密级:_____

中国科学院大学
University of Chinese Academy of Sciences

硕士学位论文

特定场景下的弱监督行人检测

作者姓名：张天亮

指导教师：叶齐祥 教授 中国科学院大学

学位类别：工程硕士

学科专业：工业工程

研究所：中国科学院大学 工程科学学院

二零一七 年 五 月

Weakly Supervised Pedestrian Detection on Specific Scenes

By
Zhang Tianliang

A Thesis Submitted to
The University of Chinese Academy of Sciences
In partial fulfillment of the requirement
For the degree of
Master of Industrial Engineering

School of Engineering Science
University of Chinese Academy of Sciences
May, 2017

中国科学院大学直属院系

研究生学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：张天亮

日　　期：2017年5月15日

中国科学院大学直属院系

学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密的学位论文在解密后适用本声明。

作者签名：张天亮　　导师签名：

日　　期：2017年5月15日　　日　　期：2017.5.15

摘　要

智能视频监控系统已经越来越普及，并且被安装到了城市的各个角落，所保留的监控视频已经成为重要的大数据资源。然而，当前对这些视频监控数据上的解读仍然停留在初级阶段，大多数系统仅仅具有视频信息存储功能，只有在发生异常时才进行人工数据搜索。提升视频智能监控系统的自动化处理能力和实现视频中目标的自动解读具有重要的意义。

行人检测是智能监控系统的基本问题，它作为视频监控、智能交通系统、车辆辅助驾驶和视频检索系统的关键技术，同时它也是解决智能监控系统的基础条件。在本文中，我们提出一个解决特定场景下的行人检测问题的弱监督自学习方法，旨在不使用任何样本标注信息的基础上进行行人目标建模。该方法利用三个渐进的学习步骤，包括目标发现、目标增强和标签传播，将每个视频帧中的目标位置视为隐变量，采用渐进的方法进行隐变量求解。与传统的隐模型相比，所提出的渐进隐模型结合了空间正则化，减少了目标候选区域的不确定性，并且增强了目标的位置，以及结合了基于图的传播方法以发现相邻视频帧中的难样本。

本文主要贡献如下：

（1）提出了一种基于纹理复杂度的冗余候选区域排序算法，在保证正例样本查全率的情况下显著减少了目标候选区域的数量；在纹理复杂度方面，采用了LBP 熵和目标完整区域轮廓数量这两个指标；

（2）提出了特定场景下渐进隐模型弱监督行人检测框架，在完全没有标注样本的静态视频数据集下训练一个行人检测器，算法包括目标发现、目标增强和标签传播几个迭代的过程；

（3）渐进隐模型采用了时间和空间正则化来减少样本的模糊性，增强自学习算法的稳定性。

关键词：智能视频监控，弱监督学习，行人检测，纹理复杂度，渐进隐模型

Abstract

Intelligent video surveillance system attracts more and more attention in recent years, which has been widely utilized all over the world. It produces big video data with 24-hour surveillance, but the content of video remains not being well utilized. Most of the surveillance system is used for information storage and retrieval, but many procedure are based on human effort. It is required to develop sophisticated techniques to manage surveillance video, particularly recognize the key objects automatically. Pedestrian detection, one of the most important and fundamental techniques in surveillance, is also a footstone for intelligent transportation system, driving assistant system, and automatic driving system. In this thesis, a self-learning approach is proposed towards solving scene-specific pedestrian detection problem without any human' annotation involved. The self-learning approach is deployed as progressive steps of object discovery, object enforcement and label propagation. In the learning procedure, object locations in each frame are treated as latent variables that are solved with a progressive latent model (PLM). Compared with conventional latent models, the proposed PLM incorporates a spatial regularization term to reduce ambiguities in object proposals and to enforce object localization, and also a graph-based label propagation to discover harder instances in adjacent frames.

The contributions of this thesis are as follows:

1) We propose a re-ranking algorithm based on texture complexity to reduce the redundant of object proposal, which is expected that as many as true object regions are preserved while the number of the object proposals is significantly reduced. Local Binary Pattern (LBP) entropy and complete contour number are used to measure texture complexity.

2) We propose a self-learning approach to train a scene-specific pedestrian model without any human' annotation involved. The self-learning approach is deployed as progressive steps of object discovery, object enforcement, and label propagation.

3) We propose a progressive latent model (PLM), which uses spatial-temporal regularization to reduce ambiguity of discovered samples, as well as addressing the stability of self-learning.

Key Words: intelligent video surveillance, weakly supervised learning, pedestrian detection, texture complexity, progressive latent model

目录

图目录

表目录

第一章 绪论

本章包含三个部分的内容：选题的背景和意义、研究内容和论文的组织结构。

1.1 研究背景与意义

大规模数据处理成为现今社会发展技术的主流，它服务于人类的各行各业，创造经济与社会价值。自进入 21 世纪以来，我国信息技术产业在生产和科研方面已经大幅度加快了发展速度，大数据信息处理已成为一个重要的产业。

视觉系统是人类信息感知的主要来源，计算机视觉系统也已经成为目前计算机智能感知的重要手段。计算机视觉不仅代表着科学技术发展的前沿，同时也是带领人们走向人工智能时代的必经之路。随着高速无线宽带的普及，信息的传播速度越来越快，通信的成本也相应减少，保证了视频和图像数据量显著提升。我国在 2016 年提出的国家十三五规划中明确指出实施国家大数据战略和积极推进云计算和物联网的发展，计算机视觉在城市监控系统中的大规模应用成为一种重要的发展趋势。

所谓视频智能监控系统就是将视频监控技术和计算机视觉相结合，提升系统自动化处理能力，从而减少人工成本，达到在视频监控中对目标的自动实时检测。城市视频智能监控系统已经逐渐渗透到城市的各个公共角落，比如交通系统人流和车流监控、居民区和公司的安保监控、医院的医护监控等。当前大多数监控系统的作用集中在以下几个方面：车辆和行人的目标发现、检测和识别，人流和车流量估计，人群密集地区拥挤程度估计与控制和一些重要物品的防盗。

面对普及的监控系统，24 小时的视频监控数据已经成为我们拥有的巨大数据资源。然而，我们当前对这些视频监控数据上的解读仍然停留在初级阶段，大多数系统仅仅具有视频信息存储功能，只有在发生异常时才人工的对数据进行搜索。为了解决这个问题，我们提出一种特定场景下的自学习目标检测算法，根据目标检测的基础和视频场中的运动信息，实现自学习的目标检测，减轻工作人员的工作量，使对监控视频这种大数据载体的解读与分析成为可能。

1.2 国内外研究进展

目标检测一直是计算机视觉中的主流方向，最近几年在该领域中的进步突飞猛进。国内外很多研究机构都致力于解决目标检测的监督学习问题，一般目标检

测可以分为几个步骤：1）目标候选区域的提取；2）提取该区域的特征表示；3）分类学习算法。我们在目标候选区域提取和弱监督目标检测算法上做了一些研究工作，下面我们介绍一下这两方面的研究现状。

1.2.1 目标候选区域提取

目标候选区域方法是发现一组候选框集合，该组候选区域的数量尽可能的少，并且能够精确地覆盖尽可能多的目标。现有的目标候选方法可以粗略地被分为基于超像素合并策略和基于对象性置信度策略两种。

基于超像素合并策略：通过求解约束参数最小割（Constrained Parametric Min-Cuts，CPMC）的序列，Carreira 和 Sminchisescu[1]提出生成 figure-ground 分割以指示目标。一幅图像可以产生 10000 个冗余区域，其随后由经过训练的回归器排序。Uijlings 等人[2][3]提出一种分层策略（Selective Search）来合并颜色相似的部分并且生成目标候选区域。他们提出使用多个低级特征和合并函数来生成冗余区域，以便尽可能的覆盖目标。这种方法已经成功应用于 R-CNN 目标检测[4]研究中。类似于 Selective Search 策略，Manen 等人[5]提出使用学习的权重作为合并超像素的函数。通过利用 CPMC 和 Selective Search 两者的优点，Rantalankila 等人[6]提出使用具有大量特征的一个合并过程，以及使用类似 CPMC 的过程产生图像分割。一个最近的研究方法 MCG[7]将多尺度分级分割区域结合到高精度的目标区域候选中。MCG 实现了高召回率，但它没有考虑计算效率的重要性。Xiao 等人[8]提出了用于超像素合并的复杂度自适应度量距离，其在不同复杂度水平实现改善的分组。这些基于超像素合并的方法中大多数都会产生大量的候选区域，并且没有为这些区域分配目标的置信度。

基于对象性置信度策略：一个基于目标区域候选的对象性度量的早期工作是 Alexe 等人[9]提出的，他提出使用对象性置信度作为检测窗口包含目标的可能性的得分。基于包括显著性、颜色对比度、边缘密度、位置和大小统计的多信息的组合来估计分数。Cheng 等人[10]提出使用滑动窗口的方法建立的二值化规范梯度（BING）用来提取目标候选区域，该方法是基于使用二进制规范梯度训练有效弱分类器。通过二进制计算的巧妙设计，保证了 BING 的低计算成本，论文中所说在 PC 平台上可以达到 300fps。EdgeBoxes[11]也使用了多尺度和多宽高比的滑动窗口方式，通过检测到的完整轮廓的数量来估计候选窗口的分数。Karianakis 等人[12]将卷积神经网络（CNN）的较低卷积层与快速增强决策树产生鲁棒的目标候选区域。与使用图像金字塔变化尺度的传统滑动窗口方法不同，用于目标候选区域生成的滑动窗口需要考虑对于不同种类的对象而变化的长宽比。然而，为了

降低计算效率，在这些方法中图像不能被紧密地滑动，并且基于对象性置信度的方法产生的目标候选区域常常不能精确地定位目标。最近，Chen 等人[13]关注目标候选区域定位偏差，并提出多阈值跨越扩展方法（Multi-Thresholding Straddling Expansion ，MTSE）使用超像素紧密性来减少定位偏差。紧密性只是一个区域的属性，所以它也产生无序的目标候选区域。在论文[14]中，CNN 学习了一个深度得分，用于更新目标区域候选的置信度。然而，这种基于深度学习的方法是数据驱动的，当与超像素合并方法组合时需要良好的训练。

1.2.2 特定场景目标检测

在行人检测方向，使用监督的方法已经被广泛的研究[15-26]。然而，我们的这项工作涉及特定场景的检测，该类问题通常会使用迁移学习、在线学习、弱监督学习和无监督目标发现等方法。

迁移学习：迁移学习的动机是利用目标域中的上下文和对象分布来提高源域中预训练检测器的性能。研究人员已经研究了上下文线索[27][28]、置信度传播[28][29]和虚拟现实世界适应[30]来实现平滑的转移。高斯过程回归[31]和超像素区域聚类[32]已经被用来探索选择目标域中的“安全”的样本。大边缘嵌入[33]和传导多视图嵌入[26]已经被用来扩展检测器的范围。研究人员还使用域适应来构建自学习相机[34]。

迁移学习可以明显减少人工标注。然而，它遇到了概念间隔问题，例如，存在于源域和目标域之间的目标外观、视角和照明的主要差异。当这个差异明显时，预训练的适应变得不平滑或者不可行。相比之下，在同一场景中自学习初始化和改善检测器自然地避免了概念间隔问题。

在线/半监督学习：在线学习和半监督学习通过利用来自目标域的连续输入数据流来改进特定场景的检测器。经典的 detection-by-tracking（DBT）[35][36]使用离线训练的检测器初始化系统，并且利用时间线索来扩展样本域并且消除检测误差。Tracking-Learning-Detection（TLD）[37]使用单个样本初始化系统，并且使用跟踪和在线学习来提升检测器。尽管 DBT 和 TLD 方法很普及，但是最近研究[38]表明，检测和跟踪简单的组合可能引入性能较差的检测器，因为来自检测和跟踪误差可以在耦合系统中被放大。在 TLD 中使用 P-N 专家来控制精度和召回率，它保证了作为线性动态系统的学习的稳定性。

弱监督学习：弱监督学习（Weakly Supervised Learning，WSL）的输入是图像或视频级的标签（对象类别），并且在学习检测器时，算法就会找到目标对象

[39][40]。WSL 的一般假设是同一类别的对象，并且它们来自一个潜在的聚类，而背景则是多样的。在这样的假设下，使用聚类[41][42]、跟踪[39]、boosting[43]、区域匹配[44]、图标记[40]和多示例学习[45][46]去发现对应的目标，抑制背景并且学习检测器。

WSL 以类似于期望最大化优化的方式，交替的进行样本标记和检测器学习。然而，由于缺少标注信息，这个优化是非凸的，因此容易陷入局部最小值并输出错误的标签[9]。Cinbis 等人[45]使用训练集合的多重剥离，而 Bilen 等人[9]使用凸聚类来防止被陷入到到错误的标记中。我们的这项工作会以一种更合理的方式通过引入关于域知识的正则化项来缓解局部最优问题，这个正则化项是帧内难样本挖掘和帧内相似度传播。

无监督视频对象发现：有一个早期的方法，Wu 等人[43]通过在线部件检测器的提升来学习特定场景的目标检测器，但是它需要离线学习一般的种子检测器。最近的研究[39][47]将无监督的视频目标发现作为两个互补步骤的组合：发现和跟踪。第一步骤是在跨视频帧的突出区域之间建立对应关系，然后第二步骤是在相同视频内关联相似的目标区域。Xiao 等人[47]提出一种完全无监督的找到视频目标候选区域的方法，首先通过聚类发现一组容易聚成组的示例，然后更新其外观模型，以通过初始检测器和时间一致性逐渐检测更难的示例。这种无监督的方法可以自动生成目标候选，但不能输出精准的检测。

1.3 本文研究内容

本课题研究的目标是构建特定场景下的弱监督行人检测系统，基于行人的先验信息和通用性目标性置信度，完成自学习检测器的任务，要同时保证模型的准确性和召回率。具体来说，我们在视频行人数据集上，不使用任何标注信息，根据行人的运动信息和一般性目标的显著性信息，通过弱监督学习方法学习一个目标检测器。与其他弱监督检测方法相似，本课题的主要任务是准确且充分的挖掘特定场景中的具有价值的样本，渐进地优化模型和筛选样本。为了实现本课题的研究目标，本课题结合了一般性目标候选区域提取、目标发现、目标增强、标签传播和渐进优化等几个过程。本文研究主要贡献有：

1）结合超像素分层合并候选区域提取方法，我们融合了纹理复杂度来衡量每个区域的优劣。我们在保证保留覆盖目标数量的同时显著较少了整体候选窗口的数量。

2）针对特定场景目标发现，本文提出了使用隐 SVM 模型来实现目标发现的过程，以选择视频中初始化的行人目标。该方法可以获得图像级别的最优解。

3）为了解决初始化样本不是目标级的准确窗口的问题，本文提出使用类似于难反例挖掘的方法使正例样本与难反例样本最大化边界距离，从而达到精选目标候选区域的目的。

4）为了获取更多的样本，我们提出使用帧内标记传播方法来达到增量学习的目的。

1.4 本文的组织结构

第一章，绪论。本章介绍了特定场景下弱监督目标检测的研究背景和研究意义。然后我们总结了当前国内外在目标候选区域提取和弱监督视频目标检测的相关研究工作，同时比较了现有方法研究的优缺点。最后对本文目标候选区域提取与弱监督特定场景下行人检测研究进行了简单的介绍和说明。

第二章，相关技术研究。本章起到了承上启下的作用，主要针对目标候选区域提取、纹理特征表示和弱监督学习方法的概念进行阐述，同时列举了几种相关的方法。

第三章，基于纹理复杂度的冗余区域排序。本章介绍了提出的基于纹理复杂度的冗余区域排序方法。在 PASCAL VOC 2007 数据集上，给出了召回率、候选窗口与目标的重叠比例和候选窗口数量等性能对比，证明了该方法可以在保证覆盖目标的前提下，明显减少候选区域的数量。

第四章，弱监督视频场景行人检测。本章提出了在特定场景下的弱监督行人检测方法。该方法使用了目标发现，目标增强和标签传播的三个渐进的步骤。在学习过程中，使用渐进隐变量模型求解。通过在几个视频行人数据集上的获得的实验结果，表明本方法获得了与监督学习方法可以相比的性能。

第五章，结论与展望。本章主要对本文的研究工作进行总结，讨论尚未解决的问题以及对下一步工作的展开进行讨论。

第二章 相关技术研究

第一章已经简要地介绍了本文的研究背景和意义、相关领域研究综述、主要研究内容和文章结构安排。本章作为上一章的扩展和延伸，将介绍一些本论文提出的相关算法和研究技术，并且该章是第三章和第四章研究工作的基础。

2.1 目标候选区域提取

目标检测一直是计算机视觉中的主流研究方向之一，最近几年在该领域中的进步突飞猛进，很多方法都取得了显著的成功。一般目标检测可以分为三个步骤：1）目标候选区域的提取；2）特征选择和表示；3）分类器的训练，如图 2-1 所示。

