计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(6):165−170 [doi: 10.15888/j.cnki.csa.006362]

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

联合语义的深度学习行人检测①

邓　炜, 刘秉瀚

(福州大学 数学与计算机科学学院, 福州 350108)
通讯作者: 邓　炜, E-mail: dengwei92@foxmail.com

摘　要: 视频行人检测是计算机视觉的一个重要应用, 本文利用深度学习检测近似垂直视角的行人, 但若单纯检测行人, 易受与行人语义相关的行人附属属性 (如背包和帽子) 的干扰, 容易造成误检. 本文提出一种基于更快区域卷积神经网络的联合语义行人检测方法: 首先调整网络模型, 增强对小目标的辨别力, 使其可以有效的检测行人和行人的语义属性; 然后利用空间关系建立行人及其语义属性的关联, 合并行人与其语义信息, 并对候选行人目标进行自适应得分调整, 结合行人语义属性判断候选行人目标. 大量的实验表明, 本文的方法精度高, 速度快, 具有实用价值, 且检出的行人与其语义属性还可用于后续的人数统计和行人行为分析.

关键词: 行人检测; 深度学习; 语义属性; 卷积神经网络

引用格式: 邓炜,刘秉瀚.联合语义的深度学习行人检测.计算机系统应用,2018,27(6):165–170. http://www.c-s-a.org.cn/1003-3254/6362.html

Deep Learning-Based Pedestrian Detection Combined with Semantics

DENG Wei, LIU Bing-Han

(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China)

Abstract: Pedestrian detection is an important application of computer vision. However, it mostly uses the methods of low-level features. Deep learning, by combining the low-level features of pedestrians, can get more abstract representation of high-level features which makes the detection more robust. In this study, we propose a Faster Region-based Convolutional Neural Networks (RCNN)-based pedestrian detection method in which semantics is jointly considered. Firstly, we modify and fine-tune the Faster RCNN for fitting in the pedestrian dataset and for making it more capable of detecting small objects. Secondly, we establish connections between the pedestrian and its semantic attributes by spatial relationship, then fuse the pedestrian and its semantic attributes, and meanwhile adaptively adjust the confidence of the target pedestrian. The adaptive adjustment strategy, based on the connections between the pedestrian and its semantic attributes, realizes the fusion of the individual information. Extensive experiments and comparison show that the proposed approach in this study is of high accuracy, acceptable speed, and practical value. What is more, the semantic attributes can be used to count people or analyze the pedestrian's behavior.

Key words: pedestrian detection; deep learning; semantic attributes; convolutional neural network

1　引言

行人检测按照视频视角可以分为直立行人检测和俯视行人检测. 行人检测常被用于基于目标检测的人数统计、行人行为分析和视频语义理解等, 在公共安防、自动驾驶、人群引流、场所规划等领域具有重要意义. 行人检测的一般方法有两步: 第一步提取行人特征, 第二步根据特征进行行人定位和判别. 2005 年 Dalal[1]提出用方向梯度直方图特征 (Histogram of

① 基金项目: 国家自然科学基金 (61473330)
收稿时间: 2017-09-15; 修改时间: 2017-10-10; 采用时间: 2017-10-17; csa 在线出版时间: 2018-05-28

Oriented Gradient, HOG) 和支持向量机 (Support Vector Machine, SVM) 分类器检测直立行人[2], 效果远超之前的方法, 再次掀起了行人检测的研究热潮. 然而多数行人检测的方法依然利用图像纹理特征、像素统计特征或人的形态特征等底层特征[3,4].

近年来, 深度学习方法在目标检测方面取得了很大成功, 深度学习可以组合行人的底层特征, 得到更抽象的高层特征表示, 检测更具鲁棒性. 其中基于区域卷积神经网络 (Region-Based Convolutional Neural Network, RCNN) 的目标检测具有更好的表现和更快的处理速度[5–9]. 基于深度学习的直立行人检测研究较多, 如文献[10–13]. 而通过深度学习检测俯视行人的研究则很少. VuT[14]通过对象之间的上下文关系, 在基于本地模型区域卷积神经网络的基础上, 提出一个全局的卷积神经网络模型来检测人头的位置, 并用一种对象成对模型来联合训练, 在人数不多的生活场景中表现较好, 但模型复杂, 处理速度较慢. Stewart[15]等人将图像解码成一组人物模型, 并直接输出一组检测假设, 改进了行人目标在拥挤场景下的检测效果, 但是如果序列中第一个目标有遮挡模糊等情况, 可能会影响后续目标的检测.

