关于目标检测中行人检测的文献综述

王起凤

摘要：

行人检测是计算机视觉研究领域中的重要领域。在智能交通、安防监控、无人车等等方面，行人检测技术发挥着重要的作用。随着深度学习的兴起，基于深度学习的行人检测技术流行起来，但是由于很多应用场景不能提供深度学习方法所需的计算资源，所以如何解决将基于深度学习的方法部署在嵌入式设备上这一问题，有着巨大的商业前景和应用价值。本文首先讨论了行人检测过去 20 多年来的经典技术[1]，阐述了传统行人检测方法，并分析了它们的优点和局限性。然后对目前主流的基于深度学习的行人检测做了对比分析，最后，本文将行人检测技术的应用平台定位于车载嵌入式视觉系统上[2][3]，分析总结目前已有的行人检测技术遇到的问题和挑战。

关键字：行人检测；人工智能；嵌入式 AI

1. 引言：

行人检测可定义为：判断输入图片（或者视频帧）中是否出现行人，如果出现，则用检测框将行人框出并给出置信度。行人检测一直是计算机视觉领域当中的难点与热点。由于人体的外观特征非常容易产生变化，非常容易受到尺度、遮挡、姿态、视角[4][5]等因素的影响，且这些特征还会相会影响，给精确检测带来了非常大的挑战。

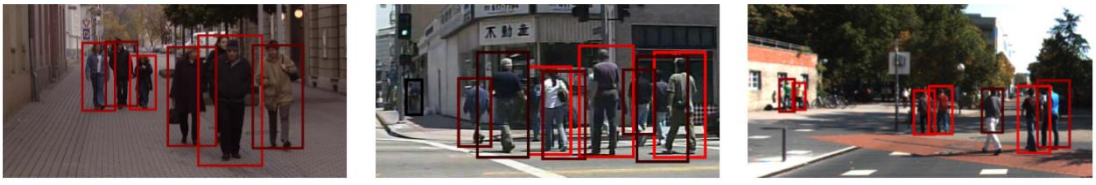

图 1.行人检测系统[6]

随着时代发展，智能交通系统概念的提出，无人驾驶研究的兴起，对于行人检测的检测精度和速度都有了更高的要求[7]。在这样的前提下，行人检测技术发展十分迅猛。特别是大数据时代深度学习技术的流行，基于深度学习的行人检测技术在精度上有了极大的提高，在企业和高校的联合推动下，数据库的质量与数量都与日俱增，且近年来强调 AI 技术落地的声音越来越大，对于行人检测技

术的要求趋向于实用化，即算法必须具备一定的鲁棒性以及实时性。

但是，在很多应用场景下，如无人驾驶汽车，车载的嵌入式系统很难满足目前深度学习方法所需要的计算能力，这导致基于深度学习的行人检测方法很难落地应用，真正的造福社会。如何将检测效果好的深度学习方法应用在资源有限的嵌入式平台上是我们目前面临的一个挑战性问题。

在这样的形势下,有必要及时地对行人检测技术的现状进行梳理。一般来说，这个技术最为重要的问题可以分为两个：第一个是特征提取，第二个是定位与分类[8]。本文以自动驾驶，车载辅助驾驶系统为场景，车载嵌入式系统为平台，对行人检测技术进行分析和讨论。文章安排如下：第 2 节讨论传统的行人检测方法；第 3 节总结目前基于深度学习的行人检测方法；第 4 节从实时性和检测精度的角度讨论基于深度学习方法的行人检测技术应用在嵌入式平台上的问题与挑战。

2. 传统的行人检测方法

本节我们将首先讨论传统的手工设计提取的底层特征的方法，随后讨论传统目标检测方法的定位方法，然后分类总结在得到感兴趣区域后传统的对其分类的方法，给出了它们应用在今天的车载辅助系统，无人驾驶中存在的不足。