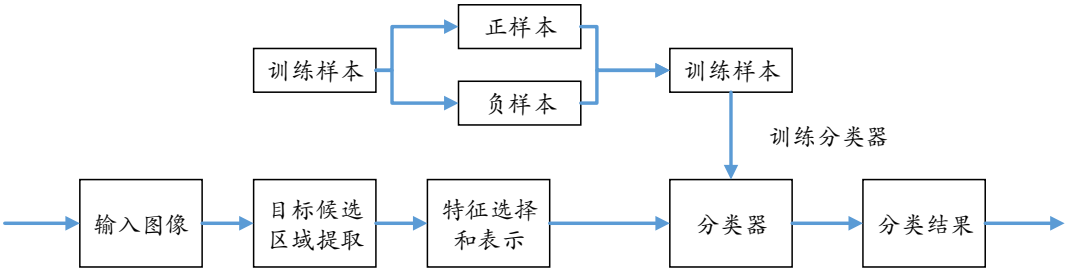


图 2-1 目标检测流程图

目标候选区域提取（Object Proposals）是目标检测中具有优秀检测率的一个保证，它作为算法输入的最前端，确保我们可以发现目标可能所在的位置，如果在这个阶段中目标丢失了，那么再优秀的检测器也不会检测到原始目标。传统的目标候选区域提取方式是使用扫窗策略（Sliding Window Strategy），如图 2-2 所示。扫窗策略是使用不同尺度、不同比例的窗口以不同的步长在目标图像上进行滑动，每一个位置记录为一个候选区域，所以该策略会生成几十万个候选区域，造成严重的冗余。因此，这些冗余的窗口会造成巨大的计算代价，并且在后端分类器使用很复杂的算法时会尤为严重。所以，如何尽可能减少目标候选区域是计算机视觉领域的研究学者们都想解决的一个研究问题。最近的部分研究致力于采用图像分割的方式提取目标候选区域，在图像中首先相似的像素会合并成为一体，之后再通过在这些分割后的结果上添加不同的策略进行合并实现目标候选区域的提取。下面我们分别介绍超像素分割技术和分层区域合并技术。

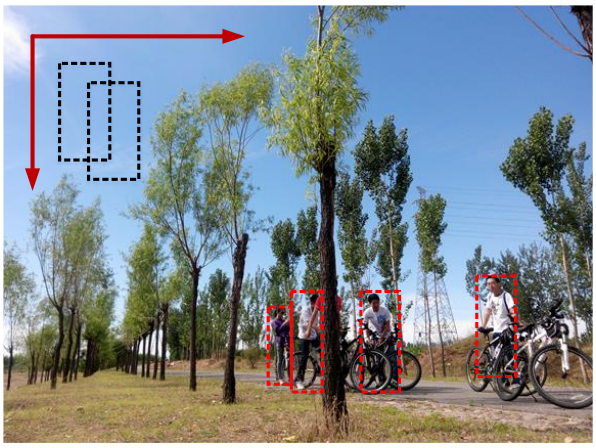


图 2-2 扫窗策略示意图

2.1.1 超像素分割与分层合并

2.1.1.1 超像素分割

超像素（Superpixel）是一种过度分割形成的小区域，如图 2-3 所示。超像素的算法有很多，其中一种产生超像素的方法是运用 Ncut 算法[48]对图像进行初始划分，而后在分割后的每个区域中在运用聚类的算法聚类生成更粗糙的分割结果。

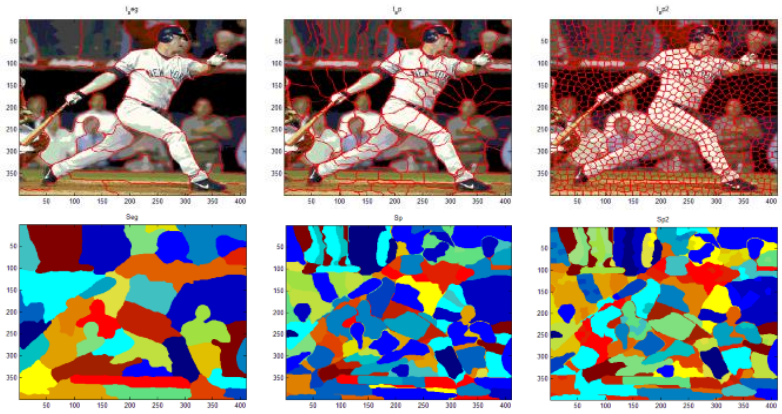


图 2-3 不同尺度超像素图像

我们使用 Felzenszwalb[49]等人提出一种高效的基于图的分割算法，该算法是一种阈值可变的算法，它以邻近的两个区域中所有点中最小的权值作为这两个相邻区域相似度，每次合并之后阈值根据当前区域尺寸重新计算。当分割后的所有区域最终满足特定尺寸和阈值时算法停止。

该方法是一种贪心的图像分割算法，但是它产生的超像素分割结果仍然满足全局的性质。通过构建局部的相邻集合来应用该算法对图像进行分割。算法的运行时间和图的边的数量近乎成线性关系，在实际中速度也非常的快。该方法的一

个重要特性就是它可以在小尺寸的区域中保护图像的细节，而在大的变化区域忽略细节影响。

图割算法是基于图的方法对图像进行分割。$G=(V,E)$ 是一个带 V 个点的无向图，这些点是要待分割的元素集，E 是对应每两个相邻节点之间的连边，每条边都会有一个对应的权重 W，这用来表示两个相邻元素之间的一个非负的不相似性的度量。在图像分割中，V 是图像的像素点，边上的权重是一些被边连接起来的两个像素之间的不相似性度量（例如亮度、颜色、位置或者其他的局部属性）。

在基于图的方法中，一个分割 S 是 V 的一部分（区域），每个部分或是区域都属于分割 S 对应在图中的一个连接的部分 $G'=(V,E')$，这里 $E'\subseteq E$。换句话说，任何一个分割都是在边 E 中的一个子集。有很多方法去测量一个分割的质量，但是通常我们都希望在同一部分的元素都是相似的，而在不同的部分中的元素是不相似的。这就意味着在同一个部分中的相邻的两个节点之间的边上的权值相对比较低，并且在不同部分上的相邻的点之间的边上的权值应该更高。

为了把图分割成不同的区域，我们首先把该分割部分的最小生成树中最大的权值定义成这个分割部分之间的内在差异度量：

$$Int(C)=\max_{e\in MST(C,E)} w(e) \tag{2-1}$$

这里 $C\subseteq V$，$MST(C,E)$ 是该分割部分的最小生成树。

接下来我们把每个分割部分之间相互连接的最小权边定义为这两个分割部分之间的差异：

$$Dif(C_1,C_2)=\min_{v_i\in C_1,v_2\in C_2,(v_i,v_j)\in E} w(v_i,v_j) \tag{2-2}$$

这里 $C_1,C_2\subseteq V$，并且如果在 C_1 和 C_2 之间没有连边，那么他们的差异为 $Dif(C_1,C_2)=\infty$。这样定义组内差异和组间差异在分割中取得了很好的结果。如果在一对分割部分中组间差异大于组内差异，证明他们两个部分确实存在一条边界将他们分开。这个控制的阈值函数我们定义为：

$$D(C_1,C_2)=\begin{cases} true\ if\ (Dif(C_1,C_2)>MInt(C_1,C_2)) \\ false \qquad\qquad otherwise \end{cases} \tag{2-3}$$

$$MInt(C_1,C_2)=\min(Int(C_1)+\tau(C_1),Int(C_2)+\tau(C_2)) \tag{2-4}$$

特别说明的是如果 $|C|=1$，$Int(C)=0$。

图 2-4 快速图割的分割结果

2.1.1.2 分层合并

超像素分层合并是提取图像冗余候选区域的主要方法之一。对于给定图片我们使用一个图像分割和合并的方法来产生冗余区域，例如 Selective Search[2]。它是典型的超像素合并方法，它可以产生具有高定位精度的冗余区域，通常是成千到上万个。基于上一节我们使用基于图的高效图像分割算法得到多张超像素底图，接下来要用这些超像素底图进行分层合并，直至把一整幅图像合并为一个区域为止。为了获得更高的召回率，该分层合并方法使用了各种不同的颜色空间、不同的相似性度量组合和不同的初始化像素块的尺寸。

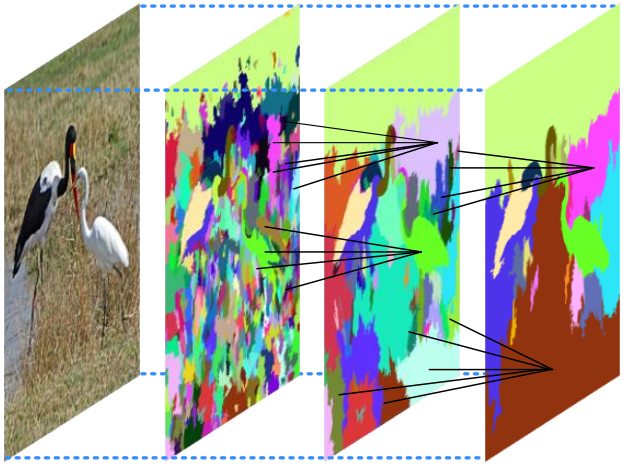


图 2-5 分层合并示意图

在定义相似性的度量中，使用了三种互补的、快速计算的测度度量。其中包括颜色相似性、纹理相似度、尺寸测度和填充区域形状测度。

1、颜色相似度

$S_{colour}(r_i,r_j)$是度量颜色的相似性。对于每个区域我们获得一维的颜色直方图，每个颜色通道被分为 25 个块，这样对于每个区域三个颜色通道会有一个 75 维的颜色直方图向量$C_i=\{c_i^1,...,c_i^n\}$，这里$n=75$。颜色直方图被L_1范数正则化。相似性的度量定义为：

$$S_{colour}(r_i,r_j)=\sum_{k-1}^{n}\min(c_i^k,c_j^k) \tag{2-5}$$

颜色直方图可以根据公式（2-6）通过该分层结构进行有效地传播

$$C_t=\frac{size(r_i)\times C_i+size(r_j)\times C_j}{size(r_i)+size(r_j)} \tag{2-6}$$

同时生成区域的尺度为$size(r_t)=size(r_i)+size(r_j)$。

2、纹理相似度

$S_{texture}(r_i,r_j)$是度量纹理的相似性。使用 fast SIFT-like 特征作为纹理的表示，在每个颜色通道上使用八个方向做高斯微分。对于每个通道上的每个方向提取一个大小为 10 的直方图。所以我们得到一个维度为 240 维的纹理直方图$T_i=\{t_i^1,...,t_i^n\}$，这里$n=240$。纹理直方图被被L_1范数正则化。纹理相似性被定义为：

$$S_{texture}(r_i,r_j)=\sum_{k=1}^{n}\min(t_i^k,t_j^k) \tag{2-7}$$

3、尺度测度

$S_{size}(r_i,r_j)$是尺度测度，它鼓励小的区域先被合并。这会强迫初始化超像素分割结果通过算法合并成尺寸相似的图片块。这是希望它在图片所有的部分创建所有尺度的目标定位。例如，这样可以阻止一个单一的区域逐个吞并所有其他的区域。区域r_i和r_j之间的$S_{size}(r_i,r_j)$定义为：

$$S_{size}(r_i,r_j)=1-\frac{size(r_i)+size(r_j)}{size(im)} \tag{2-8}$$

4、填充区域形状测度

$S_{fill}(r_i,r_j)$是度量区域r_i适合嵌入r_j的程度。这个测度倾向于填补合并中的漏洞：如果r_i包含在r_j内，这样将倾向于合并他们，从而避免在合并的区域中产生空洞。换句话说，如果r_i和r_j的相互连接的部分很小，这样如果他们合并就会产生一个奇怪的区域，并且这个区域不应该合并。BB_{ij}是围绕窗口r_i和r_j的紧窗口。$S_{fill}(r_i,r_j)$定义为：

$$S_{fill}(r_i,r_j)=1-\frac{size(BB_{ij})-size(r_i)-size(r_j)}{size(im)} \tag{2-9}$$

最后把以上四种测度联合起来作为区域r_i和r_j的最后相似度：

$$S(r_i,r_j)=a_1S_{colour}(r_i,r_j)+a_2S_{texture}(r_i,r_j)+ a_3S_{size}(r_i,r_j)+a_4S_{fill}(r_i,r_j) \quad (2\text{-}10)$$

根据最后的相似度指引每个区域完成合并，直至最后合并成为一张整图。

2.2 特征提取算法

特征提取算法是在目标候选区域提取之后的一个关键步骤。提取的特征需要对目标具有良好的表示特性。早期以手工设计的特征为主，最近几年深度特征的良好表示性被大家所认可。本小节将对经典的手工设计特征和深度学习特征进行介绍。

2.1.1 手工设计特征

早期的研究都在设计一种具有尺度不变性、光照不变性和旋转不变性等的手工设计的特征，最成功的两个特征分别为尺度不变性特征变换特征和方向梯度直方图特征。下面两个小节分别对它们进行简单的介绍。

2.1.1.1 尺度不变特征变换

尺度不变特征变换（Scale-Invariant Feature Transform，SIFT）特征是由 Lowe[50] 等人在 1999 年提出的一种图像局部特征，之后又在 2004 年由其本人得以完善。该特征对尺度缩放、旋转、亮度变化保持不变性，对视角变化、放射变换、噪声也一样保持一定程度的稳定性，成为目标表示的一种非常有效的特征。SIFT 特征提取分为四个步骤：①尺度空间极值检测；②关键点搜索与定位；③方向确定；④关键点描述。

1）尺度空间极值检测

算法需要考虑图像在多个尺度下的表现以获知感兴趣物体的最佳尺度。如果图像在不同尺度下都具有相同的关键点，那么在不同尺度的输入图像下就可以使用这些关键点进行匹配，即所谓的尺度不变性。 尺度空间极值检测首先需要构建高斯及高斯差分（Difference of Gaussian，DoG）金字塔，在对 DoG 金字塔进行极值检测，初步确定特征点的位置及所在尺度。

为了得到在不同尺度下的稳定特征点，将图像与不同尺度下的高斯核进行卷积操作，形成图像高斯金字塔。一般选择 4 组，每组有 5 层。下一组的图像由上一组按照隔点降采样得到。DoG 算子是两个不同尺度的高斯卷积核的差分，所以计算方便。

2）关键点搜索与定位

从每组图像中得到 4 幅 DoG 图像，在对中间的两幅 DoG 图像进行极大极小值像素点检测后，可以标记出近似的极大极小值点。

3）方向确定

在特征点附近，创建一个方向收集区域来控制该特征点的影响范围。方向收集区域的大小依赖于它所在的图像的尺度，尺度越大，收集区域越大。在实际计算中，用一个直方图来统计方向收集区域中像素的平均方向。在直方图中，将 360°的方向分为 36 个位（bin），每个位包含 10°。当直方图在某个柱上出现最高峰时，直方图峰值代表了该点邻域内图像梯度的主方向，也就是该关键点的主方向。在梯度方向直方图中，当存在另一个相当于主峰值 80%的峰值时，则将这个方向认为是该关键点的辅方向。

4）关键点描述

在拥有尺度不变性和旋转不变性的特征点后，接下来要为每个特征点创建一个唯一标志，称之为该特征点的 SIFT 描述子（descriptor）。将特征点周围16×16的窗口分解为 16 个4×4的子窗口，在每个4×4的子窗口中，计算出梯度的大小和方向，并用一个 8 位的直方图来统计子窗口的平均方向，这样就可以对每个特征形成一个 128 维的描述子。

2.1.1.2 方向梯度直方图

方向梯度直方图[51]（Histogram of Oriented Gradient，HOG）是由 Dalal 和 Triggs 于 2005 年针对人体目标检测提出的特征描述子。论文中提出了基于 HOG 和 SVM 的人体目标检测算法。它与 Lowe（2004）提出的尺度不变特性变换（SIFT）相似，HOG 特征通过提取局部区域的边缘或梯度信息的分布来表征局部区域内目标的形状。

在提取 HOG 特征的过程中，将8×8个像素点分成一个单元（cell），例如有64×128的训练样本，我们对其进行划分，可以将这个训练样本分为$8\times16=128$个单元。然后再将每相邻的田子形结构的 4 个单元组成一个块（block），通过滑动块得到多组田字形局部区域特征。块一次滑动 8 个像素，所以，一个64×128的训练样本可以得到$7\times15=105$个块。在对其所有像素的梯度方向进行投影，形成每个单元各自的梯度方向直方图。

这里使用的方向位的数量设定为 9，所以每 20°会对应一个位；0°到 180°与 180°到 360°的方向采用对等角相等的方法进行归类划分。然后，再将每个块中 4 个单元的梯度直方图的数据串联起来。由于每个单元的梯度直方图为一个

9 维的向量，所以每个块可以提取一个 36 维向量。再将所有的块（对于 64×128 像素的样本共 105 个块）依次串联起来，便形成了每个图像的 36×105=2780 维的特征。

