计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(4):271−276 [doi: 10.15888/j.cnki.csa.007915]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 YOLOv3 增强模型融合的人流密度估计①

孙乾宇, 张振东

(上海理工大学 机械工程学院, 上海 200093)
通讯作者: 孙乾宇, E-mail: 1305489555@qq.com

摘　要: 为了解决在复杂背景以及人流密集且互相遮挡的场景下, 对人流密度进行估计精度低的问题, 提出了基于YOLOv3 增强模型融合的方法进行人流密度估计. 首先将数据集分别进行头部标注和身体标注, 生成头部集和身体集. 然后用这两个数据集分别训练两个 YOLOv3 增强模型 YOLO-body 和 YOLO-head, 最后使用这两个模型在相同的测试数据集上推理, 将其输出结果进行极大值融合. 结果表明基于 YOLOv3 增强模型融合的方法, 与原始目标检测方法和密度图回归的方法相比精度提高了 4%, 且具有较好的鲁棒性.

关键词: YOLOv3; 模型融合; 人流密度估计; 深度学习; 目标检测

引用格式: 孙乾宇,张振东.基于 YOLOv3 增强模型融合的人流密度估计.计算机系统应用,2021,30(4):271–276. http://www.c-s-a.org.cn/1003-3254/7915.html

Crowd Density Estimation Based on YOLOv3 Enhanced Model Fusion

SUN Qian-Yu, ZHANG Zhen-Dong

(School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: The accuracy of crowd density estimation is low in complex backgrounds and the scenario with dense and mutually occluded crowds. To solve this, we propose a method based on YOLOv3 enhanced model fusion to estimate crowd density. The heads and bodies in the data set are labeled to generate head and body sets, which can then help train the two YOLOv3 enhanced models: YOLO-body and YOLO-head. Finally, the two models are reasoned on the same test data set, and their outputs are fused to the maximum value. Consequently, the method based on YOLOv3 enhanced model fusion has great robustness because its accuracy is 4% higher than that of original target detection and density map regression.

Key words: YOLOv3; model fusion; crowd density estimation; deep learning; object detection

随着消费水平的提高, 人们外出旅游, 商场购物已然成为常态, 导致区域人流量急剧增加. 一方面, 在人流密集的场景下容易产生踩踏事件以及打架, 纵火等恶劣行为造成人流恐慌. 在另一方面, 可以帮助商场等各大商铺有效的统计客流量. 因此对特定场合的人流密度进行精确的估计具有重大意义.

现有的人流密度估计方法主要有两种, 一种是针对检测的方法, 另一种是针对回归的方法. 对于检测的方法来说一般假设可以通过使用给定的对象检测器[1−3]来检测和定位人群图像上的每个人, 然后通过累积每个检测到的人来计数, 然而, 这些传统的检测方法[4−6]需要耗费很大的计算资源而且还会受到行人人为遮挡以及背景复杂的限制. 在实际情况下, 精度较低, 鲁棒性较差. 基于回归的方法是给定一张图片, 直接从图片中回归出人口的数量. Chan 等[7]使用手动制作的图像特征来将人数统计任务转变成回归任务; 文献 [8,9] 提出了更多检测人数估计任务相关的特征, 包括针对整体结构的特征和局部纹理的特征; Lempitsky

① 收稿时间: 2020-08-19; 修改时间: 2020-09-25, 2020-10-21; 采用时间: 2020-10-28; csa 在线出版时间: 2021-03-30

等人[10]提出了一种密度图回归的算法, 该算法通过对检测图像的密度图进行积分来统计人群个数. 然而这些基于回归的方法在复杂背景下准确性相对较低.

近年来随着 YOLO[11−13] 系列模型的出现, 以超高的推理速度和较高的精度在各个邻域广泛应用. 然而在人流密集和行人互相遮挡的情况下直接使用目标检测模型精度相对较低. 因此提出了基于 YOLOv3 增强模型融合的人流密度估计方法. 一方面, 使用 YOLOv3 增强模型来提高精度. 另一方面, 使用人流头部标注数据集和人流身体标注数据集分别训练两个模型进行融合来提高鲁棒性. 模型融合的方法在数据集上进行测试, 结果表明具有较高的精度和鲁棒性.

