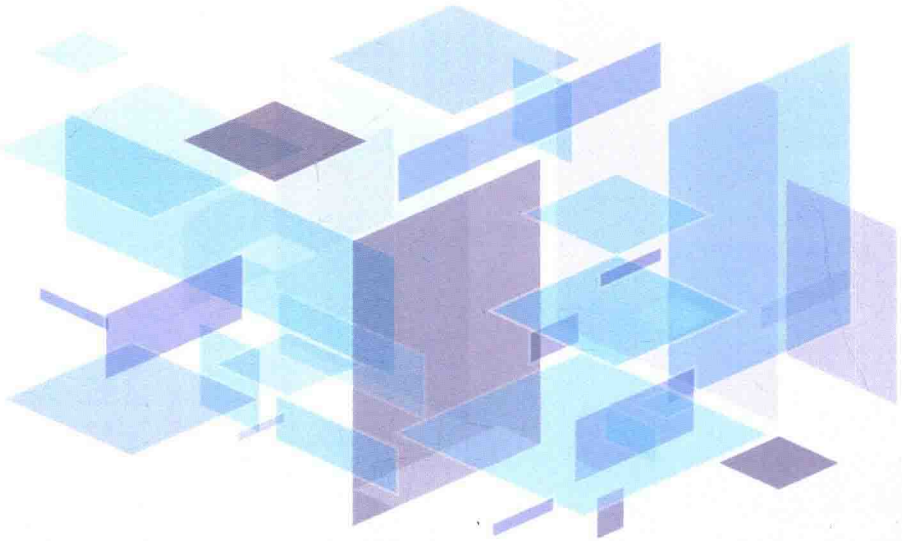
视频图像处理研究
——基于监控场景下的视觉算法

叶人珍◎著
冯亚闯◎审

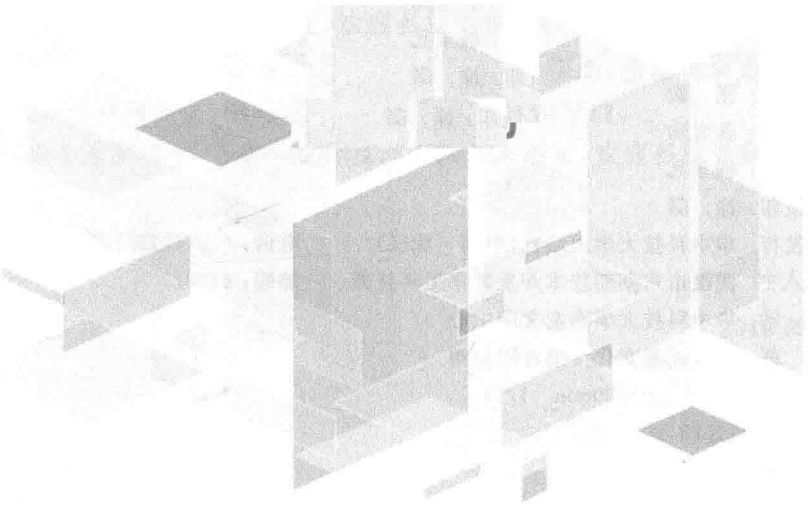


华中科技大学出版社
http://www.hustp.com

国家自然科学基金项目
“基于自上而下学习的视频异常检测（61772223）”研究成果

视频图像处理研究
——基于监控场景下的视觉算法

叶人珍◎著
冯亚闯◎审



華中科技大學出版社
http://www.hustp.com
中国·武汉

图书在版编目(CIP)数据

视频图像处理研究:基于监控场景下的视觉算法/叶人珍著. —武汉:华中科技大学出版社,2018.9

ISBN 978-7-5680-4310-6

Ⅰ. ①视… Ⅱ. ①叶… Ⅲ. ①视频信号-图象处理 Ⅳ. ①TN941.1

中国版本图书馆 CIP 数据核字(2018)第 211433 号

视频图像处理研究——基于监控场景下的视觉算法 叶人珍 著

Shipin Tuxiang Chuli Yanjiu——Jiyu Jiankong Changjing xia de Shijue Suanfa

策划编辑:张 毅

责任编辑:狄宝珠

封面设计:杨玉凡

责任监印:徐 露

出版发行:华中科技大学出版社(中国·武汉) 电话:(027)81321913

武汉市东湖新技术开发区华工科技园 邮编:430223

录 排:华中科技大学惠友文印中心

印 刷:北京虎彩文化传播有限公司

开 本:710mm×1000mm 1/16

印 张:11.25

字 数:220 千字

版 次:2018 年 9 月第 1 版第 1 次印刷

定 价:68.00 元

前　言

本书从原理、方法等角度详细介绍了作者在视频图像处理方面的一些研究，涉及图像去模糊、视频异常事件检测和视觉检索三个领域，为了保证论述的完整性，在视频异常事件检测领域中，加入了冯亚闯博士的部分研究——基于结构字典学习的异常事件检测、基于假设检验的异常事件检测与基于深度表达的异常事件检测。

