计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(10):36–43 [doi: 10.15888/j.cnki.csa.008711]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于深度学习的二维人体姿态估计算法综述①

马双双, 王　佳, 曹少中, 杨树林, 赵　伟, 张　寒

(北京印刷学院 信息工程学院, 北京 102600)
通信作者: 王　佳, E-mail: wangjia@bigc.edu.cn

摘　要: 二维人体姿态估计作为人体动作识别的基础, 随着深度学习和神经网络的流行已经成为备受学者关注的研究热点. 与传统方法相比, 深度学习能够得到更深层图像特征, 对数据的表达更准确, 因此已成为研究的主流方向. 本文主要介绍了二维人体姿态估计算法, 首先根据检测人数分为单人姿态估计与多人姿态估计两类, 其次对单人姿态估计分为基于坐标回归与基于热图检测的方法; 对多人姿态估计可分为自顶向下 (top-down) 和自底向上 (bottom-up) 的方法. 最后介绍了姿态估计常用数据集以及评价指标对部分多人姿态估计算法的性能指标进行了对比, 并对人体姿态估计研究所面临的问题与发展趋势进行了阐述.

关键词: 深度学习; 卷积神经网络; 人体姿态估计; 关键点检测

引用格式: 马双双,王佳,曹少中,杨树林,赵伟,张寒.基于深度学习的二维人体姿态估计算法综述.计算机系统应用,2022,31(10):36–43. http://www.c-s-a.org.cn/1003-3254/8711.html

Overview on Two-dimensional Human Pose Estimation Methods Based on Deep Learning

MA Shuang-Shuang, WANG Jia, CAO Shao-Zhong, YANG Shu-Lin, ZHAO Wei, ZHANG Han

(School of Information Engineering, Beijing Institute of Graphic Communication, Beijing 102600, China)

Abstract: As the basis of human motion recognition, two-dimensional human pose estimation has become a research hotspot with the popularity of deep learning and neural networks. Compared with traditional methods, deep learning can achieve deeper image features and express the data more accurately, thus becoming the mainstream of research. This study mainly introduces two-dimensional human pose estimation algorithms. Firstly, according to the number of people detected, the algorithms are divided into two categories for single-person and multi-person pose estimation. Secondly, the single-person pose estimation methods are divided into two groups based on coordinate regression and heat map detection. Multi-person poses can be estimated by top-down and bottom-up methods. Finally, the study introduces commonly used data sets and evaluation indexes of human pose estimation and compares the performance indexes of some multi-person pose estimation algorithms. It also expounds on the challenges and development trends of human pose estimation.

Key words: deep learning; convolutional neural networks (CNN); human pose estimation; key-point detection

人体姿态估计是计算机视觉领域一个基础问题, 解决这个问题是图像和视频中识别人类行为的重要步骤, 主要内容是从图像中识别身体的各个部分, 并计算其方向和位置信息. 人体姿态估计作为解决图像和视频中人体关键点 (如头部、肩部、肘部等) 坐标的重要技术, 其流行与发展得到了众多学者的广泛关注. 深度学习与卷积网络的不断发展, 人体姿态估计在动作识别[1]、动作捕捉[2]、姿态追踪[3]、手势识别[4]、图像生

① 基金项目: 北京市自然基金和北京市教委联合项目 (KZ202010015021); 北京印刷学院科研项目 (Ec202002, Eb202103); 北京印刷学院博士启动基金 (27170120003/021); 北京市教育委员会科研计划 (KM201910015003, KM201610015001)
收稿时间: 2021-12-20; 修改时间: 2022-01-18; 采用时间: 2022-02-17; csa 在线出版时间: 2022-06-24

成[5]、人机交互[6] 等方面得到了广泛应用.

人体姿态估计算法发展至今可以分为传统方法和深度学习的方法. 传统方法采用手工提取特征建立模型, 一般是基于图结构 (pictorial structures) 模型[7] 和基于形变部件模型[8], 由于遮挡严重、光线条件差和拍摄角度不同, 因此具有挑战性. 它们的准确性受到限制, 特别是在严重遮挡和复杂光照条件下. 近几年人工智能发展迅速, 学者将目光专注于研究深度学习模型, 比如深度卷积神经网络 (convolutional neural networks, CNN)[9]、生成对抗网络 (generative adversarial nets, GANs)[10]、递归神经网络等. 在图像分割、图像分类、图像融合、图像识别等领域获得了显著成果. 人体姿态估计采用深度学习的方法可以利用 CNN 提取到更加准确的特征, 有利于获取人体关节点之间的联系.