1 模型

1.1 YOLOv3 模型原理

YOLOv3 算法的基本思想可以分成两部分:

首先, 根据一定的规则在图片上生成一系列候选区域, 然后根据这些候选区域与图片上物体的真实区域之间的位置关系对候选区域进行标记. 跟真实框之间的距离小于阈值的那部分候选区域会被标注为正样本, 同时将真实框的位置坐标作为正样本的位置坐标目标值. 距真实框的距离较大的那些候选区域则会被标注为负样本, 负样本不需要预测位置坐标或者类别信息.

其次是使用卷积神经网络提炼出图片的特征, 并对候选区域的位置坐标和类别信息进行预测. 这样, 可以将每个预测框视为一个样本, 并根据真实框相对于其的位置坐标和类别信息来获取标签值. 使用网络模型来预测其位置和类别, 并比较网络预测值和标签值. 这样就可以构建损失函数来进行训练. YOLOv3 算法的思想如图 1 所示.

YOLOv3 采用的骨干网络是 DarkNet-53. DarkNet-53 网络结构没有池化层, 在前向传播过程中, 通过改变卷积核的步长代替池化层, 特征提取模型采用很多 3×3 和 1×1 的卷积层, 再加上全连接层共有 53 层. 在经过 DarkNet-53 特征提取后, 为了提高不同大小物体的检测精度, YOLOv3 在 3 个不同尺度上经行检测, 每个尺度有 3 个界限值 (bounding box), 最后由与真实框的交并比 (IOU) 最大的界限值预测目标. YOLOv3 结构如图 2 所示. 图中 Res*n* 表示一个残差块, 其中含有 *n* 个残差单元; DBL 是 YOLOv3 的基本组件, 表示卷积 (conv)+批归一化 (BN)+激活函数 leaky ReLU.

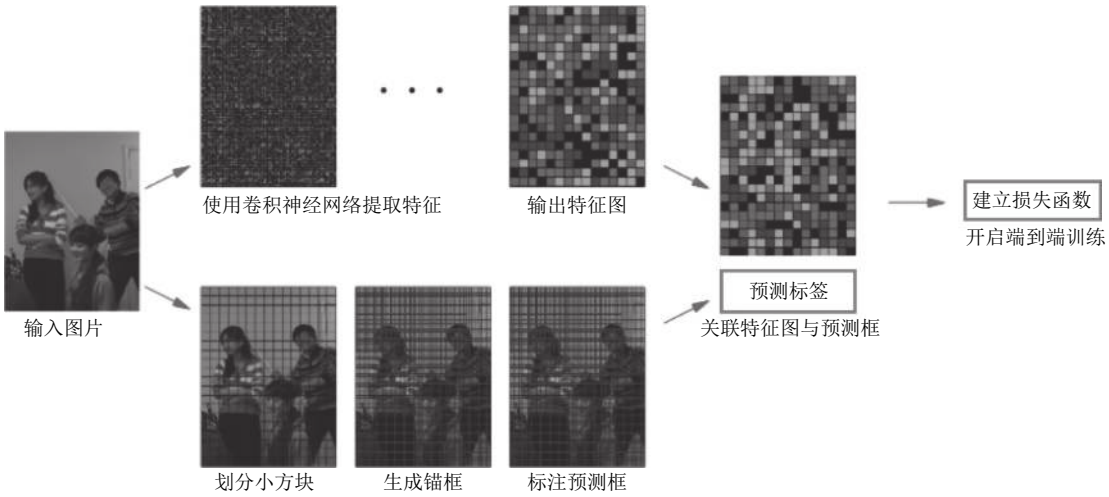

图 1 YOLOv3 算法思想

1.2 YOLOv3 增强模型

为了在保持推理速度的同时最大限度的提升检测精度, YOLOv3 增强模型在原网络的基础下做了如下改进:

(1) 骨干网络采用 ResNet50-vd 替换原有的 DarkNet-53. ResNet-vd 是 ResNet 系列的改进网络, ResNet-vd 的参数量和计算量与 ResNet 几乎一致,但是精度提升了 2%. 虽然 DarkNet-53 也使用了残差网络如图 3, 但是同 ResNet50-vd 相比, ResNet50-vd 在速度和精度上都有一定的优势, 而且选用 ResNet[14] 系列网络更加容易扩展. 可以根据不同的业务需求, 灵活选择 ResNet18、50、101 等不同的网络作为目标检测的骨干网络.

