第52卷第5期
2020年10月

南京航空航天大学学报
Journal of Nanjing University of Aeronautics & Astronautics

Vol. 52 No. 5
Oct. 2020

DOI:10.16356/j.1005-2615.2020.05.002

监控视频中异常事件检测技术研究进展

吉根林，许　振，李欣璐，赵　斌
（南京师范大学计算机科学与技术学院，南京，210046）

摘要：异常事件检测技术是当前智能监控技术研究领域关注的一个热点，作为计算机视觉的重要研究内容，其主要目标是利用计算机自动检测出可被视为异常的事件。传统方法存在低层视频特征描述能力弱，异常检测方法计算代价大，对复杂场景建模时鲁棒性差等方面的限制。本文结合国内外的研究现状和目前的主流方法，介绍了监控视频中异常事件检测涉及的基本技术，分析了各类监控视频特征提取方法、特征学习模型和异常检测方法的优缺点，整理归纳了可用于监控视频中异常事件检测的常用实验数据集，最后讨论了监控视频中异常事件检测技术的难点、挑战及未来发展趋势。

关键词：异常事件检测；监控视频分析；行为识别；计算机视觉；机器学习
中图分类号：TP301.6　　**文献标志码：**A　　**文章编号：**1005-2615(2020)05-0685-10

Progress on Abnormal Event Detection Technology in Video Surveillance

JI Genlin, *XU Zhen*, *LI Xinlu*, *ZHAO Bin*
(School of Computer Science and Technology, Nanjing Normal University, Nanjing, 210046, China)

Abstract: Abnormal event detection is a hot topic in the field of intelligent surveillance monitoring technology research currently. As an important research content of computer vision, its main goal is to use computers to automatically detect abnormal events. Traditional methods have limitations in the weakness of low-level video feature description ability, high computational cost of anomaly detection methods and poor robustness in modeling complex scenes. In recent years, how to design a high-level semantic feature extraction method, accelerate the process of abnormal event detection, and model complex scenes such as multiple cameras have become the forefront topic of current research. Based on the research situation at home and abroad and the mainstream methods, this paper introduces the basic techniques involved in abnormal event detection in surveillance videos, and analyzes the advantages and disadvantages of various types of surveillance video feature extraction methods, feature learning models and anomaly detection methods. This paper also summarizes the commonly used benchmark datasets that can be used to abnormal event detection in surveillance videos. Finally, we discuss the difficulties, challenges and future development trends of abnormal event detection in surveillance videos.

Key words: abnormal event detection; surveillance analysis; activity recognition; computer vision; machine learning

随着中国天网工程、雪亮工程等重大安防项目的战略部署以及公众对构建平安城市的迫切需

基金项目：国家自然科学基金(41971343)资助项目。
收稿日期：2020-06-20；**修订日期：**2020-08-21
作者简介：吉根林，男，教授，博士生导师，研究方向：数据库，数据挖掘，在各类权威和核心期刊上发表论文90多篇，江苏省高校“青蓝工程”中青年学术带头人。
通信作者：许振，男，硕士研究生，E-mail：182202028@njnu.edu.cn。

引用格式：吉根林，许振，李欣璐，等. 监控视频中异常事件检测技术研究进展[J]. 南京航空航天大学学报，2020，52(5)：685-694. JI Genlin, XU Zhen, LI Xinlu, et al. Progress on abnormal event detection technology in video surveillance[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2020, 52(5): 685-694.

求，智能监控视频分析（Intelligent surveillance video analytics，ISVA）技术已经渗透进每个人的生活。ISVA技术是利用计算机视觉、模式识别和机器学习等方法对监控视频序列自动分析和处理的技术。监控视频中异常事件检测技术是ISVA技术的重要分支，它解决的主要问题是：利用计算机对监控视频数据分析和处理，学习和理解视频中的动作和行为，对这些动作和行为进行异常判断和定位。监控视频中异常事件检测技术涉及对象检测与跟踪、视频特征提取、低层视频分析以及高层语义理解等多项技术。由于监控视频数据具有高度冗余、容量巨大以及分辨率低等特点，监控视频中异常事件检测技术能够大幅节省人工观察监控视频的成本、降低人工方式漏检带来的安全隐患，也可为建设智慧城市等领域提供决策支持。因此，监控视频中异常事件检测技术具有重要的研究意义和应用价值。

监控视频中异常事件检测技术发展至今，国内外已经产生了一些重要成果。学者们利用手工设计的方法，较好地表示了视频低层视觉特征，如方向梯度直方图（Histograms of oriented gradient，HOG）、光流直方图（Histograms of oriented optical flow，HOF）、混合动态纹理（Mixtures of dynamic texture，MDT）[1]等，也有学者在低层视觉特征的基础上设计中层语义特征来理解群体行为，如社会力模型[2]、粒子群直方图（Histogram of swarm，HOS）[3]。另外，以稀疏表示、自编码器（Auto-encoder，AE）[4]、循环神经网络（Recurrent neural network，RNN）[5]以及生成对抗网络（Generative adversarial network，GAN）[6]为代表的异常事件检测模型也表现出优异性能。视频分析与检索领域的国际性权威评测TRECVID（TREC video retrieval evaluation）中设置了监控视频事件检测任务，北京大学研究团队在2012年的评测中，分别采用基于对象检测与跟踪相融合的方法、基于序列立方体特征的序列分类方法和序列非均衡分类方法进行异常事件检测，在参赛队伍中成绩突出。

