计算真值和所预测人体关节点的相似度。

4 结束语

在这篇综述中,对基于深度学习的二维人体姿态估计方法进行了总结和讨论。尽管当前的人体姿态估计方法已经有了显著的改进,但是为了更好的现实应用,仍然可以被改进。

关于算法速度问题:目前的算法速度仍然很慢,不能满足实时预测的要求,因此必须进一步探索加快检测速度。虽已有一些研究网络压缩和网络加速的工作,但其不是为人体姿态检测而设计的,与分类任务和检测任务相比,人体姿态检测需要更高分辨率的输出特征图。加速方法应进一步研究。

关于数据集问题:目前的数据集非常大,但姿态分布不平衡,还没有研究探索用不平衡数据集检测罕见姿态的方法。可能的改进包括做数据扩充和设计一个特殊的训练程序。

关于数遮挡问题:遮挡和自遮挡仍然给人体姿态估计带来挑战。一些工作结合了人类先验和数据驱动的方法来解决这个问题,但是其结果不够健壮。

本文讨论了 2014 年以来发表的基于深度学习的人体姿态估计方法的里程碑式研究成果,总结了基于深度学习的人体姿态估计的数据集和度量。希望读者能从调查分析中得到启发,解决上面提到的困难,够促进提升姿态估计速度、基于不平衡和未标记数据的数据增强、解决遮挡问题等研究领域的进步。

参考文献

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [J]. Advances in neural information processing systems, 2012, 25: 1097-1105.

[2] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. arXiv preprint arXiv:1506.01497, 2015.
[3] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 3431-3440.
[4] TOSHEV A, SZEGEDY C. Deeppose: Human pose estimation via deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2014: 1653-1660.
[5] CARREIRA J, AGRAWAL P, FRAGKIADAKI K, et al. Human pose estimation with iterative error feedback[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 4733-4742.
[6] SUN X, SHANG J, LIANG S, et al. Compositional human pose regression[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 2602-2611.
[7] LUVIZON D C, TABIA H, PICARD D. Human pose regression by combining indirect part detection and contextual information[J]. Computers & Graphics, 2019, 85: 15-22.
[8] CHEN X, YUILLE A. Articulated pose estimation by a graphical model with image dependent pairwise relations[J]. arXiv preprint arXiv:1407.3399, 2014.
[9] CHEN Y, SHEN C, WEI X S, et al. Adversarial posenet: A structure-aware convolutional network for human pose estimation[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 1212-1221.
[10] WEI S E, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016: 4724-4732.
[11] NEWELL A, YANG K, DENG J. Stacked hourglass networks for human pose estimation[C]//European conference on computer vision. Springer, Cham, 2016: 483-499.
[12] CHU X, YANG W, OUYANG W, et al. Multi-context attention for human pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1831-1840.
[13] MARTINEZ J, HOSSAIN R, ROMERO J, et al. A simple yet effective baseline for 3d human pose estimation[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 2640-2649.
[14] IQBAL U, GALL J. Multi-person pose estimation with local joint-to-person associations[C]//European Conference on Computer Vision. Springer, Cham, 2016: 627-642.
[15] FANG H S, XIE S, TAI Y W, et al. Rmpe: Regional multi-person pose estimation[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 2334-2343.
[16] HUANG S, GONG M, TAO D. A coarse-fine network for keypoint localization[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 3028-3037.
[17] XIAO B, WU H, WEI Y. Simple baselines for human pose estimation and tracking[C]//Proceedings of the European conference on computer vision (ECCV), 2018: 466-481.
[18] SARAFIANOS N, BOTEANU B, IONESCU B, et al. 3d human pose estimation: A review of the literature and analysis of covariates[J]. Computer Vision and Image Understanding, 2016, 152: 1-20.
[19] Moon G, Chang J Y, Lee K M. Posefix: Model-agnostic general human pose refinement network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 7773-7781.
[20] WANG M, TIGHE J, MODOLO D. Combining detection and tracking for human pose estimation in videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11088-11096.
[21] BODLA N, SINGH B, CHELLAPPA R, et al. Soft-NMS--improving object detection with one line of code[C]//Proceedings of the IEEE international conference on computer vision, 2017: 5561-5569.
[22] BURGOS-ARTIZZU X P, HALL D C, PERONA P, et al. Merging pose estimates across space and time[C]//in British Machine Vision Conference (BMVC), Bristol, UK, 2013.
[23] PISHCHULIN L, INSAFUTDINOV E, TANG S, et al. Deepcut: Joint subset partition and labeling for multi person pose estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 4929-4937.
[24] INSAFUTDINOV E, PISHCHULIN L, ANDRES B, et al. Deepercut: A deeper, stronger, and faster multi-person pose estimation model[C]//European Conference on Computer Vision. Springer, Cham, 2016: 34-50.
[25] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2d pose estimation using part affinity fields[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 7291-7299.
[26] NIE X, FENG J, XING J, et al. Pose partition networks for multi-person pose estimation[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 684-699.
[27] KREISS S, BERTONI L, ALAHI A. Pifpaf: Composite fields for human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 11977-11986.
[28] CHENG B, XIAO B, WANG J, et al. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 5386-5395.
[29] NEWELL A, HUANG Z, DENG J. Associative embedding: End-to-end learning for joint detection and grouping[J]. arXiv preprint arXiv:1611.05424, 2016.
[30] 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]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 269-286.
[31] KOCABAS M, KARAGOZ S, AKBAS E. Multiposenet: Fast multi-person pose estimation using pose residual network[C]//Proceedings of the European conference on computer vision (ECCV), 2018: 417-433.
[32] VARAMESH A, TUYTELAARS T. Mixture dense regression for object detection and human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 13086-13095.