复杂环境下多尺度行人实时检测方法

周薇娜* 孙丽华 徐志京

(上海海事大学 上海 201306)

摘 要：作为计算机视觉和图像处理研究领域中的经典课题，行人检测技术在智能驾驶、视频监控等领域中具有广泛的应用空间。然而，面对一些复杂的环境和情况，如阴雨、雾霾、被遮挡、照明度变化、目标尺度差异大等，常见的基于可见光或红外图像的行人检测方法的效果尚不尽如人意，无论是在检测准确率还是检测速度上。该文分析并抓住可见光和红外检测系统中行人特征差异较大，但在不同环境中又各有优势的特点，并结合多尺度特征提取方法，提出一种适用于多样复杂环境下多尺度行人实时检测的方法——融合行人检测网络(FPDNet)。该网络主要由特征提取骨干网络、多尺度检测和信息决策融合3个部分构成，可自适应提取可见光或红外背景下的多尺度行人。实验结果证明，该检测网络在多种复杂视觉环境下都具有较好的适应能力，在检测准确性和检测速度上均能满足实际应用的需求。

关键词：行人检测；复杂环境；自适应提取；多尺度；决策融合

中图分类号：TN911.73 **文献标识码**：A **文章编号**：1009-5896(2021)07-2063-08

DOI: 10.11999/JEIT200436

A Real-time Detection Method for Multi-scale Pedestrians in Complex Environment

ZHOU Weina SUN Lihua XU Zhijing

(*Shanghai Maritime University*, *Shanghai* 201306, *China*)

Abstract: As a classic subject in computer vision and image processing, pedestrian detection has a wide range of applications to intelligence driving and video monitoring fields. However, most of pedestrian detection methods based on visible or infrared images have no satisfying result in some complex environments or situations, such as rain, smog, occlusion, variation of illuminance and target scales, no matter in terms of detection accuracy or speed. This paper analyzes and finds out that, pedestrians usually have quite different characteristics in visible and infrared image, and which have their own advantages in different environments. Therefore, combining fusion and multi-scale technology, a real-time multi-scale pedestrian detection algorithm suitable for complex environment named FPDNet (Fusion Pedestrian Detection Network) is proposed. The detection framework is consisted by three main modules: feature extraction backbone network, multi-scale detection network and decision-level fusion network. The proposed method is able to extract multi-scale pedestrian characteristics under visible or infrared background adaptively. Experimental results prove that the detection network has good adaptability in complex visual environments, and can meet the demands of practical applications to detection accuracy and speed.

Key words: Pedestrian detection; Complex environment; Adaptive extracting; Multi-scale; Decision-level fusion

1 引言

行人检测旨在从视频和图像中定位行人目标，是计算机视觉领域中一项传统但又极其重要的任务。随着时代的发展，行人检测技术非但没有被淘汰，反而越发在实际生活中发挥出了重要作用，获得了越来越多的关注[1–3]。现如今在大多数人工智能应用中，例如自动驾驶、人流统计、行人身份识别、行人跟踪、行人行为动作分析等方面，行人检测技术都独立或作为一个重要环节起到了独一无二，无可替代的作用。

行人检测技术一般主要由特征提取和目标分类决策两部分组成。传统行人检测算法将这两部分分开执行。首先通过手工提取行人特征(包括颜色、

收稿日期：2020-06-01；改回日期：2020-12-01；网络出版：2021-03-31
*通信作者： 周薇娜 wnzhou@shmtu.edu.cn
基金项目：国家自然科学基金(61404083, 52071200)，中国博士后科学基金(2015M581527)，专用集成电路与系统国家重点实验室开放研究课题(2021KF010)

Foundation Items: The National Natural Science Foundation of China (61404083, 52071200), China Postdoctoral Science Foundation (2015M581527), The State Key Laboratory of ASIC & System (2021KF010)

边缘和纹理等常用特征)，然后用这些特征来训练好分类器后进行检测。手工提取特征有诸如SIFT [4]或梯度方向直方图[5](Histogram of Oriented Gradients, HOG)等，而特征分类器，例如支持向量机[6](Support Vector Machines, SVM)、适应提升方法[7](Adaptive Boosting, AdaBoost)和随机森林[8](Random forest，RF)等。这种检测方法基于滑动窗口区域，没有针对性，计算量大，时间复杂度高，有窗口冗余，且手工特征对于背景复杂的情况不具有很好的鲁棒性。基于以上分析，传统方法无论在速度还是准确性方面都无法在实际应用中获得令人满意的结果。