基于深度学习的行人检测效果还有待于提升, 这是因为行人检测面临的情况复杂多样: (1) 在拥挤人群等复杂环境中很难准确分离行人个体; (2) 行人目标姿态各异、可大可小、或远或近; (3) 行人穿戴的服饰或携带的东西对行人目标的干扰. 这些问题都可能导致行人目标的误检或漏检.

视频中行人的语义属性指与行人语义上关联的行人附属属性[16], 例如行人的帽子、包等. 在行人检测的应用场景中, 人不是人们感兴趣的唯一目标, 人的语义属性同样具有意义. 随着计算机视觉研究的深入, 图像中目标语义关系的挖掘、视觉关系的提取也越来越受关注[17]. 行人语义属性也是影响行人检测的因素之一. 因此, 针对上述行人检测所存在的问题, 本文提出融合行人语义的深度学习俯视行人检测, 同时检测行人和行人的语义属性, 利用行人的语义属性来辅助检测行人, 抑制行人语义属性对行人的干扰, 增加检测精度.

2 深度学习

深度学习简而言之就是多层神经网络, 典型的深度网络有卷积神经网络、递归神经网络、深度置信网络和生成对抗网络等.

Faster RCNN[7]是一种基于卷积神经网络的目标检测模型, 它抛弃了基于区域提名的卷积神经网络一贯的选择性搜索 (selective search)[5,6], 首次提出了区域提议网络 (Region Proposal Network, RPN), 使得区域提名、分类、回归一起共享卷积特征, 网络速度加快. Faster RCNN 实质上是 RPN 和 Fast RCNN[6]的结合, RPN 和 Fast RCNN 共享卷积层, 先由 RPN 提取候选区域, 再把候选区域送到 Fast RCNN 中进行目标识别. 它主要有四个步骤. 第一步是特征提取: 输入整张图片, 通过卷积神经网络提取特征图; 第二步是区域提名: 在第一步得到的特征图上进行区域提名; 第三步是分类与回归: 对每个提名的区域进行目标或非目标的二分类, 用回归模型微调候选框位置和大小; 第四步是目标识别: 选取得分高的候选区域进行目标识别.

Faster RCNN 需要对大量候选区域先判断是否是目标, 然后再进行目标识别, 分成了两步, 这点不如不需要区域提名的端到端检测方法如 YOLO[8]和 SSD[9]. 但是 YOLO 使用 S×S 的分割策略, 如果两个目标落入同一个格子也只能识别出一个目标, 而 SSD 中定义的 Default Box 的形状以及网格大小是事先固定的, 对特定的小目标提取不够好, Faster RCNN 则更加灵活, 而且 Faster RCNN 对设备的要求不高, 更具有实际使用条件. 基于 Faster RCNN 目标检测的优势本文选择 Faster RCNN 模型设计俯视行人目标检测器.

Faster RCNN 目标检测已被证明具有很好的鲁棒性, 但对小目标检测效果却不够理想. 在大多数公共场所出入口的近垂直视角的监控视频中, 行人的尺度变化较小, 但场景中会存在很多小尺度的物件. 本文改进了 Faster RCNN 目标检测器对小目标的辨别力, 并针对行人穿戴的服饰或携带的东西的干扰, 引入行人的语义属性, 把行人的语义属性和行人联合训练, 通过目标检测器分类再聚合, 之后进行行人的辨识. 本文所提出的鲁棒行人检测方法可分为两个步骤: (1) 行人与行人语义属性目标检测 (检测出候选行人目标及若干个行人语义属性); (2) 行人与语义属性聚合 (基于空间信息建立行人与其语义属性的联系, 合并行人与其语义属性, 对候选行人目标自适应的奖励得分, 融合检测框).

