doi:10.3772/j.issn.1002-0470.2022.09.004

基于深度信息融合的密集目标检测①

王建浩② 呼子宇③ 张翮翔 代 言 郝若欣 高泽航

(燕山大学电气工程学院 秦皇岛 066004)

摘 要 针对密集行人检测精度低的问题,提出基于深度信息融合的密集目标检测方法——YOLOv4-SD。该方法通过将 single scale Retinex (SSR)与目标检测算法信息融合,增强输入图像质量,凸显图像中更多的信息元素;并对 YOLOv4 算法中特征融合层进行改进,增加原始图像特征的利用率,深度优化特征融合层的网络结构。在 VOC 2012 等数据集上进行对比实验,结果表明在保持检测速度的前提下,该算法的平均检测精度和交并比分别提高了 7.7% 和 5.2。对于数据集中边缘低像素或高重叠的行人目标,YOLOv4-SD 算法能够较为准确地检测出特殊目标具体位置。

关键词 信息融合; 图像处理; 目标检测; 深度学习; 数据聚类

0 引 言

基于特征优化的行人目标检测[1]在无人驾驶[2]、智能监控、行人行为分析与动作识别[3]以及智能机器人[4]等领域都有重要的作用。目前,深度卷积神经网络[5]广泛应用于计算机视觉领域,并且按照卷积层的结构可以将行人检测算法分为 One-stage 检测算法和 Two-stage 检测算法两类。One-stage 算法基于线性回归的思想对目标进行分类和定位,代表算法有 YOLO[6]系列算法、SSD[7]算法和 Retina-Net[8]算法。Two-stage 算法基于区域建议的思想对目标进行检测,代表算法有 Fast R-CNN[9]算法、Faster R-CNN[10]及 Pyramid Network[11]算法。对于密集行人检测问题,由于目标行人存在无序性、高重叠性和复杂性,使得现有的检测网络出现高频率的误检和漏检现象,从而降低检测精度。

近年来,文献[12]提出降低交并比(intersection of union, IOU)的阈值,对严重遮挡的目标放宽检测标准的方法。该方法提高了检测的能力,使网络能够检测出严重遮挡的目标。但是同时也降低其他非行人目标的阈值,导致网络出现大量的假阳性检测结果;通过改进 Faster R-CNN 网络的结构,对于同一个目标,网络可以同时给予可见框和标准框。通过对两个目标框的综合评判,确定目标框的正负性,进而增加了对密集目标的检测精度。该方法虽增加了检测层,但网络的参数量倍化增加,严重影响密集行人检测的速度。

为了解决密集行人检测精度低的问题,本文提出一种基于深度信息融合的密集行人检测算法——YOLOv4-SD。该算法以 YOLOv4 作为骨骼算法,联合图像处理算法对目标图像质量进行改进,削弱目标遮挡和虚化问题对检测的影响。修改 YOLOv4 的特征融合层结构,增加网络对底层信息的利用率,使网络对密集目标更加敏感。

1 SSR 图像处理算法

现有的目标检测网络中大部分使用简易的图像变换算法[6-7],这类算法通过改变图像的旋转角度、亮度等参数来扩展检测视野。但面对密集行人检测

① 国家自然科学基金(62003296)和河北省自然科学基金(F2016203249)资助项目。
② 男,1999 年生,硕士生;研究方向:图像处理,深度学习; E-mail:wjh991006@sina.com。
③ 通信作者,E-mail: hzy@ysu.edu.cn。
(收稿日期:2021-02-01)