基于深度学习的目标检测算法近年来在计算机视觉任务方面取得了重大突破[9–11]，大幅度地提高了检测算法的准确率。基于深度学习的行人检测算法主要集中为两类：

(1)两步检测算法：将检测问题分为两个阶段，第1阶段生成候选区域，第2阶段对候选区域做回归和分类。这类算法的典型代表是R-CNN[12] 系列算法，如R-CNN, Fast R-CNN[13]和Faster R-CNN[14]等。此种基于候选区域的算法虽然精度高，但存在计算量大、运算速度慢，不能满足实时应用的缺点。

(2)单步检测算法：一种基于回归的端到端的目标检测方法，具体包括特征提取网络、边界框回归和分类网络两个部分。该类算法不生成候选区域，直接将数据输入卷积神经网络。之后预测物体的边界框信息，并对这些候选边界框进行判别，其中包括：对边界框内包含的物体的类别标签进行判别、对边界框的位置进行回归。典型的该类算法有YOLO算法[15–17]和SSD[18]算法。单步检测算法打破了候选区域生成这一常规前期处理操作，从而提高了基于深度学习的目标检测类算法的速度，也推动了行人检测技术的发展。

然而，行人检测尚有两大难点是目前已有的传统和深度学习还未完全解决的。一是面对多变的复杂环境，不管是依靠可见光还是红外传感器信息都难以获得令人满意的效果。如日常生活中常会遇见的烟雾、雨水、灰尘、光照度低等情况都会造成可见光图像中行人目标不可见或模糊不清的情况，而红外图像虽可以提高此种情况下图像的质量，但其无法描述行人特征丰富的轮廓和颜色信息，也会导致一些漏检和误检情况。二是行人目标的多尺度变化和小尺寸行人目标的检测。当场景中同时存在不同尺寸的行人，或者行人目标中存在尺寸较小、分辨率较低的个体，检测器提取到的行人特征会更加容易受到环境背景中的噪声干扰，将会导致漏检、误检，这对检测结果的准确性带来了不小的挑战。因此，本文以复杂环境下的多尺度行人目标检测的难点问题为主要研究内容，融合可见光和红外图像信息，结合深度学习和多尺度信息提取技术提出了一种基于卷积神经网络的融合行人检测网络(FPDNet)。

本文所提出的方法具有以下几个方面的特色和创新：(1) 设计并搭建了一个精简的特征提取骨干网络，包括基于残差结构的骨干基础网络和金字塔池化层，提高了网络对目标的检测效果；(2) 在网络检测结构中采用多尺度检测技术，来应对行人目标的尺寸变化，尤其改善了小尺寸行人目标的检测效果；(3) 决策融合可见光和红外图像信息，为网络提供更丰富可靠的行人目标特征，使得整个网络在多变复杂环境中更具鲁棒性。

2 本文提出的算法

本文提出的融合行人检测网络FPDNet顶层框图如图1所示。可见光和红外图像由多尺度检测网络得到特征分类结果后，经过决策融合模块得到最终的检测结果。而多尺度检测网络，又主要由特征提取骨干网络和多尺度检测模块构成(详见图2)。因此整个算法网络的主要模块包括特征提取骨干网络、多尺度检测和信息决策融合3部分。

2.1 特征提取骨干网络

如上一小节所述，单步检测深度学习网络具有较高的检测率和速度。单步检测法中又以YOLO算法执行最为高效，其YOLOv3版本可以通过强大的GPU实现实时目标检测，在行人检测上占有明显的优势。但YOLOv3使用的Darknet-53网络并不适合单一类别物体的检测，过深和冗余的网络结构，造成网络收敛速度较慢，增大了训练过程的时间开销，增加了目标细节特征信息丢失的可能性。对于骨干网络，参数过多会导致检测模型训练复杂化，增加对输入数据量的需求，检测速度也会降低。YOLOv3-tiny[17] 为YOLOv3的精简版本，由于使用了更少的卷积层，检测速率得到了提高。然而，YOLOv3-tiny检测精度却下降得很厉害。针对这些不足，本文建立了一个改进的卷积神经网络框架作为行人检测模型的特征提取网络，解决了检测精度和参数冗余互为矛盾的问题。该特征提取骨干网络分为骨干基础网络和金字塔池化层[19](Spatial Pyramid Pooling, SPP)网络两个部分。