二维人体姿态估计是在图像中识别出人体关键点, 将关键点按顺序连接形成人体骨骼图. 本文主要从单人目标和多人目标两个方向对二维姿态估计进行梳理和分析, 整理了相关数据集与评价指标, 并对当前所面临的问题和未来发展趋势进行了阐述.

1 传统算法

传统方法主要用于解决姿态估计问题, 大部分采用模板匹配的方法. 基于 Fischler 等人[7] 提出的图结构模型, 首先人体部件检测器将人或物体表示为多个部件, 并使用图形模型确定部件之间的连通性. 2005 年 Felzenszwalb 等人[11] 提出了一个统计框架, 用于表示可变形结构中对象的视觉外观, 它允许对外观进行定性描述, 并假设组件与树形结构一致.

文献 [12] 提出图结构主要由表述人体部件的局部模型 (part model) 和表述空间关系的空间模型 (spatial model) 构成. 为改善局部模型表现能力差的缺点, 使用了表现力更强的图像特征, 例如 HOG 特征[13] 和 SIFT 特征[14]. 韩贵金等人[15] 提出一种基于 HOG 和颜色特征融合的外观模型, 用于图像中人体上半身的姿态估计. 前景技术可以应用到姿态估计中[16], 也可以将判别能力更强的检测器来提高姿态估计准确性[17]. 人体姿态估计会存在肢体遮挡的问题, 为解决此类问题非树形结构的空间模型被提出[18]. 传统方法已拥有较高的效率, 但无法提取图像中的充分信息并加以利用, 使得适用方法范围受到限制, 并且由于传统方法依赖于专业的摄影设备, 成本较高, 无法使用所有的应用场景.

2 基于深度学习的方法

在近几年, 受到以端到端为特征的图像识别的影响, 越来越多的研究人员引入深度学习的人体姿态估计模型, 并不断提高模型的性能. 深度学习通过训练大量的样本数据, 获取更加高效准确的特征. 相较于传统方法, 深度学习的方法鲁棒性更强、泛化能力更好. 自 2014 年首次引入深度学习以来, 基于深度学习的人体姿态估计已成为一个研究学者的主流研究领域. 根据应用场景可将二维人体姿态估计分为单人姿态估计和多人姿态估计, 二维人体姿态估计分类如图 1 所示.

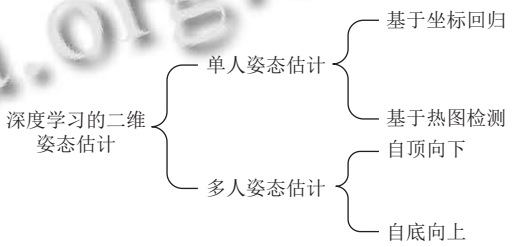

图 1 二维人体姿态估计分类

2.1 单人姿态估计

单人姿态估计作为人体姿态估计的基础尤为重要, 图像里只有单个待检测目标, 首先检测出目标的边界框图像, 在检测出目标人体的所有关节点. 大多数单人姿态估计都使用有监督的方法, 可按照真值 (ground truth) 分为基于坐标回归与基于热图检测.

2.1.1 基于坐标回归

2014 年 Toshev 等人提出的 DeepPose[19] 首先将深度学习应用在人体姿态估计领域, 它将 2D 人体姿态估计问题由原本的图像处理和模板匹配问题转化为卷积神经网络图像特征提取和关键点坐标回归问题, 它将 2D 人体姿态估计问题由图像处理和模板匹配问题转化为 CNN 图像特征提取和关键点坐标回归问题, 使用回归准则来估计被遮挡的人体关节点. 其思路是针对 CNN 学习到的特征尺度固定、回归性能差的问题, 在网络得到粗分回归的基础上增加一个阶段, 将特征图像传入 CNN 网络学习高分辨率的特征, 进行较高精度的坐标值回归. 具体 DeepPose 流程图如图 2 所示.

Geng 等人[20] 认为回归关键点坐标的特征必须集中注意关键点周围的区域才能精确回归出关键点坐标, 提出了直接坐标回归方法解构式关键点回归 (DEKR). 使用自适应的卷积激活关键点区域周围的像素, 利用这些激活的像素去学习新的特征, 并利用多分支结构,

每个分支都会针对某种关键点利用自适应卷积学习关键点周围的像素特征, 回归关键点的位置.