2.1 手工设计提取的底层特征

2.1.1 HOG(Histogram of Oriented Gradient, HOG)算子

HOG 算子[9]是非常经典的一种目标检测算子，被广泛使用在传统的行人检测算法中。由于图像的形状能够被梯度的方向密度很好的描述，HOG 就是利用了这一点，使用梯度直方图来作为图像的特征。但是对于行人检测来说，一旦目标有了遮挡，或者姿态发生改变，又或是尺度跨度大，它就很难准确检测了。并且 HOG 还有特征维度高、计算慢等缺点。很多学者针对它的不足，做了许多改进。[10][11]

2.1.2 局部二值模式(Local Binary Pattern, LBP)

LBP 算子[12]是一种描述图像局部纹理特征的目标检测算子，它具有旋转不变的性质，且它的特征维度低有利于计算。但它的缺点也很明显，当光线环境复杂多变时，该算子的缺陷问题就暴露出来了，这时的 LBP 特征也会发生变化从而影响检测结果。对于车载视觉系统上的行人检测，所处的检测环境十分复杂，这

样的缺陷是不利于行人测的，且由于算法的原因，此特征具有旋转不变性的同时，也丢失了方向信息。

2.2 定位与分类

行人检测中的分类指的是判断当前检测窗口是否包含行人,定位指的是行人在图片中的具体位置。下面介绍一下这些技术的经典方法。

2.2.1 定位

目标检测中的定位方法比较经典的就是滑动窗口法(Sliding Window)，算法中心思想如下：选取一个固定尺寸的窗口，在图像上按照一定的次序，依次滑动窗口,每次窗口都会框住图像中的一部分像素，这片像素区域就是我们需要检测的区域，到这里就完成了定位。随后就利用已经训练好的分类器检测框内是否出现了行人,通过窗口的滑动，不断的从原图像中框出一块区域来进行判断，从而给出行人在图片中出现的位置。滑动窗口的算法思想非常简单，且确实有效。但是它这种等同于遍历的方法，有着相当大的计算冗余，造成了很大的计算机资源的浪费。

在滑动窗口法的基础上，利用人体结构信息可以将定位方法进一步的划分为整体法和部位法。整体法是对滑动框取得感兴趣区域中的像素计算关键特征,得到行人的全局信息；部位法对人体的部位，诸如头部、手和脚等，进行特征提取,构建部位间的几何空间关系，从属关系。用这样的定位方法可以有效的缓解遮挡问题造成的影响[1]。

2.2.2 分类

分类一般采用机器学的方式，常用的机器学习分类器有：SVM、Boosting 和多实例学习等。依据上面的定位检测方式，下面将介绍两种定位检测方式下，采用不同分类方法的一些研究。

2.2.2.1 基于整体法的分类

Oren 等[13]最早提出基于 SVM 的行人检测方法，它将图像进行预处理后提取到的特征用 SVM 进行分类，在当时取得了不错的效果。

Viola 等[14]利用哈尔特征、AdaBoost 算法和级联分类器融合起来，实现了一个完整的人脸检测系统，并且系统具有实时性，并且成功的将这个方法整合到了智能监控领域中的行人检测中[15]，但该系统的算法鲁棒性不足，很难用于高

安全要求的车载嵌入式系统上。

2.2.2.2 基于部位法的分类

基于部位法的定位检测方法将人体看成是部位的组合,是一种基于部件的检测方法。这种方法对目标的遮挡有着一定的抵抗性。如何建立有效的部位检测模型？如何对部位间的几何关系，空间关系，从属关系进行建模？是这个方法面临的挑战性问题。

Felzenszwalb 等[16]在早些年提出了形变部位模型(Deformable Part Model, DPM)。这种基于部件的检测方式虽然能克服部分遮挡的影响，但是还是无法解决尺度变换，视角变换，旋转变换等引起的问题，算法鲁棒性比较差，且部位之间的空间关系很难精确的描述，人工设计特征工作量巨大，特征本身难于学习。

整体法的直观缺点是无法抵抗遮挡这类形变,部位法使用分治法的思想,能够抵抗检测目标的一些姿态上的形变或者是遮挡造成的形变。部位法的难点在于如何有效的对部位进行划分以及如何准确描述部位与整体的空间位置关系和从属关系。Parikh 等[17]的研究指出，部位之间的空间位置关系，几何关系对于算法检测效果的贡献很少，而部位本身的检测精度对算法检测效果贡献比较大。对于一幅图像中的多个行人，各个部位分别属于哪个行人，以及出现多个相同部位的时候会不会有冲突从而对检测结果造成影响，也是部位法需要解决的问题。

