doi:10.3969/j.issn.1001-4616.2024.04.012

残差混合注意力结合骨骼图卷积多人姿态识别

陈　斌,樊飞燕,陆天易

(南京师范大学信息化建设管理处,江苏 南京 210023)

[**摘要**]　多人姿态识别研究起步晚,成熟度低,复杂性高,因此网络深度也随之加深,梯度消失问题也随之加剧,网络性能也随之衰减,由此造成识别精度差,识别效率低等共性问题. 为解决这些问题,本文提出了一种残差混合注意力结合骨骼图卷积多人姿态识别模型. 通过自顶向下的研究路径,运用预处理干预方式对多人体图像进行检测并对单人体坐标定位及框选标定,生成骨骼关键点架构图,借助残差块对网络结构进行改进以抑制梯度弥散,加载混合注意力机制对模型赋能增效. 在 MPII 及 MSCOCO2017 两个数据集上对本文提出的模型进行了验证,结果显示该模型对多人姿态识别效果较好,在两个数据集上分布稳定,差异微小. 同时,将本文模型与对本领域各类重要文献中记载模型综合能力进行了比较,结果表明在各项精细指标上本模型都有一定程度提升,稳定性较好,分布较为均匀. 本文提出的多人姿态识别模型在跨数据集基础上表现出较好的识别效果和效率,为多人姿态识别的研究增添了动力.

[**关键词**]　多人姿态识别,残差,混合注意力机制,骨骼关键点图,图卷积
[**中图分类号**]TP394.1　[**文献标志码**]A　[**文章编号**]1001-4616(2024)04-0106-12

Skelton-based Graph Convolution with Residual Combined with Mixed Attention Mechanism for Multi-Person Posture Recognition

Chen Bin, Fan Feiyan, Lu Tianyi

(Informatization Office, Nanjing Normal University, Nanjing 210023, China)

Abstract: The research of multi-person attitude recognition started lately, with low maturity and high complexity, so the network depth is also deepened, the problem of gradient vanishing is also intensified, and the network performance is also attenuated, resulting in the common problems of poor recognition accuracy and low recognition efficiency. To solve these problems, this paper proposes a model of skelton-based graph convolution with residual combined with mixed attention mechanism for multi-person posture recognition. Through the top-down research path, the pre-processing intervention was used to detect multi-body images and select the single body coordinate frame, and the bone key point architecture map was generated. With the residual block, the network structure was improved to suppress the gradient dispersion, and the mixed attention mechanism was loaded to enable and enhance the model. The proposed model is validated on two datasets, MPII and MSCOCO2017, and has stable distribution on the two datasets with small differences. At the same time, the model in this paper is compared with the comprehensive ability of the model recorded in various important literature in this field. In various fine indicators, the model has been improved to a certain extent, with good stability and uniform distribution. The multi-person pose recognition model proposed in this paper reflects the good recognition effect and efficiency based on the cross-data sets, and adds impetus to the study of multi-person gesture recognition.

Key words: multi-person posture recognition, residual, mixed attention mechanism, skeletal key point diagram, graph convolution

伴随着信息技术的高速发展,计算机在视觉领域的应用进入到了全新的阶段,其中人体姿态识别已成为当前非常重要的研究分支[1],在智能安防、人机交互、医疗辅助、运动康复及虚拟现实等方面都有着现实的研究意义和广阔的应用前景[2]. 针对单人姿态识别研究已经有较多的研究成果,不论是典型数据集

收稿日期:2024-07-18.
基金项目:江苏省现代教育技术研究 2023 年度智慧校园专项(2023-R-107311).
通讯作者:陈　斌,博士,高级工程师,研究方向:模式识别、机器学习、大数据分析. E-mail:60167@ njnu.edu.cn

还是经典模型算法都有着比较多的积累. 典型数据集的代表有 Kinetics 数据集[3]、NTU RGB+D 数据集[4]、KTH 数据集[5]、Hollywood 数据集[6]、UCF 数据集[7]、HMDB51 数据集[8]、HiEve 数据集[9]、AVA 数据集等[10]. 经典模型算法例如基于光流特征及密集加速鲁棒特征算法运动检测模型[11]、动作相关性关键姿态特征可迁移有限动作识别模型[12]、混合 SVM 及 K 近邻分类器动作识别模型[13]、基于注意力机制的全局上下文感知骨骼关键点识别模型[14]、基于目标检测及定位的融合长短期记忆网络行为识别模型[15]、基于姿态估计的人体异常行为识别模型[16]. 然而单纯的个体姿态识别应用场景非常有限,大多数场景都会涉及多人交互的情况,于是多人姿态识别研究成为了更受关注的焦点. 但受发展时间和研究难度的限制,针对多人姿态识别的数据集和成熟方法都较少.

在姿态识别研究领域,早在 2014 年,Toshev 等[17]就提出了基于人体姿态的深度卷积网络 DeepPose,成为人体姿态识别方向的创始之作. 2016 年,由 Lin 等[18]提出的一种沙漏形状的卷积神经网络结构 Hourglass 融入了多元尺度特征对人体关节点特征进行提取,其显著地提高了识别精度. 2017 年,在 Abdel-Basset 等[19]提出的一种自顶向下的目标检测方法中,借助了 Faster R-CNN 网络,针对多人姿态识别首次提出了将人体目标检测和人体关键点检测分步实现的思想,也就是后续自顶向下方法的由来. 在此基础上,他们又进一步提出了通过增加网络深度,以及利用残差模块借助卷积路径和跳级路径的方法获得高层级特征,对网络性能进行改善[20].

在姿态识别方向上,深度神经网络图卷积为重要手段,但随着神经网络深度的增加,梯度在反向传播过程中会逐渐消失,导致网络难以训练[21]. 通过引入残差块恒等映射或跳跃连接,允许网络直接学习输入和输出之间的差异,从而缓解了梯度消失的问题. 这种结构允许网络在增加深度的同时保持性能,甚至可能进一步提升性能,由此深度神经网络中的梯度消失问题被较好控制. 对于残差块的应用,本项目组在前期的表情识别领域已有较好研究成果[22]. 在姿态识别领域,Kocabas 等[23]于 2018 年提出的 MultiPoseNet 便使用了深度残差网络 ResNet 作为基础网络,并叠加两个特征金字塔网络头对人体检测框及人体关键点进行输出,再通过姿态残差网络对所有检测到的关键点按照分割后标定的检测框进行聚类处理,最终得到单人体关键点的集合. 秦晓飞等[24]于 2021 年提出一种基于深度残差网络的多人姿态估计,该算法利用已有的人体检测器,依靠 Simple Baseline 作为单人姿态估计的主干网络,对残差块进行改进,并引入多尺度监督模块和多尺度回归模块,利用多元的多尺度特征进行特征性匹配,使关键点定位精准度和鲁棒性都有所提升. 注意力机制让模型能够动态地关注输入数据中不同部分,允许模型按照任务需求聚焦于输入数据的重要部分,忽略次要部分. 该机制模拟了人类注意力选择性地专注于特定信息而忽略其他信息的认知过程,它对模型理解复杂数据起到很好的作用[25].

