中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

基于骨架的人体异常行为识别与检测研究进展

朱红蕾, 卫鹏娟, 徐志刚

引用本文:
朱红蕾,卫鹏娟,徐志刚. 基于骨架的人体异常行为识别与检测研究进展[J]. 控制与决策, 2024, 39(8): 2484−2501.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2023.0451

基于骨架的人体异常行为识别与检测研究进展

朱红蕾, 卫鹏娟, 徐志刚†

(兰州理工大学 计算机与通信学院，兰州 730050)

摘 要: 人体异常行为识别与检测技术已广泛应用于各种领域. 由于视频中存在的物体遮挡、光照及视角变化、复杂背景等问题,使得利用轻量级人体骨架数据处理此类实时任务成为竞争性工具. 多数研究从不同角度对此任务相关方法进行综述,但缺少针对人体骨架的整理工作. 对此,立足于骨架数据,系统地综述了深度学习背景下的人体异常行为识别与检测方法. 首先,按照应用场景中目标个数的不同,分类总结了典型的人体姿态估计算法;其次,依据特征提取网络的不同,将异常行为识别方法分为5类,分别围绕CNN、RNN、GCN、Transformer以及混合模型展开对比分析;然后,从数据与标签的映射学习角度,对3类异常行为检测方法进行讨论;最后,介绍了基准数据集及其上相关算法的表现,并探讨了此任务所面临的挑战及展望,以期为本领域未来的研究提供参考.

关键词: 人体骨架；异常行为识别；异常行为检测；深度学习；姿态估计算法；注意力机制

中图分类号: TP391.4 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2023.0451

引用格式: 朱红蕾,卫鹏娟,徐志刚. 基于骨架的人体异常行为识别与检测研究进展[J]. 控制与决策, 2024, 39(8): 2484-2501.

Research progress on skeleton-based human abnormal behavior recognition and detection

ZHU Hong-lei, WEI Peng-juan, XU Zhi-gang†

(School of Computer and Communication，Lanzhou University of Technology，Lanzhou 730050，China)

Abstract: The technology of human abnormal behavior recognition and detection has been widely applied in various fields. Due to the problems such as object occlusion, illumination and visual angle changes and complex background in video, lightweight human skeleton becomes a competitive tool for processing such real-time tasks. Most researches have reviewed the methods relevant to this task from different perspectives, but there is a lack of work on human skeleton. Based on skeleton data, this paper systematically reviews the methods of human abnormal behavior recognition and detection under the background of deep learning. Firstly, according to the different number of targets in the application scenario, human pose estimation algorithms are classified and summarized. Secondly, based on the different feature extraction networks, the abnormal behavior recognition methods are divided into five categories, which are compared and analyzed around the CNN, RNN, GCN, Transformer and hybrid models. Then, from the perspective of data and label mapping learning, three types of abnormal behavior detection methods are discussed. Finally, the baseline datasets and the performance of related algorithms are introduced, and the challenges and prospects facing this task are discussed in order to provide reference for future research in this field.

Keywords: human skeleton；abnormal behavior recognition；abnormal behavior detection；deep learning；pose estimation algorithm； attention mechanism

0 引 言

视频中人体异常行为识别与检测是一项极具挑战的计算机视觉任务. 异常行为定义为在特定环境下,目标做出的一些不适宜的动作、姿态或事件[1]. 异常行为识别的目的是从原始视频中正确分类人体行为,进而判断测试样本属于哪种类型的异常行为;而异常行为检测的目的是通过分析上下文时空信息,检测与训练数据或已知行为模式不相符合的情况,从而准确定位异常行为发生的时刻.

人体异常行为识别与检测广泛应用于人机交

收稿日期: 2023-04-10；**录用日期:** 2023-08-16.
基金项目: 国家自然科学基金项目(62161020).
责任编委: 谢晖.
†通讯作者. E-mail: xzg_cn@163.com.

互、交通管控、公共安全、医疗监护、智能监控等方面[2]. 在人机交互领域,可实现对空巢老人或患者的跌倒行为的实时监护[3]. 在交通管控领域,可检测驾驶员接打电话等违规行为[4],最大限度阻止事故发生. 在公共安全领域,可检测在校园中发生学生打架斗殴事件[5],识别在商场自动扶梯中出现的攀爬、摔倒等危险行为[6-7],以及公共区域内恐怖暴力分子持枪抢劫、武装攻击等违法行为[8]. 在医疗监护方面,可监控病人在康复训练中动作是否规范[9],从而保证训练的有效性. 在智能监控方面,可监控监狱中犯人行为是否正常[10]. 在产品实现方面,海康威视(Hikvision)研发了 DeepinMind NVR,利用深度学习技术可实现对人体异常行为的检测[11];苏州爱可尔智能科技研发出玄目 AI 照护系统,通过实时重建人体 3D 骨架,可识别攀爬、跌倒、进入禁区等老年人重点行为[12]. 人体异常行为研究的最终目标是解放人眼,替代传统监控系统存在的低识别率和高漏检率并完成实时的自动预警任务,因此该项研究具有重要现实意义.

异常行为识别方法可分为传统方法和基于深度学习的方法. 传统方法利用手工特征和规则提取特征,以时空关键点、密集轨迹算法为代表,多使用支持向量机(SVM)、贝叶斯、隐形马尔科夫等传统的机器学习分类器. 该方法存在人工提取特征高度依赖设计者的先验知识、难以应用到复杂场景等问题. 基于深度学习的方法借助深度神经网络,大多融合注意力机制,通过端到端的学习来提取视频特征,具有良好的特征提取能力和泛化能力.

多数研究将异常行为检测任务转换为二值分类问题或依据异常概率、得分数值来检测异常. 在训练数据缺乏标签的情况下,利用聚类、帧重建、未来帧预测、生成网络及混合模型的方法来检测异常行为,并且取得了不错的成果. 但是,视频中复杂背景、遮挡、光照及视角变化等因素给异常检测带来困难.

随着深度传感器和人体姿态估计技术的不断成熟,获取骨架数据变得更加准确便利. 此外,骨架数据在计算和存储方面也是有效的[13],这使得人体骨架的相关工作受到了广泛关注.

已有学者在异常行为识别与检测领域开展了大量工作,但基于人体骨架这一立足点的工作仍处于探索阶段. 已有文献分别侧重于基于深度学习的视频异常检测方法[14-15]、人体行为识别数据集[16]、基于半监督学习的视频异常检测方法[17]、基于深度学习和表示学习的视频异常检测方法[18]以及时空特征提取和行为建模的角度对相关方法进行综述[19],但缺乏聚焦人体骨架的整理工作. 为进一步推动此任务的进展,本文归纳并分析了深度学习背景下基于骨架的人体异常行为识别与检测相关方法.

1　人体姿态估计

人体骨架通常建模为关节点(个数介于 10～30)位置及其连接形成的肢体部分,如图 1 所示. 也可描述为拓扑图结构,其中顶点指人体关键点部位,如头部、肩膀、手部、膝盖、脚部等;边指的是关键点的先验连接,如大小臂、左右腿等. 骨架拓扑图结构简单且灵活,很大程度上可以表征行为信息.

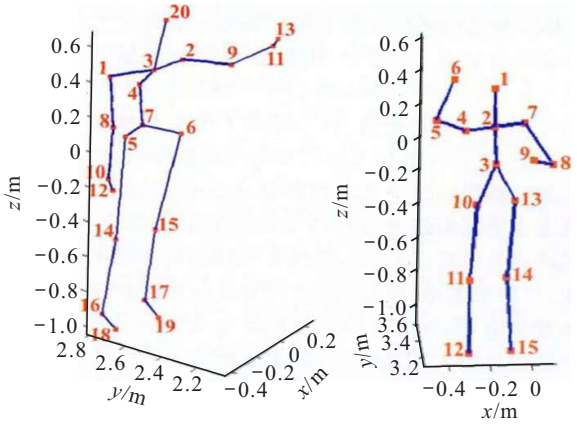


图 1　包含 15 个和 20 个关节点的人体骨架示意图[20]

在实验室环境下通过使用物理设备(如微软的深度摄像机 Kinect、嵌入各种传感器的可穿戴设备、红外摄像头等)来获取人体骨架的方式易受噪声及遮挡的影响,从而导致模型泛化能力弱,难以迁移到自然环境中.

近年来,基于深度学习的姿态估计方法陆续被提出,将对人体姿态的估计转化为对关键点的检测和识别问题. 首先定位人体各关键点的位置坐标,再根据先验知识确定其空间位置关系,最后连接各关键点得到人体骨架. 该方法的准确度很大程度上决定了人体异常行为识别与检测的精度. 按照应用场景中目标个数的不同将姿态估计方法分为单人姿态估计和多人姿态估计.

1.1　单人姿态估计

单人姿态估计算法首先检测并识别出人体各关键点,然后将其自然连接以构建人体骨架,这种算法并不适用于现实生活中的大多数场景.

Toshev 等[21]提出一种基于深度神经网络的,使用级联方式的姿态估计器 DeepPose. 该方法包含 3 阶段的处理,每一阶段都将已检测到的关键点作为后续网络的输入,逐步得到更高分辨率的图像及更高的精度. 然而,该方法不适用于原始图片分辨率较小的情

况. Chu 等[22] 采用堆叠的 Hourglass 模型[23] 来生成具有不同语义的多分辨率特征的注意力图, 融入上下文注意力机制, 使模型能够关注从局部显著区域到全局语义一致性空间的不同粒度特征, 从而提升识别精度. Yang 等[24] 以 Hourglass 为基础, 在姿态估计网络中加入多尺度特征金字塔, 利用金字塔残差模组 PRMs 来增强深度卷积神经网络尺度上的不变性, 以解决由于遮挡、相机视角变化导致关键点定位不准确的问题.

1.2 多人姿态估计

对于多人姿态估计算法, 通常可根据预测的出发点及执行顺序分为自顶向下、自底向上和单阶段 3 种方式.

1.2.1 自顶向下方式(top-down)

自顶向下的方法从高层抽象开始, 首先检测出人体并以边界框标记, 然后再分别对每个人进行姿态估计. 该方法检测精度较高但极度依赖人体边界框的检测质量, 且耗时严重. 此外, 自顶向下方法的计算成本与图像中目标的数量成正比.

Chen 等[25] 提出级联金字塔网络 CPN 来缓解遮挡场景中的关键点检测问题. 该网络包括 GlobalNet 和 RefineNet 两个阶段, 分别定位简单的关键点和整合多尺度特征, 通过关键点挖掘损失及扩大感受野来优化对关键点的检测, 该方法能够兼顾人体关节点的局部信息和全局信息. 为解决检测框的错误定位以及重复问题, Fang 等[26] 提出一种多人姿态估计框架, 利用对称空间变换网络 SSTN 提高检测框精度, 并提出参数化姿态非极大值抑制 NMS 来解决冗余问题. 该方法不适用于远距离人群及拥挤场景. Li 等[27] 通过候选点的概念, 设计对应的候选损失, 抑制非当前人体实例的关键点, 实现对拥挤人群关键点的提取, 其网络模型如图 2 所示. 该方法提升了模型的泛化能力, 但仍不可避免非当前人体的关键点权重对节点分类造成的影响.

