ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 3,May 2019,pp. 530－537
DOI:10. 16337/j. 1004-9037. 2019. 03. 016

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于融合FPN和Faster R-CNN的行人检测算法

王 飞[1] 王 林[2] 张儒良[2] 赵 勇[3] 王全红[3]

(1.贵州民族大学人文科技学院,贵阳,550025; 2.贵州民族大学数据科学与信息工程学院,贵阳,550025; 3.北京大学深圳研究生院信息工程学院,深圳,518055)

摘 要:针对多尺度行人检测的问题,本文提出一种基于融合特征金字塔网络(Feature pyramid networks,FPN)和Faster R-CNN(Faster region convolutional neural network)的行人检测算法。首先,对FPN和区域建议网络(Region proposal networks,RPN)进行融合;然后,对FPN和Fast R-CNN进行融合;最后,在Caltech数据集、KITTI数据集和ETC数据集上分别对融合FPN和Faster R-CNN的行人检测算法进行训练和测试。该算法在Caltech数据集、KITTI数据集和ETC数据集上的mAP(mean Average Precision)分别达到69.72%,69.76%和89.74%。与Faster R-CNN相比,该算法不仅提高了行人检测精度,而且在多尺度行人检测的问题上也获得了较为满意的检测效果。

关键词:特征金字塔网络;区域建议网络;Faster R-CNN;多尺度行人检测

中图分类号:TP391.41 **文献标志码**:A

Pedestrian Detection Algorithm Based on Fusion FPN and Faster R-CNN

Wang Fei[1], Wang Lin[2], Zhang Ruliang[2], Zhao Yong[3], Wang Quanhong[3]

(1. College of Humanities & Sciences, Guizhou Minzu University, Guiyang, 550025, China; 2. College of Data Science and Information Engineering, Guizhou Minzu University, Guiyang, 550025, China; 3. School of Electronic and Computer Engineering, Shenzhen Graduate School Peking University, Shenzhen, 518055, China)

Abstract: Aiming at the problem of multi-scale pedestrian detection, a pedestrian detection algorithm based on fusion feature pyramid networks (FPN) and faster R-CNN (Faster region convolutional neural network) is proposed. Firstly, FPN and region proposal networks (RPN) are fused. Secondly, FPN and Fast R-CNN are fused. Finally, the pedestrian detection algorithm with fusion FPN and Faster R-CNN is trained and tested on Caltech dataset, KITTI dataset, and ETC dataset, respectively. The mAP (mean Average Precision) of this algorithm reaches 69.72%, 69.76% and 89.74% on Caltech dataset, KITTI dataset, and ETC dataset, respectively. Compared with Faster R-CNN, this algorithm not only improves the pedestrian detection accuracy, but also obtains satisfactory detection effect on the problem of multi-scale pedestrian detection.

Key words: feature pyramid networks; region proposal networks; Faster R-CNN (Faster region convolutional neural network); multi-scale pedestrian detection

---

基金项目:贵州省教育厅创新群体重大研究项目(黔教合KY字[2018]018)资助项目;深圳市科技计划(JCYJ20160506172651253)资助项目;贵州省研究生科研基金立项课题(黔教研合KYJJ字[2016]04)资助项目;贵州民族大学人文科技学院科研基金(18rwjs016)资助项目。

