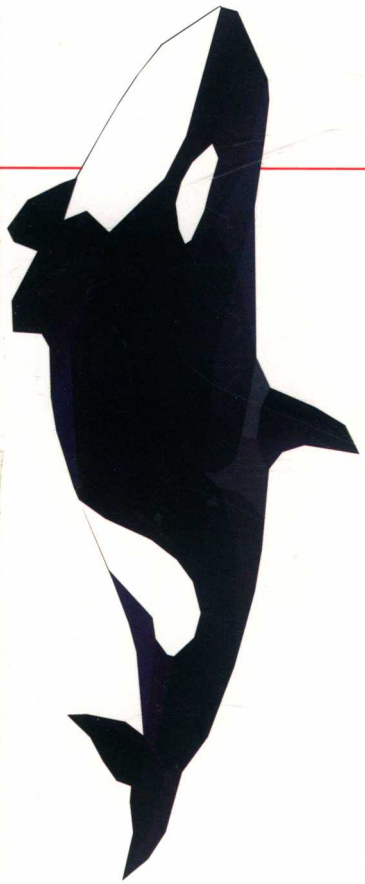
领域自适应目标检测方法与应用

叶茂 唐宋 李旭冬◎著



科学出版社

领域自适应目标检测方法与应用

叶 茂 唐 宋 李旭冬 著

科学出版社

北 京

内 容 简 介

国务院印发的《“十三五”国家科技创新规划》提出要把图像智能分析作为人工智能的关键技术。图像智能分析的关键步骤之一是目标检测，当把训练好的目标检测方法应用于新场景时，由于工作环境的变动，目标检测效果通常会迅速下降。本书围绕领域自适应目标检测方法开展了两方面的研究：面向监控场景的有监督、无监督目标检测迁移方法研究和基于记忆预测机制的目标检测方法研究。本书成果能广泛应用于智能视频监控、车辆自动驾驶、视觉人机交互等领域。

本书可供从事人工智能领域研究的高年级本科生和研究生参考，也可供科研院所的研究者使用。

图书在版编目(CIP)数据

领域自适应目标检测方法与应用/叶茂，唐宋，李旭冬著. —北京：科学出版社，2019.1

ISBN 978-7-03-057639-2

Ⅰ. ①领… Ⅱ. ①叶… ②唐… ③李… Ⅲ. ①数字图像处理-研究 Ⅳ. ①TP391.413

中国版本图书馆 CIP 数据核字(2018) 第 122277 号

责任编辑：张 展 陈丽华 / 责任校对：熊倩莹
责任印制：罗 科 / 封面设计：墨创文化

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

四川煤田地质制图印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2019 年 1 月第 一 版 开本：B5 (720 × 1000)
2019 年 1 月第一次印刷 印张：14 3/4
字数：297 千字
定价：136.00 元
(如有印装质量问题，我社负责调换)

作 者 简 介

叶茂，男，1973 年生，电子科技大学计算机科学与工程学院教授、博士生导师，中国工程物理研究院客座教授，西华师范大学兼职教授。计算机学会计算机视觉专委会委员，多媒体计算专委会委员，自动化学会混合智能专委会委员。2002 年从香港中文大学获得计算数学哲学博士学位并加入电子科技大学至今。曾于 2005~2006 年、2010 年分别在昆士兰大学和美国宾夕法利亚大学做访问学者。入选教育部新世纪优秀人才支持计划，四川省学术与技术带头人后备人选，四川省杰出青年学科带头人支持计划。目前主要研究领域为机器学习与计算机视觉，主持国家自然科学基金、四川省科技厅等国家、省部级课题，主持开发多个已取得良好经济效益的工业产品项目，已发表论文 70 余篇，其中国际一流 SCI 期刊论文 60 余篇，已授权发明专利 13 项；担任多个国际顶级学术会议程序委员会成员和分会场主席，如国际人工智能联合会 2018 (IJCAI2018)，国际神经网络研讨会(ISNN)，高级数据挖掘与应用国际会议(ADMA)等。担任中科院工程技术领域二区期刊 *Engineering Applications of Artificial Intelligence* 编委，担任多个国际一流期刊审稿人，如 *IEEE Trans. NNLS*，*IEEE Trans. Cybernetics*，*IEEE Trans. Signal Processing* 等。荣获 2012 年华为–电子科技大学优秀合作团队。

