

基于视频的人体异常行为识别与检测方法综述

张晓平^{1,2}, 纪佳慧^{1,2}, 王力^{1,2†}, 何忠贺^{1,2}, 刘世达^{1,2}

(1. 北方工业大学 电气与控制工程学院, 北京 100144; 2. 北方工业大学
城市道路智能交通控制技术北京市重点实验室, 北京 100144)

摘要: 作为计算机视觉的重要分支, 异常行为识别与检测技术已在智能安防、医疗监护、交通管控等领域获得了广泛应用. 然而对异常行为的界定及判别方法与场景因素紧密相关, 针对不同应用场景特点, 适当选择特征提取及异常行为识别与检测方法, 进而保证预警准确率, 在实际应用中至关重要. 基于此, 本文对基于视频的人体异常行为识别与检测方法进行综述, 首先给出了人体异常行为的定义、特点及分类; 其次, 对特征提取方法进行了总结, 特征提取方法的选取及提取特征的好坏直接影响后续判别结果; 之后, 从异常行为识别和异常行为检测两个角度对异常行为判别方法进行了分析和讨论, 并给出了常用异常行为检测数据集及相关算法表现. 最后, 对本领域未来研究方向提出了展望.

关键词: 人体异常行为; 行为识别; 异常行为检测; 视频监控; 特征提取; 数据集

中图分类号: TP391.4

文献标志码: A

DOI: 10.13195/j.kzyjc.2020.1428

Overview of video based human abnormal behavior recognition and detection methods

ZHANG Xiao-ping^{1,2}, JI Jia-hui^{1,2}, WANG Li^{1,2†}, HE Zhong-he^{1,2}, LIU Shi-da^{1,2}

(1. School of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China; 2. Beijing Key Laboratory of Urban Road Intelligent Traffic Control Technology, North China University of Technology, Beijing 100144, China)

Abstract: As an important branch of computer vision, abnormal behavior recognition and detection technology has been widely used in intelligent security, medical monitoring, traffic control and other fields. However, the definition and discrimination methods of abnormal behavior are closely related to the scene factors, and it is very important to appropriately choose the feature extraction as well as abnormal behavior recognition and detection methods according to the characteristics of different application scenarios, so as to improve the warning accuracy. So, this paper reviews the video based human abnormal behavior recognition and detection methods. Firstly, the definition, characteristics and classification of human abnormal behavior are given. Secondly, the feature extraction methods are summarized. The selection of feature extraction methods and the quality of extracted features directly affect the subsequent discrimination results. Then, the paper analyzes and discusses the abnormal behavior discrimination methods from two aspects: abnormal behavior recognition and abnormal behavior detection. The common abnormal behavior detection data sets and related algorithms' performance are also given. Finally, the future research directions of this field are prospected.

Keywords: human abnormal behavior; behavior recognition; abnormal behavior detection; video surveillance; feature extraction; datasets

0 引言

当前, 人体异常行为识别与检测已在社会生产及生活中得到广泛应用. 在医疗监护领域, 通过异常行为识别与检测技术, 可实现对无看护病患或老人的实时监控^[1], 判断目标是否出现跌倒或其他意

外^[2], 并及时报警呼救, 保证他们第一时间得到治疗和帮助; 在交通监管领域, 有关部门可利用人体异常行为识别与检测技术监控包括驾驶员行为在内的车内、外异常情况^[3-4], 降低事故发生风险; 在公共安全领域, 该技术可用于公共场所人员异常情况检测, 判

收稿日期: 2020-10-16; 修回日期: 2021-01-22.

基金项目: 北京市自然科学基金项目 (4204096); 国家自然科学基金项目 (61903006); 北京市长城学者培养计划项目 (CIT&TCD 20190304); 国家重点研发计划 (2017YFC0821102, 2017YFC0822504); 北京市教委基础科研计划项目; 北方工业大学青年毓优项目; 北方工业大学科研启动基金支持

†通讯作者. E-mail: wangli939@ncut.edu.cn.

断打架斗殴等违反社会治安行为^[5-6]。可见,对人体异常行为进行识别与检测具有十分重要的意义,然而,识别与检测的过程十分复杂,且与周围环境、背景复杂度、光线等因素密切相关,而针对不同应用场景,不同识别和检测方法效果存在差异,因此,需对不同的人体异常行为识别与检测方法根据其特点进行分类,以便在实际应用中方便、快捷地进行算法选取。

从目前取得的成果来看,对人体行为的识别与检测可基于监控视频、穿戴式传感器等,本文主要对视频下的异常行为识别与检测技术进行综述,大致可分为两类:(1)以行为识别为第一任务,需针对异常姿态或动作建立样本库,之后通过人体目标检测、姿态估计、动作识别等方法判别具体行为,并最终判定其是否属于异常行为样本库范畴^[7];(2)以异常检测为第一任务,较少考虑具体的异常动作,往往通过与正常场景进行相似度对比实现对视频中异常情景的判定^[8]。在效果方面,基于行为识别的方法往往对个体异常行为具有较好的识别效果,对于全局信息的关注则较少;而基于异常检测的方法对视频的全局信息具有更强的分析能力,但对局部微小的异常行为难有精确的检测效果。两类方法对异常行为的判别原理各有不同,但又紧密联系。首先,如图1所示,二者均以特征提取环节为前提;其次,在实际应用中,往往既希望检测全局异常,同时又希望获知异常的具体对象及原因,因此两类方法常联合使用,用以提升识别与检测效果。

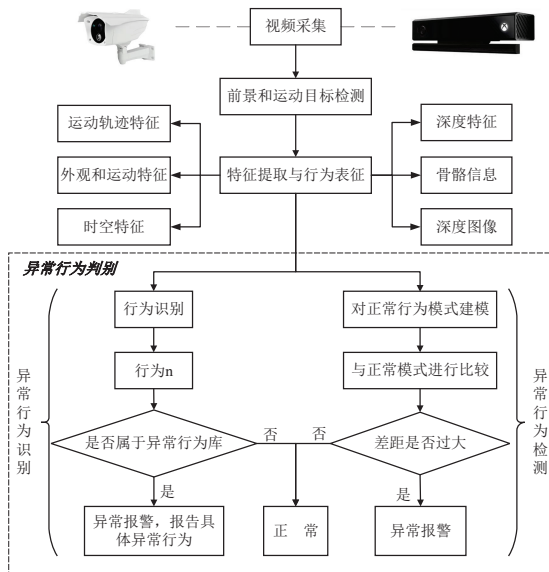


图1 基于视频的人体异常行为判别流程图

异常行为的识别与检测均需进行特征提取,而不同应用场景对特征提取方法需求不同,进而关系到后续异常行为识别和检测方法的选取。基于此,本文在给出人体异常行为定义的基础上,首先对特征

提取方法进行了整理和总结;之后,从异常行为识别和检测两个角度对异常行为判别方法进行了阐述;最后,列举了异常行为检测常用数据集及部分算法相关表现。文末对本领域未来研究方向提出了展望。

1 人体异常行为

1.1 异常行为的定义

对于“行为”这一概念,西方学者曾给出如此定义:行为是人类与周围环境做出最基本的有意义的交互^[9]。文献[10]指出,动物行为包含姿势、动作和环境三个要素。人作为一种高级动物,其行为同样具备如上要素,然而相比于动物行为,人的行为具有更深层的意义及社会性。不同环境下的同一行为会被解读成不同的含义,如在赛场挥拳可视为喝彩,而在街头挥拳则可能被理解成抗议。可见,对行为的判别既要考虑人体动作或姿态对环境造成的影响,又要兼顾环境对于行为的约束作用。

关于“异常”,一般情况下,是指不同于正常状态的现象,如文献[11]将视频场景中的小概率事件定义为异常。对于人体行为而言,违反社会道德及法律或不符合当前场景下的行为均可视为异常,如在操场上奔跑可视为正常,在狭窄的车厢内奔跑则被看作异常。

综上,对人体异常行为作出如下定义,即:当前场景下,目标做出的一切不适宜的动作、姿态或事件等。该定义下,常见的单人异常行为有:摔倒、越界、遗失物品、携带危险物品、损坏公共设施等;常见的多人异常行为有:打斗、持刀行凶、抢劫、推搡、偷窃、踩踏等^[12]。此外,表1还给出了异常行为判别技术常用场景及各场景下的特定异常行为^[13-15]。