本书一共七章：第一章为绪论，总体介绍了目前视频图像处理的有关方法；第二章基于潜在语义概念模型的图像盲去卷积方法，提出了一种基于潜在语义概念约束的模型；第三章基于协同表达的异常事件检测方法，提出了基于多种特征描述符的鲁棒协同表达；第四章基于结构字典学习的异常事件检测方法，提出了一种视频数据的时空结构建模方法；第五章基于假设检验的异常事件检测方法，提出了一种测试数据中异常模式的表达模型；第六章基于深度表达的异常事件检测方法，提出了一种视频时空特性的深度特征学习方法；第七章基于稀疏相似保持嵌入的紧致结构哈希算法，提出了一种基于结构的方法(SSBH)。

本书适合从事图像处理方面工作或研究的学生或者工作者作为参考书。

感谢冯亚闯博士在编著此书过程中的大力支持与帮助，感谢李学龙研究员、袁媛研究员、卢孝强研究员、郑向涛博士等的鼎力支持与帮助。

由于知识与水平欠缺，书中一定存在许多缺点与不足，欢迎各位专家、学者批评指正！

叶人珍

2018 年 8 月

目　录

第1章 绪 论

1.1 引 言

随着国民经济的发展，人类对社会安全提出了更高的要求。视频监控以其广泛的覆盖范围和实时记录的特点，受到广泛的关注和研究，已经成为公共社会安全的重要组成部分。目前，我国公共场所中安装的监控摄像头越来越多，覆盖范围越来越广，实现了公共安全场所摄像头的广泛覆盖。这些数量众多的摄像头通过互联网连接形成了一个覆盖各个角落的视频监控网络。因此，视频监控已经成为继数字电视、视频会议之后的又一个重大视频应用，而且日益成为体量最大的一个视频系统。

随着视频监控技术的日益成熟和监控设备的普及，视频监控应用日益广泛，监控视频数据量呈现出爆炸性的增长，已经成为大数据时代的重要数据对象。根据国际数据公司(IDC)的研究报告，2012 年全球各种数据的总量为 2.54 ZB，到 2020 年将上升到 40 ZB，IDC 称之为“数字宇宙(digital universe)”。众多的摄像头，庞大的监控网络，瞬间就会产生海量视频数据。如何从这些海量数据中高效地提取出有用的信息，就成为智能视频监控技术要解决的问题。由于视频数据的非结构化特性，海量监控下的视频数据处理和分析相对困难，已经成为制约智能视频监控能力提升的瓶颈。面对大量摄像头采集的海量监控视频和高昂的存储成本，如何根据监控场景的内容和特性对监控区域进行有效的智能化理解，已成为当前多媒体领域面临的重大挑战。随着智能视频监控的不断发展以及人们对公共安全的日益重视，这一趋势将更加突出。

尽管研究者在视频图像的学习理解上取得了很大进展，并且提出了很多有效的处理算法，但视频图像的学习理解仍存在以下挑战和问题。

(1) 对监控图像来说，摄像头在拍摄过程中不可避免会受到干扰，这使得获得的图像质量均会存在一定程度的干扰。受环境、线路、镜头、摄像机等影响，提取的图像、视频出现退化或关键部位模糊不清，从而对图像的识别、取证等造成困难，使系统无法发挥应有的作用。因此，开展模糊图像处理技术的研究和应用对于安防领域具有重要意义。

(2) 在视频监控中，由于对安全需求不断增加，自动检测和定位异常行为已经成为计算机视觉和模式识别的热门研究领域。异常事件检测的目的是检测个体或拥挤人群的异常行为。近年来，在异常事件检测方面，各国研究人员已经做了

大量工作，但由于在拥挤场景中异常检测较为困难，建立鲁棒的方法仍然是一个具有挑战性的问题。例如，对于人群来说，列出其所有的异常行为是不切实际的。

(3) 近几年，由于文件、图像和视频等网络数据的快速发展，大规模图像搜索引起了广泛的关注。例如，照片共享网站 Flickr 已经聚合了超过 60 亿张照片。YouTube 是一个受欢迎的视频分享网站，每分钟可以收到超过 48 小时的上传视频。从这样大量的图像数据库中检索相关信息非常重要。最近邻搜索是信息检索数据库和计算机科学的热门话题，已广泛应用于大规模视觉问题中，包括图像检索、物体识别以及其他的计算机视觉领域。图像检索的任务主要是在大型数据库中准确找到与查询图像相近的图片。检索相近图片的直接方法是搜索给定的数据库，并根据与查询对象的相似性进行排序。图像搜索的任务旨在获取查询图像，并在大型数据库中准确找到其最近的邻居。然而，当数据库项目较大时，穷举搜索的复杂性被放大并变得非常昂贵。此外，由于原始数据的存储具有高维度，搜索性能将下降。因此，有必要考虑使用近似最近邻(ANN)技术来实现大规模搜索。