(1)骨干基础网络：骨干基础网络以ResNet [20]为基础，基础网络基本单元如图3所示。对于传统

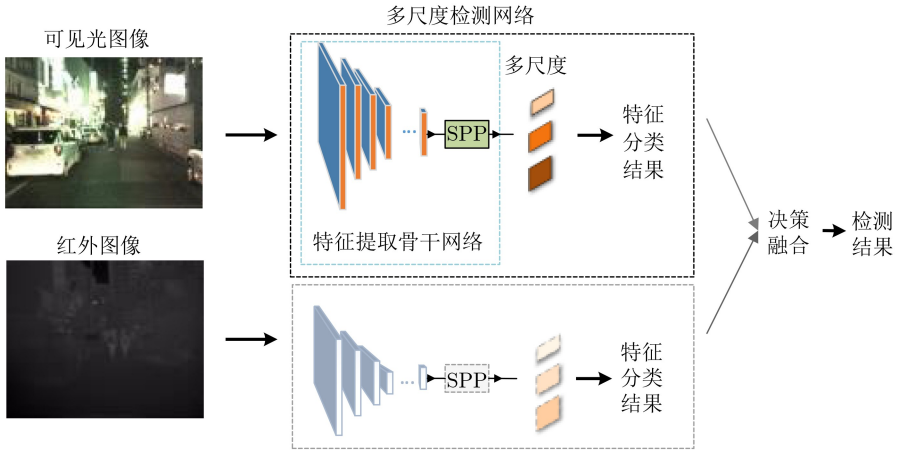


图 1 FPDNet顶层框图

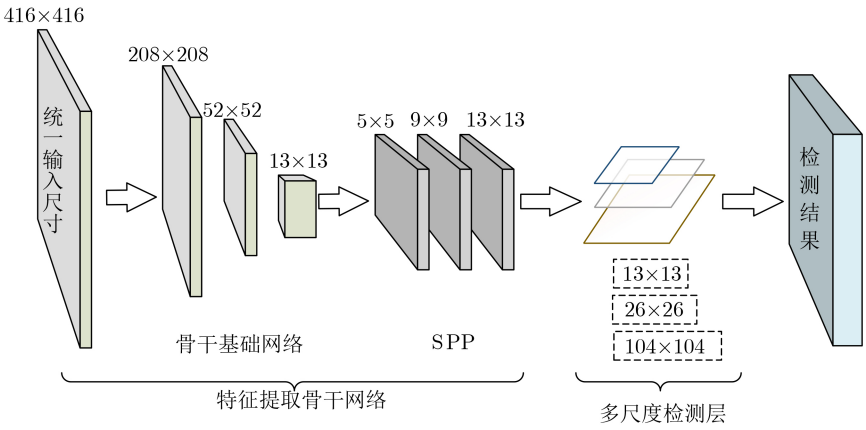


图 2 多尺度检测网络内部结构图

卷积神经网络，若输入为x，学习到的特征输出$H(x)=x$，为恒等映射。本文中骨干基础网络的核心思想是将网络的输出分为恒等映射(Identity Mapping)和残差映射(Residual Mapping)两部分，即输出 $H(x)=x+F(x)$。网络将学习目标改变，学习的将是期望输出和输入的残差$F(x)$，因此，接下来的训练方向就是要随着网络加深，准确率不下降的情况下，将残差$F(x)$逼近于零。

骨干基础网络结构如表1所示，网络输入维度为416×416×3，包括Conv(卷积层)，Max(最大池化层)和Res(残差层)。左边第1列的3, 4, 2等数字表示对应模块的重复次数，Res层执行快捷连接操作。ResNet结构可以很好地控制梯度的传播，避免训练中发生梯度消失或爆炸。其将卷积神经网络的逐层训练改为了逐阶段训练，网络被分为若干个子段，其中包含比较浅的网络，然后用快捷连接的方式对每个小段训练。这种训练方式解决了特征信息在网络中传递时信息丢失损耗的问题，一定程度上保护了信息的完整性，降低了目标特征学习的难度。并且从表1中可以看出，该骨干基础网络参数相对较少，因此能提高收敛速度，降低整体网络训练的难度。