表1 特定场景下的异常行为

场景	异常行为	场景	异常行为
公交站点	区域入侵、人车碰撞	超市	恶意损坏物品
交通工具	乞讨、滞留、携带宠物	医院走廊	聚集、跳跃
水库	轻生、溺水、游泳	工厂车间	迟到、早退、违规操作
加油站	吸烟、打电话、滞留	高速公路	行走、奔跑、骑行
银行 ATM	奔跑、徘徊、聚集	查票口	强行闯入、尾随通行
公园	踩踏草坪、开垦	考场	扭头、转身、东张西望
手扶电梯	逆行、攀爬、探头	电梯轿厢	剧烈运动、扒门

1.2 异常行为的特点与分类

对如上异常行为分析可知,异常行为通常具备如下特征:环境相关性、不可预知性、突发性、短时性、低频性、无周期性、局部时空性和全局一致性等^[16]。异常行为的上述特征一般要求异常行为识别及检测算法要具有一定的鲁棒性、实时性和适应性。

此外, 从不同角度出发, 可将异常行为划分为不同类别. 从行为的发生对象出发, 可分为个体行为异常(如倒地、持枪、持械)、群体行为异常(如踩踏、聚集、恐慌)、人物交互异常(如恶意损坏公共物品及公共设施)和人人交互异常(如打斗、推搡)等^[17]. 从发生事件的性质出发, 可分为暴力异常行为和非暴力异常行为. 从发生异常的目标空间位置出发, 又可分为区域入侵行为、区域徘徊行为等. 可见, 异常行为的类型多种多样, 特点各不相同, 因此需根据具体应用场景选择合适的特征提取及识别与检测方法.

2 特征提取方法

特征提取是指从视频数据中提取关键信息用以表征行为的过程, 提取特征的好坏直接影响异常行为识别与检测算法的速度和准确率. 对近些年特征提取方法研究成果进行总结, 具体如下.

2.1 基于人体外观和运动信息的特征提取方法

此类方法将单个人体看作目标, 提取运动人体的质心、轮廓、运动方向等信息作为特征描述目标的行为. Bobick 等^[18] 基于人体的轮廓信息, 结合时间信息构造运动能量图和运动历史图, 从中提取 Hu 不变矩特征表示人体运动, 通过模板匹配的方法实现了行为识别. Wu 等^[19] 通过提取人体最小外接矩形的宽高比实现了摔倒行为检测. 根据不同行为运动方向的差异, 胡芝兰等^[20] 通过提取块运动方向来获取视频段的行为特征, 从而实现行为识别. 上述方法计算量小, 对图像的噪声和人数变化鲁棒性好.

人体的运动信息常通过光流场^[21]、运动历史图^[18] 等方法进行表征. 此类方法在普通场景下识别准确率较高, 但在背景复杂、人群密集的场景中难以获得理想的效果. 对此, 文献 [22] 基于外观和运动信息构建混合动态纹理模型, 提出了时空异常联合检测方法, 实现了拥挤场景下的异常行为检测. 针对人群密集场景, 文献 [23] 将加速度信息与光流特征融合, 构造混合光流直方图作为特征描述子, 应用稀疏表示的方法实现了异常行为检测.

2.2 基于运动轨迹的特征提取方法

基于运动轨迹的特征提取方法通过获取物体在运动过程中的位置、长度、速度等信息构造特征. 在该方面, Wang 等^[24] 提出密集轨迹算法, 该方法密集采样特征点、提取特征点轨迹特征并编码, 通过支撑向量机进行分类. 之后, Wang 等继续改进特征正则化方式和特征编码方式, 提出改进的密集轨迹算法, 实现了人体行为识别^[25].

基于运动轨迹的异常检测, 其主要思想是在训

练阶段获得正常轨迹的模式, 之后在测试阶段将目标轨迹与正常轨迹比较, 当目标轨迹发生重大偏离时, 判断其行为异常. Junejo 等^[26] 利用轨迹的大小、位置、速度、加速度和时空曲率特征训练动态贝叶斯网络, 实现了异常行为检测. 文献 [27] 跟踪目标的运动轨迹并基于频率分析, 应用上下文感知方法实现异常行为检测. 针对视频段内的小范围异常, Yang 等^[28] 提出基于轨迹分割和多示例学习的局部异常检测方法. 文献 [29] 利用稀疏重构法重构行为轨迹, 并将最小残差法用于异常行为检测中. 文献 [30] 在目标跟踪和轨迹分析方法的基础上, 建立稀疏异常检测模型, 实现了多目标联合异常检测. 文献 [31] 利用轨迹运动模式特征生成隐马尔可夫模型, 从而实现视频中的异常行为检测. 利用轨迹分析人体行为的方法, 其识别效果很大程度上依赖于对目标跟踪的准确性, 当目标数量较小时, 此类方法一般可满足识别需求, 但对拥挤场所或复杂场景存在一定局限.

2.3 基于时空兴趣点的特征提取方法

基于时空兴趣点的特征提取方法通过探测器寻找视频内时空维度中波动最剧烈的部分, 即时空兴趣点, 并使用特征描述符描述兴趣点周围的时空信息^[32]. 此类方法从输入视频中检测到时空兴趣点后, 一般利用局部特征描述子来描述人体行为, 其中具有代表性的特征描述子有方向梯度直方图^[33]、局部二值模式^[34]、光流方向直方图^[35]、尺度不变特征变换^[36] 等. 然而, 上述传统的特征描述符不足以描述局部外观和运动信息, 对此, Chen 等^[37] 基于 SIFT(Scale-Invariant Feature Transform) 特征构造 MoSIFT(Motion Scale Invariant Feature Transform), 能够较好地描述运动强度并具有较强的区分性. 利用 MoSIFT 描述符, 文献 [38] 提取视频的底层特征, 实现了对视频中暴力行为的检测. 为利用兴趣点的全局时空分布特性, Bregonzio 等^[39] 从多个时间维度上累积兴趣点形成兴趣点云, 并从中提取整体特征, 从而实现行为识别. 基于兴趣点获取的局部时空特征可以在无需前景背景分割及运动目标精确跟踪的情况下, 通过特征编码的方式描述人体的行为和运动, 适用于背景复杂的场景.

2.4 基于二维人体骨骼信息的特征提取方法

基于二维人体骨骼信息的特征提取方法是通过姿态估计, 获取人体关键部位的位置和状态信息, 从而构建特征向量来描述人体行为. 运用此类方法时, 一般需进行人体检测和骨骼关节点检测, 并利用目标跟踪算法对关键点进行跟踪, 用以辅助人体行为

识别. Fujiyoshi 等^[40] 利用人体头部与四肢 5 个关键节点来表征人体姿态, 利用这些点与重心形成的矢量构造特征向量描述人体行为. 随着姿态估计算法研究的不断深入, 应用于行为识别的关键点数量不断增长, 现能够获取的关键点信息超过 20 个. 同时, 姿态估计方法的精度和速度也不断提升, 可识别的人数不断增加. 在对多人姿态估计的研究中, 主要有自顶向下和自底向上两种方式. 自顶向下的方式首先通过检测算法获得人形轮廓, 然后使用估计检测器检测出轮廓内的关键点, 进而连接所有关键点获取人体姿态^[41], 这种方式较为直观, 便于理解, 骨骼信息提取精度较好. 自底向上的方式则是先检测出一幅图像内所有人体部位, 然后通过聚类等方法将所有关键点进行连接并分组, 拼接成每个人的骨架图^[42], 这种方式的最大的特点是只需对图片进行一次检测, 并且检测速度不受图像内人数影响.

基于人体骨骼信息的姿态估计不易受光线和背景变化的影响, 具有较好的鲁棒性和适应性, 被广泛应用于异常行为判别技术^[43-44]. 相比图像特征, 骨骼特征更为紧凑、结构更强、对人体运动的描述更加具体. 基于骨骼信息的特征提取方法为异常行为识别与检测技术开辟了新的思路.

2.5 基于三维人体骨骼信息的特征提取方法

进入二十一世纪后, 三维数据采集技术迅猛发展^[45], 深度传感器^[46] 硬件的开发和进步使得研究者们可以更容易地获取图像中的深度信息. 基于三维人体骨骼信息的特征提取方法是从深度图像序列中获取人体关键点的三维信息并建立人体骨骼模型, 进而利用关键点构成的人体轮廓和视频帧间关键点的变化表征人体行为. Alzahrani 等^[47] 应用 Kinect2.0 获取 25 个人体骨骼关键点的三维坐标、与骨骼点同一坐标系下的房间地板平面方程及每一帧的时间戳信息, 从中提取所有可能的骨骼特征, 运用监督学习的方法实现了摔倒行为识别.

