文章编号：2095-2163(2021)08-0107-06 中图分类号：TP391 文献标志码：A

# 基于改进的 Cascade- RCNN 网络的人员检测算法

吉鹏飞
(浙江理工大学 信息学院，杭州 310018)

**摘 要**：为解决人员密集情景下行人检测存在大量的目标误检、漏检的情况，本文提出了一种改进的基于 Cascade- RCNN 的目标检测网络，提高了人员检测的准确率。对目前检测效果较好的 Cascade- RCNN 做了一些改进：选用 ResNeXt101 代替 ResNet 作为骨干网络，以便提取更加充分的特征；为了获得更好的标记框，用 kmeans 聚类算法得到更符合目标形态的 anchor 长宽比例，通过 WBF 算法融合多个模型的结果得到更精确的边界框，同时引入多尺度训练以提高对小尺度目标的检测能力。实验结果表明，在 CrowdHuman 公开数据集上，用 ResNeXt101 提取特征其得分提高了 3.7%，用 kmeans 聚类算法生成 anchor 比例和 WBF 算法融合多预测框其准确率提升了 0.7%和 1.2%，最终整体性能较基础 Cascade- RCNN 提升近 6%。

**关键词**：行人检测；Cascade- RCNN；kmeans 聚类算法；WBF 算法；多尺度训练

## Person detection algorithm based on improved Cascade- RCNN network

JI Pengfei
(Information Institute Zhejiang Sci-Tech University, Hangzhou 310018, China)

**【Abstract】** In order to solve the problem of a large number of false and missed target detections in pedestrian detection in crowded scenarios, an improved target detection network based on Cascade-RCNN is proposed to improve the accuracy of human detection. Some improvements have been made to Cascade-RCNN, which has good detection results: ResNeXt101 is used instead of ResNet as the backbone network to extract more sufficient features; in order to obtain better labeled frames, the kmeans clustering algorithm is used to obtain anchor lengths that are more in line with the target shape Wide scale, the results of multiple models are merged through the WBF algorithm to obtain a more accurate bounding box, and multi-scale training is introduced to improve the detection ability of small-scale targets. Experimental results show that on the CrowdHuman public data set, using ResNeXt101 to extract features increases the score by 3.7%, using kmeans clustering algorithm to generate anchor ratios and WBF algorithm fusion multi-prediction frame, the accuracy rate increases by 0.7% and 1.2%, and finally the overall The performance is improved by nearly 6% compared to the basic Cascade-RCNN.

**【Key words】** pedestrian detection; Cascade-RCNN; kmeans clustering algorithm; WBF algorithm; multi-scale training

## 0 引 言

行人检测目的是通过计算机自动识别当前画面中的行人并将其标定出来，是身份判定、姿态分析、目标追踪等研究的子任务。行人检测在视频监控、车辆辅助驾驶、智能交通等领域应用广泛[1]。影响行人检测准确率的主要因素有人员密集、背景繁杂、遮挡严重和目标形变等。目前行人检测主要分为基于传统机器学习和深度学习算法两种。

传统机器学习算法通过提取行人特征，如颜色特征、纹理特征等，通过分类器在图片中检测出所有目标。Vida 和 Jones 等较早的提出了 VJ 检测器并用于行人检测任务[2]；Dalai 等提出了将 HOG 结合 SVM 用于行人检测[3]；Wu 等提出了一种基于人体部件的 Edgelet 特征，能够提高遮挡情景下的准确率[4]；Felzenszwalb 等用 DPM (Deformable Parts Model)算法检测行人，该算法能减弱人员形变带来的影响[5]；P. Dollar 等提出积分通道特征(Integral Channel Feature，ICF)应用于行人检测、速度和精度有了很大提升[6]；甘玲等采用聚合支持向量机(Ensemble SVM)分类器解决了正负样本数量相差过大的问题[7]。