问题,实际检测效果并不理想。为解决图像信息质量表现问题,在原算法基础上联合单尺度 Retinex (single scale Retinex, SSR)[13]图像处理算法进行密集行人目标检测。SSR 图像处理算法是 Retinex[14]经典算法之一,可以有效解决图片光线不定、亮度不够等图像质量问题。Retinex 算法以物体颜色是由物体对不同波长光线的反射能力决定为理论基础,物体的色彩不会因为光照的均匀性而改变,即颜色恒常性。Retinex 算法在动态范围压缩、边缘增强等方面表现良好。因此 Retinex 算法能够对不同类型的图像进行自适应的增强,公式如下:

$$S(x, y) = R(x, y) \times L(x, y) \tag{1}$$

图像 $S(x, y)$ 被分解为一个反射图像 $R(x, y)$ 和一个亮度图像(或入射图像)$L(x, y)$。基于 Retinex 理论的图像增强算法的目的是估计光照 L,分解反射光 R,从而消除不均匀光照的影响,达到优化图像的目的。其增强过程的公式如下:

$$\log(S) = \log(R \times L) \tag{2}$$

$$s = \gamma + \iota \tag{3}$$

SSR 图像增强算法处理过程如下所示。

步骤 1 读取图像,若原图像为彩色图片,则将颜色分通道处理,并且将分量像素值转换为浮点数,利用上述公式转换到对数域;若为灰度图,则直接将各像素的灰度值由整数型转换为浮点数,并转换到对数域。

步骤 2 通过计算图像中像素点与邻域中像素的加权平均值对图像中照度变化做合理估计,并将其清除,只保留图像中物体的反射属性,利用高斯模板对原图做卷积,得到低通滤波后的图像,公式为

$$D(x, y) = S(x, y) \times F(x, y) \tag{4}$$

步骤 3 原图像减去步骤 2 中所得的图像就可得到高频增强图像,公式为

$$r(x, y) = \log S(x, y) - \log D(x, y) \tag{5}$$

步骤 4 如果原图为彩色图,那么每一个通道对应一个 $r(x, y)$;若为灰度图,则仅有一个 $r(x, y)$。

步骤 5 对 $r(x, y)$ 取反对数映射到实数域,得到增强后的图像 $R(x, y)$,此时的 $R(x, y)$ 可能需要进行线性拉伸并且转换成相应格式输出显示才符合相应的系统要求。

上述步骤对应的算法流程图如图 1 所示。

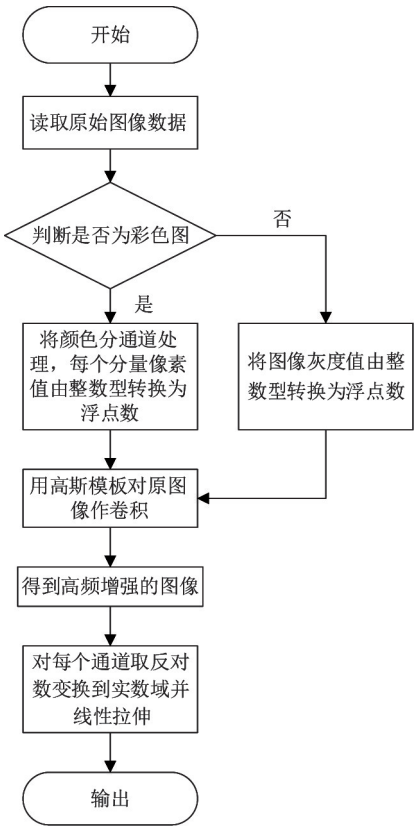


图 1 SSR 算法流程图

2 YOLOv4 特征层深度信息融合

YOLO 系列算法作为 One-stage 算法中的经典算法,因其较高的检测精度和超高的检测速度,在目标检测领域有着广泛的应用。YOLOv4 算法作为 YOLO 系列的代表算法,在 MS COCO 数据集上拥有高达 62 FPS 和 43.5% AP 值(输入图像大小为 608 × 608)。YOLOv4 的主体结构为:基础骨架网络是 CSPDarknet 53,附加网络为空间金字塔池化网络(spatial pyramid pooling, SPP)[15]和 PANet[16],检测部分是 YOLOv3。图 2 为 YOLOv4 整体结构图。

CSPDarknet 53 网络是一个高度模块化的特征提取网络,包含 725 × 725 的感受野和 2.76×10^7 的参数量。在保持检测速度为 60 FPS 左右的前提下,提高了目标检测精度。SPP 通过对 107 层网络进行尺度为 5 × 5,9 × 9,13 × 13 的最大池化,得到了网络结构中的 108 层、110 层、112 层。通过最大池化操作,图像的感受野进一步加大,从而提高了 YOLOv3