但是监控视频中异常事件检测技术也面临着诸多挑战。（1）光照、遮挡、视角等非受控条件导致手工特征性能受限，大多数视频特征描述还停留在低层特征阶段，在高层语义特征层面难以突破；（2）大量实时视频数据在线分析时，处理效率低下导致异常事件预警的高延迟；（3）针对多监控摄像头下视频智能分析缺乏有效技术。因此，监控视频异常事件检测技术的理论研究与真实场景中现实应用有很大差距，仍有许多问题亟须解决。

针对监控视频中异常事件检测技术，本文研究已有的重要研究成果，特别关注监控视频特征提取、特征学习模型和异常检测方法上的优缺点和适用场景。在最新研究成果分析和概述的基础上，展望监控视频中异常事件检测技术的未来发展趋势，并给出了监控视频中异常事件检测的实验数据集，为即将进入本领域的学者提供参考。与现有相关综述文献相比，本文更系统地介绍了监控视频异常事件检测的完整工作流程，对依据不同基本假设的异常事件检测模型进行形式化描述，并分析了导致在线检测系统处理效率低下的原因并提出解决方案。

1 异常事件检测过程

监控视频中异常事件检测过程可分为4个阶段，如图1所示。（1）视频数据预处理：对视频帧进行形态学处理、图像降噪和图像增强等。（2）视频特征提取：从输入监控视频中提取关键的表观和运动特征。（3）特征学习：利用视频特征建立正常或异常事件模式和变化规律的行为模型，以规则、模型的形式保存在数据库中。（4）异常检测：衡量测试视频中的基本事件与已建立的行为模型之间的匹配程度，以异常分数或类别标签的形式返回检测结果。

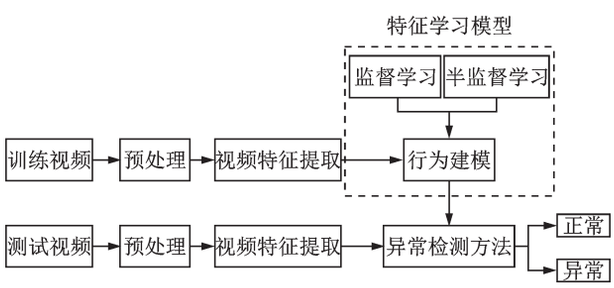

图1 监控视频中异常事件检测过程

Fig.1 Abnormal event detection process in surveillance video

2 监控视频特征提取

监控视频中异常事件检测是在提取视频特征并编码的基础上运用特定异常检测技术完成的。编码后视频特征会成为算法的输入，因此视频特征提取对于异常事件检测能力起到关键作用。文献[7]认为视频中的基本事件可以由低层视觉特征和高层语义特征来表示，基于低层视觉特征的事件表示是手工提取不包含语义信息的简单视觉特征，如光流、梯度等，而基于高层语义特征的事件表示需要对视频内容进行复杂处理以捕捉语义信息，如利用目标跟踪算法获得时空轨迹[8]、模拟

人群运动提取交互能量[2]等。随着深度学习技术的兴起，人们开始利用深度神经网络（Deep neural network，DNN）自动提取高层语义特征，将其应用于视频异常检测且表现出优于手工特征的性能。本文将用于监控视频中异常事件检测任务的特征提取方法分为两大类：基于手工设计的方法和基于深度学习的方法，对文献[7]中的事件表示方法进行补充。

2.1　基于手工设计的特征提取

基于手工设计的特征提取方法一般可获得3类低层视觉特征：表观特征、运动特征和融合特征。表观特征包括视频前景、人体骨架和空间梯度等，运动特征包括光流、时空轨迹等。这类方法提取的特征往往相对单一、在复杂场景下鲁棒性较差，但可解释性很强。

2.1.1　表观特征提取

提取视频的表观特征是为了检测出表观异常。由于异常事件通常发生在视频前景区域，张俊阳等[9]利用自适应混合高斯背景分割算法提取视频前景以判断人群异常行为，而Wei等[10]将传统的背景分割算法与DNN结合，用于交通异常行为检测。随着新技术的发展，人体骨架信息也能被快速提取，Morais等[11]提取人体骨架用于判断个体异常行为。另一种常用表观特征是空间梯度的统计信息——方向梯度直方图。表观特征提取方法不必检测视频中的移动目标，适用于拥挤人群场景，但是复杂背景干扰可能导致对象的边缘特征丢失，关于场景的特定信息较少。

2.1.2　运动特征提取

提取视频的运动特征是为了检测出运动异常。光流由美国心理学家Gibson于1950年正式提出，定义是空间运动物体在观察成像平面上像素运动的瞬时速度，是最常用的低层运动特征。光流特征有3种使用方式，（1）纯光流特征直接用于视频异常检测任务，如汪洪流等[12]构建纯光流特征图结构，利用特征图结构的迭代尺度化变化降低光流特征数量；（2）构建光流场的统计信息——光流直方图，Cong等[13]还考虑光流场矢量的模长，使用多尺度光流直方图（Multi-scale histograms of optical flow，MHOF）来捕捉视频运动信息，而Colque等[14]提出基于光流场矢量的方向与模长、像素熵值的运动特征HOFME（Histogram of optical flow orientation and magnitude and entropy）；（3）利用光流信息生成更高层次的语义特征，社会力模型就是采用这种方式，而Qasim等[3]先计算光流模长的方差，再利用蚁群优化聚类算法和捕食者-猎物优化算法寻找显著区域，最后对显著区域信息编码构建粒子群直方图。但是光流特征的缺点也很明显，在光照强度发生突然变化的情况下容易产生误判，因此基于光流特征设计的异常事件检测算法鲁棒性较低，并且相比于其他低层特征提取方法，光流法的计算量偏大。