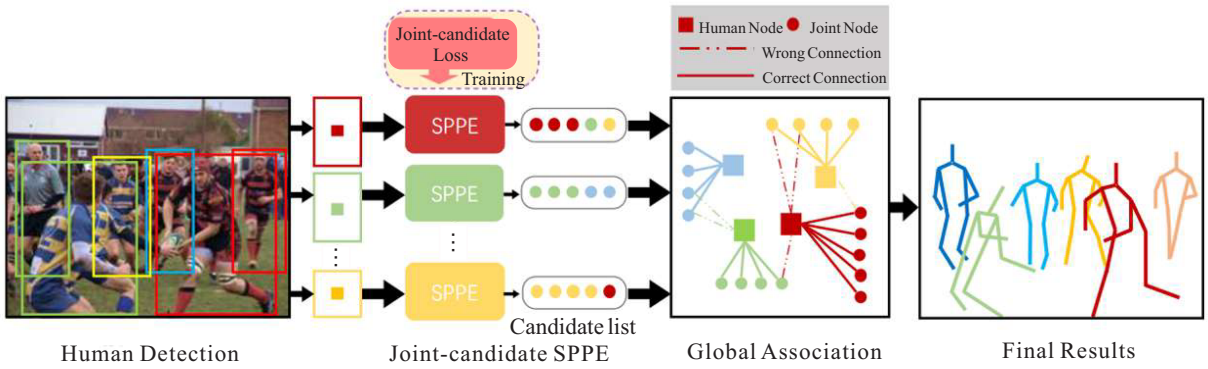


图 2 拥挤人群关键点提取模型[27]

为缓解遮挡、外观及尺度变化带来的影响, Xu 等[28] 为姿态估计任务提供了一个简单且有效的基线模型 ViTPose, 利用原始 Transformer 提取特征图, 并设计嵌入反卷积、预测层的轻量级解码器和热图回归进行姿态估计. 该模型结构较为简单、易扩展、泛化能力强. 为解决此类方法受边界检测框限制, 难以处理遮挡及大幅度姿态变化等问题, Qiu 等[29] 提出循环结构引导注意力网络 SGAN, 通过优化联合置信图和联合关联图编码多尺度表示以增强特征提取, 提高模型的检测精度.

摄像机视角的变化及人体行为的复杂多样造成关节点相对位置的多变, 导致检测精度低下, 而多尺度融合可以很好地应对这一问题. Ke 等[30] 引入多尺度监督网络 MSS-Net 和多尺度回归网络 MSR-Net, 利用结构感知损失从多尺度特征中学习人体骨架结构, 结合丰富的多尺度特征, 通过跨尺度特征匹配提高关键点定位的鲁棒性. 该方法能够适应复杂多人场景下的遮挡情况.

Li 等[31] 通过构建级联 Transformer 提出一种基于回归的姿态识别方法. 编码器将 CNN 生成的图像特征进行压缩以产生上下文特征; 解码器推理特征之间的关系, 并行化输出所有查询对象; 最后将检测目标分类为人或背景, 并预测边界框. Li 等[32] 提出一种基于 Token 的人体姿态估计方法 (TokenPose), 其中, 每个关键点被当作一个标记, 以同时学习约束关系和来自每一帧中图像的外观特征. 同时, 利用 Transformer 中的多头自注意力机制学习不同关键点之间的关联性, 极大减少了模型的参数量和计算量. Papandreou 等[33] 提出用于多人检测的 PoseNet 模型, 使用 Faster-RCNN 进行快速检测, 并通过全卷积预测每个关键点的密集热图和偏移量, 提升检测性能.

1.2.2 自底向上方式(bottom-up)

自底向上的方法首先定位输入图像中每个人的所有关键点部位,然后通过人体模型拟合或其他算法对其进行分组. 此类方法检测速度快,适用于实时性检测任务,但当图像中多人距离较近乃至重叠时,易将关键点错误分配.

Cao等[34]提出OpenPose算法,使用部分亲和域PAFs将身体部位与图像中的个体相关联. 通过自底向上的解析步骤,在增加网络深度的同时消除了对身体部位的置信度的细化,实现实时性和高精度. 该方法可以区分如手臂重叠的交叉情况,为后续工作做出了很大的贡献. 然而,当面对人体罕见姿态、物体干扰时会出现检测缺失,并且会错误识别雕像或动物.

Xie等[35]将Hourglass作为主干网络来进行特征提取,并对不同阶段中相同尺度的输出层进行信息融合,解决了人体尺度多变的问题. 为解决Hourglass在下采样过程中造成部分信息丢失的问题,Li等[36]抛弃了Hourglass的每个阶段的连接方式,提出多阶段位姿估计网络MSPN,在下采样过程中提高通道数;使用U-net连接方式,有效进行信息传递.

为充分建模关键点间丰富的语义关联,Qiu等[37]设计轻量级动态图卷积模型DGCN,通过捕捉多层次信息以动态适应不同人体骨架结构,从而增强模型对于遮挡或复杂动作的鲁棒性.

热图回归方法被广泛应用于多人姿态估计中,通过将二维高斯核作用到关键点上来构造热图,热图上的像素值即为对应像素点作为关键点的概率,热图回归的方法相较于传统的坐标回归方法具有更高的定位准确率. Zhang等[38]构建一个轻量级的空间Transformer网络STN,通过关键点热图的方式,利用快速姿态蒸馏FPD模型训练方法将潜在的知识从一个预训练的较大模型转移到构建好的轻量级网络中,使得模型具有健壮性.

Luo等[39]提出尺度自适应热图回归SAHR方法,自适应调整每个关键点的标准差,同时引入权重自适应热图回归WAHR来平衡前景目标,极大地提高了姿态估计的准确性. Yang等[40]基于Transformer和低层卷积块构建了一个TransPose模型,其中的注意力层可以捕获空间上关节点之间的远距离关系. 首先,利用CNN作为骨干网络来提取特征图并将其展开为序列;接着,Transformer编码器迭代地从序列中捕获依赖项来预测关键点热图.该方法可以有效应对遮挡情形.

自顶向下的检测方法本质上是基于检测框的单人姿态估计问题,虽精度高但实时性差,且小尺度图像受限,所需计算成本高. 而自底向上方法的复杂度不会随着人数的增多而增大,在实时性能方面具有优势,但该方法的精度不如自顶向下方式,且难以处理拥挤问题.

1.2.3 单阶段方式(single-stage)

最近,研究者提出单阶段的姿态估计算法,以端到端的方式从空间位置密集回归一组候选姿态,其中每个候选姿态由来自同一个体的关键点位置组成. Nie等[41]提出结构化的姿态表示方法来统一人体实例和人体关键点的位置信息,采用分层策略,有效划分相邻关节之间的长距离信息,从而估计多人位姿. 该模型推理速度较快且检测性能较好,可灵活泛化到2D及3D场景. Duan等[42]提出基于三元组的单阶段关键点检测方法CenterNet,将回归的关键点位置与从关键点热图中检测到的最近关键点进行匹配,提高了关键点检测精度和召回率. Miao等[43]提出将特征金字塔网络FPN作为主干网络的多人姿态回归方法SMPR. 该方法遵循密集预测,包含初始化阶段和细化阶段,分别回归初始姿态和优化姿态. 此外,姿态评分模块可解决非极大值抑制NMS中的排序问题,从而进一步提高检测性能. Shi等[44]将姿态估计问题视为分层集合预测问题,提出利用关节解码器学习人体关节间运动特征的多人姿态估计框架PETR,并引入注意力机制,自适应关注与目标关键点最为相关的特征,提高了检测性能.

为充分获取实例间的全局时空上下文信息,Qiu等[45]提出实例引导的IVT模型进行人体3D姿态估计. 首先将视频帧描述为一系列包含人体结构信息的引导标记,并将其输入至IVT来建模时空特征,然后通过坐标回归解码3D姿态. 同时,提出跨尺度实例引导的时空注意力机制来有效应对人体尺度变化. Qiu等[46]提出基于渐近式端到端的3D姿态估计模型PSVT. 该方法利用时空编码器捕获全局特征交互,并通过渐近解码机制和姿态引导注意的形状解码器逐步定位人体关节及形状,同时优化网格参数. 此外,采用循环解码过程可降低时空注意力的计算成本,增强解码能力.

2 人体骨架异常行为识别方法

基于深度学习的人体骨架异常行为识别方法是利用神经网络模型从视频帧中提取人体骨架特征,通过训练和学习模型参数进行识别. 按照特征提取骨干网络模型的不同,该方法主要分为基于卷积神经网络(CNN)、循环神经网络(RNN)、图卷积神经网络(GCN)、Transformer网络以及混合模型的方法.

2.1 基于CNN的异常行为识别方法

CNN具有自然提取特征并获取高级语义信息的能力,基于CNN的方法通常将骨架序列重构为一系列伪图像以学习特征. Caetano等[47]提出一种基于运动信息的表示方式SkeleMotion,其网络模型如图3所示. 该方法将骨架空间结构编码为图像表示,通过计算骨架关节的大小及其方向值实现对时序运动信息的捕捉. 同时,采用不同的时间尺度来计算并动态聚合运动值,从而捕获距离较远的关节间的信息. 该模型训练速度快,但由于缺乏高层语义特征,识别率较低,在NTU RGB+D60数据集CS、CV两种标准下的识别准确率分别达到76.5%、84.7%;在NTU RGB+D120数据集中C-Sub、C-Set两种标准下分别取得67.7%、66.9%的准确率.

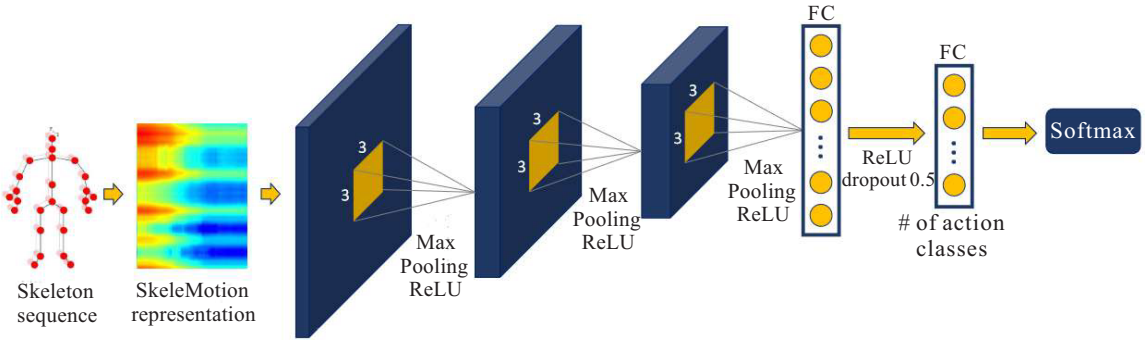


图 3 SkeleMotion网络模型[47]

为了更好地表征骨架序列的时空信息,双流网络被广泛应用于此任务中. Li等[48]利用CNN自动从骨架序列中学习分层共现特征,聚合不同层次的上下文信息,独立学习每个关节点的特征并将其组合为空间域和时间域的语义表示,采用双流网络模型整合原始骨架坐标及其时序差异,从而充分利用关节间的交互信息,最后通过Softmax分类器进行分类. 该方法在NTU RGB+D60数据集上CS、CV两种标准下的识别准确率分别达到86.5%、91.1%. 此外,引入双流框架融合骨架运动特征的方法显著提高了异常行为识别和检测任务的性能,为后续工作提供了参考.

