UAnalysis
Understanding

Video Events Analysis and Understanding

视频事件的分析与理解

裴明涛 赵 猛 著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

Video Events Analysis and Understanding

视频事件的分析与理解

裴明涛 赵 猛 著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

内 容 简 介

视频事件的分析与理解是计算机视觉领域的重要研究内容之一，具有重要的理论研究意义和实际应用价值。本书首先介绍了视频事件分析与理解所涉及的目标检测、目标跟踪以及事件识别的研究现状，分析了视频事件分析与理解中的关键问题，然后重点介绍了作者研究团队在视频事件分析与理解领域的研究工作和成果。

本书可供计算机、自动化、模式识别等领域的科研人员参考，也可作为高等院校计算机、自动化、电子信息等专业的教学参考书。

图书在版编目（CIP）数据

视频事件的分析与理解/裴明涛，赵猛著. —北京：北京理工大学出版社，2019.3

ISBN 978-7-5682-6819-6

Ⅰ.①视…　Ⅱ.①裴…②赵…　Ⅲ.①视频系统-监视控制-研究　Ⅳ.①TN948.65

中国版本图书馆 CIP 数据核字（2019）第 041430 号

出版发行／北京理工大学出版社有限责任公司
社　　址／北京市海淀区中关村南大街 5 号
邮　　编／100081
电　　话／（010）68914775（总编室）
　　（010）82562903（教材售后服务热线）
　　（010）68948351（其他图书服务热线）
网　　址／http：//www.bitpress.com.cn
经　　销／全国各地新华书店
印　　刷／保定市中画美凯印刷有限公司
开　　本／710 毫米×1000 毫米　1/16
彩　　插／4
印　　张／14　　责任编辑／陈莉华
字　　数／245 千字　　文案编辑／陈莉华
版　　次／2019 年 3 月第 1 版　2019 年 3 月第 1 次印刷　　责任校对／周瑞红
定　　价／62.00 元　　责任印制／李志强

前　言

视频事件的分析与理解由于在智能监控、智能人机交互等领域有着广泛的应用前景，成为计算机视觉领域备受关注的前沿方向之一。目前的视频事件分析与理解方法存在两个问题有待解决：一是事件模型多是人工指定，不能准确反映事件的内在特征；二是现有的事件分析方法多是针对底层的动作或者高层的事件进行分析，没有将动作、事件以及场景理解进行有机的结合。

对于视频事件的分析与理解，有一类方法是直接将视频数据作为输入，提取特征，进而进行视频事件的识别。这样做的一个问题是视频中包含了大量的与事件无关的信息，例如背景中的无关物体以及与事件无关的人的运行等，直接将视频作为输入可能会导致无关信息对事件识别产生干扰，使得事件的分析与理解无法得到满意的效果。因此本书中采取的视频事件分析与理解方法是首先检测视频中感兴趣的目标（包括人和事件涉及的物体），并对这些目标进行跟踪，得到感兴趣的目标在每一帧的位置和大小等信息，根据每一帧中人的姿态以及与感兴趣物体的位置关系来检测原子动作，进而进行视频事件的分析与理解。

本书详细介绍了作者近年来在视频事件分析与理解方面所做的工作，主要包括以下 3 个方面的内容。

（1）场景中的物体检测方法研究。检测出场景中的人以及感兴趣的物体是进行后续视频事件分析与理解的前提和基础，我们采用基于深度通道特征的行人检测方法对场景中的行人进行检测，采用特征共享的联合 Boosting 方法进行场景中其他感兴趣物体的检测。

（2）视频中的目标跟踪方法研究。检测到场景中的行人及感兴趣物体后，需要对它们进行跟踪，得到它们的轨迹，这些信息是后续的视频事件分析的基础。针对行人的特点，我们基于多分量可变部件模型对视频中的人进行跟踪，而采用半监督的基于锚点标签传播的跟踪方法对视频中其他物体进行跟踪。