时空轨迹是视频的一种高级运动特征描述，在交通异常事件检测中很有效，常用于检测违规变道、追尾以及闯红灯等交通违法行为[15]。Li等[16]为了弥补光流和表观特征的局限性，结合对象检测和跟踪方法提取远距离目标的轨迹以降低漏检率。由于获取时空轨迹依赖对象检测和跟踪技术，运动对象之间的遮挡会严重轨迹质量，因此时空轨迹在一定程度上可以有效描述视频内容，适合运动对象较少的场景，在拥挤人群场景下鲁棒性差。

2.1.3　特征融合

多种特征组合比单特征拥有更强的表达能力，许多学者开始融合表观特征和运动特征，更准确地描述正常和异常事件模式和变化规律。最简单的融合方式是柳晶晶等[17]将各特征向量连接成维度更高的新向量，Xu等[18]先对视频前景的图像块与光流场块进行像素级别的融合，形成新的图像块，再将它们输入去噪自编码网络以学习融合图像块的深度特征表示，而Fang等[19]则使用主元分析网络（Principal component analysis networks，PCANet）融合视频帧显著性信息和多尺度光流直方图。

2.2　基于深度学习的特征提取

深度学习能够学习到场景特定的语义表达，产生高层语义特征，在对象检测和跟踪、识别等领域取得了优异成绩。于是，学者们开始利用卷积神经网络、自编码器等深度学习模型来完成监控视频中异常事件检测任务。卷积神经网络（Convolutional neural network，CNN）是计算机视觉中性能优异的DNN模型之一，与应用于图像特征提取的二维CNN不同的是，应用于视频特征提取的CNN的卷积核通常是三维的，能够同时执行空间和时间两个维度的卷积操作来捕捉时空特征。学者们将这种特殊结构的CNN模型称为时空卷积神经网络（Spatial-temporal convolutional neural network，ST-CNN）[20]。受到ST-CNN模型的启发，学者们又陆续设计出双流卷积网络[21]、全卷积网络[22]、时空残差网络[23]、栈式自编码器[18]以及稀疏去噪自编码器（Sparse denoising auto-encoders，SDAE）[24]等DNN视频特征提取模型。

2.3　各种特征提取方法的比较

用于异常事件检测的特征选择与应用场景密切相关，各种特征提取方法的比较如表1所示。针对不同的应用场景，如何选择有效的特征提取方法

表1 各种特征提取方法特点和适用场景

Table 1 Characteristics and application scenarios of different feature extraction methods

特征提取方法	特点	适用场景
时空立方体[25]	依赖更少训练数据，利于稀疏表示，可实现快速异常定位、实时监测任务	室内公共场所、地铁站、码头、公共汽车和公路
时空梯度[26]	同时保留视频帧的局部信息和全局结构	户外公共场所
时空兴趣点[27]	仅对局部兴趣区域提取特征，计算代价小	地铁站、停车场、人行道、交叉路口和公路
混合动态纹理[1]	考虑表观特征的动态性，利于检测时间和空间两个维度的异常	人行道
光流方向、模长和熵直方图[14]	综合考虑方向、速度和熵，对异常事件高敏感，漏检率低	室内拥挤场所、人行道和地铁站
粒子群直方图[3]	利于检测全局异常	高密度拥挤人群
完整轨迹[28]	保留对象移动信息，利于运动模式分析，对异常模式判别能力强	交叉路口、公路、低密度人群
子轨迹段[29]	更小粒度的子轨迹被设计专门用于检测局部异常，在拥挤场景下鲁棒性好	中高密度拥挤人群
光流＋空间梯度[30]	具备对异常事件高敏感性和正常行为高鲁棒性	户外公共场所、人行道和公路
光流＋位置＋轨迹[16]	检测车辆异常行为时鲁棒性高	公路
方向梯度直方图＋光流直方图＋轨迹[31]	对多类型异常判别能力强	户外公共场所、人行道和地铁站
DNN自动提取的特征[32]	考虑时间维度平滑性和稀疏性，易于特征降维	所有场景

具有挑战性。例如，人群稀疏场景和稠密场景中的视频特征提取方法截然不同，时空轨迹只能从人群稀疏场景中提取，而无法从人群稠密场景中提取，这是因为移动对象过度拥挤会产生的遮挡问题，从而无法准确获得运动轨迹，因此在人群稠密场景中提取低层视觉特征更为有效。

从特征工程角度看，手工设计特征需要一定的领域知识，而DNN是模仿人类神经元工作方式的网络结构，能够从海量数据中自动学习特定规律，对复杂场景的鲁棒性强，因此基于深度学习的特征提取方法的更有研究前景。

从特征语义性角度看，低层视觉特征表达能力有限，高层语义特征能够更好地刻画监控视频中的各种行为。产生高层语义特征共有3种方式：(1)低层特征直接映射为高层语义特征；(2)从中层语义特征过渡到高层语义特征，中层语义特征提取建立在低层特征基础上，关注情境信息和社会属性对群体行为的影响；(3)使用DNN自动提取高层语义特征，值得注意的是，背景偏置现象[33]的存在可能会限制DNN的性能，而且DNN并非越深越好，深度过高可能导致模型训练困难和计算量显著增加。

3 异常事件检测模型

提取监控视频特征之后，需要利用这些特征建立异常事件检测模型。建模过程大致可分为两个阶段：训练特征学习模型和离群点检测。由于学术界对于异常事件并没有明确的定义，也很难罗列出所有异常事件，学者们对异常事件的基本假设不同，因此建模角度也各不相同。依据不同的基本假设和建模角度，可将异常事件检测模型分为4种类型。各种异常事件检测模型的优缺点对比如表2所示。

表2 不同类型异常事件检测模型的优缺点

Table 2 Advantages and disadvantages of different abnormal event detection models