3. 深度学习方法

传统方法虽然需要的计算资源相对较少,但是检测性能不能满足当今应用场景的要求。近年来，越来越多的学者在探索，将深度学习应用在图像分割和图像分类以及目标检测等应用中。深度学习中最广为使用的是卷积神经网络(Convolutional Neural Network, CNN)[18]，深层的神经网络在使用大量数据进行训练后，可以提取出图像的深层特征，将这些深层特征用于各类视觉任务都取得了非常好的效果，比之前的方法在精度上有非常大的提升。

采用基于 CNN 来解决提取高层特征，以及分类的问题，还必须解决定位的问题。可以使用经典的滑动窗口方法来获取检测框，在通过极大值抑制来筛选检测框，但是这个方法由于目标尺度大小不一，且当今图片分辨率普遍比较高的情况下，遍历范围非常大，有非常多的冗余计算。

现在的方法一般采用的是，利用选择性搜索（selective search，SS）方法创建目标检测的感兴趣区域(ROI),然后对这个区域内的目标进行分类。近年来，已经把这个方法做进了神经网络里，即用神经网络来生成这些区域。典型的就有RCNN[19]、Fast-RCNN[20]以及 Faster-RCNN[21]。其中 Faster-RCNN 是现在非常流行的一个基于深度学习的目标检测方法。它使用一个卷积网络提取目标候选区域，这个网络叫(Region Proposal Network, RPN)，RPN 可以提供更少并且质量更高的候选区域，并且实现了目标检测模型的端到端训练和测试。很多用深度学习方法来做行人检测的算法，所应用的基础框架大多都是 Faster-RCNN。它使用 CNN 结构来提取图像的特征图，对比 Fast-RCNN,它有一个 RPN 网络结构来完成 ROI 的提取过程，大大加速了算法的运行速度。

把应用场景放在车载系统上，在车载系统里，首先要保证算法的性能，检测效果必须要过关，检测速度必须要快。如果做不到这两点，那就无法保证驾驶员和乘客的安全性，这样的系统是没有任何意义的。要保证算法的性能，必须要考虑到车载系统是一个嵌入式系统，它的计算资源是受限的，在这样的情况下，要保证检测性能和实时性是非常具有挑战性的。

之前的深度学习算法检测行人的效果确实不错，但是它的模型参数数量巨大，需要的计算资源太多，以嵌入式设备的内存，算力是不能满足的。即使是 Faster-RCNN 也仍然达不到实时性的要求。再这种情况下，势必要对网络进行改进，以及引入一系列轻量级网络技术。

3.1 提升精度

zhang 等[22]从 anchor 出发，通过设计一种 loss，使得训练过程中多个匹配到真实目标上的 anchor 尽量靠近，从而可以更好的检测互相遮挡的行人。Wang 等人[23]也从 loss 的角度入手，设计了 Repulsion Loss，使得预测框更加接近所负责的真实目标框，而远离周围的目标。

但是对于车载视觉系统来说，行人间的相互遮挡没有那么关键，反而是障碍物与行人间的遮挡才会造成安全问题。而这些问题，上面的研究并没有解决。

Mao 等人[24]用额外的特征来提升检测器的性能，她们在原来的 faster-rcnn 的基础上加入上加入了新的网络分支，用来额外引入梯度、边缘、像素分割等额外特征，提升了网络的性能。Luo 等人[25]在 3D 数据中，引入了时间维度，

一起构成了 4D 数据，使得网络能在有检测目标能力的基础上，还具有具备跟踪和预测的能力。融合更多的上下文信息是有利于检测结果，但是也增加了计算复杂度。

Law 等人[26]用一个卷积网络预测所有同一类别的样本的左上角点的 heatmap，及右下角点的 heatmap,及用于检测到角点的 embeding vector。这个方法没有用目前非常流行的 anchor box 的方式来检测的方法，取得了不错的效果。但是这种方式较一般的神经网络方法更加缺乏可解释性。

