DOI:10.19850/j.cnki.2096-4706.2022.01.024

基于 Yolov5 和 DeepSort 的视频行人识别与跟踪探究

张梦华

（安徽理工大学计算机科学与工程学院，安徽 淮南 232001）

摘　要：视频监控在信息化时代尤其是交通系统中占据重要地位，文章提出一种基于 Yolov5 和 DeepSort 在可见光环境下将行人识别和行人跟踪两大模块相结合的多目标跨镜头跟踪算法。首先使用 Yolov5 算法通过保存视频号、行人序号和位置信息给视频中行人赋予标签，得到视频中所有行人的信息；然后根据信息用 DeepSort 实现行人跟踪。经过测试和训练可以快速准确地完成任务，有一定的理论探索意义和实用价值。

关键词：Yolov5；DeepSort；行人识别；行人跟踪

中图分类号：TP391.4　　**文献标识码：**A　　**文章编号：**2096-4706（2022）01-0089-04

Exploration of Video Pedestrian Recognition and Tracking Based on Yolov5 and DeepSort

ZHANG Menghua

(College of Computer Science and Engineering, Anhui University of Science & Technology, Huainan 232001, China)

Abstract: Video surveillance plays an important role in the informatization age, especially in traffic system. This paper proposes a multi-target cross-shot tracking algorithm, which combines two modules of pedestrian recognition and pedestrian tracking in the visible light environment based on Yolov5 and DeepSort. Firstly, Yolov5 algorithm is used to label the pedestrian in the video by saving the video number, pedestrian serial number and location information, and obtain the information of all pedestrians in the video. Then, according to the information, DeepSort is used to achieve pedestrian tracking. After testing and training, it can complete the task quickly and accurately, which has a certain theoretical exploration significance and practical value.

Keywords: Yolov5; DeepSort; pedestrian recognition; pedestrian tracking

0 引　言

计算机视觉中的目标检测是较早开始的研究方向，在智能视频监控、工业检测、航空航天等诸多领域上经过几十年的不断探索后取得了显著的发展。其中智能视频监控中的行人检测是通过计算机视觉中的方法来获取图像或视频中行人的位置。由于行人刚柔两方面的特性，穿戴、比例、遮掩物、行为等都会影响检测的准确性，因此研究行人检测变成计算机视觉领域中富有挑战价值的热门课题 [1]。

传统的方法是基于图像上的行人识别和跟踪，只包含空间特征，缺少时序信息，在复杂条件下的精度不高；而在视频序列中两者都包含进去，因此在视频行人识别的研究中有重要意义。

随着大规模视频数据集的出现，研究者设计了多种模型来实现行人识别与行人跟踪。对于行人识别的实现，文献 [2] 运用背景差法把当前图像与背景图像做差判断像素，根据建模获得的近似图像判断跟踪效果。文献 [3] 运用帧差法将邻近的两幅图像做差，二值化后获得目标，因为对噪声的敏感性导致获取的目标不完整。文献 [4] 运用光流法对光流场进行检测分割，可以轻易地检测到目标和获取背景图像，计算量较大。对于行人跟踪的实现，文献 [5] 运用基于特征的跟踪方法在原始图像中提取最明显的特征。SIFT 算法、KLT 算法、Harris 算法和 SURF 算法都有很好的鲁棒性，是典型算法 [6-9]。文献 [10] 运用基于贝叶斯的跟踪方法将行人跟踪转为贝叶斯估计。Kalman 滤波（KF）[11] 可以精准的预测行人下一个时间点的位置，是目前已成熟的方法。

根据已经提出的方法进行改进，本文提出基于 Yolov5 和 DeepSort 的视频行人识别与跟踪，在可见光的环境下实现多目标跨镜头识别与跟踪，有较高的准确性和实时性。

1 Yolov5 实现行人识别

Yolov5 是 Yolov4 工程化的版本，它有更好的灵活性和更快的速度，在模型的快速部署上具有极强优势。相比 Yolov4，该算法有以下优点：

（1）数据增强，通过随机选取训练集中四张图片的中心点，在其四角位置分别放置一张图片，可以增加 batch size。

（2）DropBlock 机制。通过 Dropout 防止过拟合，通过 DropBlock 随机去除神经元。标签平滑，使神经网络减弱。

（3）损失函数：使用 CIoU 进行边框回归；使用 BCEWithLogitsLoss 和 CIoU 进行 Objectness；使用 BCEWithLogitsLoss 进行分类损失。