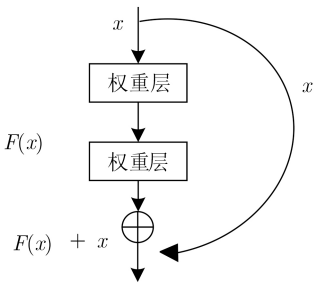


图 3 骨干基础网络基本单元

(2)金字塔池化层：区别于YOLOv3结构在特征提取网络之后直接连接预测网络，而忽略了同一层卷积上对每一张图像多方面的特征提取，本文在骨干基础网络之后，引入一个金字塔池化层(SPP)，以提高其检测精度。它可以将图像从较精细的层次划分为较粗的层次，并在所有层次上聚合局部特征。

如图4所示，SPP层由3个最大池化层组成，本文将池化层网络中使用的卷积核大小设计为5×5, 9×9和13×13。在本文的结构中，所有这些不同尺

表 1 骨干基础网络结构表

重复次数	类别	卷积核	卷积核尺寸	输出特征图大小
	Conv	64	7×7/2	208×208
	Max		2×2/2	104×104
	Conv	64	3×3/1	
3	Conv	64	3×3/1	
	Res			104×104
	Conv	128	3×3/2	
	Conv	128	3×3/1	
	Res			52×52
	Conv	128	3×3/1	
3	Conv	128	3×3/1	
	Res			52×52
	Conv	256	3×3/2	
	Conv	256	3×3/1	
	Res			26×26
	Conv	256	3×3/1	
4	Conv	256	3×3/1	
	Res			26×26
	Conv	512	3×3/2	
	Conv	512	3×3/1	
	Res			13×13
	Conv	512	3×3/1	
2	Conv	512	3×3/1	
	Res			13×13

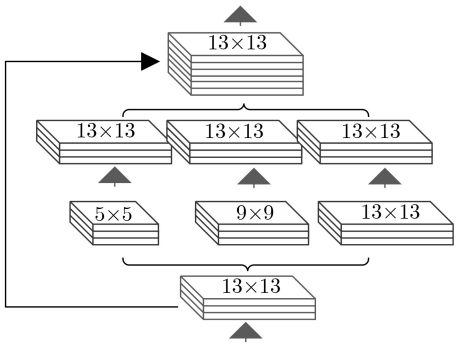


图 4 SPP层结构

寸的最大池化层会通过填充操作进行结果对齐，使具有相同大小的输出。然后将所有合并后的输出特征图连接起来作为下一个卷积层的输入。网络从不同的空间尺度对特征图进行特征提取，再组合，更多空间信息通过卷积运算被充分利用，能进一步提高行人目标的检测精度，增强了算法的鲁棒性。

2.2 多尺度行人检测

行人检测的一个难点在于不同尺寸，尤其是小尺寸行人目标的检测。由于物体细节特征的丰富性和具体程度与其所占像素数量基本成正比，小目标特征不易提取，从而会导致整体检测率的下降。在目标检测中，语义特征是指区分背景和目标对象的信息。对于卷积神经网络来说，不同的网络深度对应不同层次的特征，细粒度特征更容易在分辨率高的浅层网络学到，分辨率较低的深层网络则学到的更多的是语义特征。针对弱小行人目标，由于小尺度的特征图中无法提供所需的分辨率信息，还需联合大尺度的特征图进行判断，增加特征层之间的互补性。因此本文采用了3个不同尺度(13×13, 26×26和104×104)的特征图来进行行人检测。如图5所示，本文从网络中前两个图层中得到特征图，并对它进行两次上采样操作。再根据网络更早的图层中获得的特征图，把高低两种分辨率的特征图连接到一起，增强底层特征图对行人目标特征的描述能力。通过这种方式，上采样的特征中更多的语义信息将被利用，也能从网络中较低层的特征映射图中获得检测目标更细粒度的信息。每个检测层分配3个锚箱(anchor)来预测目标候选框信息(包括中心坐标位置，框的高度和宽度)、1个目标置信度值和它的类别概率，最后比较9种尺寸的检测结果，通过非极大值抑制方法确定最终检测结果。