1.2 图像去模糊

随着信息技术的发展和多媒体设备的普及，图像成为日益重要的信息媒介。据统计，在人类自身获取的信息中，通过人的视觉系统获得的占 75%。无论是在日常生活中，我们用智能手机拍摄照片，或医生为病人做超声波检查，还是在科学研究中，卫星对地面的遥感探测，或太空望远镜对其他天体的观测，都离不开图像的作用。为了更容易地从图像中得到有用的信息，为日常生活或其他科学研究提供便利，获取清晰的图像是一件十分重要的事情。然而，在复杂多变的环境下，图像在采集、处理、压缩、存储、传输或显示时很容易受到许多不利因素的影响。这些不利的因素包括噪声、模糊、低分辨率、缺损、遮挡和反光等，它们使得图像质量严重下降，对正常的生产生活或科学研究产生巨大的阻碍。为了避免这些不利因素的影响，提高图像的质量，就需要借助于图像恢复。其中，图像去模糊是图像恢复的一个重要方面。

在获取图像时，场景与相机传感器的相对运动、失焦或者大气湍流等都会导致图像模糊。图像模糊的数学模型是

$$Y = k \otimes X + N \tag{1-1}$$

式中：k ——模糊核；

X——清晰原图；

$\otimes$——卷积操作符；

N ——可叠加噪声。

图像去模糊是计算机视觉领域一个经典的问题。依据图像去模糊的数学模型[式(1-1)],按照模糊核 k 是否已知的情况,可将图像去模糊算法分为两大类:对于模糊核已知的情况,称为非盲去模糊问题;对于模糊核未知的情况,称为盲去模糊问题。本书主要研究盲去模糊问题,尤其是单张图像的盲去模糊问题。我们需要在仅仅已知单张模糊图像 Y 的前提条件下,求出模糊核 k 和清晰原图 X,如图 1-1所示。在这个问题中已知变量少于未知变量,所以图像去模糊是一个十分病态的问题。

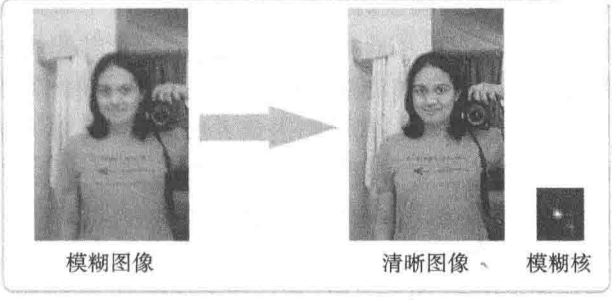


图 1-1　图像去模糊过程的简单示意图

为了降低问题的病态性,许多算法在求解的过程中对模糊核 k 或清晰原图 X 加入了先验知识。依据这些先验知识的不同类型,可以将单张图像的盲去模糊算法大致分成以下几类:基于正则化的算法;基于贝叶斯先验的算法;基于稀疏先验的算法。当然除了以上列举的这些传统的方法,随着深度学习的兴起,一部分基于深度学习的去模糊方法也出现了。

1. 基于正则化的算法

基于正则化的图像去模糊算法常对模糊核 k 或清晰原图 X 使用不同的正则项,从而使输出的结果更加稳定。这些正则项通常是 l_1 范数或者是 l_2 范数,更常用的是二次 Tikhonov 正则项、全变分正则项以及它的一些改进或变体形式。Mariana 等人提出了一种边缘探测器和更易产生锐利边缘的正则项。Krishnan 等人引入了一种全新的正则项,即 l_1/l_2 范数,这一范数可以以最小的损失值得到真实的清晰原图。基于正则化的图像去模糊算法可以使结果稳定,并且对于抑制噪声的影响有着很好的效果。但是,这些方法易产生过于平滑的结果,丢失了很多图像的细节,所以不适合实际的图像去模糊问题。

2. 基于贝叶斯先验的算法

贝叶斯模型(Bayesian model)在图像去模糊中有着广泛的应用。基于贝叶斯框架的去模糊方法通过计算清晰原图的后验概率,利用这种后验概率分布估计清晰原图和模糊核。在图像和模糊核上添加的概率模型,对估计图像的求解过程添加了限制并起到正则项的作用。为了消除贝叶斯模型的复杂性并充分利用全贝