针对多人姿态识别中的精度及效率问题,本文提出一种残差混合注意力结合骨骼图卷积多人姿态识别模型. 通过自顶向下的方式,结合人体检测手段框选多人体,利用图卷积进行骨骼关键点特征提取以定位人体姿态结构,结合残差块改进网络结构以缓解深度网络的梯度消失情况,并集成混合注意力机制改进数据处理方式来提高模型的表示能力和效率,提升模型性能.

本文主要有以下研究贡献:(1)将残差机制与注意力机制有效结合起来,从克服网络深度和聚焦全局关键点两个方向上,对识别模型发挥促进作用. (2)在自顶而下检测框选基础上,结合项目组前期单人姿态骨骼图卷积能力的研究成果,对多人姿态进行识别处理,并借助残差注意力机制提升整体识别精度和性能.

1　基本原理及相关工作

1.1　多人姿态识别基本原理和典型方法

多人姿态估计算法主要分为自顶向下以及自底向上两种类型[26],两类方法有着各自不同的原理.

自顶向下姿态估计算法从定位图像中人物全局入手,再细化到身体部位,最终计算整体姿态. 其以目标检测为基础,主要过程分为两个阶段,第一阶段通过多人目标检测手段对图像中的人体进行检测并进行框选,第二阶段对框选出的多人图像中多个人体进行分割,形成多个单人姿态对象,继而对单人图像的关键点检测并提取,最终完成每个包围框的人体姿态估计目标,进而完成多人姿态估计[27]. 该方法受遮挡和重叠等问题的影响较小,对图像分辨率依赖小,可以此提高姿态估计准确率,但其受包围框质量的影响较

大,并且执行效率受图像中人数影响明显,时间复杂度增长梯度较大.

自底向上姿态估计算法以图像中基本像素为基础,层层递进勾勒出人体姿态的关键点,对关键点聚类及分组处理,并以此作为判断人体整体姿态的依据[28]. 该方法对整幅图像进行整体重新构解,符合多人体姿态估计的逻辑,且其分类识别效率不会因人数变化而变化. 但由于该方法对整幅图像缺乏全局性控制,当面临复杂背景环境以及多人人体重叠交互和拥挤遮挡时,比较容易出现关节点位置错乱,导致识别精度降低,姿态估计效果显著下降[29].

自顶向下方法较自底向上方法的优势在于其对图像中目标进行了显性解耦,将多人体之间复杂关系简单化,只需考虑框内关键点位置,不用考虑不同框之间关键点关系. 自底向上方法虽然不依赖于检测手段,可以平铺式获得所有人体的关键点,但对姿态尺度变化及多姿态间信息冗余处理能力较弱.

多人姿态识别通常先利用人体检测器定位人体位置,再通过特征提取关键点,最后通过骨骼关键点进行姿态分类. 常用的人体检测方法有 YOLO[30],Fase-RCNN[31],常见的姿态识别方法有 Mask R-CNN[19]、DeepCut[32]、级联金字塔网络(cascaded pyramid networdk,CPN)[33]、堆叠沙漏网络(hourglass network)[34]、对称空间变换网络(spatial de-transformer network,SDTN)[35]等.

1.2 多人姿态识别主要数据集

在人工智能研究领域,数据集起着非常关键的作用,其规模和质量从一定程度上可以反映对应研究方向的成熟程度. 人体姿态识别领域发展时间比较短,特别是多人姿态识别方向,从各类文献整理情况来看,主要代表性数据集有:

1.2.1 MPII 数据集

斯坦福大学创建的 MPII 数据集是一个非常具有挑战性的多人姿态数据集[24],其包含了超过 25 000 幅多人真实行为的交互图像,涉及丰富的活动场景,例如行走、站立、跑动等. 约有 40 000 个带有标注的人体样本,包含了详细的人体姿势信息标注,例如头部、肩部、手肘、手腕、膝盖、脚踝等关键部分的局部姿态. 其中用于训练的约有 25 000 个,用于验证评估的约有 3 000 个,用于测试的约有 12 000 个. 每个人体样本标注的关键点数量是 16 个. 该数据集涵盖了多种拍摄角度及视角,为多人姿态估计的研究提供了较多的素材支持,其大量的图像数据集的详细信息标注,对相关算法的研究发展提供了有效动力.

1.2.2 COCO 数据集

COCO(common objects in COntext)数据集是微软在 2014 年生产创建的,最初创建的目的是用于图像分割机图片上下文关系方向研究[36]. 数据集按类型分为训练集、验证集和测试集,按版本分为 MSCOCO2016 版本及 MSCOCO2017 版本. MSCOCO2017 训练集包含了 57 000 幅图像,150 000 个带有标注的人体实例,这些标注用于对图像中人物的姿态、行为及动作的描述. 验证集包含了 5 000 幅图像,测试集包含了 20 000 幅图像. 每个人体样本标注的关键点数量是 17 个,这些关键点位置通过坐标的形式在图像中进行标注,标注的方法为“密集人体关键点”标注法. 该数据集对人体行为分析和动作识别有着重要意义.

1.2.3 PASCAL VOC 数据集

PASCAL(pattern analysis statical modeling and computational learning)VOC 数据集为 PASCAL VOC 挑战赛所使用的数据集[2],它是一个由欧盟资助的网络组织举办的世界级计算机视觉领域的挑战赛. 有很多业界知名的计算机视觉模型都是在 PASCAL VOC 挑战赛上推出的,例如 R-CNN、YOLO 及 SSD 等. 2012 年是该挑战赛的最后一届,所以 VOC2017 数据集成为了该数据集最后一个版本,该数据集包含了 20 个类别的行为数据,其中包含 11 540 张图像,27 450 个带有标注的人体实例. 虽然挑战赛已经结束,但研究人员依然可以通过该数据集训练模型并检验模型的预测结果.

1.2.4 UT-Interaction 数据集

UT-Interaction 数据集包含了 6 个多人交互动作分类[37]:握手、拥抱、踢腿、手指、击拳和推搡. 该数据集标注了这些交互动作的行为实况标签、时间间隔和边界框. 按照制作条件区分为 UT-1 版本和 UT-2 版本. UT-1 版本在停车场拍摄,背景静止,相机无抖动. UT-2 版本在刮风的草坪上拍摄,背景动态,相机有抖动. 两个版本各有视频 60 个,每个视频长度为 2-6 s,帧率为 30 帧/s,视频分辨率为 720 * 480,人体高度约为 200 像素. 该数据集定位为有限遮挡,固定视点,其上的实验均采用每组 10 次留 1 次交叉验证法来评

估模型的性能,即从每组的 10 个序列中留下 1 个用于测试,将其他 9 个用于训练,在迭代更改测试集的同时计算平均性能.