多阶段回归可更加精确地反映关键点坐标, 改善多阶段直接回归方法. Carrira 等人[21] 提出了自我修正模型, 通过从输入到输出的联合空间学习特征提取器, 对联合空间中丰富的结构化信息进行建模. 文章引入了自顶向下的反馈机制, 通过反馈错误预测逐步改变初始解的自校正模型, 此过程称为迭代错误反馈 (IEF).

基于坐标回归的方法, 只减少了每个关节点位置的误差, 忽略了关节点之间的相关信息, 相比于关节点骨骼信息更准确. Sun 等人[22] 提出了一种基于 ResNet-50[23] 的结构感知回归方法, 它采用重新参数化的姿势表示, 使用骨骼进行姿态表示, 对姿势进行编码.

总体而言, 关节点坐标的直接回归是非线性的, 在映射学习中存在困难, 而且不能应用于多人情况, 缺乏鲁棒性. 相较于坐标回归, 更多使用基于热图检测的方法.

图 2 DeepPose 网络结构

2.1.2 基于热图检测

热图检测的方法将人体各部位作为检测目标, 通过检测关键点热力图 (heatmap), 获得关键点的概率分布以及关键点的位置信息.

Tompson 等人[24] 采用深度卷积网络进行姿态估计, 采用 heatmap 的方式回归关键点, 将重叠感受野和多分辨率输入, 利用人体关键点之间的空间信息, 结合马尔科夫随机场的思想来优化预测结果. 该方法也为多人场景下的姿态估计中关键点聚类问题提供思路. 针对于定位的精度较低的问题, Tompson 等人[25] 在此基础上做了相应改进, 使用两个级联网络来回归人体关键点的热图, 并联合训练这两个网络, 提升模型的泛化能力.

Isack 等人[26] 提出高效轻量级模型 RePose, 将基于部件的结构和几何先验合并到分层预测框架, 利用人体运动学约束, 采用端到端的训练, 根据先验知识进行建模, 传播低分辨率特征以达到细化预测的姿势信息的目的.

Artacho 等人[27] 基于“瀑布式”的空间池架构, 提出了统一的人体姿态估计框架 UniPose, 将空洞卷积的级联方法和空洞空间金字塔模块并行. 该方法结合上下文分割和联合定位来确定关键点位置和人体边界框, 以实现人体姿势的高精度估计.

基于坐标回归的方法获取关键点信息更加直接, 能够获取丰富的特征, 但增加了复杂度, 通用性低, 精度低. 基于热图检测的方法相较于坐标回归的方法鲁棒性更好, 关节点之间的关联更加清晰, 但计算量较大, 效率低, 基于坐标回归与热图检测的方法对比如表 1 所示.

表 1 单人姿态估计方法对比

类别		优点	缺点
基于坐标回归	直接回归坐标	模型简单; 时间效率高	通用性较低; 准确率较低
	多阶段回归坐标	准确率较直接回归有提升	通用性较低
基于热图检测		准确率高; 关节连接关系逻辑清晰	网络结构复杂; 时间效率低

2.2 多人姿态估计

与单人姿态估计不同, 多人姿态估计需要检测出图像中的所有目标人体, 包含检测和定位步骤. 多人姿态估计根据检测步骤分为自顶向下 (top-down) 和自底向上 (bottom-up), top-down 的方法先检测人体目标, 在对人体进行姿态估计; bottom-up 的方法先检测图像中的所有关节点, 再将关节点进行聚类组合成人体. 同时, 多人图像场景可能会存在遮挡问题, 如何精确预测出遮挡情况下的关节点, 补齐缺失关键点是多人姿态估计中的一个重要研究方向.

2.2.1 Top-down

基于自顶向下的方法首先采用目标检测算法获取图像中的多个人体, 再对单个人体目标进行姿态估计. Iqbal 等人[28] 提出了一种多人姿态估计的方法, 利用 Faster R-CNN 进行人体目标检测, 对检测出的人体使用 convolutional pose machines (CPM) 网络进行姿态估计. 但是在对人体边界框进行姿态估计时, 并未考虑多

人图像中人体之间可能存在的遮挡情况, 有可能会使得关键点信息缺失无法与人体相关联, 从而导致姿态估计的误差降低准确度. Papandreou 等人[29] 基于复杂场景下, 没有提供人体的真实位置或比例的情况下, 提出了基于自顶向下简单有效的 G-RMI 多人姿态估计方法. 使用 Faster R-CNN 进行目标检测, 并估计目标框中包含的关节点. 对于关节点的类型, 使用全卷积 ResNet 预测关节点的热度图和偏移量. 引入热图-偏移的聚合方法来获得准确的关节点. Mask R-CNN[30] 首先检测出目标边界框, 通过特征图进行关节点检测. Mask R-CNN 的网络结构在 Faster R-CNN 分类和回归的基础上增加了一个分支进行图像的语义分割, DensePose 借用了 Mask R-CNN 的架构.