叶斯(full Bayesian)模型的优势,经过变分近似的变分贝叶斯(variational Bayesian,VB)模型被用在图像去模糊中。相比于传统的基于贝叶斯模型的方法,变分贝叶斯模型需要一个事先定义好的下降的噪声序列值才能获得更好的效果,难以对问题进行全面的分析。在大多数情况下,基于贝叶斯的方法局限于最大后验(maximum A posteriori,MAP)估计的结果,所以在许多去模糊问题中并不适用。

3. 基于稀疏先验的算法

基于稀疏表达的去模糊算法将图像的稀疏性作为先验来恢复得到清晰的原图。Cho 等人提出了基于回归模型的图像去模糊算法,这一回归模型可以根据图像内容自适应地学习稀疏先验。Cai 等人将盲去模糊问题转化为一个新的联合优化问题,从而最大化清晰原图和模糊核的稀疏性。Aharon 和 Elad 在图像去噪中引入了过完备字典(over-complete)的稀疏表达方法。Zhang 等人提出了一种基于稀疏表达先验的联合处理图像盲去模糊问题和识别问题的方法。Jia 等人提出了基于图像稀疏表达先验和图像块稀疏梯度先验的图像去模糊算法。然而,基于稀疏先验的盲去模糊算法通常采用 l_1 范数稀疏惩罚项,假设稀疏系数是独立同分布的,这没有考虑到不同系数间的关联性。传统的 l_1 范数稀疏模型已经被证明是不稳定的,尤其在图像质量发生退化的时候,这限制了基于稀疏表达的盲去模糊算法的效果。

4. 基于深度学习的算法

近几年来,随着深度学习的崛起,一些基于深度学习框架的图像去模糊算法也出现了,其中有一部分也取得了不错的效果。Xu 等人提出了由两个子模块构成的深度卷积神经网络图像非盲去模糊算法。Sun 等人将图像的动态模糊问题转化为一个基于图像块的模糊核分类问题,利用深度卷积神经网络求解出模糊核后,再用已有的非盲去模糊算法得到最终的去模糊结果。Nah 等人提出了一种基于多尺度残差网络结构的动态去模糊方法。基于深度学习的图像去模糊算法从大量数据中学习模糊图像的先验,避免了人为先验假设带来的错误,但是基于深度学习的去模糊算法需要大量的数据和训练资源、时间,算法实现成本太高。

1.3 异常事件检测

近年来,全球范围开展了视频监控项目的建设,特别是以“平安城市”“平安校园”为代表的重点项目,带来了视频监控市场的蓬勃发展。根据 IHS 的科技报告[①],2014 年全球共安装了 245,000,000 个监控摄像头,其中亚洲占了 65%。到了 2016 年,亚洲所占的比例增长到了 68%,而这其中主要的监控摄像头分布在中

① http://www.securitynewsdesk.com/how-many-cctv-cameras-are-there-globally

国。并且中国还宣布，到2020年将在全国所有公共场所和街道安装监控摄像头。如果能够有效地利用这些监控摄像头拍摄到的海量视频数据，那么对于犯罪的预防和侦破将会起到非常显著的辅助作用。

目前的视频监控技术主要依靠人力。图1-2是目前视频监控大厅的一个实例。系统将各个地点监控摄像头采集到的视频信息通过网络传输技术汇集到一个电视墙上，由相关的工作人员对这些视频画面进行不间断的观察。当发现可疑事件时，做出及时的处理。另外，工作人员不可能捕捉到所有的可疑情况，往往会有所遗漏。针对这些情况，监控视频通常会作为事发后的线索和证据。这种基于人力的视频监控技术存在以下两个问题：①人力资源的浪费。一方面，统计信息显示，人工不能有效监控多个电视屏幕，操作人员盯着电视监控屏幕超过10分钟后将漏掉90%的视频信息。另一方面，视频中的可疑事件是非常少的，99.9%是正常的，这就造成了大量人力资源的浪费。而在事件发生后，通过人工来检索事件也会变得困难而低效，大部分事件稍一疏忽就会遗漏。“伦敦七七爆炸案”中，有100多位安保人员花费了70多个工时才在大量磁带中找到了需要的信息。尽管在今天视频捕捉设备的价格越来越便宜，但监视和分析这些视频数据所需的人力却是十分昂贵的。②存储空间的浪费。以1080P的监控摄像头为例，在4 Mbps的码率下，一个摄像头一天所拍摄的视频数据为4 Mbps×60秒/分钟×60分钟/小时×24小时/天=345,600 Mb≈42.19 GB。根据公安部要求，录像数据在系统中需要保存30天以上，那么一个摄像头就需要配备42.19 GB/d×30 d≈1.24 TB的存储空间。由于监控视频中几乎都是正常的事件，可疑事件所占的比重非常少，也就是说，这些存储空间大部分是没有用的。通过以上两个方面的分析，利用计算机自动地从大量的监控视频中检测出可疑行为是目前安防领域的一个重要方向。