1.2.5　SBU 数据集

SBU 数据集是一个较为早期的大规模图像描述数据集[37]. 它包含了约 1 万张多人动作交互图像,每张图像对应 5 个描述,这些描述均通过 Amazon Mechanical Turk 上的人工收集而来. 在数据收集前,先使用对象、属性、动作、物品及场景查询词对图片分享网站进行查询,得到大量携带相关图文的照片,再根据描述相关性及视觉描述性进行过滤,最后保留至少两个拟定术语作为其描述. 其包括了靠近、离开、推搡、踢腿、击打、交接、拥抱及握手 8 类交互动作,一共 282 个动作短视频,每个视频约 2-3 s,帧率为 15 帧/s. 该数据集可以用于图像标注,多模态数据集训练等任务,采用 5 折交叉实验方法验证评估.

1.3　多人姿态识别面临的主要问题和解决思路

人体姿态识别从类型区分和发展方向上分为单人姿态识别和多人姿态识别,前者在一定程度上又是后者的研究基础和发展动力. 单人姿态识别领域随着近十年来人工智能技术力量总体提升和众多科研学者的努力追求,已有了较丰硕的科研成果,其中不乏很多高质量数据集和模型算法. 多人姿态识别领域由于起步较晚,难度较大,复杂性较高,仍然面临很多的关键问题需要解决. 在多人姿态识别中,随着复杂性的提升,网络深度也随之加深,梯度消失问题也随之加剧,网络性能也随之衰减. 另外,多人图像中因为存在互相遮挡和相互交叠,对关键点的提取产生一定影响,所以识别精度也因受到干扰而降低. 对于这些问题,本文主要解决思路是在骨骼图卷积框架基础上,通过引入残差处理机制解决多级网络导致梯度消失引发的精度效率衰减问题,并引入注意力机制解决噪声导致的特征提取能力下降问题.

2　基础理论及总体框架

2.1　多人骨骼关键节点特征提取

对于多人图像的骨骼关键点提取,通常先使用已完成预训练的人体目标检测模型进行人体检测及框选标定工作. 利用边框数据转换成位置矩阵,继而对框选图进行映射处理,按照统一规格输出单人图像. 本文采用了自顶而下的处理方式,选取了 YOLOv8s-Pose 作为人体目标检测工具,对输入的多人体图像进行检测,单独框定所有单人体实例,并对边界框进行比例调整后裁剪适配处理,将裁剪的图像作为单人姿态识别对象输出到姿态识别模型.

针对裁剪后的单人体图像,以动作涉及的局部区域活动单元为基础,通过对动作序列中所有关键点统计坐标向量,并结合时域空域双流通道建立关键点之间对应关系,以邻域节点之间的邻接矩阵表述关键点之间的互联关系. 本文继续在之前研究成果“骨骼双流注意力增强图卷积人体姿态识别”的基础上向前推进,依然利用 OpenPose 的 25 个关键点中的 18 个作为输入单人体对象的对标节点,这 18 个关键点定义关系如表 1 所示. 这些关键点作为骨骼特征提取的标本,通过前期对骨骼节点流及骨骼架构流双通道相结合机制的研究经验,分别对双流通道中的流信息进行交互分析,并按帧提取骨骼关键点. 在此基础上,构建运动信息矩阵,完成骨骼关键点时间线关系构造,以及运动信息关联模型构建,从而完成骨骼关键节点的提取和运动信息表征工作. 人体目标检测处理及骨骼特征点提取流程示意如图 1 所示.

表 1　骨骼关键节点定义表

Table 1　Bone key node definition table

编号	关节名	编号	关节名
1	鼻子	10	左肩
2	右耳	11	左肘
3	右眼	12	左手
4	左眼	13	右胯
5	左耳	14	右膝
6	脖子	15	右踝
7	右手	16	左胯
8	右肘	17	左膝
9	右肩	18	左踝

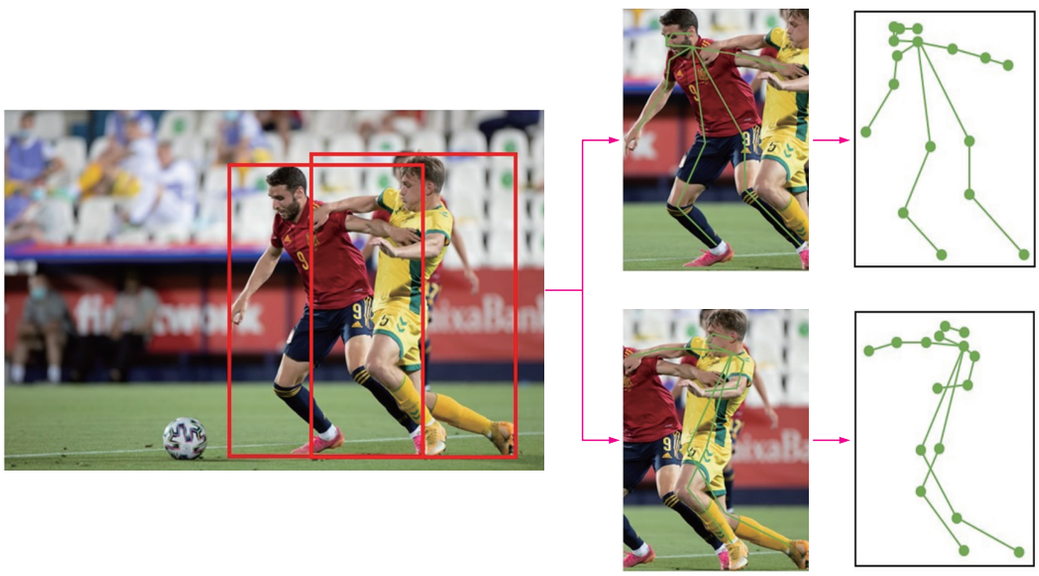

图 1　目标检测处理及骨骼特征点提取过程示意图

Fig. 1　Schematic diagram of target detection processing and bone feature point extraction process

对于骨骼关键节点运动信息表示,本质上是通过对不同邻域节点之间坐标向量变化为基础数据的反映,以相邻骨骼关键节点之间尺寸关系集合构成了整个骨骼关键点链的关联. 对邻接关键点之间关联定位关系定义如表 2 所示.

表 2　邻接关键点之间关联定位关系

Table 2　Association relationships between adjacent key points

邻接关节编号	邻接关节	邻接关节编号	邻接关节
(1,6)	鼻子-脖子	(1,3)	鼻子-右眼
(6,9)	脖子-右肩	(6,10)	脖子-左肩
(6,13)	脖子-右胯	(6,16)	脖子-左胯
(8,9)	右肘-右肩	(7,8)	右手-右肘
(10,11)	左肩-左肘	(11,12)	左肘-左手
(13,14)	右胯-右膝	(14,15)	右膝-右踝
(16,17)	左胯-左膝	(17,18)	左膝-左踝

2.2　姿态识别中残差机制的梯度弥散抑制

关键点检测通常会通过对基础图像的循环迭代信息提取手段,以达到特征对象构成的目标. 但受到网络深度的影响,提取的信息抽象级别也越来越高,特征图空间尺寸也迅速缩小,也就是通常所说的深度神经网络梯度弥散问题,该问题会导致网络训练难度加大,特征表征能力变弱,过拟合现象频发,性能极度退化,收敛效果不理想,对整体预测情况产生负面效应.

