基于 Anchor-free 架构的行人检测方法

张庆伍,关胜晓
(中国科学技术大学 微电子学院,安徽 合肥 230026)

摘　要:使用无预选框(Anchor-free)的检测框架,设计了一种行人检测算法。将深度残差网络(ResNet)作为特征提取网络,与特征金字塔网络(FPN)结构相结合,使用了多尺度预测的方式进行预测。把目标中心点和尺寸作为一种高级的语义特征,将含有更多细节信息的浅层特征图和含有更多语义信息的深层特征图进行融合。在 Citypersons 数据集上进行了实验验证,相较现有行人检测算法,提出的算法在轻微遮挡、一般遮挡和严重遮挡情况下漏检率分别提升了 1.11% ~3.01%,0.15% ~6.55% 和 0.59% ~6.39%,检测效果更好。

关键词:无预选框;行人检测;特征融合;多尺度检测
中图分类号:TP183　　**文献标识码**:A　　DOI: 10.19358/j.issn.2096-5133.2020.04.008
引用格式:张庆伍,关胜晓. 基于 Anchor-free 架构的行人检测方法[J]. 信息技术与网络安全,2020,39(4):43-47,52.

Pedestrian detection algorithms based on Anchor-free architecture

Zhang Qingwu, Guan Shengxiao
(School of Microelectronics, University of Science and Technology of China, Hefei 230026, China)

Abstract: This paper designed a pedestrian detection algorithm based on the Anchor-free detection framework. The deep residual network (ResNet) was used as a feature extraction network, combined with the feature pyramid structure (FPN), and finally multi-scale prediction was used for prediction. This paper also regarded the target center point and size as an advanced semantic feature, and combined the shallow feature map with more detailed information and the deep feature map with more semantic information. The experiments were verified on the Citypersons dataset. Compared with the existing pedestrian detection algorithms, the detection results were respectively improved by 1.11% ~3.01%, 0.15% ~6.55% and 0.59% ~6.39% in the case of slight occlusion, general occlusion and severe occlusion, and the detection effect is better.

Key words: anchor-free; pedestrian detection; feature fusion; multi-scale detection

0　引言

行人检测是智能安防和车辆辅助驾驶等实际应用的关键技术。随着计算机视觉技术的快速发展,目标检测算法的性能也不断提升。目前基于深度学习的目标检测方法按照是否提出预选框可以分为两大类:一类是基于预选框的检测算法,该类算法首先预先设置预选框,然后通过预选框和真实目标进行匹配,最终选出合适的预选框进行训练,这类算法以 Faster-RCNN[1] 和 SSD[2] 为代表;另一类是不使用预选框的检测算法,该类算法首先对预测目标的关键点进行标注,然后将深度神经网络的输出设置成相同的格式,直接进行训练,这类算法以 YOLO[3] 和 DenseBox[4] 为代表。其中,Anchor-free 的算法框架结构简洁,更加适用于计算资源较少的实际应用场景。本文在 Anchor-free 算法的基础上,首先使用不同的基础网络构建检测算法,然后选出性能稳定的基础网络,利用特征金字塔结构[5]对不同卷积层上的特征图进行融合,提升检测效果,最后使用多尺度预测的方法,通过不同尺度的预测图生成了更多的检测结果,再次提升了检测效果。本文算法在 Citypersons[6] 数据集上进行了验证,其检测精度相较其他行人检测算法有一定提升。

1　相关工作

1.1　行人检测

到目前为止,行人检测的发展大致分为三个阶段:早期的图像处理阶段(基于背景建模的方法)、特征模型分类阶段(基于统计学习的方法)和深度学习方法阶段。DALAL N 等人[7]提出了梯度方向