深度学习算法主要是通过神经网络对大量的数据进行训练得到一个模型，图片输入后能直接找到所有目标。基于深度学习的行人检测算法可分为 3 类：

(1)基于深度置信网络(Deep Belief Networks，DBN)，通过训练神经元间的权重，让整个神经网络按照最大概率来生成训练数据[8]；

(2)基于卷积神经网络(Convolutional Neural

基金项目：国家自然科学基金(6207050141)；浙江省自然科学基金(LQ20F050010)。

作者简介：吉鹏飞(1996-)，男，硕士研究生，主要研究方向：计算机视觉与图像处理。

收稿日期：2021-05-23

Network, CNN)。双阶段 CNN 算法有 Faster-RCNN、MaskRCNN、CascadeRCNN 等,需要用算法先生成一定数量的候选框后才进行分类和回归,因此 two-stage 准确率更高。单阶段 CNN 算法如 SSD, YOLO 等则直接进行分类和回归;

(3)基于循环神经网络(Recurrent Neural Network,RNN),是一类在其演进方向进行递归且所有节点按链式连接的神经网络[9];LSTM(Long ShortTerm Memory networks)是最常见的循环神经网络。

虽然目前行人检测准确率已经较高,但仍然存在许多问题:

(1)预测目标框不佳,尽管能够检测出目标,但包围框有较大冗余或和实际目标框存在一定程度偏移;

(2)漏检,当目标在检测画面中较小或遮挡较多,难以查询出所有目标;

(3)误检,将一些类人目标错误的判别成行人。

## 1 本文算法

### 1.1 Cascade RCNN 网络

Faster-RCNN 是一种双阶段目标检测网络,其结构如图 1 所示。Faster-RCNN 与单阶段网络不同的是在通过由多个卷积层、激励层和池化层组成的骨干特征提取网络后,不是直接进行分类和回归,而是要经过一个 RPN 网络,图 1 中红线标注的部分即是 RPN 网络,其目的是为后续精确的分类和回归网络提供一定数量的候选框。ROI Pooling 的作用是将 RPN 生成的候选框转变成某一特定大小的框,为之后更细致的分类和回归任务提供方便。

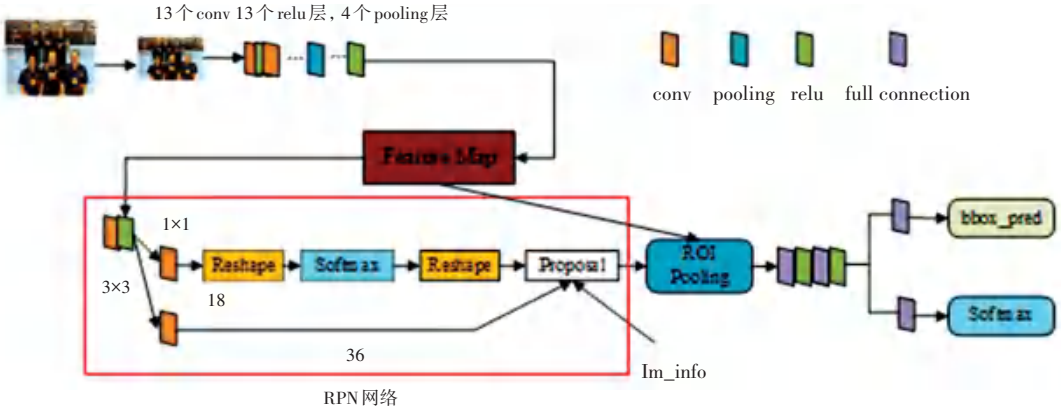


**图 1 Faster-RCNN 网络图**

**Fig. 1 Faster-RCNN network diagram**

Cascade RCNN 是由 Faster-RCNN 改进而来的,其结构如图 2 所示。与 Faster-RCNN 相比其最大的改进是级联多个不同 *IOU*(Intersection over Union,预测框和实际框的交并比)的分类回归网络。通常 *IOU* 值较低时,会学习到很多背景特征信息,降低模型预测的准确率;而 *IOU* 值较高时尽管能够减小匹配的错误率,但这样会造成有效样本占比太小,出现过拟合的问题。Cascade R-CNN 是一种 stage-by-stage 的结构,上一个检测网络输出是后一个检测模型的输入,并且每个阶段的 *IOU* 阈值依次增加。与 Faster-RCNN 一样,输入图片经过一个特征提取网络 CNN 后,通过 RPN 网络产生一定数量的 proposals,图 2 中 $B_0$ 就是 proposals,与 faster-Rcnn 一样要对这些 proposals 进行精细的分类和回归,C 和 B 分别代表分类和回归网络,H 代表网络头部。

