DOI:10.16799/j.cnki.csdqyfh.2024.06.058

# 改进 YOLO 算法的行人检测和慢行规划应用

刘丰嘉,王 桀,马 柱
(浙江省城乡规划设计研究院,浙江 杭州 310030)

**摘 要:** 行人是物体检测应用中非常重要的一类目标,在检测行人的同时对人流数量进行实时精准统计在诸多领域有着很大的实用意义。通过搭建深度学习环境,设定行人数据集并对其进行重新标注、训练和测试后,对统计视频中行人目标的正确率、漏检率和误检率行了统计。结果表明:基于改进 YOLO v4 的行人检测模型,能够更加准确、高效地识别密集场景下的行人目标,达到了预期目标,从而能够为城市慢行系统规划和商业区规划等领域提供应用价值。

**关键词:** 交通监控;深度学习;YOLO;行人检测
**中图分类号:** U491.1 **文献标志码:** A **文章编号:** 1009-7716(2024)06-0245-04

## 0 引 言

近年来,随着我国经济社会的发展和城市化进程的推进,公共场所聚集的人口数量不断增加。在规划和设计人行通道设施(例如人行过街天桥、交通枢纽、购物中心或体育场等)时,人行流量及其特征成为重要的规划和设计依据。特别是在倡导慢行和公交优先的背景下,慢行系统应基于人流特征与城市发展有机融合,推动城市交通向绿色、减少碳排放和可持续性发展。

常规的人流量计算方法如人工统计和压力传感器计数存在多种不足,如耗时、速度慢、精确度低、使用寿命短、安装调试不方便等。面对智慧城市的快速发展和大数据在城市规划领域的广泛应用,应重点研究慢行交通的数据,深入挖掘其理论和应用技术,充分发挥大数据的价值,使城市慢行系统真正实现数字化。

在全国智慧城市和安防项目大规模建设的背景下,公共场所摄像头的覆盖密度非常高。利用视频数据进行技术分析将成为交通规划和设计的新常态。基于视频的实时人流数据统计将有效地改善交通设计和管理。一方面,可以提高交通枢纽和道路交叉口等交通混合区域的运行效率;另一方面,有助于在大型公共场所合理组织行人流线,避免人流冲突和踩踏事故的发生。在新的需求推动下,如何从海量视频中提取有效的实时人流统计数据成为亟待解决的关键技术问题。

行人检测技术的发展已有很长时间,早期的算法主要采用传统的机器学习方法,如 SIFT[1]和 HOG[2]。随着计算机技术的进步,后期的算法主要分为两类:一类是基于候选区域的算法,如 R-CNN[3]、Fast R-CNN[4]和 Faster R-CNN[5];另一类是基于回归的算法,如 SSD[6]和 YOLO[7]。这些算法各有特点,其中基于回归的 YOLO 算法将物体分类和物体检测网络合并,通过全连接层完成。YOLO 算法的运算速度达到 45 帧 /s,完全符合实时性要求(人眼识别的图像连续性阈值为 24 帧 /s)。

YOLO 网络结构图见图 1。YOLO 算法能够减少目标检测的时间,经过多个版本的改进,其实时性得到了大幅提升,更适用于实时检测。

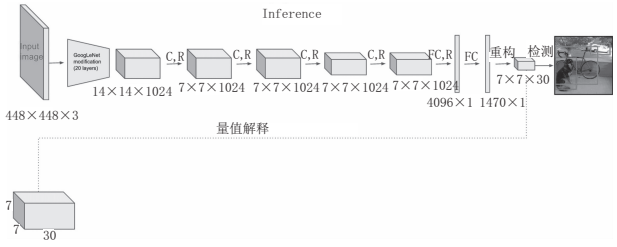


图 1 YOLO 网络结构图

然而,YOLO 算法也存在一定的缺点,例如对靠近物体和较小物体的检测效果较差。因此,在视频行人检测场景下,当行人密度高或相互遮挡时,原生的

收稿日期:2023-06-26
作者简介:刘丰嘉(1994—),男,硕士,工程师,从事交通规划工作。
通信作者:王桀(1988—),男,硕士,高级工程师,从事交通规划工作。电子信箱:759469490@qq.com

