计算机视觉前沿发展

——目标检测专题

田　彦◎主编

浙江工商大学出版社
ZHEJIANG GONGSHANG UNIVERSITY PRESS

计算机视觉前沿发展

——目标检测专题

主　编

田　彦

副主编

程国华　杨柏林　包翠竹　李建元

编　委

吴佳辰　杨　涛　余路阳

江腾飞　赵晓波　赵绪然

刘丹丹　虞世豪　季红丽

浙江工商大学出版社

ZHEJIANG GONGSHANG UNIVERSITY PRESS

·杭州·

图书在版编目(CIP)数据

计算机视觉前沿发展：目标检测专题 / 田彦主编.
— 杭州 ：浙江工商大学出版社，2020.5
ISBN 978-7-5178-3629-2

Ⅰ. ①计… Ⅱ. ①田… Ⅲ. ①计算机视觉—研究
Ⅳ. ①TP302.7

中国版本图书馆 CIP 数据核字(2019)第 280647 号

计算机视觉前沿发展——目标检测专题

JISUANJI SHIJUE QIANYAN FAZHAN——MUBIAO JIANCE ZHUANTI

田 彦 主 编

责任编辑 杨 戈
封面设计 雪 青
出版发行 浙江工商大学出版社
(杭州市教工路 198 号 邮政编码 310012)
(E-mail:zjgsupress@163.com)
(网址:http://www.zjgsupress.com)
电话:0571-88904980,88831806(传真)
排 版 杭州朝曦图文设计有限公司
印 刷 浙江全能工艺美术印刷有限公司
开 本 787mm×1092mm 1/16
印 张 9.5
字 数 243 千
版 印 次 2020 年 5 月第 1 版 2020 年 5 月第 1 次印刷
书 号 ISBN 978-7-5178-3629-2
定 价 40.00 元

序

本书起源于2014年年初，当时，海康威视邀请作者总结计算机视觉领域的技术发展和未来趋势。那个时候，深度学习刚刚在计算机视觉领域崭露头角，在ImageNet数据库上将识别准确率提升了十个百分点，但是在其他计算机视觉上的任务并未取得显著成果，其计算复杂度、显存占用率也是令人头疼的缺点。公司迫切需要知道，这种新技术是否能够成功应用到自己的产品中，以及是否有其他颠覆性的技术，可用于提升产品的竞争力，以便在未来的销售过程中各种碾压所谓的“友商”。

不光是智能视频监控领域，其他诸如电子商务、3D数字化与3D打印、智慧医疗、新媒体等领域的公司，都和计算机科学发展结合紧密，有着将学术界最新科技成果应用到工业界的动力。在生产实践过程中，工业界逐渐发现计算机视觉发展迅速，能够不断帮助公司改善产品性能，从而在产品竞争中处于优势地位。于是，这些公司希望时刻能关注、了解计算机视觉技术的最新发展，但是，学术界的“灌水”风气又使得工业界特别的不习惯。每年计算机视觉领域的学者和学生都会制造数以万计的学术论文，其中，97%以上的论文由于不关注算法实时性、内存占用率、数据标定的代价、公开库数据和真实数据间的巨大鸿沟等实际因素，使得论文上的算法无法在实际工程中部署和使用，工业界迫切地希望能在海量的论文中快速地寻找到适合自己产品的算法，用于验证和试错。

近年来，各种大数据集上的竞赛帮助工业界解决了不少问题，如ImageNet和KITTI，有效地检索出性能和效率都适合技术落地的算法，但是，只关注单个竞赛或数据库，仅能得到部分有用算法，难免遗漏在其他竞赛或数据库上性能强劲的算法或方法。

为此，针对工业界工程师的痛点需求，作者根据多年的研究和工程实践经验，在本书撰写过程中，加入了一些其他书籍不包含但一线工程人员又特别感兴趣的内容：

1)在每章的起始处，将该章所枚举的算法按照算法论文提供的、在广泛比较的公开数据库下的性能进行排名，迫切需要短期内了解特定问题目前做到什么程度、什么是效果最好算法的读者，可以直接通过阅读该表格获取所需信息。该表格同时收集多个主流对比公开库下算法的效果，方便读者尽可能充分地了解算法的性能。截止本书定稿时，从CVPR2014一直对比到ICCV2019的部分目标检测算法。

2)每种的列出算法是否有源代码可以下载，如果有代码可以下载，是何种语言或平台的代码，如C表示该算法提供C/C++语言代码，Matlab或m表示该算法提供Matlab语言代码，Python或p表示该算法提供Python语言代码，—表示算法作者在论文中声称会在未来的某个时间公布算法代码。由于近年来深度学习在计算机视觉的几乎各个任务都取得