异常事件检测模型	优点	缺点
基于概率推断的模型	理论完备，模型简单、易于存储，计算代价最小，可在线更新模型	在高维样本空间中，需要大量训练样本来拟合概率生成模型
基于重构的模型	最灵活，适合于较少的高维训练样本，抗噪能力强	计算代价最高
基于分类的模型	受训练样本分布影响小，抗噪能力强	分类器参数的选择依赖经验
基于聚类的模型	无须数据分布的先验知识，可在线更新模型	高维特征空间下聚类计算代价高

3.1　基于概率推断的模型

异常事件发生的概率远远小于正常事件。依据这一基本假设,有学者认为异常事件检测模型的建立是对正常视频行为建模,挖掘正常行为规律后找出与这些规律存在较大偏差的行为判定为异常。按照这一研究思路形成了基于概率推断的异常事件检测模型,它的核心思想是假定正常数据由随机过程产生,学习一个拟合正常数据的概率生成模型,然后识别该模型低概率区域中的数据对象,形式化描述为:设x为视频中局部时空区域内的特征,$p_X(x)$是正常数据观测值$\boldsymbol{X}$的分布,存在假设检验模型:H_0: x服从于概率分布$p_X(x)$;H_1: x服从于$p_X(x)$以外的未知分布,当$p_X(x)<\epsilon$时,拒绝H_0,接受H_1,ϵ为未知分布的标准化常数。

按照需要拟合的概率生成模型参数是否已知,建模方法又可以分为参数方法和非参数方法。常用的参数方法有高斯混合模型(Gaussian mixture model, GMM)、隐马尔科夫模型(Hidden Markov model, HMM)、混合概率主元分析(Mixture of probabilistic principal component analysis, MP-PCA)等。文献[27]使用最大似然参数估计方法,对视频前景和光流能量建立GMM模型,对光流直方图建立HMM模型,Feng等[34]则提出一种堆叠多个GMM的模型对正常行为模式建模,而Li等[35]假设正常事件的表观和运动特征深度表示都服从多变量高斯分布,结合自编码器和生成对抗网络联合训练以拟合这种分布。非参数方法中最经典是Saleemi等[36]提出的核密度估计(Kernel density estimator, KDE)方法。

3.2　基于重构的模型

这类模型假定正常数据能够嵌入更低维度子空间中,并且能以很小的误差进行重构,而异常数据的重构误差较大。依据这一基本假设,学者们常用方法是利用正常行为数据训练一组模型,在测试阶段,若发现样本无法使用这组模型进行重构,则将其判定为异常。最常用的方法是稀疏表示和自编码器。

稀疏表示的基本思想是所有正常视频数据都可以用一组基底线性表示且表示系数稀疏,无法通过这组基底线性稀疏表示的样本即为异常,形式化描述为:利用正常样本$\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$训练一个过完备字典$\boldsymbol{D}\in\boldsymbol{R}^{d\times k}$,其中$d\ll k$,$k$为字典中基底个数,优化目标为

$$\min_{D,A}\ \|\boldsymbol{X}-\boldsymbol{DA}\|_F^2+\lambda\|\boldsymbol{A}\|_{M_1} \tag{1}$$

式中:$\boldsymbol{A}=\{\alpha_1,\alpha_2,\cdots,\alpha_m\}\in\boldsymbol{R}^{k\times n}$为$\boldsymbol{X}$的稀疏表示,对测试样本$\boldsymbol{y}\in\boldsymbol{R}^d$,利用字典$\boldsymbol{D}$中基底的稀疏线性组合对其重构,即

$$\alpha^*=\arg\min\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{D\alpha}\|_2^2+\lambda\|\boldsymbol{\alpha}\|_1 \tag{2}$$

式中:$\alpha^*\in\boldsymbol{R}^k$是样本$\boldsymbol{y}$的重构系数,$y$的稀疏重构代价为

$$\mathrm{SRC}=\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{D}\alpha^*\|_2^2+\lambda\|\alpha^*\|_1 \tag{3}$$

可作为异常事件的度量准则。Yuan等[37]在式(1)中加入视频结构信息,Jardim等[38]在式(1)中加入域转移约束条件并改进优化方法,用于检测移动摄像机拍摄的异常事件,而Liu等[39]提出一种包含两个动态更新稀疏表示过程的异常事件检测方法,一个判断区域是否正常,另一个判断区域是否异常。稀疏表示方法的缺点是数量过多的基底导致范数优化问题计算代价极高,为克服这一缺点,Lu等[26]提出构造多个包含少量基底的小字典,将每组基底上的系数表示问题转化为最小化均方误差问题,大大减少计算代价,而胡正平等[40]则从约简字典学习输入的角度来降低计算代价,他们利用光流和非均匀元胞分割对视频中运动目标进行提取,字典学习时采用AP聚类将训练样本中具有代表性的特征作为字典,极大地降低了字典维度。

自编码器的异常事件检测过程是先利用反向传播算法训练正常视频的自编码网络,再将未知视频输入网络,根据前向传播算法获得网络输出,计算输出与输入之间的重构误差,当误差高于某一阈值时,判定该视频中存在异常事件。Hasan等[31]设计的时空卷积自编码网络(Convolutional auto-encoders, Conv-AE)有效地学习了视频序列中的正常模式,而袁静等[41]在稀疏去噪自编码器的基础上,增加梯度差约束条件改进了自编码网络的编解码效果。