唐宋，男，1982 年生，四川人，上海理工大学机器智能研究院副教授，深圳市行者机器人技术有限公司人工智能研发总监。2017 年毕业于电子科技大学，获计算机科学博士学位，2017~2018 年，任德国汉堡大学信息科学系助理研究员。目前研究领域为：迁移学习，跨模态学习，深度学习。自 2011 年来，先后主研国际联合研究项目，国家自然科学基金项目 7 项；发表论文 20 余篇，其中被 SCI 检索 16 篇，申请软件著作权 3 项。

李旭冬，男，1988 年生，四川成都人。2017 年毕业于电子科技大学，获得工科博士学位，研究方向为深度学习、模式识别、图像处理等。现任腾讯科技(成都)有限公司研究员，从事游戏智能的相关研究。自从 2011 年来，参与国家自然科学基金课题和横向研究课题 5 项，以第一作者发表 3 篇 SCI 论文、2 篇国际会议论文和 1 篇中文核心期刊论文，申请 3 项国家专利，荣获 2017 年 IEEE ICME 大会最佳学生论文奖。

前　言

机器人技术是战略高技术，是各国必争的前沿技术。在我国，智能机器人是《中国制造 2025》规划和国家“十三五”规划重点支持的研究领域，2015 年我国科学技术部把机器人产业列为国家新兴战略与重点扶持产业。随着机器人制造技术与应用的日益成熟，针对高端制造、医疗康复、国防安全等领域的需求，开发适应于动态场景的服务机器人是一个必然趋势。动态场景通常是非结构化的，是指作业无法在事先布置好的条件下进行，而且在作业过程中环境会发生变化。

基于视觉的任务是指将某类目标对象从输入图像的背景中分开，是机器视觉系统的一个重要组成部分。若采用窗口方法检测目标，此时目标检测转化为 0-1 分类问题。基于视觉的目标检测算法 (简称目标检测) 根据深度学习的兴起可以分为两个阶段，之前的方法主要是精心设计分类器或者新的特征来实现目标检测，如基于支持向量机的方法、基于级联分类器的方法、基于 Hough 变换的方法、基于部件形变模型的方法、基于结构化学习的方法、基于上下文的方法等。目前，基于深度学习的方法成为主流，如基于深度卷积神经网络的方法、基于区域预测及掩码的方法、基于回归的方法、基于记忆的方法等。有相当一部分研究成果已经在某些领域得到了一定程度的实际应用，如人脸检测等。

当人们将这些目标检测方法应用于服务机器人时，由于工作环境的变动，源场景与目标场景数据分布不一致，目标检测效果通常会迅速下降。原因有两个：① 训练源目标检测器时，有足够多的训练样本，足够多的训练样本能让算法更泛化，然而，由于训练集包含了太多不同情形下的目标样本，而新场景中目标只是源训练集的一个小子集，充分训练后的目标检测器会在新场景中产生较高的误报，也就是负迁移情况；② 人们不可能收集所有情形下的样本，而且可能缺失与新应用场景相关的目标样本，此时的目标检测器将产生很多漏报。

为解决这个问题，自适应目标域的迁移学习方法引起了许多学者注意，也产生了不少优秀成果。考察服务机器人应用于新场景，发现这些方法工作的条件是：在迁移阶段保留所有源训练集样本，或者需要少量的目标域样本标签。鉴于存储空间、计算能力的约束及经常改变的应用环境，服务机器人很难满足这些条件。由

此，激发我们进行本课题的研究，即将预先训练好的目标检测器无监督地迁移到新场景，尽量不保留或者少保留源训练样本，且没有新场景的样本标签，同时迁移后的检测器应具有更高的检测准确率与召回率。