直方图特征(HOG)。HOG 基于梯度信息并允许块间相互重叠,用于刻画人体的边缘特征。FELZENSZWALB P 等人[8]提出了 DPM(Deformable Parts Model)算法,这是一种基于部件的检测方法,对目标的形变具有很强的鲁棒性,采用了改进后的 HOG 特征、SVM 分类器和滑动窗口的检测思想。Zhang Liliang 等人[9]将 Faster RCNN 用在了行人检测领域,通过在 RPN 网络后添加随机森林,对 RPN 网络筛选出来的候选区域做分类,提高了网络对于小目标的检测能力。Zhou Chunluan 等人[10]为了解决行人检测中的遮挡问题,提出训练两个卷积神经网络分支网络来分别做全身估计和可见部分估计的方法,并将两个分支网络的输出进行互补,以减少遮挡对检测的影响。Liu Wei 等人[11]提出了一种级联结构,并将其使用在单步检测框架中,在保证检测速度的情况下,进一步提升了检测精度。

1.2 Anchor-free 检测框架

早期的深度学习检测框架多是基于预选框(Anchor-base)的,在使用时,需要在特征图上密集平铺大量预选框,但其中仅有少量样本是正样本,最终导致计算资源花费在无用的样本上。另外,对预选框的预处理进一步增加了算法的计算量。同时,预选框的各项参数都需要人为设置,这使得算法的检测性能会很大程度上受到预设参数的影响,而且调参的过程十分复杂。基于上述原因,Anchor-free 算法开始受到关注。DenseBox 构建了一种 Anchor-free 检测框架,即特征提取网络和检测头组合的方式。FSAF[12]则将 Anchor-base 和 Anchor-free 两种检测方式相结合,在原有的 FPN 网络上添加一个分支网络,同时利用两种检测方式生成的结果,在 COCO[13]数据集上做出了有竞争力的结果。随后 Anchor-free 框架逐渐被使用到行人检测的任务中,Song Tao 等人提出了基于 Anchor-free 架构的行人检测方法 TLL[14],将人体中轴线作为标注。为了提升检测效果,又采用了基于马尔可夫随机场的后处理方案。Liu Wei 等人设计了基于 Anchor-free 的检测方法 CSP[15],将行人的中心点作为标注,其主要关注的是如何用尽量简单的结构完成检测并获得好的效果,其网络结构相较其他工作更加简洁。

Anchor-free 框架抛弃了提出预选框的步骤,将算法的流程简化为特征提取、检测头和后处理三个部分。但一味地简化会造成检测精度的损失,所以本文在 Anchor-free 架构的基础上将算法流程增加为特征提取、特征融合、检测头、后处理四个部分,按照该流程重新设计了卷积神经网络,并使用中心点和高度作为标注训练神经网络,最终在 Citypersons 数据集上进行验证。

2 算法设计

2.1 模型设计

模型的结构共包括三个部分:特征提取网络、特征融合网络、检测头。模型为全卷积的网络结构。通常将特征提取网络中生成的第 N 层特征图命名为 Stage$N$,特征提取网络可以描述为:

$$\phi_i = F(\phi_{i-1}) \tag{1}$$

ϕ_i 表示第 i 层的特征图,ϕ_{i-1} 表示第 $i-1$ 层的特征图,F 表示两个特征图之间的卷积、池化、归一化等操作。可以将 N 层的特征图表述为一个集合:

$$\Phi = \{\phi_1, \phi_2, \cdots, \phi_N\} \tag{2}$$

式中,ϕ_i 表示第 i 层的特征图,Φ 表示所有特征图的集合。然后使用这些特征层构建特征融合模块。构建的方式可以表述为:

$$\begin{cases} \varphi_i = \phi_{i+1} + \phi_i & i = 4 \\ \varphi_i = f(\varphi_{i+1}) + \phi_i & i = 2 \text{ or } i = 3 \end{cases} \tag{3}$$