3 基于联合语义的行人检测

3.1 基于 Faster RCNN 的目标检测

在基于深度学习的行人检测中, 常把行人身体的显著特征作为感兴趣检测目标. 俯视的行人检测的场景多为近似垂直视角监控视频, 行人的头部和肩部是最显著特征. 本文以行人的头部和肩部作为候选行人

检测目标, 而把帽子、提包和背包等易造成混淆的行人语义属性, 作为辅助检测目标.

在 Faster RCNN 中, 图片在输入网络后, 依次经过若干卷积层和池化层的特征提取后, 得到一个高维的特征图. 然后把这个特征图送到 RPN 网络中, 进行候选区域提名. RPN 网络使用滑动窗口策略. 输入到 RPN 的特征图, 被划分成 $n\times n$ 个矩形窗口 (滑动窗口), 把每个矩形窗口的中心点当成一个基准点, 围绕这个基准点选取 k 个不同尺度、不同长宽比的矩形框 (Anchor) 的对应区域作为候选区域 (如图 1 中的虚线矩形框). 文献[7]取 n=3, 并定义 3 种基准尺度框: 128×128、256×256、512×512, 对每种基准尺度框进行 3 种长宽比率变倍 (1:1、1:2、2:1), 这样就得到有 9 个 Anchor(k=9). 然后把候选区域送到两个全连接层: 分类层和窗口回归层, 进行目标或非目标的判别和矩形窗口位置的微调. 最后选取得分最高的前 300 个候选区域到后续的 ROI Pooling 层和全连接层中进行目标分类. 由于在高维的特征层, 有效感受野很大, 文献[7]中的 Anchor 可以感知一个很大范围的目标. 但是小尺度的目标在这个特征层上的特征不明显, 易导致漏检.

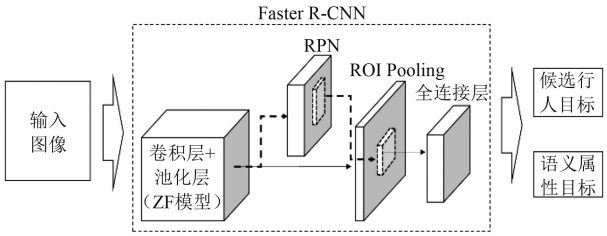


图 1 基于 Faster RCNN 的目标检测器

本文基于 Faster RCNN 的目标检测模型 (如图 1), 结合本文行人检测的特点: 近似垂直视角、多小尺度目标、行人语义属性联合训练和特定的行人数据集, 对 Faster RCNN 做了适当调整. 在行人数据集中, 行人及其与语义属性的大小大部分集中在 60×60 到 300×300 像素之间, 为了在不同尺度上检测行人, 本文使用 64×64、128×128 和 256×256 作为 Anchor 的基准尺度框, 以此增加对小目标的鲁棒.

在得到输入 RPN 的特征图前, 有若干个卷积层和池化层, 这里可以使用预训练的图像分类网络, 比如在 ImageNet 数据集上训练好的 VGG-16[18]和 ZF-net[19], 来初始化网络的参数 (权重和偏移值). VGG-16 是一个很深的网络, 有 13 个卷积层和 3 个全连接层, 速度慢, 精度较高. 而行人检测对实时性的要求高, 所以本文选择了 ZF-net 作为预训练的模型, ZF-net 只有 5 个卷积层和 3 个全连接层, 速度快, 深度虽不如 VGG-16, 但本文用行人的语义属性辅助检测, 弥补了精度的差距. 此外, 较浅的 ZF-net 的特征层维度比 VGG-16 低, 因此小目标的特征在 ZF-net 中会更明显.

行人检测的场景多为监控摄像头捕捉的视频图像, 因此通常距离行人较远, 不存在很大的目标, 因此本文改进了 Faster RCNN 对小目标的辨别力, 虽然大目标的检测会受到影响, 但在实际应用中, 多为中小目标, 所以影响非常小.

本文根据行人数据集的规模和大量实验的训练效果, 调整了网络的学习率、迭代次数和批次大小等. 使得网络模型有效拟合了行人数据, 并避免了用小样本的俯视行人数据集训练时出现过拟合.