AlphaPose 由 Fang 等人[31] 提出, 此研究认为虽然当前最先进的人体检测已经达到较好的效果, 但人体目标的定位和识别仍会产生误差, 提出了区域多人姿态估计 (RMPE) 框架, 由空间变换网络 (SSTN)、参数姿态非最大抑制 (NMS) 和姿势引导区域生成器 (PGPG) 组成. SSTN 主要作用是在不精准的边界框中提取出高质量的人体区域, NMS 用来解决人体目标被重复检测的问题, 使用 PGPG 来进行数据增强, 根据检测结果生成的训练样本. AlphaPose 利用 RMPE 框架对不准确的人体目标边界框进行准确的姿态估计, 减少了因为人体目标检测不准确而导致的误检. 文献 [32] 提出一种用于人体姿态估计的无偏的数据处理方法 (UDP), 以减少训练和推理过程中的计算增量.

HRNet[33] 在 2019 年被提出, 主要是为保持高分辨率的特征图信息, 现有方法大多是从低分辨率特征中恢复高分辨率特征, HRNet 通过并行化多分辨率子网络保持高分辨率特征, 并通过多尺度融合来增强高分辨率特征. Zhang 等人[34] 在 HRNet 的基础上提出了一种新型的注意力模块, 去规范化注意力 (DNA) 来解决传统注意力模块的特征衰减问题.

总体而言, 自顶向下的方法思路清晰, 精度较高, 在检测人体边界框时不会出现漏检、误检; 但实时性较差, 对于每次检测, 都要运行单人姿态估计, 检测的人数越多, 计算成本越高. 虽然相较于先前的方法检测精度得到了很大提升, 但发生检测错误还是不可避免的, 比如边界框定位错误, 会阻碍自顶向下方法精度的提高.

2.2.2 Bottom-up

基于自底向上的方法步骤包含关节点检测和聚类, 首先检测出图像中的所有关节点, 通过相应策略将关节点聚类成人体, 实现姿态估计. 自底向上的方法摆脱了首先对个体进行进行检测的前提.

Pishchulin 等人[35] 提出了基于 Fast R-CNN 检测器的 DeepCut, 首先提取图像中的所有关键点, 将关键点作为节点组成密集连接图, 将同一个体的关键点采用非极大值抑制聚类为完整个体. Insafutdinov 等人[36] 改进 DeepCut 提出了基于 ResNet 的 DeeperCut. 该算法采用 ResNet 来获取人体关节点, 提升检测精度; 提出图像条件成对项 (ICPT) 减少候选区域的关节点, 减少网络的计算量.

DeeperCut 相比于 DeepCut, 精确度提升了, 并且减少了运行时间, 从时间效率依旧无法达到实时检测. 为了提高实时检测效率, Cao 等人[37] 提出了基于 CPM 的 OpenPose 方法, OpenPose 的网络模型如图 3 所示.

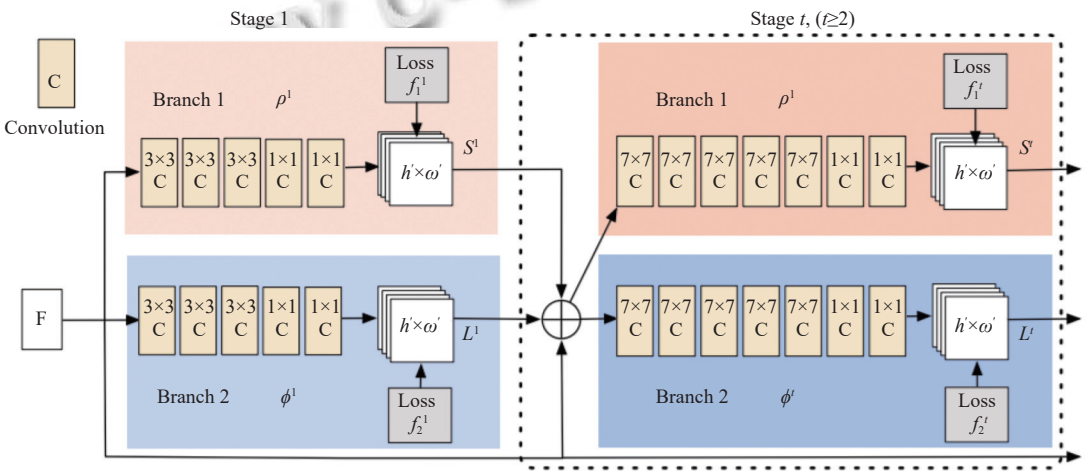

图 3 OpenPose 网络架构图

该方法利用 VGG-19[38] 的前 10 层为输入图像创建特征映射, 网络框架分为两个并行分支, 一个分支预测关节点的置信度; 另一分支预测部分亲和域场 (PAFs), PAF 表示部件之间的关联程度; 利用匈牙利算法进行最优化匹配将同一个体的关节点进行聚类, 得到人体姿态信息.