收稿日期:2018-02-23;**修订日期**:2019-04-19

引　言

近年来,随着深度学习(Deep learning,DL)这股浪潮的兴起,计算机视觉领域逐渐采用深度学习算法来研究目标检测。2013年,Sermanet等[1]提出一种基于卷积神经网络(Convolutional neural networks,CNN)的OverFeat算法,该算法主要采用滑窗来实现对目标的定位检测。2014年,Girshick等[2]提出R-CNN(Region convolutional neural network)算法。该算法首先采用选择性搜索(Selective search,SS)[3]从图像中提取2 000个可能包含目标的候选区域,即感兴趣区域(Region of interest,RoI);然后将这些RoI压缩到统一大小(227 × 227),并传递给CNN进行特征提取;最后把提取到的特征送入支持 向量机(Supported vector machine,SVM)分类器,以获得该RoI的种类。2015年,Girshick等[4]提出Fast R-CNN算法,该算法首先采用SS算法从原始图像中提取2 000个RoI,然后对整幅图像进行卷积计算,得到卷积特征图,最后使用感兴趣区域池化层从卷积特征图中提取每个候选框的特征向量。2015年,Ren等[5]提出Faster R-CNN算法,该算法采用区域建议网络(Region proposal network,RPN)替代SS算法进行候选框选择,使整个目标检测实现了端到端的计算。深度学习目标检测也可以采用回归的思想。最具有代表性的是YOLO[6](You only look once)和SSD[7](Single shot multibox detector)两种算法。

行人检测是通用目标检测的一种特例。针对行人检测自身的特性,研究人员提出一些基于深度学习的行人检测算法。2013 年,Ouyang等[8]提出Joint deep算法,该算法将行人检测的特征提取、变形处理、遮挡处理和分类组合成一个联合的深度学习框架。2016年,Zhang等[9]提出一种结合RPN和RF(Roosted forests)的行人检测算法,该算法有效地克服了R-CNN用于行人检测的两个限制:处理小实例的特征映射的分辨率不足和缺乏用于挖掘难例的引导策略。2017年,Mao等[10]提出一种HyperLearner网络架构,其联合学习行人检测以及给定的额外特征。通过多任务训练,其能够利用给定特征的信息来提高检测性能,而不需要额外的推理输入。

针对多尺度行人检测的问题,本文对特征金字塔网络[11](Feature pyramid networks,FPN)和Faster R-CNN进行融合,提出一种基于融合FPN和Faster R-CNN的行人检测算法。

1　基于Faster R-CNN的行人检测

基于Faster R-CNN的行人检测的整体网络结构图如图1所示,该网络结构的输入是一幅包含行人的图像,输出是行人的概率得分和边界框。RPN生成300个候选区域输入给行人检测网络Fast R-CNN,考虑到RPN和Fast R-CNN 网络的前部分都采用若干卷积层来计算特征图。因此,此网络结构

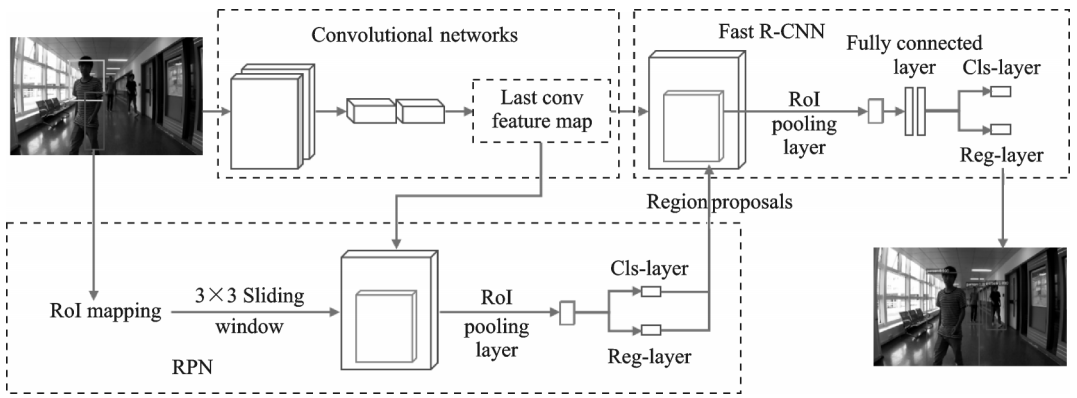

图1　基于Faster R-CNN的行人检测的整体网络结构图

Fig.1　Overall network structure diagram of pedestrian detection based on Faster R-CNN