对于上述梯度弥散导致的影响,较成熟的做法是引入恒等映射或跳跃连接机制,将输入和输出之间的差异直接投向网络. 残差块由于携带了卷积路径并使用了跳级路径的方式,不仅在特征信息层面上能有更高层次的反馈,并且对原始层面的初始信息也可以完整的封存,这对于整个姿态识别网络的性能提升有非常大的助力作用. 该做法在单人姿态识别中残差块已有一定应用,在多人姿态识别中网络深度复杂度进一步加深,通过对残差块的改进,引入多元监督模块及多元回归模块,融入多尺度特征,并对这些多尺度特征进行匹配,达到对关键点定位精度提升的目标.

对残差块的具体改造策略,是针对 1×1 的卷积结果进行拆分,再将其通过 3×3 的深度独立化卷积进行处理,通过这种方式减少带入参数的数量. 通过在特征图的拆分图之间融入跳跃因子,大幅提升了残差块的特征提取能力. 同时,通过减少残差网络的 2 至 5 层进出通道数量,降低通道复杂度和计算量. 残差块优化改进过程如图 2 所示,蓝色区域为基础残差块卷积处理,橙色区域为独立化变形跳跃卷积处理. N 代表输入图像的人体数量,d 代表单个体特征向量的维度,设定初始残差块参数数量为 X,通道数降低并深度独立化卷积后的参数数量为 Y,I 为进向通道数量,O 为出向通道数量,按照 $O=2\times I$ 设置,x_i 和 y_i 分别代表拆分图独立跳跃变形卷积前后. 上述两种参数数量可按照下式计算.

初始残差块参数数量计算如下式(1):

$$X=I\times\frac{O}{2}+(3\times3)\times\left(\frac{O}{2}\right)^2+\frac{O}{2}\times O+I\times O=14I^2. \tag{1}$$

通道数降低并深度独立化卷积后的参数数量计算如下式(2):

$$Y=\frac{I}{2}\times\frac{O}{4}+(3\times3)\times\frac{O}{4\times4}\times3+\frac{I}{4}\times\frac{O}{2}+\frac{I}{2}\times\frac{O}{2}=3I^2+\frac{27}{8}I. \tag{2}$$

按照计算结果,得出变换前后参数量 X 比 Y 减少了约 3/4. 这种方式较好的解决了多人姿态识别情况下姿态复杂,关键节点变化尺度大的问题,通过增大感受野范围以适应整体人体图像复杂度的提高,通过保存小范围的局部卷积区域以保留微细关键信息. 如上所述,仅将 1/4 的特征信息单元投入独立化变形感受野,通过跳跃因子的连接作用降低反复卷积带来的感受野不收敛及网络退化副作用,并在卷积过程中借助批归一化(batch normalization,BN)进行批量标准化训练处理,从而达到梯度弥散抑制的目标.

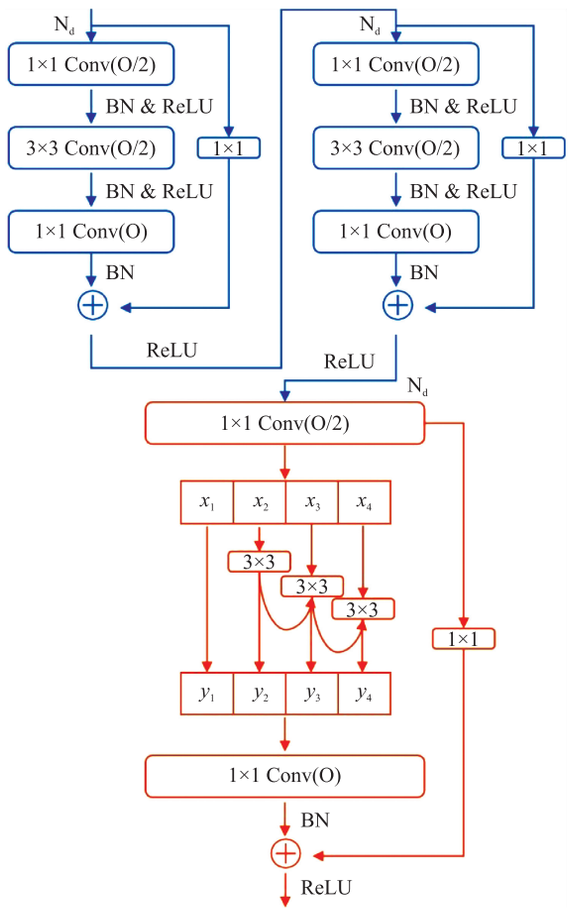

图 2　残差块的优化改进过程

Fig. 2　The optimization and improvement process of the residual block

2.3　多人姿态识别中的混合注意力增强机制

与单人姿态识别相比,多人姿态识别过程中由于存在人体样本数不确定问题,必须首先采用样本遍历器进行扫描检测,这不仅对执行效率产生影响,对精度提升也带来更大的困难和要求. 而且多人在图像中的不同分布、距离远近、图像尺寸比例千差万别,互相遮挡嵌套,都会对识别效率和精度结果构成巨大障碍.

注意力机制模仿了人类注意事物的过程,让模型排除干扰,聚焦于输入信息中关键部分,从而提高模型的效率和精度. 它的核心在于,输入信息中关键信息是什么,以及到哪里去寻找这些关键信息. 对应到多人姿态识别方面,也就是每一个人体在图像的什么位置,这些单人体的关键节点又在何处. 空间注意力机制(spatial attention mechanism,SAM)从广度上来解决图像关键信息在何处的问题,通道注意力机制(channel attention mechanism,CAM)从深度上来解决人体关键点位于何处的问题. 混合注意力机制(convolutional block attention module,CBAM)将空间及通道注意力机制相融合,对输入特征图进行混合处理,并得到优化后的结果. 通过对关键信息的集中处理,注意力机制可以促进模型的准确性提高及效率提升.