Yolov5 算法中的四种网络结构 Yolov5s、Yolov5m、Yolov5l 和 Yolov5x 在原理和内容上基本一样，但在宽度和深度上不同。网络深度通过 depth_multiple 参数控制，网络宽度通过 width_multiple 参数控制。CSP1 和 CSP2 是 Yolov5 的两种 CSP 结构，Backbone 主干网络储存 CSP1，Neck 网络储存 CSP2，四种网络中每个 CSP 结构的深度都不相同，且随着网

收稿日期：2021-12-06

络层数的加深网络的特征提取和融合能力也不断升高。网络宽度中特征图第三维度受卷积核数影响，核数越多，特征图越宽，网络提取特征能力越强。各部分具有的主要功能结构为：

输入端：Mosaic 数据增强、自适应锚框计算，以及自适应图片缩放。

主干网络：Focus 结构、CSP 结构。

Neck 网络：FPN+PAN 结构。

输出端：GIOU_Loss。

1.1　输入端

1.1.1　Mosaic 数据增强

在输入端选择 Mosaic 数据增强方式，首先可以增加数据集的复杂度，其次可以减少 GPU 的内存使用。数据集的复杂性体现在对多张图片进行随机裁剪缩放，提高训练后的精度。由于训练的图片数量不需要设置的非常大，因此可以减少 GPU 的内存使用。

1.1.2　自适应锚框计算

在 Yolov5 算法中，所有视频中的行人都使用默认的标签框距，训练时会在此基础上输出一个预测框，方便将初始框与预测框对比计算差值。

1.1.3　自适应图片缩放

对于数据集中一帧一帧的图片尺寸不同的现象，都会在初始时设置固定的尺寸，在处理完成后可以对其进行缩放裁剪，提高精度。

1.2　主干网络

1.2.1　Focus 结构

在提取视频行人特征的过程中，方便对其进行切片处理，对不同层的特征图有不同的切片选择，最终卷积后形成特征图。

1.2.2　CSP 结构

在视频行人识别中使用 CSP 结构，可以使网络模型轻量化，便于数据集的训练，减少了 GPU 内存的使用，还降低了计算的时间，使效率提高。

1.3　Neck 网络

首先使用自顶向下的 FPN 层可以使语义特征顺利传达下去，通过 PAN 结构可以有效定位特征，使每一个主干层中的检测层完成参数聚合。

1.4　输出端

输出端中的损失函数由分类损失函数（Classificition Loss）和回归损失函数（Bounding Box Regeression Loss）组成。

由初始框与预测框对比，A 为交集，B 为并集，C 为最小外接集合，可以计算差值得到 IOU 的 Loss：

$$IOU\_Loss=1-IOU=1-A/B \tag{1}$$

然后得到 GIOU_Loss 的值：

$$GIOU\_Loss=1-GIOU=1-(IOU-|\text{差集}|/|C|) \tag{2}$$

2　DeepSort 实现行人跟踪

DeepSort 是在 Sort 目标跟踪基础上进行的改进。其优点为：

（1）增加 Deep Association Metric：可以实现行人检测，是在学习卡尔曼滤波和匈牙利算法的基础上改进的。

（2）添加外观信息：通过卡尔曼滤波算法和匈牙利算法对行人进行识别和目标分配，添加外观信息对行人跟踪有更好的效果。

由于存在多目标跟踪中一个目标覆盖多个目标或多个检测器检测一个目标的情况，DeepSort 算法使用八维状态空间（u，v，γ，h，x，y，γ，h）定义跟踪场景。根据算法可知马氏距离计算公式为：