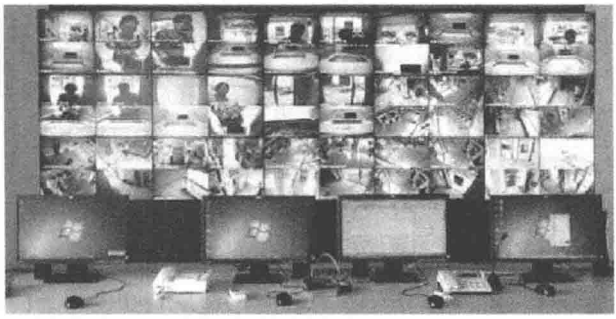


图1-2 监控大厅的一个实例①

① 图片来源于 http://news.c-ps.net/article/201409/212838.html

视频中的异常事件检测是指利用机器学习和计算机视觉技术，对视频序列中包含的事件自动分析，并统计视频事件中的规律。当有异常（可疑）情况发生时，系统能准确并及时地检测和报警，以及提供相关的分析结果。该技术可以协助监控人员有效地发现和处理危险事件，并在最大程度上降低漏检和误检的概率。异常事件检测在帮助监控工作人员更高效率工作的同时，还可以显著地缩小需要保存视频数据的体积，降低监控系统在人力和存储上的需求。

与行为识别不同，视频中的异常事件检测没有带标注的异常样本用来学习异常事件的规律。这主要是因为异常视频事件的以下四个特性：①稀有性。相比于正常视频事件，异常事件的数量可谓是沧海一粟。有时一个月的监控视频都不存在一个异常事件，这就导致了大量的人力和存储上的资源浪费。但它们又不是不发生，这使得传统监控系统不得不浪费。②定义不明确。异常视频事件的定义是与场景相关的，同样的行为在不同的场景下可能具有完全不同的异常属性。比如，奔跑这一行为，在人群中奔跑是异常的，然而在比赛中奔跑却是正常的。③不可预知性。在一个场景中，异常事件的种类多种多样，不可能列举所有的异常情况。比如，在人群中可能会出现逃跑、摔跤、丢包等行为。这一特性使得研究者们很难采集所有的异常事件类型，并研究其内在特性，也就给异常事件数据库的构建带来了困难。④与正常视频事件的差异性。与正常事件相比，异常视频事件总存在或多或少的差异，正是这些差异促进了异常事件检测领域的蓬勃发展。

由于异常视频事件的稀有性和不可预知性，很难一一列举特定视频场景中所有的异常事件，因此研究者们难以构建有效的异常视频事件数据库来学习异常事件的模式规律。同时，异常事件具有定义不明确性，即不同场景下异常事件的定义不一样，在某一场景下的异常事件在另一个场景下可能是正常的。图 1-3 所示是目前通用的异常视频事件检测流程图。在训练过程中，首先对视频事件进行特征提取，要求提取的事件特征能够有效地区分正常和异常视频事件。然后，基于提取的视频事件表达，利用机器学习技术学习出正常的视频事件模型。在测试过程中，采用与训练过程相同的的方式提取视频事件特征。最后，计算测试视频事件与训练得到的正常事件模型间的匹配程度。如果匹配度较高，则认为是正常事

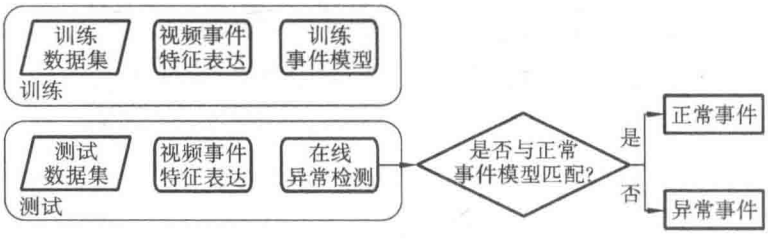


图 1-3 异常视频事件检测流程图

件，否则判断为异常事件，同时发出警报。

随着智能设备的发展，视频数据在呈爆炸式地增长。有数据显示，在知名网站 YouTube 上，每分钟上传的视频时长超过 100 小时，这需要一个网民 24 小时不间断观看 4 天。如此海量的视频数据，如果直接对原始视频进行操作会造成不可估量的计算代价。同时，由于视频数据的时间相关特性，原始视频数据中存在大量的信息冗余，如果直接处理会带来不必要的计算量。此外，由于视频数据受拍摄条件的影响，不同光照、视角等条件下得到的视频数据具有明显的差异，此时直接对视频数据进行处理将降低算法的性能。因此，如何从原始视频数据中提取出紧致的、能够有效区分异常和正常视频事件的特征表达是决定异常事件检测精度的关键步骤。从图 1-3 中可以看出，接下来需要对提取的视频事件特征进行建模。视频事件建模的目的是从大量的视频事件特征中挖掘出其统计或者时间变化上的规律，而对正常视频事件的准确建模是决定异常事件检测精度的另一关键步骤。综上所述，决定视频中异常事件检测算法精度的两大关键步骤分别是视频事件特征表达与视频事件模型构建。