Liu等[8]通过YOLOv4提取人体3D骨架关节点,使用多目标跟踪技术和双目摄像机提取每帧中目标的3D骨架信息,并将其转换为RGB信息,再输入到多层CNN中进行识别. 该模型主要针对特殊安全场景,识别持枪、武装攻击、投掷、攀爬等异常行为的平均准确率达到89%,mAP达到68.9%,识别帧率达到13FPS,即可实现实时、准确地检测.

基于3D-CNN的方法可以同时学习骨架的时空特征信息,解决了使用双流网络在时空特征融合时信息丢失的问题. Duan等[49]提出PoseC3D来识别摔倒等异常行为,通过姿态估计算法提取2D人体骨架,并沿时间维度堆叠关节或肢体的热图,生成3D热图体,最后使用3D-CNN进行识别. 该方法在NTU RGB+D60数据集CS、CV两种标准下的识别准确率分别达到94.1%、97.1%;在NTU RGB+D120数据集中C-Sub、C-Set两种标准下分别取得86.9%、90.3%的准确率.

基于CNN的方法可以提取多尺度的特定局部模式,但存在参数量大、计算要求高的问题,且无法充分捕捉空间全局信息及时序运动信息.

2.2 基于RNN的异常行为识别方法

基于RNN的方法通过将每帧关键点的坐标拼接成向量,并将其串联起来实现对上下文信息的建模. 由于RNN建模长序列时易出现梯度消失的问题,有研究将RNN扩展至长短时记忆网络(LSTM)[50]. 该网络在建模序列数据的依赖性和动态性方面具有优势,常被用来捕捉视频时域上的运动信息.

多数研究者融合LSTM与注意力机制以充分学习特征. Liu等[51]构建全局上下文感知注意力LSTM网络GCA-LSTM,包括两个全局情景记忆单元和两个LSTM层. 该模型提出的关节点级的细粒度注意力机制和身体部分级的粗粒度注意力机制可经多次迭代逐步提高模型捕捉关键信息的能力. 该方法在NTU RGB+D60数据集CS、CV评价标准下的识别准确率分别为76.1%、84.0%.

为消除相机视角变化造成的影响,Zhang等[52]以LSTM为基本单元,设计出一种视图自适应神经网络,通过视图自适应模块调整每一帧的观察视点来学习并确定最适合的视点,然后将各种视图的骨架转换为更一致的虚拟视点,从而学习具体的行为特征并简化训练,其网络模型如图4所示. 但由于不同行为在经过处理后会转变为朝向一致的视图,丢失了部分有价值的运动信息,从而会造成误判. 该方法在NTU

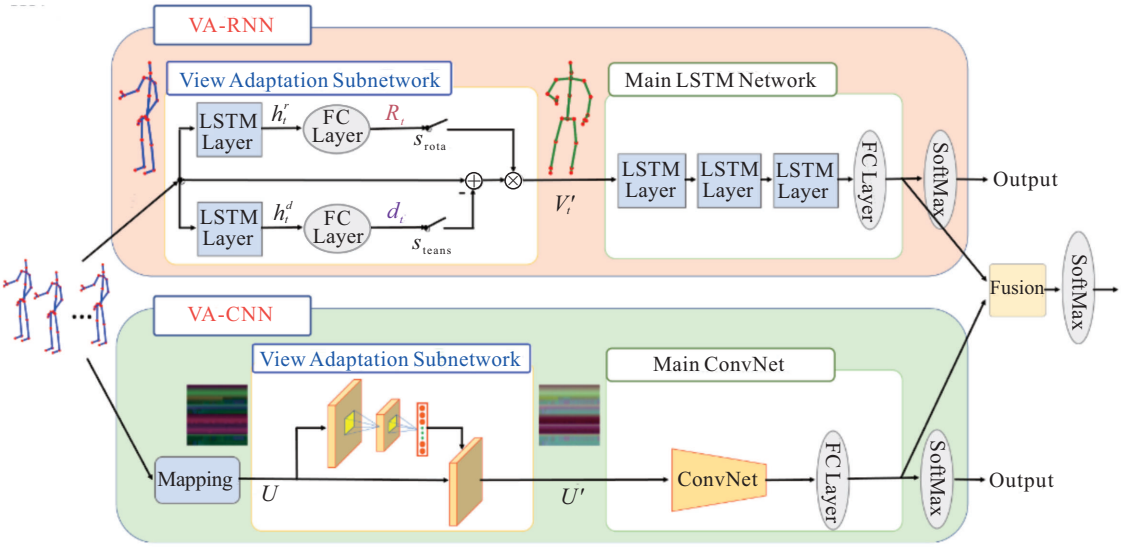


图 4　视图自适应网络模型[52]

RGB+D60 数据集 CS、CV 评价标准下的识别准确率分别为 89.4 %、95.0 %.

与 CNN 相比, LSTM 网络可以更好地建模运动信息及时间依赖性, 但没有充分利用骨架的先验知识, 且计算成本较高.

2.3　基于 GCN 的异常行为识别方法

由于人体骨架是一种非欧式的拓扑图结构, 近年来, 多数研究通过图卷积网络 GCN 来捕捉以骨架信息表征的人体行为特征. 基于 GCN 的方法主要分为基于频域和基于空域的方法[53]. 基于频域的方法将处于空间域的图信号变换到频域, 并对频域属性进行滤波, 然后再恢复到原来的图信号所在的空域, 从而完成特征提取[54]. 该方法灵活性低、普适性差、运行效率低. 而基于空域的方法通过信息聚合, 将顶点及其邻居节点的特征进行卷积, 从而更新顶点特征. 该方法中图卷积操作本质上是沿着边传播顶点信息, 增强了泛化能力, 降低了计算复杂度, 提高了运行效率, 逐渐成为主流的方法. 基于 GCN 的方法通常将骨架结构中的关节和骨骼视为顶点和边, 自然地保持了人体骨架的先验结构.

由于连续帧中存在过多的冗余信息, Tang 等[55] 提出一种基于骨架的深度渐近式强化学习方法 DPRL, 通过强化学习从输入序列中提取固定数量的关键帧, 再将其输入到 GCN 中来识别行为. 该方法减少了计算量并简化了模型, 但提取关键帧的过程中仍可能会丢失某些重要的时域信息. 该方法在 NTU RGB+D60 数据集 CS、CV 评价标准下的识别准确率分别达到 83.5 %、89.8 %.

Yan 等[56] 首次将基于图的神经网络应用于人体骨架的行为识别研究, 提出时空图卷积网络 ST-GCN (其网络模型如图 5 所示), 由空间和时间卷积模块组成, 将人体骨架建模为骨架序列时空图. 该方法利用图卷积的局部性和时间动态来隐式地学习局部信息, 具有较强的表达能力和泛化能力. 由于图卷积的感受野有限, 该方法只能建模相邻关节点的空间特征与短时运动特征, 限制了对全局时空特征的学习. 该方法在 NTU RGB+D60 数据集 CS、CV 评价标准下的识别准确率分别达到 81.5 %、88.3 %.

受 ST-GCN 提取骨架时空信息的启发, 后续许多工作以其为基础并进行了优化. 为增强模型对全局上下文时序信息的学习能力, 曹毅等[57] 在 ST-GCN 基础上做出改进, 提出时空自适应图卷积神经网络 ST-AGCN, 实现了骨架的空间结构特征的提取和全局上下文时间信息的建模, 并构建残差结构来确保模型的

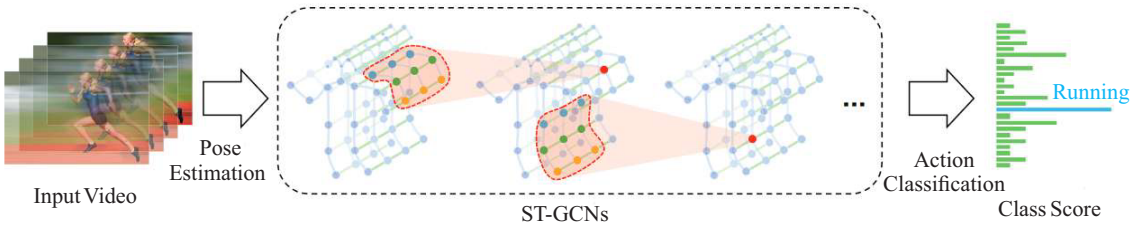


图 5　ST-GCN 网络模型[56]

稳定收敛. 该方法在NTU RGB+D60数据集CS、CV评价标准下的识别准确率分别达到86.4%、92.1%.

为增强所捕获信息的有效性,注意力机制与图卷积网络相结合的策略被广泛应用. 曹毅等[58]提出一种三维图卷积与注意力增强的异常行为识别模型,该三维图卷积核具有时间与空间两个采样维度,可实现对骨架序列中时空信息的有效提取;引入注意力增强结构提高了对于特定关节点的关注度,使模型聚焦于重要的、关键的行为信息. 该方法在NTU RGB+D60数据集CS、CV评价标准下的识别准确率分别为89.43%、93.3%.

除在骨架空间关节点上使用注意力机制,部分研究者将其作用于时域及通道域以增强特征提取. 为充分挖掘时空特征,曹毅等[59]提出一种图卷积行为识别模型STFE-GCN,分别在空域和时序上使用图注意力机制及混合池化模型,并改进通道注意力网络,实现时空特征及全局上下文特征的提取. 该方法在NTU RGB+D60数据集CS、CV评价标准下的识别准确率分别为89.8%、96.0%;在NTU RGB+D120数据集C-Sub、C-Set评价标准下的识别准确率分别为84.1%、86.3%.

上下文信息对于视频中异常行为识别至关重要. Ye等[60]将上下文编码网络嵌入图卷积层,通过堆叠多个上下文编码网络构建动态GCN,实现端到端的学习. 然而,将图卷积神经网络进行深度堆叠易导致过平滑问题. 该方法在NTU RGB+D60数据集的CS、CV评价标准下分别达到91.5%、96.0%的准确率.

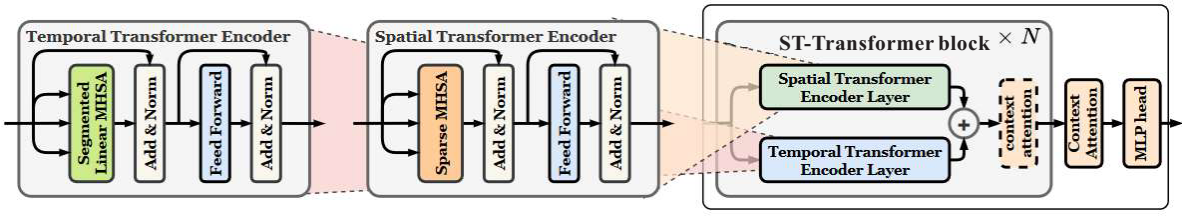
深度信息作为二维信息的补充,也被应用于此研究中. Ding等[61]提出一种基于时空关系的骨架行为识别模型以检测学生在使用自动扶梯时是否发生攀爬、摔倒等异常行为. 该方法采用基于单目相机的深度估计方法获取关节点的深度信息,并将其与通过OpenPose提取的二维骨架坐标相结合形成三维骨架数据;之后,利用图卷积神经网络及滑动窗口投票法对采集的人体骨架特征进行分类. 该方法在自建数据集上的识别准确率达到99.5%,但应用场景较为单一.