骨骼特征提取的方法主要有基于人工设计的特征手动提取方法和基于深度学习的特征自动提取方法. 在手动提取骨骼特征方面, 文献 [48] 利用关节的高度、速度、位置等特征识别了人体的摔倒行为. 文献 [49] 提取单个关节的运动特征和多个关节的关系特征作为人体运动识别的综合特征, 从运动学和空间几何学的角度发掘了人体运动时的关节特征, 获得较好的识别效果. 这一类利用关节之间角度和运动特征识别人体行为的方法能够较好地反应人体运动的实际规律, 易于理解和表达, 但在识别过程中建

立的模型较为复杂, 计算量较大. 在自动提取骨骼特征方面, Pham 等^[50] 构造了由骨骼姿势及其运动组成的紧凑图像表示的骨骼位置运动特征, 利用自适应直方图均衡化算法对特征进行增强, 并应用基于 DenseNet 结构的深度卷积神经网络, 实现了骨架序列与其动作标签之间端到端的映射. 类似地, 文献 [51] 将三维骨骼序列中的时空信息编码成三幅二维图像, 并将其动态特性编码成图像中的颜色分布, 即关节轨迹图, 三幅关节轨迹图相互提供补充信息, 实现了高效的行为表征.

人体骨骼序列具有较好的行为时空特征, 对时空特征的区分性充分利用, 有利于快速、准确地实现行为判别. Song 等^[52] 从骨骼序列中提取时空特征识别人类行为, 选择性地关注了输入帧的关键骨骼关节, 对不同的关节赋予不同程度的注意力. Li 等^[53] 将原始的骨骼坐标和骨骼运动直接输入网络中进行行为标签预测, 同时构建了能够对重要骨架关键点自动重排和选择的骨架变换模块. Yan 等^[54] 提出 ST-GCN (Spatial Temporal Graph Convolutional Networks), 该模型根据骨骼序列构建时空图, 并使用图卷积网络提取其特征, 具有较强的表达能力和泛化能力. 基于三维人体骨骼信息的行为识别方法在特殊场景下数据规模较小. 针对该问题, 文献 [55] 提出了一种名为样本融合网络的数据增强网络, 该网络利用长短期记忆自动编码器生成新样本, 同时将样本融合网络与人体动作识别网络级联, 有效地提高了分类的准确性.

基于三维人体骨骼特征进行人体异常行为判别时, 不易受到人体外形差异的影响, 特别是在光线变化、出现阴影等情况下, 基于三维骨骼信息的特征可提高智能视频监控系统的识别能力和检测精度, 在背景复杂、噪声较多时同样具有较好的鲁棒性.

2.6 基于深度学习的特征提取方法

基于深度学习的特征提取方法是利用深度神经网络直接从图像中学习深度特征^[56], 在使用时需根据特征提取的规则设计网络结构并通过训练和学习获得网络参数. 相比于人工设计特征如时空特征、外观及运动特征等, 深度神经网络提取的特征可解释性较差, 但它对于数据库的依赖较小, 提取特征较为客观, 对于不同视频数据中的光线变化、遮挡、视角转换等问题具有更好的普适性. 一些情况下, 可将原始视频和图像直接传入深度神经网络并输出结果, 实现端到端的异常行为判别. 常用的深度神经网络主要涉及卷积神经网络、递归神经网络等.

2.6.1 基于三维卷积神经网络的特征提取

卷积神经网络 (Convolutional Neural Networks, CNN) 是基于卷积计算的多层神经网络, 各层之间稀疏连接且同一通道内像素权重共享. 这种结构大大削减了模型参数数量, 提高了网络的训练速度和泛化能力, 降低了过拟合的风险. 同时, 卷积神经网络能够提取平移不变特征, 因此被广泛应用于图像识别任务中. 为提取视频帧间的时空特征, Ji 等^[57] 在输入数据中增加时间维度信息, 使用三维卷积核对连续多帧图片进行卷积, 实现了人体的行为识别. 利用三维卷积神经网络能够较好提取时空特征的优势, Tran 等^[58] 在文献 [57] 的基础上提出 C3D (Convolutional 3D) 网络, 通过反复实验, 确定了合适的三维卷积核尺寸, 同时提取输入视频的外观和运动特征, 并将之输入到多分类线性支持向量机中, 实现简单高效的行为识别. 为充分获取视频及图像的显著特征, 文献 [59] 在行为识别框架中引入注意力机制, 通过在特征映射中增加与前景区域相关联的值来构造剩余注意单元, 以减少背景运动对识别过程产生的不利影响. 由于三维卷积网络计算量较大、模型参数训练困难, 文献 [60] 使用空间二维卷积和时间一维卷积代替三维卷积网络, 并以多种组合方式植入残差网络, 设计了伪三维残差神经网络结构 (Pseudo-3D ResNet, P3D), 有效降低了模型的复杂度, 在行为检测和场景识别等方面取得了较好的效果.

2.6.2 基于双流卷积神经网络的特征提取

双流卷积神经网络是将输入视频分为时间流和空间流两个部分, 提取多帧稠密光流信息作为时间流的输入, 将单帧 RGB 图像作为空间流的输入, 利用深度卷积神经网络分别对两种信息流进行处理, 最后将结果进行融合实现行为识别^[61]. 然而, 原始的双流结构无法实现两个卷积流之间的信息交互, 为充分利用时间信息和空间信息, Feichtenhofer 等^[62] 基于双流网络提出了一种时空融合架构, 在不同层级对两个网络的特征图进行融合, 之后利用三维卷积神经网络对融合后的特征进行处理, 从而更好地实现时间网络与空间网络的交互. 与图像数据集相比, 现有的视频数据集规模较小, 对此, Wang 等^[63] 对原有双流架构加以改进, 利用 ImageNet 对时空流进行预训练, 使用更小的学习率, 结合数据增强技术, 防止了因数据量太小而产生的过度匹配情况. 之后, 基于长时间的视频序列建立模型, 构造时域分段网络 (Time Segment Network, TSN), 利用稀疏时间采样方法从较长的视频序列中随机抽取短片段输

入不同的双流网络中, 并采用段共识函数对不同的片段得分进行融合, 实现视频级的行为识别^[64]. 由于整段视频内还可能存在标签以外的其他动作, Lan 等^[65] 用带有视频级标签的局部视频片段训练 TSN, 提取视频的局部特征, 将局部特征聚集成全局特征并映射到视频级的动作标签. Zhou 等^[66] 基于 TSN 提出 TRN (Temporal Relation Network), 能够对多个时间尺度的视频帧进行时序推理, 获取多帧之间的时间依赖关系. 文献 [67] 在 TSN 稀疏采样策略基础上, 提前对时空信息进行融合, 构建了一套完整的在线视频分析框架, 实现了更加快速的行为检测. Carreira 等^[68] 基于双流框架和三维卷积网络提出 I3D (Inflated 3D Convolutional Network), 该网络将输入的 RGB 图像和光流信息分别训练, 取两者预测结果的平均值输出. 为充分利用时间、空间及跨通道维度的特征, 文献 [69] 在 I3D 基础上添加了通道-时空注意力块, 提出了细粒度动作识别的多视角注意力机制, 提高了行为识别准确率.