对于注意力的计算通常可以通过注意力函数来完成. 注意力函数可以按 AM(*Query*,*Key*,*Value*)表示,其中 *Query* 表示原始向量序列,*Key* 与 *Value* 表示目标向量序列键值对,使用 *Query* 及(*Key*-*Value*)键值对描述输出映射,实际上注意力机制本质上就是计算 Query 及 Key 之间的相似关系,并通过这种相似关系确定 *Query* 及 *Value* 之间的注意力关系. 在式(3)中,首先求得 *Query* 与全部 I 个迭代的 *Key* 的积,再对第 i 个迭代的维度 *dimension* 开根号,将两者相除以防止乘积过快增长,再通过 softmax 进行归一化指数处理以保持与 *Value* 相应的权重.

$$\mathrm{AM}(Query,Key,Value)=\mathrm{softmax}\left(\frac{Query\times Key^{I}}{\sqrt{dim_i}}\right)\times Value. \tag{3}$$

2.4　残差混合注意力结合骨骼图卷积多人姿态识别模型

在以上章节中介绍了多人姿态识别的基本原理、方法、涉及的主要数据集、面临的主要问题和解决思路. 处理的基本手段包括目标检测和骨骼关键点提取,残差跳跃以及混合注意力机制的作用. 本节将通过在骨骼图卷积基础上渗透带入残差信息并借力空间和通道注意力混合机制,在此基础上提出了残差混合注意力结合骨骼图卷积多人姿态识别模型(skelton-based graph convolution with residual combined with mixed attention mechanism,SGCRA).

该模型将人体骨骼关节点图的结构特征和卷积神经网络特征提取能力相结合(并将残差块信息处理机制和混合注意力机制相结合),对复杂多人体图像识别处理. 首先进行预处理,利用人体检测器对图像中所有人体进行标注框定,再利用骨骼图特征提取算法,对框定人体进行深度卷积处理,得出骨骼关键点坐标信息和上下文关系信息,并由此生成关键点初步特征图. 由于在对局部信息捕获的过程中对部分全局信息的丢失,在特征学习过程中,利用残差块对深度网络梯度消失问题的有效抑制能力,在卷积层间添加残差跳跃因子进行连接,加深了网络全局信息学习能力,优化了信息流动的效果和模型训练的效率. 同时融入混合注意力机制,利用空间注意力机制的空间位置权重自适应调整能力,聚焦于人体关键姿态特征的位置分布,利用通道注意力机制的特征权重调整能力,聚焦于各个通道中重要信息的捕获. 在混合注意力机制的作用下,增强了模型抗干扰能力和关键信息提取能力. 在模型对局部特征信息和全局特征信息处理的基础上,利用多层感知机(multilayer perceptron,MLP),再经过 Sigmoid 激活函数和全连接处理,最终对特征进行分类. 残差混合注意力结合骨骼图卷积多人姿态识别模型框架如图 3 所示.

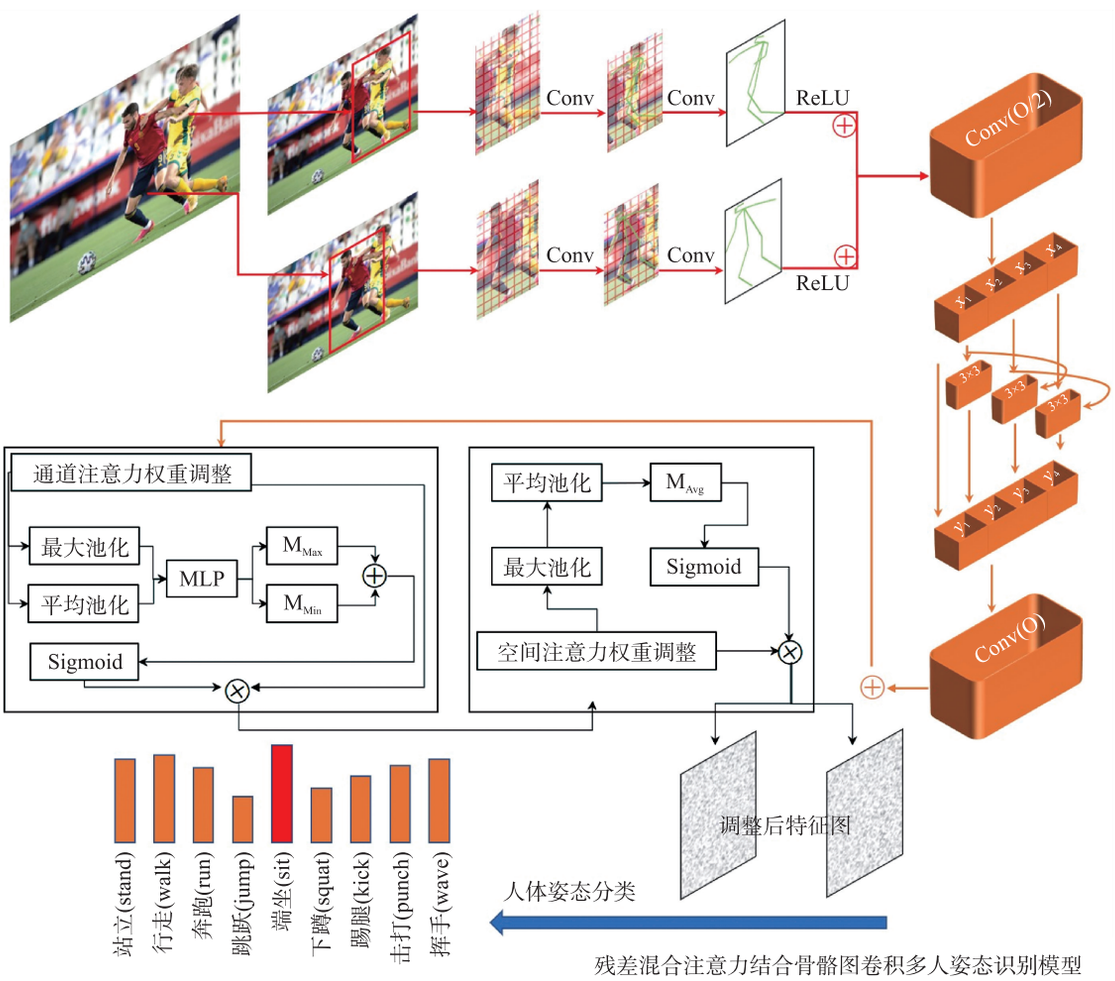

图 3　残差混合注意力结合骨骼图卷积多人姿态识别模型

Fig. 3　Model framework for skelton-based graph convolution with residual combined with mixed attention mechanism for multi-person posture recognition

3　实验及分析

3.1　实验环境介绍

基础硬件及操作系统环境:品牌型号为浪潮 NF5280M6 机架式服务器,配置了两块 Intel Xeon Gold 5318Y CPU,其内核总数 48 核,线程总数为 96 条,内存为 128 GB DDR4 SDRAM,显卡为两块 NVIDIA GeForce A10 GPU 加速卡,显存为 48 GB DDR6. 操作系统采用的是基于 Linux 的 Ubuntu 平台,版本号为 20.04.2. 并行计算和编程平台采用的是 NVIDIA 的 CUDA11.6,CuDnn8.6.