图卷积网络因其参数量小,同时高度拟合人体骨架图结构而被广泛应用. 虽然此类方法在性能上取得了突破,但仍存在一些缺陷:卷积操作无法收集不直接相连的全局关节信息;重复使用图卷积可以得到多跳依赖,但计算复杂度和优化难度随之增加. 除此之外,图卷积网络直接处理坐标导致其识别能力易受到坐标分布偏移的影响.

2.4 基于Transformer的异常行为识别方法

Transformer网络是一种基于注意力机制、可并行化处理数据的深度神经网络[62]. 自2017年由Google[63]提出以来,在自然语言处理领域取得了巨大成功,引起了视觉领域研究者的注意,并将其应用于图像分类[64-65]、目标检测[66]、语义分割[67]、视频理解[68]等多类计算机视觉任务. 同时,Transformer及其衍生模型VTN[69](video transformer network)也逐渐被应用于行为识别任务[70-75].

在视频异常行为识别领域,Transformer需要同时学习骨架的空间先验信息和时序动态信息. 针对可变长骨架序列,Shi等[72]提出一种基于稀疏Transformer的行为识别模型STAR,其网络模型如图6所示. 该模型由空间编码器和时间编码器组成,分别应用稀疏注意和分段线性注意机制以减少计算量和内存资源. 该方法在NTU RGB+D60数据集的CS、CV评价标准下分别达到83.4%、89.0%的准确率;在NTU RGB+D120数据集的C-Sub、C-Set评价标准下分别获得78.3%、80.2%的准确率.



(a) Temporal Transformer Encoder Layer: The attention module is a Segmented Linear MHSA

(b) Spatial Transformer Encoder Layer: The attention module is a Spare MHSA

(c) The architecture of our model: Consisits of N ST-Transformer block

图6 STAR网络模型[72]

针对传统卷积中感受野小的局限性,Sun等[76]使用Transformer替代图卷积,以较低的复杂度保持图的拓扑结构,分别将4种骨架序列输入相应的时空相对Transformer网络ST-RT中进行特征提取、融合及预测. 该模型结合不同尺度的运动信息,提高了泛化程度. Plizzari等[77]提出一种新的时空Transformer

网络 ST-TR, 包含空间自注意力模块 SSA 和时间自注意力模块 TSA, 分别建模帧内不同关节之间的相互作用及连续帧间同一关节的运动特征. 该方法在 NTU RGB+D60 数据集的 CS、CV 评价标准下分别达到 88.7 % 和 95.6 % 的准确率; 在 NTU RGB+D120 数据集的 C-Sub、C-Set 评价标准下分别达到 85.1 % 和 87.1 % 的准确率.

针对在空间和时间上关节语义信息不一致的问题, Qiu 等[78] 提出一种时空元组 Transformer, 利用时空注意力捕获每个元组中关节的相关特征, 并利用帧间特征聚合模块将其集成. 该方法可建模连续帧中不同关节之间的关系, 利于区分相似行为, 从而提高识别准确率. 其在 NTU RGB+D60 数据集的 CS、CV 评价标准下分别达到 92.3 % 和 96.5 % 的准确率; 在 NTU RGB+D120 数据集的 C-Sub、C-Set 评价标准下分别达到 88.3 % 和 89.2 % 的准确率.

Transformer 借助多头自注意机制增强了特征提取的全面性, 在建模时序行为时具有显著优势, 但缺乏人体各关节间的先验关系且依赖于大规模数据集, 计算量大. SwinTransformer 网络[79] 整合了空间局部性的归纳偏置、层次和平移不变性, 通过移动窗口的方式学习图像多尺度特征, 可有效降低计算复杂度. Video-SwinTransformer 模型[80] 基于 SwinTransformer 架构, 将局部注意力的计算范围从空间域扩展至时空域, 使得其在处理下游视频任务上更灵活. 因此, 使用 Transformer 及其衍生模型开展此研究具有前景意义.

2.5　基于混合模型的异常行为识别方法

CNN 处理骨架数据时, 通常需结合 LSTM 网络以充分获取时空信息. GCN 得益于对非欧式数据建模的优势, 可充分提取人体骨架拓扑图节点的特征. Transformer 网络用以捕捉远距离关节点的空间信息及长时间序列的时序信息. 现有研究结合多个网络的优势, 以捕捉视频中丰富的时空特征及上下文语义特征, 往往可取得比单一模型更好的效果.

将 LSTM 提取的时序上下文运动信息与 CNN 或 GCN 提取的空间结构信息相结合以充分提取时空信息, 有效解决了训练过程不易收敛的问题. Zhou 等[10] 利用目标深度估计算法提取人体在三维空间中的深度信息, 将深度信息和二维骨架坐标组合成三维信息; 然后使用基于时空卷积和注意力 LSTM 提取三维骨架序列信息并进行分类, 其在 NTU RGB+D60 数据集的 CS、CV 评价标准下分别达到 74.1% 和 81.5% 的准确率.

由于 GCN 模型可高效建模骨架空间结构信息, 而 Transformer 网络可捕获视频中的时域动态信息, 最近很多研究将两者与双流网络结合以增强时空特征. Liu 等[81] 提出一种核注意自适应图变网络 KA-AGTN (见图 7), 利用多头自注意力来建模关节间的高阶依赖. 此外, 时间核注意块利用时序特征生成全局通道级注意分数, 增强了时间运动信息. 与传统的特征提取器相比, 骨架图 Transformer 模块可有效捕捉时空信息及远距离依赖关系, 并缓解超平滑问题.

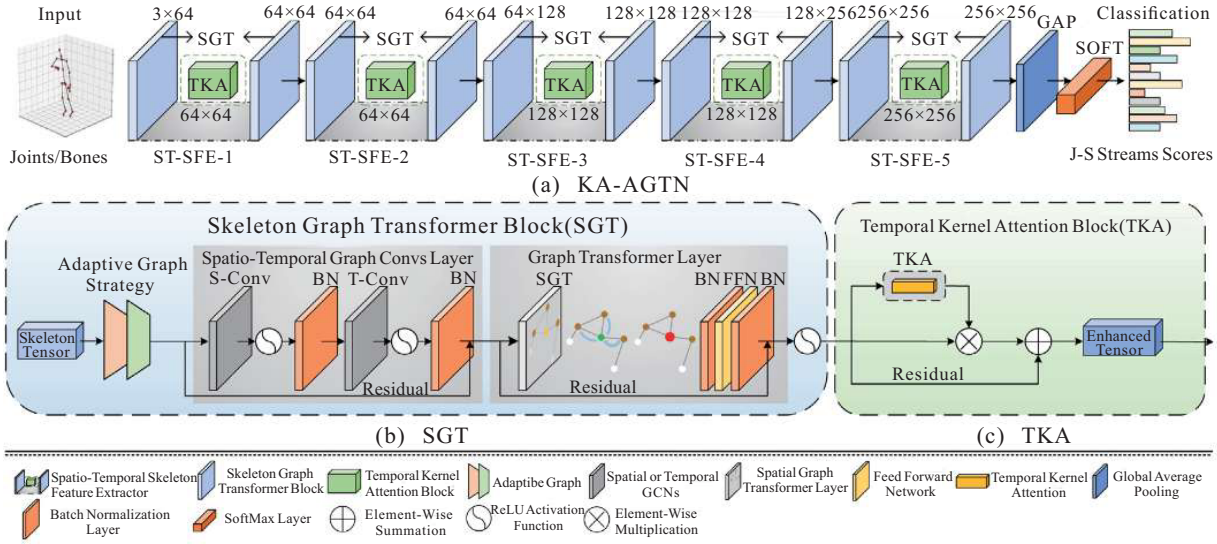


图 7　KA-AGTN 网络模型[81]

3　人体骨架异常行为检测方法

在异常行为识别任务中, 需预先对视频序列进行分割, 并定义好哪一行为发生的起始帧和结束帧. 通过训练模型学习每一类行为的特征, 使其能够准确辨别输入行为的类型, 基于此判断该输入行为是否异常. 而异常行为检测任务则需解决更困难的问题, 即

在未分割的视频序列中准确定位到异常行为. 在行为识别领域,以人体骨架为研究对象的方法得到了广泛应用,并将其扩展到检测任务中.

基于骨架的人体异常行为检测方法结合行为发生的背景环境及上下文时空信息计算异常分数,并依据所设定的阈值准确定位异常行为. 按照训练过程是否涉及数据与标签对应关系的学习,将其分为有监督、弱监督、无监督3种.

3.1 有监督异常行为检测方法

有监督异常行为检测方法旨在使用标记数据分别建模正常和异常行为,通常用于检测训练阶段预定义的特定行为. 在训练网络之前需赋予样本标签,模型通过学习样本和标签之间的关联进行分类.

田联房等[6]通过结合可变形组件模型特征的支持向量机SVM检测乘客人脸并用核相关滤波器KCF对其跟踪,得到乘客在扶梯中的运动轨迹. 通过CNN提取轨迹中乘客的人体骨架序列,并应用骨架方向余弦通过模板匹配检测出异常行为的骨架序列. 杜启亮等[7]使用图卷积神经网络GCN对单帧骨架的异常行为进行分类,然后对同一个乘客进行滑动窗口投票,判断该乘客是否发生异常行为. 相比于使用余弦相似度特征进行模板匹配,GCN通过学习可得到更有利的深度姿态特征,从而提高识别准确率. 该方法适用于人员稀疏场景,在处理由拥挤造成的遮挡问题时效果不好.

Yu等[9]将异常行为检测定义为二值分类问题,设计了堆叠三层时空图卷积网络的模型,每层由空间、时间卷积层以及Drop-out层组成. 首先,通过姿态估计算法提取每一帧中的骨架结构,利用图卷积建模骨架运动数据;然后,从Softmax层生成的概率分布中提取异常得分,并根据设定的阈值来推断是否属于异常. 该方法通过空间自注意力增强模块捕捉关节的局部和全局特征.

由于OpenPose算法使用的VGG卷积层中包含大量的计算,Qiu等[82]提出了用MobileNet作为骨干网络来提取特征,将残差连接块代替原来的卷积操作,从而增强了模型的特征提取能力,然后通过轻量级KNN分类器来判定异常行为.

为提高模型的场景感知能力,Sun等[83]将人体骨架作为运动增强表示,并结合外观特征及背景信息设计场景感知自编码器,通过二分类器来检测异常行为. 该模型有助于处理正常行为模式的多样性,同时可有效区分与场景相关的正负样本.

基于有监督的异常行为检测方法需要从大量有标签数据中学习出一般规律特征,具有较好的效果,但缺乏标注良好的数据集.

3.2 弱监督异常行为检测方法

弱监督方法是指训练集中仅提供正常和异常行为的视频级标签,在测试时根据预测到每一帧的异常得分来判断其是否为异常帧.

多示例学习MIL可应对部分正负样本过于相似的问题. 该方法将视频数据定义为包含多个示例的包,异常视频生成的包称为正包,正常视频生成的包称为负包,正包中至少包含一段异常行为,而负包中不包含异常片段. Feng等[84]提出了一个多示例自训练框架,其中,伪标签生成器采用稀疏连续采样策略产生更可靠的视频片段标签,使用自注意特征编码器自动关注帧中的异常区域,同时提取特定的任务表示,从而提高检测率.