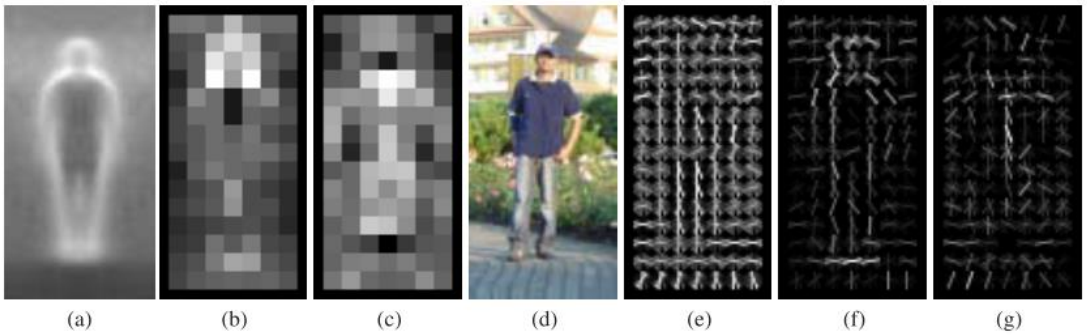


图 2-6 使用梯度直方图的行人检测算法

图 2-6 来自于 HOG 论文中，图 2-6（a）训练样本的平均梯度图像；（b）每个像素显示了以它为中心的块的最大的正 SVM 权重；（c）同样，在负 SVM 权重上的情况；（d）测试图像；（e）计算的 R-HOG（矩形梯度直方图）描述子；（f）用正的 SVM 权重加权了的 R-HOG 描述子；（g）用负的 SVM 权重加权了的 R-HOG 描述子。我们可以看到，在人的头、肩膀、躯干和脚等位置有很多正的响应。

2.2.2 深度学习特征

卷积神经网[52]（Convolutional Neural Network，CNN）最近几年受到了广泛的关注，它有着出色的特征表达能力，在图像分类、目标检测和图像检索等各个领域都表现非常出色。CNN 来源于经典的人工神经网络（Artificial Neural Network，ANN），它利用权值共享策略，降低了网络模型的复杂度，并且减少了权值的数量。在网络的输入是图像时，权值共享的优点显现的更为明显，它使得图像可以直接作为网络的输入，并且显著地降低了待求网络权值的个数。网络的结构一般包含卷积层、池化层、Relu 层和全连接层，如图 2-7 所示。图像的局部感受区域作为最底层输入到网络中，依次经过不同的层。卷积层通过卷积滤波器和感受区域做卷积计算出图像的显著特征；池化层对特征做下采样，降低每一层特征的维度数量；Relu 层可以对特征进行非线性变换；全连接层对特征进行重组和选择。

通过这种逐层操作的结构，逐渐把初始的“底层”特征转化为“高层”特征，特征的表达能力进一步增强，之后在通过简单的模型即可以完成分类任务。

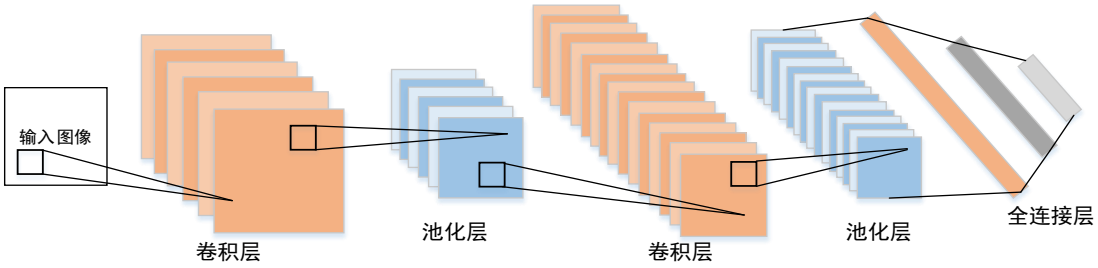


图 2-7 卷积神经网络结构示意图

2.2 弱监督学习算法

早期的机器学习和模式识别算法主要集中于监督学习，主要解决有明确标注的数据学习问题。随着该领域的发展和扩展，弱监督学习逐渐受到了广泛的关注和探索。如今随着互联网的普及，图像、视频、语音和文本数据等越来越容易获取，但是让这些数据具有专家级的标注就显的非常困难，这项工作非常浪费人工成本。对海量规模的数据进行标注，完全是一件不可能完成的事情。在这种背景下，弱监督学习便发挥了更大的作用。弱监督学习是对监督学习算法进行改进，对数据进行部分标注或者对示例包进行标注。它主要可以分为半监督学习、多示例学习和多标签学习等。本节首先介绍监督学习算法，为后文介绍的弱监督学习算法作为基础。后续介绍几种与我们工作相关的弱监督学习算法。

2.2.1 监督学习

全监督学习算法是弱监督和半监督学习方法的基础，如图 2-8 所示，监督学习方法中的数据包含样本和每个样本对应的标签，经典的监督学习方法有支持向量机、Boosting 和 Adaboost。

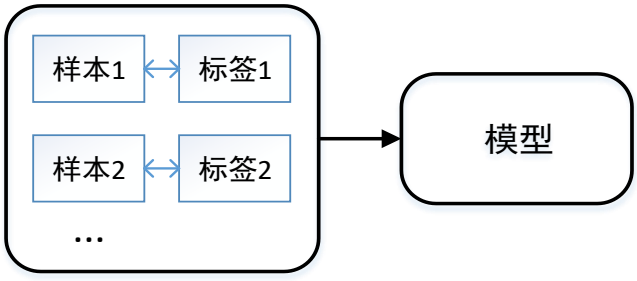


图 2-8 监督学习示意图

2.2.1.1 支持向量机

支持向量机[53]（Support Vector Machine，SVM）是由 Vapnik 等人提出的方法，它是基于最大边界的学习方法。最大边界的同时也就是最小化结构风险，它的学习策略是间隔最大化，所以我们可以通过求解凸二次规划的最优化算法对 SVM 进行求解。

假设在二维空间中有两类样本点，线性支持向量机需要分开两类样本，找到最优的分割平面，我们也称之为超平面。该平面不光要分开两类样本，同时还要使得两类样本之间的间隔最大化。假设样本为$\{x_i, y_i\}$，$i=1,...,N$，其中x_i是第 i 个样本的特征向量，$y_i=\{+1,-1\}$是二分类的类别标签。设定线性超平面的公式为：

$$g(x)=w^T\cdot x_i+b \tag{2-10}$$

式中 w 是线性超平面的法向量，b 是超平面的偏移量。

对于线性可分的情况，可以求得两个线性超平面使得对正例样本有$g_1(x)=w^T\cdot x_i+b\geq 1$，对反例样本有$g_2(x)=w^T\cdot x_i+b\leq 1$。其中两个平行超平面之间的间隔为$2/\|w\|_2$，如果想获得间隔最大化，需要使$\|w\|_2$最小。因此，优化模型可以表示为：

$$\begin{aligned}&\min_{w,b}\frac{1}{2}\|w\|^2\\&s.t.\ y_i(w^T\cdot x_i+b)-1\geq 0, i=1,...,N\end{aligned} \tag{2-11}$$

对于线性不可分的情况，则无法求得其法向量 w 和偏移量 b 的解析式。而我们的数据往往都是线性不可分的情况，此时上式是无解的，也就是说不存在一个线性超平面可以将正反例样本全部分对，因此需要改进模型为下面的形式，即：

$$\begin{aligned}&\min_{w,b}\frac{1}{2}\|w\|^2+c\sum_i\xi_i\\&s.t.\ y_i(w^T\cdot x_i+b)\geq 1-\xi_i, i=1,...,N\\&\xi_i\geq 0,\forall i\end{aligned} \tag{2-12}$$

2.2.1.2 Adaboost

Adaboost[54]算法是将弱学习算法提升成为强学习算法的一种方法。其中的弱学习算法也就是我们所说的弱分类器，强学习算法也就是强分类器。它采用的机制是加权投票，将具有投票权的弱分类器线性组合起来形成强分类器。Adaboost

是一种迭代式的算法，每次迭代选择一个最好的弱分类器，最后将所有迭代中选择出的弱分类器进行线性加权组合，因此形成一个强分类器。Adaboost 在每次迭代中，对每个训练样本赋予一个权重，选择分类误差最小的弱分类器构建强分类器，并且调整每个样本的权重。主要是根据更加重视被误分的样本的原则，被误分的样本获得较大的权重，这样随着迭代次数的增加，算法会更加关注难以分类的样本上面。最后将每次迭代选择出的弱分类器进行加权组合，形成强分类器。值得说明的是一个弱分类器对应一个特征，在选择哪些弱分类器形成的强分类器的同时也完成了特征选择功能。

2.2.2 多示例学习

多示例学习（Multiple-Instance Learning，MIL）[55]是监督分类学习方法的一般化，它的类别标记不再是针对每个样本，而是示例包（bag）的标记，每个示例包包含多个样本，同时该示例包具有一个标签。针对二值分类问题{-1,1}，如果一个示例包中每个样本都是反例样本，则这个示例包的标记为反例示例包（标记为-1）；如果一个示例包中至少包含一个正例样本，则这个示例包标记为正例示例包（标记为 1）。在根据内容的图像索引中，一个图像可以被视为一个示例包，它包含一些局部图像区域，这些区域可能具有这幅图像的表示性物体，比如说一幅牧场的图像可能包含牛、羊、草地等。之前的方法需要对每个图像中这些样本进行详细的标注，而在多示例学习中，我们只需标注整幅图像。这样图像级别的标注相比于图像区域级别的标注减少了很多成本，所以多示例学习问题未来会有着更广阔的应用前景。

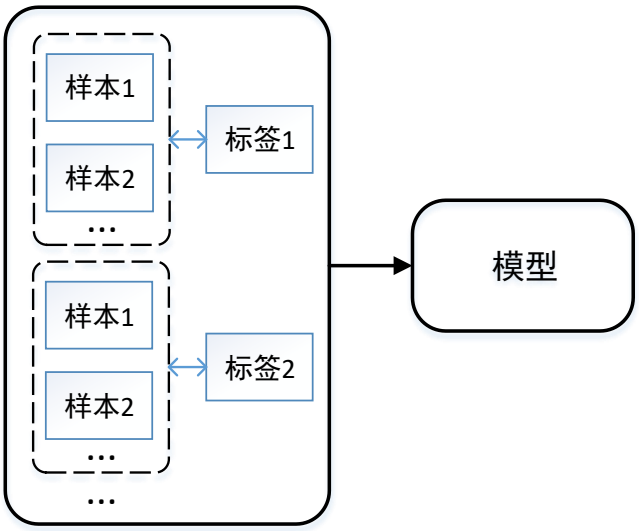


图 2-9 多示例学习示意图

下面介绍两种经典的基于支持向量的多示例学习方法：最大化样本间隔的多示例学习方法 mi-SVM 和最大化示例包间隔的多示例学习方法 MI-SVM。下面我们分别对这两种方法进行介绍。

2.2.2.1 mi-SVM

mi-SVM 是最大化样本间隔的多示例学习方法。它的原始形式可以写成如下：

$$\min_{\{y_i\}} \min_{w,b,\xi} \frac{1}{2}\| w\|^2 + C\sum_I \xi_I \qquad (2\text{-}13)$$

$$s.t.\ \ \forall i : y_i(\langle w, x_i\rangle + b) \geq 1-\xi_I,\ \xi_I \geq 0,\ y_i \in \{-1,1\}$$

在标准的分类设定中，训练样本 x_i 的标签 y_i 是给定的。而在公式 2-13 中不属于任何负包的样本 x_i 的标签 y_i 被视为未知的整数变量。在 mi-SVM 中最大化软间隔标准需要同时优化可能的标签分配和优化超平面。我们寻找一个 MI-separating 线性判别式使得每个正包中至少有一个样本位于正半空间中，而所有负包中的所有负样本都位于负半空间。同时，我们希望根据公式（2-13）对正包中的样本估计标签实现（完整的）数据集达到最大边界。mi-SVM 算法的求解过程如下表所示。

表 2-1 mi-SVM 优化启发式伪码

初始化：$y_i = Y_I$ for $i \in I$

REPEAT:

使用具有估计标签的数据集计算 SVM 的解 w 和 b

对于所有正例包中的 x_i 计算器输出 $f_i = \langle w, x_i\rangle + b$

对每一个 $i \in I, Y_I = 1$，设定 $y_i = \mathrm{sgn}(f_i)$

FOR （每个正包 B_I）：

IF （$\sum_{i\in I}(1+y_i)/2 == 0$）：

计算 $i^* = \arg\max_{i\in I} f_i$

设定 $y_{i^*} = 1$

END

END

WHILE （估计标签不再改变）

输出：(w, b)

2.2.2.2 MI-SVM

MI-SVM 是最大化示例包间隔的多示例学习方法。该方法中对于包的标签预测形式为 $\hat{Y}_I = \text{sgn}\max_{i\in I}(\langle w, x_i\rangle + b)$。对于一个正包，边缘由分值最高的样本所决定。对于以样本为中心的 mi-SVM 公式，正包中的每个样本的边界都是很重要的，尽管可以自由地设置其标签变量已获得最大边界。而在以包为中心的公式中，每个正包中只有一个样本重要，因为它将决定包的边缘。一旦确定了“witness”样本，其他正包中的样本的位置对于分类边界的贡献就毫无意义了。使用上述包边界的概念，定义一个 MIL 版本的软边界分类器：

$$\min_{w,b,\xi} \frac{1}{2}\| w\|^2 + C\sum_I \xi_I \tag{2-14}$$

$$s.t.\ \ \forall I : Y_I \max_{i\in I}(\langle w, x_i\rangle + b) \geq 1 - \xi_I,\ \xi_I \geq 0.$$

算法通过迭代过程实现，分步优化分类器和优化训练样本。MI-SVM 算法如表 2-2 所示。

表 2-2 MI-SVM 优化启发式伪码

初始化：对于每个正例包 B_I： $x_I = \sum_{i\in I} x_i / \| I \|$
REPEAT： 对于具有正例样本 $\{x_I : Y_I = 1\}$ 的数据集，计算二次规划的解求 w 和 b 计算所有正例包中的所有 x_i 输出 $f_i = \langle w, x_i\rangle + b$ 设定 $x_I = x_{s(I)}$， $s(I) = \arg\max_{i\in I} f_i$，对于每一个 I， $Y_I = 1$ WHILE （选择器变量 $s(I)$ 不再更新）
输出：(w, b)

2.2.3 半监督学习

半监督学习（Semi-Supervised Learning）顾名思义是介于无监督学习和监督学习之间。事实上，大多数半监督学习策略是基于无监督或监督学习方法进行扩展。我们定义训练样本集 $D_l = \{(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)\}$，这 l 个样本的类别标记已知，成为有标记的样本；此外，还有 u 个未标记的样本 $D_u = \{x_{l+1}, x_{l+2}, ..., x_{l+u}\}$，其中 $l \ll u$。当我们使用传统的监督学习方法，直接使用 D_l 数据训练模型，则其

他数据被浪费掉了，并且学的的模型由于数据不足所以泛化能力较差，所以产生了半监督学习问题。半监督学习方法可以不依赖外界交互、自动地利用未标记的样本提升学习性能。

在我们的实际应用中，许多问题都往往很容易地可以收集到大量的未标记的样本，这些未标记的数据和我们标记的数据是满足同源同分布的关系，因此这使得把这些数据利用起来提升模型的性能是非常可行的。在利用这些未标记的信息之前，我们需要做一些必要的假设来揭示数据分布信息和类别标记的相互关联。1）聚类假设（Cluster Assumption），它假设这些数据存在簇的结构，同一个簇属于同一个类别。2）流形假设（Manifold Assumption），即假设数据分布在一个流形结构上，邻近的样本拥有相似的输出值。“邻近”的程度常常用“相似”程度来刻画，因此流形假设可以看作是聚类假设的推广。两者的本质都是“相似的样本拥有相似的输出”这个基本假设。

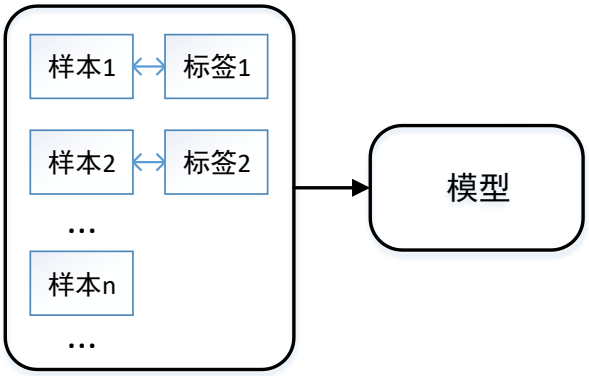


图 2-10 半监督学习示意图

2.3 本章小结

本章主要介绍了目标检测中的几个关键步骤，其中包括目标候选区域提取、特征提取和分类学习算法。

候选区域提取主要涉及图像的超像素分割和分层合并策略，该方法与第三章的研究有着直接的联系，我们在第三章对这一类候选区域提取方法做了一些改进，减少了冗余窗口的数量。本章介绍的内容是第三章冗余区域产生的预备知识，可以更好的理解基于分割的候选区域提取方法。

特征提取方法我们只是简单介绍了基于手工设计的特征和基于深度学习的特征。手工设计的特征介绍了尺度不变性特征变换特征和方向梯度直方图特征，

然后我们简单介绍了卷积神经网络。在第四章的研究中我们用到了方向梯度直方图特征。

弱监督学习方法是我们研究特定场景弱监督行人目标检测的一个重要部分，我们介绍了最大化样本间隔的多示例学习方法 mi-SVM 和最大化示例包间隔的多示例学习方法 MI-SVM。