其中,f 表示反卷积操作,φ 表示特征融合网络的特征图,ϕ 表示特征提取网络的特征图。算法网络结构如图 1 所示。特征提取网络将输入图像转化成不同分辨率的特征图。以 ResNet 为例,ResNet 的所有中间层可以被划分成 5 个部分,本文取其中的后四层,分别下采样至原图大小的 1/16,1/64 和 1/256,然后经过特征融合网络生成特征金字塔结构,最后通过检测头生成预测图。训练时使用 640 × 1 280 大小的图像,将其下采样至 160 × 320,80 × 160,40 × 80 这三种大小,然后分别用检测头生成预测结果。

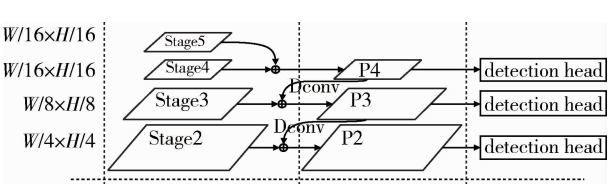

图 1 算法网络结构

相比 CSP 和 TLL 两种方法,本文的方法增加了特征融合网络,代替了原有的对特征图直接反卷积然后拼接的方式。

2.2 标注方法和损失函数设计

2.2.1 标注方法

使用 Anchor-free 框架的检测算法的标注方式有检测框的角点、目标中轴线、目标中心点等。为了减少计算量,本文使用目标中心点和目标高度作为标注。通过该种标注方法,生成一张高斯热图,该图可以表述为:

$$M_{ij} = \max_{k=1,2,\cdots,k} G(i,j,x_k,y_k,\sigma_{w_k},\sigma_{h_k}) \tag{4}$$

$$G(i,j,x,y,\sigma_w,\sigma_h) = e^{-\left(\frac{(i-x)^2}{2\sigma_w^2}+\frac{(j-y)^2}{2\sigma_h^2}\right)} \tag{5}$$

其中,x,y 是目标中心点在高斯热图中的坐标,(i,j) 表示高斯热图中的位置,k 表示目标的标号,(σ_w,σ_h) 为目标的宽和高。生成的高斯热图如图 2 所示。

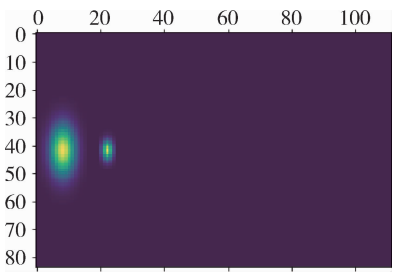

图 2 高斯热图可视化

可以看出,越靠近目标中心点,标注的值越接近 1,越远离目标中心点,标注的值越接近 0。

2.2.2 损失函数

总的损失函数由两个部分组成:中心点损失和目标尺度损失。对于中心点的预测,使用 Focal Loss[16]。为了平衡正负样本对于总的损失函数的贡献,增加了 α_{ij} 项。

$$L_{center} = -\frac{1}{k}\sum_{i=1}^{W/r}\sum_{j=1}^{H/r}\alpha_{ij}(1-P_{ij})^{\gamma}\log(P_{ij}) \tag{6}$$

$$P_{ij} = \begin{cases} P_{ij} & y_{ij}=1 \\ 1-P_{ij} & \text{otherwise} \end{cases}$$

$$\alpha_{ij} = \begin{cases} 1 & y_{ij}=1 \\ (1-M_{ij})^{\beta} & \text{otherwise} \end{cases}$$

式中,W 和 H 为原图的宽和高;r 为预测图的缩放比例;k 表示该原图中目标的个数;P_{ij} 表示预测图中坐标为 (i,j) 的点上的预测值;y_{ij} 表示预测图中坐标为 (i,j) 的点上是否存在中心点;γ 和 β 是两个可以设置的超参数,分别用于控制训练速度和正负样本平衡;M_{ij} 的含义如式(4)所示,表示预测图中坐标为 (i,j) 的点上的标注值。