把这两个网络统一成一个网络，使得RPN和Fast R-CNN卷积层参数共享，最终形成一个端到端的目标检测网络结构。

1.1 特征提取

当前，用于图像特征提取的主流网络有AlexNet，GoogLeNet，VGGNet，ResNet和DenseNet。综合考虑网络的复杂度以及分类精度，本文采用VGG16网络来提取图像特征，即采用在ImageNet数据集上预训练的权值作为初始值进行网络训练。

1.2 区域建议网络

Faster R-CNN使用CNN直接生成候选区域，该网络称之为RPN，如图2所示。在RPN中，最后一个卷积层有512个卷积核，因此，特征图有512个，特征维度为512维，每个特征图的大小约为40 × 60。采用3 × 3的滑窗来滑动特征图，当滑窗滑到每个位置时，预测输入图像3种尺度{128，256，512}和3种长宽比{1∶1，1∶2，2∶1}的候选区域，因此，每个滑动的位置有$k=9$个候选区域，一幅图像会生成约40 × 60 × 9个候选区域。在卷积层之后接有两个全连接层，一个为分类层(Cls-layer)，其输出$2k$个得分，用于判定候选区域是行人还是背景；另一个为边界回归层(Reg-layer)，其输出$4k$个坐标，用于对候选区域的边界进行微调。虽然由RPN选取的候选区域约有2 000个，但是，该网络结构依据候选区域的得分高低筛选出前300个输入到行人检测网络中。

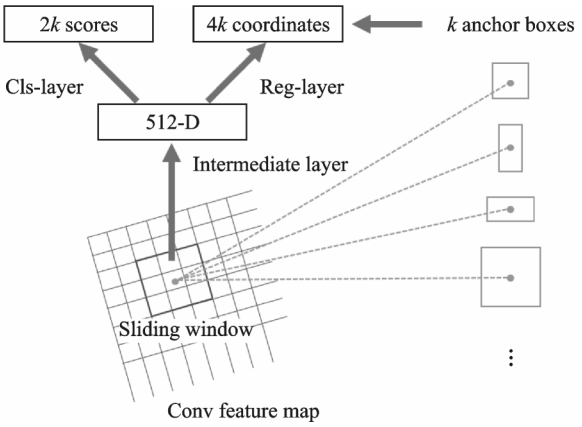

图2 区域建议网络

Fig.2 Region proposal network

1.3 感兴趣区域池化层

R-CNN是将提前生成好的每个RoI作为一幅图像输入到网络中进行后续操作。本文先对整幅图像提取一次卷积层特征，接着把RoI在原始图像的位置映射到卷积层的特征图上，以获得各个RoI的特征图。由RoI在原始图像的位置到特征图映射任务的层，称之为感兴趣区域池化层(RoI pooling layer)，如图3所示。在映射出每一个RoI的特征图后，需要把它们输入给全连接层，但是，全连接层要求大小一样的特征输入，而RoI的大小却是不相同的，为了通过该层映射输出大小一样的特征，本文网络结构对文献[12]中SPP-Layer进行了改进，采用单尺度输出7 × 7的特征图，若输入的候选区域为(r,c,h,w)，RoI pooling layer首先产生7 × 7个$r \times c \times (h/7) \times (w/7)$的块，然后利用Max pooling方式求出每一个块的最大值，这样输出的都是7 × 7的特征图。