目标检测器最终输出候选的行人目标框 p 和语义属性目标框 a. p 和 a 都带有一个目标得分 $score$. 本文在全连接层先对候选目标做一次粗筛选, 把大量的 $a.score<\mu_1$及$p.score<\mu_2$的低分无意义目标去除, 以便于加速后续的处理. 阈值 μ_1 的设置要保证语义属性 a 高置信低误检以辅助候选行人 p 的判断. μ_2 的设置要保证高敏感以避免漏检行人 p.

3.2 行人与语义属性聚合

检出候选目标之后, 需进行目标聚合. 融合行人及其语义属性, 先要建立行人候选目标与其语义属性的联系. 实际场景中的行人及其语义属性最显而易见的联系体现在空间距离, 且语义属性与行人属于单一的从属关系, 所以本文直接用距离贪心的策略建立行人和语义属性的联系: 计算语义属性与行人的重叠率, 重叠率大于 0 视为有联系, 如果语义属性与多个行人都有联系, 则计算该语义属性中心点到各行人中心点的距离, 将其划归于最近的行人. 为体现语义属性属于行人的概率大小, 本文重叠率定义为重叠区域的面积除以语义属性目标框面积.

设: 候选目标框左上角和右下角坐标为: (x_1, y_1, x_2, y_2), 每个候选目标框对应一个检测得分为 $score$. 一帧图像中检测出行人候选框集合为: P=$\{p_1, p_2,\cdots,p_n\}$, 行人语义属性框集合为: A=$\{a_1, a_2,\cdots, a_m\}$. 则按式 (1) 和 (2) 求重叠率为 o 与距离 d:

$$o=\begin{cases}\frac{w\cdot h}{a.area}, & \text{if } w>0 \text{ and } h>0\\ 0, & \text{else}\end{cases} \tag{1}$$

$$d=\frac{\sqrt{(p.x_1+p.x_2-a.x_1-a.x_2)^2+(p.y_1+p.y_2-a.y_1-a.y_2)^2}}{2} \tag{2}$$

其中, $w=\min(a.x_2,p.x_2)-\max(a.x_1,p.x_1)$, $h=\min(a.y_2,$

$p.y_2)-\max(a.y_1,p.y_1),a.area=(a.x_2-a.x_1)\cdot(a.y_2-a.y_1)$.

由于附属物只会出现在行人的附近, 所以当检测到语义属性时, 语义属性附近有行人的概率加大. 基于此, 本文对聚合了 a 的行人候选框 p 进行自适应加分奖励, 奖励原则为: 与 a 距离近、a 的检测分值高, 则奖励力度大, 见式 (3). 由于 sigmoid 函数具有平滑渐进、在零点附近的导数高值性的特点, 本文选择 sigmoid 函数归一化, 并按式 (4) 进行自适应加分.

$$\omega=\frac{a.score}{d} \tag{3}$$

$$p.score=\min\left(1,p.score+\alpha\cdot\frac{1}{1+e^{-\beta(\omega-\tau)}}\right) \tag{4}$$

其中, β、τ 为 sigmoid 范围调整参数, α为最高加分阈值参数.

本文行人与语义属性聚合的具体步骤如下:

输入: 行人集合 P 中的所有 p_i, i=1,⋯,n; 语义属性集合 A 中的所有 a_j, j=1,⋯,m.
输出: 联合语义属性的行人集和遗留语义属性集.

1) 把每个行人框 (p_i, i=1,⋯,n) 初始化为单元素集合 pa_i, 然后置 j=1;
2) 对语义属性 a_j, 按式 (1)、(2) 计算其与 P 集中所有行人的重叠率和距离, 根据最小距离找出其所归属的行人, 不失一般性, 设归属行人框为 p_k;
3) 将 a_j 移出 A 集加入 pa_k 集, 并按式 (4) 调整 $p_k.score$;
4) j=j+1, 若 $j\leqslant m$ 转 2); 否则结束.

合并后的结果 (联合语义的行人集) 如图 2. 经过融合和得分奖励之后, 图像中检出目标为: 联合语义的行人、行人和未与行人合并而独自存在的语义属性. 本文不去除孤立的语义属性目标框, 因为语义属性目标在行人检测的应用场景下, 可能是遗失物品, 可能是危险品, 可以给人提供信息, 具有应用价值.

