学校编码：10384　　分类号_____密级_____
学号：31520121153000　　UDC_____

厦门大学

硕 士 学 位 论 文

基于深度学习的特定场景下的行人检测方法研究

Research on Scene-specific Pedestrian Detection Method Based on Deep Learning

袁德东

指导教师姓名： 苏松志　助理教授

专 业 名 称： 人 工 智 能 基 础

论文提交日期： 2015　年　5　月

论文答辩时间： 2015　年　5　月

学位授予日期： 2015　年　　月

答辩委员会主席：_____

评　　阅　　人：_____

2015 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

行人检测一直是计算机视觉中，尤其是目标检测领域的研究重点。而由于环境的时变性和多样性，不同场景之间样本不满足同分布，普通场景下训练得到的检测器直接应用到某一特定场景时性能会急剧下降。基于这样的背景，本文依据场景的复杂程度不同提出了两种对应的解决方法，主要工作和创新点如下：

1）当背景比较简单时，容易获取样本的标定信息，本文提出了一种快速的行人检测算法，首先基于结构化局部边缘模式计算样本特征，然后通过积分图技术快速计算分类器分数，从而使得检测过程更加高效。实验表明该方法在保证检测精度的情况下提高了行人检测的速度。

2）当场景比较复杂时，我们必须要有足够的样本支撑去训练一个鲁棒性的行人检测器，然而对所有特定场景中数据集都进行人工标定，是一个耗时耗力的工作。针对这个问题，本文提出一种在不需要任何人工标定的情况下的特定场景行人检测算法：在特征学习阶段，多层卷积稀疏编码(Multi-stage convolutional sparse coding)用来学习样本的深层信息，一方面这些信息可以作为先验知识，以重构误差的方式权重化普通场景中的训练样本，指导后面迁移学习的进行；另一方面无监督预训练扮演着正则化的角色，当初始化点限制在较小范围时，其作为先验知识克服了后面有监督训练时候的弥散问题，从而使得优化的时候更加准确快速。在分类器训练阶段，基于置信度编码 MLP(Confidence-encoded MLP)，利用新设计的目标函数将普通场景中的样本迁移到特定场景中进行分类器训练，当样本适用特定场景，则赋予其较高的权重，不适应的则权重较低。通过在 INRIA、MIT 交通数据集、Caltech 和 TUD-Brussels 数据集上实验验证了该方法针对特定场景时行人检测器训练的有效性。

关键词：行人检测，特定场景，多层卷积稀疏编码，置信度编码 MLP

Abstract

Pedestrian detection is a key problem for surveillance, automotive safety and robotics applications, however due to the environmental degeneration and diversity, The performance of a generic pedestrian detector may drop significantly when it is applied to a specific scene due to the mismatch between the source training set and samples from the target. In order to solve this problem, we propose two corresponding solutions:

1) When the background of the scene is simple, it is easy to obtain the labeled samples, so a simple Structured Local Edge Pattern (SLEP) is proposed to extract and encode local edge cues, and an integral image based acceleration is proposed toward fast classifier score computation by transforming the classifier score into a linear sum of weights. Experimental results on CASIA gait recognition dataset show that our proposed method is highly efficient than most existing detectors.

2) When the background of the scene is complex, in order to get a robust pedestrian detector, we must get enough samples to support the training stage. However, all training data in specific scene manual labeled are time-consuming. So we propose a deep model to automatically learn scene-special features in static video surveillance without any manual labels. Multi-stage convolutional sparse coding are used to excavate the deep information of the samples, it based on unsupervised learning to pre-train the filters from target training set, followed by supervised fine-tuning from target samples. This method reduces the redundancy between feature vectors at neighboring locations and improves the efficiency of the overall representation compared with patch based method. In the classifier training stage, the source samples are weighted by confidence scores. Target samples with higher scores have larger influence on training scene-specific detectors. All these considerations are formulated under a single objective function called confidence-encoded MLP, The effectiveness is demonstrated through experiments on INRIA, MIT Traffic, Caltech and TUD-Brussels data sets.