软件开发及支撑环境:机器学习库采用的是开源的 PyTorch,它是拥有自动求导功能的强大深度神经网络. 代码集成开发环境采用了由 Jetbrains 发布的 PyCharm2023.2.1,它是一种适用于科学数据和 Web 开发的 Python 集成开发环境,带有一整套可以帮助用户在使用 Python 语言开发时提高其效率的工具,可以简化困难的编码任务,助力开发者提高工作效率. 数据科学包管理平台采用的是 Anaconda23.07.180256. 人体姿态估计基础框架采用的是开源的 OpenPose 工具,它是世界上首个基于深度学习的实时多人二维姿态估计应用,由美国卡耐基梅隆大学开发研制,底层逻辑是卷积神经网络和监督学习,可以实现面部表情、人体动作、局部运动等姿态的估计,且支持单人及多人模式,鲁棒性较好.

3.2　实验数据集、参数调校及评价指标

本文的实验基于两个在多人识别研究方面具有代表性的数据集 MPII 数据集和 COCO 数据集进行对比性验证. 之所以选择这两个数据集,是因为网上公开的数据集大多数都属于综合类数据集,针对多人且以姿态行为作为主题的比较少,为数不多的多人数据集中,带有丰富标注的更为少见. 如果用综合数据集来进行训练使用,对算法效果非但不理想,还可能产生负面效果. 基于两个数据集上比对的结果,进行了本文提出的残差混合注意力结合骨骼图卷积多人姿态识别模型的验证,并在后续章节中将本模型算法的效果与其他多人姿态识别算法进行了对比. 前面章节已经对两个数据集有过简要的介绍. MPII 数据集由斯坦福大学创建,在多人姿态库中属于数据量较大、场景丰富、姿态分类较多的数据集,且标注详细,便于使用. COCO 数据集由微软创建,本文使用的是 2017 版本,该数据集数据量与前者数量集相似,姿态分类数量级相当,而且采用了“密集人体关键点”标注法,精确标注了关键点坐标,效用更加精准,差异更加弱化.

为提升算法运行效率,增强并行执行率,在实验图像输入之前,对其进行了处理. 首先对原始图像进行了缩放和尺寸统一化,都按照 512×512 像素进行了处理,并在训练阶段通过本实验组之前在人脸表情识别领域采用过的数据增强手段,即通过动态变形、动态翻转、动态拉伸、动态旋转等方式进行数据扩容,以增加同类数据类别的数据量,增强训练相同姿态不同角度和形变的处理能力,提升模型识别效果的鲁棒性. 在使用人体目标检测工具 YOLOv8s-Pose 时,为了提高模型执行效率并保证模型训练效果,将训练参数 batch_size 提升设置为 128,但将切片值 subdivisions 降低设置为 4,意义在于通知模型如果一次性无法处理 128 幅图像,就减小队列中大批量数据个数,将大数据切片成 4 幅图像一组以减轻模型压力,循环处理,一旦能力提升可再次吞吐大数据包. 为防止梯度弥散,将模型初始学习率 learning_rate 设置为 0.001,学习衰减率 scales 设置为 0.1,最大批次数 max_batches 设置为 200 000.

多人姿态识别通常使用的衡量标准有关键点相似度(object keypoint similarity,OKS)、10 个 OKS 的阈值基础上的平均精度均值(mean average precision,mAP),以及平均召回率(average recall,AR). OKS 类似于目标检测的交并比(intersection over union,IOU),用于计算预测关键节点和真实值间的差距,按照式(4)归一化尺度计算. 其中 k 表示多人图像中人物编号,i 表示关键点编号,E_{ki} 表示预测关键点和实际关键点之间的欧氏距离,S_k 表示第 k 个人图像的尺寸因子,等于人像面积的平方根,σ 函数用于计算可见点,σ_i 表示第 i 个关键点的归一化因子,v_{ki} 表示第 k 个人的第 i 个关键点的可见值.

$$OKS_k = \frac{\sum_i \exp[-E_{ki}^2/(2S_k^2\sigma_i^2)]\sigma(v_{ki}=1)}{\sum_i \sigma(v_{ki}=1)}. \tag{4}$$

平均精度均值 mAP 表示 OKS 取值为 0.50 至 0.95 间以 0.05 为增幅的平均值,其计算基础为平均精度(average precision,AP),计算式如(5). 其中 s 表示图片中人体总数,M 为人工设置的最大值,一个最大值 M 就对应一个 AP 值,例如 AP^{50} 表示 OKS 为 0.50 的 AP 值,所以平均精度均值 mAP 表示 OKS 取值为 0.50 至 0.95 之间以 0.05 为增幅得到的平均值. AP^S 为尺寸较小目标的检测精度,AP^L 为尺寸较大目标的检测精度. 召回率用于衡量模型在检测任务中识别到所有目标的程度,平均召回率 AR 作为目标检测任务中重要指标,通常与 AP 结合使用,用于对模型性能全面评估. AP 聚焦于不同置信度阈值条件下的精度,而 AR 聚焦于模型在不同条件下的召回率.

$$AP = \frac{\sum_s \sum_k \sigma(OKS_k > M)}{\sum_s \sum_k 1}. \tag{5}$$

3.3　实验结果及分析

本实验沿用了骨骼双流注意力增强图卷积人体姿态识别中采用的场景,也就是在受控情况下的去噪场景,排除背景复杂化干扰因素,同时保证给予充足室内光照. 同样采用正机位方式进行实验,即多人直面机位,完成动作姿态的解析. 对姿态的分类,本模型提供的为站立(stand),行走(walk),奔跑(run),跳跃(jump),端坐(sit),下蹲(squat),踢腿(kick),击打(punch),挥手(wave)九类,与之前实验不同的是,增加了多人处理机制. 将实验结果与本实验组先前所提出的针对单人的骨骼双流注意力增强图卷积人体姿态识别模型识别结果和效率进行比对,可以发现受多人因素影响,改进后的多人姿态模型识别率和速度较之前单人识别率有一定下降. 究其原因,首先在多人姿态识别中需要加入自顶而下的多人体检测器对图像中多个人体位置进行检测和框定. 另外多人图像往往存在肢体交互和彼此遮挡干扰,这增加了识别难度和计算复杂度,也增加了定位和识别的误差概率. 再者,多人姿态识别数据集质量远不如单人数据集的高,可用于训练的代表性数据集数量也相差甚远.

实验比对结果如表 3 所示. 第一列表示实验识别的九种姿态. 第二列表示本项目组前期研究成果——骨骼双流注意力增强图卷积人体姿态识别(skeleton double-flow attention enhance graph convolution, SDFAEGC)模型的最高识别率,该结果基于 NTU RGB+D 数据集. 第三列和第四列分别表示在 MPII 数据集和 MSCOCO2017 数据集基础上残差混合注意力结合骨骼图卷积多人姿态识别模型的最高识别率. 从实验结果可以看出,SGCRA 识别率不论在 MPII 数据集还是 MSCOCO2017 数据集上均较 SDFAEGC 有所下降. SGCRA 在 MSCOCO2017 数据集上训练效果一定程度上优于 MPII 数据集,虽然在个别姿态识别率上存在反向差异,但总体上在两个数据集上各类姿态分类识别情况分布较为一致,这说明模型跨数据集稳定性较好. 实验现场情况及模型测试情况示例如图 4 所示,其中融入了两人及三人的场景,且加入了男性及女性多种差异测试组合的情况. 同时在遮挡及重叠等情况下进行了实验,识别效果骤减,所以该模型还不适用于各类复杂场景,这也是未来有待继续攻克的研究方向.