3.3　基于分类的模型

当监控视频的训练数据中包含正常和异常事件的标注时,异常事件的稀疏性导致训练数据高度有偏,即异常样本数量远远低于正常样本数量,因此有学者认为监控视频中异常事件检测是典型的不平衡分类问题。依据这一基本假设,在某个特征空间下,正常与异常事件之间存在明显边界,建模过程就是训练一个分类器,如图2(a)所示。Xu等[18]按照这一研究思路使用单类支持向量机(One-class support vector machine, OCSVM)对深度融合表示后的特征进行分类,Sun等[42]将OCSVM整合进CNN以构建端到端模型,而Xu等[43]提出的自适应帧内分类网络将原先的单分类问题发展为多分类问题。更高级的分类方法是Zhou等[44]使用的时空卷积神经网络,对视频块执行多

次卷积和池化操作后，归一化全连接层的输出以计算异常概率，但是这类方法的难点是如何利用不均衡的样本集训练分类网络，Chu等[45]利用手工特征的稀疏编码结果来指导网络的无监督训练，美国中佛罗里达大学计算机视觉研究中心提出的弱监督算法框架[27]则利用深度多实例排序的方法训练分类网络。

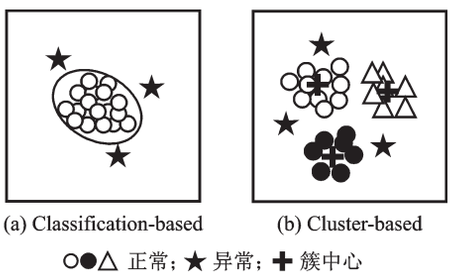

图2 基于分类的模型和基于聚类的模型示意图

Fig.2 Schematic diagrams of classification-based model and cluster-based model

3.4 基于聚类的模型

由于正常事件种类较少、特征相似，而异常事件种类繁多，不胜枚举，因此有学者认为，在某个特征空间下，正常事件的分布是紧致的，且与异常事件是可区分的。依据这一基本假设，当训练数据中只包含正常事件时，正常事件的特征数据会被聚类成若干个大而紧凑的簇，而异常事件的特征数据则远离这些簇的聚类中心，即属于小而稀疏的簇或不属于任何簇，如图2(b)所示。黄鑫等[15]使用这类模型来检测监控视频中的车辆异常行为，他们利用均值漂移算法对车辆速度值和角度值进行聚类，使用运动特征量到聚类中心的欧式距离作为异常判别准则。值得一提的是，Nawaratne等[46]提出了结合模糊聚类的增量时空学习器，可以根据进入系统的新样本在线更新模型，具有较强的场景自适应能力。

4 异常事件类型和实验数据集

4.1 异常事件类型

监控视频中大多数异常事件具有特性：

(1)稀疏性——异常事件相比于正常行为而言，它们是偶尔发生的。

(2)特殊性——异常事件与正常行为是与众不同的，它们有十分明显的特征。

(3)可变性——异常事件并非一成不变，随着时间的推移，异常事件和正常行为之间可以相互转化。

异常也称为离群点，可分为3类：全局离群点、情境离群点和集体离群点。全局离群点是最简单的异常形式，并且最容易检测。如果一个数据对象显著地偏离数据集中的其他对象，那么它对应为全局离群点。情境离群点是关于数据对象的特定情境显著偏离其他对象。例如，如果一个骑车者与其他骑车者相比骑得更快，若交通十分拥挤，将该骑车者判定为异常；若交通很通畅，则判定为正常。在情境离群点的检测中，情境必须作为问题定义的一部分加以说明。数据对象的一个子集形成集体离群点，尽管个体数据对象可能不是离群点，这些对象作为整体显著地偏离整个数据集。例如，公共场所短时间内聚集的一群人可被视为集体异常事件。

按照异常事件在场景中发生规模的大小和持续时间的长短，也可分为局部异常和全局异常。局部异常事件关注如下异常：在时间和空间关系中较周围邻域剧烈的异常活动，通常发生在场景的特定区域，如局部反方向运动和快速驶入画面的货车。全局异常事件关注如下异常：在一定时间段发生的异常大于其他时间段，可以存在于视频的某个片段或某一帧中，而不指出其发生的空间确切位置，如人群聚集、恐慌和暴乱等剧烈的异常活动。目前已经有Vagia等[47]提出能够在交通场景中同时检测全局异常和局部异常的先进方法。

4.2 实验数据集

近年来学者们开发了一些监控视频异常事件检测数据集，如表3所示。人群稠密场景的特点是人流量大，比如地铁站、医院和公园等公共场所，代表数据集是UMN，UCSD，Avenue和Hockey Fight。人群稀疏场景的代表性数据集是CAVIAR，PETS 2009，ShanghaiTech和Subway Exit & Entrance。交通场景是指交叉路口、高速公路和停车场等，异常事件主要由车辆产生，如车辆超速、闯红灯和越过停止线等，代表性数据集是UCF-Crime，QMUL Junction，ViF和LV。

4.2.1 检测算法性能评价

异常事件检测的性能评价通常分为帧级别和像素级别两种粒度。在帧级别评价中，一旦检测出某一帧包含异常事件，则将该帧判定为异常帧。帧级别评价仅仅关注异常事件在时间维度上定位的准确性，不考虑空间维度定位，可能会出现假阳性情况，即将真实异常帧中的正常事件误判为异常事件，因此同时关注异常事件空间和时间定位的像素级别准则更加可靠。在像素级别评价中，如果检测到某一帧的异常像素块占真实异常像素块的比例达到某一阈值，则该帧被判定为异常帧。常用的算法性能指标是受试者工作特性(Receiver operating characteristic，ROC)曲线下的面积，即AUC(Area