（3）基于时序与或图的视频事件分析与理解方法。在得到了场景中人及其他物体的轨迹后，我们采用基于人的姿态以及人与物体的位置关系的一元及二元关系来表示原子动作，使用时序与或图来建模事件，表现子事件以及原子动作之间的层次关系和原子动作之间的时序关系，研究事件时序与或图模型的自动学习方法、基于时序与或图模型的事件解析方法以及基于环境上下文信息的事件解析方法。

本书结构如下。

第1章介绍了视频事件分析与理解的研究现状，主要包括目标检测的研究现状、目标跟踪的研究现状以及视频事件分析与理解的研究现状。第2章介绍了视频中的目标检测方法，包括基于深度通道特征的行人检测方法以及基于特征共享和联合 Boosting 方法的物体检测方法。第3章介绍了基于多分量可变部件模型的行人跟踪方法和基于锚点标签传播的物体跟踪方法。第4章介绍了事件的时序与或图模型及其自动学习方法。第5章介绍了基于时序与或图模型的事件分析方法。第6章介绍了基于关键原子动作和上下文信息的事件识别方法。

本书总结了作者和研究组成员在视频事件分析与理解这一研究领域所取得的学术成果，其中包括贺洋、刘钊在行人检测方面的研究成果（第2章），刘钊在行人跟踪方面的研究成果和武玉伟在一般目标跟踪方面的研究成果（第3章），裴明涛在事件与或图模型学习及事件解析方面的研究成果（第4、5章），赵猛、王亚菲在基于关键原子动作和上下文信息的事件识别方面的研究成果（第6章）。

本书中的工作得到了国家自然科学基金（No. 61472038）以及中国博士后科学基金（No. 2018M642680）的资助。

本书是对视频事件分析与理解这一重要问题所涉及的理论和方法的研讨与总结。可以给读者提供一个有益的参考，以普及对于视频事件分析与理解的认知和理解，进而推广其应用。本书对同行研究者，以及对目标检测、跟踪和事件识别相关领域的研究者和爱好者，也具有一定的参考意义。

由于作者水平所限，书中难免存在疏漏和不当之处，敬请读者不吝指教！

裴明涛

2018 年 11 月于北京理工大学

目 录

第 1 章

引　言

1.1　视频事件分析与理解的背景和意义

视频事件的分析与理解，是指让计算机能够像人类一样通过视觉感知外部环境，自动对视频中发生的事件进行分析与理解，知道周围环境中发生了什么事件、事件持续了几个阶段以及每阶段分别发生了哪些行为，从而帮助或辅助人类完成许多重要的任务，如智能视频检索、智能视频监控、高级人机交互、智能环境构建等。

视频事件的分析与理解是计算机视觉和模式识别的重要研究内容，涉及人工智能、计算科学和认知科学等多个学科领域，属于多学科交叉前瞻性研究。此研究具有重要的学术意义。

从人工智能的角度看，了解智能的本质并制造出与人类智能相似的智能机器是人工智能的最终目标，根据底层视频处理的结果获得人类可以理解的语义符合人类智能的一般过程，而视频事件的分析与理解正是解决如何跨越底层视觉信息到高层语义信息之间“语义鸿沟”的问题。

从计算科学的角度看，进行数学模型的构建和定量分析以及利用计算机分析和解决科学问题是其关注的重点。视频事件的分析与理解研究不同层次不同粒度的事件模型构建，通过建立数学模型解析不同层间的信息传递关系，用定量优化算法进行模型的学习与推理，可以丰富计算科学领域的建模理论与数值计算方法，促进计算科学的发展。

从认知科学的角度看，使计算机利用类似人类视觉感知的方式，对多个视频分段进行处理和分析，在很大程度上实现底层数据结构化、语义化的表达，这与人的认知过程相呼应。因此，从事件分析与理解的角度探索人在认知过程

中进行信息分析与处理的机理，将为进一步认知心理学，探索人类的感知和心理活动提供新的思路和方法。同时，对于视频事件分析与理解的研究具有重要的应用价值，在智能视频监控、智能视频检索、高级人机交互等方面都具有广阔的应用前景。