了最具有竞争力的结果，因此，所有深度学习算法还专门细分为算法运行平台，如 Caffe、Tensorflow、Torch、Pytorch、Theano、MxNet、Keras 等。需要注意的是，某些算法可能会在本书作者统计之后的某段时间新提供了源代码或提供了其他语言或平台下的源代码，所以本书显示没有提供源代码的算法，也具有本书定稿后能够找到公开源代码的可能性。

3)算法的计算复杂度也是算法能够成功落地的关键因素，在表格的最右边，同样说明了算法提出者在论文中描述的该算法在特定输入分辨率、特定硬件条件支持下的计算时间。

本书适用于致力于计算机视觉研究和开发的本科高年级学生和研究生，以及活跃在开发第一线的广大计算机视觉算法工程师们。

感谢国家自然科学基金(61602407，61972351，61976188)、国家重点研发计划(2018YFB1403200，2016YFB1000400)、国家卫生计生委(现为国家卫生健康委员会)科学研究基金(WKJ－ZJ－1814)、浙江省自然科学基金(LY19F030005)、浙江省重点研发计划(2019C03002)、杭州市重大科技创新专项(20172011A038)参与资助本书的出版。感谢浙江工商大学、杭州健培科技有限公司、杭州先临三维科技股份有限公司、银江股份有限公司的同事和好友，在本书的撰写过程中提出的许多宝贵的建议。感谢家人在作者撰写和申请各种基金、备课和上课、完成各种纵向和横向课题、撰写论文、专利、软件著作权、著作、指导本科生和研究生过程中，给予的理解和支持！

由于编者水平有限，书中难免存在纰漏，欢迎广大读者批评指正。在阅读过程中，如果发现问题，请发送电子邮件告知，以便今后重印时加以订正。

杭州钱塘新区

2019 年 8 月

目 录

第一章 概 述

1.1 什么是目标检测

目标检测(Object Detection)问题是计算机视觉领域的经典任务之一,学术界已有二十多年的研究历史。图像识别通过模拟人类的感知过程实现对自然场景的理解,包括四大类任务:一是分类(Classification),主要解决"what"问题;二是定位(Location),主要解决"where"问题;三是检测(Detection),主要解决"what & where"问题;四是分割(Segmentation),它又分为实例分割(Semantic segmentation)和语义分割(Instance-level),解决"每一个像素属于哪个实例或哪一类"的问题。其中,目标检测的任务是找出图像中所有感兴趣的目标,确定它们的位置(where)和类别(what)。

由于各类物体有不同的外观、形状、姿态,加上成像时光照、遮挡等因素的干扰,目标检测也一直是计算机视觉领域最具有挑战的问题。其中,小尺度目标,主要是因为经过多层卷积、池化处理后,特征图的分辨率大幅下降,小目标的局部信息已经完全淹没在全局信息中。动态目标(多姿态目标),主要是因为卷积层、池化层具有固定几何结构,缺乏对物体形变的内部建模能力(如同层的所有激活单元具有相同感知域)。部分被遮挡目标,主要是因为目标被遮挡后局部信息丢失、训练数据和测试数据的遮挡模式不完全一致。

以交通场景为例,目标检测需要达到的目的如图 1-1 所示,给定一张场景图片,目标检测算法以物体外接框的形式确定物体所在区域,并给出物体的所属类别,如人、车、交通标识符等。小尺寸的交通标识符、被遮挡的车辆、发生动态形变的行人等,都是目标检测算法难以解决的问题。

图 1-1 目标检测效果图

1.2 2D 目标检测简史

最近的十几年里,各类计算机视觉任务取得了显著的进步,在深度学习之前,研究人员们

提出了各种类型的特征提取器与分类器，例如 SIFT、Haar、HOG、Strip 等经典图像特征。但这些方法存在两个难以克服的问题：(1)由于缺乏高级语义信息，手动提取的特征并不可靠；(2)即使提供了大量带标注的数据，从特征空间到标签空间的多模型和非线性映射也非常复杂，以至于在有限的参数下很难很好地学习。由于目标检测问题是计算机视觉领域的经典问题，二十年间被提出的方法的种类非常多，本章仅以使用最为广泛的集成学习(Ensemble Learning)架构、支持向量机(Support Vector Machine，SVM)架构、动态部件模型(Dynamic Part Model，DPM)、实例学习架构(Exemplar Learning)描述目标检测经典方法。