Liu 等人[27]提出了一种改进的行人检测器，享受 SSD 网络的速度，同时保持更快的 R-CNN 家族所具有的准确性。具体而言，提出了一种结构简单但有效的模块，称为渐近定位拟合（ALF），它叠加了一系列预测器，可以逐步直接演化 SSD 的默认锚框，从而改进检测结果。

上面这些网络，由于模型参数体量过大的问题，都难以应用在嵌入式设备上。

3.2 提高实时性

Kim 等人[28]提出了一种轻量级网络结构 PVANet，糅杂了 C.ReLU[29]，HyperNet[30]，Inception 模块。C.ReLU 模块压缩了参数，Inception 模块由于具有多种感受野的卷积核组合，因此能够适应多尺度目标的检测，HyperNet 相当于融合了多尺度的信息，增大了信息量。这些技术在保证了检测效果的同时，满足了实时性。但是 PVANET 仍然基于 anchor 的形式来检测目标，这种方式需要找到合适的尺度和形状，模型泛用性差。

Howard 等[31]提出了 mobileNet 网络，作者运用一种新的方法深度可分离卷积(depthwise separable convolutions)大大减低了运算量。但其基于 VGG16 的网络结构，512 层的深度以及 1024 通道的运算仍然有很大的参数量。

Zhang 等[32],提出了 shufflenet，这是一个可用于移动设备的目标检测网络。它提出了一个 Group Convolution，这是一个更高效的网络结构，它可以实现模型变小和变快。

Sandler 等[33]改进了 mobileNet 网络，提出了 mobileNet v2，引入残差结构，对 feature map 进行升维随后再降维，减小了因为层数过多引起的梯度减小。放弃了 ReLU 层，转而采用 Linear 的方式，这样可以保留特征的多样性，防止特征的破坏。

上面两个方法，都减小了模型参数量和 FLOPS 运算量，但是有相关研究指出这两个指标降低并不能表明最终的运行速度就快。

Ma 等[34]提出了高效网络架构设计应该考虑的两个基本原则：第一，应该用直接的评估标准替换间接标准，例如用实际的检测速度来替换浮点数计算量；第二，这些标准应该在规定的真实物理平台上进行评估。在这项研究中，作者遵循这两个原则，并提出了一种更加高效的网络架构 shufflenet v2。但是仍然面临着检测精度不足的问题。

4. 总结展望

能用于车载嵌入式视觉系统的行人检测技术必须满足下面二个要求：其一，实时性，一个可靠的无人驾驶系统或者车载辅助驾驶系统必须能对实时采集到的传感器数据进行实时处理；其二，可靠性，系统的检测功能的检测必须具备良好的鲁棒性，能够适应各种极端状况，保证检测结果的准确性。

可靠性：总的来说，目前行人检测技术的检测效果在不是特别极端的场景下已经达到可用的标准。但是，如果是应用在汽车上，由于汽车对于安全的要求极高，汽车必须要最大限度保障乘客的人身安全，目前检测效果还远远达不到应用标准。算法仍然不能很好的解决行人的各种形变带来的系统性能下降的问题。

实时性：当汽车上搭载的多个摄像头同时工作时，将产生巨额的数据量。且深度学习算法的模型参数是巨大的，对单帧图像而言，都要经过一段时间的计算才能得到结果。这对于车载辅助驾驶系统，以及无人车这些实时系统而言是灾难性的。如何设计出一种轻量级的行人检测神经网络，能够降低检测精度，甚至提高检测效果的同时，保证在资源受限的嵌入式平台上的实时性是一个迫切需要解决的问题。

总体来说，目前的行人检测技术的检测效果已经不错了。但是行人检测技术应用领域广泛，有许多问题还没有解决。比如，行人的尺度变换，形变（姿态，外貌）等变换都会让目前的检测器的检测效果下降不少。这要是放在汽车上，是不能满足它的安全可靠性这一要求的。且现在流行的深度学习方法也没有一个足够好的网络模型，可以同时满足检测效果以及实时性的要求，可以在资源受限的嵌入式设备上好的完成检测任务。如何压缩神经网络模型，设计一个整体的轻量

级网络结构也是一个值得研究的方向。

主要参考文献

[1] 苏松志，李绍滋，陈淑媛, et al. 行人检测技术综述[J]. 电子学报, 2012, 40(4):814-820.