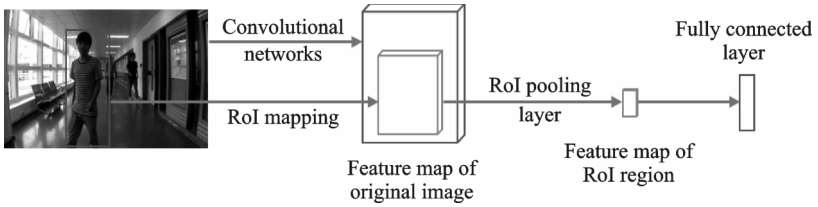

图3 感兴趣区域池化层

Fig.3 Region of interest pooling layer

2　特征金字塔网络

对于多尺度行人检测的问题,传统的方法是先构建一幅图像的金字塔,然后通过提取特征来形成特征金字塔,最后在每一层特征图上进行预测(如图4(a)),其缺点是计算开销比较大。神经网络的方法仅需要在最后一层特征图上进行预测(如图4(b)),也能够获得较好的效果,但是,对于小的行人,其表现还是不太好,因此,仍然需要考虑金字塔。SSD采用从多层的特征图上进行预测(如图4(c)),为了避免利用太低层的特征,SSD选择从CNN中的高层开始构建金字塔,同时还往后增加了几层,以分别抽取每一层特征进行综合利用。SSD未再使用高分辨率的低层特征,而这些层对于检测小的行人却是十分重要的。FPN使用自下而上路径(Bottom-up pathway)、自上而下路径(Top-down pathway)以及横向连接(Lateral connections)的方式(如图4(d))将低分辨率、高语义信息的高层特征与高分辨率、低语义信息的低层特征结合在一起,使每一个尺度下的特征均拥有十分丰富的语义信息。

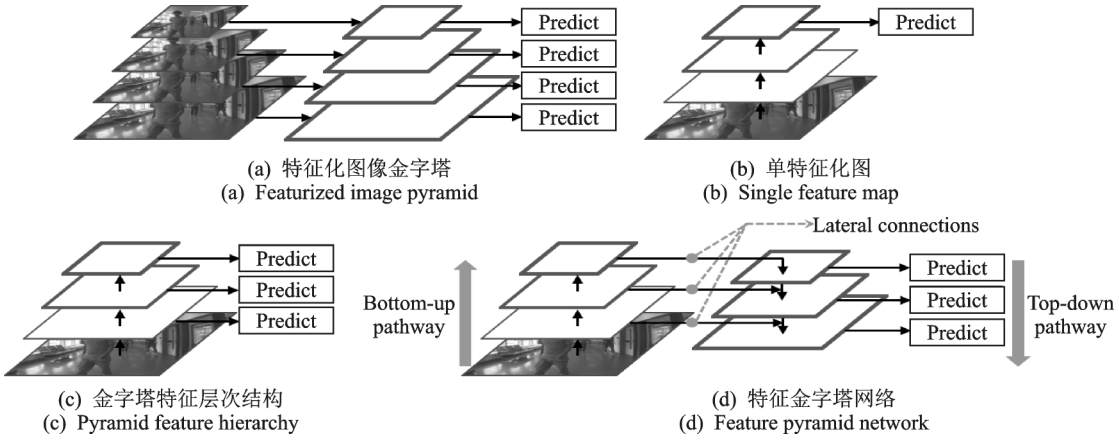

图4　4种特征金字塔

Fig.4　Four feature pyramids

2.1　自下而上路径

卷积神经网络的前馈计算便是自下而上路径,特征图经卷积核计算后,往往越变越小。不过,也存在一些特征层的输出与原来大小相同,称之为相同网络阶段。对于本文的特征金字塔,首先为每个阶段定义一个金字塔等级;然后把每个阶段的最后一层的输出当成特征图的参考集。具体而言,对于VGG16网络,首先使用每个阶段的最后一层的特征激活输出;然后把它们表示为$\{C_2,C_3,C_4,C_5\}$,其与Conv2,Conv3,Conv4以及Conv5的输出相对应,同时相对于输入图像具有$\{4,8,16,32\}$像素的步长。由于内存占用比较大,故没有将Conv1包含于金字塔。

2.2　自上而下路径和横向连接

自上而下路径,即先将更抽象、语义更强的高层特征图进行上采样,然后将此特征横向连接到前一层特征。因此,高层特征得以加强。横向连接的两层特征在空间尺寸上要相同,主要是为了利用低层的定位细节信息。

构建自上而下路径和横向连接的示意图如图5所示。首先,使用较粗糙的分辨率特征图将空间分辨率上采样为2倍;然后,按照元素相加的方式将上采样图和自下而上图合并。重复迭代此过程,直到产生最精细的特征图。为了使迭代开始,在C_5后附加一个1×1的卷积核层,用于产生最粗糙的分辨率图。最后,需要在每个合并的图上附加一个3×3的卷积,用于产生最终的特征图,以减少上采样的