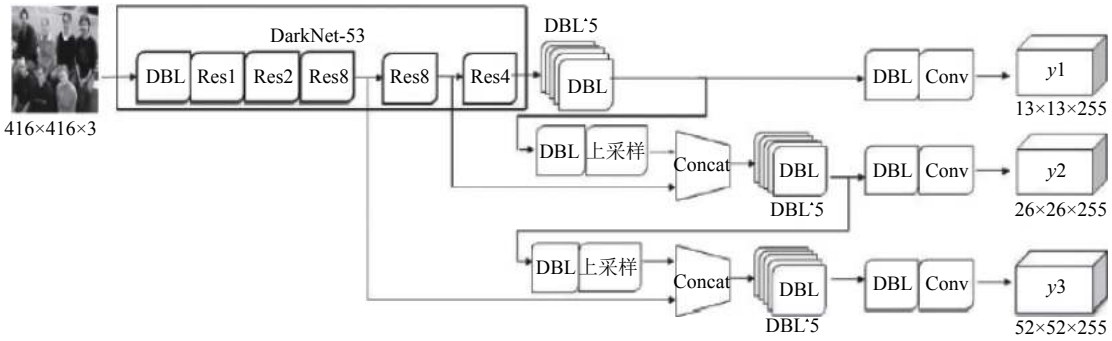

图 2 YOLOv3 网络结构图

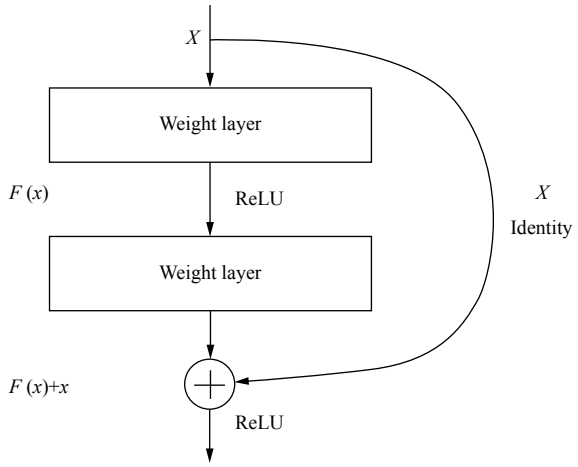

图 3 残差结构

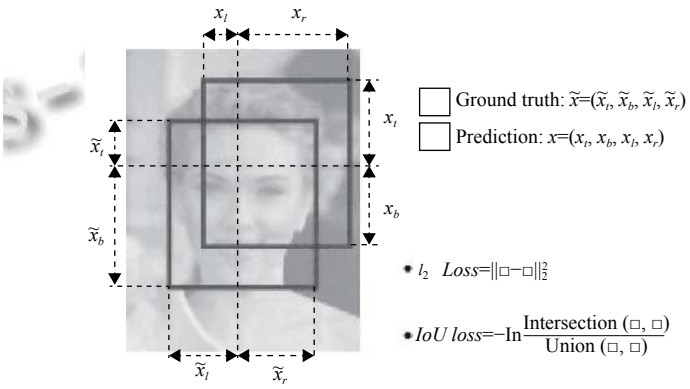

图 4 L2 损失和 IoU 损失

(2) 引入可变形卷积 (Deformable Convolution, DCN)[15], 替代原始卷积操作. 可变形卷积已经在各个邻域的视觉任务中广泛验证过其效果. 在考虑到保持速度与提升精度平衡的前提条件下, YOLOv3 增强模型使用可变形卷积替换了主干网络中第 5 阶段部分的 3×3 卷积.

(3) 由于 YOLOv3 作为单阶段目标检测模型, 在定位精度上相比 Faster RCNN、Cascade RCNN 等两阶段目标检测模型有着其天然的劣势, 所以 YOLOv3 增强模型增加了 IoU 损失[16] 分支, 可以一定程度上提高边界框的定位精度, 缩小单阶段目标检测网络和两阶段检测网络精度的差距.