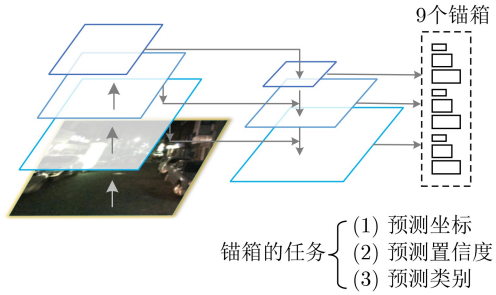


图 5 多尺度检测模块

2.3 基于可见光和红外图像的决策融合检测

可见光图像在光线良好、目标不被遮挡的情况下，能较为准确地反映目标的特征信息；而红外图像不易受光照度和烟雨、雾霾等复杂环境的影响，可以适应日夜连续监测。本文充分利用可见光和红外图像各自的优势特点，采用图像决策融合的方法来实现两种图像中行人特征信息的互补，充分利用多个信息源的细节特征，从更周全、准确和可靠的角度来描述检测目标，以此来提高行人目标在复杂环境变化中的识别准确度。决策融合检测方法对前期数据要求低，并且具有良好的实时性和容错性，可以更好地克服单一传感器造成的误差。本文基于决策融合的行人目标检测方法的具体流程如图6所示。

从图6中可以看到，当前期特征提取网络训练

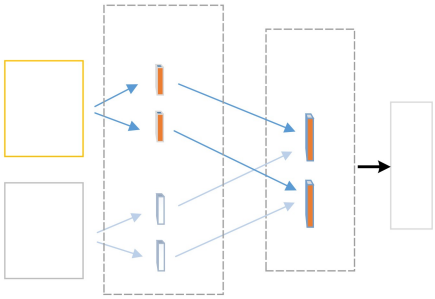


图 6 基于决策融合的目标检测流程

得到两个独立的区域特征分类模块后，在决策融合层对两个独立的分类模块预测的置信度采用Softmax[21]分类，如式(1)所示。将具有最高预测概率的类别作为包含在预测边界框中物体的类别，最后输出决策融合的检测类别结果。式中p为对应的预测概率，x为分类模块的预测值，i表示序号，j表示分类数目。基于决策级融合的检测算法效率较高，可以实现多波段图像的信息互补，比基于单一类别的可见光或红外图像的行人检测器更具稳健性，而且方便实现，运算速度较快，能满足实时检测的需求。这些都为行人检测在多变复杂环境中的实际应用奠定了基础。

$$p=\frac{\mathrm{e}^{x_i}}{\sum_j \mathrm{e}^{x_j}} \tag{1}$$

3　实验

3.1　实验环境

本文所做实验的硬件配置为Intel i7 8700k处理器、NVIDIA TITAN XP 显卡、64 GB RAM的服务器，软件环境为Ubuntu16.04系统、Darknet框架。

3.2　数据集

实验所采用的训练和测试数据主要来源于KAIST[22]行人数据库，该数据库包含了校园、街道等各种常见生活和交通场景，并且包含了丰富的白天和夜晚各种天气下的环境，能代表绝大多数实际生活中会碰到的复杂情况。行人与镜头的距离也各有不同，行人目标被分为3种尺度，即大于80像素的大尺度行人目标，30～80像素的中等尺寸行人目标和不在少数的小于30像素的小尺寸行人目标。该数据库分为set01～set09 共9个子集针对复杂环境下多尺度行人检测的研究需求，在保留训练样本多样性的前提下，为缩短训练时间，精简训练样本，本文选取了set05和set09文件夹中的可见光和红外图像作为数据样本，并剔除了其中空对象。最终本文的数据集总共包括2783对图像，训练集和测试集样本的分配比例为7:3。

3.3　训练参数的选择

本文在FPDNet检测模型中使用了3个尺度的目标检测层，尺寸分别为13像素×13像素、26像素×26像素和104像素×104像素。对于每个检测层，分别使用3种尺寸锚箱，即共9种尺寸的锚箱，通过聚类输入图像的真实边界框参考值来检测行人目标。在训练模型前，首先根据已经标注了边界框的数据集，用K-means聚类算法[23]得出初始的锚箱尺寸和比例，具体候选框的高度和宽度如表2所示。

表 2　候选框的宽度和高度表

检测层尺寸(像素)	(宽度，高度)	(宽度，高度)	(宽度，高度)
13×13	(41,103)	(53,138)	(77,205)
26×26	(30,74)	(30,94)	(35,84)
104×104	(20,30)	(20,51)	(27,61)

训练过程中本文使用的是随机梯度下降法[24]，设定初始学习率为0.001，动量(momentum)为0.9，权重衰减(decay)为0.005，批大小(batchsize)为16，使用0.001 的学习率迭代训练数据12000次，然后再用0.0001的学习率迭代到18000 次，最后用0.00001 的学习率迭代到20000 次。正负样本的IOU[25](Intersection Over Union)设置为0.5，即大于此阈值就设定为正样本，否则设定为负样本。