1.3.1 视频事件特征表达

根据特征提取方式的不同，视频事件特征表达的方法可以分为目标层的特征表达和像素层的特征表达两类。

基于目标的特征表达具有高层的语义信息，通常采用目标提取及跟踪技术来获取目标的信息，包括目标的大小、形状、轮廓、纹理、速度、方向等。在这类特征表达方法中，最常用的是目标的轨迹信息。比如，文献[61,62]直接提取视频帧中运动目标所在的位置信息作为目标的特征表示。该类特征表示的维度与视频序列的长度一样，由于视频事件的长度不一致现象，导致了视频事件特征表达的维度不一致。为了解决该问题，文献[63,64,65,66]采用线性差值和重采样等技术生成维度固定的特征向量，提升了轨迹特征的可适用性。

在视频空间分辨率低、存在遮挡等情况下，单个运动目标的提取非常困难，此时通常采用基于特征点或者粒子追踪的方式提取轨迹特征。例如，Wu 等人采用拉格朗日(Lagrangian)粒子的轨迹对视频场景进行建模，而 Cui 等人通过追踪感兴趣点来计算交互作用能势(interaction energy potential)，并以此模拟人群间的相互作用关系。Tran 等人提出时空路径优化算法，从三维时空块中找到监控目标最优的时空轨迹。此外，利用轨迹信息衍生的特征也被广泛使用。比如，Zhang 等人利用追踪算法得到运动目标在每一视频帧中的位置，然后计算目标的瞬时运动信息、连续 T 帧的平均速度信息、二阶能量特征等。尽管基于目标轨迹的特征表达方法具有明确的语义信息，而且在异常视频事件检测中取得了较高的检测精

度，然而对运动目标以及特征点追踪的时间复杂度非常高。同时，随着场景中目标数量的逐渐增加，运动目标间的遮挡越来越严重，大多数追踪算法都将失效，提取精确的目标轨迹已不再可能。此时，基于目标的视频事件特征表达方法将不再可靠。

在拥挤的场景中，由于目标间不可避免的遮挡现象，造成了追踪算法的不精确，也使得目标层视频事件特征表达在异常事件检测中不再适用。为此，研究者们提出采用像素层的特征表达方法来对视频事件进行描述，主要包括底层的视觉特征以及相应的统计信息。根据引起异常视频事件的原因，异常事件大致可以分为两种类型：目标的异常行为模式和异常目标的侵入。而根据两种异常事件所采用底层视觉特征的不同，像素层的视频事件特征表达方法可以分为以下两类：运动特征表达和表观特征表达。

1. 运动特征表达

在犯罪视频中，罪犯通常会进行一些伪装，从表观上很难与正常的目标区分，此时运动特征就成了异常事件检测中最重要且最具有判别性的信息。在目前常用的运动特征表达中，光流是用于描述像素级瞬时运动最重要的工具之一。光流的概念由美国实验心理学家 Gibson 于 1950 年正式提出，目前已经成为运动图像分析的重要方法。它是图像中亮度模式的表观运动，由观察者和目标的相对运动引起，而图像中的光流场是所有像素点位置上瞬时速度的集合。传统的光流计算方法均基于亮度一致性模型（brightness constancy model，BCM），即通过相邻视频帧之间亮度的一致性对应关系来确定运动的起点和终点。在光流的计算方法中，最具有代表性的两类方法为 Horn 和 Schunk 提出的全局法（H-S 方法）与 Lucas 和 Kanade 提出的局部方法（L-K 方法）。H-S 方法在光流亮度一致性约束的基础上对整个光流场加以平滑约束，使得光流场不仅满足亮度一致性假设，同时满足光流场的全局平滑性假设。而 L-K 方法假定光流在局部区域上是恒定的，使得该算法对噪声比较鲁棒。H-S 方法和 L-K 方法是光流计算中比较经典的方法，后续的光流计算方法多是在此基础上发展起来的。