Cascade RCNN 的损失函数(1)为:

$$L = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i A_i^* L_{reg}(t_i, t_i^*), \tag{1}$$

其中,$p_i$ 和 $p_i^*$ 分别表示预测类别的概率和正负样本标签;$t_i$ 和 $t_i^*$ 分别表示预测框和实际框的坐标;$\lambda$ 是回归损失占整个损失函数的比重。分类损失函数(2)为:

$$L_{cls} = y_i \log(h(x_i)) + (1 - y_i)\log(1 - h(x_i)), \tag{2}$$

其中:

$$h(x_i) = \frac{e^i}{\sum_j e^j}, \tag{3}$$

位置回归函数(4):

$$L_{reg} = \begin{cases} 0.5 * x^2, & |x| < 1, \\ |x| - 0.5, & other. \end{cases} \tag{4}$$

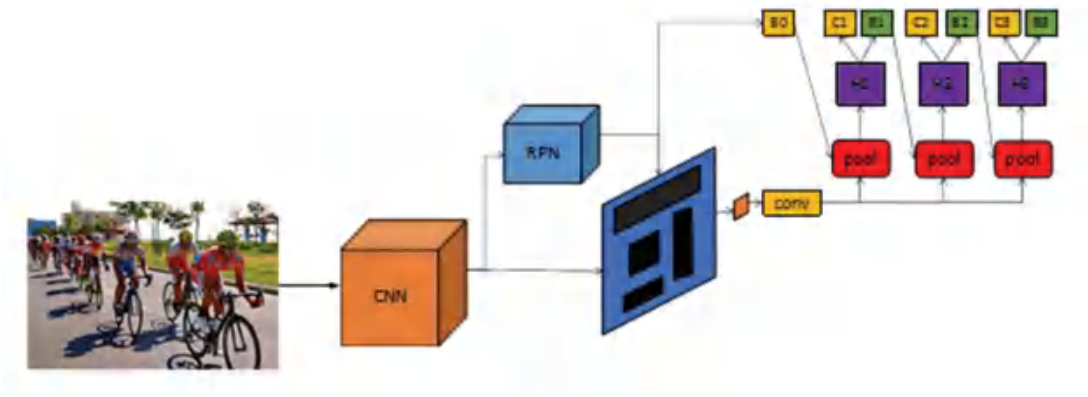


图 2　Cascade RCNN 网络结构

Fig. 2　Cascade RCNN network structure

## 1.2　ResNeXt 特征提取网络

通常为了提升检测精度,会选择更加复杂的骨干特征提取网络,但这样会引入过多的参数,增加计算量。ResNet 是一种常用的特征提取网络,其结构如图 3 左侧所示,学习的目标是目标值和输入的“差值”,这样能有效解决网络加深带来的梯度消失问题[10]。本文选用 ResNeXt101 代替 ResNet 作为最终模型的骨干特征提取网络,其结构如图 3(b)所示,ResNeXt 瓶颈结构是在 ResNet 基础上改进而来的。ResNext 用多路与 ResNet 类似的拓扑结构的 blocks 提取特征,最后融合多路特征,这样不仅可以减少计算量,而且模型拟合能力也得到了进一步提升。实践已经证明增加分组卷积的分支数比加深或加宽网络对准确率的提升更大。