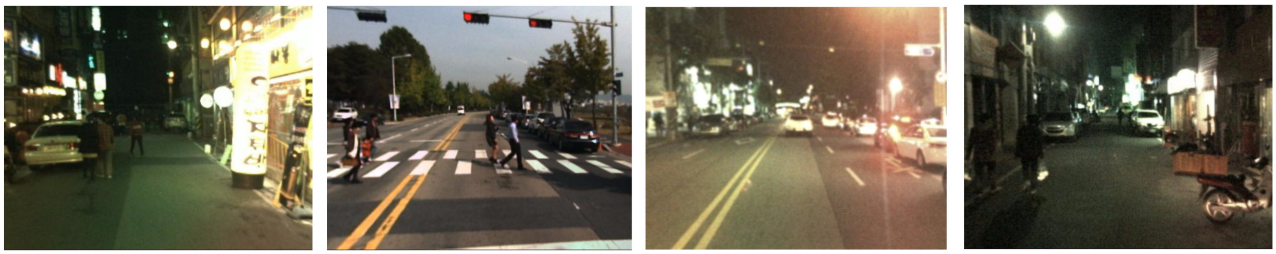
3.4　实验结果与分析

3.4.1　与基于单可见光或红外图像的算法在检测效果上的直观视觉比较

为了更直观证明本文提出的融合思想的有效性，将决策融合后的检测结果与基于单传感器的检测结果进行对比。我们在数据集中选取了4幅图像来代表4类比较典型的生活场景下多尺度行人的图像。4幅图像如图7所示，其中图7(a)、图7(b)中的行人较为密集，大、小尺寸的目标兼有，并有相互遮挡的情况。图7(c)中行人目标较小，目标特征不明显，难以辨认。图7(d)中行人目标尺寸相对正常，作为对比。

图8显示了对图7中几幅图像或其相对红外图像进行行人检测的结果，其中图8(a1), (b1), (c1), (d1)为基于可见光图像的行人检测结果图，图8(a2), (b2), (c2), (d2)为基于相对红外图像的行人检测结果图，图8(a3), (b3), (c3), (d3)为基于本文提出的FPDNet的检测结果图。

由图8对同一幅图像的处理效果对比得出，单一行人检测器对行人目标均有一定程度的丢失，融合检测能够适应多种检测环境，取得较好的效果。如图8(a1)可见光图像中未检测出人群中的小目



(a) 目标密集 (b) 目标尺寸不一 (c) 目标偏小 (d) 目标尺寸正常

图 7 4幅行人检测实验图

(a1) 图7(a)的可见光检测结果 (b1) 图7(b)的可见光检测结果 (c1) 图7(c)的可见光检测结果 (d1) 图7(d)的可见光检测结果

(a2) 图7(a)的红外检测结果 (b2) 图7(b)的红外检测结果 (c2) 图7(c)的红外检测结果 (d2) 图7(d)的红外检测结果

(a3) 图7(a)的本文检测结果 (b3) 图7(b)的本文检测结果 (c3) 图7(c)的本文检测结果 (d3) 图7(d)的本文检测结果

(a) 图7(a)检测结果 (b) 图7(b)检测结果 (c) 图7(c)检测结果 (d) 图7(d)检测结果

图 8 融合检测效果对比图

标。图8(b2)和图8(c2)红外图像中均漏检了一个小目标，而只有目标尺度较大，较为明显的图7(d)图的检测结果，不论是可见光还是红外图像均检测出了目标，但红外的结果更为显著。

3.4.2 与其他基于融合的算法在检测指标上的比较分析

本文采用mAP(mean Average Precision)指标反映目标检测精度，fps(frame per second)为每秒的检测帧数，代表目标检测速度。式(2)和式(3)为mAP的计算公式，其中AP表示单一类别的检测精度；R表示数据集中某个类别所有目标的数量；n表示数据集中所有目标的数量；j表示目标的序号，若与真实值相关，I_j为1，否则I_j为0；而R_j是前j个目标中相关目标的数量。q表示某一类别；Q_R表示总的类别数量。mAP值介于0～1之间，值越大表示该算法的检测精度越高。

$$\mathrm{AP}=\frac{1}{R}\sum_{j=1}^{n}I_j\frac{R_j}{j} \tag{2}$$

$$\mathrm{mAP}=\frac{1}{|Q_R|}\sum_{q\in Q_R}\mathrm{AP}(q) \tag{3}$$

表3显示了ACF+T+THOG[26], HalFus+TSD-CNN[27], TSDCNN+Ada[28], SSD, YOLOv3和YOLOv3-tiny算法与本文的FPDNet多尺度融合行