第三章 基于纹理复杂度的冗余区域排序

上一章介绍了一些目标候选区域的基础知识和弱监督学习的几种算法。本章将介绍本文提出的基于纹理复杂度的冗余区域排序算法和具体实现步骤。

3.1 问题描述

对于视觉目标检测，目标定位是第一个步骤，同时也是最重要的一个步骤。传统的方法使用滑动窗口策略[51][56][57]来定位目标位置，这会倾向于产生百万级别的候选窗口。在随后的检测步骤中对这些窗口的分类是及其耗费计算资源的，特别是当使用复杂特征或复杂的分类方法使用时。最近，已经研究了替代方式，即目标候选区域方法，这可以提高目标的定位效率。目标候选区域方法往往产生比滑动窗口策略少的多的候选窗口，大约会少 2 个数量级，这将毫无疑问有助于提高计算效率，以及保持目标检测速率。在最近的检测候选区域提取方法中，Hosang 等人[58]对其重复性和召回率进行了比较。Selective Search[2][3]和 EdgeBoxes[11]都被认为是具有高召回率和高效率的两种方法。一方面，Selective Search 是典型的超像素合并方法，其产生具有高定位精度的标准相似的区域。但是它往往会产生数万个窗口，并导致严重的冗余问题。为了缓解这个问题，由于对于每个区域没有对象性的策略以准确的作为目标的置信度，通常使用伪随机算法来选择最终的候选窗口。另一方面，EdgeBoxes 是一种典型的对象性置信度的策略方法，它假定对象通常拥有更完整的轮廓，并且使用滑动窗口策略来定位具有完整轮廓的区域作为目标候选区域。使用完整的轮廓有助于减少候选区域的冗余，然而，稀疏的滑动窗口导致宽高比和定位精度的损失。以这两种方法的互补性激发我们将它们集成在一起去生成高准确性和高可信度的目标候选区域。

我们提出了一种基于纹理复杂度的冗余区域排序（Texture Complexity based Redundant Regions Ranking，TCR）策略，用于提取目标候选区域。我们的方法是首先使用超像素分层合并的方式产生冗余区域，然后使用完整轮廓数和局部二值模式（LBP）熵计算每个区域的纹理复杂度（TC）得分。由于这些冗余区域都是基于颜色信息生成的，所以我们提出使用纹理复杂度信息作为互补。通过 TC 得分的排序，保留该区域的召回，并且显著减少候选区域的数量。

3.2 方法概述

为了充分挖掘超像素合并和对象性度量的互补性，提出一种策略，即基于纹理复杂度的冗余区域排序用于目标区域候选。TCR 算法的概要如图 3-1 所示。对

于输入图像，使用颜色分割方法和分层超像素合并的过程（3.2.1 节）生成成千上万个冗余区域。然后计算纹理复杂性（TC）分数以衡量每个区域作为对象的置信度（3.2.2 节）。基于 TC 得分的排序可以减少候选区域的冗余（3.2.3 节）。

我们方法的创新之处总结如下：

1、提出一个 TC 分数的对象性衡量指标，其包含完整轮廓数和 LBP 熵。

2、区域候选过程中集成超像素合并策略和对象性衡量策略。使用 Selective Search 生成冗余区域，并使用它 TC 分数对这些区域排序。

3、与其他超像素合并方法组合提升，无需额外参数调整。

3.2.1 冗余区域的产生

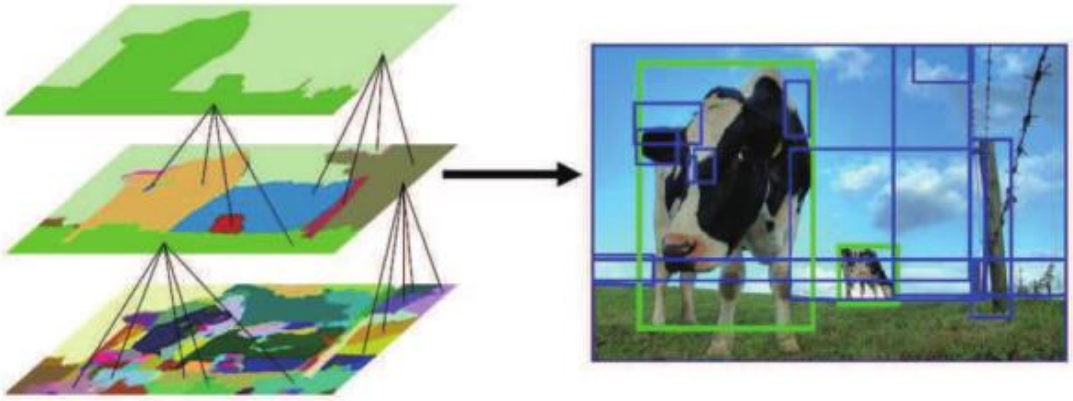


图 3-1 Selective Search 的分层结构

给定一幅图像，使用图像分割方法和分层区域合并过程生成冗余区域，例如 Selective Search[2][3]。图像分割过程首先使用图割算法将图像初始化为彩色均匀的像素块，即超像素。然后执行合并过程，直到整个变为单个区域，如图 3-1 所示。当每两个团块从分层合并结构中合并时，释放当前区域。为了获得高的召回率，使用了多种颜色空间、不同的相似度度量组合和初始化像素块大小，这样做可以增加窗口的数量，使得候选窗口不会丢失任何目标。具体来说，使用了五个颜色空间：HSV、Lab、归一化 RGB 的 RG 通道加上亮度、来自 HSV 的 Hue 通道 H 和亮度。四个相似性度量是颜色、纹理、大小和面积。

因为上述过程包括多种合并过程，所以它生成的区域是极其冗余的。对于自然场景图像会产生数万个候选区域。一些窗口彼此高度重叠，并且一个对象可以覆盖多次。据我们所知，这种基于超像素合并的方法不能测量区域的置信度。为了减少区域数量，使用了伪随机算法，这毫无疑问会损害目标的召回率。

3.2.2 目标候选区域的纹理复杂性

基于以上算法，我们通过超像素的分层合并完成了冗余区域的产生，提出了一种新的算法来给冗余区域添加新的度量，从而来衡量该区域是否为一个可能的目标。通过该度量我们可以对产生的冗余区域进行排序。由于冗余区域产生主要基于颜色线索，所以使用纹理复杂性来互补颜色线索所产生的冗余区域。

3.2.1.1 目标候选区域中完整的轮廓数量

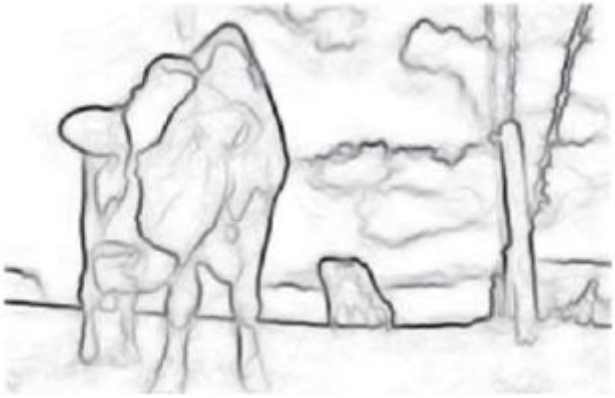


图 3-2 基于结构森林方法获得的边缘图

基于结构森林的方法[59]，可以获得图像的边缘图如图 3-2 所示。一个完成的轮廓是带有相似方向的相邻边缘点的集合。完全包含在一个候选窗口的中完整轮廓的数量大体可以表示这个窗口是否有目标的存在。我们定义 $X=\{x_n=(m_n,o_n)\}_{n=1}^{W\times H}$ 表示一幅大小为 $W\times H$ 的图片边缘图，这里 m_n 和 o_n 分别表示像素点的边缘的强度和方向。假如在一幅图像中边缘组的集合是 $S=\{s_i\}$，在候选框中的边缘组集合就是 $S_b\subset S$。在候选框中完整轮廓的数量计算公式为

$$w_e=\frac{\sum_{s_i\in S_b} f(s_i)m_i}{2(h_b+w_b)^{\kappa}} \tag{3-1}$$

m_i 是在边缘组 s_i 中所有边缘 p 的大小 m_p 的和，h_b 和 w_b 是候选框的宽度和高度。窗口的周长起到归一化的作用，并且 κ 是一个参数。起到抵消更大的窗口包含更多的轮廓的偏置作用。权值函数 $f(s_i)$ 定义为：

$$f(s_i)=\begin{cases}1-\max\limits_{P}\prod\limits_{j=1}^{|P|-1}a(\bar{o}_j,\bar{o}_{j+1}) & if\ \ s_i\in S_b\\ 0 & else\end{cases} \tag{3-2}$$

这里 $a(\bar{o}_j,\bar{o}_{j+1})$ 是轮廓 t_j 和 t_{j+1} 平均方向的亲和度， P 是长度是 $|P|$ 的有序路径。一旦一条完整的轮廓重叠窗口，那么它基本不可能是目标的一部分。

在预处理的边缘图基础上，通过公式 3-2 就算每个窗口的得分。更高的得分说明窗口中包含一个目标，也是对应的在窗口中有更多的完整轮廓。

3.2.1.2 目标候选区域中 LBP 熵

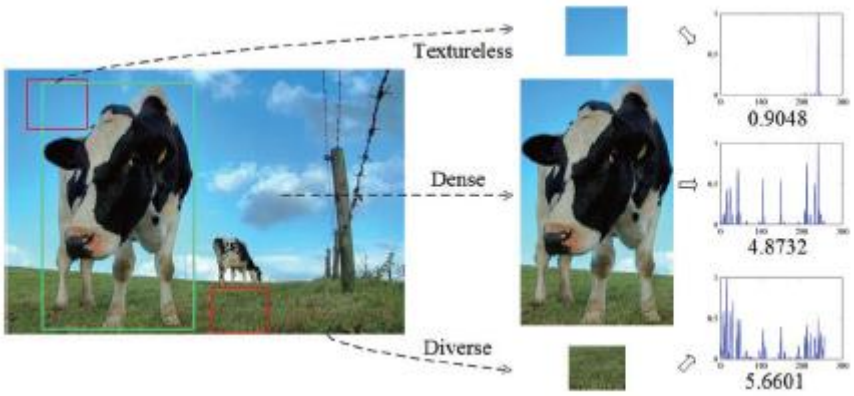


图 3-3 LBP 直方图和熵

纹理特征使用 LBP[60]（Local Binary Pattern，局部二值模式）特征进行表示，它是一种用来描述图像局部纹理特征的很好的算子，并且具有旋转不变性和灰度不变性等显著的优点。

基于我们提出的假设，目标区域通常有着密集的纹理特征，然而背景区域通常是无纹理的或者是多样的，如图 3-3 所示。所以 LBP 熵能够捕获目标与非目标之间的差异性。最后将 LBP 特征可视化如图 3-4 所示。

（a）原图像　　（b）LBP 特征图

图 3-4 LBP 特征可视化

LBP 在 3×3 的图像块上计算如图 3-5 所示。原始的 LBP 算子定义为：在 3×3 的窗口内，以窗口中心像素为阈值，将相邻的 8 个像素的灰度值与其进行

比较，若周围像素值大于中心像素值，则该像素点的位置被标记为 1，否则为 0。这样，3×3 邻域内的 8 个点经比较可产生 8 位二进制数（通常转换为十进制数即 LBP 码，共 256 种），即得到该窗口中心像素点的 LBP 值，并用这个值来反映该区域的纹理信息。

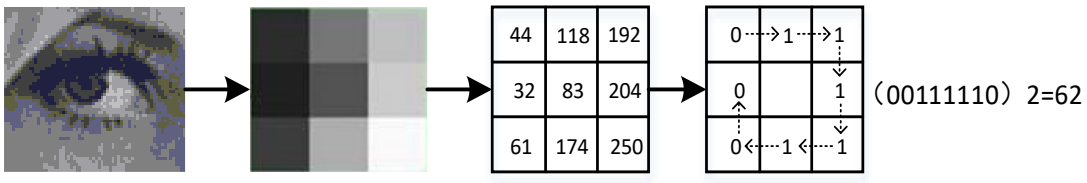


图 3-5 LBP 计算流程图

特别说明 LBP 熵的计算公式是：

$$w_t = -\sum_{i=0}^{255} p_i \log p_i \tag{3-3}$$

这里 p_i 是模式的概率，计算公式如下：

$$p_i = \frac{N_i}{\sum N_i} = \frac{N_i}{h_b + w_b} \tag{3-4}$$

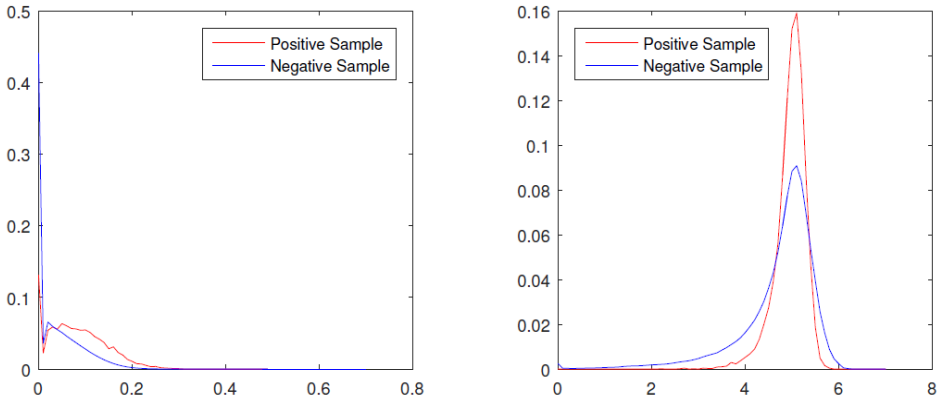
这里 N_i 是第 i 个模式的数量，h_b 和 w_b 分别是区域的宽度和高度。像图 3-3 中所示，一个目标区域的 LBP 熵既不会太大也不会太小。

计算 LBP 熵是非常高效的，因为在预处理的图片特征图上，区域的 LBP 特征非常容易获取，同时 LBP 熵的计算没有尺度的校准。

3.2.3 基于冗余区域的纹理复杂性排序

在 PASCAL VOC2007 数据训练集上，计算了完整区域轮廓的数量和 LBP 熵如图 3-6 所示。根据图 3-6（a），目标候选是在召回率和窗口数量之间的一个权衡。伴随着完整轮廓数量阈值的增加，冗余窗口的数量也在减少，但是这样正样本的召回率就会下降。为了获得一个高的召回率，会要求使用一个很小的阈值。然而，这意味着会有更高的误报率。我们结合 LBP 熵和完整轮廓数量来减少误报率。根据图 3-6（b），正例样本 LBP 熵的方差比反例样本的方差要小。因此我们提出一个 LBP 熵的门函数：

$$g(w_t) = \begin{cases} 1 & if \quad w_t \in (T_{ml}, T_{mr}) \\ 0.5 & if \quad w_t \in (T_l, T_{ml}] \cup [T_{mr}, T_r), \\ 0 & else \end{cases} \tag{3-5}$$



（a）完整轮廓的数量　　（b）LBP熵

图 3-6 纹理复杂性表示

这里T_l，T_{ml}，T_{mr}和T_r是四个阈值。如果一个窗口的 LBP 熵接近分布的峰值，则这个窗口就会被提取出来。这意味着它更可能是一个对象。我们删除具有太大或者太小的 LBP 熵值的候选框，这些候选框通常是无纹理或背景区域。利用等式（3-1）中的完整轮廓数和等式（3-3）中的 LBP 熵，一个区域的组合 TC 分数被定义为

$$o = w_e \cdot g(w_t). \tag{3-6}$$

我们使用 TC 分数对冗余区域进行排序以减少目标候选区域的数量。

3.3 实验

3.3.1 数据集和度量标准

数据集：我们选择在 PASCAL VOC 2007 数据集[61]上评估我们的方法。这个数据集包含了 4501 幅训练图像，2510 幅验证图像和 4952 幅测试图像。我们在训练数据集上获取我们的 TC 分数的参数，在验证数据集上显示我们 TCR 方法的效率，并且在测试集上将我们的方法与现有最好技术进行比较。

评估程序：遵循与其他方法[11]相同的评估程序，使用召回率、候选窗口的数量和候选窗口与目标的重叠比例（Intersection over Union，IoU）。

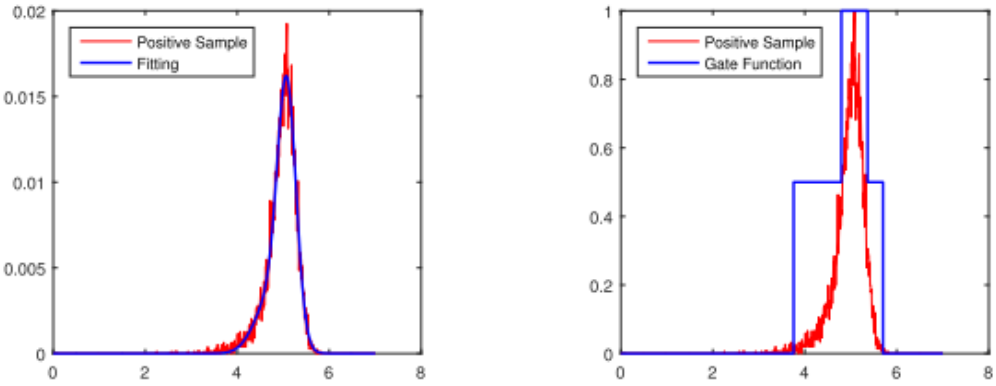
召回率：只有获得更高的召回率，分类器才更有可能获得高检测精度。如果一个目标在目标候选区域提取过程中丢失，分类器将无法再次检测到这个对象。