基于计算得到的光流图，Reddy 等人采用时域平均的方法去除其中的噪声，然后利用视频块中前景的平均光流作为视频事件的运动特征表达。文献[81]和文献[82]在去除背景的视频中使用光流，提高了算法的抗噪性能。文献[83]中，作者发现由于人行走时四肢做周期性运动，四肢周围的光流幅度会发生周期性的变化，而汽车或自行车在运动的过程中光流幅值是均匀的，因此文中利用灰度共生矩阵（grey level co-occurrence matrix，GLCM）对光流的纹理特征进行描述。由于光流只利用了相邻两帧图像的信息，Mehran 等人提出在视频帧上放置粒子，计算粒子在连续多帧图像上的移动路径，构成光流的纹线（streakline），用来表示视频事件长时间的运动特征表达。

除了直接利用光流信息之外，基于光流的统计特征也被广泛采用，其中最为常用的是光流直方图(histogram of optical flow，HOF)特征。例如，Chaudhry 等人利用光流直方图描述视频事件的运动特征，直方图中的每个柱代表该方向上光流的统计。文献[86]提出一种选择性光流直方图(selective histogram of optical flow，SHOF)特征，在光流方向直方图特征的基础上考虑在不同场景中自适应地判别运动方向的异常或者运动幅值的异常。Cong 等人提出一种多尺度光流直方图(multi-scale histogram of optical flow，MHOF)特征，根据光流幅值的大小对光流进行分类，在每个尺度上统计光流的方向直方图信息。文献[87]构建了一种运动滤波方法，将运动不明显的疑似运动目标平滑掉，得到多尺度下的滤波光流直方图。文献[63]在利用可见光图像计算光流场后，分别对位置、速度和方向进行量化，使得每个像素都具有位置、速度和方向特征，最后利用词袋(bag of words，BoW)模型计算视频块的特征表达。此外，基于目标在前后帧的光照不变性假设，文献[88]使用多尺度的时间梯度来作为视频事件的运动特征表达。

2. 表观特征表达

针对视频场景中异常目标侵入的事件，表观特征表达是一种非常有效的事件描述方法。比如，在拥挤的人群中出现一辆手推车，它具有行人的运动速度和方向，却在视觉属性上具有明显不同的表观特性。目前，很多异常检测算法都采用表观特征对视频事件进行描述。比如，文献[2]中，Reddy 等人利用二维 Gabor 特征和目标尺寸对目标的表观进行描述，方法中的目标均是针对视频帧中的块，并不涉及目标的提取和追踪。相比于 Gabor 特征，空间梯度特征显得更加简单、快速。为了降低特征的维度和计算复杂度，对空间特征的描述通常计算其统计特性。比如 Zhao 等人采用梯度直方图(histogram of gradient，HoG)计算每一个空间-时间感兴趣点(spatio-temporal interest point，STIP)的描述。梯度直方图是用来描述目标表观特性的统计特征，其中直方图的每个柱代表方向，而各柱的值是视频块内梯度幅值的累加。Cong 等人首先利用 Sobel 模板对视频帧滤波，提取图像中的边缘，然后统计得到边缘方向直方图(edge orientation histogram，EOH)对视频块中的外观特征进行描述。

此外，一些特征同时考虑了视频事件的运动和表观描述。比如，空时梯度(3D 梯度)特征表达中，时间梯度描述了运动信息，而空间梯度描述了表观信息。由于既能够描述运动，又能够描述外观信息，空时梯度成了一种重要的事件描述工具。Kratz 等人首先对视频帧图像进行高斯滤波去除噪声，然后计算视频块内各像素的空时梯度共同表示立方体内的事件模式特征。Lu 等人在 3D 梯度的基础上，利用主成分分析(principal component analysis，PCA)算法进行降维，得到的特征表达既有效又紧致，取得了目前最快的异常检测速度。Hu 等人提出一种局部最近邻距离(local nearest neighbor distance，LNND)描述子，该方法利用像素的空时

梯度计算局部运动模式(local motion pattern,LMP)。这种基于空时梯度的描述子能够很好地刻画动态场景的运动和表观特征。然后,计算视频块与其 K 个空间最近邻和 N 个时间最近邻的测地距离(earth mover's distance,EMD),并将该距离向量作为视频块的特征表达。文献[101,102]对视频进行密集采样,得到若干相同尺度的视频时空块(spatio-temporal video volumes,STV),然后利用邻接图模型从全局和局部对 STV 之间的关系进行建模,将 STV 在底层表示成直方图的形式进行编码,并结合词袋模型进行高层的事件模式构建。

1.3.2 视频事件模型构建

视频中的异常事件检测是典型的一分类问题,即训练数据集中只有正样本(正常事件),没有负样本(异常事件)。这是由于正常的视频事件很容易获取,而异常视频事件的获取代价太大,而且发生概率太低。目前通常采用的方法是利用训练数据中的视频事件对正常事件进行建模,测试视频中与正常事件模型不匹配的样本即被判定为异常事件。根据所采用视频事件模型的不同,异常事件检测算法大致可分为:基于概率的方法基于距离的方法、基于重构的方法和基于域的方法。下面将分类进行介绍。