2.6.3 基于递归神经网络的特征提取

与前馈网络相比, 递归神经网络 (Recurrent neural networks, RNN) 能够存储信息并处理时序数据, 具有对输入信息的记忆能力, 能够反映时间序列数据的关系. 然而, 在解决长序列问题时, 递归网络容易出现梯度消失问题, 为解决这一问题, Hochreiter 等^[70] 将 RNN 拓展至长短期记忆单元 (Long Short-Term Memory, LSTM), 用存储单元代替神经元, 并添加输入门、输出门、遗忘门, 其中输入门决定需要保留的当前输入信息, 输出门决定需要输出到下一时刻隐藏层的信息, 遗忘门决定需要抛弃的上一时刻的信息, 模型实现了系统状态的整体更新和结果输出, 对于学习长序列数据的特征有不错效果. 文献 [71] 将卷积神经网络与 LSTM 单元相结合, 利用较长视频进行训练, 得到了较好的识别效果. 由于长期递归神经网络模型可以直接将可变长度的输入映射到可变长度的输出, Donahue 等^[72] 提出了 LRCN (Long-term Recurrent Convolutional Networks), 利用 CNN 提取特征, 将之作为 LSTM 的输入, 实现了端到端的行为识别. 为充分利用视频的空间相关性, 文献 [73] 提出一种基于运动的注意力机制, 将注意力引导到相关的时空位置. 由于卷积长短期记忆网络能够较好地获取局部时空特征, 文献 [74] 将相邻帧之间的差异作为输入, 利用卷积长短期记忆网络进行编码, 基于视频实现了端到端的暴力行为检测.

相比于其他的特征提取方法, 基于深度学习的

特征提取方法不必定义提取的具体特征, 可以通过较少的预处理、借助自身的多层隐藏节点直接从原始数据中学习有用特征^[75]. 同时, 基于深度学习的特征提取方法易与大数据进行结合, 在多种场景下取得较好的应用效果.

2.7 基于深度图像的特征提取方法

随着三维视觉传感器的飞速发展, 深度图像逐渐广泛应用于行为识别领域. 深度图像中包含三维空间的深度信息, 相比于传统色彩图像, 深度图像去除颜色和纹理信息, 使用空间几何信息和目标结构信息, 背景变化和遮挡等因素的影响被极大削弱, 具有较好的鲁棒性^[76]. 文献 [77] 将深度数据中的运动信息和形状信息结合, 生成基于运动的兴趣点和基于形状的兴趣点, 从而高效地描述行为的局部外观和时空分布. 文献 [78] 采用空间拉普拉斯和时间能量金字塔方法, 将深度图像序列分解为不同时空位置的特定频带, 提取其中低频和低频特征并将之融合, 实现高效的行为分类. 在一定程度上, 利用深度图像进行人体行为分析弥补了二维图像部分信息缺失的问题, 但深度图像本身存在较多噪声, 且缺少外观和纹理信息. 此外, 基于深度图像的特征提取方法计算量较大, 目前还有很大的研究和发展空间.

3 异常行为判别

提取合适的特征表征行为信息后, 需根据提取特征的特点及应用场景选择合适的异常行为判别方法. 如引言所述, 对异常行为的判别可大致分为以异常行为识别为主和以异常行为检测为主两类.

3.1 异常行为识别方法

在某些特定目标和特定场景下, 正常状态与异常状态的差别较小, 因此在实际应用中, 需要对异常行为和姿态进行划分, 并建立异常行为样本库, 利用动作识别方法识别具体行为, 即提前对感兴趣的事件或行为进行定义, 结合标签对网络进行训练, 通过识别目标的具体动作和姿态, 判断其行为是否异常. 此种方法与场景具有相关性, 有助于了解具体发生了何种异常行为, 以及是否出现虚假报警^[79]. 同时, 当目标人数较少或对特定人员进行行为检测时, 将此类动作识别方法与目标跟踪算法结合, 在判断目标是否出现异常行为方面能够取得较好的效果. 如图 2 所示, 给出了异常行为识别的基本框架, 一般来说, 在检测人体目标后, 需进行特征提取和行为分类, 最终判别结果. 在智能家居、医疗监护等应用领域, 实现异常行为的具体识别对后续所需采取的应对策略具有十分重要的指导意义.

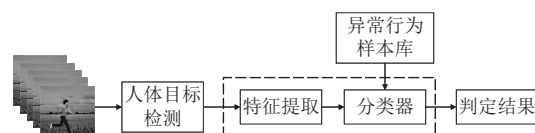


图2 异常行为识别方法框图

异常行为识别有时基于已有视频, 对既发异常进行判别, 属于离线范畴. 更多场景中, 需要在线对异常行为进行识别并报警, 此时输入数据是以未进行时域分割的序列形式输入到识别网络. 此类数据往往存在较多干扰, 且异常行为是否发生和发生的时间不可预知, 常通过逐帧检测法、滑动窗口法等实现在线行为分析^[80-81]. 此外, 递归神经网络等可捕捉行为之间的长期依赖关系, 同样被用于在线行为判别^[82-83].

3.2 异常行为检测方法

异常行为识别在实际应用中存在以下问题: (1) 实际应用中单个目标异常行为样本较少; (2) 当视频中人数众多、遮挡严重时, 单个人体目标行为识别的精度下降, 同时人数的增加使得计算量增大; (3) 在拥挤场景中, 对每种行为进行标记成本较高, 同时无法保证标记能够覆盖全部异常行为. 此时, 异常行为检测方法就显得格外重要. 异常行为检测方法仅将行为分为正常和异常两个类别, 从大量的视频数据中学习经验, 实现像素级、帧级或视频级的异常行为判别. 针对不同场景, 采集的视频数据及标签的制作各不相同, 基于不同的数据形式, 异常行为检测方法可分为有监督、半监督和无监督三种.

3.2.1 有监督异常行为检测方法

有监督异常行为检测方法在模型训练之前需要对所有数据进行标注, 通过样本数据与标签的映射关系建立起网络模型, 之后进行特征提取并设置分类器, 最终实现数据的分类, 常用的分类器有支持向量机、贝叶斯网络等. 此外, 卷积神经网络作为一种有监督方法在异常行为检测领域同样有广泛应用. 针对异常行为的环境相关性, 文献 [79] 提出一种监控视频中异常事件的联合检测与重述方法, 利用动作、对象及其属性三个视觉任务训练卷积神经网络, 获取事件的语义信息, 之后将它们与预先为这些语义设定的评分规则结合, 作为新的特征输入到异常检测器中继续训练, 进而获得更高层次的异常事件语义信息.

在分析具体的异常行为时, 有监督的分类方法具有较好的效果. 然而, 监督学习对于像素级和帧级视频标注的成本较高, 部分研究人员将目光投入以视频级异常标签为训练数据的异常行为检测技术

中, 利用数据的弱标签实现数据包与标签的映射. 由于在视频中准确标定异常行为的时间位置较为繁琐, 一些研究利用多示例学习方式来解决其中的弱监督异常检测问题^[84], 所谓弱监督学习是指在训练数据只具有粗粒度标签的情况下建立模型, 此过程中视频被视为包, 视频的每一帧图像便是示例, 通过标记包的正负注释训练数据的异常标签. 文献 [85] 基于多示例学习提出了一种新的异常行为检测方法, 只标定视频当中是否存在异常, 并利用深度多示例排序框架学习异常, 利用 C3D 提取视频特征, 从特征向量中对异常分数进行回归, 实现了弱监督学习下的异常行为检测与定位. 文献 [86] 以视频包为输入, 通过基于隐向量的注意力机制对视频包的时空特征进行加权处理, 强化特征中的重要部分, 最终获得视频包的异常得分, 实现了端到端的异常行为检测. 由于多示例学习方法中被标注的异常视频内可能包含许多正常片段, 文献 [87] 对训练数据中的标签噪声进行清除, 将弱监督异常检测问题转化为噪声标签下的监督学习任务, 结合 C3D 和 TSN 两种动作分类器实现了异常行为检测.

3.2.2 半监督异常行为检测方法

半监督学习是通过学习一部分已知标签的样本和一部分未知标签的样本, 从而将输入数据映射到标签. 相比于异常数据, 获取正常数据相对容易, 因此, 半监督方法在异常行为检测任务中使用较多. 在半监督异常行为检测方法中, 仅需对正常样本进行标记, 在训练阶段从视频中学习正常模式, 在测试阶段计算当前模式与正常模式的相似度, 将偏离正常模式的检测视为异常模式. 从模型的原理出发, 半监督异常行为检测方法可分为基于重构模型的异常行为检测和基于预测模型的异常行为检测.