候选窗口数量：少量的候选框才能保证接下来分类器的工作效率。

IoU：较大的 IoU 意味着更准确的定位，使得接下来的特征提取更为准确。更好的目标候选区域提取方法是在保证召回率的前提下，使用更少的候选窗

口数量，同时具有更大的 IoU。三个常用的实验设置：在给定 IoU 的前提下，对比召回率和候选窗口的数量；在给定候选窗口的数量前提下，对比召回率和 IoU；在给定召回率和 IoU 的前提下，最小化候选窗口的数量。

参数设定方面，在 TC 分数的轮廓数项中，我们设置大于 1 以拒绝大窗口，其余设置和原始方法[11]设定一致。在 LBP 熵项中，我们从 PASCAL VOC 2007 训练集中获得参数。有正样本分布的两边具有不同的斜率，因此使用两项高斯来拟合分布，如图 7（a）所示。通过实验确定高斯参数，$\mu_1 = 5.080$，$\delta_1 = 0.285$，$\mu_2 = 4.728$，$\delta_2 = 0.486$，我们设定 $T_L = \mu_2 - 2\delta_2$，$T_{M1} = \mu_1 - \delta_1$，$T_{M2} = \mu_1 + \delta_1$，$T_R = \mu_2 + 2\delta_2$。门函数如图 7（b）所示。

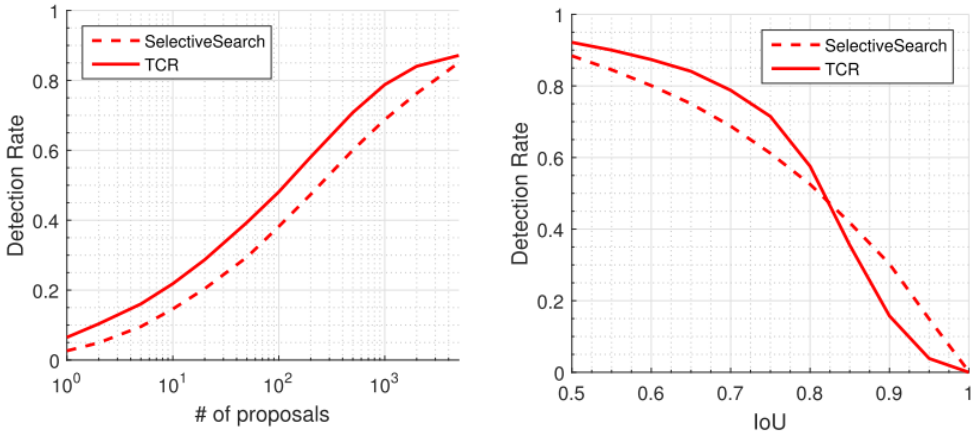


（a）高斯拟合　　（b）门函数

图 3-7 计算 LBP 熵的参数

3.3.2 和基线的性能对比

Selective Search[2][3]用作我们的 TCR 方法的基线。给定一个窗口，可以使用公式 3-6 计算 TC 分数。我们在 PASCAL VOC 验证数据集上评估 TCR 的效率。图 3-8 显示了 TCR 方法和基线之间的对比。



（a）召回率和窗口数量　　（b）召回率和 IoU

图 3-8 和基线对比结果

当 IoU 为 0.7 时，召回率和候选窗口数量曲线如图 3-8（a）所示。可以看出，当时用 100 或 1000 个窗口时，TCR 显著提高了召回率超过 10%。召回率被提升到了 0.87，而 Selective Search 只有 0.85。此外，TCR 只需 720 个检测窗口就可以达到 75%的召回率，而 Selective Search 需要 1777 个检测窗口。

当使用 1000 个检测候选窗口时，召回与 IoU 曲线如图 3-8（b）所示。在图 8(b)中，可以看出，当 IoU 在 0.5 到 0.81 的范围内时，TCR 优于 Selective Search。当 IoU 大于 0.81 时，TCR 的召回率低于 Selective Search，这是由 TCR 还采用了非极大值抑制（NMS）过程。然而，大于 0.5 的 IoU 通常对于检测和分类任务已经足够好了，从图 3-8（b）中可以得出结论，TCR 在召回与 IoU 方面都优于基线。

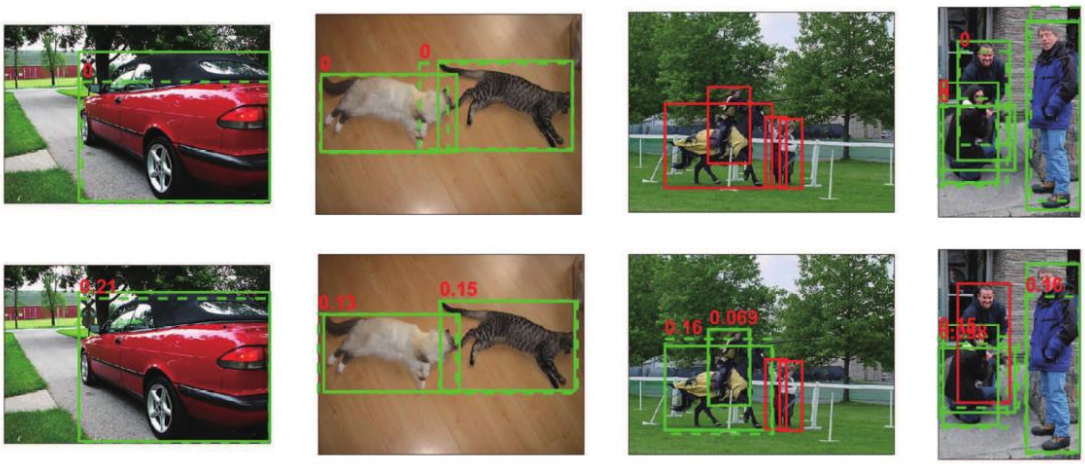


图 3-9 PASCAL VOC 2007 验证数据集目标候选窗口示例

目标候选区域的示例在图 3-9 中给出。在 IoU0.7 的情况下，选择 1000 个检测窗口来说明和真值样本所匹配的结果。从前两列的结果中可以表明我们的 TCR 可以进一步提供更加准确的位置候选窗口。第三列表明 TCR 把潜在的目标窗口寻找出来了。最后一列表明我们的方法错过了一个对象，即穿着黑衣服的男人。这种检测丢失的原因是因为它的 LBP 的分布是规则的，由于它的熵太小，所以它是无效的。

3.3.3 和其他技术对比

我们将 TCR 与最近的研究方法做比较，包括 Objectness[9]、MCG[7]、CPMC[1]、BING[10]、Endres[63]、Rigor[64]、RandomizedPrims[5]、Rahtu[65]、Rantalankila[6]、Selective Search[2][3]、complexity-adaptive（CA）[8]和 EdgeBoxes[11]。这些比较的结果由 Hosang 等人[58]提供，并且使用结构化边缘检测工具箱 V3.0[11]生成曲线。

图 3-10 中说明了召回与候选窗口数量之间的关系，我们比较了最近的几种方法，使用的 IoU 阈值为 0.5、0.6 和 0.7。红色的曲线显示了 TCR 的召回性能。从图 10 可以看出，TCR 方法的召回优于当前最近水平，特别是在 IoU=0.7 时。召回与 IoU 曲线如图 11 所示。候选窗口数量分别设置为 100、500 和 1000。当给定 100 个候选窗口时，从 0.5 到 0.75 变化的 IoU 中，Endres、CPMC 和 MCG 性能略好于 TCR。然而，当给定 500 和 1000 个候选窗口时，TCR 实现了最好的召回率。

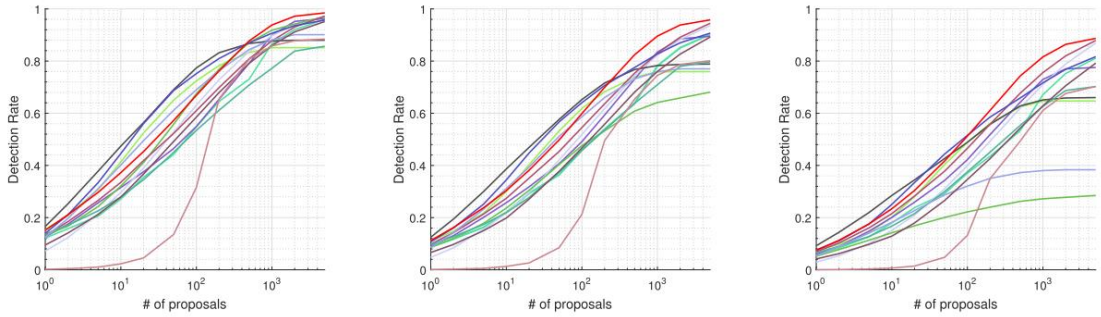
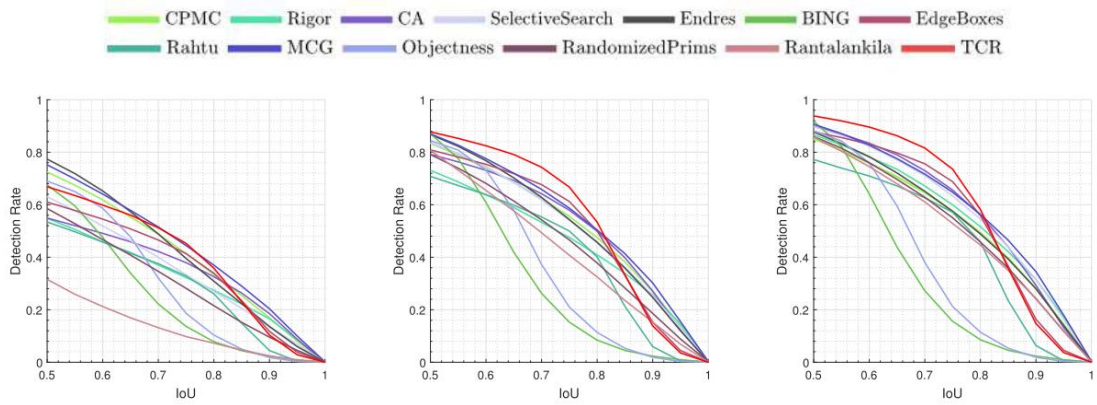


图 3-10 召回率与窗口数量对比



(a) 100 proposals per image (b) 500 proposals per image (c) 1000 proposals per image

图 3-11 召回率和 IoU 对比

在表 3-1 中，我们在 25%、50%和 75%的召回率和 IoU=0.7 条件下比较了每种方法所要的候选窗口数量。可以看出，当使用数千个窗口时，TCR 方法具有 0.89 的最好召回率。TCR 只需要 655 个检测窗口就可以实现 75%的召回率，这是所有比较方法中最少的数量。TCR 仅使用 12 和 91 个检测窗口就可以分别实现 25%和 50%的召回率，这在所有对比方法中都是足够少的窗口数量。在表 3-1 中可以看出，TCR 的 AUC 为 0.48，这是在所有比较方法中最高的性能。表 3-1 中的所有比较均证实了 TCR 改善了现有的技术水平。

表 3-1 使用 TCR 与相关方法的性能比较结果

Method	AUC	N@25%	N@50%	N@75%	Recall
BING[10]	0.20	302	-	-	0.28
Rantalankila[6]	0.25	146	520	-	0.70
Objectness[9]	0.27	28	-	-	0.38
RandomizedPrims[5]	0.35	42	358	3204	0.79
Rahtu[65]	0.36	29	310	-	0.70
Rigor[64]	0.38	25	367	1961	0.81
Selective Search[2][3]	0.39	29	210	1416	0.87
CPMC[1]	0.41	17	122	-	0.65
Endres[63]	0.44	**07**	122	-	0.66
MCG[7]	0.46	10	**86**	1562	0.82
EdgeBoxes[11]	0.47	12	96	6558	0.88
TCR(our approach)	0.48	12	91	**655**	0.89

3.4 本章小结

目标候选方法可以把数百万的目标候选窗口减少至数千个。我们的动机就是将颜色和纹理复杂性信息适当整合可以有助于提升目标候选区域的质量。为了完全整合基于颜色的超像素合并方法的精度定位和基于对象性衡量策略方法的排序置信度，我们提出了一种基于纹理复杂度的冗余区域排序（TCR）策略，进一步提高目标候选区域性能。Selective Search 用于输出冗余区域，并且计算 TC 分数以测量目标区域的置信度。TC 分数由两项组成，完整轮廓数和 LBP 熵，它们都可以使用预先计算的边缘和 LBP 图有效地计算。通过一个门函数，将两项融合在一起以减少区域的冗余，这基本上提高了整个目标检测系统的准确性和效率。此外，TCR 很容易扩展到一些其他区域生成的方法，并较少目标候选区域的数量。

第四章 弱监督特定视频场景行人检测

4.1 问题描述

随着监控摄像机的广泛的使用，对自动检测目标（例如行人）的需求量已经大幅度的增加。最近的方法[56][57][4][66]在图像中的目标检测问题上已经取得了令人鼓舞的进展。然而，它们在视频场景中的性能受到以下几个主要原因的限制：1）对于不同场景下的检测器的监督学习需要重复的人力；2）离线训练的检测器通常会随着场景和相机的变换而退化；3）场景中包含的特定的线索，包括目标的分辨率、遮挡和背景的结构都没有被利用到检测器的学习中[59]。学习特定场景的检测器已经被越来越多地研究[67][38][68]，这个检测器的目的在于通过并入特定场景的判别信息来在视频场景中建模目标。为了在较少的人为监督下学习特定场景的检测器，通常使用迁移学习和半监督学习[67][38][68]。迁移学习使预先训练的检测器适应新的特定域，这样就减少了对标记信息的需要并且提高了检测器的性能[69][44][45]。半监督学习通过使用少量的具有标记的样本来初始化训练检测器来节省人工成本，并通过扩展样本域来递增地改进检测器[23][38][71]。然而，当目标域中的对象和源域中的对象出现明显的差异时，迁移学习就会受到挑战；而半监督模型在给定噪声或不相关样本的情况下可能会偏离预期的目标[38]。最重要的是这两种方法都需要部分目标级的标注，因此，不能消除人的监督。

作为一个有前途的方向，最近无监督的视频目标发现技术[39][72][47]已经有了显著的改善，这应该打破了在实际应用中的自学习的瓶颈。本文讨论了在特定和动态变化的场景（例如城市广场）中以完全无监督的方式下建立的自学习行人检测器的可能性，我们需要给定一段视频序列一些附加的从网上随机收集到的反例样本，在这段视频中行人是主要移动的目标，如图 4-1 所示。自学习问题被分解为三个主要的组成部分：目标发现、目标增强和标签传播。目标发现使用一个隐 SVM 方法[73]实现，通过最小化帧级分类误差，来输出粗糙的模型和标注信息。目标增强的目的在于通过利用空间的正则化来强制目标的位置和减少其不确定性，即通过目标本身来区分目标的部分。标签传播优化一个基于图的目标函数来逐渐发现视频帧中的更难的正例。它也使得自学习框架能够找到更复杂的样本域，例如包括多姿态和多视图目标的流行空间[25]、在具有凸差分（DC）目标函数的差的进行隐模型（Progressive Latent Model，PLM）中形成的三个过程，其以渐进的方式用于凹凸程序优化。

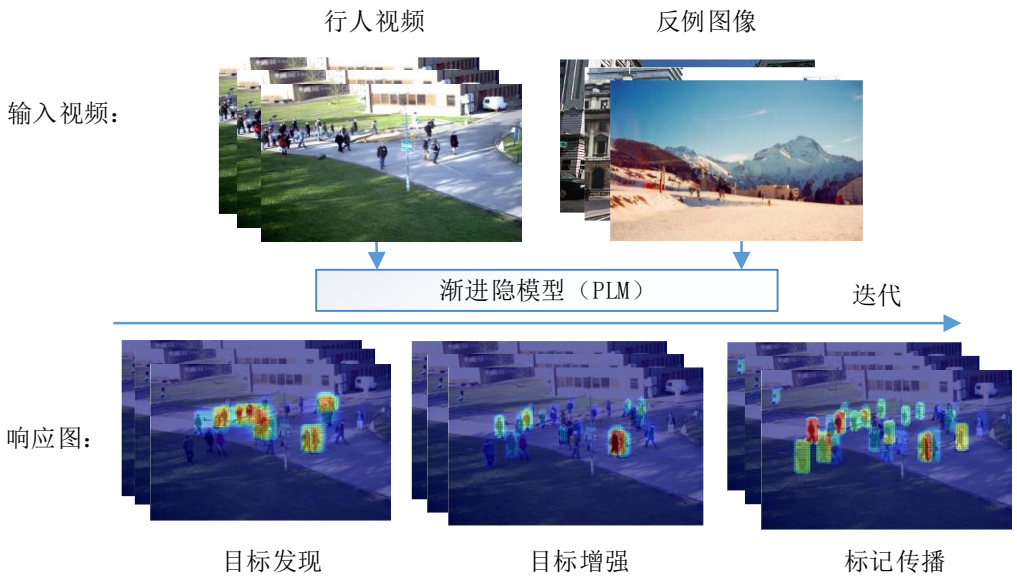


图 4-1 自学习框架示意图

主要贡献包括：（1）提出自学习行人检测框架，其包括目标发现、目标增强和标签传播几个迭代的过程，在（无监督的）目标检测领域中构成了一个新的方向；（2）渐进隐模型使用了时间和空间上的正则化来较少发现的样本的模糊性，这可以解决自学习的稳定性；（3）对 PETS2009、Towncenter、PNN-Parking-Lot2/Pizza、CUHK Square 和 24 小时数据集进行了广泛的实验，以验证提出方法的性能。