智能视频监控系统可以自动分析摄像机捕捉的视频数据，实时监测并理解目标行为和事件意义，是社会安全的重要保障，在国民经济建设和人民生活中得到了广泛的应用，如银行、机场、停车场等。

智能视频检索是未来多媒体和网络的重要组成部分，面对海量的视频数据，如何快速准确地找到所需的视频信息成为急需解决的问题。对于视频事件分析与理解的研究可以为智能视频检索系统提供有效的方法。

高级人机交互将是未来计算机与人类自然、友好、便捷的交流方式。人机交互的一个重要前提就是要让计算机能“看懂”周围环境中的人与事物，得到类似于人类表达方式的语义，以便能做出正确合理的反应。对于视频事件分析与理解的研究正是要解决计算机的“看懂”问题。

视频事件的分析与理解具有重要的理论价值和广阔的应用前景，因此，受到了越来越多科研人员的关注，国内外许多研究机构都投入了巨大的人力、物力，就此问题展开了研究。

在国外方面，得州大学奥斯汀分校计算机视觉研究中心设立了人体行为识别项目，目标是建立行为识别的整体方法论，重点研究行为的语义分析以及复杂的时空逻辑关系[7,8,157]。马里兰大学的自动控制研究中心计算机视觉实验室就视频监控过程中的目标检测、事件建模、多目标事件识别等内容展开了广泛的研究[130,132,181]。南加州大学视觉实验室的研究范围则包括行人、车辆等移动目标的检测与跟踪，基于分层结构的事件建模，网络视频事件的分类等[36,75,128]。麻省理工学院计算机科学与人工智能实验室的视觉研究组分别对自然场景分析、目标识别以及事件预测等问题展开了研究[42,116,214]。卡耐基梅隆大学机器人研究所对动作识别以及视频事件检测等问题进行了研究[80,89,127]。佛罗里达中央大学视觉实验室在自然环境下的动作识别、复杂视频事件的检测与识别等方面取得了很多的研究成果[9,82,208]。法国国家计算机科学与控制研究所 INRIA 也建立了 PERCEPTION、VISTA 以及 QGAR 等多个项目组对基于时空信息的行为分析、图像分析和识别、图像理解、视频检索等问题展开研究[95]。加州大学洛杉矶分校[21-22]、英国伦敦大学玛丽皇后学院视觉实验室[23]、多伦多大学计算机视觉实验室[24]、瑞典皇家理工学院计算机视觉与感知实验室[25]等诸多研究机构都在这个领域进行了大量的研究工作。

在国内方面，中国科学院模式识别国家重点实验室成立了智能视觉监控研究组，对智能视频监控领域展开了研究，取得了一定的成果[26]，如基于粒子滤波器的目标跟踪算法[107,205]、动作识别[76]、行为识别[58]以及复杂的高层语义事件的表示和识别[223]等。北京理工大学媒体计算与智能系统实验室在动作识别[65,196]、复杂视频事件的表示与识别[151]等方面展开了深入的研究。国内其他高校也正在进行相同或相似方向的研究工作，如清华大学、北京交通大学、山东大学、吉林大学等。

国际上很多主流的学术期刊和学术会议也将视频事件的分析与理解作为热点研究议题，如 IJCV（International Journal of Computer Vision）、T - PAMI（IEEE Transactions on Pattern Analysis and Machine Intelligence）、CVIU（Computer Vision and Image Understanding）、IVC（Image and Vision Computing）、PR（Pattern Recognition）、MVA（Machine Vision and Applications）、T - IP（IEEE Transactions on Image Processing）等国际期刊，以及 ICCV（International Conference on Computer Vision）、CVPR（Computer Vision and Pattern Recognition）、ECCV（European Conference on Computer Vision）、ICPR（International Conference on Pattern Recognition）和 IWVS（IEEE International Workshop on Visual Surveillance）等国际会议。