在基于重构模型的方法中, 首先在训练阶段, 利用大量的正常片段得到自动编码器的权重; 训练结束后将原始时间序列输入自动编码器网络, 得到重构后的时间序列; 最后, 设计样本的估计异常分数, 用来表示重构前后视频序列之间的整体和局部误差, 当测试样本得分超过阈值则判定此时出现异常. 作为一种常见的统计推理方法, 主成分分析法^[88](principal component analysis, PCA) 能够将高维数据映射到低维特征空间, 降维后的特征可以由低维空间重新映射到原空间, 实现对原始数据的重构, 从而将离群点从数据中分离出来. 根据这一思路, 以 PCA 为基础的一系列算法被应用于异常检测任务^[89]. 为获取高维数据中的非线性结构, 文献 [90] 扩

展了鲁棒主成分分析法, 提出运用卷积自编码器模型实现异常检测. 类似地, 文献 [91] 提出一种基于限制玻尔兹曼机的视频异常检测框架, 利用重构数据的误差识别偏离正常行为的情况.

基于预测模型的异常检测方法也是目前应用较为广泛的方法之一. 不同于重构模型, 这种方法考虑了视频序列的时空相关性, 其目的在于学习一个能够较好地重构视频帧的生成模型, 从而利用过去的多帧图像信息建模当前帧及未来帧, 通过比较预测帧与实际帧之间的差异实现异常检测^[92]. 由于长短期记忆网络能够学习时间特征并预测时序数据, 因此常被用于检测视频序列内的行为异常. Medel 等^[93] 基于复合卷积长短期记忆网络框架实现视频的异常检测, 通过构造编码器和解码器来重构视频序列并预测其未来帧, 从而计算视频序列的异常得分实现异常行为判别. Chong 等^[75] 提出一种时空自编码器来学习训练视频中的正常模式, 其空间自编码器利用卷积神经网络获取视频的空间特征, 时间自编码器利用卷积长短期记忆网络学习空间特征随时间的变化. 相似地, 文献 [94] 将 CNN 和卷积长短期记忆网络与自动编码器结合, 学习视频的外观和运动信息, 相比于基于三维卷积自动编码器的方法, 结合了 LSTM 的深度框架能够更好地识别出外观和运动变化, 从而更有效地实现异常检测.

总体而言, 半监督异常行为检测方法的优点在于训练时只需要提供正常样本, 而对异常的检测则是通过测量样本对正常模式的偏离程度, 即样本的异常 (或正常) 得分, 来判定视频中是否出现异常, 并从获得异常得分的时间序列中分离出低于阈值分数的片段, 从而实现异常行为在时间维度上的定位^[95].

3.2.3 无监督异常行为检测方法

无监督学习是在无需任何数据标签的情况下, 仅依靠样本之间的内部关联进行建模并将全部数据映射到多个标签. 由于大多数样本中异常行为出现的概率较低, 因此, 在无监督异常行为检测过程中, 通常将与数据主体相似度较低的行为视为异常行为. 生成对抗网络 (Generative adversarial networks, GAN) 由一个生成器和一个判别器构成, 通过对抗学习的方式估测数据样本在特征空间中的潜在分布并生成新的数据样本, 在无监督异常行为检测任务中获得了成功. 文献 [96] 提出一种利用生成对抗网络进行异常检测和定位的方法, 利用正常帧及其对应的光流图训练网络, 通过比较真实数据与重构数据的外观和运动信息的局部差异实现异常判别和异常区

域的确定. 此外, 随着人工神经网络的发展, 基于深度学习技术的无监督异常检测方法层出不穷. 文献 [97] 提出一种基于深度神经网络的双融合框架, 利用堆叠去噪自编码器学习视频的外观和运动特征, 并基于学习到的深度特征采用多个一类支持向量机预测视频序列的异常得分. 文献 [98] 设计了三维卷积神经网络与深度自编码器结合的级联深度神经网络, 用于拥挤场所的异常检测与定位. 这类方法无需数据标注等繁琐步骤, 运算方便快捷, 但需要大量数据作为支撑, 才能够获得较好的准确率.

4 异常行为检测数据集

4.1 数据集

在异常行为判别过程中, 针对异常行为检测, 已有非常多的公开数据集可供研究者使用. 为方便异常行为相关研究学者更好地开展工作, 此处对常用的异常行为检测公共数据集进行介绍, 包括视频内容、视频场景及特点等.

(1) USCD^[99]: 该数据集利用固定高度摄像机采集行人视频, 训练集中只包含正常行为, 测试集中包含部分异常行为, 其中异常行为包括非人实体闯入和行人行为异常两大类. USCD 共包含 98 段视频, 根据视频画面中人群的移动方向和场景干扰因素的不同, 构造了 USCD Ped1 和 USCD Ped2 两部分, 其中 USCD Ped1 主要包含人群垂直方向的移动, 目标分

辨率较低; USCD Ped2 主要包含人群水平方向的移动, 运动目标一般存在遮挡. 该数据集主要用于人群中个体异常行为研究, 属于局部异常行为数据集.

(2) UMN^[100]: 该数据集包含 11 段视频, 涉及一个室内场景和两个室外场景, 其中人的行走或游荡属于正常行为, 人群奔跑和四处逃窜属于异常行为. 每段视频以正常行为起始, 从某一帧开始人群发生异常, 直至在画面中消失, 该段视频结束. 该数据集主要研究视频中人群异常行为, 属于全局异常行为数据集.

(3) CUHK Avenue^[101]: 该数据集由 16 段训练视频和 21 段测试视频组成, 涉及投掷物体、游荡和奔跑共 47 个异常事件, 每一帧的异常行为具有像素级标注.

(4) Subway entrance and exit^[102]: 该数据集包含两个视频段, 一段为地铁入口视频, 时长 96 分钟, 另一段为地铁出口视频, 时长 43 分钟, 包含 19 种异常行为, 如走错方向、徘徊等. 视频中的每一帧图片具有帧级的异常标签, 属于室内场景中的异常行为.

(5) ShanghaiTech^[103]: 该数据集视频数据在 13 个光照条件复杂及不同相机角度场景下采集, 包含 330 个训练视频和 107 个测试视频, 共 130 个异常事件, 其中异常行为也复杂多样, 是一个规模较大的异常行为检测数据集.

表 2 部分算法的帧级 AUC 比较 (%)

方法	USCD Ped1	USCD Ped2	UMN	ShanghaiTech	CUHK Avenue	Subway Entrance	Subway Exit
Social Force ^[100]	67.5	55.6	96.0	-	-	-	-
MPPCA ^[89]	66.8	69.3	-	-	-	-	-
Permutation-based ^[104]	-	-	91.0	-	78.3	69.1	82.4
Conv AE ^[105]	81.0	81.1	-	-	70.2	94.3	80.7
ConvLSTM AE ^[94]	75.5	88.1	-	-	77.0	93.3	87.7
Spatiotemporal AE ^[75]	89.9	87.4	-	-	80.3	84.7	94
GrowingGas ^[106]	93.8	94.1	99.7	-	-	-	-
GAN ^[96]	97.4	93.5	99.0	-	-	-	-
Multi-task Fast R-CNN ^[79]	-	92.2	-	-	89.8	-	-
Unmasking ^[107]	68.4	82.2	-	-	80.6	71.3	86.3
Stacked RNN ^[103]	-	92.2	-	68	81.7	-	-
U-Net ^[108]	83.1	95.4	-	72.8	85.1	-	-
AMDN ^[97]	92.1	90.8	-	-	-	-	87.9
Plug-and-play CNN ^[109]	95.7	88.4	98.8	-	-	-	-
Narrowed Normality Clusters ^[110]	-	-	99.3	-	88.9	93.5	95.1
MPED-RNN ^[44]	-	-	-	75.4	86.3	-	-
Object-centric AE ^[111]	-	97.8	99.6	84.9	90.4	-	-
Adversarial 3D Conv AE ^[112]	95.7	96.0	-	84.0	91.2	90.5	98.8

4.2 一些方法在典型数据集上的表现

针对不同的应用场景, 异常行为数据的正负样本分布可能存在偏移, 同时, 不同方法对同一场景下同一异常行为的判定结果不尽相同. 算法在典型公开数据集上运行获得的通用异常检测性能指标可作为衡量算法可靠性和实用性的重要依据. 在异常检测领域, ROC(Receiver Operating Characteristic Curve) 曲线因其不受正负样本分布的影响, 常被用于比较算法性能的优劣. 其中, ROC 曲线下的面积被定义为 AUC(Area Under Curve), 是一个 0-1 之间的值, 该值越大表明算法性能越好. 如表 2 所示, 给出了部分算法在典型异常检测数据集上的性能表现, 此处评价指标即为算法的帧级 AUC.