与有监督检测方式相比,弱监督方法仅需视频级的标签,可避免标注训练数据耗时问题.

3.3 无监督异常行为检测方法

无监督方法的目的是只使用正常数据进行训练,从而学习正常行为的特征表示,并根据测试样本与正常样本的偏离程度来检测异常. 多数研究通过聚类、帧重构、未来帧预测、生成网络及混合模型的方法检测异常行为.

3.3.1 基于聚类的异常行为检测方法

聚类模型学习正常行为样本的特征聚类,通过判断输入样本与簇中心的距离来检测异常. Markovitz等[85]首先利用姿态估计算法来提取人体骨架,并通过时空图卷积编码器ST-GCAE捕捉序列中关节间依赖关系,之后使用深度聚类层将潜在向量软分配到聚类,根据异常分数判断该行为是否正常. 该算法既适用于细粒度异常检测来检测单个动作的变化,也适用于粗粒度异常检测来区分正常和异常行为.

Liu等[5]提出利用时空自注意增强图卷积自编码器提取每帧内骨架的局部和全局特征,将其生成潜在向量并嵌入聚类,通过狄利克雷(Dirichlet)过程混合模型以对数概率对每种行为进行评分. 该方法难以区分相似行为,且不适用于处理遮挡场景及交互行为. Hirschorn等[86]设计基于骨架的轻量级时空图归一化流模型STG-NF,可以直接学习数据分布与潜在高斯分布之间的双向映射,并根据其似然度对样本进行评分.

3.3.2 基于预测模型的异常行为检测方法

预测模型通常根据视频序列中前几帧来预测下一帧,即从输入长度L的视频帧中提取行为信息并根

据训练时所学到的正常样本特征来预测第$L+1$帧的骨架信息,之后在预测序列和真实序列之间附加损失函数. 在训练时,最小化预测帧与真实帧间的误差,测试时根据误差及阈值进行判断.

Li 等[87] 提出一种多尺度图神经网络 DMGNN. 该模型包含多尺度图编码器及循环预测图解码器,编码器利用多尺度图计算单元 MGCU 提取单尺度上的特征并将其融合,解码器是基于图的门控循环单元 GRU,根据估计的姿态序列预测下一帧. 该方法在 H3.6M 数据集上的识别率达到 89%. Luo 等[88] 将多个 ST-GCN 进行叠加并引入 Resnet 机制建模关节的时空关系,将其在空间维度和时间维度上的信息累加,最后通过全连接层中的预测模块来预测未来帧,根据预测误差判断异常.

Fan 等[89] 提出一种基于人体骨架的门控循环单元-前馈网络 GRU-FFN. 在前馈网络中引入 GRU 单元,建立具有记忆和权值共享能力的反馈回路,并级联多个 GRU 单元,利用消息传递机制最小化预测帧与其真实帧之间的差异.

由于多数模型缺乏对不同场景的理解,Zeng 等[90] 使用时空图卷积网络提取人体骨架特征,并集成低层与高层图表示即结合骨架序列和运动特征来检测异常,其网络模型如图 8 所示. 低层图表示专注于编码高分辨率视频中人体关节点的时空信息;高层图表示捕获低分辨率视频中个体间的相互作用. 同时,通过边界框及光流来区分密集和稀疏人群,以此适应不同的场景. 该方法易受视角变化影响,且无法区分具有相近移动速度的自行车与行人,易造成误检与漏检.

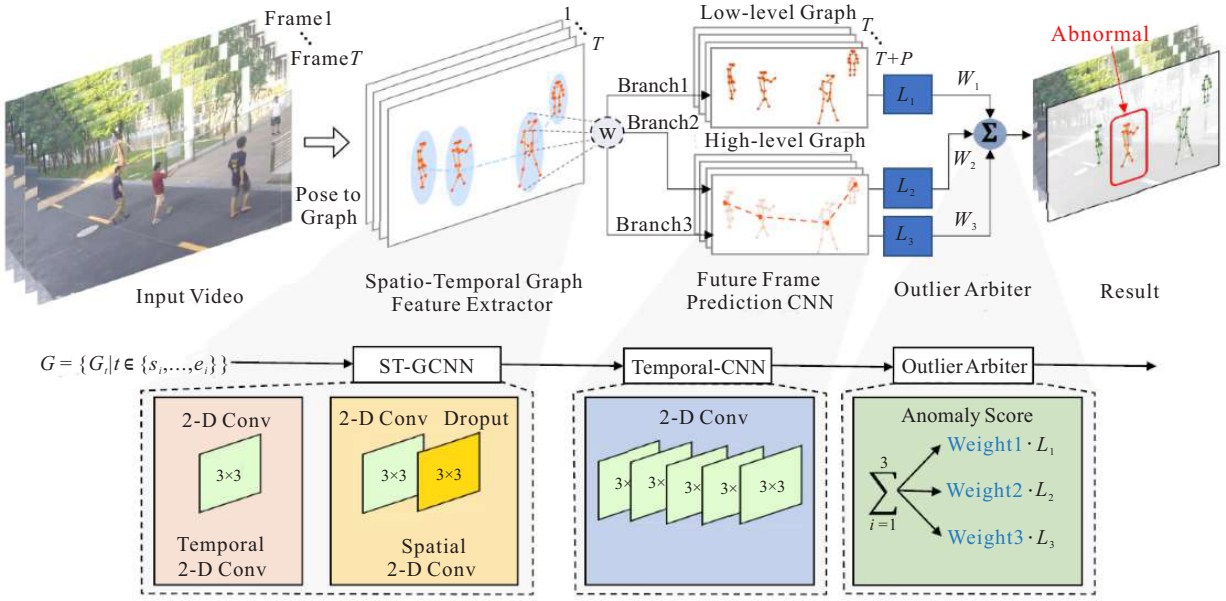


图 8　HSTGCNN 网络模型[90]

为处理正负样本的时序尺度多样性,Rodrigues 等[91] 提出一种多时间尺度模型来捕获不同时间尺度上的动态特征. 该模型对给定的输入姿态轨迹在不同时间尺度上进行未来和过去预测,通过组合预测结果来检测异常行为.

为捕获不同骨架关节间的全局依赖关系,Pang 等[92] 引入 Transformer 网络,根据预测和真实骨架序列之间的误差检测异常. 将原始骨架分解为局部和全局姿态分量,并利用位置嵌入为其添加时序信息. 该方法中多头自注意力机制 MSA 可以从不同角度捕捉任意成对姿态组件之间的依赖关系,时间卷积层 TCL 可以增强模型提取局部时间信息的能力.

为充分捕获个体之间的交互性特征,Huang 等[93] 将骨架特征编码为层次图表示,提出时空图-Transformer (STGformer) 模型来建模不同骨架之间的交互信息和每一骨架中关节点之间的相关性,并通过预测骨架全局及局部部件来检测异常行为.

3.3.3　基于重构模型的异常行为检测方法

基于重构的异常行为检测方法利用输入帧与重建帧之间的误差作为异常检测评分和定位的基础. 该方法多利用自编码器,编码器将输入压缩成一个低维特征表示,解码器从中重构出尽可能接近输入帧的输出,通过重构误差将异常事件和正常事件区分开来,若重构误差大于给定阈值,则认为该输入行为

是异常的.

由于缺乏RGB信息,基于骨架的方法难以获取复杂的动态特征. Yu等[94]提出一种运动先验规则学习器MoPRL来解决此问题. 该模型由运动嵌入器ME和时空TransformerSST两个模块组成,其中,ME将骨架的时空信息建模为概率分布,而SST学习来自ME的骨架时空特征并对骨架轨迹的规律性进行重建.

3.3.4 基于生成模型的异常行为检测方法

近年来,生成对抗网络(GAN)由于能够应对数据稀疏及不平衡问题而被广泛应用于视频异常检测,并且多结合光流来捕捉个体运动特征[95-97]. 生成对抗网络由生成器和鉴别器组成,生成器用来学习正常行为模式并以最小误差重构正常行为,鉴别器用来区分输入骨架序列属于真实序列或生成序列.

为缓解单一GAN提取特征的不准确性,Fan等[98]结合重构思想,首次将CycleGAN用于视频异常检测. 该模型仅使用正常行为样本及相应的骨架特征来训练生成器和鉴别器,使得周期一致性损失尽可能小,以达到学习正常行为特征的目的. 在测试阶段,通过设定阈值过滤重构误差较大的异常行为. 但该方法对慢动作变化不敏感,可能造成误检.

使用GAN模型的方式保留了动态骨架特征,然而,如何自适应调节该模型的泛化能力仍值得深入探究.

3.3.5 基于混合模型的异常行为检测方法

由于训练数据只包含正常行为样本,基于预测和重构的方法在测试时输入异常行为的误差大于正常行为. 由于自编码器具有强泛化能力,异常帧有时能被很好地重建,从而造成漏检或误检. 而结合预测方法与重构方法可以缓解这一问题,或在预测的前提下进行重构以扩大异常误差,或将预测误差和重构误差加权后依据异常分数进行判别.

Morais等[99]结合稀疏光流与人体骨架,并使用匈牙利算法求解相邻帧间骨架的相似性评分. 该模型由双循环编码器-解码器网络(重构解码器和预测解码器)组成,通过跨分支消息传递机制建模全局和局部组件间依赖关系,最后结合重构误差和预测误差来判断异常. 该模型缓解了自编码器易过于泛化的问题.

Li等[100]提出采用单编码-双解码结构的时空图卷积自编码器ST-GCAE-LSTM模型,并嵌入LSTM以充分建模潜在向量的时序特征. 其中,双解码器为重构解码器和预测解码器,分别用于重建输入骨架序列以及预测未来骨架序列,实现在重构输入的同时预测未来帧.

为增强GAN模型在训练过程中的稳定性,Li等[101]将人体骨架分解为全局和局部组件,提出带有梯度惩罚的记忆增强Wasserstein生成对抗网络MemWGAN-GP,通过反向学习拟合生成样本分布和真实样本之间的Wasserstein距离,最终联合重构误差和预测误差检测异常行为.

无监督异常行为检测方法只学习正常行为的分布,假设未知的异常行为具有很高的重建和预测误差. 但由于缺乏关于异常行为的先验知识,无法学习所有正常的行为模式,模型在不同场景下具有一定偏差.

4 性能评估指标及相关数据集

在异常行为识别任务中,常用的指标是精确度(ACC),对应于正确分类行为的比例. 在检测领域,两种常用的标准是等错误率(EER)及ROC曲线下面积(AUC),这两个标准来自受试者工作特征(ROC)曲线,该曲线通过对异常分数或异常概率取不同的阈值进行绘制,被广泛应用于性能比较. EER是ROC曲线上假阳性率(正常行为被判定为异常行为)与假阴性率(异常行为被判定为正常行为)相等的点. 对于一个好的异常检测算法,EER应尽可能小,AUC尽可能大.

4.1 异常行为识别数据集

异常行为识别方法的大量提出表明了对异常行为识别领域的广泛研究,同时,对评估各类方法的公共数据集的需求日益增长. 已有学者对人体行为相关数据集做出总结[16],此任务中与人体骨架相关的典型数据集有:SBU交互[102]数据集、Kinetics[103]数据集、NTU RGB+D 60[104]数据集、NTU RGB+D 120[105]数据集、Human3.6M[106]数据集和KTH[107]数据集. 部分算法在典型骨架数据集上的ACC(%)表现对比分析如表1所示.