$$d^{(1)}(i,j)=(d_j-y_i)\,TS_i^{-1}\,(d_j-y_i) \tag{3}$$

在设置运动状态关联成功后，可以得到示性函数为：

$$b_{i,j}^{(1)}=1[d^{(1)}(i,j)]\leqslant t^{(1)} \tag{4}$$

由此类推可以得到 $d^{(2)}(i,j)$ 和 $b_{i,j}^{(2)}$，最终得到 2 种度量方式线性加权的度量：

$$C_{i,j}=\lambda d^{(1)}(i,j)+(1-\lambda)\,d^{(2)}(i,j) \tag{5}$$

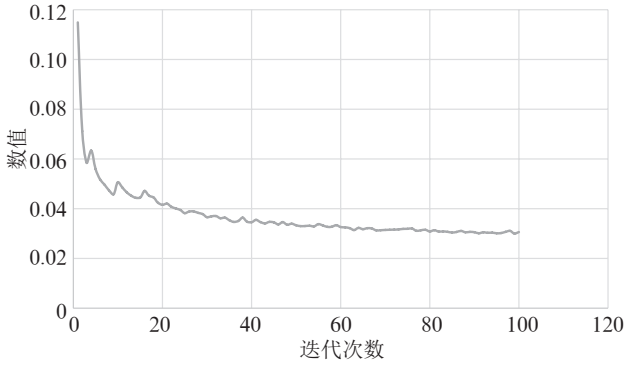
当 $C_{i,j}$ 位于 2 种度量阈值交集内，则认为实现了正确的关联。

为了实现行人跟踪，使用神经网络对视频行人识别数据集训练。通过 DeepSort 算法，在行人特征提取后得到一帧一帧的图像，完成对行人的跟踪。此方法可以有效改善遮挡问题。

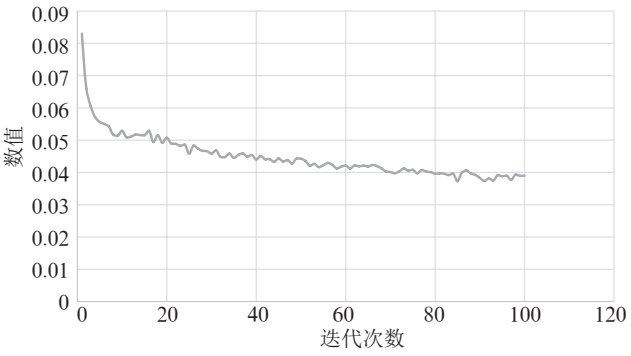
3　实验结果及分析

为了验证 Yolov5 和 DeepSort 对视频中行人识别和跟踪的效果，本文选取了一段交通环境下的行人视频，该视频在 AMD Ryzen 5 4600U with Radeon Graphics 2.10 GHz 处理器、16 GB 内存、Windows 10 操作系统的电脑上完成。

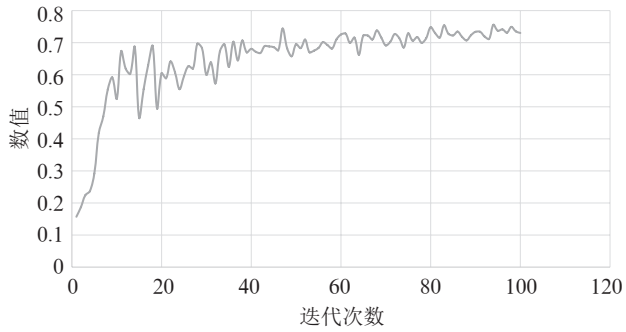
训练过程的各种数值随着迭代次数的增加而变化，本次实验迭代次数 100 次，各种数值的变化如图 1 所示。