YOLO 算法无法准确检测。本文通过改进 YOLO 算法的参数设置，提高了检测的准确率。

## 1　YOLO 模型概况

### 1.1　卷积神经网络

YOLO 是一种基于卷积神经网络的目标检测算法。该算法将输入的图像划分为 1 个固定大小的网格，并对每个网格预测出 2 个边界框，每个边界框包含目标的置信度和在多个类别中的概率。通过这种方式，YOLO 能够同时检测出图像中的多个目标。

YOLO 的检测流程示意图见图 2。

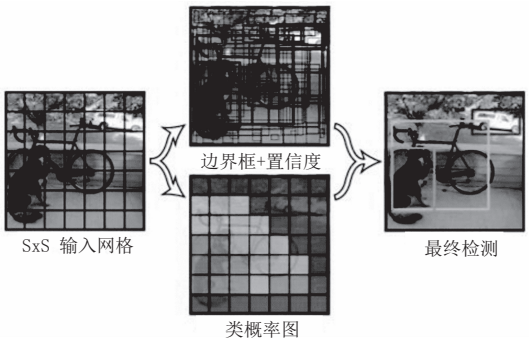


图 2　YOLO 的检测流程示意图

与传统的目标检测算法相比，YOLO 具有更快的检测速度，能够达到实时检测的要求。它通过将目标检测任务转化为一个回归问题，将物体分类和位置检测同时进行，从而减少了多次扫描图像的过程。此外，YOLO 还采用了多尺度训练和数据增强等技术，提高了算法的准确性和鲁棒性。

YOLO 算法的实现主要包括网络结构设计、损失函数定义和训练过程。网络结构采用卷积层、池化层和全连接层组成，以提取图像特征并输出目标的位置和类别信息。损失函数用于评估预测结果与真实标签之间的差异，并通过反向传播算法来更新网络参数。训练过程通过使用大量标注数据进行迭代优化，使得模型能够更好地适应目标检测任务。

总体而言，YOLO 卷积神经网络实现了高效准确的目标检测，在计算机视觉领域具有广泛的应用前景。

### 1.2　损失函数

YOLO 目标检测算法使用损失函数来评估模型的预测性能。该损失函数包括位置误差、置信度误差和类别误差 3 个方面。在设计损失函数时，需要合理地平衡这 3 个方面的重要性。

然而，YOLO 采用的是平方和误差损失函数，导致位置误差和类别误差的重要性无法区分。此外，不包含物体的目标窗口的置信度误差趋向于 0，会导致网络不稳定。为了解决这些问题，本文对 YOLO 的损失函数进行了调整。

经过测试，将位置误差的权重设置为 5，类别误差的权重设置为 1。同时，将不包含物体的目标窗口的置信度误差权重调整为 0.5，而包含物体的权重调整为 1。通过这些调整，提高了 YOLO 算法的检测准确率，使其更适用于实时检测场景。

## 2　视频片段的行人目标检测测试

### 2.1　实验环境配置

本文搭建的硬件环境为：Intel 酷睿 i7 11 代系列 CPU、16G 内存、Ge Force GTX1050Ti 显卡；软件平台为：Windows10 操作系统、CUDA10.0、CuDNN7.5.0，Python3.6、OpenCV 3 视频开源库、PyTorch1.0 深度学习框架。

### 2.2　行人数据集

YOLO v4 默认的类别识别结果基于 VOC 和 COCO 2 个数据集得到。在其默认的行人数据识别中，对于密集场景下人群重叠等情况的识别效果不太理想，因此需要重新通过标注数据集来提高算法检测的准确率。

新建的密集人群数据集主要来源于 2 个渠道：(1)上海科技大学开源的 ShanghaiTech 行人数据集；(2)道路监控卡口拍摄的包含人群的视频数据集。为了对这些数据集进行整理和标注，将它们分成了 4 000 张测试图片和 8 000 张训练图片。为了进行分类和标注自定义目标，采用了常用的深度学习标注工具 LabelImg。标注完成后，按照 YOLO v4 的配置规定，对图片进行了更名和路径保存。

在密集人群的视频场景中，通常只能看到行人的头部和肩部。因此，为了提高算法的检测效率，采用 3 种尺寸的矩形标注框，对行人肩部以上的特征区域开展预标定，最后通过将图像调整为固定的大小(数值为 416×416)，确保输入图像具有统一的尺寸，并最终形成能够应用于算法训练和推理的数据集。