本书在介绍传统基于深度学习目标检测方法的基础上，系统地描述有监督与无监督领域自适应目标检测方法。第 1 章系统地描述有监督和无监督目标检测方法。第 2 章详细介绍卷积神经网络和基于卷积神经网络的车辆检测方法。第 3 章针对现有车辆检测器缺乏场景自适应性的问题，研究一种基于网络调整和结构优化的车辆检测器迁移方法。第 4 章为了构建具有场景自适应性的目标检测器，研究一种基于网络迁移和上下文学习的目标检测器构建方法。第 5 章研究一种基于分类器回归的迁移方法，应用于行人检测迁移问题。第 6 章研究在不保留源域样本集、目标域信息未知、源域样本足够的条件下，利用调控网络实现行人检测器迁移的方法。第 7 章进一步研究特征调控迁移方法。第 8 章基于人脑记忆和预测机制的启发，研究一种基于回复式神经网络的行人检测方法。第 9 章为了使行人检测器模拟人类的记忆过程，研究一种基于序列次序交换和序列模式记忆的行人检测器设计方法。第 10 章进行全书总结，并对未来研究方向进行展望。

近几年来，作者在研究中获得多项研究基金和项目的资助，特别是国家自然科学基金项目 (61375038，61773093) 与四川省科学技术厅应用基础重点项目 (2016JY0088)，作者在此表示感谢。本书第 3、4、8、9 章主要由李旭冬博士撰写，第 5~7 章主要由唐宋博士撰写，叶茂教授提供了主要研究思路和问题，并参与了全书撰写、修订和整合。

由于作者水平所限，疏漏与不足之处在所难免，殷切希望广大读者批评指正。关于我们更多目标检测研究方面的进展，欢迎访问网站 http://faculty.uestc.edu.cn/yemao/zh_CN/index.htm。

叶茂　唐宋　李旭冬

2018 年 5 月

目　录

第1章 绪 论

1.1 背景及意义

随着图像处理技术的发展、数码产品的推广以及互联网的普及，图像日益成为人类生活中获取信息的重要来源和传递信息的重要载体。目前，全球互联网用户每天上传的图像达到数亿幅。面对这种爆炸式的数据增长，已经不可能通过人力分析每幅图像，迫切需要利用计算机准确、快速地捕获图像的有效信息，实现图像智能分析。国务院印发的《“十三五”国家科技创新规划》提出要把图像智能分析作为人工智能的关键技术重点支持。

图像智能分析的关键步骤之一是目标检测(object detection)[1, 2]。作为计算机视觉领域中一个重要的课题，目标检测主要利用图像处理和机器学习的方法定位图像中感兴趣的目标，准确地判别每个目标的标签 (label)，并且给出每个目标在图像中的边界框 (bounding box)。精确的目标检测是后续图像智能分析 (跟踪、识别、验证、匹配、检索等) 顺利进行的必要条件。因此，目标检测在智能视频监控、车辆自动驾驶、机器人环境感知、医学图像分析、大规模图像检索、视觉人机交互等领域都有广泛的应用。

目标检测非常有挑战性，主要难点在于两方面：① 对于目标，姿态改变、局部遮挡、尺度不一等因素，引起目标的外观发生较大形变，导致目标检测产生漏报；② 对于场景，光照变化和视角变更等因素，同样会使目标的外观发生一定形变，并且场景越复杂，区别目标与非目标的难度就越大，导致目标检测产生漏报和误报。目前，在光照、视角和姿态相对固定的简单场景中，传统目标检测方法可以取得较好的检测结果。然而，在实际场景中，由于存在以上影响因素，传统目标检测方法的准确度达不到实际需求。其主要原因在于传统目标检测方法采用手工制作的特征 (hand-crafted feature)，对目标的表达能力不足，抗形变能力差，导致难以区分目标特征和非目标特征。

为了解决上述问题，LeCun 等于 2015 年提出了深度学习(deep learning)方法[3]，利用深度神经网络从大规模数据中自动地学习特征。相比于手工制作的特