4.3 算法比较与分析

分析以上内容可知, 当处理具有多种异常类型的视频数据时, 仅利用单类分类方法对视频帧建模, 难以取得理想效果^[89, 97]. 利用深度学习获取视频帧的高级特征往往能获得更好的检测效果, 如文献 [75, 94, 105, 111, 112] 基于自编码器的异常检测方法, 在多个数据集上获得了不错的效果. 此外, 基于 GAN 等的无监督检测方法在异常检测方面同样有较好的表现^[96, 110]. 然而, 这类方法模型可解释性较差, 一般难以直观了解异常得分的预测规则. 如文献 [79] 利用对象、属性和动作标签解释异常发生的原因, 提高了算法的可解释性, 在某些场景中取得了成功, 但必须借助标签全面的数据进行训练. 文献 [44] 采用更低维度的语义特征, 通过对骨骼轨迹信息的学习, 判断个体行为是否异常, 但对人人交互、人物交互行为的异常判别还有待提升. 可见异常行为相关研究还有待深入.

5 未来展望

人体异常行为识别与检测技术涉及人体行为识别、人体行为检测、异常行为检测和异常事件检测等多个方面, 被应用于医疗看护、智能家居、人机交互、智能监控等各种场景, 具有较高的应用价值和实际意义. 同时, 相关技术对于医疗诊断、入侵检测、事件分析、视频检索及多媒体语义标注和索引等研究具有一定的参考意义. 综合基于视频的人体异常行为识别与检测技术研究现状及表现出的问题, 对本领域未来研究方向提出以下展望.

(1) 用于人体异常行为识别与检测的多特征融合. 特征的提取对于行为的识别与检测至关重要, 对多特征进行融合能够发挥不同类型数据的优势, 克服部分特征的缺陷, 有助于更加准确地描述行为并

实现异常判别. 但过多的特征会使得数据处理规模增大, 特征维度升高, 计算复杂度增加. 如何在保证特征全面性和有效性的同时提高计算速度, 是进行异常行为判别前的一个重要科学问题.

(2) 在线行为识别. 行为的发生往往包含时间信息, 相比于离线行为识别方法, 如何从数据流中准确判别出行为发生的始末, 对最终识别具体行为具有指导意义, 其中包含动作分割、异常行为预测、行为在线检测等多方面问题, 均可开展更为深入的研究.

(3) 异常检测方面: 大多数异常行为检测方法中, 需预先设定参数, 如异常行为得分阈值, 参数的取值将决定模型对异常情况的敏感度, 然而目前针对参数的设定标准尚不清晰, 需开展更加深入的研究.

(4) 异常行为的环境关联性. 异常行为的界定与环境紧密相关, 在异常检测的同时获取目标行为的状态、属性以及与场景的交互信息等, 对于深入理解具体发生的行为和行为的异常程度并准确预警具有指导意义.

(5) 异常行为预测. 对异常行为的识别与检测多基于既有数据对已发生事件进行检测和判别, 如能在事件发生前实现对异常的预测并报警, 将极大扩展该技术的应用领域和范围, 目前该方面还具有较大的研究空间.

6 结束语

人体异常行为识别与检测技术在生产生活中具有较高的应用价值, 对此, 本文对计算机视觉下的人体异常行为识别与检测研究现状进行了分析与概述. 首先, 给出了人体异常行为的定义和特点, 并列出了当前异常行为技术常用场景及对应的常见异常行为. 在此基础上, 对异常行为判别方法进行了分析和讨论, 首先对特征提取方法进行了分类和总结, 该环节直接影响后续行为判别方法的选取及准确度; 之后, 从行为识别和异常检测两个角度论述了异常行为判别方法; 此外, 文末还给出了现有的常用于异常行为检测的公开数据集、列举了部分代表性算法在公开数据集上的表现、给出了相关分析. 最后, 在对现有技术的分析和讨论基础上对本领域未来研究发展方向提出了展望.

参考文献 (References)

- [1] Lentzas A, Vrakas D. Non-intrusive human activity recognition and abnormal behavior detection on elderly people: a review[J]. Artificial Intelligence Review, 2019: 1-47.
- [2] 于乃功, 柏德国. 基于姿态估计的实时跌倒检测算法[J]. 控制与决策, 2020, 35(11): 2761-2766.

- (Yu N G, Bai D G. Research on real-time fall detection algorithm based on pose estimation[J]. Control and Decision, 2020, 35(11): 2761-2766.)
- [3] Jiang Q Y, Li G M, Y J W, et al. A model based method of pedestrian abnormal behavior detection in traffic scene[C]. 2015 IEEE First International Smart Cities Conference (ISC2). Guadalajara: IEEE, 2015: 1-6.
 - [4] Hu Y, Lu M, Lu X, et al. Feature refinement for image-based driver action recognition via multi-scale attention convolutional neural network[J]. Signal Processing-image Communication, 2019, 81(115697).
 - [5] Hatirnaz E, Sah M, Direkoglu C. A novel framework and concept-based semantic search Interface for abnormal crowd behaviour analysis in surveillance videos[J]. Multimedia Tools and Applications, 2020, 79(25): 17579-17617.
 - [6] Tripathi V, Mittal A, Gangodkar D, et al. Real time security framework for detecting abnormal events at ATM installations[J]. Journal of Real-Time Image Processing, 2019, 16(2): 535-545.
 - [7] Dhiman C, Vishwakarma D K. A review of state-of-the-art techniques for abnormal human activity recognition[J]. Engineering Applications of Artificial Intelligence, 2019, 77(JAN.): 21-45.
 - [8] Wang T, Snoussi H. Detection of abnormal visual events via global optical flow orientation histogram[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(6): 988-998.
 - [9] Herath S, Harandi M, Porikli F. Going deeper into action recognition: A survey[J]. Image and Vision Computing, 2017, 60(APR.): 4-21.
 - [10] 蒋志刚, 李春旺, 彭建军, 胡慧建. 行为的结构、刚性和多样性 [J]. 生物多样性, 2001(03): 265-274.
(Jiang Z G, Li C W, Peng J J, Hu H J. Structure, elasticity and diversity of animal behavior[J]. Biodiversity Science, 2001(03): 265-274.)
 - [11] 胡正平, 张乐, 李淑芳, 孙德纲. 视频监控系统异常目标检测与定位综述 [J]. 燕山大学学报, 2019, 43(01): 1-12.
(Hu Z P, Zhang L, Li S F, Sun D G. Review of abnormal behavior detection and location for intelligent video surveillance systems[J]. Journal of Yanshan University, 2019, 43(01): 1-12.)
 - [12] 付路瑶. 场景约束下的视频数据人体异常行为识别研究 [D]. 南京: 南京师范大学地理科学学院, 2015: 37-38, 97-100.
(Fu L Y, Research on human abnormal behavior recognition based on scene constraint[D]. NanJing: School of Geography, NanJing Normal University, 2015: 37-38, 97-100.)
 - [13] 欧阳惠卿, 舒文华, 李行, 李杨. 基于双目深度图像的自动扶梯乘客危险行为识别与预警系统 [J]. 中国电梯, 2020, 31(14): 36-39+42.
(OuYang H Q, Shu W H, Li X, Li Y. A passenger dangerous behavior recognition and early warning system of escalator based on RGB-D sensor[J]. China Elevator, 2020, 31(14): 36-39+42.)
 - [14] 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).
(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).)
 - [15] Hendryli J, Fanany M I. Classifying abnormal activities in exam using multi-class Markov chain LDA based on MODEC features[C]. 2016 4th International Conference on Information and Communication Technology (ICoICT). Bandung: IEEE, 2016: 1-6.
 - [16] 杜鉴豪, 许力. 基于区域光流特征的异常行为检测 [J]. 浙江大学学报 (工学版), 2011, 45(07): 1161-1166.
(Du J H, Xu L. Abnormal behavior detection based on regional optical flow[J]. Journal of Zhejiang University (Engineering Science), 2011, 45(07): 1161-1166.)
 - [17] Zhang H, Zhang Y, Zhong B, et al. A Comprehensive Survey of Vision-Based Human Action Recognition Methods[J]. Sensors, 2019, 19(5): 1005.
 - [18] Bobick A F, Davis J W. The recognition of human movement using temporal templates[J]. IEEE Transactions on pattern analysis and machine intelligence, 2001, 23(3): 257-267.
 - [19] Wu X, Gong H, Chen P, et al. Intelligent household surveillance robot[C]. 2008 IEEE International Conference on Robotics and Biomimetics. Bangkok: IEEE, 2009: 1734-1739.
 - [20] 胡芝兰, 江帆, 王贵锦, 等. 基于运动方向的异常行为检测 [J]. 自动化学报, 2008(11): 1348-1357.
(Hu Z L, Jiang F, Wang G L, et al. Anomaly detection based on motion direction[J]. Acta Automatici Sinica, 2008(11): 1348-1357.)
 - [21] He H, Li Y, Tan J. Relative motion estimation using visual - inertial optical flow[J]. Autonomous Robots, 2018, 42(3): 615-629.
 - [22] Li W, Mahadevan V, Vasconcelos N. Anomaly detection and localization in crowded scenes[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 36(1): 18-32.
 - [23] Wang Q, Ma Q, Luo C H, et al. Hybrid histogram of oriented optical flow for abnormal behavior detection in crowd scenes[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2016, 30(02): 1655007.
 - [24] Wang H, Kläser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition[J]. International journal of computer vision, 2013, 103(1): 60-79.
 - [25] Wang H, Schmid C. Action recognition with improved trajectories[C]. Proceedings of the IEEE international conference on computer vision. Sydney: IEEE, 2013:

- 3551-3558.
- [26] Junejo I N. Using dynamic Bayesian network for scene modeling and anomaly detection[J]. *Signal, Image and Video Processing*, 2010, 4(1): 1-10.
- [27] Jiang F, Yuan J, Tsafaris S A, et al. Anomalous video event detection using spatiotemporal context[J]. *Computer Vision and Image Understanding*, 2011, 115(3): 323-333.
- [28] Yang W, Gao Y, Cao L. TRASMIL: A local anomaly detection framework based on trajectory segmentation and multi-instance learning[J]. *Computer Vision and Image Understanding*, 2013, 117(10): 1273-1286.
- [29] Li C, Han Z, Ye Q, et al. Visual abnormal behavior detection based on trajectory sparse reconstruction analysis[J]. *Neurocomputing*, 2013, 119: 94-100.
- [30] Mo X, Monga V, Bala R, et al. Adaptive sparse representations for video anomaly detection[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2013, 24(4): 631-645.
- [31] Kang K, Liu W, Xing W. Motion pattern study and analysis from video monitoring trajectory[J]. *IEICE TRANSACTIONS on Information and Systems*, 2014, 97(6): 1574-1582.
- [32] Dollar P, Rabaud V, Cottrell G, et al. Behavior recognition via sparse spatio-temporal features[C]. 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. Beijing: IEEE, 2005: 65-72.
- [33] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). San Diego: IEEE, 2005: 886-893.
- [34] Ojala T, Pietikainen M, Harwood D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions[C]. *Proceedings of 12th International Conference on Pattern Recognition*. Jerusalem: IEEE, 1994: 582-585.
- [35] Dalal N, Triggs B, Schmid C. Human detection using oriented histograms of flow and appearance[C]. *European conference on computer vision*. Graz: Springer, Berlin, Heidelberg, 2006: 428-441.
- [36] Dawn D D, Shaikh S H. A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector[J]. *The Visual Computer*, 2016, 32(3): 289-306.
- [37] Chen M, Hauptmann A. Mosift: Recognizing human actions in surveillance videos[J]. *Carnegie Mellon University*, 2009.
- [38] Xu L, Gong C, Yang J, et al. Violent video detection based on MoSIFT feature and sparse coding[C]. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence: IEEE, 2014: 3538-3542.
- [39] Bregonzio M, Gong S, Xiang T. Recognising action as clouds of space-time interest points[C]. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 1948-1955.
- [40] Fujiyoshi H, Lipton A J, Kanade T. Real-time human motion analysis by image skeletonization[J]. *IEICE TRANSACTIONS on Information and Systems*, 2004, 87(1): 113-120.
- [41] Fang H S, Xie S, Tai Y W, et al. Rmpe: Regional multi-person pose estimation[C]. *Proceedings of the IEEE International Conference on Computer Vision*. Venice: IEEE, 2017: 2334-2343.
- [42] Cao Z, Simon T, Wei S E, et al. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 1302-1310.
- [43] 王恬, 李庆武, 刘艳, 周亚琴. 利用姿势估计实现人体异常行为识别 [J]. *仪器仪表学报*, 2016, 37(10): 2366-2372.
(Wang T, Li Q W, Liu Y, Zhou Y Q. Abnormal human body behavior recognition using pose estimation[J]. *Chinese Journal of Scientific Instrument*, 2016, 37(10): 2366-2372.)
- [44] Morais R, Le V, Tran T, et al. Learning regularity in skeleton trajectories for anomaly detection in videos[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 11996-12004.
- [45] Aggarwal J K, Lu X. Human activity recognition from 3D data: A review[J]. *Pattern Recognition Letters*, 2014, 48(1): 70-80.
- [46] Zhang Z. Microsoft kinect sensor and its effect[J]. *IEEE MultiMedia*, 2012, 19(2): 4-10.
- [47] Alzahrani M S, Jarraya S K, Ben-Abdallah H, et al. Comprehensive evaluation of skeleton features-based fall detection from Microsoft Kinect v2[J]. *Signal, Image and Video Processing*, 2019, 13(7): 1431-1439.
- [48] Nizam Y, Mohd M N H, Tomari R, et al. Development of Human Fall Detection System using Joint Height, Joint Velocity, and Joint Position from Depth Maps[J]. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 2016, 8(6): 125-131.
- [49] Tian X M, Fan J Y. Joints kinetic and relational features for action recognition[J]. *Signal Processing: The Official Publication of the European Association for Signal Processing (EURASIP)*, 2018, 142(JAN.): 412-422.
- [50] Pham H H, Salmane H, Khoudour L, et al. Spatio - Temporal Image Representation of 3D Skeletal Movements for View-Invariant Action Recognition with Deep Convolutional Neural Networks[J]. *Sensors*, 2019, 19(8): 1932.
- [51] Wang P, Li W, Li C, et al. Action recognition based on joint trajectory maps with convolutional neural networks[J]. *Knowledge-Based Systems*, 2018, 158: 43-53.
- [52] Song S, Lan C, Xing J, et al. An end-to-end

- spatio-temporal attention model for human action recognition from skeleton data[C]. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2017: 4263-4270.
- [53] Li C, Zhong Q, Xie D, et al. Skeleton-based action recognition with convolutional neural networks[C]. 2017 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). Hong Kong: IEEE, 2017: 597-600.
- [54] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]. Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018, 32(1).
- [55] Meng F Y, Liu H, Liang Y S, et al. Sample fusion network: an end-to-end data augmentation network for skeleton-based human action recognition[J]. IEEE Transactions on Image Processing, 2019, 28(11): 5281-5295.
- [56] 朱煜, 赵江坤, 王逸宁, 郑兵兵. 基于深度学习的人体行为识别算法综述 [J]. 自动化学报, 2016, 42(06): 848-857.
(Zhu Y, Zhao J K, Wang Y N, Zheng B B. A Review of Human Action Recognition Based on Deep Learning[J]. Acta Automatica Sinica, 2016, 42(06): 848-857.)
- [57] 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.
- [58] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]. Proceedings of the IEEE international conference on computer vision. Santiago: IEEE, 2015: 4489-4497.
- [59] Liao Z, Hu H, Zhang J, et al. Residual attention unit for action recognition[J]. Computer Vision and Image Understanding, 2019, 189: 102821.
- [60] Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3d residual networks[C]. proceedings of the IEEE International Conference on Computer Vision. Santiago: IEEE, 2017: 5533-5541.
- [61] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. Advances in neural information processing systems, 2014, 27: 568-576.
- [62] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas: IEEE, 2016: 1933-1941.
- [63] Wang L, Xiong Y, Wang Z, et al. Towards good practices for very deep two-stream convnets[J]. arXiv, 2015: arXiv: 1507.02159.
- [64] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]. European conference on computer vision. Amsterdam: Springer, Cham, 2016: 20-36.
- [65] Lan Z, Zhu Y, Hauptmann A G, et al. Deep local video feature for action recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017: 1-7.
- [66] Zhou B, Andonian A, Oliva A, et al. Temporal relational reasoning in videos[C]. Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer, 2018: 803-818.
- [67] Zolfaghari M, Singh K, Brox T. Eco: Efficient convolutional network for online video understanding[C]. Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer, Cham, 2018: 695-712.
- [68] Carreira J, Zisserman A, Vaidis Q. Action recognition? A new model and the kinetics dataset[C]. IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4724-4733.
- [69] Zhu Y, Liu G. Fine-grained action recognition using multi-view attentions[J]. The Visual Computer, 2020, 36(9): 1771-1781.
- [70] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
- [71] Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. Boston: IEEE, 2015: 4694-4702.
- [72] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. Boston: IEEE, 2015: 2625-2634.
- [73] Li Z, Gavriluk K, Gavves E, et al. Videolstm convolves, attends and flows for action recognition[J]. Computer Vision and Image Understanding, 2018, 166(C): 41-50.
- [74] Sudhakaran S, Lanz O. Learning to detect violent videos using convolutional long short-term memory[C]. 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Lecce: IEEE, 2017: 1-6.
- [75] Chong Y S, Tay Y H. Abnormal event detection in videos using spatiotemporal autoencoder[C]. International Symposium on Neural Networks. Hokkaido: Springer, Cham, 2017: 189-196.
- [76] Chen C, Zhang B C, Hou Z J, et al. Action recognition from depth sequences using weighted fusion of 2D and 3D auto-correlation of gradients features[J]. Multimedia Tools and Applications, 2017, 76(3): 4651-4669.
- [77] Liu M, Liu H, Chen C. Robust 3D action recognition through sampling local appearances and global distributions[J]. IEEE Transactions on Multimedia, 2017, 20(8): 1932-1947.
- [78] Ji X P, Cheng J, Tao D P, Wu X Y, et al. The spatial