混叠效应。最终的特征映射集为$\{P_2,P_3,P_4,P_5\}$,对应于具有相同空间大小的$\{C_2,C_3,C_4,C_5\}$。

与传统的图像特征金字塔一样,本文金字塔的每一层均采用共享的分类器或回归器,需要在每一个特征图中固定特征维度(即通道数,记为d)。本文中,设定$d=256$,因而所有额外的卷积层具有256个通道的输出。

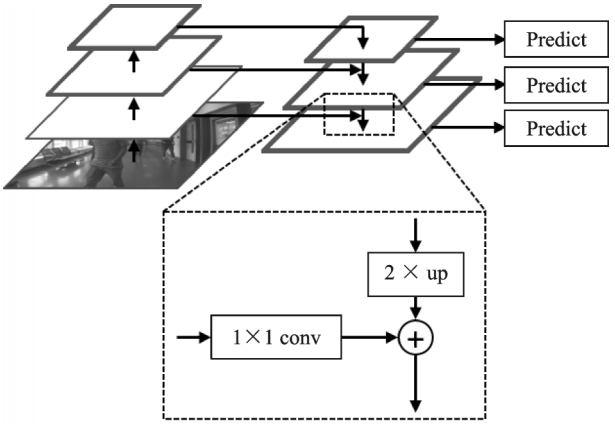

图5 构建自上而下路径和横向连接的示意图

Fig. 5 Diagram of building top-down pathway and lateral connections

3 融合FPN和Faster R-CNN的行人检测

Faster R-CNN主要由RPN和Fast R-CNN构成,对FPN和Faster R-CNN进行融合时,可以将FPN分别融合到RPN和Fast R-CNN中。

3.1 融合FPN和RPN

RPN是在密集的3 × 3的滑窗上评估一个小的子网络,同时在单尺度的卷积特征图的顶部执行目标或非目标二进制分类以及边界框回归。它是使用一个3 × 3卷积层后跟两个用于分类、回归的1 × 1卷积实现的,称为网络头部。目标或非目标标准以及边界框回归目标是相对于一组称为锚的参考框定义的。锚具有多个预定义的比例尺和纵横比,以覆盖不同形状的目标。

本文使用FPN替代单尺度特征图以适应RPN。在特征金字塔的每个等级上附加一个一样设计的头。由于头部在每一个金字塔等级的每一个位置密集地滑动,故不需要在特定水平上有多尺度的锚点。相反,本文为每一个等级分配一个单尺度的锚点。在形式上,$\{P_2,P_3,P_4,P_5\}$所对应的锚点的尺度为$\{64^2,128^2,256^2,512^2\}$。本文仍然使用这3种长宽比$\{1:1,1:2,2:1\}$,因此在金字塔中共有12个锚点。

训练过程中,本文将重叠率(Intersection over union,IoU)大于0.7的当作正样本,小于0.3的当作负样本。由于FPN之间有参数共享,故使所有层级具有相似的语义信息。

3.2 融合FPN和Fast R-CNN

Fast R-CNN采用RoI pooling layer来提取特征,其通常在单尺度特征图上执行。如果要将其与FPN一起使用,那么需要把不同尺度的RoI分配给金字塔等级。在形式上,可以通过式(1)将宽度w和高度h(在输入图像上的网络)的RoI分配给特征金字塔的等级P_k,即有

$$k=\left\lfloor k_0+\log_2(\sqrt{wh}/224)\right\rfloor \tag{1}$$

式中:224表示ImageNet数据集预训练图像大小,k_0表示映射到$w\times h=224^2$的RoI的目标水平,本文中将k_0设置为4。式(1)意味着若RoI的尺度变小(如:224的1/2),则其应该被映射到更精细的级别(如:$k=3$)。