检测模块对于图像特征的感受能力。PANet 代表特征融合网络，具体结构如图 3 所示，主要作用是将 56 层、86 层、116 层的特征图进行多层次融合，将融合后的特征图输入到最后的目标检测模块中。

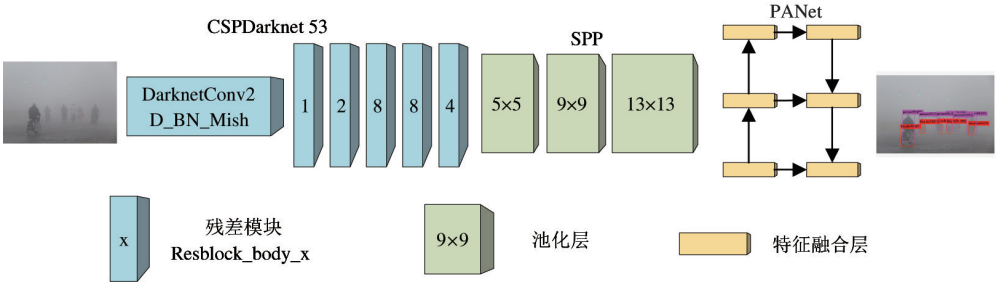


图 2　YOLOv4 结构图

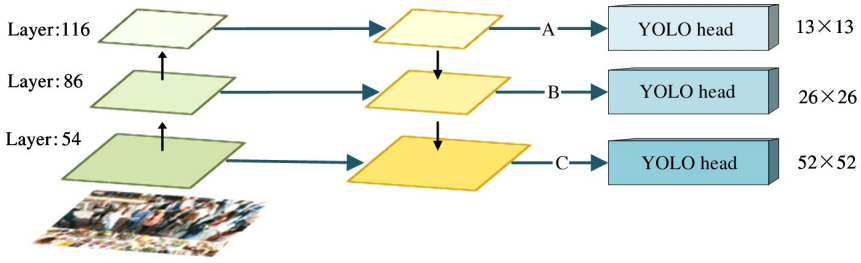


图 3　原 PANet 网络结构图

密集行人图像由于行人与特征环境间存在着严重遮挡的现象，导致被检测行人出现大量残缺和行人与行人间的肢体干扰特征。相比于其他领域的目标检测，密集行人检测中行人与行人间的关系更加复杂。现有的 YOLOv4 网络面对复杂的行人特征关系表现较差，出现高频率的漏检误判现象。为达到检测速度和检测精度的相对平衡，本文提出了一种特征层深度信息融合的网络结构（YOLOv4-D）。

YOLOv4-D 算法选用 YOLOv4 算法作为密集行人检测基础骨架网络，并对 PANet 结构进行改进。通过对特征融合层中增加特征提取层初始成果的前向通道，增加网络对于原始特征的敏感性。由于原始特征的利用率增加，使得网络对于密集行人中小目标的检测能力有所提升。

增加信息融合层[17-18]，使初始的特征图和多次卷积融合后的特征图进行二次融合。从具体结构展开图 4 中可观察到，该部分的结构为信息连接点和卷积层组。信息连接层通过 Concatenation 对特征图