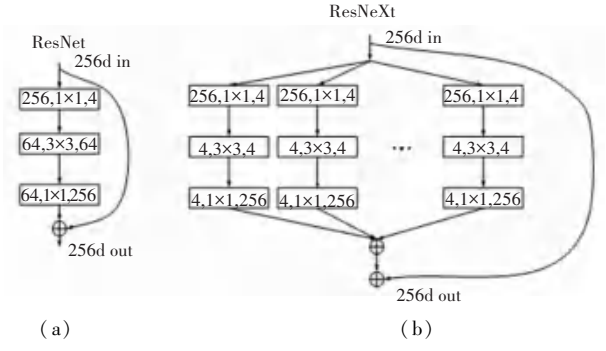


图 3　ResNet 和 ResNeXt 网络结构

Fig. 3　ResNet and ResNeXt network structure

## 1.3　生成更精确的预测框

与 faster-rcnn 一样,为了更好地适应图片中目标形状、大小的变化,Cascade RCNN 也引入了 anchor 机制,如图 4 所示。在特征图上每个位置生成多个不同比例、不同尺度的 anchor,每个 anchor 都对应着原图一定大小形状的区域。Cascade RCNN 有默认的 anchor 尺度、长宽比,但如果用默认参数可能难以生成与实际需要相匹配的目标框。为此本文采用 kmeans 聚类算法得到更适应实验数据集的 anchor 长宽比例,其详细步骤为:

(1)将训练集中 bounding box 对角坐标转变成高和宽的数据;

(2)在训练集中挑选 $k$ 个 bounding boxes 对 $k$ 个 anchor-box 初始化;

(3)算出所有的 bounding box 和每个 anchor-box 的 $IOU$,将所有的 bounding box 分类给与其误差 $d$ 最小的 anchor-box。用 $d(n,k)(d=1-IOU)$ 来刻画第 $n$ 个 bounding box 和第 $k$ 个 anchor-box 间的误差;

(4)根据分类的结果找出每个 anchor box 对应的 bounding box 的长宽的中值,并用其来更新 anchor box;

(5)重复上面的步骤,直到 bounding box 的分类已经不再更新。

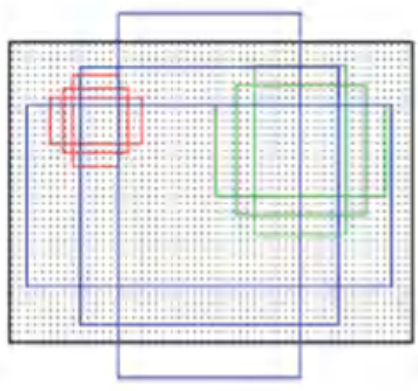

图 4　anchor 机制

Fig. 4　Anchor mechanism

本文采用了多尺度训练和测试的方法,能够学习不同尺度目标下的特征,提高了模型的适应能力。实验中将训练集和测试集同时放大一定比例(双线性插值)一定程度上可以提高小尺度目标的检测准

确率。

对于同一目标,根据模型可能会在其周围生成若干重叠率较高的预测框,如图 5 所示。通常本文用 NMS 只留下置信度最高的框,将其余的框排除。本文通过 WBF(Weight BoxFusion)方式融合多个模型的结果,提高了目标检测的准确率,其可以修正单个模型预测不精确的问题,消除冗余的边界框,最终预测框坐标是多个模型预测框的坐标加权和,权重为相应的边界框置信度[11]。

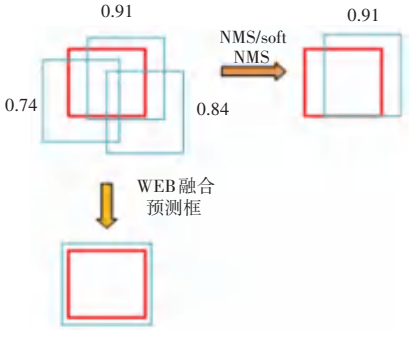


图 5 NMS 和 WBF 处理重叠框

**Fig. 5 NMS and WBF processing overlapping boxes**

WBF 的主要步骤:

(1)将 $N$ 个模型的所有预测框按得分降序输入到 List B 中;

(2)新建两个空的 List $L$ 和 List $F$。遍历 B,在 $F$ 中寻找和其匹配的框($IOU$ > 阈值),如果找到,就将其插入到 $L[i]$ 中,$i$ 表示该框在 $F$ 中的下标;如果未找到,直接将其插入到 $L$ 和 $F$ 的末尾;