对于目标高度的预测,使用 Smooth$L1$[17] 损失函数使训练过程更加平滑,函数表述为:

$$L_{scale} = \frac{1}{K}\sum_{k}^{K}\text{Smooth}L1(s_k,t_k) \tag{7}$$

$$\text{Smooth}L1(x) = \begin{cases} 0.5x^2 & |x|<1 \\ |x|-0.5 & \text{otherwise} \end{cases} \tag{8}$$

其中,s_k、t_k 分别是目标高度的预测值和真实值,将 s_k、t_k 的差值作为 Smooth$L1$ 函数的输入。k 是目标的标号,K 表示所有目标的个数。

总的损失函数可以表述为:

$$L = \lambda_c L_{center} + \lambda_s L_{scale} \tag{9}$$

其中,L 表示总的损失函数,L_{center} 和 L_{scale} 的含义如式(6)和式(7)所示。λ_c 和 λ_s 分别设置为 0.01 和 1,用于控制两部分损失函数的比重。

3 实验

为了对本文提出的方法进行有效的评估,使用 Citypersons 数据集进行验证。该数据集共有 2 975 张图片,包含 19 238 个行人,平均每张图片上含有行人 6.47 个。其中被轻微遮挡的目标占整个数据集的 55.9%,被一般遮挡的目标约占整个数据集的 20%,检测难度相对较大,更加适合验证算法的性能。实验均在 Ubuntu16.04 操作系统,GTX080Ti 的 GPU 集群上使用 Keras 框架完成。

为了更好地评估算法的性能,将检测指标分成三类,各种遮挡情况下的数据示例如图 3 所示。严重遮挡示例(遮挡率 60% 以上)如图 3(a)所示,普通遮挡示例(遮挡率 30% ~60%)如图 3(b)所示,轻微遮挡示例(遮挡率 30% 以下)如图 3(c)所示。其中遮挡率是被遮挡部分与目标大小的比值。

(a) 严重遮挡示例

(b) 普通遮挡示例

(c) 轻微遮挡示例

图 3 验证集部分样本图

3.1 不同基础网络的实验效果

特征提取网络是 Anchor-free 算法的基础网络，为了选择合适的基础网络，首先对各种基础网络的性能进行测试，也就是分别利用不同的基础网络进行特征提取，然后比较检测效果。比较的结果如表 1 所示，评估指标使用漏检率（Miss Rate）。

表 1 各种基础网络性能比较 （%）

Epoch	DenseNet	ResNet-50	MobileNet
0 ~ 50	14.80	12.86	16.35
50 ~ 100	13.86	12.50	15.83
100 ~ 150	12.83	12.83	16.86

经过比较，ResNet 的性能最为稳定，MobileNet 的速度更快，但是检测精度略有下降。由于显存爆炸问题，本文将 DenseNet 的批尺寸（Batchsize）设置为 4，但其整体性能仍略逊于 ResNet。所以本文使用 Resnet-50 进行后续实验。

3.2 特征融合策略

为了能将含有更多细节信息的浅层特征图和含有更多语义信息的深层特征图进行融合，本文在算法中加入了特征金字塔网络（FPN）结构。加入 FPN 后的实验结果如表 2 所示，结果从 100 到 150 Epoch 中选取。

表 2 加入 FPN 后的实验效果 （%）

FPN	Reasonable	Bare	Partial	Heavy
NO	12.86	8.82	11.96	51.11
YES	12.21	8.11	11.37	48.84

从表 2 可以看出，加入 FPN 后，轻微遮挡和一般遮挡情况下的检测效果并没有明显变化，遮挡严重情况下的检测效果提升较为明显，检测结果如图 4 所示。产生该现象的原因是被遮挡目标的信息在池化过程中会不断丢失，而浅层特征图能保留这些信息，深层特征图能更好地描述目标的位置信息。所以使用 FPN 融合两种特征图能够提升严重遮挡情况下的检测效果。