1.2.1 集成学习架构

传统的集成学习目标检测算法，通过(带一定滑动步长的)滑动窗遍历输入图像，在每个滑动窗覆盖范围内的图像块提取诸如 SIFT 的低层特征，从而将图像块像素内容转化为特征向量，然后，将该特征向量放入级联分类器中进行训练，得到该图像块的最终判断结果(分类，是否是所需目标的图像块)。对于传统的目标检测算法，该方法通常存在以下问题：(1)基于滑动窗口的区域选择策略没有针对性，时间复杂度高，窗口冗余；(2)手工设计的特征对于多样性的变化没有很好的鲁棒性。以上两点影响了传统目标检测算法的速度和精度。随着训练数据的增加，该传统算法效果并未有显著提升。

1.2.1.1 低层特征设计

除了被广泛使用的 SIFT、LBP、HOG、Haar 等低层特征，历史上也出现过其他一些针对具体问题进行特殊人工设计的低层图像特征，并且取得了一些效果。

Li 等在 CVPR 2013[1] 采用级联检测器完成车辆检测任务，特征使用多维 SURF(Speeded Up Robust Features，SURF)，弱分类器采用逻辑回归(logistic regresstion)。在侧面车辆检测任务中，PACSAL VOC2005 数据集的检出率 70%，FPPW (False Positive per Window，FPPW)为 2×10^{-6}。多维 SURF 特征能够很好地平衡特征表达能力与计算效率。具体而言，Li 在特征选择时使用稠密局部采样，特征提取是使用 8 通道的 T2 描述子。给定检测窗口的模板大小(40×40)，定义 4 个空间单元的图像块，允许图像块大小变化(从 12×12 到 40×40 像素)，同时允许图像块有不同的长宽比，如 4 个空间单元不仅可以配置为 2×2，还可以配置为 4×1 和 1×4。按照这种方式，在 40×40 检测模板内可以生成 450 个图像块。并对每个图像块提取 32 维的 SURF 特征(8 直方图区间×4 空间单元)。当 SURF 特征提取特征的通道数为 8 时不利于速度优化，为了在保留通道数的同时提升效率，Li 提出了 T2 描述子，并将 4－直方图区间 T2 描述符扩展到 8－直方图区间，从对角线和反对角线方向的梯度连接另外 4 个直方图区间，给出了 2D 滤波核，从而分别得到了主对角梯度图像和反对角梯度图像。8 通道的 T2 描述子，具有相似于原始 SURF 特征、甚至优于 HoG 特征的表示能力，且在特征提取速度上占优势地位。

1.2.1.2 算法效率提升

1)训练效率

算法训练效率的提升对测试时的实时运行没有帮助，但是可以帮助进行算法参数的调试，即同等时间可以完成更多次的参数调整实验。Cascade 检测器的性能取决于用于估计

拒绝门限的验证集，CVPR 2013[2]加速这种分类器验证过程，在每个中间响应处保持中间过程分类标记的概率估计，并在相应的不确定性足够小时停止运算。根据观察到的响应顺序，评估提前终止。此外，它是独立于使用的集成分类器类型或它的训练方式来实现的。实验证明，与当时最先进的方法相比，该方法可以将速率提升 2 到 10 倍，在许多对象分类任务上几乎没有精度损失。

2）测试效率

测试效率的改善，直观的结果就是可以提高预测时的处理速度，使得原来仅能用于非实时运算场景的算法实现实时的处理速度。Lampert 等在 CVPR 2008[3]针对滑动窗效率低，提出一个简单但功能强大的分支定界方案，即有效子窗口搜索（Efficient Subwindow Search，ESS）。该方法允许在所有可能的子图像上有效地最大化一大类分类器函数，通常在次线性时间内收敛到全局最优解。这种方法可以适用于不同的对象检测和检索场景，实现的加速允许使用分类器进行本地化。在 UIUC 汽车数据集、PASCAL VOC 2006/2007 数据集上进行测试，取得了 2007 年当时最先进的性能。

Benenson 等在 CVPR 2012[4]通过处理尺度问题并将部分计算从测试端移到训练端，从而提高检测运算速度。在处理图像时，系统能够实现 50fps 的高质量检测。而在一台 CPU＋GPU 台式机上，从校正输入到检测输出可以达到 135fps 的处理速度，成为目标检测算法追求实时性的经典方法。

Dean 等在 CVPR 2013[5]针对每个候选位置、每个分类器都判定一遍速度太慢的问题，将卷积变为局部敏感哈希（locality-sensitive hashing），用固定数量的散列表探针代替卷积中的点积核运算符，该探针可以有效地对所有的过滤器响应进行时间采样，而不受过滤器组大小的影响。具体方法为：事先计算得到 C×P 个滤波器对应的哈希值，1）计算多尺度的边缘强度和边缘方向图像；2）对所有窗口进行遍历，对于每个窗口，计算其高斯加权 HOG 直方图特征，计算特征对应的哈希值，分别计算 HOG 特征哈希值和 C 类 P 个滤波器的哈希值的汉明距离；3）将具有局部最大响应的窗口作为候选，得到可能的物体中心的分布累积，综合得到最终的物体检测结果。这篇文章参考的基本算法是下文将要提到的 DPM 模型，分类 100000 类别，mAP 为 0.16，需耗费时间 20s，比 DPM 模型的计算速度提高了整整 20000 倍。