表 3　SDFAEGC 模型在 NTU RGB+D 数据集及 SGCRA 模型在 MPII 数据集和 MSCOCO2017 数据集上的实验结果比较

Table 3　Comparison of experimental results of model SDFAEGC on NTU RGB+D and SGCRA on MPII and MSCOCO2017 datasets

姿态类别	NTU RGB+D 数据集 SDFAEGC	MPII 数据集 SGCRA	MSCOCO2017 数据集 SGCRA
站立(stand)	100.00%	86.25%	89.10%
行走(walk)	100.00%	77.08%	76.93%
奔跑(run)	100.00%	73.21%	78.05%
跳跃(jump)	100.00%	64.49%	70.11%
端坐(sit)	100.00%	82.13%	86.44%
下蹲(squat)	100.00%	76.52%	69.73%
踢腿(kick)	100.00%	78.36%	79.67%
击打(punch)	100.00%	60.87%	71.47%
挥手(wave)	100.00%	83.45%	88.69%

3.4　实验结果与主流方法的比较

在 MSCOCO2017 数据集基础上,本项目组将实验结果与其他相关多人姿态识别模型进行了比较.

根据各类文献记载情况,收集到的基于 MSCOCO2017 数据集的算法有:根据各类文献记载情况,收集到的基于 MSCOCO2017 数据集的算法有:CPN、HRNet-W32、HigherHRNet、DB-Pose、SSR-Pose、CMU-Pose、CHED、S-ViPNAS+、TransPose-H-A6+、TFPose+、PRTR、CRNet、Mask-RCNN、G_RMI、SimpleBaseline、FAIR、oks 和 Bangbangren,其中最后三个是 COCO 官方竞赛榜列表中的方法. 对各个算法及本文算法从输入图像尺寸、参数量、计算量、mAP、AP^{50}、AP^{75}、AP^{S}、AP^{L}、mAR 等指标进行了详细的比较.

在 MSCOCO2017 数据集上对各模型算法的各项主流明细指标的验证比对结果如表 4 所示. 根据比较情况,以往模型中 CRNet 算法在各类指标上均为领先(在表 4 中以下划线标注),而本文模型在各指标上较 CRNet 都有一定提升(在表 4 中以粗体标注). 具体来说,在 mAP 指标上提升了 0.9%,在 AP^{50}指标上提升了 0.5%,在 AP^{75}指标上提升了 2.2%,在 AP^{S} 指标上提升了 2.7%,在 AP^{L} 指标上提升了 0.4%,在 mAR 指标上提升了 0.3%,其中 AP^{S} 指标提升较为明显,这表明针对小尺度图像切片,本文模型对其形变适应性及全局感受野感知性都有着较优的表现. 由实验比对数据结果可以得出,本文算法通过将残差块信息与

图 4　实验环境及结果图例

Fig. 4　Experimental environment and results legend

时域空域混合注意力相结合以提升骨骼图卷积多人姿态识别效果,提高了人体姿态预测的精确度,在各个精细指标上相较于优秀算法都有一定程度的提升,总体而言稳定性较好,分布较为均匀.

表 4　本文及各代表性算法在各项主流明细指标上的验证比对结果

Table 4　The verification and comparison results of this paper and the representative algorithm in the mainstream detailed indicators

模型算法	骨干网络	年份	尺寸	参数量	计算量	mAP	AP^{50}	AP^{75}	AP^{S}	AP^{L}	mAR
CPN	ResNet-50	2018	256×192	27.0 M	6.2 G	68.6	—	—	—	—	—
HRNet-W32	HRNet-W32	2019	384×288	28.5 M	16.0 G	76.7	91.9	83.6	73.2	83.2	81.6
HigherHRNet	HRNet-W32	2020	512×512	—	—	65.7	85.9	71.3	60.0	74.3	70.6
DB-Pose	SSR-W32	2021	—	—	—	66.8	86.1	72.8	61.4	75.8	71.7
SSR-Pose	SSR-W32	2020	—	—	—	67.8	86.8	73.3	62.4	76.1	81.7
CMU-Pose	—	—	—	—	—	61.8	84.9	67.5	57.1	68.2	66.5
CHED	HRNet-W32	—	—	—	—	67.0	86.2	72.7	61.4	75.5	71.7
S-ViPNAS+	HRNet-W32	2021	256×192	16.3 M	5.64 G	74.7	89.9	82.0	71.0	81.5	81.2
TransPose-H-A6+	HRNet-W48	2021	256×192	17.5 M	21.8 G	75.8	—	—	—	—	80.8
TFPose+	ResNet-50	2021	384×288	—	20.4 G	72.4	—	—	—	—	—

续表 4 Table 4 continued

模型算法	骨干网络	年份	尺寸	参数量	计算量	mAP	AP^{50}	AP^{75}	AP^{S}	AP^{L}	mAR
PRTR	HRNet-W32	2021	512×384	57.2 M	37.8 G	73.3	89.2	79.9	69.0	80.9	80.2
CRNet	HRNet-W32	2021	384×288	28.6 M	16.1 G	78.2	92.2	84.5	74.8	84.4	82.9
Mask-RCNN	ResNet-50-FPN	—	—	—	—	63.1	87.3	68.7	57.8	71.4	—
G_RMI	ResNet-101	2017	353×257	—	—	64.9	85.5	71.3	62.3	70.0	69.7
SimpleBaseline	ResNet-152	2020	384×288	—	—	73.7	91.9	81.1	70.3	80.0	79.0
FAIR	ResNetXt-101-FPN	—	—	—	—	69.2	90.4	77.0	64.9	76.3	75.2
oks	—	—	—	—	—	72.0	90.3	79.7	67.6	78.4	77.1
Bangbangren	ResNet-101	—	—	—	—	72.8	89.4	79.6	68.6	80.0	78.7
SGCRA(本文)	ResNet	2024	512×512	36.7 M	12.4 M	**79.1**	**92.7**	**86.7**	**77.5**	**84.8**	**83.2**

4 结论

针对多人姿态识别研究方向存在的识别率低和执行效率差的问题,本文结合项目组前期经验,在骨骼双流注意力增强图卷积人体姿态识别、双流增强融合网络微表情识别以及基于残差整流增强卷积神经网络的表情识别等研究成果的基础上,选用典型的自顶向下多人姿态识别研究途径,采用预处理人体检测对单人体进行框定,采用骨骼图卷积进行人体关键点定位,采用残差块机制缓解网络深度诱发的梯度弥撒症状,采用混合注意力机制提升模型能力和效率. 对本文所提出模型在 MPII 以及 MSCOCO2017 两个数据集上进行了实验验证,结果显示该模型对九类预设姿态均有较好识别效果,鲁棒性良好. 同时,也将本文算法与同类主流算法进行了横向比较,根据比对结果,本文算法在各类指标上都有着较好的表现.