[2] Menze M, Geiger A. Object scene flow for autonomous vehicles[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3061-3070.

[3] Mukhtar A, Xia L, Tang T B. Vehicle detection techniques for collision avoidance systems: A review[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(5): 2318-2338.

[4] Kosaka N, Ohashi G. Vision-based nighttime vehicle detection using CenSurE and SVM[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(5): 2599-2608.

[5] Wang C, Fang Y, Zhao H, et al. Probabilistic inference for occluded and multiview on-road vehicle detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 17(1): 215-229.

[6] Benenson R, Omran M, Hosang J, et al. Ten years of pedestrian detection, what have we learned?[C]//European Conference on Computer Vision. Springer, Cham, 2014: 613-627.

[7] Ranft B, Stiller C. The role of machine vision for intelligent vehicles[J]. IEEE Transactions on Intelligent vehicles, 2016, 1(1): 8-19.

[8] Engel J I, Martin J, Barco R. A low-complexity vision-based system for real-time traffic monitoring[J]. IEEE Transactions on Intelligent Transportation Systems, 2016, 18(5): 1279-1288.

[9] Dalal N , Triggs B . Histograms of Oriented Gradients for Human Detection[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005.

[10] Zhu Q, Yeh M C, Cheng K T, et al. Fast human detection using a cascade of histograms of oriented gradients[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, 2006, 2: 1491-1498.

[11] Wojek C, Schiele B. A performance evaluation of single and multi-feature people detection[C]//Joint Pattern Recognition Symposium. Springer, Berlin, Heidelberg, 2008: 82-91.

[12] Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on featured distributions[J]. Pattern recognition, 1996, 29(1): 51-59.

[13] Oren M, Papageorgiou C, Sinha P, et al. Pedestrian detection using wavelet templates[C]//cvpr. 1997, 97: 193-199.

[14] Viola P, Jones M J. Robust real-time face detection[J]. International journal of computer vision, 2004, 57(2): 137-154.

[15] Viola P, Jones M J, Snow D. Detecting pedestrians using patterns of motion and appearance[J]. International Journal of Computer Vision, 2005, 63(2): 153-161.

[16] 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, 2009, 32(9): 1627-1645.

[17] Parikh D, Zitnick C L. Finding the weakest link in person detectors[C]//CVPR 2011. IEEE, 2011: 1425-1432.

[18] 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.

[19] 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.

[20] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[21] 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.

[22] Zhang S, Wen L, Bian X, et al. Occlusion-aware R-CNN: detecting pedestrians in a crowd[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 637-653.

[23] Wang X, Xiao T, Jiang Y, et al. Repulsion loss: Detecting pedestrians in a crowd[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7774-7783.

[24] Mao J, Xiao T, Jiang Y, et al. What can help pedestrian detection?[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3127-3136.

[25] Luo W, Yang B, Urtasun R. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 3569-3577.

[26] Law H, Deng J. Cornernet: Detecting objects as paired keypoints[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 734-750.

[27] Liu W, Liao S, Hu W, et al. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 618-634.

[28] Kim K H, Hong S, Roh B, et al. Pvanet: Deep but lightweight neural networks for real-time object detection[J]. arXiv preprint arXiv:1608.08021, 2016.

[29] Shang W, Sohn K, Almeida D, et al. Understanding and improving convolutional neural networks via concatenated rectified linear units[C]//international conference on machine learning. 2016: 2217-2225.

[30] Kong T, Yao A, Chen Y, et al. Hypernet: Towards accurate region proposal generation and joint object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 845-853.

[31] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[32] Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural

network for mobile devices[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6848-6856.

[33] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4510-4520.

[34] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 116-131.