Costea 等在 CVPR 2017[6]提出一种基于 boosting 的滑动窗方法，如图 1-2 所示，使用颜色、运动、深度多种特征，引入了信号强度、梯度幅度和方向通道的多模态多分辨率滤波，以便在多个尺度和方向上捕获数据结构。为了实现尺度不变的分类特征，分析了尺度变化对不同滤波器类型特征的影响，提出了基于理论和经验的尺度校正方案。为了提高识别能力，提出了几种上下文特征通道，如：二维上下文、对称通道、三维上下文、三维几何通道。实验结果表明，在 KITTI 等标准数据集上测试，其速度是深度学习方法的 10－100 倍。这种方法的性能表明，尽管深度学习方法可能在目标检测领域占主导地位，但传统的滑动窗口方法可以在具有竞争力的情况下提供一种低成本的替代方法。

1.2.1.3 多姿态（视角）问题

动态目标、多姿态目标、多视角问题，是目标检测的常见问题，主要是因为卷积层、池化层具有固定几何结构，缺乏对物体形变的内部建模能力（如同层的所有激活单元具有相同感知域），从而引起检测过程性能的下降。

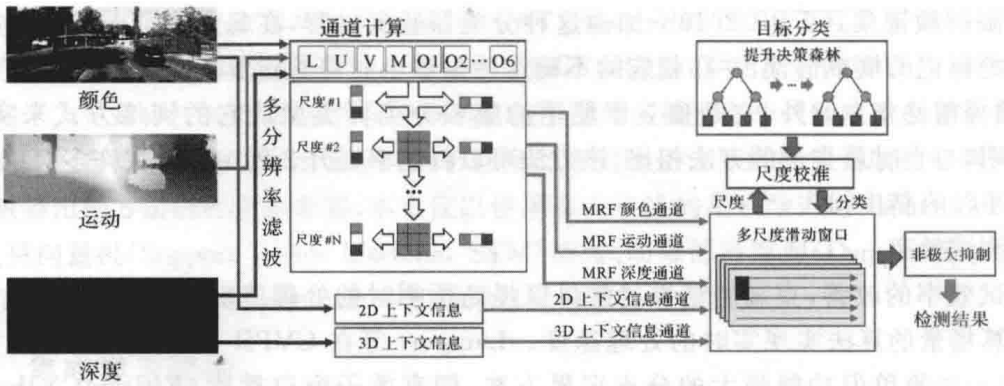

图 1-2 基于尺度不变多模多分辨率滤波特征的快速增强检测算法

1)特征提取角度

Zheng 等在 CVPR 2009[7]为进行多视角车辆检测，提出一种 strip 特征，代表了各种类型的线条、弧线以及带有边缘形状和脊状的条形图案，显著丰富了 haar－like 特征和 edgelet 特征等简单特征。同时开发了一种复杂度感知的 realboost，以平衡所选特征的识别能力和效率。在 UIUC 数据集上的实验表明该方法速度快且性能良好，EPR 率达到 96%。

Wang 等在 ICCV 2013[208]为处理目标形变的问题，使用级联增强(cascaded boosting)的 Regionlets 特征进行检测，如图 1-3 所示，即滑动窗或选择性搜索(selective search)得到候选窗，区域里的 Regionlets 表示 part 可能的位置。在训练时，遍历位置和尺寸得到区域，随机位置固定尺寸得到 regionlet 特征，boosting 选择特征维度(得到 1D 特征)；而测试时，每个区域内的所有 regionlet 特征中选择激活值最大特征(max－pooling)作为整个区域的 1D 特征(对形状变化鲁棒)。在 PASCAL VOC 2007 的 mAP 为 41.7(bus 55.5，car 68.7，person 43.4)，PASCAL VOC 2010 的 mAP 为 39.7(bus 56.1，car 54.5，person 43.5)，KITTI Vision Benchmark 达到 Moderate 75.58 %。