(3)利用 $L[i]$ 处所有的 $m$ 个框($L$ 与 $F$ 可能是一对多的关系)加权计算 $F[i]$ 框的坐标($x_1$,$y_1$,$x_2$,$y_2$)和得分 $C$,式(5) ~ 式(7):

$$C = \frac{\sum_{i=1}^{m} C_i}{m}, \tag{5}$$

$$x_{1,2} = \frac{\sum_{i=1}^{m} C_i * x_{1i}, x_{2i}}{\sum_{i=1}^{m} C_i}, \tag{6}$$

$$y_{1,2} = \frac{\sum_{i=1}^{m} C_i * y_{1i}, y_{2i}}{\sum_{i=1}^{m} C_i}, \tag{7}$$

(4)如果 B 中每个框都处理完了,再次对 $F$ 得分更新,式(8):

$$C = C \times \frac{\min(m, N)}{N}. \tag{8}$$

### 1.4 SWA 融合多模型

为了提高训练出的模型的稳定性和泛化能力,本文采用了 SWA(Stochastic Weight Averaging)方法融合多个训练周期的模型,该方法可以在一定程度上提高目标检测的准确率,并且不会增加额外的计算量[12],融合第 $i$ 轮模型后的公式(9):

$$SWA\_model_i = \frac{1}{\frac{i}{2}+1} \times SWA\_model_{i-1} + \left(1 - \frac{1}{\frac{i}{2}+1}\right) \times model_i. \tag{9}$$

本文一共训练 20 轮得到 20 个模型。根据式(9),本文采用第 9~19 轮阶段参数模型加权融合为最终模型。

## 2 实验设计及结果分析

### 2.1 实验环境及数据集

实验基于 CentOS 操作系统,Python3.6,pytorch1.4,cuda10.1,GPU 型号为 tesla T4,显存为 15G,CPU 型号为 Intel(R) Xeon(R) Silver 4110,2.10GHz。实验用的人体检测数据集是 Crowd-Human,训练和测试的图片有 1.5w 和 5k 张,共 340k 个人体目标,并且有场景多样、尺度各异,部分目标还存在一定的遮挡。

为了增加数据集的多样性,提高模型的泛化能力,本文尝试了一些数据增强方法,如水平翻转、GridMask、高斯模糊、颜色抖动等。实验学习率初始值为 $12\times10^{-4}$,一共经历 20 轮训练。

### 2.2 实验设计及实验评价指标

CrowdHuman 数据集采用 JI(Jaccard Index)评测,被定义为式(10):

$$\mathrm{JI}(D,G) = \frac{|\ Match_{iou}(D,G)\ |}{|\ D\ | + |\ G\ | - |\ Match_{iou}(D,G)\ |}. \tag{10}$$

其中,$|\ D\ |$ 和 $|\ G\ |$ 分别表示预测框和标注框的数量,$|\ Match_{iou}(D,G)\ |$ 是两者匹配的数量(评判标准是 $IOU > Threshold$ ),本文 $IOU$ 阈值为 0.5。

为了证明提出的改进之处的实用性,本文主要设计了 3 个实验:ResNeXt101 代替 ResNet 实验、kmeans 聚类算法生成 anchor 长宽比实验和引入 WBF 算法实验。

### 2.3 kmeans 生成 anchor 比例前后对比实验

通过观察训练集,发现大多数对象的纵横比都大于 1.0,为了得到更好的纵横比的 anchor,本文采用 kmeans 聚类算法得到 anchor 的长宽比为[1.05,

1.84,3.21]，为了使用方便，本文将其取整[1.0,2.0,3.0]（Cascade- RCNN 默认的比例为[0.5,1,2]）。由于实验数据集中的图片尺度差异较大，其短边几百像素至几千像素不等。为此，本文通过多尺度训练的方式将训练集中的图片短边做一定程度的缩放或放大，长边按相同比例缩小或放大，结果见表 1。

**表 1　kmeans 得到 anchor 比例前后对比结果**

**Tab. 1　Comparison results before and after the anchor ratio obtained by kmeans**

| Backbone | Training scales | Test scales | Anchor Ratio | Score |
|---|---|---|---|---|
| Res50 | 704-1 024 | 1 024 | [0.5,1,2] | 0.777 2 |
| Res50 | 1 440-1 760 | 1 600 | [0.5,1,2] | 0.813 8 |
| Res50 | 1 440-1 760 | 1 600 | [1,2,3] | 0.820 2 |