表 3　不同应用场景下的实验数据集

Table 3　Experiment datasets in different application scenarios

实验数据集	场景	异常事件	分辨率/像素	总时长
QMUL Junction	公路交通	违反交通规则行为	360×288	1 h
CAVIAR	商场、球场等室内公共场所	行人单独行走、休息、昏厥、多人会面、进出商场、打架和留下可疑物品	384×288	2 h
Hockey Fight		曲棍球比赛中的暴力斗殴行为	360×288	16.7 min
Subway Exit & Entrance	地铁站	乘客走错路线和跳过检票口	720×576	3 h
ShanghaiTech	学校	骑自行车、推婴儿车、翻越栏杆、奔跑、追逐和争吵	856×480	3.01 h
Avenue		行人奔跑、行走方向错误、出现非行人实体和留下可疑物品	640×360	30 min
UCSD Ped1 & Ped2	人行道	人行道上出现非行人实体和异常行人运动模式	238×158 360×240	10 min
PETS 2009		行人奔跑、人群迅速分散和人群聚集	768×576	8 min
ViF	以上多种场景组合	拥挤人群的暴力行为	320×240	14.76 min
LV		室内、街道、公路等公共区域内32种异常事件	176×144 1 280×720	3.93 h
UCF-Crime		逮捕、纵火、交通事故、盗窃、爆炸、打架、枪击、入店行窃和故意破坏等13种犯罪案件	320×240	128 h
UMN		遗落物品、异常人群活动、摄像机遮挡、受限制区域内运动和街头行人闲逛	320×240	27 min

under curve)。假定ROC曲线是由坐标为$\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$的点按序连接而形成,则AUC可估算为

$$\mathrm{AUC}=\frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)(y_i+y_{i+1})$$

如果一个异常事件检测算法的AUC越大,则表明该算法性能越好。

4.2.2　典型异常检测算法性能

近年来提出的典型异常事件检测方法在不同实验数据集上的AUC实验结果如表4所示。

表 4　异常事件检测方法在实验数据集上的AUC值

Table 4　AUC values of the abnormal event detection methods on the experiment datasets

特征提取方法	异常事件检测模型	AUC/%								
		Ped1		Ped2		Entrance	Exit	Avenue		UMN
		f	*p*	*f*	*p*	*f*	*f*	*f*	*p*	*f*
HOFME[14]	KNN	—	72.7	—	87.5	81.5	84.9	—	—	—
HOS+HOG[48]	OCSVM	85.0	74.0	78.0	74.0	—	—	—	—	98.1
MDT[1]	CRF	81.8	66.2	85.0	82.9	89.7	90.8	—	—	**99.5**
SDAE[18]	OCSVM	92.1	67.2	90.8	—	—	—	71.7	33.2	—
HOG+HOF+MBH	判别学习[49]	—	—	—	—	69.1	82.4	**89.6**	—	91.0
时空梯度	快速字典学习[21]	91.8	63.8	—	—	—	—	—	—	—
时空立方体	结构字典学习[37]	93.2	71.6	—	—	—	—	—	—	—
PCANet	深度GMM[34]	92.5	69.9	**97.9**	55.8	—	—	75.4	—	—
时空立方体	Deep-cascade[50]	**99.6**	—	95.3	—	—	—	—	—	—
视频序列	Stan[51]	82.1	—	96.5	—	—	—	87.2	—	—
	sRNN[5]	—	—	—	92.2	—	—	87.2	81.7	—
光流+空间梯度	U-Net+GAN[25]	83.1	—	95.4	—	—	—	84.9	—	—
	ConvAE+GAN[52]	—	—	—	—	—	—	—	—	99.6
	C3D-AE[53]	92.3	—	91.2	—	—	—	80.9	—	—
视频序列	ConvAE[26]	81.0	**89.9**	90.0	87.4	**94.3**	**94.0**	70.2	80.0	—
	ConvLSTMAE[54]	89.9	75.5	87.4	88.1	93.3	87.7	80.3	77.0	—
	AnoPCN[55]	—	—	—	**96.8**	—	—	—	**86.2**	—

注:*f*表示帧级别,*p*表示像素级别。

5 结 论

目前,监控视频中异常事件检测技术已经在低层视觉特征提取方法、特征学习模型和异常检测方法等方面取得重要成果。但是,学术界的大多数研究成果尚且无法实际应用。为了设计出符合工业产品级要求的原型系统,学者们需要在高层语义特征提取、复杂场景下行为建模以及实时数据处理等方面深入研究。

监控视频中异常事件检测的挑战很大程度上来源于场景复杂性。(1)异常事件种类丰富,基本事件的特征表示对异常检测效果影响巨大,而低层视觉特征难以准确地对复杂场景下的基本事件进行表示,目前学术界缺少由低层视觉特征构建高层语义特征的研究成果,探索丰富多样的语义特征表达形式值得学者进一步研究。(2)随着长时间的监控,异常事件变化多端,先前的异常事件可能会成为正常事件,无专家交互很难辨别异常行为,离线训练的特征学习模型无法适应这种变化,因此必须采用具备模型自动更新能力的在线特征学习模型。(3)为了增强模型的鲁棒性,减小环境噪声对检测结果的影响同样值得关注。

目前国内大多数城市的公共区域都已经安装视频监控系统,收集的视频能够为众多犯罪案件提供重要线索,但是这些摄像头只能被动记录视频,作为事后调查的依据,而无法实时预警。异常事件实时检测系统的研究正好满足智能监控中自动实时预警的迫切需求,典型的系统如中科院自动化所谭铁牛院士主持的实时智能视频监控预警系统。监控视频中异常事件检测任务既是数据密集型,也是计算密集型,非结构化、高度冗余的视频数据增加了实时处理分析的难度,目前大多数异常事件检测算法难以满足实时预警的要求。在线处理实时产生的监控视频大数据不仅需要充分利用视频编解码信息,甚至需要高性能计算机的支持或发挥云计算平台并行计算的优势。另外,开发多摄像头框架下的实时检测系统也值得进一步探索。

参考文献:

[1] WEIXIN LI, 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.