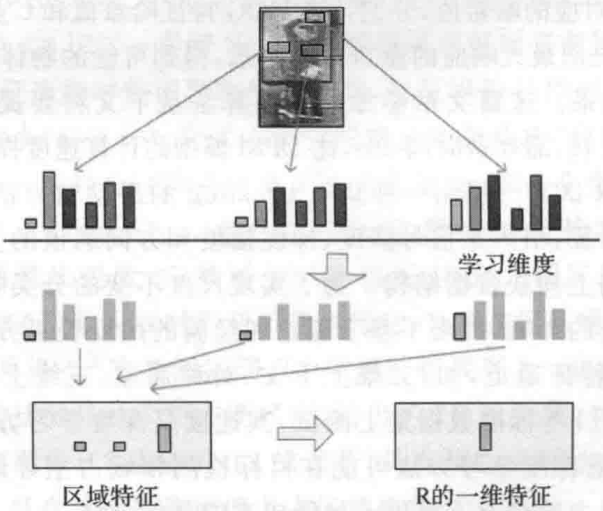

图 1-3 regionlet 特征中对形状变化鲁棒的特征选择

2)分支定界

Wu 等在 ICCV 2007[8]使用非监督的分支定界处理多视角多姿态问题，即 boost 寻找每个辨别特征后，通过验证最近 3 次选择特征的辨别力(与固定门限比较)，如果辨别力不足，则用 k-means 将样本聚成子类，重新开始学习分类树，在 UIUC 数据集上 equal-precision-

recall 率为 92.8%。

Ohn-Bar 等在 CVPR 2014[521c]针对 DPM 检测速度慢不适合移动设备，提取 LUV 和 HOG 特征，使用 DSC 聚类，cascade boosting 进行检测，如图 1-4 所示。在这项研究中，视觉子分类是一种捕捉外观变化的方法。具体而言，使用颜色和渐变特性对训练数据进行聚类，而聚类用于学习一组模型，这些模型捕获由于方向、截断和遮挡程度不同而产生的视觉变化，然后利用积分图像特征和像素查找特征来实现快速目标检测。这一方法实现了快速检测，同时保持与 DPM 相当的结果，在 KITTI 数据集上主观效果更好。

图 1-4 基于聚类的分支定界检测方法

3)联合检测和对齐

Chen 等在 ECCV 2014[9c+p]针对 cascade 的 SVM 后处理费时的问题，(人脸形状用 27 个点标记的 54 维向量表示)，使用决策树同时实现人脸检测和对齐，这种联合学习大大提高了级联检测的能力，并保持了其实时性。实验表明，FDDB 和 AFW 数据集测试好于 fastDPM 和 Boosted Exemplar，VGA 图像检测大于 80×80 图像，用 16 核 CPU 进行 3 天的训练，测试单线程 2.93 GHz CPU 上 0.0286s/帧，占用 15M 内存。

4)个体检测器

Hall 等在 CVPR 2014[10c]针对(检测、跟踪、重验证中)目标在视频中的变化，提出了一种在线实时学习单个目标检测器的方法，先使用类别检测器，然后在线实时训练基于 boosting 的个体检测器。通过对弱分类器阈值的基本操作，使用与类别检测器相同的特征级联，得到单个检测器。在 Fifty People One Question (FPOQ) 人脸数据集、Caltech Roadside 行人数据集、VIPeR 数据集和 ETHZ 数据集验证性能的过程中，该方法能够实现在线交互式学习。

1.2.1.4 遮挡

物体如果全部被遮挡，那检测器是无法进行准确的定位和分类的，本文这里描述的是物体部分被遮挡的情况。部分被遮挡目标，主要是因为目标被遮挡后局部信息丢失、训练数据和测试数据的遮挡模式不完全一致，从而造成的检测器性能下降问题。

Zhou 等在 ICCV 2017[11]采用多标记学习(multi-label learning)、联合学习 part 检测器和遮挡模式，part 检测器共享 boosting 决策树来挖掘 part 关系和降低复杂度。在具体实现时，学习决策树使用 AdaBoost. MH 方法，特征使用通道有限元和 CNN 特征，由于决策树的共享，利用了 part 相关性，降低了应用这些 part 检测器的计算成本，所学习的决策树可以捕获所有 part 的总体分布。在加州理工学院的行人检测数据集上的实验证明了该方法的有效性，检测严重拥挤的行人时取得了当年的最佳效果。

1.2.2 SVM 架构

支持向量机是一种基函数可调的分类模型，通过找到对分界曲面影响最大的支持向量来学习复杂的分类曲面，相对于深度学习模型，SVM 可以看作是一种 2 层的浅网络模型。

Benenson 等在 CVPR 2013[12] 比较各种基于 HOG+SVM 框架的特征池建立、特征选择、预处理、训练办法，并用于行人检测，证明了通过适当地设计特征池、特征选择、预处理和训练方法，可以使用单一的刚性组件达到最佳质量，而此最佳检测器是完全前馈的，具有单一的统一架构。