传统的 L2 损失将检测目标的位置坐标信息当作互相独立的 4 个变量来进行训练. 而 IoU 损失直接使用预测的边界框与基本真实值之间的最大重叠, 并将所有绑定变量作为一个整体进行回归, 将位置坐标信息当作一个整体进行训练. 所以使用 IoU 损失能够获得更加精准的训练效果和检测结果. L2 损失和 IoU 损失说明如图 4 所示.

1.3 改进的 YOLOv3 模型与原模型对比

使用在 Object365 数据集上训练的模型作为预训练模型, 在 COCO 数据集上进行训练和验证, 用 TensorRT 进行部署推理. 不同改进变量的模型验证精度和推理速度如表 1.

表 1 不同改进变量的模型验证精度

模型	预训练数据集	验证集 mAP	预测速度 (ms)
YOLOv3_DarkNet	ImageNet	38.9	42.5
YOLOv3_ResNet50_vdDCN	ImageNet	39.1	35.2
YOLOv3_ResNet50_vdDCN	Object365	42.5	35.2
YOLOv3_ResNet50_vdDCN IOULoss	Object365	43.2	35.2

2 实验过程与模型融合分析

实验环境为 NVIDIA Tesla V100 16 GB 显存 GPU, 用 Tensorflow 搭建的 YOLO 目标检测模型进行训练.

2.1 数据集

为了验证所提出方法的精度和鲁棒性. 实验所用

的训练和测试数据集来自公开数据集 ShanghaiTech. ShanghaiTech 数据集是一个大规模的人群统计数据集, 其包含 1198 幅图像. 为了提高精度, 还添加了公开数据集 UCF-CC-50 的部分数据. 实验原始数据集共包含 2000 张训练图片数据和 1000 张测试图片数据. 将 2000 张训练数据用 LabelImg 数据标注工具分别进行人流头部标注和人流身体标注, 从而构建了两个训练数据集: 头部集和身体集.

为了防止由于训练数据不足而导致模型训练过程发生过拟合, 在数据处理阶段采用了随机反转、移动、改变饱和度、改变亮度、添加噪声等图像增强技术. 图像增强效果图如图 5 所示.

图 5 图像增强效果图

2.2 模型训练

原始的数据集分别经过头部标注和身体标注生成两个训练数据集: 头部集和身体集. 使用这两个数据集分别训练两个 YOLOv3 增强模型: YOLO-head 和 YOLO-body. 其中, 为了提高模型的精度和提升训练速度, YOLO-head 和 YOLO-body 模型都选用了在旷世公开数据集 Object365 上训练好的 YOLOv3 增强模型预训练参数. 模型具体训练参数如表 2.

表 2 训练参数

参数名称	大小
输入尺寸	[3,256,256]
批次大小	8
学习率	0.001
训练轮数	60 000
优化器	Adam
动量参数	0.9
L2正则	0.0005
Iouloss	DiouIoss
Anchor mask	[[6,7,8],[3,4,5],[0,1,2]]
Anchors	[[10,13],[16,30],[33,23]] [30,61],[62,45],[59,119] [116,90],[156,198],[373,326]]

2.3 评估指标

采用平均错误率 (AER) 来对所提出的方法的人流密度估计精度进行评估. 平均错误率如下:

对任意一张测试图像I_i, 设总人数真值为G_i, 预测值为P_i. 则这张图片的评估错误率为:

$$E_i = \frac{|P_i - G_i|}{G_i}$$

所有测试数据平均错误率为:

$$\sum_{i=1}^{N} \frac{E_i}{N}$$

平均错误率越低, 模型预测效果越好.

2.4 模型融合

用头部集和身体集分别训练的两个 YOLOv3 增强模型 YOLO-head 模型和 YOLO-body 模型. 在使用相同的测试数据集测试时发现在人流密集人体互相遮挡的情况下, YOLO-body 模型会漏选而 YOLO-head 模型表现更好, 实验结果如图 6 所示. 图 6(a) 为 YOLO-head 模型检测结果, 图 6(b) 为 YOLO-body 模型检测结果.

在背景复杂以及行人头部为不完全裸露的情况下, YOLO-head 模型会漏选而 YOLO-body 模型表现更好, 实验结果如图 7 所示. 图 7(a) 为 YOLO-head 模型检测结果, 图 7(b) 为 YOLO-body 模型检测结果.