4 实验分析

本文方法检出行人目标之后, 可进行帧间目标跟踪, 从而得到更精确的检测结果. 由于篇幅限制跟踪细节无法展开, 因此, 实验仅用$p_k.score<\mu_3$过滤低分行人目标.

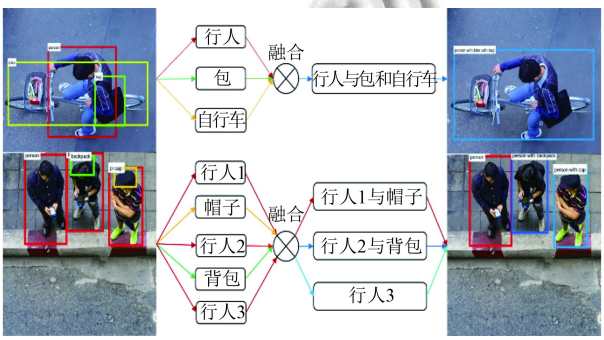


图 2 行人与语义信息聚合

本文用一个近似垂直视角的行人数据集来评估本文方法的性能. 数据集中的图像和视频采集自多个不同的场景. 包括 750 张的照片和多段视频共 20 分钟. 我们从视频中选取 2000 帧图像用来训练, 剩下的用来测试. 照片和视频没有统一的分辨率, 但长边小于等于 1000 像素. 本文定义行人的语义属性包括: 帽子、手提包 (袋子)、背包、箱子、行李箱、自行车、购物车和婴儿车. 表 1 是数据集详细信息.

表 1 训练数据集

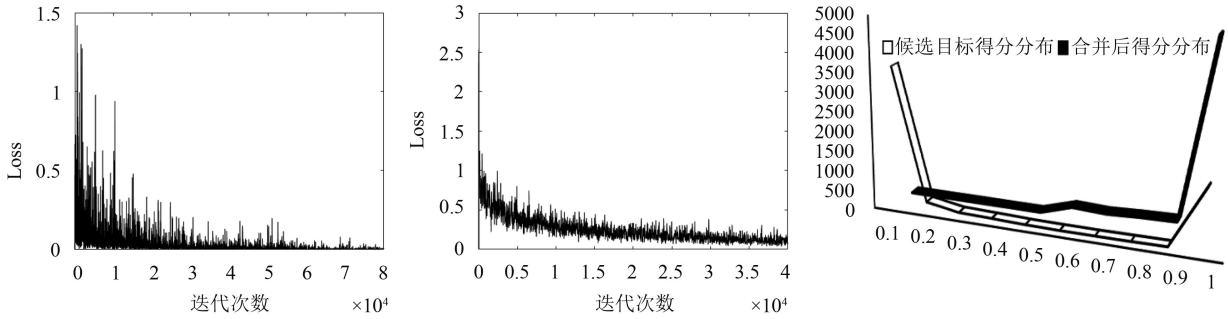
训练集 (张)	行人	语义属性								
		帽子	自行车	购物车	提包	背包	箱子	行李箱	婴儿车	总数
2750	6753	285	150	117	733	1120	63	60	68	2596

为了补充数据和平衡各类语义属性的数目, 我们挑选了部分训练数据进行数据增广. 第一, 对选取的训练图像按倍率 0.6、1 和 1.2 进行缩放; 第二, 旋转 90 度; 第三, 镜像翻转; 第四, 在图像上加上高斯噪声. 经过这样处理,挑选出来的每一张图像都能得到额外的 23 张图像.

本文在 Windows7 系统中使用 GTX1050Ti4G 的 GPU, 在 Caffe 框架完成实验.

我们通过反复实验调整了网络模型的参数, 使网络模型拟合数据, 最终确定网络参数: 学习率: 0.005, 优化算法: 随机梯度下降 (Stochastic Gradient Descent , SGD), 梯度更新的权重 (momentum): 0.9, 权重衰减 (weight decay): 0.0005, 批次大小 (batch size): 128. 最终网络模型的 loss 如图 3(a) 和图 3(b).