1.2.2.1 遮挡

Li 等在 ECCV 2014[13c] 提出使用分层 And-Or 模型并结合环境上下文和遮挡信息进行车辆检测，在具体实现时，0 层 root Or-node 表示不同的配置（每个配置是一个 And-node）；1 层 And-node 表示不同的车（每个车是一个 Or-node）；2 层 Or-node 表示不同的视角遮挡模式（每个模式是一个 And-node）；3 层 And-node 由 part 组成，part 用车辆 3D CAD 仿真或 heuristic method (DPM)得到。训练分两个阶段：1）学习分层 And-Or 模型结构，形成有向无环图 DAG 并用 DP 预测；2）使用弱标签结构 SVM 学习（外观、形变、偏差）模型参数。在 KITTI 数据集取得 AP 为 Easy80.26%，Moderate 取得 AP 为 67.03%，Hard 取得 AP 为 55.60%；PASCAL VOC 2007 Car 上 Ap 为 60.6%。

Hu 等在 CVPR 2012[14] 从不同标定图像学习（由外观和几何定义的）3D 对象模板（可形变的平面部件模板组成），分层的 Gabor 过滤器得到线段和几何形状表示的外观。AND-OR 树量化各个部件模板的几何和外观。AND-OR 树通过自底向上和自顶向下的信息增益得到最佳的 3D 模板。预测时，三维可变形模板可以投影到二维可变形模板上，滑动窗口使用这些可变形的二维模板在该视图中执行检测。在每个窗口中，动态编程用于推断所有可能变形的最大二维模板得分。

Lan 等在 ICCV 2013[15] 基于弱监督的方式，输入基本层目标类别（如人、车），用基于 exemplar-SVM 的聚类自动得到低层目标子类别（如方向、外观）和复合类别，如图 1-5 所示，提出了一种基于范例支持向量机的聚类方法，该方法具有隐支持向量机（latent SVM）细化功能，可以发现每个对象类的一组可变长度的区分子类别。然后，开发了一个用于对象检测的结构化模型（structured model），该模型捕获对象子类别之间的交互，并自动发现语义上有意义的和有区别的高层视觉环境组合（子类别组合，如人骑车）。在 UIUC car 上的实验取得的 mAP 为 38.1，取得了当时的最佳性能。

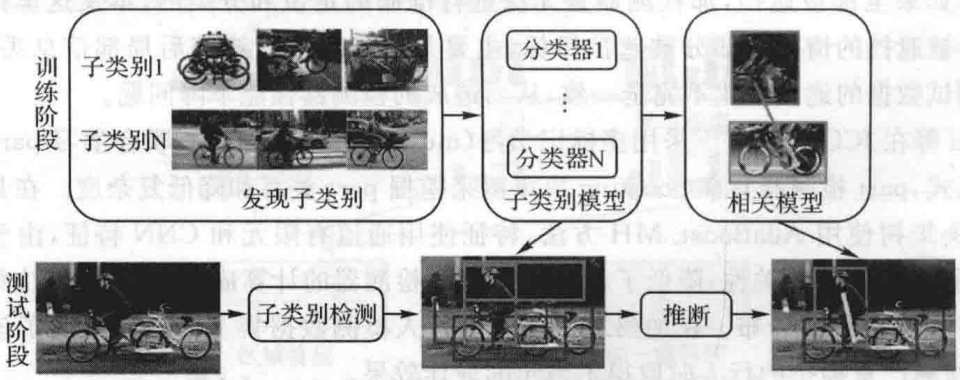

图 1-5 基于 exemplar-SVM 的弱监督目标检测

1.2.3 DPM 架构

形变部分模型（Deformable Part Model，DPM）是目标检测任务中的一个里程碑级算

法，通过将物体拆分为部件从而很好地处理了物体遮挡问题，同时，巧妙地将分类函数表达为结构化学习问题，基于支持向量机的思想学习复杂的分类曲面。

Felzenszwalb 等在 CVPR 2008[16c] 和 PAMI 2010[17c] 提出形变部件模型 DPM，将边缘敏感的数据挖掘难负样本和潜在支持向量机(latent SVM)相结合。潜在的支持向量机像一个隐藏的条件随机场，导致非凸训练问题。然而，latent SVM 是半凸训练，一旦为正例指定了潜在信息，训练问题就变成凸训练。这种训练方法能够有效利用更多潜在信息，例如层次(语法)模型和包含潜在三维姿势的模型。实验表明，在 PASCAL VOC 2006 上平均精度 bus 为 0.502，car 为 0.631，pers 为 0.401；PASCAL VOC 2007 上平均精度 bus 为 0.397，car 为 0.516，pers 为 0.368；PASCAL VOC 2008 上平均精度 bus 为 0.251，car 为 0.334，pers 为 0.431。