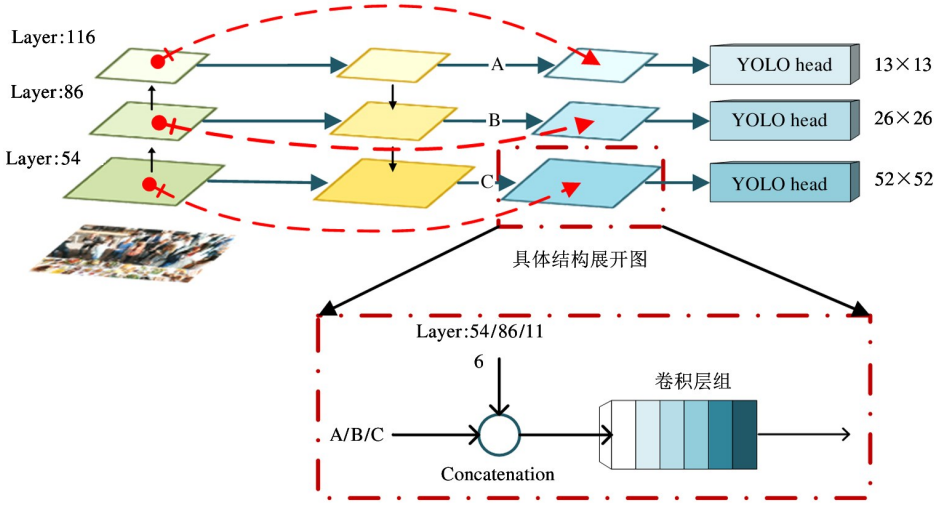


图 4　改进后的 PANet 网络结构图

进行二次融合,卷积层组对融合后的特征图进行信息提取。改进后的 PANet,对于原始图像的利用率提高,从而强化网络对于密集目标的敏感性,最终达到提高检测精度的效果。

3 高密度人群数据的重聚类

YOLOv4 的算法使用 anchor boxes 对目标框进行估计,此机制可以有效地解决进行密集行人检测时多目标干扰的问题。当目标的边界与特定的 anchor 有较高交并比(intersection-over-union, IOU)时,则与该 anchor 相关的区域负责将预测框回归到最终边界,经过网络训练后,每个中心点在空间上都是相对独立的,不会受到其他 anchor 的目标影响。K-means 聚类算法中是常用的聚类算法,其特点为算法流程简易、运算速度快。具体算法过程如下所示。

步骤 1 设定函数值 K,K 表示最终聚类的中心点数量。

步骤 2 从数据集中随机选择 K 个数据点作为初始质心(Centroid)。

步骤 3 对集合中除质心之外的每一个子目标,使用欧几里得距离(Euclidean distance)或者曼哈顿距离(Manhattan distance),计算其与每一个质心的距离,并将其分到距离最近的质心分类中。IOU 的计算方式对 bounding box 进行分析,其公式为

$$IOU = \frac{|B \cap C|}{|B \cup C|} \tag{6}$$

其中 B 为 box,C 为 centroid。box 为样本聚类结果,centroid 为簇的中心,IOU 为所有簇心与所有聚类框的交并比。

步骤 4 当质心周围有一个或多个子目标,通过上述方式再次计算出新质心。

步骤 5 如果新质心和原质心的距离小于某一设置的阈值时,说明新质心收敛。

步骤 6 如果新质心和原质心距离过大,则迭代步骤 3 ~ 5。图 5 为 K-means 计算流程图。表 1 为其详细数据。在不影响其他参数的前提下,文章选用当 K 值为 9 时的输出值作为网络的 anchor 进行网络训练。

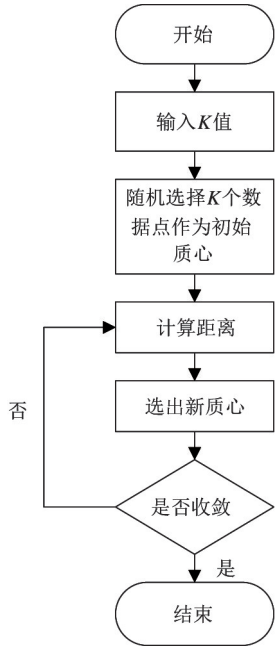


图 5 K-means 算法流程图

表 1 不同 K 值对应的数据结果表

K 值	聚类 1	聚类 2	聚类 3	聚类 4	聚类 5	聚类 6	聚类 7	聚类 8	聚类 9
6	[24,56]	[30,68]	[40,86]	[57,109]	[111,201]	[101,204]	-	-	-
7	[22,49]	[32,59]	[41,84]	[60,122]	[84,190]	[92,203]	[103,220]	-	-
8	[21,45]	[33,68]	[39,81]	[49,104]	[52,110]	[63,135]	[70,169]	[90,204]	-
9	[24,50]	[32,67]	[40,86]	[46,98]	[54,106]	[60,128]	[72,165]	[89,210]	[90,207]