对视频事件进行自动分析与理解需要对视频中的人以及事件所涉及的感兴趣物体进行检测和跟踪，在得到视频中人及感兴趣物体的跟踪结果后，可以有针对性地提取相应的特征进行事件的分析与理解。因此对视频事件的分析与理解涉及目标检测、目标跟踪以及事件识别 3 个方面的内容，下面我们将分别对这 3 个方面的研究现状进行综述。

1.2 目标检测的研究现状

视觉目标检测（Visual Object Detection），是根据目标的特征，利用最优化方法、机器学习等技术，检测出图像中目标所在的位置和大小。在视频事件的分析与理解中，人是最重要的目标，首先要将视频中的人检测出来，才可以进行后续的事件分析与理解。同时事件中涉及的其他感兴趣物体，如桌子、椅子、电话、水杯等物体也对事件的分析与理解有着重要的作用。下面我们将以行人检测为例来介绍目标检测的研究现状，行人检测的方法可以用于对其他感兴趣的物体进行检测。

行人检测技术是指计算机在一张图像里标示出行人区域的位置、行人区域

所占的大小并给出一定的置信度。行人检测算法主要包含行人表观建模和目标定位两个部分。其中表观建模主要包括行人的视觉特征，如颜色、纹理、部件模型等，以及如何度量视觉特征之间的相似度和区分度；目标定位主要是通过分类器对行人所在的位置进行确定。行人检测得到了广泛而深入的研究，从早期的基于 HOG/SVM 的行人检测[246]发展到近期的基于行人外观恒定性和形状对称性（Appearance Constancy and Shape Symmetry）[247]的行人检测算法，对行人检测算法的研究主要集中在寻找可以提供更高区分度的外观表现方式和更好的分类器上[245]。

1.2.1 基于 HOG/SVM 的行人检测

早期比较著名的行人检测算法是 Dalal 和 Triggs[246]提出的基于梯度方向直方图（Histograms of Oriented Gradients，HOG）和支持向量机（Support Vector Machine，SVM）的行人检测及相关的衍生算法。

HOG 的主要思想为在一幅图像中，局部目标的表观和形状能够被梯度或边缘的方向密度分布很好地描述。其本质为使用梯度的统计信息。而梯度主要存在于边缘的位置，提取 HOG 特征需要首先将图像分成小的连通区域，这些连通区域叫作细胞单元。然后采集细胞单元中各像素点的梯度或边缘的方向直方图，最后把这些直方图组合起来，就可以构成特征描述符。将这些局部直方图在图像的更大范围内（叫作区间）进行对比度归一化，可以提高该算法的性能。所采用的方法是：先计算各直方图在这个区间中的密度，然后根据这个密度对区间中的各个细胞单元做归一化。归一化后，能对光照变化和阴影获得更好的效果。

与其他的特征描述方法相比，HOG 具有以下优点。

（1）由于 HOG 是在图像的局部方格单元上进行操作，所以它对图像几何的和光学的形变都能保持很好的不变性。

（2）在粗略的空域抽样、精细的方向抽样以及较强的局部光学归一化等条件下，只要目标大体上能够保持直立的姿势，可以容许目标有一些细微的形变，这些细微的变化可以被忽略而不影响检测效果。

基于 HOG/SVM 的行人检测的主要算法步骤如下：首先提取正负行人样本的 HOG 特征，并训练一个 SVM 分类器，生成初步的检测器；其次，利用训练出的检测器检测负样本，从中得到难例（Hard Example）；将难例的 HOG 特征和最初的特征一起投入 SVM 进行训练，得到最终的检测器。图 1－1 汇总了 HOG 特征提取和基于 HOG 和 SVM 的行人检测过程。