1.2.3.1 遮挡

多个行人距离比较近时，检测框合并和遮挡都会引起漏检，Ouyang 等在 CVPR 2013[18m] 将附近行人作为检测线索，使用混合 DPM(每个成份由聚类得到)的多行人检测得到附近行人视觉特征，概率建模单检测器和多检测器结果关系。该方法是通过多行人检测器为单行人检测提供丰富的补充信息，能对单行人检测器进行有效改善，这种方法非常灵活，它可以集成任何单个行人探测器，而不会显著增加计算负荷。与当时最先进的方法进行比较，加州理工学院测试数据集的平均改善率为 9%，都柏林布鲁塞尔数据集的平均改善率为 11%，ETH 数据集的平均改善率为 17%；加州理工学院测试数据集的最低未命中率从 48%降至 43%，都柏林－布鲁塞尔数据集的最低未命中率从 55%降至 50%，ETH 数据集的最低未命中率从 51%降至 41%。

Yan 等在 CVPR 2012[19] 利用混合模型将每个行人设为特定子类并用子模型描述(人体部分表示外观，二次核表示空间交互)，将参数估计问题转为排序问题并使用隐排序 SVM 学习模型参数和隐子类标记，最后检测问题转为最大后延概率问题并使用贪婪算法得到近似解。该方法将识别参数学习定义为一个学习排序问题，并提出了利用弱标记数据学习的隐秩支持向量机。

针对密集人群中的人体检测问题，Ouyang 等在 CVPR 2013[20m] 中提出了一种联合估计多个共存行人可见度状态的互可见度深度模型。文中认为重叠行人的遮挡/可见性状态为可见性估计提供了有用的相互关系，即一个行人的可见性估计有助于另一个行人的可见性估计，因而提出了一个共同的能见度深度模型，共同估计重叠行人的能见度状态，从识别共存行人的深层模型出发，研究行人之间的可见度关系。该方法在 Caltech 列车数据集、Caltech 测试数据集和 ETH 数据集上实现了最低的漏检率。

1.2.3.2 多姿态(视角)问题

Kokkinos 等在 NIPS 2011[21] 使用分支定界和形变部分模型 DPM 快速检测形变目标，其中使用对偶树(Dual Trees)确定部分形变的边界。在此基础上，Kokkinos 等在 ICCV 2013[22] 使用稀疏编码学习形变部分模型 DPM 的共享基，从而挖掘不同目标种类的结构，结合分支定界对偶树和 cascade 用于目标检测。

1.2.3.3 效率提升

Pedersoli 等在 CVPR 2011[23] 认为检测成本很可能由将每个部件(part)与图像匹配的

成本决定，而不是由通常假定的部件最佳配置计算成本决定。为了尽可能减少部分与图像的比较，提出了一个基于多分辨率层次部件模型和一个相应的粗到细(Coarse-to-fine)的推理过程，该过程递归地由低精度的部分位移得到高精度的部分位移，并迭代去除无潜力的部分位移。该方法比标准的动态规划方法速度提升了 10 倍，并且在某些情况下，结合部件级联方法速度可以提升 100 倍。在 Pascal VOC 和 INRIA 数据集上的测试结果表明，速度得到了很大的提升，精度降低很小。

Yan 等在 CVPR 2014[24]针对 DPM 速度慢的问题，1)将根滤波器限制为低秩，这样 2D 内积可以用 1D 内积表示；2)邻域感知 cascade 去除附近有更高得分的半正样本，去除附近有很低得分半负样本；3)使用查表计算 HOG；Pascal VOC 2007 上平均精度 bus 为 45.2%，car 为 54.1%，person 为 41.5%，VGA 图单核 CPU 效率为 0.25－0.33s/图。

1.2.4 Exemplar 架构

Malisiewicz 等在 ICCV 2011[25c]提出 Exemplar-SVMs 集成并用于目标检测，为训练集中的每个范例(和数百万负样本)训练一个线性 SVM。虽然每个检测器都是特定于其范例的，但这种 Exemplar-SVMs 的集合得到了良好泛化。这种核心方法的核心优势是，在每个检测和单个训练示例之间创建了一个明确的关联。由于大多数检测都显示出与相关范例的良好对齐，因此可以将任何可用的示例元数据(分段、几何结构、3D 模型等)直接传输到检测上，然后将其用作整体场景理解的一部分。

1.2.4.1 复杂背景

Monroy 等在 ECCV 2012[26]解决标定框内背景影响检测的问题，1)首先数据驱动的聚类，组合 CMPC 得到前景分割的假设的分割结果，分割加权得到密度图，用 mean-shift 聚类和连接性得到组，从而用于建立包(bag)；2)然后基于多示例学习同时学习检测函数和目标分割；在 Pascal VOC 2007 comp3 上结果为 43.7mAP(car 59.8，bus 51.6，person 41.9)。