由于多人姿态识别的复杂性,其对算法能力及执行效率都有着更高的要求,故而本文实验依然采取在简单场景中进行的方式,且多人之间并无交互性. 但在实际应用场景中,更多情况是在错综复杂的环境下多人存在交互、遮挡、重叠等背景下. 在后续的研究中,将针对复合场景或真实场景,结合多人交互行为识别方面开展研究,为与多人姿态识别相关的社会应用服务提供更有价值的动力和支持.

[参考文献]

[1] 马境远,刘鲲,傅慧源. 一种融合多模态特征的视频暴力检测方法[J]. 重庆邮电大学学报(自然科学版),2021,33:861.

[2] 蔡哲栋,应娜,郭春生,等. YOLOv3 剪枝模型的多人姿态估计[J]. 中国图像图形学报,2021,26(4):837-846.

[3] XIA R J,LI Y S,LUO W H. LAGA-Net:local-and-global attention network for skeleton based action recognition[J]. IEEE transactions on multimedia,2022,24:2648-2661.

[4] YANG H Y,GU Y Z,ZHU J C,et al. PGCNTCA:pseudo graph convolutional network with temporal and channel-wise attention for skeleton-based action recognition[J]. IEEE access,2020,8:10040-10047.

[5] SULTANI W,CHEN C,SHAH M,et al. Real-world anomaly detection in surveillance videos[C]//Computer Vision and Pattern Recognition. Piscataway:IEEE,2018:6479-6488.

[6] PENG W,SHI J G,ZHAO G Y. Spatial temporal graph deconvolutional network for skeleton-based human action recognition[J]. IEEE signal processing letters,2021,28:244-248.

[7] YU W,YANG K,YAO H,et al. Exploiting the complementary strengths of multi-layer CNN features for image retrieval[J]. Neurocomputing,2017,237:235-241.

[8] LIU J,SHAHROUDY A,WANG G,et al. Skeleton based online action prediction using scale selection network[J]. IEEE transactions on pattern analysis and machine intelligence,2020,42(6):1453-1467.

[9] PENG W,SHI J,VARANKA T,et al. Rethinking the ST-GCNs for 3D skeleton-based human action recognition[J]. Neurocomputing,2021,454:45-53.

[10] ZHANG S Y,YANG Y,JUN X,et al. Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks[J]. IEEE transactions on multimedia,2018,20(9):2230-2343.

[11] 王佳铖,鲍劲松,刘天元等. 基于工件注意力的车间作业行为在线识别方法[J]. 计算机集成制造系统,2021,27(4):1099-1107.

[12] 苏江毅,宋晓宁,吴小俊,等. 多模态轻量级图卷积人体骨架行为识别方法[J]. 计算机科学与探索,2021,15(4):733-742.

[13] JI S,XU W,YANG M,et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence,2012,35(1):221-231.
[14] 黄海新,王瑞鹏,刘孝阳. 基于3D卷积的人体行为识别技术综述[J]. 计算机科学,2020,47(S2):139-144.
[15] ZHANG B,WANG Y,HOU W,et al. Flexmatch:boosting semi-supervised learning with curriculum pseudo labeling[J]. Advances in neural information processing systems,2021,34:18408-18419.
[16] CHEN P,GAO Y,MA A J. Multi-level attentive adversarial learning with temporal dilation for unsupervised video domain adaptation[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Piscataway,NJ:IEEE Press,2022:1259-1268.
[17] TOSHEV A,SZEGEDY C. DeepPose:human pose estimation via deep neural networks[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus:IEEE,2014:1653-1660.
[18] LIN G,LI Q,LI M,et al. A novel bottleneck-activated feedback neural network model for time series prediction[J]. IEEE transactions on neural networks and learning systems,2021,32(4):1621-1635.
[19] ABDEL-BASSET M,HAWASHH,CHAKRABORTTYR K,et al. ST-DeepHAR:deep learning model for human activity recognition in loHT applications[J]. SENSORS,2021,8(6):4969-4979.
[20] HE K M,ZHANG X Y,REN S Q,et al. Delving deep into rectifiers:surpassing human-level performance on ImageNet classification[C]//2015 IEEE International Conference on Computer Vision. Santiago:IEEE,2016:1026-1034.
[21] WEI S E,RAMAKRISHNA V,KANADE T,et al. Convolutional pose machines[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas:IEEE,2016:4724-4732.
[22] 陈斌,朱晋宁,东一舟. 基于残差整流增强卷积神经网络的表情识别[J]. 液晶与显示,2020,35(12):1299-1308.
[23] KOCABAS M,KARAGOZ S,AKBAS E. MultiPoseNet:fast multi-person pose estimation using pose residual network[C]//Proceedings of the 15th European Conference on Computer Vision. Munich:Springer,2018:417-433.
[24] 秦晓飞,郭海洋,陈浩胜,等. 基于深度残差网络的多人姿态估计[J]. 光学仪器,2021,43(2):39-47.
[25] HU J,SHEN L,SUN G. Squeeze-adn-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City:IEEE,2018:7132-7141.
[26] CAO Z,GINES H,SIMON T,et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//Proceedings of the Conference on Computer Vision and Pattern Recognition. Washington,D. C.,USA:IEEE Press,2017:7291-7299.
[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. Long Beach,CA:IEEE,2019:11969-11978.
[28] CHENG B,XIAO B,WANG J,et al. HigherHRNet:scale-aware representation learning for bottom-up human pose estimation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,WA:IEEE,2020:5385-5394.
[29] 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.
[30] REDMON J,DIVVALA S,GIRSHICK R,et al. You only look once:unified,real-time object detection[C]//Computer Vision & Pattern Recognition. Las Vegas,NV,USA:IEEE,2016,1(2):779-788.
[31] WANG X,SHRIVASTAVA A,GUPTA A. A-fastrcnn:hard positive generation via adversary for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii,USA:IEEE,2017:2606-2615.
[32] 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. Las Vegas,NV:IEEE,2016:4929-4937.
[33] CHEN Y,WANG Z,PENG Y,et al. Cascaded pyramid network for multi-person pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lakecity:IEEE,2018:7103-7112.
[34] NEWELL A,YANG K Y,DENG J. Stacked hourglass networks for human pose estimationproceedings of the european[C]//Conference on Computer Vision.Berlin,Germany:Springer,2016:483-499.
[35] FANG H S,XIE S,TAI Y W,et al. RMPE:regional multi-person pose estimation[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice,Italy:IEEE,2017:2353-2362.
[36] LIN T Y,MAIRE M,BELONGIE S,et al. Microsoft COCO:common objects in context[C]//European Conference on Computer Vision. Zurich:Springer,2014,8693:740-755.
[37] 成科扬,吴金霞,王文杉,等. 融合时空图卷积的多人交互行为识别[J]. 中国图像图形学报,2021,26(7):1681-1691.

[责任编辑:杜忆忱]