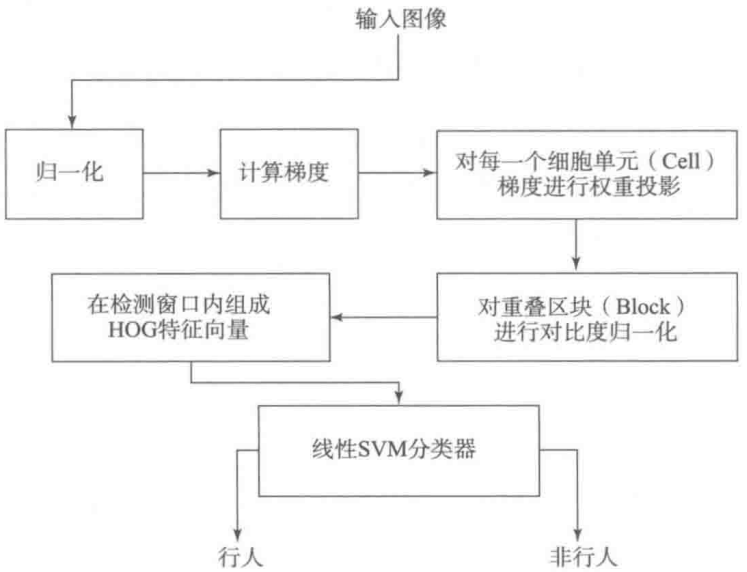


图 1-1 HOG 特征提取和基于 HOG 和 SVM 的行人检测过程

图 1-2 为 HOG 特征在行人边缘信息上的表现，图 1-2（a）为训练样本的平均梯度图，图 1-2（b）为区域内正 SVM 的最大权重，图 1-2（c）为区域内负 SVM 的最大权重，图 1-2（d）为测试样本，图 1-2（e）为计算得到的 R-HOG 述子，图 1-2（f）和（g）为正负 SVM 分别加权后的 R-HOG 算子。

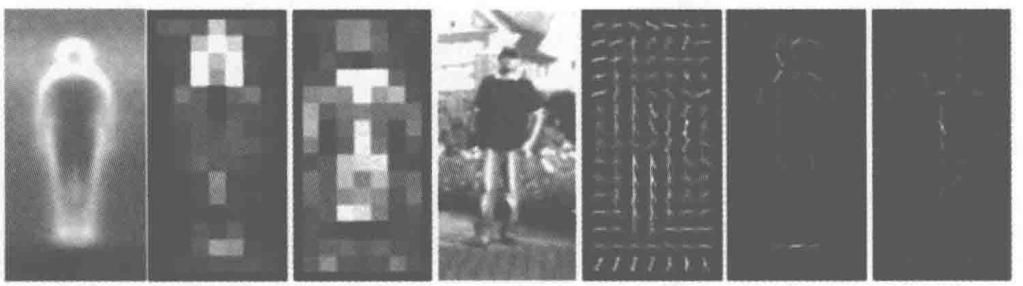


图 1-2 HOG 特征在行人边缘信息上的表现[246]

（a）训练样本的平均梯度图；（b）区域内正 SVM 的最大权重；（c）区域内负 SVM 的最大权重；（d）测试样本；（e）计算得到的 R-HOG 述子；（f）、（g）正负 SVM 加权后的 R-HOG 算子

HOG 特征的提取在行人检测和跟踪中具有重要的作用。HOG 特征与 SVM 结合的行人检测算法被提出后，许多相应的衍生算法围绕 HOG 和分类器结合

这一思路，对行人检测进行了研究，进一步提高了行人检测的效果。

1.2.2 基于可变形部件模型的行人检测

Felzenszwalb 等人[248]利用可变形部件模型对基于 HOG/SVM 的行人检测算法进行了扩展。可变形部件模型将目标划分为若干区域（部件），并且各部件之间允许有一定的位移。可变形部件模型可以解决部分遮挡对目标检测造成的不良影响，也对行人的姿态变化具有一定的拟合能力。Felzenszwalb 等人在全局模型之下计算各部件的 HOG 特征和各部件相对于全局模型的位移来进行行人检测。