Keyword: Pedestrian detection; Scene-special; Multi-stage convolutional sparse coding; Confidence-encoded MLP

目 录

Table of Contents

第一章 绪 论

1.1 研究背景及意义

行人兼具刚性物体和柔性物体的特性，是一类典型的目标，因此行人检测[1-27]一直是计算机视觉，尤其是目标检测领域中的难点和热点[1]。行人检测从二十世纪九十年代中期开始，主要可以分两个阶段：第一阶段从开始至 2002 年，研究者主要采用一些图像处理技术，如图像匹配、如图像分割、边缘提取、光流法等等，以及简单的分类算法，如神经网络、SVM 等对行人进行检测。第二阶段从 2002 年至今，检测算法方面变得更加灵活，大体可以理解为是基于分类的行人检测框架，其中包含特征提取和分类器设计两部分。特征提取包括对新的特征表发方法的研究、如何对特征降维、如何对特征进行选择等技术，分类器设计变得更加复，包含串联分类器、树状组合分类器以及并联分类器等。

最近几年，行人检测取得了非常显著的进步[8-26]，常用的行人特征有梯度方向直方图（Histogram Of Gradient, HOG）特征[2]、基于层级的部位模型匹配特征(Hierarchical Part-Template Matching)[3]，然而基于这些特征的检测器性能在很大程度上依赖于训练集。例如，当用 Caltech 或者 TUD-Brussels 数据集训练得到的行人检测器拿到 INRIA[4]和 MIT 交通数据集[5]上进行测试，得到的结果中会包含很多的误报（False Alarms）和漏检(Missing Rate)，导致性能急剧下降。这是因为两种场景下的表观特征会因为图像分辨率、视角、场景变化的不同而产生很大的差异，旧场景下训练得到的检测器并不能满足新场景下的检测要求。从机器学习的角度看，即训练样本和测试样本之间不是同分布的。

现实中，随着科技的发展，人们渴望技术改变我们的生活。智能监控、智能辅助驾驶，老年人和残疾人照看等领域受到越来越多的认可，这带来了机遇的同时，也带来了挑战：在某个场景下训练的分类器，如何在跨场景的情况下稳定工作，是未来行人检测研究中的一个重要方向，其本质上可视为是跨领域的目标检测。虽然跨领域下的分类问题在自然语言处理和语音识别等领域已有不少的研究，但在计算机视觉的目标检测领域中，目前相关的研究较少。而构建一个“万能”的检测器，使其工作在任何环境下是一个不合实际的想法。

本文以特定场景下的行人分类检测方法作为选题，所谓**特定场景**，指的是由于分辨率、视角、光照条件、背景等的不同所产生的新的某一具体的场景。

特定场景与普通场景相比，共同点是：（1）特征空间相同（2）由于视角、分辨率、光照条件等存在一部分相似的样本。不同点是：两种场景之间样本分布特性不同。本文以这些场景之间的共性和差异做为切入点，基于深度学习研究特定场景下的行人检测方法，具有重要的理论意义和应用价值。

1.2 研究现状

特定场景下的行人检测是该领域的一个新问题，也是一个关键技术难题，它的解决自然是以行人检测技术为基础。所以本章首先介绍行人检测的研究现状，接下来针对特定场景下的行人检测方法做一些简单的梳理，最后对所面临的技术挑战作一个总结。

1.2.1 行人检测研究现状

Papageorgiou[6]等人是第一个提出采用滑动窗口进行行人检测的，他们采用 SVM 和多尺度 Haar 小波过完备基结合的方式进行行人检测。而 Viola 和 Jones[7]则基于这种思路，用积分图来达到快速特征计算的目的，利用 AdaBoost 算法来进行自动特征筛选。上述这些思路都构成了如今行人检测算子的基石。

受到 SIFT(Scale-invariant feature transform)算子的启发，Dalal 和 Triggs[2]等人提出了梯度直方图特征用于行人的特征描述，并通过实验证明了 HOG 比基于灰度的特征更富有信息。而 Shahua[8]等人也提出了一种类似的方法来刻画行人。自此以后，基于 HOG 的变种方法开始急剧增加，而所有的这些变种，几乎都在一定程度上采用了 HOG 算子的一些思想。形状特征也是一个对行人检测有效的特征描述方法。Gavrila[9]等人利用 Hausdorff 距离变换和一种分层模板匹配方法来快速检测行人。Wu 和 Nevatia[10]则利用大量的线段和曲线，构成一种称之为“Edgelet”的特征来局部的表达形状特征。有研究人员还利用 Boosting 方法来学习头部、躯干、腿部以及全身的检测算子。类似的，有研究人员提出一种称之姿态子的特征，它是一种基于局部图像区域的梯度来刻画形状特征的。

运动则是行人检测中的另一个重要线索。然而，在摄像机运动的情况下，有

效的利用运动特征则是一个具有挑战性的课题。在相机固定的情况下，Viola 等人提出通过计算不同图像的 Haar-like 特征，可以获得较好的性能提升。而对于摄像机不固定的情况，则需要将运动分类进行分解。Dalal[11]等人利用光流场来对图像内部的运动进行统计建模，然后在图像局部区域内进行一定的运动补偿。