此外,为进一步说明基于人体骨架的异常行为识别方法的实时性,本文对部分算法提及的参数量、浮点运算次数FLOPs及模型搭载硬件设备的对比分析如表2所示. 其中,参数量和FLOPs表征模型的计算复杂度,硬件设备型号表征算力要求. 分析表2可知,模型的参数量及FLOPs总体呈降低趋势,表明模型趋于轻量化,更适用于此类实时性识别任务.

表 1　部分异常行为识别算法在典型骨架数据集上的ACC (%)表现

方法	年份	NTU60		NTU120		
		X-Sub	X-View	X-Sub	X-View	X-Set
GCA-LSTM[51]	2018	76.1	84.0	–	–	–
SkeleMotion[47]	2019	76.5	84.7	67.7	66.9	
VA-RNN[52]	2019	79.8	88.9	–	–	–
ST-GCN[56]	2018	81.5	88.3	81.9	–	71.3
STAR[72]	2021	83.4	89.0	78.3	–	80.2
DPRL+GCNN[55]	2018	83.5	89.8	–	–	–
SR-TSL[108]	2018	84.8	92.4	–	–	–
ST-GCN+PM-STFGCN[109]	2020	85.4	89.7	–	–	–
STA-GCN[110]	2021	86.3	93.7	–	–	–
ST-AGCN[57]	2020	86.4	92.1	–	–	–
HCN[48]	2018	86.5	91.1	–	–	–
CA-GCN[111]	2020	86.5	94.1	–	–	–
ATT+ST-GCN[112]	2021	87.2	94.7	–	–	–
VA-CNN[52]	2019	88.7	94.3	–	–	–
ST-TR[77]	2021	88.7	95.6	85.1	–	87.1
AGC-LSTM[113]	2019	89.2	95.0	90.2	–	96.2
JT-AGCN[114]	2021	89.3	95.5	82.9	83.7	–
ATT+3DGCN[58]	2021	89.4	93.3	–	–	–
PTF-SGN[115]	2022	89.7	95.2	81.3	–	83.5
SGN+GCN[116]	2021	89.9	94.7	82.1	–	83.8
KA-AGTN[81]	2022	90.4	96.1	86.1	–	88.0
Shift-GCN[117]	2020	90.7	96.5	85.9	–	87.6
DynamicGCN[60]	2020	91.5	96.0	87.3	88.6	–
2s-AGCN+PM-STFGCN[109]	2020	91.7	96.2	–	–	–
STTFormer[78]	2022	92.3	96.5	88.3	–	89.2
2s-AAGCN[118]	2022	92.3	97.5	–	–	–
CTR-GCN[119]	2021	92.4	96.8	88.9	–	90.6
ST-GAT[120]	2022	92.8	97.3	88.7	–	90.4
PoseC3D[49]	2022	94.1	97.1	86.9	–	90.3

表 2　部分异常行为识别算法参数量、浮点运算次数及硬件设备对比

方法	年份	参数量(M)	FLOPs(G)	硬件设备
VA-CNN[52]	2019	24.09	–	–
AGC-LSTM[113]	2019	22.90	54.40	–
DynamicGCN[60]	2020	3.60	1.99	–
ST-GCN[56]	2018	3.10	3.56	8 NVIDIA GTX TITAN X GPUs
KA-AGTN[81]	2022	2.70	–	NVIDIA Tesla V100 SXM2 32GB GPU
ST-AGCN[57]	2020	–	–	NVIDIA GTX 1080Ti
CA-GCN[111]	2020	1.87	–	–
ST-TR[77]	2021	1.74	–	–
CTR-GCN[119]	2021	1.46	–	–
STAR[72]	2021	1.26	–	2 NVIDIA GTX TITAN X GPUs;TITAN RTX GPU
STTFormer[78]	2022	–	–	2 NVIDIA GTX 3090 GPUs
PTF-SGN[115]	2022	0.79	–	–
SGN+GCN[116]	2021	0.77	–	–
VA-RNN[52]	2019	0.47	–	–
PoseC3D[49]	2022	0.24	0.60	–

4.2　异常行为检测数据集

由于视频异常行为检测是一个相对较新的研究领域, 公开数据集较少. 此外, 在基于骨架的相关任务中, 提取骨架工作要求视频分辨率较高的大规模数据集. 然而, 在现实场景中, 异常行为具有多样性且大多需关联背景信息, 故异常行为难以被准确定义. 因此, 目前可用的基准数据集较少, 常用数据集包括UCSD[121]数据集、CUHK Avenue[122]数据集及ShanghaiTech[123]数据集. 其中, UCSD数据集因其分辨率较低, 不适用于提取骨架. 部分算法在典型的视频异常检测数据集上的帧级AUC (%)表现对比分析如表3所示, 部分算法搭载硬件设备型号如表4所示.

表 3 部分异常行为检测算法在典型数据集上的帧级AUC (%)表现

方法	年份	ShanghaiTech	HR-ShanghaiTech	CUHK Avenue	HR-Avenue
MPED-RNN[99]	2019	73.40	75.40	–	86.30
Normal Graph[88]	2021	74.10	76.50	87.30	–
Markovitz[85]	2020	75.20	–	–	–
MemWGAN-GP[101]	2023	75.30	78.00	83.70	88.50
STGCAE-LSTM[100]	2022	75.60	77.20	83.70	86.30
MSTA-GCN[124]	2023	75.90	76.80	87.60	–
Rodrigues[91]	2020	76.03	77.04	82.85	88.33
Skeleton-Transformer[92]	2022	–	77.65	–	86.70
GraphConv-Clustering[5]	2022	78.90	79.30	88.40	–
MoPRL[94]	2021	81.26	82.38	–	–
HSTGCNN[90]	2023	81.80	83.40	87.51	88.65
GRU-FFN[89]	2021	82.60	–	91.70	–
CycleGAN[98]	2022	–	–	87.80	90.20
STGformer[93]	2022	82.86	86.97	88.83	89.67
HSC[83]	2023	83.40	–	93.70	–
STG-NF[86]	2022	85.90	87.40	–	–

表 4 部分异常行为检测算法搭载硬件设备对比

方法	年份	硬件设备
Ganokratanaa[95]	2019	NVIDIA GeForce GTX 1080Ti
田联房[6]	2019	NVIDIA GTX 1080Ti GPU
杜启亮[7]	2020	NVIDIA GTX 1080 GPU
Liu[8]	2020	NVIDIA RTX2080
Rodrigues[91]	2020	NVIDIA RTX 2070
Zhou[10]	2021	8 NVIDIA GTX 1080Ti GPUs
GRU-FFN[89]	2021	NVIDIA GTX 3080
HSTGCNN[90]	2021	NVIDIA GTX 2070
Qiu[82]	2022	NVIDIA GTX 1660Ti
STGformer[93]	2022	NVIDIA GTX 1080Ti
STG-NF[86]	2022	NVIDIA TITAN XP GPU
Fan[97]	2022	NVIDIA GeForce GTX TITAN
STGCAE-LSTM[100]	2022	NVIDIA RTX TITAN GPU
MSTA-GCN[124]	2023	NVIDIA RTX 2080Ti GPU

5 总结与展望

目前,基于人体外观及光流等特征的人体异常行为识别与检测方法虽然取得了较好的表现,但由于增加了模型的参数量和训练负担,从而限制了实时检测的能力. 此外,随着研究趋于背景更为复杂、异常种类更多的视频数据集,迫切需要新的行为表示以及特征建模方式. 相较于其他模态数据,骨架能够高效地表征人体运动信息,且更加轻量化,对噪声更具鲁棒性,适用于处理实时任务. 为推动聚焦人体骨架的相关识别与检测方法的研究,本文概述了深度学习背景下的相关工作. 首先,介绍了提取人体骨架的姿态估计算法,该算法的精度直接影响后续识别及检测任务的性能. 在此基础上,对异常行为识别与检测方法进行了整理和分析,并对比了部分方法在相关数据集上的表现.

然而,由于现有公开数据集较少、视频数据源的高维性和不稳定性、人体行为的类内差异性和类间相似性、视频中异常行为难以定义且范围难以界定、缺乏已标注的训练数据等因素,此研究仍存在如下的难点和热点.

1) 多模态融合:随着硬件设备及深度学习技术的发展,多模态信息逐渐成为异常行为识别与检测任务中的研究热点. 其中,RGB视频、骨架信息、红外信息、光流信息等各类模态信息层出不穷. 不同模态数据之间具有互补性,如何综合运用多种模态数据来提升模型的泛化能力,从而更准确地识别和检测异常行为,是今后研究的重点.

2) 实时性检测:实时性是人体异常行为检测任务中一个关键考量因素. 与已发生事件的离线检测不同,实时在线检测需要更高的速度和准确度. 轻量级模型可以有效地提高模型的实时性,减少计算负担,因此设计轻量级模型更具现实意义.

3) 模型可扩展性:Transformer模型的兴起,给计算机视觉领域带来新的发展机遇. Transformer模型相比于传统的卷积神经网络模型具有更高的检测效率和更好的可扩展性. Transformer及其衍生模型(如Video SwinTransformer)能够有效地提取视频序列中的时空特征,为识别与检测异常行为提供更具有前瞻性的解决方案.

4) 注意力机制优化:骨架序列间的时空相关性是人体异常行为识别与检测任务中的重要特征. 各种注意力机制能够有效地捕捉关节间的时空相关性,提高检测的准确度和效率. 然而,注意力机制也会增加计算代价,需要在效率和精度之间进行平衡,才能达到更好的检测性能. 因此,如何优化注意力机制以提高检测的效率和精度,仍然是该领域的研究热点.

参考文献(References)

[1] 张晓平, 纪佳慧, 王力, 等. 基于视频的人体异常行为识别与检测方法综述[J]. 控制与决策, 2022, 37(1): 14-27.
(Zhang X P, Ji J H, Wang L, et al. Overview of video based human abnormal behavior recognition and detection methods[J]. Control and Decision, 2022, 37(1): 14-27.)

[2] 刘云, 薛盼盼, 李辉, 等. 基于深度学习的关节点行为识别综述[J]. 电子与信息学报, 2021, 43(6): 1789-1802.
(Liu Y, Xue P P, Li H, et al. A review of action recognition using joints based on deep learning[J]. Journal of Electronics & Information Technology, 2021, 43(6): 1789-1802.)

[3] Ali Gul M, Yousaf M H, Nawaz S, et al. Patient monitoring by abnormal human activity recognition based on CNN architecture[J]. Electronics, 2020, 9(12): 1993.

[4] 孙晓虎, 余阿祥, 申栩林, 等. 混合注意力机制的异常行为识别[J]. 计算机工程与应用, 2023, 59(5): 140-147.
(Sun X H, Yu A X, Shen X L, et al. Abnormal behavior recognition based on hybrid attention mechanism[J]. Computer Engineering and Applications, 2023, 59(5): 140-147.)

[5] Liu C M, Fu R H, Li Y H, et al. A self-attention augmented graph convolutional clustering networks for skeleton-based video anomaly behavior detection[J]. Applied Sciences, 2021, 12(1): 4.