4 实验结果与分析

4.1 Caltech数据集的实验结果与分析

Caltech数据集是一个使用车载摄像机所拍摄的行人视频数据集,视频时长约为10 h,分辨率是

640 × 480,帧率是30帧/s。该数据集标注了约250 000帧,350 000个矩形框,2 300个行人,其包含11个子集(Set00～Set10),6个子集(Set00～Set05)用作训练集,剩余的子集用作测试集,训练集、测试集都给定了标注信息。考虑到该行人数据集是视频形式,连续两幅图像之间的相似度比较大,故本文分别在训练集、测试集上间隔3帧选取1帧当作训练图像、测试图像。本文的训练图像为42 782幅,测试图像为4 024幅。

在Caltech数据集的训练集上进行训练时,将初始学习率设定为0.000 2,当迭代次数达到50 000次时,学习率变为0.000 02,继续迭代20 000次;动量为0.9,权重衰减为0.000 1。

本文算法在Caltech数据集的测试集上的mAP如表1所示。在相同训练集、测试集的情况下,本文算法与Faster R-CNN,SSD和PVANET[13]等算法进行了比较。由表1可知,本文算法在Caltech数据集的测试集上的mAP比PVANET算法高0.62%,比SSD算法高10.13%,比Faster R-CNN算法高1.15%。

表1　Caltech数据集测试对比实验结果

Tab. 1　Caltech dataset test comparison experiment results

Algorithm	Number of test sets	mAP/%
Faster R-CNN	4 024	68.57
SSD	4 024	59.59
PVANET	4 024	69.10
本文算法	4 024	69.72

4.2　KITTI数据集的实验结果与分析

KITTI数据集由7 481幅训练图像和7 518幅测试图像构成,其主要包含汽车(Car)、行人(Pedestrian)和自行车(Cyclist)等目标类别。由于测试图像没有给定标注信息,故在进行训练、测试时,需要将7 481幅训练图像分为训练和测试两部分。本文的训练图像为3 740幅,测试图像为3 741幅。

在KITTI数据集的训练集上进行训练时,将初始学习率设定为0.000 25,当迭代次数达到50 000次时,学习率变为0.000 025,继续迭代40 000次;动量为0.9,权重衰减为0.000 1。

本文算法在KITTI数据集的测试集上的mAP如表2所示。在相同训练集、测试集的情况下,本文算法与Faster R-CNN,SSD和PVANET等算法进行了比较。由表2可知,本文算法在KITTI数据集的测试集上的mAP比PVANET算法高1.92%,比SSD算法高9.53%,比Faster R-CNN算法高2.49%。

表2　KITTI数据集测试对比实验结果

Tab. 2　KITTI dataset test comparison experiment results

Algorithm	Number of test sets	mAP/%
Faster R-CNN	3 741	67.27
SSD	3 741	60.23
PVANET	3 741	67.84
本文算法	3 741	69.76

4.3　ETC数据集的实验结果与分析

ETC数据集是一个由多个数据集组合而成的数据集,其包含ETH数据集、TudBrussels数据集和部分Caltech数据集。本文对多个数据集进行组合的原因:(1)ETH数据集和TudBrussels数据集给定标注信息的图像数量比较少,ETH数据集给定标注信息的图像有1 804幅,而TudBrussels数据集给定标注信息的图像有508幅,若单独进行训练,则可能会产生过拟合现象。(2)对多个数据集进行组合,可以增加数据集的多样性,这样有利于CNN进行特征学习,但是,由于Caltech数据集的图像分辨率比较低,检测复杂度比较高,因此,本文只选择Caltech数据集的1个子集(Set00)进行组合,并且间隔30帧选取1帧。本文的训练图像为2 646幅,测试图像为515幅。