征，深度神经网络学习的特征抽象层次更高，内容更丰富，表达能力更强。常用的深度学习模型包含限制玻尔兹曼机 (restricted Boltzmann machine，RBM)[4]、自编码器 (autoencoder，AE)[5] 和卷积神经网络 (convolutional neural network，CNN)[6]。随着深度学习的不断发展，研究者发现卷积神经网络更适合用于目标检测，准确度可以获得较大的提高。一方面是因为卷积神经网络提取了高层特征，提高了特征的表达能力；另一方面是因为卷积神经网络将目标检测的关键步骤融合在同一模型中，通过端到端的训练，进行整体的功能优化，增强了特征的可分性。因此，基于卷积神经网络的目标检测成为当前计算机视觉、人工智能、模式识别等领域的研究热点之一，无论是在学术界还是在工业界都得到了广泛的关注。

在学术界，许多国内外顶级科研团队把基于卷积神经网络的目标检测作为主要研究方向，如多伦多大学的 Hinton 团队①、纽约大学的 LeCun 团队②、斯坦福大学的 Li 团队③、加州大学伯克利分校的 Darrell 团队④、密歇根大学安娜堡分校的 Lee 团队⑤、法国国家信息与自动化研究所的 Sivic 团队⑥、香港中文大学多媒体实验室⑦、清华大学智能技术与系统国家重点实验室⑧、上海交通大学计算机视觉实验室⑨。在工业界，许多知名企业设立专门的团队开展基于卷积神经网络的目标检测方法研究，如 Google DeepMind⑩、微软亚洲研究院视觉计算组⑪、Facebook 人工智能研究中心⑫、百度深度学习研究院⑬、腾讯优图⑭、海康威视研究院⑮、商汤科技⑯等。并且每年举办的 ImageNet大规模视觉识别竞赛 (imagenet large scale visual recognition challenge，ILSVRC)[7] 为目标检测提供了一个公平的比赛平台。自从 2013 年设立目标检测任务以来，每次竞赛的冠军队伍都对卷积神经网络做出

① http://www.cs.toronto.edu/~hinton/.
② http://yann.lecun.com/.
③ http://vision.stanford.edu/.
④ https://people.eecs.berkeley.edu/ trevor.
⑤ http://www.eecs.umich.edu/.
⑥ http://www.di.ens.fr/ josef/.
⑦ http://mmlab.ie.cuhk.edu.hk/.
⑧ http://www.csai.tsinghua.edu.cn/.
⑨ http://www.visionlab.sjtu.edu.cn/.
⑩ https://deepmind.com/.
⑪ http://www.msra.cn/zh-cn/default.aspx.
⑫ https://research.fb.com/.
⑬ http://idl.baidu.com/.
⑭ http://open.youtu.qq.com/#/open.
⑮ http://www.hikvision.com/cn/index.html.
⑯ https://www.sensetime.com/.

了改进，因此该竞赛成为研究基于卷积神经网络目标检测的风向标。

同时，众多科研力量的投入使基于卷积神经网络的目标检测发展迅速，每年都有相关的科研成果发表于计算机视觉、机器学习和人工智能领域的国际顶级会议和国际顶级期刊上。主要的国际顶级会议有国际计算机视觉大会 (International Conference on Computer Vision，ICCV)、国际机器学习大会 (International Conference on Machine Learning，ICML)、IEEE 计算机视觉与模式识别 (IEEE Conference on Computer Vision and Pattern Recognition，CVPR)、神经信息处理系统年会 (Annual Conference on Neural Information Processing Systems，NIPS) 等。主要的国际顶级期刊有 IEEE 模式分析与机器智能学报 (IEEE Transactions on Pattern Analysis and Machine Intelligence，TPAMI)、IEEE 图像处理学报 (IEEE Transactions on Image Processing，TIP)、计算机视觉国际期刊 (International Journal of Computer Vision，IJCV) 等。

尽管如此，当前基于卷积神经网络的目标检测方法仍然不能满足现实生活的应用需求，许多关键问题没有得到圆满解决。例如：

(1) 现有目标检测器缺乏场景自适应性。当监控场景的光照和视角发生变化时，目标检测的准确度将急剧下降。为此，可以利用迁移学习的思想调整目标检测器，使目标检测器获得针对监控场景的自适应能力。

(2) 现有目标检测器缺乏记忆和预测机制。单纯在卷积神经网络中堆加层次将难以提高目标检测的准确度。为此，可以根据人类的认知过程，通过模拟人眼的视觉系统以及大脑的记忆和预测机制设计全新的目标检测器，增强目标检测器的学习能力、记忆能力和预测能力。