图 1－3 显示了用 Felzenszwalb 等人的方法得到的多部件行人模板及其检测结果。这一多部件模型通过低分辨率下的根滤波器、高分辨率部件滤波器和部件滤波器相对根滤波器的位移模型三部分来定义。其中低分辨率根滤波器覆盖行人整体，高分辨率部件滤波器覆盖行人的各部位。低分辨率根滤波器主要检测低分辨率下行人整体的边缘信息，高分辨率滤波器主要检测行人各部分具体结构，同时赋予部件滤波器相对根滤波器的位移形变惩罚函数来约束行人外观的形变。

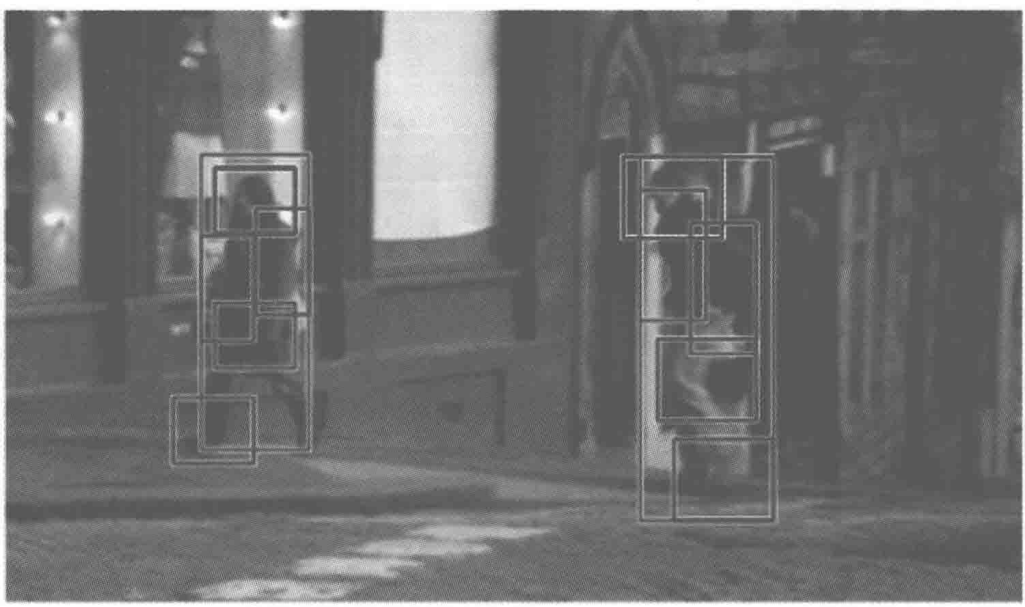
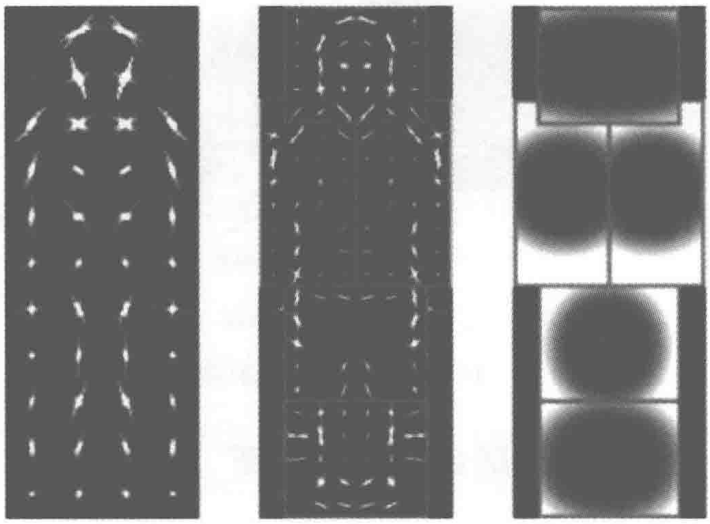


图 1－3 多部件行人模板及其检测结果[248]



(a) (b) (c)

图 1-3 多部件行人模板及其检测结果（续）[248]