对于以上问题, 采用 YOLO-head 与 YOLO-body 检测结果进行极大值融合的方法如图 8 所示, 即将两个模型对人流密度的估计结果取最大值输出, 从而能够改善由于背景复杂以及行人相互遮挡等场景下单模型漏测的情况. 因此, 能够有效地提高模型对人流密度估计的精度和鲁棒性.

(a) YOLO-head 模型检测结果

(b) YOLO-body 模型检测结果

图 6 人流密集人体互相遮挡的情况下检测结果

(a) YOLO-head 模型检测结果

(b) YOLO-body 模型检测结果

图 7 背景复杂以及行人头部为不完全裸露的情况下检测结果

2.5 实验结果

在实验时, 将模型融合的方法与原始的单模型检测方法和高斯密度图回归方法进行了比较, 结果如表 3, 表明所提出的模型融合的方法具有较高的精度和鲁棒性.

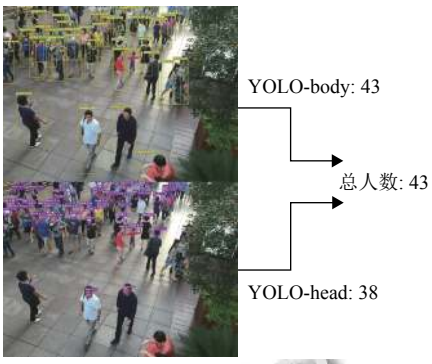

图 8 极大值融合

表 3 不同算法的错误率

方法	平均错误率(AER)
VGG16高斯密度图回归	0.347
MCNN高斯密度图回归	0.174
YOLO-body	0.225
YOLO-head	0.183
模型融合	0.133

3 结束语

提出了一种 YOLOv3 增强模型融合的方法用于人流密度估计, 通过使用 YOLOv3 增强模型来提高检测精度同时保证检测速度. 通过使用不同标注的数据集训练 YOLO-head 和 YOLO-body 模型进行融合的方法来提高精度和鲁棒性. 实验表明所提出的方法有较高的精度和较好的鲁棒性.

参考文献

1 Lin SF, Chen JY, Chao HX. Estimation of number of people in crowded scenes using perspective transformation. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2001, 31(6): 645 –654. [doi: 10.1109/3468.983420]

2 Dalal N, Triggs B. Histograms of oriented gra dients for human detection. CVPR 2005. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA. 2005, 1. 886–893.

3 Wang M, Wang XG. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. CVPR 2011. Providence, RI, USA. 2011. 3401–3408.

4 Ge WN, Collins RT. Marked point processes for crowd counting. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA. 2009. 2913–2920.

5 Idrees H, Soomro K, Shah M. Detecting humans in dense

crowds using locally-consistent scale prior and global occlusion reasoning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(10): 1986–1998. [doi: 10.1109/TPAMI.2015.2396051]

6 Lin Z, Davis LS. Shape-based human detection and segmentation via hierarchical part-template matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(4): 604–618. [doi: 10.1109/TPAMI.2009.204]

7 Chan AB, Liang ZSJ, Vasconcelos N. Privacy preserving crowd monitoring: Counting people without people models or tracking. 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008. Anchorage, AK, USA. 2008. 1–7.

8 Chan AB, Vasconcelos N. Bayesian poisson regression for crowd counting. 2009 IEEE 12th International Conference on Computer Vision. Kyoto, Japan. 2009. 545–551.

9 Chen K, Loy CC, Gong SG, *et al*. Feature mining for localised crowd counting. Proceedings British Machine Vision Conference 2012. Surrey, UK. 2012. 1–3.

10 Lempitsky V, Zisserman A. Learning to count objects in images. Proceedings of the 23rd International Conference on Neural Information Processing Systems. Red Hook, NY, USA. 2010. 1324–1332.

11 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 779–788.

12 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. 2017. 7263–7271.

13 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv preprint arXiv: 1804.02767v1, 2018.

14 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. 2016. 770–778.

15 Dai JF, Qi HZ, Xiong YW, *et al*. Deformable convolutional networks. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy. 2017. 764–773.

16 Yu JH, Jiang YM, Wang ZY, *et al*. Unitbox: An advanced object detection network. Proceedings of the 24th ACM International Conference on Multimedia. New York, NY, USA. 2016. 516–520.