[6] 田联房, 吴啟超, 杜启亮, 等. 基于人体骨架序列的手扶电梯乘客异常行为识别[J]. 华南理工大学学报: 自然科学版, 2019, 47(4): 10-19.
(Tian L F, Wu Q C, Du Q L, et al. Recognition of passengers' abnormal behavior on the escalator based on human skeleton sequence[J]. Journal of South China University of Technology: Natural Science Edition, 2019, 47(4): 10-19.)

[7] 杜启亮, 黄理广, 田联房, 等. 基于视频监控的手扶电梯乘客异常行为识别[J]. 华南理工大学学报: 自然科学版, 2020, 48(8): 10-21.
(Du Q L, Huang L G, Tian L F, et al. Recognition of passengers' abnormal behavior on escalator based on video monitoring[J]. Journal of South China University of Technology: Natural Science Edition, 2020, 48(8): 10-21.)

[8] Liu Y C, Zhang S N, Li Z Y, et al. Abnormal behavior recognition based on key points of human skeleton[J]. IFAC-PapersOnLine, 2020, 53(5): 441-445.

[9] Yu B X B, Liu Y, Chan K C C. Skeleton-based detection of abnormalities in human actions using graph convolutional networks[C]. The 2nd International Conference on Transdisciplinary AI. Irvine, 2020: 131-137.

[10] Zhou K, Hui B, Wang J F, et al. A study on attention-based LSTM for abnormal behavior recognition with variable pooling[J]. Image and Vision Computing, 2021, 108: 104120.

[11] Hikvision. DeepinMind NVR[EB/OL]. [2023-8-12]. https://www.hikvision.com/europe/products/IP-Products/Network-Video-Recorders/DeepinMind-Series.

[12] Aicaresyou. Xuanmu AI[EB/OL]. [2023-8-12]. https://www.aicaresyou.com/caring.html.

[13] Ren B, Liu M Y, Ding R W, et al. A survey on 3D skeleton-based action recognition using learning method[J/OL]. 2020, arXiv: 2002.05907.

[14] Nayak R, Pati U C, Das S K. A comprehensive review on deep learning-based methods for video anomaly detection[J]. Image and Vision Computing, 2021, 106: 104078.

[15] Yadav R K, Kumar R. A survey on video anomaly detection[C]. 2022 IEEE Delhi Section Conference. New Delhi, 2022: 1-5.

[16] 朱红蕾, 朱昶胜, 徐志刚. 人体行为识别数据集研究进展[J]. 自动化学报, 2018, 44(6): 978-1004.
(Zhu H L, Zhu C S, Xu Z G. Research advances on human activity recognition datasets[J]. Acta Automatica Sinica, 2018, 44(6): 978-1004.)

[17] Baradaran M, Bergevin R. A Critical study on the recentdeep learning based semi-supervised video anomaly detection methods[J/OL]. 2021, arXiv: 2111.01604.

[18] Anoopa S, Salim A. Survey on anomaly detection in surveillance videos[J]. Materials Today: Proceedings, 2022, 58: 162-167.

[19] Mu H Y, Sun R Z, Yuan G, et al. Abnormal human behavior detection in videos: A review[J]. Information Technology and Control, 2021, 50(3): 522-545.

[20] 丁重阳, 刘凯, 李光, 等. 基于时空权重姿态运动特征的人体骨架行为识别研究[J]. 计算机学报, 2020, 43(1): 29-40.
(Ding C Y, Liu K, Li G, et al. Spatio-temporal weighted posture motion features for human skeleton action recognition research[J]. Chinese Journal of Computers, 2020, 43(1): 29-40.)

[21] Toshev A, Szegedy C. DeepPose: human pose estimation via deep neural networks[C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, 2014: 1653-1660.

[22] Chu X, Yang W, Ouyang W L, et al. Multi-context attention for human pose estimation[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 5669-5678.

[23] Newell A, Yang K Y, Deng J. Stacked hourglass networks for human pose estimation[C]. European Conference on Computer Vision. Cham: Springer, 2016: 483-499.

[24] Yang W, Li S, Ouyang W L, et al. Learning feature Pyramids for human pose estimation[C]. 2017 IEEE International Conference on Computer Vision. Venice, 2017: 1290-1299.

[25] Chen Y L, Wang Z C, Peng Y X, et al. Cascaded pyramid network for multi-person pose estimation[C].

2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 7103-7112.

[26] Fang H S, Xie S Q, Tai Y W, et al. RMPE: Regional multi-person pose estimation[C]. 2017 IEEE International Conference on Computer Vision. Venice, 2017: 2353-2362.

[27] Li J F, Wang C, Zhu H, et al. CrowdPose: Efficient crowded scenes pose estimation and a new benchmark[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2020: 10855-10864.

[28] Xu Y F, Zhang J, Zhang Q M, et al. ViTPose: Simple vision transformer baselines for human pose estimation[J/OL]. 2022, arXiv: 2204.12484.

[29] Qiu Z W, Qiu K, Fu J L, et al. Learning recurrent structure-guided attention network for multi-person pose estimation[C]. 2019 IEEE International Conference on Multimedia and Expo. Shanghai, 2019: 418-423.

[30] Ke L P, Chang M C, Qi H G, et al. Multi-scale structure-aware network for human pose estimation[J/OL]. 2018, arXiv: 1803.09894.

[31] Li K, Wang S J, Zhang X, et al. Pose recognition with cascade transformers[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 1944-1953.

[32] Li Y J, Zhang S K, Wang Z C, et al. TokenPose: Learning keypoint tokens for human pose estimation[C]. 2021 IEEE/CVF International Conference on Computer Vision. Montreal, 2022: 11293-11302.

[33] Papandreou G, Zhu T, Kanazawa N, et al. Towards accurate multi-person pose estimation in the wild[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 3711-3719.

[34] Cao Z, Simon T, Wei S H, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 1302-1310.

[35] Li J, Su W, Wang Z F. Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11354-11361.

[36] Li W B, Wang Z C, Yin B Y, et al. Rethinking on multi-stage networks for human pose estimation[J/OL]. 2019, arXiv: 1901.00148.

[37] Qiu Z W, Qiu K, Fu J L, et al. DGCN: Dynamic graph convolutional network for efficient multi-person pose estimation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11924-11931.

[38] Zhang F, Zhu X T, Ye M. Fast human pose estimation[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2020: 3512-3521.

[39] Luo Z X, Wang Z C, Huang Y, et al. Rethinking the heatmap regression for bottom-up human pose estimation[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 13259-13268.

[40] Yang S, Quan Z B, Nie M, et al. TransPose: keypoint localization via transformer[C]. 2021 IEEE/CVF International Conference on Computer Vision. Montreal, 2022: 11782-11792.

[41] Nie X C, Feng J S, Zhang J F, et al. Single-stage multi-person pose machines[C]. 2019 IEEE/CVF International Conference on Computer Vision. Seoul, 2020: 6950-6959.

[42] Duan K W, Bai S, Xie L X, et al. CenterNet: Keypoint triplets for object detection[C]. 2019 IEEE/CVF International Conference on Computer Vision. Seoul, 2020: 6568-6577.

[43] Miao H X, Lin J Q, Cao J J, et al. SMPR: Single-stage multi-person pose regression[J]. Pattern Recognition, 2023, 143: 109743.

[44] Shi D H, Wei X, Li L Q, et al. End-to-end multi-person pose estimation with transformers[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 11059-11068.

[45] Qiu Z W, Yang Q S, Wang J, et al. IVT: An end-to-end instance-guided video transformer for 3D pose estimation[C]. Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, 2022: 6174–6182.

[46] Qiu Z W, Yang Q S, Wang J, et al. PSVT: End-to-end multi-person 3D pose and shape estimation with progressive video transformers[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 21254-21263.

[47] Caetano C, Sena J, Brémond F, et al. SkeleMotion: A new representation of skeleton joint sequences based on motion information for 3D action recognition[C]. The 16th IEEE International Conference on Advanced Video and Signal Based Surveillance. Taipei, 2019: 1-8.

[48] Li C, Zhong Q Y, Xie D, et al. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation[J]. Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, 2018: 786-792.

[49] Duan H D, Zhao Y, Chen K, et al. Revisiting skeleton-based action recognition[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 2959-2968.

[50] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[51] Liu J, Wang G, Duan L Y, et al. Skeleton-based human action recognition with global context-aware attention LSTM networks[J]. IEEE Transactions on Image Processing, 2018, 27(4): 1586-1599.

[52] Zhang P F, Lan C L, Xing J L, et al. View adaptive neural networks for high performance skeleton-based human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1963-1978.

[53] 孙琪翔, 何宁, 张聪聪, 等. 基于轻量级图卷积的人体骨架动作识别方法[J]. 计算机工程, 2022, 48(5):

306-313.
(Sun Q X, He N, Zhang C C, et al. Human skeleton action recognition method based on lightweight graph convolution[J]. Computer Engineering, 2022, 48(5): 306-313.)

[54] Shuman D I, Narang S K, Frossard P, et al. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains[J]. IEEE Signal Processing Magazine, 2013, 30(3): 83-98.

[55] Tang Y S, Tian Y, Lu J W, et al. Deep progressive reinforcement learning for skeleton-based action recognition[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 5323-5332.

[56] Yan S J, Xiong Y J, Lin D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[J/OL]. 2018, arXiv: 1801.07455.

[57] 曹毅, 刘晨, 黄子龙, 等. 时空自适应图卷积神经网络的骨架行为识别[J]. 华中科技大学学报: 自然科学版, 2020, 48(11): 5-10.
(Cao Y, Liu C, Huang Z L, et al. Skeleton-based action recognition based on spatio-temporal adaptive graph convolutional neural-network[J]. Journal of Huazhong University of Science and Technology: Nature Science Edition, 2020, 48(11): 5-10.)

[58] 曹毅, 刘晨, 盛永健, 等. 基于三维图卷积与注意力增强的行为识别模型[J]. 电子与信息学报, 2021, 43(7): 2071-2078.
(Cao Y, Liu C, Sheng Y J, et al. Action recognition model based on 3D graph convolution and attention enhanced[J]. Journal of Electronics & Information Technology, 2021, 43(7): 2071-2078.)

[59] 曹毅, 吴伟官, 李平, 等. 基于时空特征增强图卷积网络的骨架行为识别[J]. 电子与信息学报, DOI: 10.11999/JEIT220749.
(Cao Y, Wu W G, Li P, et al. Skeleton-based action recognition based on spatio-temporalFeature enhanced graph convolutional network[J]. Journal of Electronics & Information Technology, DOI: 10.11999/JEIT220749.)

[60] Ye F F, Pu S L, Zhong Q Y, et al. Dynamic GCN: Context-enriched topology learning for skeleton-based action recognition[C]. Proceedings of the 28th ACM International Conference on Multimedia. Seattle, 2020: 55-63.

[61] Ding Y R, Bao K, Zhang J Z. An intelligent system for detecting abnormal behavior in students based on the human skeleton and deep learning[J]. Computational Intelligence and Neuroscience, 2022, 2022: 1-11.

[62] Han K, Wang Y H, Chen H T, et al. A survey on vision transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 87-110.

[63] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems. Long Beach, 2017: 5998-6008.

[64] Parmar N, Vaswani A, Uszkoreit J, et al. Image transformer[J/OL]. 2018, arXiv: 1802.05751.

[65] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J/OL]. 2020, arXiv: 2010.11929.

[66] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[J/OL]. 2020, arXiv: 2005.12872.