表 3 网络模型的对比结果表

模型	mAP(%)	FPS
ACF+T+THOG	71.49	32
HalFus+TSDCNN	88.24	2.5
TSDCNN+Ada	89.03	1.3
SSD	88.01	42
YOLOv3	91.35	45
YOLOv3-tiny	80.57	155
FPDNet	91.29	68

人检测网络的检测性能比较结果。7种模型均使用了同一种数据集数据进行训练和测试。实验结果表明，本文的检测网络在准确性方面与拥有最高准确度的YOLOv3模型相差不大，但其检测速度却比YOLOv3有了明显提高。在速度方面仅次于YOLOv3-tiny，但比YOLOv3-tiny的检测精度提高了超过10%。而其他模型的检测效果与本文算法相比，在检测率和速度上均无法抗衡。综合在检测精度和速度两方面的表现，本文采用的FPDNet与其他几种常用的目标检测模型相比，具有最佳的检测效果。考虑到实际生活中对行人检测算法在对检测速度和准确性上的双重要求，本文算法显然更有利于应用在实际相关领域中。

4　结束语

本文提出了一种能在复杂背景环境中对多尺度行人目标进行高效检测的算法——融合的行人目标检测网络(FPDNet)。本文算法提出的特征提取骨干网络，其精简的参数量改善了检测算法的检测效率；针对行人目标的尺度变化，在网络结构中采用了多尺度检测技术，尤其改善了小尺寸行人目标的检测效果；此外，算法利用可见光图像和红外图像中目标特征的差异，决策融合了从可见光和红外图像中提取的行人目标信息，使得检测更具鲁棒性。实验结果表明，相对于单一红外或可见光行人检测器，以及其他常见的目标检测算法，FPDNet不仅在检测精确度上表现良好，且具备实时应用性，可以更好地应对多变复杂场景和环境下的多行人目标检测问题。

参 考 文 献

[1] SAGAR U, RAJA R, and SHEKHAR H. Deep learning for pedestrian detection[J]. *International Journal of Scientific and Research Publications*, 2019, 9(8): 66–69. doi: 10.29322/IJSRP.9.08.2019.p9212.

[2] PRISCILLA C V and SHEILA S P A. Pedestrian detection - A survey[C]. Proceedings of the 1st International Conference on Innovative Computing and Cutting-edge Technologies, Istanbul, Turkey, 2020: 349–358. doi: 10.1007/978-3-030-38501-9_35.

[3] CHEN Runxing, WANG Xiaofei, LIU Yong, *et al*. A survey of pedestrian detection based on deep learning[C]. Proceedings of the 8th International Conference on Communications, Signal Processing, and Systems, Singapore, 2020: 1511–1516.

[4] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. *International Journal of Computer Vision*, 2004, 60(2): 91–110. doi: 10.1023/B: VISI.0000029664.99615.94.

[5] 孙锐, 陈军, 高隽. 基于显著性检测与HOG-NMF特征的快速行人检测方法[J]. 电子与信息学报, 2013, 35(8): 1921–1926. doi: 10.3724/SP.J.1146.2012.01700.

SUN Rui, CHEN Jun, and GAO Jun. Fast pedestrian detection based on saliency detection and HOG-NMF features[J]. *Journal of Electronics & Information Technology*, 2013, 35(8): 1921–1926. doi: 10.3724/SP.J.1146.2012.01700.

[6] 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*, 2010, 32(9): 1627–1645. doi: 10.1109/TPAMI.2009.167.

[7] HASTIE T, ROSSET S, ZHU Ji, *et al*. Multi-class AdaBoost[J]. *Statistics and its Interface*, 2009, 2(3): 349–360. doi: 10.4310/SII.2009.v2.n3.a8.

[8] BREIMAN L. Random forests[J]. *Machine Learning*, 2001, 45(1): 5–32. doi: 10.1023/A: 1010933404324.

[9] 陈勇, 刘曦, 刘焕淋. 基于特征通道和空间联合注意机制的遮挡行人检测方法[J]. 电子与信息学报, 2020, 42(6): 1486–1493. doi: 10.11999/JEIT190606.