1.2.4.2 效率提升

Li 等在 CVPR 2014[27]针对基于样例的方法空间和时间复杂度高的问题，提出一种 boosted 基于样例的方法(包括负样本)，如图 1-6 所示，首先测试图像和 exemplar 得到相似度投票图，然后再与 domain-partitioned classifier 得到置信图，将图像分为重叠的堆(tile)以降低内存占用，使用图像与模板双金字塔以降低检测时间，使用 500 个 exemplars，速度 0.9s/帧(1480×986)，内存占用 150MB。

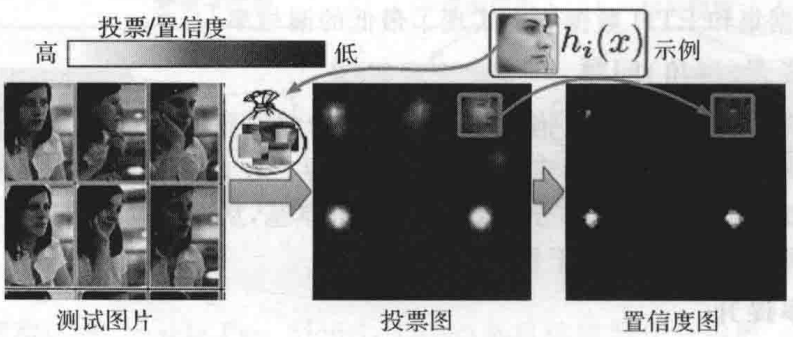

图 1-6 boosted 基于样例的检测方法

1.2.4.3 多类别检测

Razavi 等在 CVPR 2011[28] 提出了一种可扩展的多类检测算法，该算法在不影响精度的情况下，以类的数量次线性地扩展。为此，对所有类进行了特征外观的共享识别码本的联合训练，并对所有类进行了联合检测。使用全连接凝聚聚类同时学习所有类别共享的字典，检测时使用外观和位置信息得到字典权重同时检测所有类别目标。然后利用分类法进一步降低多类对象检测的成本。该方法具有线性训练和子线性检测的复杂性。Pascal VOC 2006 和 Pascal VOC 2007 数据集实验表明缩放系统不会导致精度损失。

1.2.5 深度学习架构

Hinton 在 2006 年提出了深度学习，它使用多层映射从大量数据中学习判别表示和非线性映射。与手动设计的特征相比，多层拓扑获得了更强大的辨别能力，并且可以从特征图中可视化语义信息。但当时由于计算机算力等原因，并未引起特别大的关注。直到 2012 年，Hinton 的学生 Krizhevsky 利用卷积神经网络(CNN)在 ImageNet 大型视觉识别挑战(ILSVRC)中取得了突破性的进展，打败了 Google，顿时让学术界和工业界哗然，深度学习开始走向热潮。

区域卷积神经网络(Region Convolutional Neural Network，R-CNN)首次成功将深度卷积神经网络引入目标检测问题，利用多层非线性处理逐渐提取目标高层语义特征，刷新集成学习和形变部分模型(Deformable Part Model，DPM)在目标检测问题上的效果。近年来深度学习架构在 Pascal VOC、COCO 等数据集上也取得了重大突破。

1.2.6 数据库

深度学习的关键在于实践，从图像处理到语音识别，每一个细分领域都有着细微差别和独特的解决方法。数据的获取是一个重要问题，一部分研究论文使用专有数据集，这些专有数据集通常不会公开。那么，想实践最新的理论方法就成了难题，好在目前公布了许多开源数据库，如 Pascal VOC、COCO 等数据集，深度学习爱好者可以使用这些数据集来提高自己的实践能力。

MNIST 是最受欢迎的深度学习数据集之一。它是一个手写数字数据集，包含 6 千个样本的训练集和 1 万个样本的测试集。这是一个很不错的数据集，它可用于在实际数据中尝试学习技术和深度识别模式，并且在数据预处理上它花费的时间和精力极少。

COCO 是一个可用于目标检测、分割和图像描述生成的大型数据集，包括 3 万 3 千张图像，80 个对象类别，每个图像 5 个描述，25 万人的标记结果。

ImageNet 是基于 WordNet 层次结构组织的图像数据集。WordNet 包含约 10 万个短语，ImageNet 平均提供了约 1 千个图像来说明每个短语。图像总数约 100 万张，每个都有多个边界框和相应的类标签。

Open Images Dataset 是一个包含超过 900 万个链接图像的数据集。其中包含 900 万张图像的训练集，4 万张图像的验证集以及 12 万张图像的测试集。它的图像种类跨越数千个类别，且有图像层级的标注框进行注释。

The Street View House Numbers (SVHN)是一个为训练目标检测算法而“真实”存在