YOLO 模型训练的主要参数设置见表 1。

### 2.3　训练结果及分析

针对密集行人检测，检测准确率很重要。本文选取平均准确率(mAP)作为评价指标，其准确率(Precision，以 $P$ 表示)和召回率(Recall，以 $R$ 表示)的定义为：

$$P=\frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{1}$$

表 1　训练的主要参数设置

| 训练参数 | 数值 | 训练参数 | 数值 |
|---|---|---|---|
| Batch | 64 | Max_batchs | 45 000 |
| Subdivisions | 32 | Momentum | 0.9 |
| Widths | 416 | Decay | 0.000 6 |
| Heights | 416 | Learning_rate | 0.002 |
| Channels | 3 | Burn_in | 1 500 |

$$R=\frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{2}$$

式中：TP 为正确检测的数量；FP 为误检的数量，即将背景检测为行人的数量；FN 为误把行人检测为背景的数量。Precision（$P$）表示的是分类器认为是正类并且确实是正类的部分占所有分类器认为是正类的比例；Recall（$R$）表示的是分类器认为是正类并且确实是正类的部分占所有确实是正类的比例。同时以 $F1$ 作为对 Precision 和 Recall 的综评指标，越接近 1，则效果越好，$F1$ 的表达式见式（3）。为求取 mAP 之值，以召回率 $R$ 为横坐标，准确率 $P$ 为纵坐标，绘制 $P$–$R$ 曲线，mAP 值根据式（4）得到。

$$F1=\frac{2PR}{P+R}=\frac{2\mathrm{TP}}{2\mathrm{TP+FN+FP}} \tag{3}$$

$$\mathrm{mAP}=\int_0^1 P(R)\,\mathrm{d}(R) \tag{4}$$

实验过程中，使用重新标定生成的权重文件，对数据集进行测试，改进前后 YOLO v4 算法对密集人群目标的测试效果如图 3 所示。

（a）原生模型

（b）改进模型

图 3　改进模型的检测图

观察图 3（a）可见，其右上方的目标受到了空间立柱的遮挡。由于未改进的原生 YOLO v4 算法在特征泛化能力方面较弱，对整体图片没有很好的全局感知能力，因此无法准确识别被遮挡的目标特征，导致右侧目标被遗漏。而改进后的 YOLO v4 算法能精准识别出图片中几乎所有行人。

此外，经过改进后的算法在处理画幅最上方目标时展现出较强的泛化能力。由于这些目标较小且受到遮挡，原生的 YOLO v4 算法划分图像的尺寸相对较大，可能会将小目标的特征与其他目标的特征包含在同一个检测网格单元内，导致很难准确捕捉该目标的特征，因此在原生的 YOLO v4 算法中可能无法识别出它们，这就造成了一些“应检未检”。然而，通过对检测结果的观察可以发现，改进后的模型在检测小目标方面表现出色，并且在存在遮挡、低能见度等恶劣观测条件下，对行人目标的检测也表现出良好的准确性。

为了检验模型的实战效果，对一段手机拍摄的地铁换乘通道的 15 min 左右视频进行识别，图像采集的分辨率为 1 028 × 1 920 dpi，如图 4 所示。分别基于原生 YOLO v4 模型和改进 YOLO v4 模型进行识别效果的评价。改进前后检测效果对比结果见表 2。

（a）原生模型　　（b）改进模型

图 4　改进模型的实测图

表 2　改进前后检测效果对比

| 模型 | 准确率 | 召回率 | $F1$ | 漏检率 | mAP |
|---|---|---|---|---|---|
| 原生模型 | 0.596 5 | 0.447 9 | 0.511 9 | 0.488 3 | 0.487 0 |
| 改进模型 | 0.777 6 | 0.654 9 | 0.711 0 | 0.289 2 | 0.596 3 |

从表 2 可以看出，基于改进的 YOLO v4 模型相对原生模型的检测准确率提高了 30%，且对比测试视频后发现，行人目标的检出率比较理想；在输入尺寸为 416 × 416 的图像上 mAP 值提高了 22%左右；

相较于原生的 YOLO v4，改进的 YOLO v4 在对检测目标的定位精度和漏检率方面都有显著提升，在实际工程中具有较强实用性。

## 3　视频行人目标检测的规划应用

在视频片段检验的基础上，进一步探索本算法在城市慢行系统规划编制中的应用。选取绍兴市柯桥区 6 个关键路段的视频数据进行分析，同时提取路段机动车流量，将视频人行流量和中心城区路段流量进行叠加，进行过街设施必要性的综合判定。视频选择位置及检测结果分布图见图 5。