[67] Zheng S X, Lu J C, Zhao H S, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 6877-6886.

[68] Arnab A, Dehghani M, Heigold G, et al. ViViT: A video vision transformer[C]. 2021 IEEE/CVF International Conference on Computer Vision. Montreal, 2022: 6816-6826.

[69] Neimark D, Bar O, Zohar M, et al. Video transformer network[C]. 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal, 2021: 3156-3165.

[70] Girdhar R, João Carreira J, Doersch C, et al. Video action transformer network[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2020: 244-253.

[71] Zheng C, Zhu S J, Mendieta M, et al. 3D human pose estimation with spatial and temporal transformers[C]. 2021 IEEE/CVF International Conference on Computer Vision. Montreal, 2022: 11636-11645.

[72] Shi F, Lee C, Qiu L. STAR: Sparse transformer-based action recognition[J/OL]. 2021, arXiv: 2107.07089.

[73] Liu Z Y, Luo S, Li W B, et al. ConvTransformer: A convolutional transformer network for video frame synthesis[J/OL]. 2020, arXiv: 2011.10185.

[74] Feng X, Song D, Chen Y. Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection[C]. Proceedings of the 29th ACM International Conference on Multimedia. New York, 2021: 5546-5554.

[75] Zhang Y Y, Li X Y, Liu C H, et al. VidTr: Video transformer without convolutions[C]. 2021 IEEE/CVF International Conference on Computer Vision. Montreal, 2022: 13557-13567.

[76] Sun Y, Shen Y X, Ma L Y. MSST-RT: Multi-stream spatial-temporal relative transformer for skeleton-based action recognition[J]. Sensors, 2021, 21(16): 5339.

[77] Plizzari C, Cannici M, Matteucci M. Spatial temporal transformer network for skeleton-based action recognition[C]. International Conference on Pattern Recognition. Cham: Springer, 2021: 694-701.

[78] Qiu H L, Hou B, Ren B, et al. Spatio-temporal tuples transformer for skeleton-based action recognition[J/OL]. 2022, arXiv: 2201.02849.

[79] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[J/OL]. 2021, arXiv: 2103.14030.

[80] Liu Z, Ning J, Cao Y, et al. Video swin transformer[C].

2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 3192-3201.

[81] Liu Y N, Zhang H, Xu D, et al. Graph transformer network with temporal kernel attention for skeleton-based action recognition[J]. Knowledge-Based Systems, 2022, 240: 108146.

[82] Qiu J F, Yan X L, Wang W, et al. Skeleton-based abnormal behavior detection using secure partitioned convolutional neural network model[J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(12): 5829-5840.

[83] Sun S Y, Gong X J. Hierarchical semantic contrast for scene-aware video anomaly detection[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 22846-22856.

[84] Feng J C, Hong F T, Zheng W S. MIST: Multiple instance self-training framework for video anomaly detection[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 14004-14013.

[85] Markovitz A, Sharir G, Friedman I, et al. Graph embedded pose clustering for anomaly detection[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 10536-10544.

[86] Hirschorn O, Avidan S. Normalizing flows for human pose anomaly detection[J/OL]. 2022, arXiv: 2211.10946.

[87] Li M S, Chen S H, Zhao Y H, et al. Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 211-220.

[88] Luo W X, Liu W, Gao S H. Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection[J]. Neurocomputing, 2021, 444: 332-337.

[89] Fan B, Li P, Jin S, et al. Anomaly detection based on pose estimation and GRU-FFN[C]. 2021 IEEE Sustainable Power and Energy Conference (iSPEC). Nanjing, 2022: 3821-3825.

[90] Zeng X L, Jiang Y L, Ding W R, et al. A hierarchical spatio-temporal graph convolutional neural network for anomaly detection in videos[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(1): 200-212.

[91] Rodrigues R, Bhargava N, Velmurugan R, et al. Multi-timescale trajectory prediction for abnormal human activity detection[C]. 2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass, 2020: 2615-2623.

[92] Pang W F, He Q H, Li Y X. Predicting skeleton trajectories using a skeleton-transformer for video anomaly detection[J]. Multimedia Systems, 2022, 28(4): 1481-1494.

[93] Huang C, Liu Y B, Zhang Z, et al. Hierarchical graph embedded pose regularity learning via spatio-temporal transformer for abnormal behavior detection[C]. Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, 2022: 307-315.

[94] Yu S B, Zhao Z Y, Fang H S, et al. Regularity learning via explicit distribution modeling for skeletal video anomaly detection[J/OL]. 2021, arXiv:2112.03649.

[95] Ganokratanaa T, Aramvith S, Sebe N. Anomaly event detection using generative adversarial network for surveillance videos[C]. 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Lanzhou, 2020: 1395-1399.

[96] Atghaei A, Ziaeinejad S, Rahmati M. Abnormal event detection in urban surveillance videos using GAN and transfer learning[J/OL]. 2020, arXiv: 2011.09619.

[97] Fan Y X, Wen G J, Xiao F, et al. Detecting anomalies in videos using perception generative adversarial network[J].Circuits, Systems, and Signal Processing, 2022, 41(2): 994-1018.

[98] Fan Z Y, Yi S H, Wu D, et al. Video anomaly detection using CycleGan based on skeleton features[J]. Journal of Visual Communication and Image Representation, 2022, 85: 103508.

[99] Morais R, Le V, Tran T, et al. Learning regularity in skeleton trajectories for anomaly detection in videos[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2020: 11988-11996.

[100] Li N J, Chang F L, Liu C S. Human-related anomalous event detection via spatial-temporal graph convolutional autoencoder with embedded long short-term memory network[J]. Neurocomputing, 2022, 490: 482-494.

[101] Li N J, Chang F L, Liu C S. Human-related anomalous event detection via memory-augmented Wasserstein generative adversarial network with gradient penalty[J]. Pattern Recognition, 2023, 138: 109398.

[102] Yun K, Honorio J, Chattopadhyay D, et al. Two-person interaction detection using body-pose features and multiple instance learning[C]. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Providence, 2012: 28-35.

[103] Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset[J/OL]. 2017, arXiv: 1705.06950.

[104] Shahroudy A, Liu J, Ng T T, et al. NTU RGB D: A large scale dataset for 3D human activity analysis[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 1010-1019.

[105] Liu J, Shahroudy A, Perez M, et al. NTU RGB D 120: A large-scale benchmark for 3D human activity understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(10): 2684-2701.

[106] 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 and Machine Intelligence, 2014, 36(7): 1325-1339.

[107] Schuldt C, Laptev I, Caputo B. Recognizing human actions: A local SVM approach[C]. Proceedings of the 17th International Conference on Pattern Recognition.

Cambridge, 2004: 32-36.

[108] Si C Y, Jing Y, Wang W, et al. Skeleton-based action recognition with spatial reasoning and temporal stack learning[J/OL]. 2018, arXiv: 1805.02335.

[109] 钟秋波, 郑彩明, 朴松昊. 时空域融合的骨架动作识别与交互研究[J]. 智能系统学报, 2020, 15(3): 601-608.
(Zhong Q B, Zheng C M, Piao S H. Research on skeleton-based action recognition with spatiotemporal fusion and human-robot interaction[J]. CAAI Transactions on Intelligent Systems, 2020, 15(3): 601-608.)

[110] 李扬志, 袁家政, 刘宏哲. 基于时空注意力图卷积网络模型的人体骨架动作识别算法[J]. 计算机应用, 2021, 41(7): 1915-1921.
(Li Y Z, Yuan J Z, Liu H Z. Human skeleton-based action recognition algorithm based on spatiotemporal attention graph convolutional network model[J]. Journal of Computer Applications, 2021, 41(7): 1915-1921.)

[111] Zhang X K, Xu C, Tao D C. Context aware graph convolution for skeleton-based action recognition[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 14321-14330.

[112] 李炫烨, 郝兴伟, 贾金公, 等. 结合多注意力机制与时空图卷积网络的人体动作识别方法[J]. 计算机辅助设计与图形学学报, 2021, 33(7): 1055-1063.
(Li X Y, Hao X W, Jia J G, et al. Human action recognition method based on multi-attention mechanism and spatiotemporal graph convolution networks[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(7): 1055-1063.)

[113] Si C Y, Chen W T, Wang W, et al. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2020: 1227-1236.

[114] 田志强, 邓春华, 张俊雯. 基于骨骼时序散度特征的人体行为识别算法[J]. 计算机应用, 2021, 41(5): 1450-1457.
(Tian Z Q, Deng C H, Zhang J W. Human behavior recognition algorithm based on skeletal temporal divergence feature[J]. Journal of Computer Applications, 2021, 41(5): 1450-1457.)

[115] 曾胜强, 李琳. 基于姿态校正与姿态融合的2D/3D骨架动作识别方法[J]. 计算机应用研究, 2022, 39(3): 900-905.
(Zeng S Q, Li L. 2D/3D skeleton action recognition based on posture transformation and posture fusion[J]. Application Research of Computers, 2022, 39(3): 900-905.)

[116] 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.
(Su J Y, Song X N, Wu X J, et al. Skeleton based action recognition algorithm on multi-modal lightweight graph convolutional network[J]. Journal of Frontiers of Computer Science & Technology, 2021, 15(4): 733-742.)

[117] Cheng K, Zhang Y F, He X Y, et al. Skeleton-based action recognition with shift graph convolutional network[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 180-189.

[118] 杜启亮, 向照夷, 田联房, 等. 用于动作识别的双流自适应注意力图卷积网络[J]. 华南理工大学学报: 自然科学版, 2022, 50(12): 20-29.
(Du Q L, Xiang Z Y, Tian L F, et al. Two-stream adaptive attention graph convolutional networks for action recognition[J]. Journal of South China University of Technology: Natural Science Edition, 2022, 50(12): 20-29.)

[119] Chen Y X, Zhang Z Q, Yuan C F, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]. 2021 IEEE/CVF International Conference on Computer Vision. Montreal, 2022: 13339-13348.

[120] Hu L Y, Liu S L, Feng W. Spatial temporal graph attention network for skeleton-based action recognition[J/OL]. 2022, arXiv: 2208.08599.

[121] Li W X, Mahadevan V, Vasconcelos N. Anomaly detection and localization in crowded scenes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(1): 18-32.

[122] Lu C W, Shi J P, Jia J Y. Abnormal event detection at 150 FPS in MATLAB[C]. 2013 IEEE International Conference on Computer Vision. Sydney, 2014: 2720-2727.

[123] Luo W X, Liu W, Gao S H. A revisit of sparse coding based anomaly detection in stacked RNN framework[C]. 2017 IEEE International Conference on Computer Vision. Venice, 2017: 341-349.

[124] Chen X Y, Kan S C, Zhang F H, et al. Multiscale spatial temporal attention graph convolution network for skeleton-based anomaly behavior detection[J]. Journal of Visual Communication and Image Representation, 2023, 90: 103707.

作者简介

朱红蕾 (1977–), 女, 副教授, 硕士, 从事计算机视觉、视频理解等研究, E-mail: zhuhllut@139.com;

卫鹏娟 (1999–), 女, 研究生, 从事计算机视觉的研究, E-mail: 2317091872@qq.com;

徐志刚 (1977–), 男, 教授, 博士, 硕士生导师, 从事计算机视觉、机器学习等研究, E-mail: xzg_cn@163.com.