Osokin[39] 改进 OpenPose 提出了 Lightweight OpenPose, 使用 MobileNet v1[40] 代替 VGG-19 进行特征提取, 通过权重共享来减少计算量, 为解决感受野较小而造成的效果不佳采用空洞卷积优化算法. Kreiss 等人[41] 提出了与 OpenPose 相似的 PiPaf 网络, 主要包含部分强度场 (PIF) 和部分关联场 (PAF), 分别提升热图在高分辨率下的精度和确定关节点的连接, 得到人体关节点, 与 OpenPose 相比性能有明显提升, 该算法适用于低分辨率图像. 针对高分辨率网络, Cheng 等人[42] 在高分辨率网络 HRNet 基础上提出了更高分辨率网络 (HigherHRNet), 提出了一种高分辨率特征金字塔, 通过反卷积得到更高分辨率的特征来提高准确度, 使用多分辨率监督让不同层的特征能学习不同尺度的信息, 解决多人姿态估计中的尺度变化. Luo 等人[43] 为解决人体尺度的变化和人体关键点标签的模糊这两大挑战, 提出了尺度自适应热图回归 (SAHR) 方法和权重自适应热图回归 (WAHR) 方法共同作用以提高人体姿态估计的准确性. Varamesh 等人[44] 设计了一种使用混合密度网络进行空间回归的框架, 提高对象检测和人体姿态估计的速度和精度.

目前已经有方法可以实现预测. Newell 等人[45] 提出了关联嵌入标签算法, 应用在监督学习卷积神经网络中, 可以同时检测和分组. Papandreou 等人[46] 提出了多任务网络 PersonLab, 使用模型对多人图像中的人体进行关键点检测和实例分割.

与自顶向下的方法相比, 自底向上的方法受人数增加影响较小, 处理速度较快. 但复杂背景和人体遮挡情况会对性能产生较大影响. 在复杂的背景和遮挡干扰情况下, 缺失人体关节点在将关节点聚类到不同个体上时可能会出现误判、匹配错误等问题, 如何处理背景干扰和遮挡情况是将来研究的重点和难点. 多人姿态估计方法对比如表 2 所示.

表 2 多人姿态估计方法对比

类比	优点	缺点	算法
自顶向下	关节点定位精度较高; 思路清晰; 减少漏检、误检	内存需求较大; 实时性较差; 计算成本高	CPM、G-RMI、AlphaPose、HRNet
自底向上	受图像中人数影响小; 实时性较高	复杂背景、遮挡情况影响较大, 容易出现误判、匹配错误等问题	DeepCut、DeeperCut、OpenPose、PiPaf、HigherHRNet

3 数据集与评价指标

3.1 数据集

目前主流的人体姿态估计数据集可分为单人数据集和多人数据集, 单人数据集包含 LSP[47]、FLIC[48], 多人数据集包含多人数据集 COCO[49]、MPII[50]、AI Challenger[51]、PoseTrack[52]. 表 3 对各个数据集的样本数目、类型、关节点数目以及来源场景进行对比.

LSP 数据集是一个体育姿势数据集, 收录的运动场景下的人体图像, 图像中只包含一个人体, 定义了 14 个关节点, 样本数大约 2 000 张, 图像大部分与体育有关, 该数据集中人体姿势较复杂. FLIC 数据集来源于好莱坞电影片段, 人工对电影片段截图的图像进行标注, 图像中包含多人时, 只对一个人的关节点进行标注, 此数据集不包含人体被遮挡或者清晰度过低的图像. COCO 数据集由微软构建, 来源于谷歌、Flicker 等下载的图像, 图像分为训练集、验证集和测试集, 定义了 17 个关节点, 包含 20 万张图像和 25 万个被标注的人体. MPII 数据集来源于 YouTube 的日常生活场景, 手动检测包含人的画面. 该数据集包含 2.5 万张图像, 定义了 16 个关节点, 标注了 4 万个人体目标. AI Challenger 数据集来源于网络爬取的日常片段, 包括训练集、验证集、测试集共 30 万张图像.