4.2 自学习框架

4.2.1 渐进隐模型

在弱监督目标检测设置中，将简单地给出训练样本的位置，而在自学习中，对象位置的标注信息不可以使用。自学习的主要目标是将缺失的标注引导到将对象样本与有噪声对象的候选框区分的解决方案，如图 4-2 所示，a）第一轮学习的响应图（b）红色窗口为发现的候选目标（c）第五轮学习的响应图（d）红色窗口为候选目标，红色窗口为难例样本。

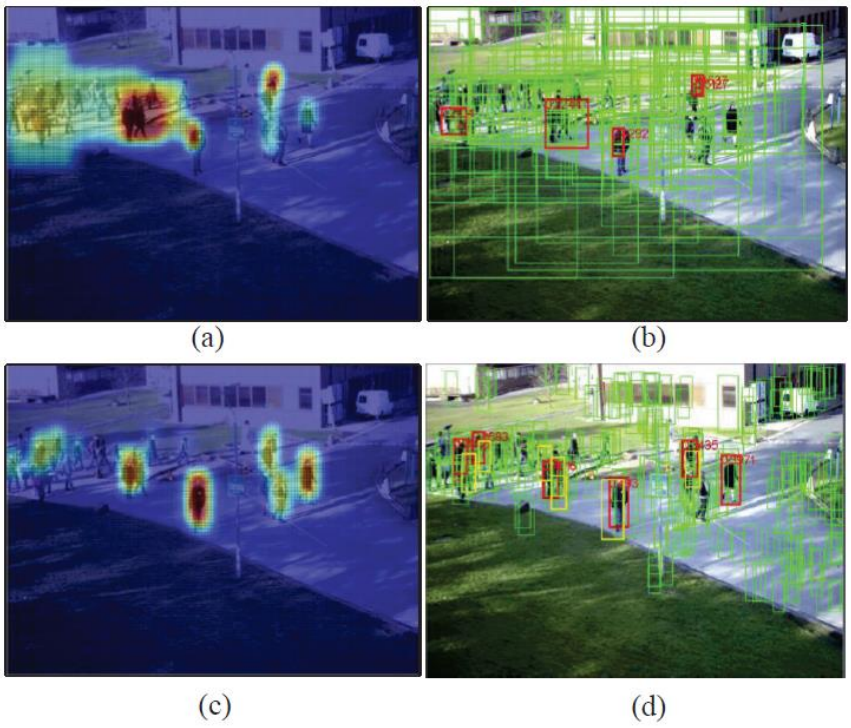


图 4-2 从噪声中发现目标

建模：自学习框架被分解为三个基本过程：目标发现，目标增强和标签传播。给定一组具有明显目标外表轮廓和运动信息的目标候选窗口，如图 4-2（a）和 4-2（b）所示，目标发现阶段旨在从最具有判别能力的视频帧中找到目标的窗口。目标增强发现难反例，这些难反例有助于减少错误的目标局部定位，从而改善整体目标定位。标签传播阶段会挖掘相应目标的更难的样本并且贯穿整个视频，如 4-2（a）和 4-2（b）所示。三个程序迭代进行，直到满足错误率的稳定性准则才终止。

定义 $x \in X$ 为一个视频帧或者是一幅反例图像，$y \in Y$，$Y = \{0,1\}$ 是标签，表示 x 中是否包含一个行人目标。$y = 1$ 表示在这一帧中至少有一个行人，而 $y = 0$ 表示这一帧中没有行人目标或者这是一个反例图像。这个自学习用一个多目标函数来表示，它可以在一个渐进优化程序中联合地决定隐目标 h 和一个隐模型 β。

$$\begin{aligned}\{h^*,\beta^*\} &= \min_{\beta,h} F_{(X,Y)}(\beta,h) \\ &= \min_{\beta,h} F_l(\beta,h) - \lambda F_s(\beta) + \gamma F_g(\beta,h),\end{aligned} \tag{4-1}$$

$F_l(\beta,h)$，$F_s(\beta)$ 和 $F_g(\beta,h)$ 分别是目标发现的目标函数，空间正则化和分数传播，在下边会给出具体定义。λ 和 γ 是正则因子。

目标发现：使用隐 SVM（Latent SVM，LSVM）模型来实现目标发现过程，以选择最佳区分正例帧图像和反例图像的目标的候选区域。

$$\{y^*, h^*, \beta^*\} = \arg\max_{y \in Y, h \in H, \beta} \beta^T \cdot v(x, y, h), \tag{4-2}$$

这里 $v(x, y, h)$ 表示归一化的特征向量，例如 HOG 特征。H 表示目标候选区域的集合，它由若干个 H_i 构成，$i = 1, ..., N$ 是视频帧的索引序号。基本上，求解公式 4-2 对于每个正例视频帧（y=1）会得到一个高得分 $\beta^T \cdot v(x, y, h)$，而对于每个反例图像（y=0）会得到一个低的分数。具体来说，在视频帧集合和反例图像 $X = \{(x_i, y_i), i = 1, ..., N\}$ 上通过下式学到一个模型 β：

$$\min_{\beta, h} F_l(\beta, h) = \min_{\beta, h} \frac{1}{2} \| \beta \|^2 + C \sum_{i=1}^{N} l(\beta, x_i, y_i, h), \tag{4-3}$$

这里 C 是一个正则化因子，l 是一个凸差分损失函数，定义如下：

$$\begin{aligned} l(\beta, x_i, y_i, h) = & \max_{y, h} (\beta^T \cdot v(x_i, y, h) + \Delta(y_i, y)) \\ & - \max_{h} \beta^T \cdot v(x_i, y_i, h) \end{aligned} \tag{4-4}$$

如果 $y = y_i$，定义 $\Delta(y_i, y) = 0$ ，否则 $\Delta(y_i, y) = 1$。等式 4-3 和等式 4-4 旨在从其他配置中选择和区分最高评分的候选窗口 h，定义最大边界（max-margin）公式以度量图像、标签和候选区域之间的不匹配程度。

目标增强：目标发现过程旨在优化图像级别的分类，而不是样本级别的分类。一旦图像级别分类目标函数达到最优化，则不论样本级别分类是否被优化，学习过程都会停止[73]。针对所有正例图像（不包含反例图像）包含的目标部分区域，LSVM 会错误地选择目标的局部作为正例样本，这是因为等式 4-3 的非凸性，并且很容易陷入到局部最小。

受难反例挖掘[74]的成功的启发，我们提出使用空间正则化来强制目标的位置和模型。用 H_i 表示在帧 i 中的候选窗口，并且用 h' 表示与视频帧中目标 h 的对应的难反例，我们定义一个函数以最大化隐对象与其空间上邻居之间的距离，

$$\max_{\beta} F_s(\beta) = \sum_{i=1}^{N} \sum_{\substack{h \in H_i \\ h' \in \Omega_{H_i, h}}} \| \beta^T \cdot (v(x_i, h) - v(x_i, h')) \|^2 \tag{4-5}$$

$\Omega_{H_i, h}$ 表示在 H_i 中 h 的空间邻域。空间邻域是有些一些高得分的目标局部窗口和周围的图像块，这些图像块与 h 的 IoU（Intersection of Union）在 0.0 到 0.25 之间。等式 4-5 使用固定的 h 来优化模型，因此是凸正则化函数。这样的函数加强了隐模型，在渐进的学习过程中产生了目标定位的一致性，并且得到显著提升。

标签传播：在目标发现过程中对每一个帧仅输出一个样本。为了挖掘到更多的正例和反例，我们提出使用帧内标签传播算法来用于增量学习。

假设有来自先前学习迭代获得的 l 个有标记的样本。选取 $u = l \times (r - 1.0)$ 个高

分的候选框作为没有标记的样本，其中 $r>1.0$ 是学习率，它与预期的行人密度有关。给定一些有标记的样本 $\{h_i\}$，$i=1,...,l$，和一些没有标记的样本 $\{h_j\}$，$j=l,...,l+u$，我们首先在特征空间构建一个 kNN 图。这个图的顶点定义样本的最近邻顶点。如果 h_i 和 h_j 中有一个在其他的 kNN[34]之中，h_i 和 h_j 就会被连接在一起。基于图的标签传播过程被定义为 $g(\beta,h_j)=\frac{\sum_{k=1}^{l+u} w_{jk} g(\beta,h_k)}{\sum_{k=1}^{l+u} w_{jk}}$，这里 $j=l,\ldots,l+u$，w_{ik} 表示 h_i 和 h_j 之间的欧氏距离上的高斯函数定义的边的权重。这相当于一个凸优化问题[34]，

$$\min_{g(\beta,h)} F_g(\beta,h)=\min_{g(\beta,h)}\sum_{i=1}^{l}\sum_{j=l}^{l+u} w_{ij}(g(\beta,h_i)-g(\beta,h_j))^2 \qquad (4\text{-}6)$$
$$s.t. \quad g(\beta,h_i)=y_i, i=1,...,l,$$

其中 $g(\beta,h_i)$ 是是候选窗口 h_i 的传播分数，并且 y_i 是 h_i 所属的帧或图像的标签。

渐进优化：在学习的过程中，$F_s(\beta)$（目标增强）和 $F_s(\beta)$（标签传播）的优化取决于 $F_l(\beta,h)$ 的结果。等式 4-1 是一个渐进模型，其中 F_l、F_s 和 F_g 是被交替优化的。根据等式 4，F_l 可以被写成 $A(x)-B(x)$，并且 F 可以被写成 $A(x)-B(x)+C(x)-D(x)$。这也就是说，目标函数等式 1 可以被写成凸函数的差。这允许我们使用两步 Concave-Convex 过程（CCCP）[73]。F_l 的第一步 CCCP 发现帧中的潜在的行人对象，并初始化隐模型，$\gamma F_g-\lambda F_s$ 的第二步 CCCP 执行目标增强和标签传播。两步 CCCP 渐进优化 PLM，直到估计的采样误差率的变化可忽略。CCCP 算法保证了凸目标函数收敛到局部最小值或鞍点的优化[73]。因此，迭代使用两步 CCCP 算法并保持样本误差率的降低可以保证学习的稳定性。

4.2.2 自学习检测器

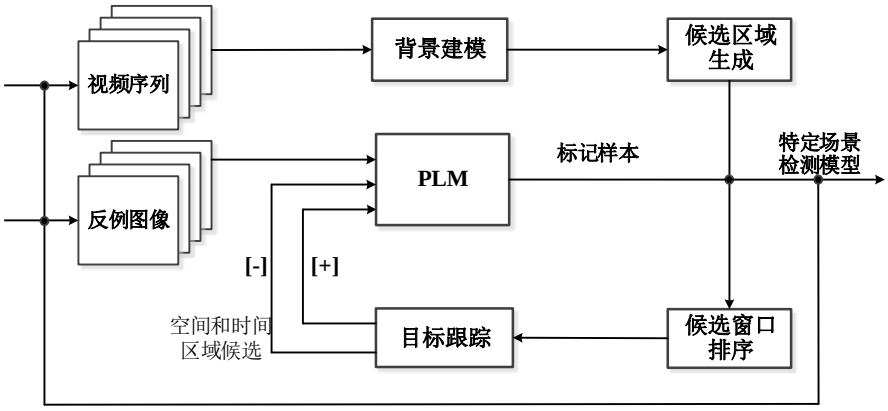


图 4-3 自学习方法流程图

使用提出的 PLM 方法，如图 3 所述实现一个自学习的方法。候选区域生成模块通过使用目标、运动和外观线索定位潜在对象。候选区域排序模块通过选择高排名的候选框作为正例候选，并且选择低排名的候选框作为反例。候选区域跟踪模块有助于在连续视频帧中发现候选窗口。PML 根据给定的候选窗口确定正例和难反例。利用挖掘到的正例样本，训练 DPM 检测器 $f_\beta(g)$ 以执行检测行人。

给定一个有静态背景的视频，使用背景建模算法为每个视频帧计算运动分数图。在运动评分图上，使用 EdgeBoxes 方法提取检测候选区域，如图 4-2（b）所示，根据该方法通过聚合高相似度的边缘获得轮廓，在根据轮廓计算边缘图。在轮廓上，使用位置、尺度和宽高比的滑动窗口策略来提取置信度较高的区域作为目标区域候选。从第二次迭代，利用初始化的检测器，使用滑动窗口策略来生成目标的候选，如图 4-2b 所示。为了在时间域上扩展候选窗口，采用 KLT 跟踪算法从 t 到 $t+\tau$ 帧中来跟踪和收集候选区域，这里 τ 根据经验设定为 10。在将这些空间-时间上的候选区域送入到学习算法之前，它们的宽高比江北归一化为平均的宽高比。为了防止在稀疏行人的视频中错误地选取静态背景，候选区域的平均背景概率需要大于阈值，在实验中通过经验设置为 0.20。

我们提出使用一个组合的分数，即 $f(h)=\alpha^T\cdot(f_\beta(h),f_m(h),f_o(g))$，去选择一个高排名的候选窗口，其中 α^T 是排序的权重向量。$f_\beta(h)$ 、$f_m(h)$ 和 $f_o(g)$ 分别是检测、运动和目标分数。一个候选区域的运动分数 $f_m(h)$ 被定义为其图像中所有像素的平均运动分数。通过计算候选区域的轮廓来定义目标分数 $f_o(g)$ [11]。较大的分数提高了候选区域是目标的可信度。由学习到的检测器从第二轮迭代计算的检测分数 $f_\beta(h)$。从此迭代中，将候选区域的中心设置为根位置，我们使用滑动窗口对候选区域进行定位。

在每次学习迭代中，使用零空间回归方法[69]来更新排序权重向量 α^T，其不使用输出值来执行学习。它基本上最小化所有样本的回归误差，以及最大化超平面到原点的距离。这导致权重向量捕获输入样本空间中找到数据的概率密度区域，并且使得候选窗口排名能够是自适应的。

4.3 实验

4.3.1 数据集

在监控摄像机获取的 5 个实际数据集（6 个视频序列）上评估了我们提出的方法。数据集涉及目标遮挡、低分辨率和移动干扰等挑战。

PETS2009[75]：在公共场所拍摄的拥挤视频序列，分辨率为 720x576 。

Towncenter[76]：一个城市中心的中等拥堵的视频序列，1920x1080 分辨率。

PNN-Parking-Lot2/Pizza[32]：中等拥堵的视频序列，包含多组具有复杂路线和相似外观的行人人群，具有 1920x1080 分辨率。由于大量的姿态变化和遮挡，所以这是一个具有挑战性的数据集。

CUHK Square[28]：长达 60 分钟的稀疏行人视频，并且具有其他运动的干扰物体，如移动的车辆。视频的分辨率为 704x576。行人的分辨率远低于其他数据集。由于相机具有大约 45 度的俯视视角，所以物体会有透视形变。

24Hours：一个 24 小时长的行人数据集，既包含稀疏行人，同时也包含密集行人，除此之外还有 24 小时照明变化和其他运动干扰的物体，例如移动的车辆，允许评估模型漂移。该视频的分辨率是 704x576。从视频中均匀采样 6000 帧用于学习，并且采样 2600 帧用于测试。

对于除 24Hours 数据集以外的所有数据集，一般视频帧用于学习，而其他具有标注的视频帧用于测试。对于提出的方法进行估计，并于以下监督学习、迁移学习和弱监督学习方法进行比较。

Offline-DPM[57]：在 PASCAL VOC 人类别上训练的 DPM 检测器。

Supervised-DPM：在特定场景下用有标注的行人样本训练的监督 DPM 检测器，并从反例图像中采集额外的反例样本。

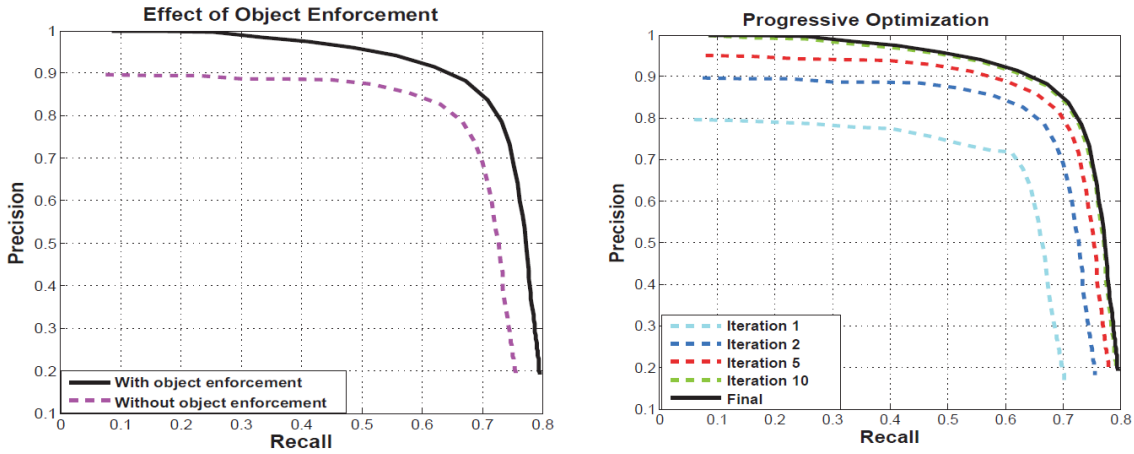
Supervised-SLSV[67]：考虑到在特定场景中模拟外观的虚拟行人上学到的最先进的行人检测器。没有公开可用的源代码，仅在 Towncener 数据集上与 SLSV 使用作者报告的结果进行比较。