因此，本书针对以上两个问题开展自适应目标检测方法研究，具有重要的理论价值和实际意义。

1.2　目标检测研究现状

本节将详细介绍目标检测方法的研究现状。首先，沿着模板匹配和图像分类两条技术路线回顾目标检测方法。然后，从分类和回归两个角度论述如何利用深度神经网络进行目标检测。

1.2.1　目标检测方法概述

经过十多年的探索，研究者提出了各式各样的目标检测方法。根据不同的技术

路线，目标检测方法大致可以分为两类：基于模板匹配的目标检测方法和基于图像分类的目标检测方法。

1. 基于模板匹配的目标检测方法

利用模板匹配进行目标检测的主要思路是首先把少量的目标图像制作为模板图像，其次将检测图像中的子图像与模板图像进行匹配，再次计算两幅图像之间的相似度，最后把相似度超过一定阈值的子图像判定为目标。

如图 1.1 所示，基于模板匹配的目标检测方法流程主要分为五个步骤：预处理、窗口滑动、特征提取、模板匹配和后处理。第一步，预处理的过程是对检测图像进行图像去噪、图像增强、色彩空间转换等操作。第二步，窗口滑动的过程是在检测图像中滑动一个固定大小的窗口，将窗口中的子图像作为候选区。第三步，特征提取的过程是利用特定的算法从子图像中提取特征。第四步，模板匹配的过程是利用特定的距离度量算法计算子图像特征与模板图像特征之间的相似度，根据一定阈值判定子图像是否与模板图像匹配，若子图像与模板图像匹配，则保留该候选区，否则排除该候选区。第五步，后处理的过程是合并与同一类模板相交的候选区，计算出每个目标的边界框，完成目标检测。

图 1.1　基于模板匹配的目标检测方法流程示意图

虽然基于模板匹配的目标检测方法在早期就获得较多的关注 [8−11]，但是随着研究的深入，基于模板匹配的目标检测方法在应用场景上受到一定限制，主要有以下三点原因。第一，因为该类方法通常选择刻画目标外形的特征作为模板匹配的特征，如边缘特征 [12, 13]、纹理特征 [14, 15]、梯度特征 [16, 17] 等，所以对内部纹理复杂的目标匹配效果较差。第二，因为该类方法严格按照距离度量方法 [18, 19] 计算两幅图像之间的相似度，所以发生局部形变的目标匹配效果较差。第三，因为该类方法需要针对每个类别收集一些具有代表性的图像制作为模板图像，如果增加检测目标的类别，会相应地添加模板图像，导致目标检测的速度变慢，并且在一定程度上提高了产生误报的可能性。因此，基于模板匹配的目标检测方法只适用于简单的场景。

2. 基于图像分类的目标检测方法

为了在复杂的场景中进行目标检测，研究者提出了基于图像分类的目标检测方法，主要思路是首先收集大量的目标图像和非目标图像来训练图像分类器，然后利用分类器逐个判断检测图像中每个子图像的类别，最后把输出值超过一定阈值的子图像判定为目标。

如图 1.2 所示，基于图像分类的目标检测方法流程主要分为六个步骤：预处理、窗口滑动、特征提取、特征选择、特征分类和后处理。第一步，预处理的过程是对检测图像进行图像去噪、图像增强、色彩空间转换等操作。第二步，窗口滑动的过程是在检测图像中滑动一个固定大小的窗口，将窗口中的子图像作为候选区。第三步，特征提取的过程是利用特定的算法从候选区中提取特征。第四步，特征选择的过程是从特征向量中挑选出具有代表性的特征，提高特征的鲁棒性，同时降低特征的维数。第五步，特征分类的过程是利用特定的分类器对特征进行分类，若输出值超过一定阈值，则候选区为目标并判定目标的类别，否则候选区为背景。第六步，后处理的过程是合并判定为与同一类别相交的候选区，计算出每个目标的边界框，完成目标检测。