- Laplacian and temporal energy pyramid representation for human action recognition using depth sequences[J]. Knowledge-Based Systems, 2017, 122:64-74.
- [79] Hinami R, Mei T, Satoh S. Joint detection and recounting of abnormal events by learning deep generic knowledge[C]. Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 3619-3627.
- [80] De Geest R, Gavves E, Ghodrati A, et al. Online action detection[C]. European Conference on Computer Vision. Amsterdam: Springer, Cham, 2016: 269-284.
- [81] 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, 2019, 42(6): 1453-1467.
- [82] Liu J, Li Y, Song S, et al. Multi-modality multi-task recurrent neural network for online action detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(9): 2667-2682.
- [83] Xu M, Gao M, Chen Y T, et al. Temporal recurrent networks for online action detection[C]. Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 5532-5541.
- [84] He C, Shao J, Sun J. An anomaly-introduced learning method for abnormal event detection[J]. Multimedia Tools and Applications, 2018, 77(22): 29573-29588.
- [85] Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 6479-6488.
- [86] 肖进胜, 申梦瑶, 江明俊, 雷俊峰, 包振宇. 融合包注意力机制的监控视频异常行为检测 [J/OL]. 自动化学报. <https://doi.org/10.16383/j.aas.c190805>
(Xiao J S, Shen M Y, Jiang M J, et al. Abnormal Behavior Detection Algorithm with Video-Bag Attention Mechanism in Surveillance Video[J/OL]. ACTA AUTOMATICA SINICA. <https://doi.org/10.16383/j.aas.c190805>)
- [87] Zhong J X, Li N, Kong W, et al. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 1237-1246.
- [88] De la Torre F, Black M J. Robust principal component analysis for computer vision[C]. Proceedings Eighth IEEE International Conference on Computer Vision. Vancouver: IEEE, 2001, 1: 362-369.
- [89] Kim J, Grauman K. Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates[C]. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 2921-2928.
- [90] Chalapathy R, Menon A K, Chawla S. Robust, deep and inductive anomaly detection[C]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Skopje: Springer, Cham, 2017: 36-51.
- [91] Vu H, Nguyen T D, Travers A, et al. Energy-based localized anomaly detection in video surveillance[C]. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Jeju: Springer, Cham, 2017: 641-653.
- [92] Kiran B R, Thomas D M, Parakkal R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos[J]. Journal of Imaging, 2018, 4(2): 36.
- [93] Medel J R, Savakis A. Anomaly detection in video using predictive convolutional long short-term memory networks[J]. arXiv preprint arXiv: 1612. 00390, 2016.
- [94] Luo W, Liu W, Gao S. Remembering history with convolutional lstm for anomaly detection[C]. 2017 IEEE International Conference on Multimedia and Expo (ICME). Hong Kong: IEEE, 2017: 439-444.
- [95] Dhiman C, Vishwakarma D K. A review of state-of-the-art techniques for abnormal human activity recognition[J]. Engineering Applications of Artificial Intelligence, 2019, 77: 21-45.
- [96] Ravanbakhsh M, Nabi M, Sangineto E, et al. Abnormal event detection in videos using generative adversarial nets[C]. 2017 IEEE International Conference on Image Processing (ICIP). Beijing: IEEE, 2017: 1577-1581.
- [97] Xu D, Yan Y, Ricci E, et al. Detecting anomalous events in videos by learning deep representations of appearance and motion[J]. Computer Vision and Image Understanding, 2017, 156: 117-127.
- [98] Sabokrou M, Fayyaz M, Fathy M, et al. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes[J]. IEEE Transactions on Image Processing, 2017, 26(4): 1992-2004.
- [99] Mahadevan V, Li W, Bhalodia V, et al. Anomaly detection in crowded scenes[C]. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010: 1975-1981.
- [100] Mehran R, Oyama A, Shah M. Abnormal crowd behavior detection using social force model[C]. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Florida: IEEE, 2009: 935-942.
- [101] Lu C, Shi J, Jia J. Abnormal event detection at 150 fps in matlab[C]. Proceedings of the IEEE international conference on computer vision. Sydney: IEEE, 2013: 2720-2727.
- [102] Adam A, Rivlin E, Shimshoni I, et al. Robust real-time unusual event detection using multiple fixed-location monitors[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 30(3): 555-560.
- [103] Luo W, Liu W, Gao S. A revisit of sparse coding based anomaly detection in stacked rnn framework[C]. Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 341-349.
- [104] Del Giorno A, Bagnell J A, Hebert M. A discriminative

- framework for anomaly detection in large videos[C]. European Conference on Computer Vision. Amsterdam: Springer, Cham, 2016: 334-349.
- [105] Hasan M, Choi J, Neumann J, et al. Learning temporal regularity in video sequences[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas: IEEE, 2016: 733-742.
- [106] Sun Q, Liu H, Harada T. Online growing neural gas for anomaly detection in changing surveillance scenes[J]. Pattern Recognition, 2017, 64: 187-201.
- [107] Tudor Ionescu R, Smeureanu S, Alexe B, et al. Unmasking the abnormal events in video[C]. Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2895-2903.
- [108] Liu W, Luo W, Lian D, et al. Future frame prediction for anomaly detection-a new baseline[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 6536-6545.
- [109] Ravanbakhsh M, Nabi M, Mousavi H, et al. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection[C]. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe: IEEE, 2018: 1689-1698.
- [110] Ionescu R T, Smeureanu S, Popescu M, et al. Detecting abnormal events in video using narrowed normality clusters[C]. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa Village: IEEE, 2019: 1951-1960.
- [111] Ionescu R T, Khan F S, Georgescu M I, et al. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7842-7851.
- [112] Sun C, Jia Y, Song H, et al. Adversarial 3D convolutional auto-encoder for abnormal event detection in videos[J]. IEEE Transactions on Multimedia, 2020, doi: 10.1109/TMM.2020.3023303.

作者简介

张晓平(1991—), 女, 讲师, 博士, 从事人工智能、图像处理、机器学习与智能机器人等研究, E-mail: zhangxiaoping369@163.com;

纪佳慧(1997—), 女, 硕士生, 从事图像处理、机器学习、异常行为分析等研究, E-mail: JiJiaHui10@163.com;

王力(1978—), 男, 教授, 博士生导师, 从事智能交通控制、道路交通工程等研究, E-mail: wangli939@ncut.edu.cn;

何忠贺(1982—), 男, 副教授, 博士, 从事混杂系统控制、复杂系统协作控制以及交通系统建模与控制等研究, E-mail: zhonghehe@ncut.edu.cn;

刘世达(1988—), 男, 讲师, 博士, 从事人工智能、无模型自适应控制、异常行为分析等研究, E-mail: lsdshiwo@hotmail.com.