Transfer-DPM[32]：基于迁移学习的特定场景检测方法。检测最初通过在 PASCAL VOC 人类别上离线训练的 DPM 检测器，然后使用基于超像素聚类和分类进行改进。

Transfer-SSPD[28]：具有迁移学习的特定场景行人检测器。

Weakly-MIL[45]：广泛使用的基于多示例学习的弱监督方法。然后从标注的正例样本中学习 DPM 学习器。

4.3.2 模型各个部分的影响



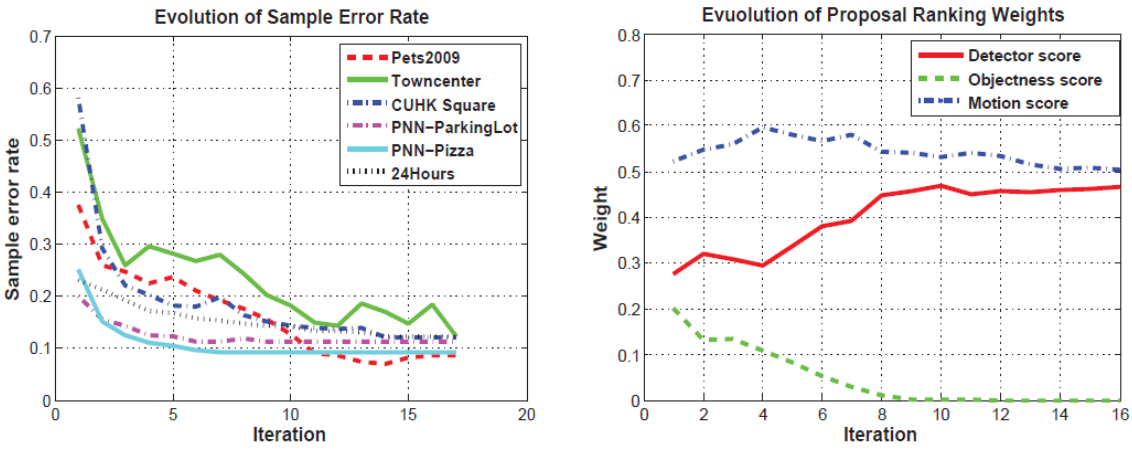
(a) (b)

图 4-4 模型的影响

在图 4-4（a）和图 4-4（b）中，我们分别评估了目标增强和标签传播的影响，表明 PLM 比我们传统的 LSVM 模型更有效。其中图 4-5（a）为样本误差率单调下降，4-5（b）为候选区域排名权重的演变。

目标增强：考虑到公式中的目标函数 4-3 是非凸的，学习倾向于在优化过程中陷入到局部最小值。通过使用目标增强过程，即公式 4-5，学习到的检测器的性能会显著提高，如图 4-4（a）。原因是行人可以更准确地被定位，并且大多数被错误地检测到的目标的部分会被抑制。在 0.7 召回率的时候，使用这种正则化项时，精度提高了 10%以上，这表明凸目标函数确实有助于非凸优化从较差的局部最小值中逃离出来。

标记传播：结合候选区域排序策略，标记传播可以无监督地增量式地标注行人样本。图 4-4（b）清楚地表明检测模型在迭代过程中得到了改进，显示了基于增量学习的图传播的有效性。经过数十次学习迭代，没有额外的被标记的正例样本，而且观察到的性能是稳定的。



(a) (b)

图 4-5 验证学习的稳定性

稳定性：图 4-5（a）显示了标记样本的误差率基本上单调下降，表明了提出自学习方法的稳定性。图 4-5（b）显示了 PETS2009 数据集的学习过程中候选区域排名权重的演变。目标分数的权重迅速衰减为 0，这意味着目标分数不如检测和运动评分那么具有判别性。检测分数的权重在学习中不断增加，这表明检测器逐渐改进。运动信息的权重减小到与检测信息类似的值，这意味着运动特性也是有区分性的。

表 4-1 显示了四个数据集的最大的 γ 值。Towncenter 数据集的 γ 是最大的，而 CUHK 数据集的 γ 最小。较大的 γ 意味着目标候选区域具有较小的噪声。Towncenter 数据集是一个具有很小照明方差和很少移动干扰的视频，因此使用较大的 γ。CUHK 和 24Hours 数据集由很多移动干扰因素，所以需要较小的 γ。

表 4-1 不同数据集上的标签传播参数

Dataset	PETS	Towncenter	PNN	CUHK	24Hours
γ	0.50	0.70	0.60	0.30	0.30

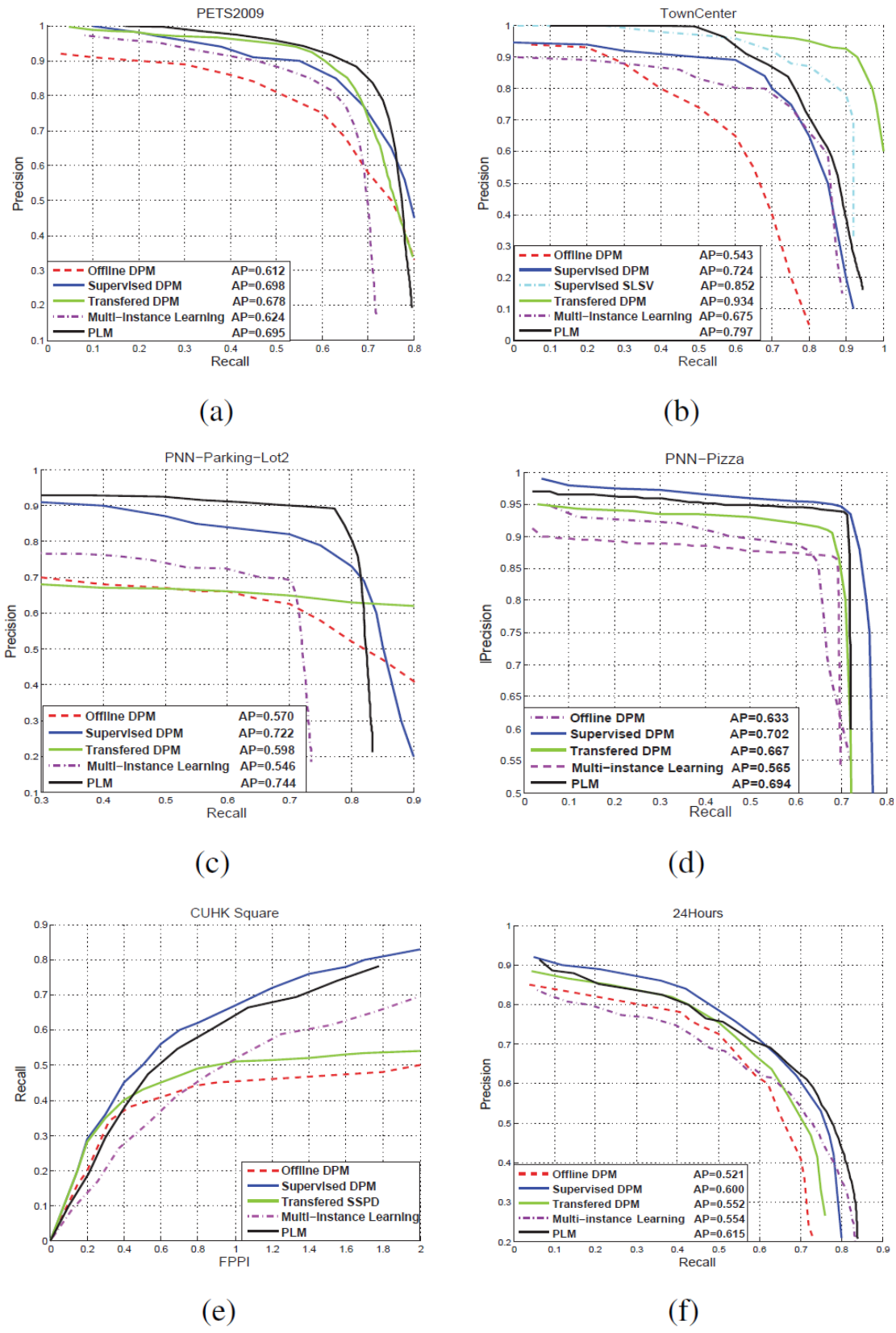


图 4-6 我们的方法的性能与弱监督、监督和迁移学习的比较

在五个数据集上，采用了 Precision-Recall 度量来评估方法并与其他方法进行比较。在 CUHK 数据集中，采用了 FPPI-Recall 度量，与现有技术的场景特异性检测方法保持一致[28]。

4.3.3 实验结果和性能

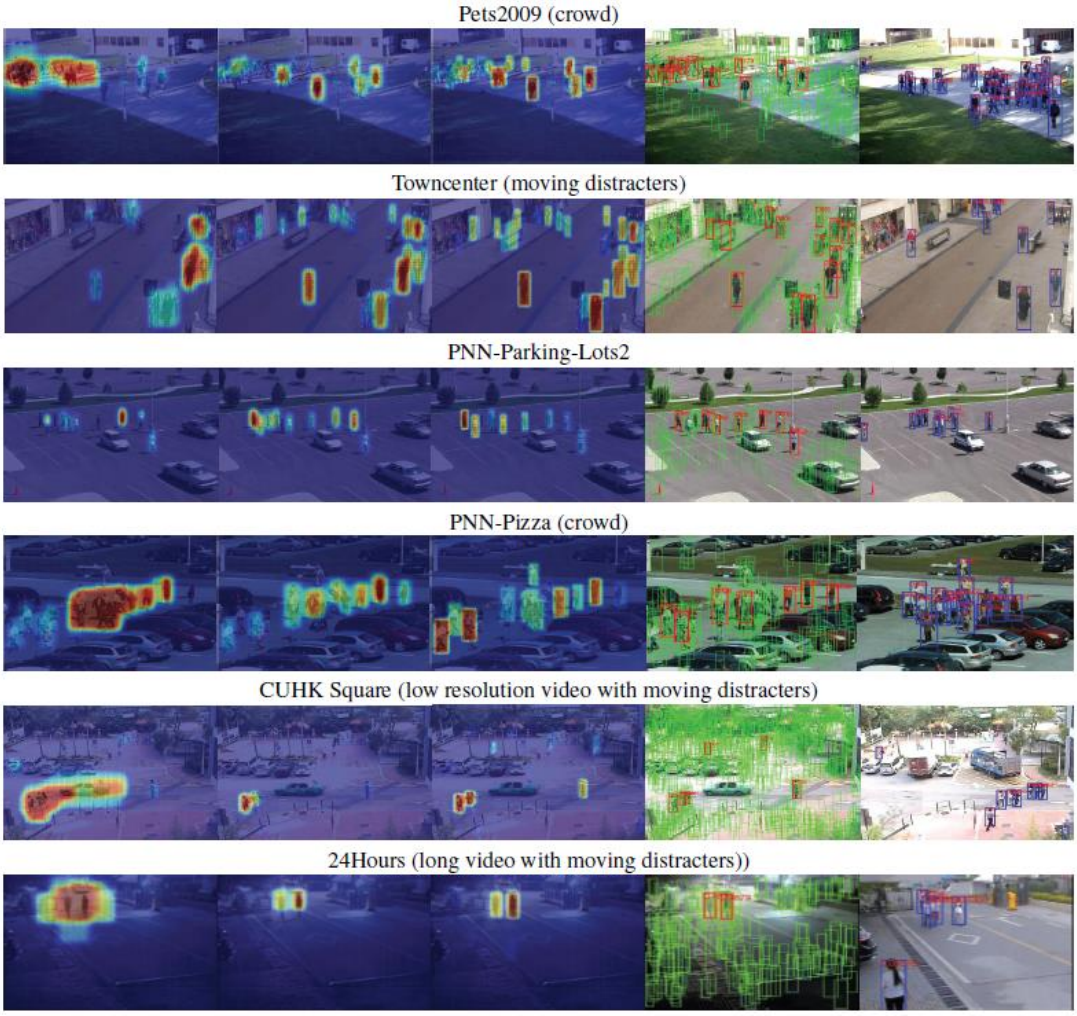


图 4-7 学习和检测展示图

在图 4-7 中，前三列：第 1、5 和 10 次学习迭代中的分数图。第四列：标注正例样本（红色框）。最后一列：测试集中的检测示例。图 4-6 中的 PR 和 FR 曲线显示我们的方法在所有数据集上都显著优于离线学习的 DPM 检测器。它还显著优于 Weakly-MIL 方法。在 PETS2009 和 PNNParking-Lot2 数据集上，我们的方法胜过所有比较的方法。在 CUHK 数据集上，我们的方法显著优于具有迁移学习的特定场景方法[28]，该迁移学习方法显示了在该数据集的最先进的性能。我们的结果甚至于监督的学习方法（Supervised-DPM）相当。在 Towncenter 数据集中，我们的方法也超过了 MIL 方法。然而，它显示比完全监督的方法 SLSV[49]和

迁移学习[32]更低的性能。原因可能是该视频场景中的行人稀少，我们的方法不能标注足够多的正例样本。但是，需要强调的是我们的方法没有使用任何有标注的训练样本。

在 24Hours 数据集中，我们的方法的 AP（平均精度）在所有比较方法中都最高，如图 4-6（e）所示。它比迁移学习方法高出约 6%，这验证了我们之前的分析：迁移学习受到概念差距问题的困扰，例如将训练在白天捕获的图像上的模型适应到 24 小时照明变化的视频中。相比之下，提出的自学习方法只是适应于相同场景中的学习检测器，自然避免了概念间隔问题。更令人惊讶的是，使用额外的运动线索之后，所提出的方法胜过该数据集中的完全监督的方法。

在图 4-7 中，我们使用每一行的关键帧来说明增量学习过程。可以看出，正样本被递增地标记，并且噪声样本逐渐减少。在拥挤的 PES2009 数据集和 PNN-Pizza 数据集中，我们的方法准确地标记样本，证明所学习的检测器已经纳入了特定场景的判别信息。在 Towncenter 数据集和 CUHK 数据集中，虽然存在移动干扰物，例如自行车和车辆，但是我们提出的方法正确地定位行人，从而在嘈杂的环境中表现出其鲁棒性。在 24Hours 数据中，一些视频帧具有密集的行人（白天），但是其他视频帧则在行人稀少（夜间）。从清晨到午夜的学习，我们的方法可以逐步提高其表现，没有模型漂移。在图 4-7 的最后一列中，检测结果表明所学习的特定场景的检测器是具有判别性的，显示了对遮挡、低分辨率和外形变化的鲁棒性。如图 4-8 所示，可以看出自学习方法适应于视角差异和 24 小时照明变化，但迁移学习会受到这些因素的影响，在图 4-8 中左图为自学习检测结果图，右图为迁移学习结果图。自学习检测可以从白天（左）到夜晚（右）正确检测所有行人，但迁移学习存在丢失和错检。

图 4-8 24Hours 数据集检测结果

4.4 本章小结

对所有场景的检测器的监督学习需要大量的人力对数据标注的努力。常用的迁移学习和半监督学习不能消除人们的监督信息，因为它们需要部分对象级标注。本文的研究结果表明，通过利用非常弱的标注的视频数据，可以自动学习特定视频场景的自定义行人检测模型。通过结合辨别和增量学习函数提出渐进潜伏模型。

通过在空间-时间对象建议上优化模型来实现自学习方法。实验表明，自学检测器与监督的检测器相当，这向自学习相机迈进了一步。

第五章 结论与展望

目标检测是计算机视觉需要解决的核心任务之一，它是让计算机获得人工智能的基础条件。目标检测也一直是计算机视觉领域的研究热点问题，而行人检测作为目标检测的一个分支同时也受到大家的关注。它作为视频监控、智能交通系统、车辆辅助驾驶和视频检索系统的关键技术，对这些众多领域的研究和发展具有着重要的实际意义。本文主要研究了目标检测中候选区域提取和特定场景弱监督行人目标检测等问题，提出了基于纹理复杂度的冗余候选区域排序算法和渐进隐模型行人检测框架，有效地解决了特定场景自学习行人目标检测问题。

本节对我们的研究工作进行总结和归纳，最后对分析其不足之处并且展望未来的研究方向。文本主要完成的研究工作：

1）提出一种基于纹理复杂度的冗余候选区域排序算法，在保证覆盖率的情况下大量减少了目标候选区域的数量。在纹理复杂度方面，我们提出了 LBP 熵和使用了目标完整区域轮廓数量这两个指标，并且通过结合实验分析了这两个测度的有效性。与现有众多候选区域进行对比，证明了我们提出的方法具有很好的性能。除此之外，我们的方法使用完全无监督的算法，可以与许多目标候选区域方法相结合，改善了目标候选窗口的质量。

2）提出特定场景下渐进隐模型弱监督行人检测框架，完全在没有标注样本的视频数据集下训练一个行人检测器。传统弱监督目标建模方法在学习过程中容易陷入局部最优，导致弱监督方法与全监督方法具有显著性能差距。我们通过研究多目标规划的渐进隐变量模型，有效抑制了导致模型陷入局部最优的关键因素如目标部件、目标类别相关性等。同时，有效加入目标的全局一致性约束，提高了样本的准确性。所提出的渐进隐变量模型在学习过程中更容易接近全局最优解，为弱监督视觉目标检测提供了更有效的解决方案，体现出方法的创新性。随着智能监控的广泛普及和智能监控技术的发展，自学习目标检测器会逐步成熟，并且嵌入到监控设备的前端，因此本研究具有实际应用价值。

虽然我们的研究完成了基本的特定场景自学习检测任务，不过还依然存在着许多可以研究的方面和一些不足：

1）在目标候选区域提取中，现在有些方法使用监督学习的方法训练目标候选区域提取算法，使得这些算法在性能上有所提升，大家也对这些监督学习的候选区域提取算法非常认可，未来可以结合监督学习方法对本算法进行改进，比如结合深度学习方法自主学习纹理复杂性特征表示。

2）本研究使用了一些特定场景行人数据集，并且标定了 24 小时行人数据集。但是这些行人数据集仍然缺乏多样性，例如在雨雪天气等，在这些数据集下我们的自学习检测算法的鲁棒性并没有在文中有所探讨。

3）所提出的自学习行人检测框架只考虑了行人这种单一的目标，而特定场景的多目标检测问题并没有进行研究，未来的工作中，特定场景多目标检测可以进行更多的研究和讨论。

4）目前监督学习中目标检测的最优秀的方法都涉及使用深度卷积特征，然而在行人检测中，许多论文表示单单使用深度卷积特征行人检测的性能并没有非常出色，所以许多方法使用深度卷积特征结合一些手工设计的特征。由于时间有限，我们的研究并没有在深度学习方面进行扩展，接下来的研究中可以进行尝试。

希望在未来的工作中，可以把特定场景弱监督行人检测问题进行深入的探索。弥补现有方法存在的不足。

参考文献

[1] Carreira J, Sminchisescu C. Cpmc: Automatic object segmentation using constrained parametric min-cuts[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 34(7): 1312-1328.