（a）低分辨率下的根滤波器；（b）高分辨率部件滤波器；

（c）部件滤波器相对根滤波器的位移模型

可变形部件模型在一定程度上解决了行人姿态变化和遮挡的问题，很好地与行人这一检测目标的属性相符合，是行人检测中的经典方法之一。

1.2.3 基于深度神经网络的行人检测

随着深度学习的出现和兴起，很多人工智能和机器学习领域的问题与应用都得到了很好的解决或提升。深度学习在目标检测和分类等领域取得了非常好的效果[249,250,251]，在行人检测上也取得了很好的表现。Sermanet 首次将深度卷积神经网络（CovNet）应用于行人检测[252]。CovNet 通过非监督学习得到多阶段神经网络用于行人检测。图 1-4 为用于行人检测的 CovNet 结构。CovNet 网络既保留了局部细节（底层表示），又学习了整体特性（中层表示）。

深度神经网络在行人检测问题上取得了很好的效果，并引起了进一步的研究，Tian 等人提出了基于深度语义学习的行人检测算法[253]，通过同时优化行人检测与场景语义，包括行人的属性和场景的属性，来检测行人。

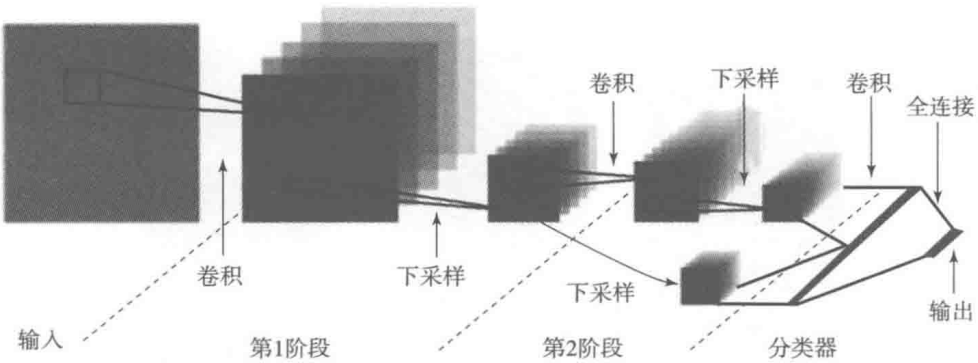


图 1-4 用于行人检测的 CovNet 结构[252]

1.2.4 基于特征融合的行人检测

如何提取具有判别力的特征是行人检测的重要研究内容之一，良好的特征将会有效地提升行人检测的效果。

从特征的角度来看，大部分检测算法都将选择合适的特征作为提升行人检测效果的主要方法。行人检测算法的特征包括 haar 或者类 Haar 特征、运动块特征、局部二值模式（LBP）特征及梯度分布特征等。其中使用最广泛的是 HOG 特征及其相关的扩展。很多其他特征也不断被尝试用于行人检测，有的特征与 HOG 特征类似，以边缘信息作为其主要特征[258,259]，有的特征包含色彩信息[260]，有的特征考虑材质信息[254]、局部外形信息[261]或者协方差特征[262]等。

通道特征通过对原图像进行多个通道变换得到，如 LUV、HOG 等。在行人检测中，LUV 通道、梯度强度通道、梯度直方图通道都是非常有效的通道。

特征是原始图像的某种特性的抽象反映。单一的特征通常只能描述目标某一方面的特性，如 HOG 特征反映的是图像的梯度统计特性，而 LBP 特征反映了图像的纹理统计特性等。因此，将多种特征进行组合，特别是组合具有相互弥补特性的特征，对目标进行描述，可以有效地提高特征对图像的描述能力。Wang 等人提出 HOG-LBP 描述子，利用 LBP 特征来估计出遮挡区域，提升了检测的效果[254]。Wojek 等人利用 HOG、Haar 和 HOF（Histogram of Flow）特征进行检测，提升了车载环境中的行人检测效果[255]。

1.2.5 行人检测中的分类器

分类是行人检测的重要组成部分，分类器根据取得的特征进行分类。行人