[2] MEHRAN R, OYAMA A, SHAH M. Abnormal crowd behavior detection using social force model[C]//IEEE Conference on Computer Vision and Pattern Recognition.[S.l.]: IEEE Computer Society, 2009: 935-942.

[3] QASIM T, BHATTI N. A hybrid swarm intelligence based approach for abnormal event detection in crowded environments[J]. Pattern Recognition Letters, 2019, 128: 220-225.

[4] RIBEIRO M, LAZZARETTI A E, LOPES H S. A study of deep convolutional auto-encoders for anomaly detection in videos[J]. Pattern Recognition Letters, 2018, 105: 13-22.

[5] 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.[S.l.]: IEEE Computer Society, 2017: 341-349.

[6] VU H, NGUYEN T D, LE T, et al. Robust anomaly detection in videos using multilevel representations[C]//Proceedings of the AAAI Conference on Artificial Intelligence.[S.l.]: IEEE Computer Society, 2019, 33(1): 5216-5223.

[7] 吴新宇, 郭会文, 李楠楠, 等. 基于视频的人群异常事件检测综述[J]. 电子测量与仪器学报, 2014, 28(6): 575-584.
WU Xinyu, GUO Huiwen, LI Nannan, et al.Survey on the video-based abnormal event detection in crowd scenes[J]. Journal of Electronic Measurement and Instrumentation, 2014, 28(6): 575-584.

[8] 张静, 高伟, 刘安安, 等. 基于运动轨迹的视频语义事件建模方法[J]. 电子测量技术, 2013, 36(9): 31-40.
ZHANG Jing, GAO Wei, LIU An'an, et al.Modeling approach of the video semantic events based on motion[J]. Electronic Measurement Technology, 2013, 36(9): 31-40.

[9] 张俊阳, 谢维信, 植柯霖. 基于运动前景效应图特征的人群异常行为检测[J]. 信号处理, 2018, 34(3): 296-304.
ZHANG Junyang, XIE Weixin, ZHI Kelin. Abnormal crowd behavior detection based on motion effect map features[J]. Journal of Signal Processing, 2018, 34(3): 296-304.

[10] WEI J, ZHAO J, ZHAO Y, et al. Unsupervised anomaly detection for traffic surveillance based on background modeling[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops.[S.l.]: IEEE Computer Society, 2018: 129-136.

[11] MORAIS R, LE V, TRAN T, et al. Learning regularity in skeleton trajectories for anomaly detection in videos[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.[S.l.]: IEEE Computer Society, 2019: 11996-12004.

[12] 汪洪流, 郭春生. 图结构多尺度变换的视频异常检测[J]. 中国图象图形学报, 2017, 22(11): 1544-1552.

WANG Hongliu, GUO Chunsheng. Video anomaly detection of multiscale transformation of graph structure[J]. Journal of Image and Graphics, 2017, 22(11): 1544-1552.

[13] CONG Y, YUAN J, LIU J. Sparse reconstruction cost for abnormal event detection[C]//Proceedings of Computer Vision and Pattern Recognition. [S. l.]: IEEE Computer Society, 2011: 3449-3456.

[14] COLQUE R V H M, CAETANO C, DE ANDRADE M T L, et al. Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(3): 673-682.

[15] 黄鑫, 肖世德, 宋波. 监控视频中的车辆异常行为检测[J]. 计算机系统应用, 2018, 27(2): 125-131.
HUANG Xin, XIAO Shide, SONG Bo. Detection of vehicle's abnormal behaviors in surveillance video[J]. Computer Systems & Applications, 2018, 27(2): 125-131.

[16] LI X, LI W, LIU B, et al. Object-oriented anomaly detection in surveillance videos[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE Computer Society, 2018: 1907-1911.

[17] 柳晶晶, 陶华伟, 罗琳, 等. 梯度直方图和光流特征融合的视频图像异常行为检测算法[J]. 信号处理, 2016, 32(1): 1-7.
LIU Jingjing, TAO Huawei, LUO Lin, et al. Video anomaly detection algorithm combined with histogram of oriented gradients and optical flow[J]. Journal of Signal Processing, 2016, 32(1): 1-7.

[18] 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.

[19] FANG Z, FEI F, FANG Y, et al. Abnormal event detection in crowded scenes based on deep learning[J]. Multimedia Tools and Applications, 2016, 75(22): 14617-14639.

[20] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proceedings of IEEE International Conference on Computer Vision. [S.l.]: IEEE Computer Society, 2015: 4489-4497.

[21] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Proceedings of Annual Conference on Neural Information Processing Systems. [S. l.]: IEEE Computer Society, 2014: 568-576.

[22] SABOKROU M, FAYYAZ M, FATHY M, et al. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes[J]. Computer Vision and Image Understanding, 2018, 172: 88-97.

[23] QIU Z, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3D residual networks[C]//Proceedings of IEEE International Conference on Computer Vision. [S. l.]: IEEE Computer Society, 2017: 5534-5542.

[24] NARASIMHAN M G, SOWMYA KAMATH S. Dynamic video anomaly detection and localization using sparse denoising autoencoders[J]. Multimedia Tools and Applications, 2018, 77(11): 13173-13195.

[25] ROSHTKHARI M J, LEVINE M D. An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions[J]. Computer Vision and Image Understanding, 2013, 117(10): 1436-1452.

[26] LU C, SHI J, WANG W, et al. Fast abnormal event detection[J]. International Journal of Computer Vision, 2019, 127(8): 993-1011.

[27] LEYVA R, SANCHEZ V, LI C T. Video anomaly detection with compact feature sets for online performance[J]. IEEE Transactions on Image Processing, 2017, 26(7): 3463-3478.