[2] Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition[J]. *International journal of computer vision*, 2013, 104(2): 154-171.

[3] Van de Sande K E A, Uijlings J R R, Gevers T, et al. Segmentation as selective search for object recognition[C]. *Computer Vision* (ICCV), *2011 IEEE International Conference on*. IEEE, 2011: 1879-1886.

[4] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014: 580-587.

[5] Manen S, Guillaumin M, Van Gool L. Prime object proposals with randomized prim's algorithm[C]. *Proceedings of the IEEE International Conference on Computer Vision*. 2013: 2536-2543.

[6] Rantalankila P, Kannala J, Rahtu E. Generating object segmentation proposals using global and local search[C]. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014: 2417-2424.

[7] Arbeláez P, Pont-Tuset J, Barron J T, et al. Multiscale combinatorial grouping[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014: 328-335.

[8] Xiao Y, Lu C, Tsougenis E, et al. Complexity-adaptive distance metric for object proposals generation[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015: 778-786.

[9] Alexe B, Deselaers T, Ferrari V. Measuring the objectness of image windows[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2012, 34(11): 2189-2202.

[10] Cheng M M, Zhang Z, Lin W Y, et al. BING: Binarized normed gradients for objectness estimation at 300fps[C]. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014: 3286-3293.

[11] Zitnick C L, Dollár P. Edge boxes: Locating object proposals from edges[C]. *European Conference on Computer Vision*. Springer International Publishing, 2014: 391-405.

[12] Karianakis N, Fuchs T J, Soatto S. Boosting convolutional features for robust object proposals[J]. arXiv preprint arXiv:1503.06350, 2015.

[13] Chen X, Ma H, Wang X, et al. Improving object proposals with multi-thresholding straddling expansion[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015: 2587-2595.

[14] Kuo W, Hariharan B, Malik J. Deepbox: Learning objectness with convolutional networks[C]. *Proceedings of the IEEE International Conference on Computer Vision*. 2015: 2479-2487.

[15] 郭劲智. 视频图像行人检测方法研究[D]. 华南理工大学, 2012.

[16] 吕敬钦. 视频行人检测及跟踪的关键技术研究[D]. 上海: 上海交通大学, 2013.

[17] 石娟峰. 基于视频的行人检测和跟踪 [D][D]. 北京: 北京邮电大学, 2012.

[18] 汤义. 智能交通系统中基于视频的行人检测与跟踪方法的研究[D]. 广州：华南理工大学, 2010.
[19] 邹依峰. 智能视频监控中的行人检测与跟踪方法研究[D]. 合肥：中国科学技术大学, 2011.
[20] 鲁帅. 视频监控中目标检测与跟踪算法研究[D]. 复旦大学, 2012.
[21] Cai Z, Saberian M, Vasconcelos N. Learning complexity-aware cascades for deep pedestrian detection[C]. *Proceedings of the IEEE International Conference on Computer Vision*. 2015: 3361-3369.
[22] Dollar P, Wojek C, Schiele B, et al. Pedestrian detection: An evaluation of the state of the art[J]. *IEEE transactions on pattern analysis and machine intelligence, 2012*, 34(4): 743-761.
[23] Donahue J, Hoffman J, Rodner E, et al. Semi-supervised domain adaptation with instance constraints[C]. *Proceedings of the IEEE conference on computer vision and pattern recognition. 2013*: 668-675.
[24] Tian Y, Luo P, Wang X, et al. Deep learning strong parts for pedestrian detection[C]. *Proceedings of the IEEE International Conference on Computer Vision. 2015*: 1904-1912.
[25] Yu C N J, Joachims T. Learning structural SVMs with latent variables[C]. *Proceedings of the 26th annual international conference on machine learning. ACM, 2009*: 1169-1176.
[26] Zhu X, Goldberg A B. Introduction to semi-supervised learning[J]. *Synthesis lectures on artificial intelligence and machine learning, 2009,* 3(1): 1-130.
[27] Wang M, Wang X. Automatic adaptation of a generic pedestrian detector to a specific traffic scene[C]. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011*: 3401-3408.
[28] Wang X, Wang M, Li W. Scene-specific pedestrian detection for static video surveillance[J]. *IEEE transactions on pattern analysis and machine intelligence, 2014*, 36(2): 361-374.
[29] Zhang S, Benenson R, Omran M, et al. How far are we from solving pedestrian detection?[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016*: 1259-1267.
[30] Vázquez D, Lopez A M, Marin J, et al. Virtual and real world adaptation for pedestrian detection[J]. *IEEE transactions on pattern analysis and machine intelligence, 2014*, 36(4): 797-809.
[31] Yang Y, Shu G, Shah M. Semi-supervised learning of feature hierarchies for object detection in a video[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013*: 1650-1657.
[32] Shu G, Dehghan A, Shah M. Improving an object detector and extracting regions using superpixels[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013*: 3721-3727.
[33] Kuznetsova A, Ju Hwang S, Rosenhahn B, et al. Expanding object detector's horizon: incremental learning framework for object detection in videos[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015*: 28-36.
[34] Gaidon A, Zen G, Rodriguez-Serrano J A. Self-learning camera: Autonomous adaptation of object detectors to unlabeled video streams[J]. arXiv preprint arXiv:1406.4296, 2014.
[35] Andriluka M, Roth S, Schiele B. People-tracking-by-detection and people-detection-by-tracking[C]. *Computer Vision and Pattern Recognition*, 2008. CVPR 2008. IEEE Conference on.

IEEE, 2008: 1-8.
[36] Mao Y, Yin Z. Training a scene-specific pedestrian detector using tracklets[C]. *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on. IEEE, 2015*: 170-176.
[37] Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection[J]. *IEEE transactions on pattern analysis and machine intelligence, 2012*, 34(7): 1409-1422.
[38] Misra I, Shrivastava A, Hebert M. Watch and learn: Semi-supervised learning for object detectors from video[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015*: 3593-3602.
[39] Kwak S, Cho M, Laptev I, et al. Unsupervised object discovery and tracking in video collections[C]. *Proceedings of the IEEE International Conference on Computer Vision. 2015*: 3173-3181.
[40] Song H O, Girshick R B, Jegelka S, et al. On learning to localize objects with minimal supervision[C]. ICML. 2014: 1611-1619.
[41] Divvala S K, Farhadi A, Guestrin C. Learning everything about anything: Webly-supervised visual concept learning[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014: 3270-3277.
[42] Wang C, Ren W, Huang K, et al. Weakly supervised object localization with latent category learning[C]. *European Conference on Computer Vision. Springer International Publishing, 2014*: 431-445.
[43] Wu B, Nevatia R. Improving part based object detection by unsupervised, online boosting[C]. *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE*, 2007: 1-8.
[44] Cho M, Kwak S, Schmid C, et al. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015: 1201-1210.
[45] Cinbis R G, Verbeek J, Schmid C. Weakly supervised object localization with multi-fold multiple instance learning[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2017, 39(1): 189-203.
[46] Ren W, Huang K, Tao D, et al. Weakly supervised large scale object localization with multiple instance learning and bag splitting[J]. *IEEE transactions on pattern analysis and machine intelligence, 2016*, 38(2): 405-416.
[47] Xiao F, Jae Lee Y. Track and segment: An iterative unsupervised approach for video object proposals[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016*: 933-942.
[48] Shi J, Malik J. Normalized cuts and image segmentation[J]. *IEEE Transactions on pattern analysis and machine intelligence*, 2000, 22(8): 888-905.
[49] Felzenszwalb P F, Huttenlocher D P. Efficient graph-based image segmentation[J]. *International journal of computer vision*, 2004, 59(2): 167-181.
[50] Lowe D G. Object recognition from local scale-invariant features[C]. Computer vision, 1999. *The proceedings of the seventh IEEE international conference on*. IEEE, 1999, 2: 1150-1157.
[51] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. *Computer Vision and Pattern Recognition*, 2005. CVPR 2005. *IEEE Computer Society Conference on*. IEEE, 2005, 1: 886-893.

[52] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. *Proceedings of the IEEE*, 1998, 86(11): 2278-2324.
[53] Cortes C, Vapnik V. Support-vector networks[J]. *Machine learning*, 1995, 20(3): 273-297.
[54] Schapire R E, Singer Y. Improved boosting algorithms using confidence-rated predictions[J]. *Machine learning*, 1999, 37(3): 297-336.
[55] Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning[J]. *Advances in neural information processing systems*, 2003: 577-584.
[56] Dollár P, Appel R, Belongie S, et al. Fast feature pyramids for object detection[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36(8): 1532-1545.
[57] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2010, 32(9): 1627-1645.
[58] Hosang J, Benenson R, Schiele B. How good are detection proposals, really?[J]. arXiv preprint arXiv:1406.6962, 2014.
[59] Dollár P, Zitnick C L. Structured forests for fast edge detection[C]. *Proceedings of the IEEE International Conference on Computer Vision*. 2013: 1841-1848.
[60] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. *IEEE Transactions on pattern analysis and machine intelligence*, 2002, 24(7): 971-987.
[61] Everingham M, Eslami S M A, Van Gool L, et al. The pascal visual object classes challenge: A retrospective[J]. *International Journal of Computer Vision*, 2015, 111(1): 98-136.
[62] Carreira J, Sminchisescu C. Cpmc: Automatic object segmentation using constrained parametric min-cuts[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 34(7): 1312-1328.
[63] Endres I, Hoiem D. Category-independent object proposals with diverse ranking[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2014, 36(2): 222-234.
[64] Humayun A, Li F, Rehg J M. RIGOR: Reusing inference in graph cuts for generating object regions[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014: 336-343.
[65] Rahtu E, Kannala J, Blaschko M. Learning a category independent object detection cascade[C]. *Computer Vision* (ICCV), *2011 IEEE International Conference on*. IEEE, 2011: 1052-1059.
[66] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]. *Advances in neural information processing systems. 2015*: 91-99.
[67] Hattori H, Naresh Boddeti V, Kitani K M, et al. Learning scene-specific pedestrian detectors without real data[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015*: 3819-3827.
[68] Tian Y, Luo P, Wang X, et al. Deep learning strong parts for pedestrian detection[C]. *Proceedings of the IEEE International Conference on Computer Vision. 2015*: 1904-1912.
[69] Chang C C, Lin C J. LIBSVM: a library for support vector machines[J]. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2011, 2(3): 27.
[70] Fu Y, Hospedales T M, Xiang T, et al. Transductive multi-view embedding for zero-shot recognition and annotation[C]. *European Conference on Computer Vision. Springer International*

Publishing, 2014: 584-599.

[71] Ye Q, Han Z, Jiao J, et al. Human detection in images via piecewise linear support vector machines[J]. *IEEE transactions on image processing, 2013*, 22(2): 778-789.

[72] Papazoglou A, Ferrari V. Fast object segmentation in unconstrained video[C]. *Proceedings of the IEEE International Conference on Computer Vision. 2013*: 1777-1784.

[73] Zeng X, Ouyang W, Wang M, et al. Deep learning of scene-specific classifier for pedestrian detection[C]. *European Conference on Computer Vision. Springer International Publishing, 2014*: 472-487.

[74] Girshick R. Fast r-cnn[C]. *Proceedings of the IEEE International Conference on Computer Vision. 2015*: 1440-1448.

[75] Ferryman J, Shahrokni A. Pets2009: Dataset and challenge[C]. Performance Evaluation of Tracking and Surveillance (PETS-Winter), *2009 Twelfth IEEE International Workshop on. IEEE*, 2009: 1-6.

[76] Benfold B, Reid I. Stable multi-target tracking in real-time surveillance video[C]. *Computer Vision and Pattern Recognition (CVPR)*, 2011 IEEE Conference on. IEEE, 2011: 3457-3464.

[77] Ke W, Zhang Y, Wei P, et al. Pedestrian detection via pca filters based convolutional channel features[C]. *Acoustics, Speech and Signal Processing (ICASSP)*, 2015 IEEE International Conference on. IEEE, 2015: 1394-1398.

致　　谢

时光荏苒，岁月如梭。雁栖湖畔的硕士生涯即将告一段落，迎接我的是新的挑战。在此感谢实验室尊敬的各位导师、亲爱的实验室兄弟姐妹以及敬爱的家人，是你们的鼓励与支持伴我一路前行。

本人的学位论文是在我的恩师叶齐祥教授的殷切关怀和耐心的指导下完成的，衷心地感谢我的恩师对我研究生期间的谆谆教诲和悉心关怀。感谢叶老师给我一个步入中国科学院大学学习和研究的机会，我非常荣幸自己是国科大的一员，在此我接触了前沿的科学研究，认识了可爱的兄弟姐妹。叶老师严谨的治学精神和对学术的敬业之心深深地感染着我，并激励着我。除此之外，我还要感谢实验室的焦建彬教授，焦老师为人和蔼，身为实验室的带头人，为实验室的建设和发展付出了许多努力，同时也感谢焦老师对我的指导和帮助。感谢韩振军老师，韩老师不单肩负学生的科研任务同时也负责实验室的基本运行事务，感谢韩老师对我的帮助和鼓励。感谢秦飞老师，秦老师严谨的思维和精益求精的工作作风也深深地令我佩服。

感谢实验室的兄弟姐妹与我一起在雁栖湖畔度过快乐难忘的硕士生涯。首先，感谢柯炜师兄带领我完成第一个科研研究方向，在此过程中让我受益匪浅。然后还要感谢李策师姐、陈孝罡师兄、高山师兄、邹佳凌师兄、魏鹏旭师姐、张晓丹师姐以及实验室每一位同学对我的帮助。特别是在我受伤生病期间，大家的关爱和鼓励让我永远也不能忘记。

还要感谢几个快乐的小伙伴，其中有李兆举、万方、戴蔚群和王忻雷，是你们让这几年的学习变得更加舒心和快乐。

特别要感谢我的家人以及崔千，感谢你们的陪伴与鼓励，你们是我最坚强的后盾，在我最需要帮助和鼓励的时候，总是你们挺身而出。

最后，我要感谢参与开题、中期和本人论文答辩的每一位老师，感谢你们提出的中肯的建议和研究的指导。

个人简历、在学期间发表的论文与研究成果

姓名：张天亮　性别：男　出生日期：1990.09.04　政治面貌：中共党员

教育经历

- 2009 年 9 月至 2013 年 7 月　武汉理工大学　电子信息工程　学士
- 2013 年 9 月至 2017 年 7 月　中国科学院大学　工业工程　硕士

曾获荣誉：

- 第五届全国大学生节能减排社会实践与科技竞赛　全国一等奖（2012 年）
- 中国科学院大学　三好学生（2014-2015 学年）
- 中国科学院大学　三好学生（2016-2017 学年）

已发表（录用）论文

1. Ke W, **Zhang T**, Chen J, et al. Texture Complexity based Redundant Regions Ranking for Object Proposal[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016: 10-18.

已接收论文

1. Ye Q, **Zhang T**, Qiu Q, et al. Self-learning Scene-specific Pedestrian Detectors using a Progressive Latent Model. Accepted to 2017 IEEE Conference on Computer Vision and Pattern Recognition.

软件著作权

1. **张天亮**，叶齐祥，韩振军，焦建彬，基于图割和区域对比度金字塔图像融合软件，软件著作权，登记号：2016SR040845，计算机软件著作权