图 4 加入 FPN 后的部分检测结果

3.3 多尺度预测

为了获得更好的检测效果，本文采用了多尺度预测的方法。分别使用 FPN 三个不同尺度的特征图生成预测结果。预测得到的高斯热图分别为原图大小的 1/16，1/64，1/256，再将所有的检测结果整合。将结果在 Citypersons 数据集上与其他算法对比，结果如表 3 所示。

表 3 行人检测算法效果对比 （%）

使用方法	基础网络	Reasonable	Heavy	Partial	Bare
FRCNN	VGG-16	15.4	—	—	—
OC-CNN	VGG-16	12.8	55.7	15.3	6.7
ALF-Net	ResNet	12.0	51.9	11.4	8.4
TLL	ResNet	15.5	53.6	17.2	10.0
TLL + MRF	ResNet	14.4	52.0	15.9	9.2
CSP	ResNet-50	11.4	49.9	10.8	8.1
Ours	ResNet-50	11.34	49.31	10.65	6.99

与其他算法的实验结果相比，本文算法在目标遮挡严重和普通遮挡的两个指标上有更好的表现，原因是由于使用了多尺度的检测结构，同一目标的周围产生了更多的高质量预测框，且多尺度预测针对不同大小的目标有着不同的预测效果，小尺度的预测图更加适合预测大型目标，大尺度的预测图对小目标的预测效果更好。

使用多尺度检测方法后，在严重遮挡情况下的漏检率有所上升，但是普通遮挡和轻微遮挡时漏检率均有所下降，检测结果如图 5 所示。产生该现象的原因是由于在严重遮挡情况下，在目标周围生成的预测结果更多，但后处理过程使用了非极大值抑制（NMS），并没有处理这些无效预测框，反而将正确结果抑制删除。

图 5 使用多尺度预测方法后的部分检测结果

从上述结果可以看出，检测失败的情况多为无效检测或重复检测。造成该现象的原因是多尺度预测策略同时增加了正确和无效的检测结果，且无

效检测结果的数量更多。针对该问题，本文尝试了不同的输出阈值，即为预测得到的高斯热图设置一个阈值，若预测值大于该阈值则认为该点存在目标中心点。实验结果如图 6 所示。

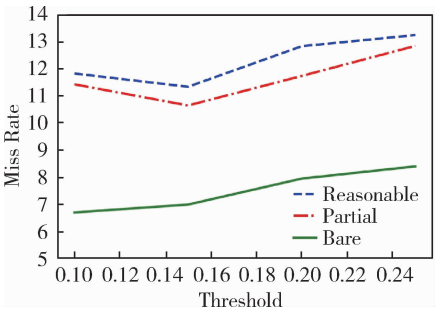

图 6　漏检率和输出阈值的对应关系

从图 6 可以看出，随着输出阈值的提高，漏检率先降低后升高。算法在输出阈值设为 0.15 时检测效果最好。这是由于输出阈值过小会导致算法输出大量的无效预测，而过高的输出阈值又会抑制算法输出正确结果。

本文使用的测试用例共 500 张，测试图片大小为 1 024 ×2 048，测试 10 次取平均值。平均每张图片耗时 0.36 s。将图片缩放至 480 ×640 大小后，平均每张图片耗时约 61 ms，基本满足应用要求。

4　结束语

本文基于 Anchor-free 架构，设计了一种特征提取、特征融合加多尺度检测头的网络结构。融合了不同特征图的信息并对同一目标提出了更多的有效预测，实验证明该方法提高了检测效果。

由实验结果还发现，在使用多尺度预测方法后，目标周围产生的部分无效预测会抑制正确结果，导致检测失败。解决这个问题需要对检测结果做进一步后处理。Anchor-free 架构往往采用 NMS 进行后处理，所以下一步可以通过优化 NMS 进一步提升检测效果。

参考文献

[1] REN S Q, HE K M, 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.

[2] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]. Proceedings of the European Conference on Computer Vision, 2016: 21-37.