我们同时做了大量的测试来确定实验中用到的 3 个阈值 μ_1、μ_2 和 μ_3 的取值, 如图 3(c) 是 1000 张图像中累计的候选目标的得分分布. μ_1 是语义属性的筛选阈值, 由于只对语义属性做一次筛选, 且需用语义属性得分自适应调整行人的得分, 综合考虑了 μ_1 取值实验的结果, 本文取 μ_1 =0.7. μ_2 对行人做粗筛选, 可以让尚有争议的候选目标通过, 本文取 μ_2 =0.5. μ_3 是对最终行人目标的筛选, 所以 μ_3 可以根据人数统计的场景和训练数据集手动调整, 本文取 μ_3=0.8.



(a) Fast R-CNN部分的loss曲线 (b) RPN部分的loss曲线 (c) 抽样图像中的候选目标得分分布

图 3 参数取值实验

式 (4) 对行人的得分进行了自适应的调整. 其中, 从式 (3) 可得$\omega < 1$, 我们取$\beta = 10$, $\tau = 0.5$得ω趋于 0 时, $1/\left(1+e^{-\beta(\omega-\tau)}\right)$的值趋于 0, ω趋于 1 时, $1/1+e^{-\beta(\omega-\tau)}$的值趋于 1. 为了让加分不超过 p 自身的得分, 我们取$\alpha = \mu_2$.

为了展示实验结果和评估本文的方法, 我们对比了文献[14]、文献[15]和原始 Faster RCNN[7]的方法. 实验结果的对比见图 4、图 5 和表 2. 其中部分方法要求把图像大小缩放到一定尺度.

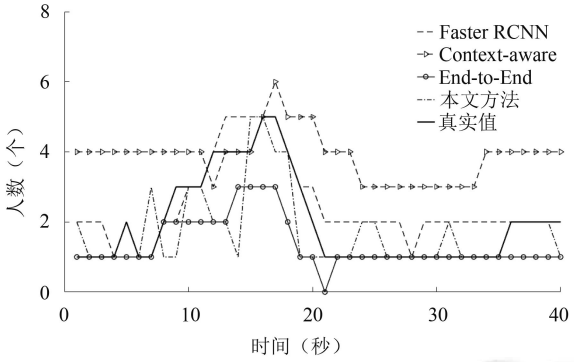


图 4 视频场景中检测到的行人数目对比

(a) 真实值: 6 (b) 本文方法: 6

(c) Faster RCNN[7]:5 (d) Context-aware[14]:5 (e) End-to-End[15]:7

图 5 其中一帧检测到的人数对比

本文使用常规指标来评估我们的方法: 均方误差和平均绝对误差. 均方误差: $mse = E\left(\left(k_j - k_j'\right)^2\right)$, 平均绝对误差: $mae = E\left(\left|k_j - k_j'\right|\right)$, 其中$k_j$是第 j 帧图像中的实际行人数目, k_j'是检测到的行人数目.

表 2 实验数据表

方法	*mae*	*mse*	图像大小 (像素)	平均时间 (毫秒/帧)
Context-aware[14]	3.15	7.76	480×480	1598
Faster RCNN[7]	2.15	5.86	长边<1000	175
End-to-end[15]	1.98	4.27	640×480	**113**
本文方法	**1.81**	**4.18**	长边<1000	180

从表 2 可以看出, 在俯视行人检测中, 本文方法准确率比只用 Faster RCNN 的方法更高, 而且耗时增加很少, 对比其他一些方法也有不错的竞争力.

此外, 本文使用在表 1 训练数据集中训练的网络模型, 从普通场景中随机抽取了 100 帧图像进行了测试. 测试结果见图 6 和表 3. 从图表中可以看出, 在普通场景的测试效果一般, 想要取得更好的结果需要在普通场景的行人数据集中训练网络模型. 但是在普通场景中, 使用本文联合语义的方法比不使用联合语义的方法精度更高, 说明联合语义的方法也具有改善其他场景行人检测的潜力.