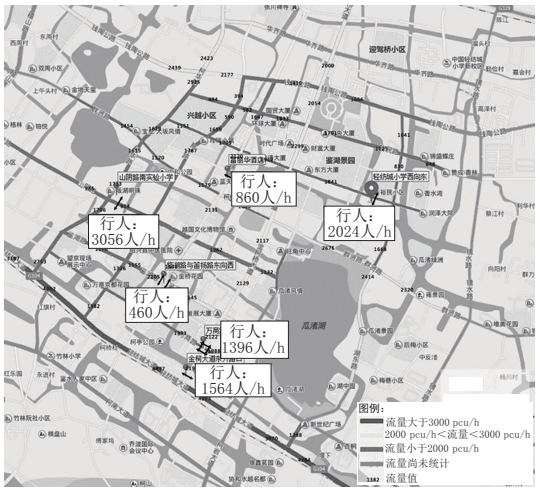


图 5　视频选择位置及检测结果分布图

该视频位置为某小学门口路段上，选择下午 15:50—16:05 的高峰放学峰期进行计数，得到横穿兴越路的人流量为每 15 min 有 506 人。放学时跨越兴越路的人流密集，学生放学高峰期采用交管现场人工疏导行人过街。现状兴越路为城市主要干道，双向机动车交通流量大，高峰期人车冲突严重，因此该路段也可以考虑设置行人过街设施，相关过程如图 6 所示。

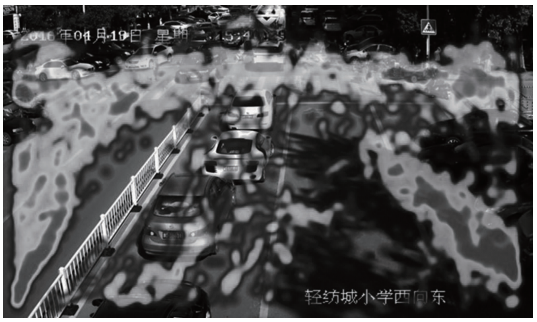

图 6　视频检测识别的人流密集点热力分布图

在视频分析的基础上，可判定上下学高峰时期小学生以及接送学生的家长人流密集程度极高，人车冲突加剧，交通安全风险上升。上下学高峰时段道路两侧均被路边停车占用，严重影响兴越路道路通行能力。非机动车由于缺乏专属路权，高峰时期机非混行，也存在交通安全隐患。由此可以得出结论，该节点的主要交通矛盾是上下学高峰期横跨兴越路的人流和兴越路上东西向的车流冲突。因此对其制定的改善方案是彻底人车分离，横跨兴越路设置过街地道，两侧梯道为步行梯道，如图 7 所示。

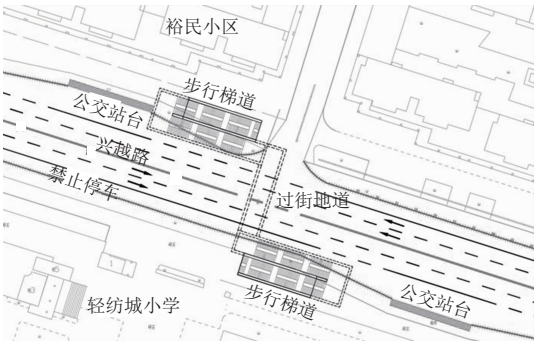


图 7　基于视频检测识别结论制定的某小学人行过街设施方案

## 4　结　语

本文介绍了基于改进 YOLO 算法的行人检测和慢行规划应用案例。通过深度学习环境搭载，行人数据集的重新标注、训练和测试，成功地提高了密集场景下行人检测模型的精度和效率。此外，应用视频检测识别结论制定的某小学人行过街设施方案，对于行人检测和慢行规划领域的研究也具有一定的参考价值。

**参考文献：**

[1] LOWE D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[2] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C]//The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Diego, CA, USA: [s.n.], 2005: 886-893.

[3] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA: [s.n.], 2014: 580-587.

[4] GIRSHICK R. Fast R-CNN [C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: [s.n.], 2015:1440-1448.

[5] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster RCNN: Towards real -time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.

[6] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector [C]//European Conference on Computer Vision.Cham: Springer, 2016: 21-37.

[7] 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 (CVPR). Las Vegas, NV, USA: [s.n.], 2016: 779-788.