表 3 人体姿态估计数据集

数据集	样本数 (万)	类型	关节点数目	来源场景
LSP	0.2	单人	14	运动场景
FLIC	2	单人	10	好莱坞电影截取片段
COCO	20	多人	17	谷歌、Flicker等下载的图片
MPII	2.5	单人, 多人	16	YouTube的日常生活场景
AI Challenger	30	多人	14	网络爬取的日常片段
PoseTrack	0.05	多人	15	对MPII扩展

3.2 评价指标

不同数据集因自身特点采用的评估指标也不同. 常用的二维人体姿态估计指标主要有以下几种:

(1) 部位正确估计百分比 (*PCP*): 关节点正确估计的比例, 用于评估人体关节点的定位精度.

(2) 目标关节点相似度 (*OKS*): 计算关节点位置距离, 检测关节点的相似度. *OKS* 的计算方式为:

$$OKS = \frac{\sum_i \left[\exp\left(-\frac{d_i^2}{2s^2k_i^2}\right)\delta(v_i>0)\right]}{\sum_i \delta(v_i>0)} \tag{1}$$

其中, i为标注的关节点编号; d_i^2为检测到的关节点位置与真实关节点位置的欧氏距离的平方; s^2为检测人体在图像中面积; k_i^2为归一化因子表示标注关节点位移的标准差; v_i为正整数是可见关节点.

(3) 平均精度 *AP* (average precision): 每一个关节点在整个测试数据集上, 检测结果的平均准确率:

$$AP@s = \frac{\sum_p \delta(OKS > S)}{\sum_P 1} \tag{2}$$

其中, p 为人体检测框编号. AP_{50}、AP_{75}为交并比 (intersection over union) 分别取值为 0.5、0.75 时 *AP* 的值, AP_M、AP_L分别为中等目标和大目标的 *AP* 值.

(4) 关节点正确定位百分比 (*PCK*): 用于评估关节点定位的准确度, 检测关节点在标注关节点的阈值内, 则该关节点为准确的.

(5) 关节点平均精度 (*APK*): 将预测的人体姿态与真实姿态评估后, 通过 *APK* 得出每个关节点定位准确的平均精度.

表 4 列出了多人姿态估计部分算法在 COCO 数据集上 *AP* 的性能对比.

4 发展趋势及难点

深度学习和卷积神经网络的飞速发展, 使得人体姿态估计领域不断前进, 在计算机视觉领域突出重要性和发展前景已被学者认可, 但依旧存在一些难点与挑战.

(1) 提高检测精度和效率, 虽然有些算法已经取得了较大的进步, 但是真正将人体姿态估计应用在无人驾驶、监控检测等领域还需要更高检测精度的算法, 需要简化网络结构, 文献 [53] 提出可采用轻量级的网络优化姿态估计算法, 保证精度的同时提高效率.

(2) 算法受复杂环境影响较大, 在实际应用中光照和遮挡情况容易对算法效率产生影响, 重叠和遮挡的关节点会导致关节点的误检和漏检. 另一方面人体在不同视角会产生信息压缩的情况, 例如仰视或俯视条件会导致无法获取到正确人体比例. 因此如何解决遮挡问题是重要研究方向, 文献 [54] 提出对于肢体遮挡修复算法的研究非常重要, 文献 [55] 提出研究姿态连续性信息, 可以还原姿态失真.

(3) 数据集分布不均匀, 目前常用数据集足够大, 但分布不平衡, 现有数据集无法对罕见姿态进行检测, 难以满足人体姿态变化复杂与多样性, 例如存在遮挡情况、角度压缩的数据集较少, 丰富扩充数据集样本仍然是人体姿态估计研究的重点.

表 4 多人姿态估计算法在 COCO 数据集上的性能对比

算法	*AP*	AP_{50}	AP_{75}	AP_M	AP_L
G-RMI	60.5	82.2	66.2	57.6	66.6
Mask R-CNN	63.1	87.3	68.7	57.8	71.4
AlphaPose	61.8	83.7	69.8	58.6	67.6
HRNet	76.3	90.8	82.9	72.3	83.4
OpenPose	61.8	84.9	67.5	57.1	68.2
PifPaf	66.7	—	—	62.4	72.9
SAHR+WAHR	68.9	—	—	63.0	77.5
HigherHRNet	70.5	89.3	75.4	64.1	75.5
Associative Embedding	65.5	86.8	72.3	60.6	70.2
PersonLab	68.7	89.0	75.4	64.1	75.5

5 总结

人体姿态估计由传统方法发展至深度学习的方法, 模型和算法性能不断得到优化和提升, 人体姿态估计在电影动画、无人驾驶、虚拟现实和智能监控等方面都取得了丰硕的研究成果. 基于图结构的传统方法可为后续的算法研究提供先验知识, 基于深度学习的人体姿态估计方法必然是未来的发展方向. 在当前大量图像数据的背景下, 应当充分利用视频数据, 将人体姿态估计应用于更多领域. 二维姿态估计作为计算机视觉众多任务的基础, 具有广阔的研究前景.