基于概率的异常事件检测模型假设视频事件特征空间内的概率代表事件的正常程度。概率越小,说明属于正常事件的可能性越低,而属于异常事件的概率越大。该模型首先从训练视频数据中估计视频事件的概率密度函数,然后设定一个阈值,使得正常视频事件的概率都大于该阈值。在测试的过程中,检验测试样本在该模型下的概率值是否大于该阈值。根据模型中是否限制概率分布类型,基于概率的方法可以分为两类:参数方法和非参数方法。

参数方法对数据的分布类型进行假设,如果该假设与数据分布差别较大,该模型会造成较大的误差。目前最常用的分布类型为高斯分布,其参数可以通过最大似然估计(maximum likelihood estimate,MLE)算法求解。例如,文献[96]认为三维梯度服从高斯分布,并使用散度作为距离度量准则,估计其参数。Thida 等人利用多个高斯模型对正常视频事件进行建模。而文献[83,105,106]采用更加复杂的高斯混合模型(Gaussian mixture model,GMM)模拟视频事件的分布。考虑到视频事件的时间特性,文献[107,108]利用隐马尔科夫模型(hidden Markov model,HMM)对视频事件进行建模。隐马尔科夫模型对于时间序列的建模以及变量间隐藏关系的捕捉非常有效。在 HMM 模型中存在两种变量:隐藏变量和观测变量。两者间的关系如下:①隐藏变量只与前一时刻的隐藏变量有关,随时间做有规律的线性变化;②观测变量只与当前时刻的隐藏变量相关。因此,在给定当前时刻的隐藏变量时,后续的观测变量彼此独立。Kratz 等人在每个视频空间

位置构建一个隐马尔科夫模型。为了获得时间上的统计信息，他们使用基于分布的 HMM 在时间方向上对数据进行建模，为了获得在空间上的关系信息，他们提出了耦合 HMM，并用此模型对正常事件进行建模以识别异常事件。此外，Kim 等人构建混合主成分分析(mixture of probabilistic principal component analysis，MPPCA)模型对正常的光流运动特征进行投影，并利用马尔科夫随机场(Markov random field，MRF)中空间邻域的状态转移概率来表征空间层面区域之间的链接性。

非参数方法需要很少的假设，因此可以构建比较灵活的模型。该模型会不断扩大规模以适应数据的复杂特性，但这需要大量的样本来拟合出更可靠的模型。在该类方法中，最为经典的方法是核密度估计(kernel density estimation，KDE)。其中，视频事件特征空间中每一个点的概率密度都依赖于其与邻域点之间的关系。核密度估计方法将核放置于每个数据点处，然后对来自每个局部邻域的贡献进行累加。这种核密度估计方法通常也称作 Parzen 窗估计方法。Saleemi 等人利用核密度估计技术对视频中目标的轨迹进行建模，学习目标正常的运动模式。Ramezani 等人利用柯西核代替通常使用的高斯核，并提出一种基于递归的核密度估计方法，提高了计算效率。该算法能够逐步更新背景模型并通过一种无监督在线学习的方式更新异常判断的准则。

基于概率的异常检测方法具有严密的数学理论基础，一旦获得了准确的概率密度函数，就可以实现快速有效的异常事件检测。然而，概率密度的准确估计需要大量的数据样本，如果训练数据集中样本量较少，那么异常检测的精度将大大降低，特别是在视频事件特征表达维度较高的情况下。

基于距离的异常事件检测模型假设正常视频事件紧密聚集，而异常样本往往与其最近邻距离较远。该类模型主要包括最近邻和聚类两种方法。

最近邻方法假设正常事件在正常训练集中具有较近的邻居，而异常事件则远离这些正常样本。如果一个样本点远离其邻域，则认为该点是一个异常点。其中的距离度量方式通常采用欧式(Euclidean)距离和马氏(Mahalanobis)距离等。例如，Cong 等人将异常事件检测问题看作样本点的匹配问题，从而克服了基于概率模型在样本量很少时不稳定的缺点。在文献[119，120]中，研究者们利用 Hausdorff 距离在线衡量轨迹间的差异程度，然后利用适性预测方法度量轨迹最小邻域内的相似度，从而进行异常轨迹的判别。在测试过程中，判断一个视频事件是否正常时，搜索其在训练集中最匹配的样本。如果两者间的距离较小，则认为是正常的，否则为异常。在实际应用中，当训练数据集规模较大或者事件特征表达维度较高时，这种穷举搜索的方式将十分耗时，为此他们采用随机投影技术来加速搜索。

如果从聚类的角度考虑异常事件检测，训练阶段将相似的视频事件描述聚合