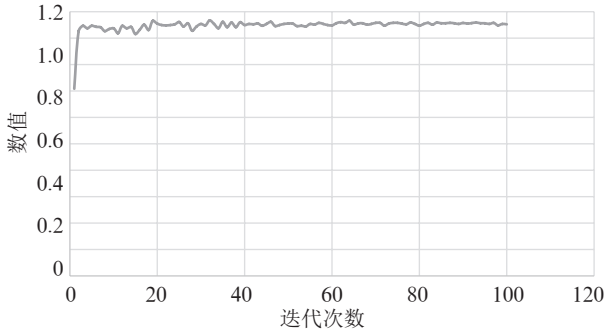
（a）Glou



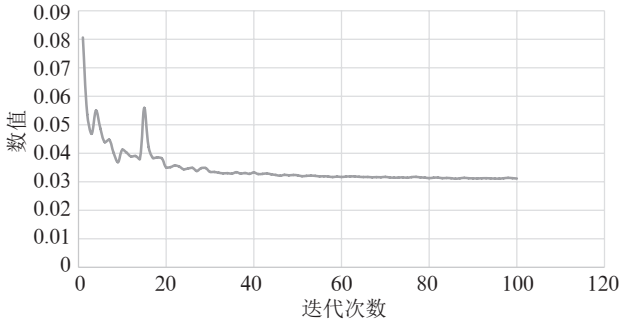
（b）Objectness



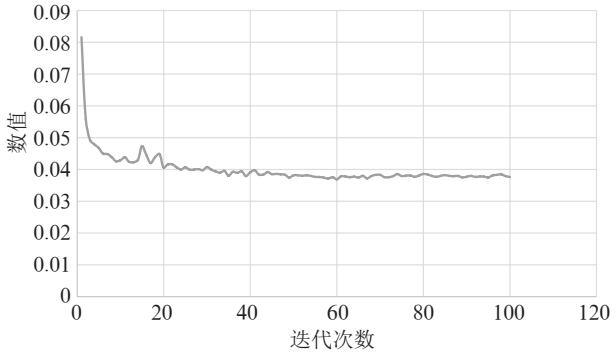
（c）Precision



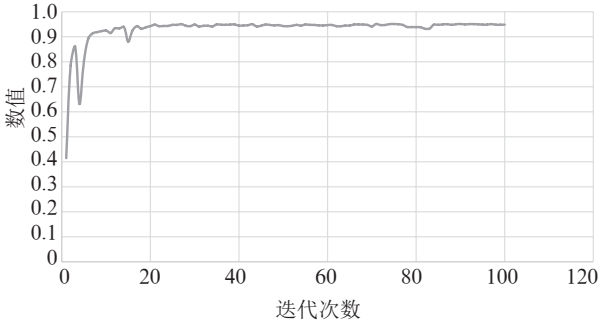
（d）Recall



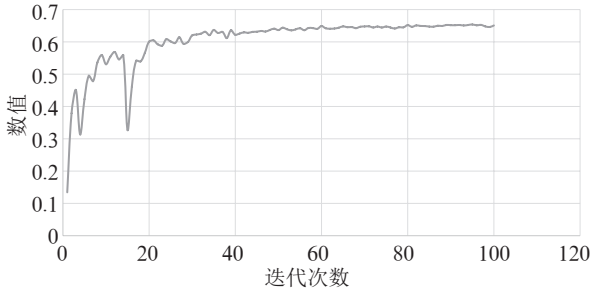
（e）val Glou



（f）val Objectness



（g） mAP@0.5



（h）mAP@0.5：0.95

图 1　训练参数变化图

GIoU 和 val Glou：数值越接近 0，目标框画的越准确。

Objectness 和 val Objectness：数值越接近 0，对行人识别得越准确。

Precision：准确率（标记的正确个数除以标记的总个数）越接近 1 越高。

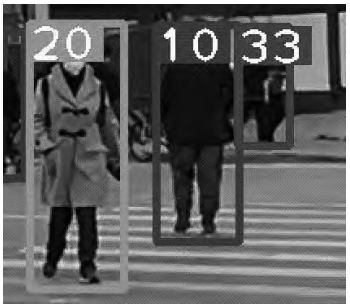
Recall：召回率（标记的正确个数除以需要标记的总个数）越接近 1 越高。

mAP@0.5 和 mAP@0.5：0.95：AP （以 Precision 和 Recall 为坐标轴作图围成的面积）越接近 1，准确率越高。

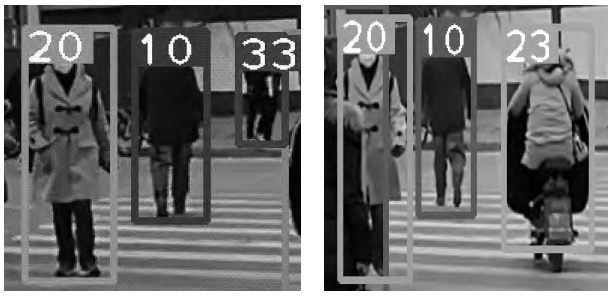
从图 1 可以看出，训练迭代次数越接近 100，各项数值变化越趋于平稳。

为了验证视频中行人的识别与跟踪效果，这里随机截取了几帧行人图片，如图 2 所示。

从图中可以看到，本次截取了第 80 帧，第 97 帧和第 115 帧的图片，可以清楚地看到视频中序号为 10，20，23 和 33 的行人被 label 标签准确的框起来，并且实现了对序号为 10 的行人和序号为 20 的行人的跟踪，从图 2 中可以准确地看到运动轨迹。使用 Yolov5 算法保存视频号、行人序号和位置信息给视频中行人赋予了标签，得到了视频中所有行人的信息，实现行人识别。然后根据行人特征信息用 DeepSort 算法实现了行人跟踪。经过测试和训练后快速准确的完成了行人识别与跟踪任务。



（a）第 80 帧



（b）第 97 帧　　（c）第 115 帧

图 2　行人识别与跟踪样例图

4　结　论

由于 Yolov5 在目标检测上有更好的灵活性和更快的速度，DeepSort 在目标跟踪过程中可以改善有遮挡情况下的目标追踪效果，减少了目标 ID 跳变的问题，本文将两者相结合，实现视频行人识别与跟踪。实验结果表明，结合后的 Yolov5 和 DeepSort 可以快速有效地实现行人识别与跟踪。但是，在行人有重叠或被遮挡的情况下不能准确的识别出来，还需进一步的改进。

参考文献：

[1] 宋艳艳，谭励，马子豪，等 . 改进 YOLOV3 算法的视频目标检测 [J]. 计算机科学与探索，2021，15（1）：163-172.