参考文献

1 王新文, 谢林柏, 彭力. 跌倒异常行为的双重残差网络识别方法. 计算机科学与探索, 2020, 14(9): 1580–1589. [doi: 10.3778/j.issn.1673-9418.1906054]

2 Nie BX, Xiong CM, Zhu SC. Joint action recognition and pose estimation from video. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston:

IEEE, 2015. 1293–1301.
3 Cho NG, Yuille AL, Lee SW. Adaptive occlusion state estimation for human pose tracking under self-occlusions. Pattern Recognition, 2013, 46(3): 649–661. [doi: 10.1016/j.patcog.2012.09.006]
4 宋一凡, 张鹏, 刘立波. 基于视觉手势识别的人机交互系统. 计算机科学, 2019, (S2): 570–574.
5 黄友文, 赵朋, 游亚东. 融合反馈机制的姿态引导人物图像生成. 激光与光电子学进展, 2020, 57(14): 141011.
6 Shotton J, Sharp T, Kipman A, *et al*. Real-time human pose recognition in parts from single depth images. Communications of the ACM, 2013, 56(1): 116–124. [doi: 10.1145/2398356.2398381]
7 Fischler MA, Elschlager RA. The representation and matching of pictorial structures. IEEE Transactions on Computers, 1973, C-22(1): 67–92. [doi: 10.1109/T-C.1973.223602]
8 Yang Y, Ramanan D. Articulated pose estimation with flexible mixtures-of-parts. Computer Vision and Pattern Recognition. Colorado Springs: IEEE, 2011. 1385–1392.
9 LeCun Y, Boser B, Denker JS, *et al*. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1989, 1(4): 541–551. [doi: 10.1162/neco.1989.1.4.541]
10 Goodfellow I. NIPS 2016 tutorial: Generative adversarial networks. arXiv: 1701.00160, 2016.
11 Felzenszwalb PF, Huttenlocher DP. Pictorial structures for object recognition. International Journal of Computer Vision, 2005, 61(1): 55–79. [doi: 10.1023/B:VISI.0000042934.15159.49]
12 冯晓月, 宋杰. 二维人体姿态估计研究进展. 计算机科学, 2020, 47(11): 128–136. [doi: 10.11896/jsjkx.200700061]
13 Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 886–893.
14 Lowe DG. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91–110. [doi: 10.1023/B:VISI.0000029664.99615.94]
15 韩贵金, 朱虹. 基于 HOG 和颜色特征融合的人体姿态估计. 模式识别与人工智能, 2014, 27(9): 769–777. [doi: 10.3969/j.issn.1003-6059.2014.09.001]
16 Nägeli T, Oberholzer S, Plüss S, *et al*. Flycon: Real-time environment-independent multi-view human pose estimation with aerial vehicles. ACM Transactions on Graphics, 2018, 37(6): 182.
17 Achilles F, Ichim AE, Coskun H, *et al*. Patient MoCap: Human pose estimation under blanket occlusion for hospital monitoring applications. Proceedings of the 19th International Conference on Medical Image Computing and Computer-assisted Intervention. Athens: Springer, 2016. 491–499.
18 Wang JB, Qiu K, Peng HW, *et al*. AI coach: Deep human pose estimation and analysis for personalized athletic training assistance. Proceedings of the 27th ACM International Conference on Multimedia. Nice: ACM, 2019. 2228–2230.
19 Toshev A, Szegedy C. DeepPose: Human pose estimation via deep neural networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2013. 1653–1660.
20 Geng ZG, Sun K, Xiao B, *et al*. Bottom-up human pose estimation via disentangled keypoint regression. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 14671–14681.
21 Carreira J, Agrawal P, Fragkiadaki K, *et al*. Human pose estimation with iterative error feedback. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2015. 4733–4742.
22 Sun X, Shang JX, Liang S, *et al*. Compositional human pose regression. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 2621–2630.
23 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 770–778.
24 Tompson J, Jain A, LeCun Y, *et al*. Joint training of a convolutional network and a graphical model for human pose estimation. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 1799–1807.
25 Tompson J, Goroshin R, Jain A, *et al*. Efficient object localization using convolutional networks. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 648–656.
26 Isack H, Haene C, Keskin C, *et al*. RePose: Learning deep kinematic priors for fast human pose estimation. arXiv: 2002.03933, 2020.
27 Artacho B, Savakis A. UniPose: Unified human pose estimation in single images and videos. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 7033–7042.
28 Iqbal U, Gall J. Multi-person pose estimation with local joint-to-person associations. European Conference on Computer Vision. Amsterdam: Springer, 2016. 627–642.
29 Papandreou G, Zhu T, Kanazawa N, *et al*. Towards accurate multi-person pose estimation in the wild. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 3711–3719.
30 He KM, Gkioxari G, Dollár P, *et al*. Mask R-CNN. 2017 IEEE International Conference on Computer Vision (ICCV).