CHEN Yong, LIU Xi, and LIU Huanlin. Occluded pedestrian detection based on joint attention mechanism of channel-wise and spatial information[J]. *Journal of Electronics & Information Technology*, 2020, 42(6): 1486–1493. doi: 10.11999/JEIT190606.

[10] REN Jing, REN Rui, GREEN M, *et al*. Defect detection from X-ray images using a three-stage deep learning algorithm[C]. Proceedings of 2019 IEEE Canadian Conference of Electrical and Computer Engineering, Edmonton, Canada, 2019: 1–4. doi: 10.1109/CCECE.2019.8861944.

[11] PAN Meiyan, CHEN Jianjun, WANG Shengli, *et al*. A novel approach for marine small target detection based on deep learning[C]. Proceedings of the IEEE 4th International Conference on Signal and Image Processing, Wuxi, China, 2019: 395–399. doi: 10.1109/SIPROCESS.2019.8868862.

[12] GIRSHICK R, DONAHUE J, DARRELL T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 580–587. doi: 10.1109/CVPR.2014.81.

[13] GIRSHICK R. Fast R-CNN[C]. Proceedings of 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 1440–1448. doi: 10.1109/ICCV.2015.169.

[14] REN Shaoqing, HE Kaiming, GIRSHICK R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137–1149. doi: 10.1109/TPAMI.2016.2577031.

[15] REDMON J, DIVVALA S, GIRSHICK R, *et al*. You only look once: Unified, real-time object detection[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 779–788. doi: 10.1109/CVPR.2016.91.

[16] REDMON J and FARHADI A. YOLO9000: Better, faster, stronger[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6517–6525. doi: 10.1109/CVPR.2017.690.

[17] REDMON J and FARHADI A. YOLOv3: An incremental improvement[J]. arXiv: 1804.02767, 2018.

[18] LIU Wei, ANGUELOV D, ERHAN D, *et al*. SSD: Single shot multibox detector[C]. Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 2016: 21–37. doi: 10.1007/978-3-319-46448-0_2.

[19] 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. doi: 10.1109/tpami.2015.2389824.

[20] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778. doi: 10.1109/CVPR.2016.90.

[21] LIU Weiyang, WEN Yandong, YU Zhiding, *et al*. Large-margin Softmax loss for convolutional neural networks[C]. Proceedings of the 33rd International Conference on Machine Learning, New York, USA, 2016: 507–516.

[22] HWANG S, PARK J, KIM N, *et al*. Multispectral pedestrian detection: Benchmark dataset and baseline[C]. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 1037–1045. doi: 10.1109/CVPR.2015.7298706.

[23] KANUNGO T, MOUNT D M, NETANYAHU N S, *et al*. An efficient K-means clustering algorithm: Analysis and implementation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, 24(7): 881–892. doi: 10.1109/TPAMI.2002.1017616.

[24] BOTTOU L. Stochastic gradient descent tricks[M]. Neural Networks: Tricks of the Trade. 2nd ed. Berlin Germany: Springer, 2012: 421–436. doi: 10.1007/978-3-642-35289-8_25.

[25] RAHMAN M A and WANG Yang. Optimizing intersection-over-union in deep neural networks for image segmentation[C]. Proceedings of the 12th International Symposium on Advances in Visual Computing, Las Vegas, USA, 2016: 234–244. doi: 10.1007/978-3-319-50835-1_22.

[26] KROTOSKY S J and TRIVEDI M M. On color-, infrared-, and multimodal-stereo approaches to pedestrian detection[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2007, 8(4): 619–629. doi: 10.1109/TITS.2007.908722.

[27] LIU Jingjing, ZHANG Shaoting, WANG Shu, *et al*. Multispectral deep neural networks for pedestrian detection[C]. Proceedings of 2016 British Machine Vision Conference, York, UK, 2016: 73.1–73.13. doi: 10.5244/C.30.73.

[28] KÖNIG D, ADAM M, JARVERS C, *et al*. Fully convolutional region proposal networks for multispectral person detection[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, USA, 2017: 243–250. doi: 10.1109/CVPRW.2017.36.

周薇娜：女，1982年生，副教授，研究方向为图像处理、电路和嵌入式系统、人工智能.

孙丽华：女，1995年生，硕士生，研究方向为模式识别与图像处理.

徐志京：男，1972年生，副教授，研究方向为海上智能交通系统、信息获取与智能处理.

责任编辑：陈 倩