4 实验平台和数据集

表 2 展示了实验平台的特定硬件和软件配置。所有实验训练均在该平台上进行。初始参数设置如下:学习率为 0.001,批次为 64,步长为(80 000, 90 000),最大批次为 10 000,分支数量为 16。本文采用 2000 个行人图像作为行人数据集,其中大多数行人图像来自标准数据集(VOC 2012),其余部分则

在工作期间收集。本文收集的行人图像中行人外观丰富、背景复杂、个体差异明显、重叠程度不同，符合行人数据集的要求。对于数据集中的 2000 张图像，训练集为 1400 张图像，验证集为 400 张图像，其余 200 张图像为测试集。行人数据集的示例图像如图 6所示。

图 6　行人数据集示例图

表 2　实验平台软硬件配置表

软件硬件配置	参数
操作平台	Ubuntu 18.04
CPU/GHz	Intel(R)Xeon(R)E5-2667,2.90
GPU	CUDA 10.0, CUDNN 7.6
RAM	16 GB
深度学习	Darknet, caffe

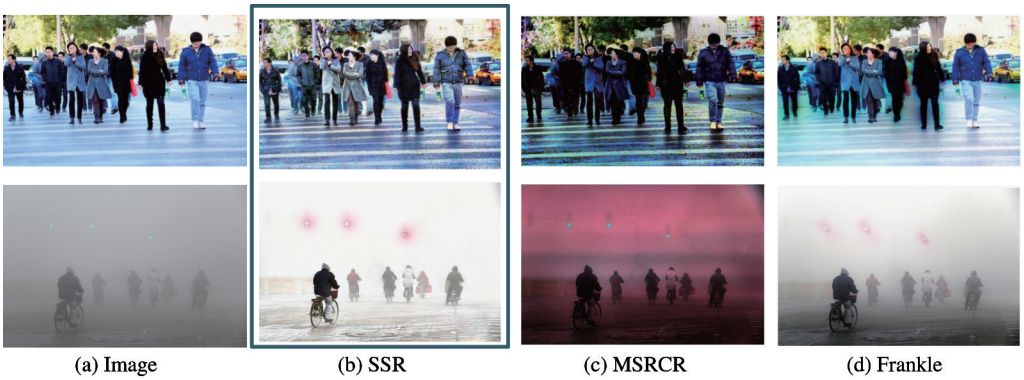
4.1　图像处理算法对比实验

为探索图像处理算法对图像的影响，将 SSR 算法、MSRCR[19] 和 Frankle-McCann[20] 算法进行图像增强效果的比较。实验通过对比平均值、标准偏差、信息熵和颜色熵的参数指标对算法做出客观评价。平均值代表图像整体亮度平均值，平均值越高，图像亮度越高。标准偏差代表图像的对比度。信息熵表示图像中的信息量。颜色熵用于测量图像的颜色增强程度。实验结果如表 3 所示，具体图像实验示例如图 7 所示。

通过表 3 可知，SSR 算法在平均值、信息熵和颜色熵等方面的数据优于其他算法。图 7 实验示例表明，SSR 算法在亮度对比度等方面有较好的优化效果。SSR 图像增强算法通过高斯模块做卷积后，可以极大地提高图像的亮度、颜色和信息量，从而提高了输入到网络中的图像质量。由表 4 的综合实验数据可以得出，这种提升图像质量的方式能够在保证检测速度的前提下，提高目标检测的精度和效果。

表 3　不同图像增强算法数据结果表

图像处理算法	平均值	标准偏差	信息熵	颜色熵
YOLOv4	60.905	48.615	5.016	15.284
YOLOv4-SSR	108.436	40.823	7.641	19.817
YOLOv4-MSRCR	40.735	21.984	4.343	10.376
YOLOv4-Frankle-McCann	105.139	68.872	6.735	17.463