在ETC数据集的训练集上进行训练时,将初始学习率设定为0.000 25,当迭代次数达到50 000次

时,学习率变为0.000 025,继续迭代30 000次;动量为0.9,权重衰减为0.000 1。

本文算法在ETC数据集的测试集上的mAP如表3所示。在相同训练集、测试集的情况下,本文算法与Faster R-CNN,SSD和PVANET等算法进行了比较。由表3可知,本文算法在ETC数据集的测试集上的mAP比PVANET算法高0.42%,比SSD算法高0.72%,比Faster R-CNN算法高8.44%。

表3 ETC数据集测试对比实验结果

Tab. 3 ETC dataset test comparison experiment results

Algorithm	Number of test sets	mAP/%
Faster R-CNN	515	81.30
SSD	515	89.02
PVANET	515	89.32
本文算法	515	89.74

本文给出了3个数据集上的部分行人检测效果图,如图6所示。由图6可知,本文算法在多尺度行人检测的问题上获得了较为满意的检测效果,并且将目标判定为行人的概率得分也比较高。

(a) 在Caltech数据集上的行人检测效果图 (a) Pedestrian detection effect diagram on the Caltech dataset

(b) 在KITTI数据集上的行人检测效果图 (b) Pedestrian detection effect diagram on the KITTI dataset

(c) 在ETC数据集上的检测效果图 (c) Pedestrian detection effect diagram on the ETC dataset

图6 本文算法的行人检测效果图

Fig.6 Pedestrian detection effect diagram of this algorithm

5 结束语

本文针对多尺度行人检测的问题,将FPN和Faster R-CNN进行融合,提出一种基于融合FPN和Faster R-CNN的行人检测算法。该算法在Caltech数据集、KITTI数据集和ETC数据集上的mAP分别达到69.72%,69.76%和89.74%。在多尺度行人检测的问题上,该算法获得了较为满意的检测效果。但是,本文算法对于遮挡的行人还存在漏检的问题。因此,下一步准备将高质量、高精度的双目视差图与RGB图像进行融合,实现基于深度学习的3D行人检测以解决该问题。

参考文献:

[1] Sermanet P, Eigen D, Zhang X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks [EB/OL]. https://arxiv.org/abs/1312.6229, 2013-1-16.

[2] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2014: 580-587.

[3] Uijlings J R R, Sande K E A V D, Gevers T, et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154-171.

[4] Girshick R. Fast R-CNN[C]// IEEE International Conference on Computer Vision. New York: IEEE, 2015: 1440-1448.

[5] Ren Shaoqing, He Kaiming, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2015: 91-99.

[6] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]// IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 779-788.

[7] Liu Wei, Anguelov D, Erhan D, et al. SSD: Single shot multiBox detector[C]// European Conference on Computer Vision. Cham: Springer Press, 2016: 21-37.

[8] Ouyang Wanli, Wang Xiaogang. Joint deep learning for pedestrian detection[C]// IEEE International Conference on Computer Vision. New York: IEEE, 2013: 2056-2063.

[9] Zhang Liliang, Lin Liang, Liang Xiaodan, et al. Is faster R-CNN doing well for pedestrian detection [C]// European Conference on Computer Vision. Cham: Springer Press, 2016: 443-457.

[10] Mao Jiayuan, Xiao Tete, Jiang Yuning, et al. What can help pedestrian detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 6034-6043.

[11] Lin Tsung-Yi, Dollar P, Girshick R, et al. Feature pyramid networks for object detection[C]// IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 2117-2125.

[12] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[13] Kim K H, Hong S, Roh B, et al. Pvanet: Deep but lightweight neural networks for real-time object detection[EB/OL]. https://arxiv.org/abs/1608.08021, 2016-1-7.

作者简介:

王飞(1989-),男,硕士研究生,研究方向:图像处理、模式识别,E-mail: wangfei10248@163.com。

王林(1965-),男,教授,研究方向:图像处理、模式识别。

张儒良(1963-),男,教授,研究方向:图像处理、模式识别。

赵勇(1963-),男,博士,副教授,研究方向:机器学习、计算机视觉和视频分析。

王全红(1991-),男,硕士研究生,研究方向:机器学习、计算机视觉和视频分析。

(编辑:刘彦东)