表 3 普通场景行人检测实验数据表

方法	*mae*	*mse*	图像大小 (像素)	平均时间 (毫秒/帧)
改进的 Faster RCN	3.74	8.83	长边<1000	177
改进的 Faster RCNN 和联合语义	3.24	8.02	长边<1000	184

5 结论与展望

本文提出了一种有效减少误检和漏检的俯视行人检测方法. 基于 Faster RCNN 框架进行行人目标检测,

在兼顾处理速度的情况下, 可以有很好的鲁棒性. 我们把容易造成误检和漏检的行人语义属性作为辅助检测目标, 和行人联合训练, 然后分别检测, 再反过来利用行人的语义属性辅助判别行人目标, 自适应地调整行人检测得分, 融合行人及其语义属性, 增加了行人目标的可靠性. 实验证明, 本文的检测方法错误率小, 处理速度快, 适合应用于商场或超市出入口等行人情况复杂、混淆目标多的监控场景. 下一步研究方向为优化深度网络速度、把语义属性辅助与损失函数结合、目标跟踪、滞留物检测机制和异常徘徊检测机制等.

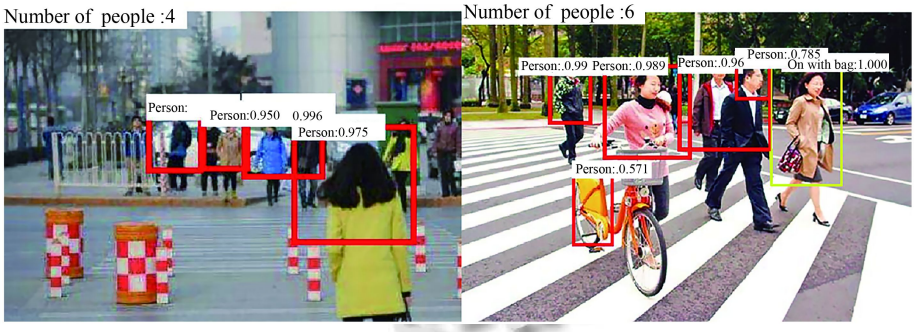


图 6 普通场景行人检测实验结果

参考文献

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

2 徐渊, 许晓亮, 李才年, 等. 结合 SVM 分类器与 HOG 特征提取的行人检测. 计算机工程, 2016, 42(1): 56–60, 65.

3 甘玲, 邹宽中, 刘肖. 基于 PCA 降维的多特征级联的行人检测. 计算机科学, 2016, 43(6): 308–311. [doi: 10.11896/j.issn.1002-137X.2016.06.061]

4 刘璨, 孟朝晖. 基于改进型 LBP 特征的监控视频行人检测. 电子设计工程, 2016, 24(21): 48–50. [doi: 10.3969/j.issn.1674-6236.2016.21.015]

5 Girshick R, Donahue J, Darrell T, *et al*. Region-based convolutional networks for accurate object detection and segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142–158. [doi: 10.1109/TPAMI.2015.2437384]

6 Girshick R. Fast R-CNN. Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. 2015. 1440–1448.

7 Ren SQ, He K, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2015. 1137.

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

9 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands. 2016. 21–37.

10 Hu YC, Chang H, Nian FD, *et al*. Dense crowd counting from still 25-6362s with convolutional neural networks. Journal of Visual Communication and Image Representation, 2016, 38: 530–539. [doi: 10.1016/j.jvcir.2016.03.021]

11 Wang C, Zhang H, Yang L, *et al*. Deep people counting in extremely dense crowds. Proceedings of the 23rd ACM International Conference on Multimedia. Brisbane, Australia. 2015. 1299–1302.

12 芮挺, 费建超, 周遊, 等. 基于深度卷积神经网络的行人检测. 计算机工程与应用, 2016, 52(13): 162–166. [doi: 10.3778/j.issn.1002-8331.1502-0122]

13 左艳丽, 马志强, 左宪禹. 基于改进卷积神经网络的人体检测研究. 现代电子技术, 2017, 40(4): 12–15.

14 Vu TH, Osokin A, Laptev I. Context-aware CNNs for person head detection. Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. 2015. 2893–2901.

15 Stewart R, Andriluka M, Ng AY. End-to-end people detection in crowded scenes. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. 2016. 2325–2333.

16 Zhang HW, Kyaw Z, Chang SF, *et al*. Visual translation embedding network for visual relation detection. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. 2017.

17 顾广华, 韩晰瑛, 陈春霞, 等. 图像场景语义分类研究进展综述. 系统工程与电子技术, 2016, 38(4): 936–948.

18 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale 25-6362 recognition. arXiv:1409.1556, 2014.

19 Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland. 2014. 818–833.