[3] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[4] HUANG L C, YANG Y, DENG Y F, et al. Densebox: unifying landmark localization with end to end object detection[J]. arXiv preprint arXiv: 1509.04874, 2015.

[5] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2117-2125.

[6] ZHANG S S, BENENSON R, SCHIELE B. Citypersons: a diverse dataset for pedestrian detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3213-3221.

[7] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015.

[8] FELZENSZWALB P, MCALLESTER D, RAMANAN D. A discriminatively trained, multiscale, deformable part model[C]. Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008: 1-8.

[9] ZHANG L L, LIN L, LIANG X D, et al. Is faster R-CNN doing well for pedestrian detection? [C]. Proceedings of the European Conference on Computer Vision, 2016: 443-457.

[10] ZHOU C L, YUAN J S. Bi-box regression for pedestrian detection and occlusion estimation[C]. Proceedings of the European Conference on Computer Vision, 2018: 135-151.

[11] 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, 2018: 618-634.

[12] ZHU C C, HE Y H, SAVVIDES M. Feature selective anchor-free module for single-shot object detection[J]. arXiv preprint arXiv: 1903.00621, 2019.

[13] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]. Proceedings of the European Conference on Computer Vision, 2014: 740-755.

[14] SONG T, SUN L Y, XIE D, et al. Small-scale pedestrian detection based on somatic topology localization and temporal feature aggregation[J]. arXiv preprint arXiv: 1807.01438, 2018.

[15] LIU W, LIAO S, REN W, et al. High-level semantic feature detection: a new perspective for pedestrian detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 5187-5196.

（下转第 52 页）

参考文献

[1] 徐玉如,李彭超. 水下机器人发展趋势[J]. 自然杂志,2011,33(3):125-132.

[2] 李晔,常文田,孙玉山,等. 自治水下机器人的研发现状与展望[J]. 机器人技术应用,2007(1):25-31.

[3] 庞硕,纠海峰. 智能水下机器人研究进展[J]. 科技导报,2015,33(23):66-71.

[4] 张荣敏,陈原,高军. 无鳍舵矢量推进水下机器人纵向稳定性研究[J]. 哈尔滨工程大学学报,2017,38(1):133-139.

[5] 罗庆生,刘星栋,弓瑞,等. 矢量喷水推进式水下机器人的构建仿真与验证[J]. 应用科技,2017,44(2):7-14.

[6] BUGROVA A I,BUGROV G E,BISHAVE A M,et al. Experimental investigation of thrust-vector deviation in a plasma thruster [J]. Technical Physics Letters,2014,40(2):161-163.

[7] 逯玉明. 水下探测机器人设计与定位导航方法研究[D]. 扬州:扬州大学,2017.

[8] 米歇尔·麦克罗伯茨. Arduino 从基础到实践[M]. 刘瑞阳,郎咸蒙,刘炜,译. 北京:电子工业出版社,2017.

[9] 李永华. Arduino 案例实战[M]. 北京:清华大学出版社,2017.

(收稿日期:2020-02-24)

作者简介:

汪淑贤(1987 -),女,硕士,讲师,主要研究方向:模式识别及图像处理。

李明(1997 -),男,本科,主要研究方向:通信工程。

王旬(1985 -),女,博士研究生在读,高级实验师,主要研究方向:无线通信。

(上接第 47 页)

[16] LIN T Y,GOYAL P,GIRSHICK R,et al. Focal loss for dense object detection[C]. Proceedings of the IEEE International Conference on Computer Vision,2017:2980-2988.

[17] GIRSHICK R. Fast R-CNN[C]. Proceedings of the IEEE International Conference on Computer Vision,2015:1440-1448.

(收稿日期:2020-01-16)

作者简介:

张庆伍(1994 -),男,硕士研究生,主要研究方向:计算机视觉、行人检测。

关胜晓(1964 -),通信作者,男,博士,副教授,主要研究方向:智能机器人、模式识别、嵌入式系统。E-mail:guanxiao@ustc.edu.cn。

版权声明