[28] BERA A, KIM S, MANOCHA D. Realtime anomaly detection using trajectory-level crowd behavior learning[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops. [S.l.]: IEEE Computer Society, 2016: 1289-1296.

[29] LIN W, ZHOU Y, XU H, et al. A tube-and-droplet-based approach for representing and analyzing motion trajectories[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(8): 1489-1503.

[30] LIU W, LUO W, LIAN D, et al. Future frame prediction for anomaly detection—A new baseline[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE Computer Society, 2018: 6536-6545.

[31] HASAN M, CHOI J, NEUMANN J, et al. Learning temporal regularity in video sequences[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE Computer Society, 2016: 733-742.

[32] SULTANI W, CHEN C, SHAH M. Real-world anomaly detection in surveillance videos[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE Computer Society, 2018: 6479-6488.

[33] LIU K, MA H. Exploring background-bias for anomaly detection in surveillance videos[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York, NY, USA: ACM, 2019: 1490-1499.

[34] FENG Y, YUAN Y, LU X. Learning deep event models for crowd anomaly detection[J]. Neurocomputing, 2017, 5(29): 548-556.

[35] LI N, CHANG F. Video anomaly detection and localization via multivariate Gaussian fully convolution adversarial autoencoder[J]. Neurocomputing, 2019, 369: 92-105.

[36] SALEEMI I, SHAFIQUE K, SHAH M. Probabilistic modeling of scene dynamics for applications in visual surveillance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(8): 1472-1485.

[37] YUAN Y, FENG Y, LU X. Structured dictionary learning for abnormal event detection in crowded scenes[J]. Pattern Recognition, 2018, 73: 99-110.

[38] JARDIM E, THOMAZ L A, DA SILVA E A B, et al. Domain-transformable sparse representation for anomaly detection in moving-camera videos[J]. IEEE Transactions on Image Processing, 2020, 29: 1329-1343.

[39] LIU P, TAO Y, ZHAO W, et al. Abnormal crowd motion detection using double sparse representation[J]. Neurocomputing, 2017, 269: 3-12.

[40] 胡正平, 张乐, 尹艳华. 时空深度特征AP聚类的稀疏表示视频异常检测算法[J]. 信号处理, 2019, 35(3): 386-395.
HU Zhengping, ZHANG Le, YIN Yanhua. Video anomaly detection by AP clustering sparse representation based on spatial-temporal deep feature model[J]. Journal of Signal Processing, 2019, 35(3): 386-395.

[41] 袁静, 章毓晋. 融合梯度差信息的稀疏去噪自编码网络在异常行为检测中的应用[J]. 自动化学报, 2017, 43(4): 604-610.
YUAN Jing, ZHANG Yujin. Application of sparse denoising auto encoder network with gradient difference information for abnormal action detection[J]. Acta Automatica Sinica, 2017, 43(4): 604-610.

[42] SUN J, SHAO J, HE C. Abnormal event detection for video surveillance using deep one-class learning[J]. Multimedia Tools and Applications, 2019, 78(3): 3633-3647.

[43] XU K, SUN T, JIANG X. Video anomaly detection and localization based on an adaptive intra-frame classification network[J]. IEEE Transactions on Multimedia, 2020, 22(2): 394-406.

[44] ZHOU S, SHEN W, ZENG D, et al. Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes[J]. Signal Processing: Image Communication, 2016, 47: 358-368.

[45] CHU W, XUE H, YAO C, et al. Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos[J]. IEEE Transactions on Multimedia, 2019, 21(1): 246-255.

[46] NAWARATNE R, ALAHAKOON D, DE SILVA D, et al. Spatiotemporal anomaly detection using deep learning for real-time video surveillance[J]. IEEE Transactions on Industrial Informatics, 2020, 16(1): 393-402.

[47] KALTSA V, BRIASSOULI A, KOMPATSIARIS I, et al. Multiple hierarchical dirichlet processes for anomaly detection in traffic[J]. Computer Vision and Image Understanding, 2018, 169: 28-39.

[48] KALTSA V, BRIASSOULI A, KOMPATSIARIS I, et al. Swarm intelligence for detecting interesting events in crowded environments[J]. IEEE Transactions on Image Processing, 2015, 24(7): 2153-2166.

[49] DEL GIORNO A, BAGNELL J A, HEBERT M. A discriminative framework for anomaly detection in large videos[C]//Proceedings of the 14th European Conference on Computer Vision. Cham: Springer, 2016: 334-349.

[50] 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.

[51] LEE S, KIM H G, RO Y M. Stan: Spatio-temporal adversarial networks for abnormal event detection[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE Computer Society, 2018: 1323-1327.

[52] SABOKROU M, POURREZA M, FAYYAZ M, et al. AVID: Adversarial visual irregularity detection[C]//Proceedings of Asian Conference on Computer Vision. [S.l.]: Springer, 2018: 488-505.

[53] ZHAO Y, DENG B, SHEN C, et al. Spatio-temporal autoencoder for video anomaly detection[C]//Proceedings of the ACM on Multimedia Conference. Mountain View, C A, USA: ACM, 2017: 1933-1941.

[54] LUO W, LIU W, GAO S. Remembering history with convolutional LSTM for anomaly detection[C]//Proceedings of IEEE International Conference on Multimedia and Expo. [S.l.]: IEEE Computer Society, 2017: 439-444.

[55] YE M, PENG X, GAN W, et al. AnoPCN: Video anomaly detection via deep predictive coding network[C]//Proceedings of the 27th ACM International Conference on Multimedia. [S.l.]: ACM, 2019: 1805-1813.

(编辑: 刘彦东)