Venice: IEEE, 2017. 2980–2988.
31 Fang HS, Xie SQ, Tai YW, *et al*. RMPE: Regional multi-person pose estimation. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 2353–2362.
32 Huang JJ, Zhu Z, Guo F, *et al*. The devil is in the details: Delving into unbiased data processing for human pose estimation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 5699–5708.
33 Sun K, Xiao B, Liu D, *et al*. Deep high-resolution representation learning for human pose estimation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 5686–5696.
34 Zhang K, He P, Yao P, *et al*. DNANet: De-normalized attention based multi-resolution network for human pose estimation. arXiv: 1909.05090, 2019.
35 Pishchulin L, Insafutdinov E, Tang SY, *et al*. DeepCut: Joint subset partition and labeling for multi person pose estimation. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 4929–4937.
36 Insafutdinov E, Pishchulin L, Andres B, *et al*. DeeperCut: A Deeper, stronger, and faster multi-person pose estimation model. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 34–50.
37 Cao Z, Simon T, Wei SE, *et al*. Realtime multi-person 2D pose estimation using part affinity fields. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 1302–1310.
38 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015. 1–14.
39 Osokin D. Real-time 2D multi-person pose estimation on CPU: Lightweight OpenPose. Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods. Prague: SciTePress, 2019. 744–748.
40 Howard AG, Zhu ML, Chen B, *et al*. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv: 1704.04861, 2017.
41 Kreiss S, Bertoni L, Alahi A. PifPaf: Composite fields for human pose estimation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 11969–11978.
42 Cheng BW, Xiao B, Wang JD, *et al*. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 5385–5394.
43 Luo ZX, Wang ZC, Huang Y, *et al*. Rethinking the heatmap regression for bottom-up human pose estimation. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2020. 13259–13268.
44 Varamesh A, Tuytelaars T. Mixture dense regression for object detection and human pose estimation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 13083–13092.
45 Newell A, Huang ZA, Deng J. Associative embedding: End-to-end learning for joint detection and grouping. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. Long Beach: NIPS, 2017. 2277–2287.
46 Papandreou G, Zhu T, Chen LC, *et al*. PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 282–299.
47 Johnson S, Everingham M. Clustered pose and nonlinear appearance models for human pose estimation. British Machine Vision Conference. Aberystwyth: British Machine Vision Association, 2010. 1–11.
48 Sapp B, Taskar B. MODEC: Multimodal decomposable models for human pose estimation. 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 3674–3681.
49 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014. 740–755.
50 Andriluka M, Pishchulin L, Gehler P, *et al*. 2D Human pose estimation: New benchmark and state of the art analysis. 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 3686–3693.
51 Wu JH, Zheng H, Zhao B, *et al*. AI challenger: A large-scale dataset for going deeper in image understanding. arXiv: 1711.06475, 2017.
52 Andriluka M, Iqbal U, Insafutdinov E, *et al*. PoseTrack: A benchmark for human pose estimation and tracking. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5167–5176.
53 周燕, 刘紫琴, 曾凡智, 等. 深度学习的二维人体姿态估计综述. 计算机科学与探索, 2021, 15(4): 641–657. [doi: 10.3778/j.issn.1673-9418.2008088]
54 田元, 李方迪. 基于深度信息的人体姿态识别研究综述. 计算机工程与应用, 2020, 56(4): 1–8. [doi: 10.3778/j.issn.1002-8331.1910-0445]
55 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. 计算机工程与应用, 2019, 55(19): 22–42. [doi: 10.3778/j.issn.1002-8331.1906-0113]

(校对责编: 孙君艳)