[2] 张咏，李太君，李枚芳 . 利用改进的背景差法进行运动目标检测 [J]. 现代电子技术，2012，35（8）：74-77.

[3] 杨阳，唐慧明 . 基于视频的行人车辆检测与分类 [J]. 计算机工程，2014，40（11）：135-138.

[4] SUN S J，HAYNOR D，KIM Y M. Motion estimation based on optical flow with adaptive gradients [C]//Proceedings 2000 International Conference on Image Processing （Cat. No.00CH37101）. Vancouver：IEEE，2002：852-855.

[5] 王亮，胡卫明，谭铁牛 . 人运动的视觉分析综述 [J]. 计算机学报，2002（3）：225-237.

[6] 侯跃恩，李伟光 . 时间连续贝叶斯分类目标跟踪算法 [J]. 计算机工程与设计，2016，37（8）：2125-2131.

[7] DAVID G L. Distinctive Image Features from Scale-Invariant Keypoints [J].International Journal of Computer Vision，2004，60（2）：91-110.

[8] 杨陈晨，顾国华，钱惟贤，等 . 基于 Harris 角点的 KLT 跟踪红外图像配准的硬件实现 [J]. 红外技术，2013，35（10）：632-637.

[9] HARRIS C，STEPHENS M. A Combined Corner and Edge Detector [C]//Proceedings of the 4th Alvey Vision Conference. Manchester：Alvety Vision Club，1988：147-151.

[10] KASHIF M，DESERNO T M，HAAK D. Feature description with SIFT，SURF，BRIEF，BRISK，or FREAK? A general question answered for bone age assessment [J].Computers in Biology and Medicine，2016，68（C）：67-75.

[11] 梁锡宁，杨刚，余学才，等 . 一种动态模板匹配的卡尔曼滤波跟踪方法 [J]. 光电工程，2010，37（10）：29-33.

作者简介：张梦华（1996—），女，汉族，山西临汾人，硕士在读，研究方向：计算机视觉。

（上接 88 页）

（a）中性　　（b）高兴

（c）高兴

图 3　智慧课堂真实环境中的应用效果

考虑到智慧课堂中的表情分布不平衡这一问题，比如厌恶、悲伤、恐惧之类的表情较少。未来，笔者将尝试进一步扩大数据库多样性和规模，同时拓展对反映学生学习状态的微表情识别的研究。

参考文献：

[1]K ELTNER D，EKMAN P. Facial Expression of Emotion [M]. Handbook ofEmotions 3rd. New York：The Guilford Press，2010：173-183.

[2] Lancet T. Communication without Words[J]. University of East London，1968，24（23）：1084-5.

[3] 魏为民，孟繁星等 . 人脸表情识别综述 [J]. 上海电力大学学报，2021（12）：597-602.

[4] 景晨凯，宋涛等 . 基于深度卷积神经网络的人脸识别技术综述 [J]. 计算机应用与软件，2018（1）：223-231.

[5] 靳显智，林霏等 . 基于 CNN 的面部表情识别算法 [J]. 齐鲁工业大学学报，2021（6）：64-69.

[6] 程换新，王雪等 . 基于 CNN 和 LSTM 的人脸表情识别模型设计 [J]. 电子测量技术，2021（9）：160-164.

[7] 周丽芳，刘俊林等 . 深度二值卷积网络的人脸表情识别方法 [J]. 计算机辅助设计与图形学学报，2022（1）：1-12.

[8] 陈汤慧，高美凤 . 基于 ME-Xception 卷积神经网络的微表情识别 [J]. 信号处理，2021（12）：1-12.

[9] 王涛，彭欣荣等 . 基于几何特征和 LBP 特征融合的笑脸识别算法的研究 [J]. 电子测试，2021（12）：52-54.

[10] 吕秀丽，黄兆昊等 . 基于改进 LBP 和 DBN 的人脸识别算法研究 [J]. 工业仪表与自动化装置，2021（5）：80-82.

作者简介：刘锦峰（1982—），女，汉族，湖南娄底人，副教授，硕士，主要研究方向：高职教育、智能教育、电子商务。