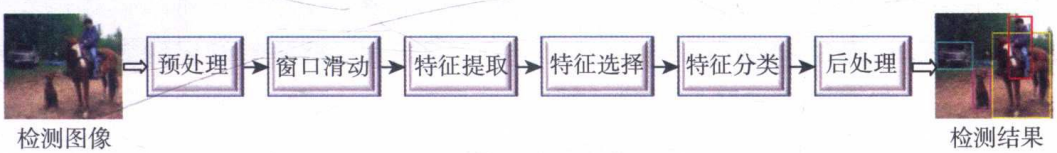


图 1.2　基于图像分类的目标检测方法流程示意图

基于图像分类的目标检测方法重点研究如何在特征提取时提高特征的表达能力和抗形变能力，以及如何在特征分类时提高分类器的准确度和计算速度。由此，研究者提出了种类繁多的特征和多种形式的分类器，代表性的特征有尺度不变特征变换(scale invariant feature transform，SIFT)[20]、加速鲁棒特征 (speeded-up robust features，SURF)[21]、Haar 特征 [22]、方向梯度直方图 (histogram of oriented gradient，HOG)[23]、Strip 特征 [24] 等。另外，代表性的分类器有 K 最近邻 (K nearest neighbours，KNN)[25]、支持向量机 (support vector machine，SVM)[26]、形变部件模型 (deformable parts model，DPM)[27]、随机森林 (random forest，RF)[28]、贝叶斯模型 (Bayesian model，BM)[29] 等。

虽然基于图像分类的目标检测方法取得了一定效果，但是由于使用了手工制作的特征 [20−24]，该类方法即使运用分类能力较强的非线性分类器，目标检测的准

确度也达不到实际需求。这是因为手工制作的特征存在以下缺点：第一，手工制作的特征属于低层特征，对目标的表达能力不足；第二，手工制作的特征可分性较差，导致分类的错误率较高；第三，手工制作的特征具有针对性，很难选择单一特征应用于多目标检测，例如，Haar 特征主要用于人脸检测，HOG 特征主要用于行人检测，Strip 特征主要用于车辆检测。

随着对特征提取研究的深入，研究者发现卷积神经网络可以从大规模数据中学习更好的特征，克服手工制作特征的缺点。第一，因为卷积神经网络通常对输入图像交替进行卷积操作和池化操作，特征经过多次非线性变换，逐步从低层特征抽象为高层特征，所以特征对目标具有较强的表达能力。第二，因为卷积神经网络融合了特征提取、特征选择和特征分类的功能，并且采用端到端的训练方式，所以特征具有一定程度的可分性。第三，因为训练卷积神经网络的数据集包含了多个目标类别的样本，所以特征不针对某个特定的目标类别。综上所述，卷积神经网络提取的特征克服了手工制作的缺点，使基于卷积神经网络的目标检测方法受到广泛关注，成为当前计算机视觉领域的研究热点之一。

实际上，基于卷积神经网络的目标检测方法并不是近几年才提出的，早在 1994 年卷积神经网络就成功地应用于目标检测 [30]。但是，由于训练数据的缺乏、硬件性能的限制、过拟合 (overfitting) 等问题，基于卷积神经网络的目标检测方法在很长一段时间里没有取得进展。与当时的传统目标检测方法相比，无论是在检测准确度上还是在检测速度上，基于卷积神经网络的目标检测方法都没有太大的优势，因此这类目标检测方法逐渐被研究者忽视。直到 2012 年，卷积神经网络 AlexNet模型 [31] 在 ImageNet 大规模视觉识别竞赛的图像识别任务上取得了重大的突破，研究者才开始重新审视卷积神经网络，讨论如何将卷积神经网络更加有效地应用在目标检测中。如今，基于卷积神经网络的目标检测方法已经超越传统的目标检测方法，成为当前目标检测的主流方法。

为了清晰地阐述基于卷积神经网络的目标检测方法的研究现状，本章根据卷积神经网络在目标检测中的使用方式，将基于卷积神经网络的目标检测方法分为两大类，即基于分类卷积神经网络的目标检测方法和基于回归卷积神经网络的目标检测方法，并在以下两个小节中进行详细介绍。

1.2.2　基于分类卷积神经网络的目标检测方法

由于卷积神经网络本身具有特征提取、特征选择和特征分类的功能，并且可以