(a) Image　(b) SSR　(c) MSRCR　(d) Frankle

图 7　图像增强算法对比图

表 4　不同目标检测网络实验数据表

目标检测算法	*mAP*/%	*IOU*	*Recall*/%	time/s
Faster R-CNN	94.2	90.01	93	0.052
YOLOv3	83.9	81.25	78.32	0.015
YOLOv4	89	86.33	86.91	0.019
YOLOv4-D	92.1	87.71	90.74	0.021
YOLOv4-SD	93.8	89.32	92.11	0.025
Efficientnet	94.7	90.08	95.02	0.09
ATSS	88.1	86.42	86.9	0.068
Center Mask	92.4	89.19	91.24	0.072

4.2　目标检测网络实验

本节实验内容为研究改进后的 YOLOv4 网络结构对密集行人目标检测的实际检测效果。为了方便描述,基于深度信息融合的密集行人检测命名为 YOLOv4-SD。将 Faster R-CNN、YOLOv3、YOLOv4、YOLOv4-D、YOLOv4-SD、Efficientnet[21]、ATSS[22] 和 Center Mask[23] 算法进行比较实验。mAP、IOU、Recall 和 time 作为网络评价指标。为了使实验更加全面,将是否使用 K-means 算法来作为对照指标。如果使用 K-means 算法对目标进行聚类,则标记为 Y,反之标记为 N。相关参数定义如式(7)、(8)、(9)所示。

$$mAP = \frac{\sum AP}{NC} \tag{7}$$

$$IOU = \frac{box_T \cap box_P}{box_T \cup box_P} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

式(7)中,$\sum AP$ 代表检测的各个类别正确率之和,NC 代表检测类别总数。式(8)中 box_T 代表了真实框的面积,box_P 表示预测框的面积。式(9)中,TP 表示真阳性,即网络判断为真,实际行人目标也为正样本;FN 表示假阴性,即网络判定为假,实际目标也为负样本的目标。表 4 为不同目标检测网络在同等实验环境下的评价指标数据表。根据表中数据可知,YOLOv4-SD 算法相较于 YOLOv4 算法将 mAP 值提高了 7.7%,检测速度仅增加 0.006 s。改进后的算法在 IOU,Recall 等指标上均有提升。通过消融实验可知,在保证检测速度的前提下,联合图像增强算法和深化特征层融合结构,检测网络对于行人整体的敏感性得到综合提升,间接证明改进方法的有效性,具体数据见表 5 和表 6。本文希望在基本不增加时间成本的前提下提高检测精度,所以仅与经典网络 Faster R-CNN 以及 YOLO 系列做可视化对比。如图 8 所示,图中仅有 YOLOv4-D 与 YOLOv4-SD 能够检测出图像边缘的 2 个行人,其他网络则出现了漏检现象。