就单个手工设计的特征而言，目前还没有其他特征描述算子比 HOG 算子更加有效。当然，可以将其它特征跟 HOG 特征结合起来，达到补充的作用。Wojek 和 Schiele[12]研究发现，通过将 Haar-like、shapelets、形状上下文、HOG 特征进行组合，将会比任何其它单独特征描述算子更加有效。而 Walk[13]等人在此基础上考虑了颜色自相关(CSS)和前面提到的运动特征。类似的，Wu 和 Nevatia[10]将 HOG、Edgelet 和协方差特征进行结合。Wang[14]等人则提出将基于 LBP 的纹理特征和 HOG 算子相互结合，此外，还将 SVM 分类器进行改进，以便使其更加适用于遮挡的情况。也有人提出将局部三值模式、颜色信息、隐式分割等同 HOG 进行结合。Dollar[15]等人在 Viola 和 Jones 的基础上进行扩展，提出在多个通道上进行 Haar-like 特征提取，包括 LUV 颜色通道，灰度，梯度幅值等，该方法可谓一个多种特征的大杂烩。当然，上述方法相比单纯的 HOG 而言，在性能上都有一定程度的提升。

目前最新的研究成果中，Zhang S[16]提出了一种过滤通道特征（Filtered Channel Features）框架，主要思想为：对输入图像进行线性或非线性变换得到不同的底层特征图（Feature Maps），接下来利用滤波器与之卷积后得到相应的卷积映射图，获得的特征向量输入到决策森林中进行 boosting 训练，boosting 方法起到了特征选择的作用。作者的实验表明仅利用 HOG+LUV 作为底层特征，使用合适的滤波器就能获得非常好的性能。

1.2.2 特定场景下的行人检测研究现状

相对于传统的目标检测研究，对于特定场景下的行人、目标检测研究相对较少。但尽管如此，对于特定场景下的行人检测依旧取得了不错的进步。Rosenberg 和 Hebert[28]最开始提出了一种半监督训练模型，该模型利用部分完整标定的数据样本再加上弱标定（weakly labeled）后的样本通过迭代训练获得一个适用于特定场景下的行人检测器，Viola 和 Levin A [29]则仅仅只需要部分样本被标定，然

后基于两种不同形式的特征训练得到一对检测器，其中一个检测器用于扩充另一个检测器所需要的训练样本，这种模型必须保证这两种形式的特征相互独立。Nair 和 Clark[30]提出了基于运动信息的无监督训练模型，方法是利用背景减法获得的结果标定训练数据集。针对特征场景下的行人检测，目前最新的研究方向主要是基于迁移学习，Pang J[31]提出了半监督的迁移学习方法，基于普通场景与特定场景之间样本概率分布相同这点出发，利用少量标定样本训练一个针对视角和场景自适应的分类器。Wang[32]等人首先提出了一种基于多线索信息指导的自适应训练框架，将普通场景中的训练样本迁移到特定场景中，但这其中采用的是硬阈值来对样本进行选择。为了提高迁移的可靠性，[33]引入了基于置信度编码的 SVM(Confidence-encoded SVM)，通过一个目标函数，权重化每一个训练样本的贡献度。Xingyu Zeng[34]在这个基础上做了进一步的提升，利用深度学习学习到的特征代替手工设计特征，采用无监督来学习特征场景下的分布特性，一定程度上避免了容易出现的过拟合问题。

1.3 存在问题

特定场景在这里可以分为两种，一种是简单背景的，一种是复杂背景。

简单场景下的行人检测，因为场景的光照、背景等比较简单，我们只需要少量的数据集就可以获得很好的效果，这种情况下就要提高检测的效率，如何在保证检测效果的前提下提高速度是我们所要解决的问题。

针对复杂背景的特定场景下的行人检测，如：雨雪天气、低分辨率、俯拍、监控场景、拥挤场景等，所面临的最大问题在于：场景的多样性导致我们不可能训练出一个“万能”的检测器，可以对所有的场景都有很好的检测效果。复杂背景的场景，特别是规模比较大的时候，就要求我们保证有足够多的数据集，但如果对所有数据集都进行人工标定，是一个耗时耗力的工作，并不可取。而现有公开的数据集已经非常多了，充分利用这些已有的数据集来提供我们特定场景下的检测器训练可以是我们研究的一个方向。但新场景与旧场景中训练分类器的数据来源往往不同，我们不能直接将旧场景下训练得到的检测器用到新场景，因此带来两个问题：1）普通场景与特定场景中的样本分布特性往往不同，旧场景训练得到的检测器因为不能适应新的场景从而导致检测的结果不好；2）普通场景与