从表 1 中对比发现 1 440~1 760 的训练尺度较较小的 704~1 024 尺度得分大幅提高，提高了 4.7%（数据集中有较多的小目标），用 kmeans 聚类算法能够获得更好的包围框，其准确率提升了 0.7%。

### 2.4　ResNeXt 作为骨干网络及引入 WBF 实验

为测试不同骨干特征提取网络的性能，本文选择了 Res50、Res101 和 ResNeXt101 作为特征提取网络。同时为了提高准确率，将 Res101 和 ResNeXt101 得到的模型通过 WBF 算法融合预测框，最后加入 kmeans 聚类算法，得到最终目标检测结果，见表 2。

**表 2　不同特征提取网络对比实验**

**Tab. 2　Comparison experiments of different feature extraction networks**

| Backbone | Score |
|---|---|
| Res50 | 0.804 2 |
| Res101 | 0.810 3 |
| ResNeXt101 | 0.840 8 |
| ResNeXt101+WBF | 0.851 1 |
| ResNeXt101+WBF+kmeans | 0.859 6 |

从表 2 可以看出，更深层次的 Res101 比浅层的 Res50 特征提取网络效果要好，融合多路特征的 ResNeXt101 网络比 Res101 得分提高 3.7%，使用 WBF 融合后得分提高约 1.2% 。实验结果显示 ResNeXt 作为骨干网络比 ResNet 性能有很大提升，用 ResNeXt101 和 Res101 做 WBF 融合后较未融合前准确率也有一些提升，最后加入 kmeans 聚类算法，较基础 Cascade RCNN 整体性能提升近 6%。提升前后的对比如图 6 所示，图 6(a)标注的 1 和 2 框均要比左侧的更好，同时发现当遮挡较严重时，本文的算法也很难检测出来。

(a)　　(b)

**图 6　算法提升前后对比图**

**Fig. 6　Comparison chart before and after algorithm upgrade**

## 3　结束语

在人员密集的场景下，难免会发生遮挡、较大程度的形变等问题，本文提出的基于改进的 Cascade-RCNN 的目标检测网络能够有效提高目标检测的准确率，减小漏检的概率。实验结果表明用 ResNeXt101 网络作为基础骨干网络，在不增加计算量的基础上融合多路特征，能够获得更细致的特征进而提高检测的准确率。用本文所述的 kmeans 聚类算法得到的 anchor 比例能得到更适合当前数据集的目标框，通过 WBF 算法融合多个模型的检测结果能够得到更加准确的预测框。整体而言，改进的 Cascade-RCNN 算法较基础的 Cascade- RCNN 算法的人员检测性能有很大提升，但对于一些重叠较大、尺度过小的目标，本文的方法依旧无能为力，这也是后续研究的难点。

## 参考文献


[1] 陈宁，李梦璐，袁皓，等. 遮挡情形下的行人检测方法综述[J]. 计算机工程与应用，2020，56(16)：13-20.

[2] VIOLA P，JONES M，SNOW D. Detecting Pedestrians Using Patterns of Motion and Appearance[J]. International Journal of Computer Vision.2005.63(2)：153-161.

[3] DALAL N，TRIGGS B. Histograms of Oriented Gradients for Human Detection[C]// IEEE Computer Society Conference on Computer Vision & Pattern Recognition. IEEE，2005：886-893.

[4] WU B，NEVATIA R. Detection of multiple，partially occluded humans in a single image by Bayesian combination of edgelet part detectors[C]//10th IEEE International Conference on Computer Vision，2005：740-757.

[5] FELZENSZWALB P F，MCALLESTER D A，RAMANAN D. A discriminatively trained，multiscale，deformable part model[C]// 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2008.

[6] DOLLAR 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.

[7] 甘玲，杨梦. 聚合支持向量机分类器的行人检测方法[J]. 计算机工程与应用，2019，55(7)：194-198.


（下转第 117 页）