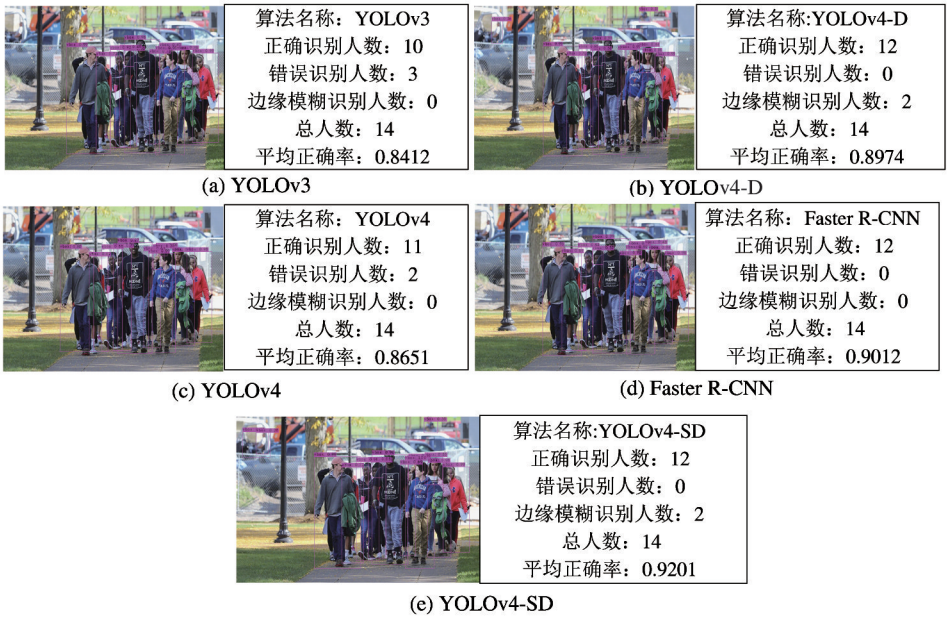


图 8　各个网络实际检测效果图

表 5 图像处理算法消融分析实验数据表

图像处理算法	平均值	标准偏差	信息熵	颜色熵
YOLOv4	60.905	48.615	5.016	15.284
YOLOv4-SSR	108.436	40.823	7.641	19.817

表 6 目标检测算法消融分析实验数据表

目标检测算法	K-means	*mAP*/%	*IOU*	*Recall*/%	time/s
YOLOv4	N	86.1	84.12	83.71	0.019
YOLOv4-D	N	90	86.25	87.89	0.021
YOLOv4-SD	N	91.2	88.21	89.09	0.025
YOLOv4	Y	89	86.33	86.91	0.019
YOLOv4-D	Y	92.1	87.71	90.74	0.021
YOLOv4-SD	Y	93.8	89.32	92.11	0.025

5 结 论

在处理密集行人检测问题时,由于图像画质和特征融合网络结构在检测过程中均具有重要作用,故采用 SSR 图像增强算法,提高图像质量;改进 YOLOv4 特征层结构,增强网络对于原图像特征信息的利用率,提高了检测原始目标的能力。综合分析本文算法与 Faster R-CNN、YOLOv3、YOLOv4 等算法在实验数据集上的比较,YOLOv4-SD 在 mAP、Recall、IOU 等数值上相对于 YOLO 系列算法均有所提升,具体数值分别为 7.7%、8.4%、5.2。相对于 Faster R-CNN 和 Efficientnet 算法,YOLOv4-SD 虽然在检测精度评价指标上稍有劣势,但在检测速度上明显优于其他高精度算法且对设备性能要求较低。通过实际的检测效果实验,YOLOv4-SD 能够准确地检测出边缘低像素和高遮挡的目标行人,而其他算法对于该类特殊的行人目标检测效果不佳。消融实验与实验数据表明本文算法平衡了检测精度与检测速度。本文基于计算机视觉技术,利用深度学习算法对密集目标检测问题展开研究,解决了检测精度和检测速率过低的问题。

未来的工作将研究和探寻以下几个方面。

(1)行人检测与跟踪相结合。检测到目标后如何更准确地在下一帧图像中匹配到目标,对于行人检测系统十分重要。

(2)复杂背景分离。预检测的图像能否使检测目标与复杂背景分割是提高检测精度的途径。

(3)准确标注数据集。数据集对于行人检测来说是十分关键的,但是目前的数据集的标注框并没有统一尺寸,这就导致在训练时模型学习到的特征不准确从而影响检测效果。

参考文献

[1] 刘琼. 导引概率图与显著特征相结合的行人目标检测[J]. 高技术通讯, 2016, 26(5): 464-474

[2] ZHANG X, GAO H, GUO M, et al. A study on key technologies of unmanned driving[J]. *CAAI Transactions on Intelligence Technology*, 2016,1(1): 4-13

[3] 汤春明,卢永伟. 基于改进的稀疏重构算法的行人异常行为分析[J]. 计算机工程与应用, 2017, 53(8): 165-169

[4] 龙慧,朱定局,田娟. 深度学习在智能机器人中的应用研究综述[J]. 计算机科学, 2018, 45(11A): 43-47, 52

[5] 周飞燕,金林鹏,董军. 卷积神经网络研究综述[J]. 计算机学报, 2017,40(6):1229-1251

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

[7] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multi-box detector[C]// European Conference on Computer Vision, Amsterdam, Netherlands, 2016: 21-37

[8] WANG Y, WANG C, ZHANG H, et al. Automatic ship detection based on RetinaNet using multi-resolution Gaofen-3 imagery[J]. *Remote Sensing*, 2019,11(5):531

[9] GIRSHICK R. Fast R-CNN[C]//Proceedings of IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 1440-1448

[10] REN S, HE K, 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*, 2016, 39(6): 1137-1149

[11] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 2881-2890

[12] JIANG B, LUO R, MAO J, et al. Acquisition of localization confidence for accurate object detection[C]//Proceedings of European Conference on Computer Vision, Munich, Germany, 2018: 784-799

[13] SI L, WANG Z, XU R, et al. Image enhancement for surveillance video of coal mining face based on single-scale retinex algorithm combined with bilateral filtering [J]. *Symmetry*, 2017, 9(6): 93

[14] FUNT B, CIUREA F, MCCANN J. Retinex in Matlab [C]//Color and Imaging Conference on Society for Imaging Science and Technology, Scottsdale, USA, 2000,13(1):112-121

[15] HE K, ZHANG X, REN S, 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

[16] WANG K, LIEW J H, ZOU Y, et al. Panet: few-shot image semantic segmentation with prototype alignment[C]//Proceedings of IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 2019: 9197-9206

[17] 李国友,纪执安,张凤煦. 融合多层深度特征的核相关滤波跟踪算法[J]. 高技术通讯, 2020, 30(2): 126-133

[18] 杨海清,许倩倩,唐怡豪,等. 结合卷积神经网络多特征融合的相关滤波跟踪[J]. 高技术通讯, 2020, 30(10): 1085-1092

[19] GAO Y, YUN L, SHI J, et al. Enhancement MSRCR algorithm of color fog image based on the adaptive scale[C]//The 6th International Conference on Digital Image Processing, Athens, Greece, 2014: 91591B1-91591B7

[20] ZHANG H, HU Z, HAO R. Joint information fusion and multiscale network model for pedestrian detection[J]. *The Visual Computer*, 2021, 37: 2433-2442

[21] TAN M, LE Q. Efficientnet: rethinking model scaling for convolutional neural net-works[C]//International Conference on Machine Learning, Long Beach, USA, 2019: 6105-6114

[22] BIFFI L J, MITISHITA E, LIESENBERG V, et al. ATSS deep learning-based approach to detect apple fruits [J]. *Remote Sensing*, 2021, 13(1): 54-73

[23] LEE Y, PARK J. Centermask: real-time anchor-free instance segmentation[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020:13906-13915

Dense target detection based on deep information fusion

WANG Jianghao, HU Ziyu, ZHANG Hexiang, DAI Yan, HAO Ruoxin, GAO Zehang

(School of Electrical Engineering, Yanshan University, Qinhuangdao 066004)

Abstract

Aiming at the problem of low accuracy of dense pedestrian detection, YOLOv4-SD a dense target detection method based on deep information fusion, is proposed. The method combines the information of single scale Retinex (SSR) and the target detection algorithm to improve the quality of the input image and highlight more information elements in the image. By improving the feature fusion layer in the YOLOv4 algorithm, the network structure of the feature fusion layer is deeply optimized to achieve the purpose of improving the utilization of original image features. Comparison experiments are conducted on data sets VOC 2012, the experimental results show that the average detection accuracy and intersection ratio of the algorithm are increased by 7.7% and 5.2, while maintaining the detection speed. For predestrian targets with low pixels or high overlap at the edge of the data set images, YOLOv4-SD algorithm can accurately detect the specific location special targets.

Key words: information fusion, image processing, target detection, deep learning, data clustering