

## 融合多模型和帧间信息的行人检测算法

王 斌<sup>1,3)</sup>, 刘 洋<sup>2)\*</sup>, 唐 胜<sup>1)</sup>, 郭俊波<sup>1)</sup>

<sup>1)</sup> (中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

<sup>2)</sup> (国家计算机网络应急技术处理协调中心 北京 100029)

<sup>3)</sup> (中国科学院大学 北京 100049)

(liuyang195753@sina.com)

**摘 要:** 行人检测在车辆辅助驾驶、视频监控、智能机器人等领域具有重要的应用价值. 针对当前行人检测算法在视频分辨率低和背景复杂的情况下存在很多误检及漏检的问题, 提出一种融合多模型和帧间信息的行人检测算法. 首先融合 Fast R-CNN 和 Faster R-CNN 模型的互补检测结果获取精准的检测窗口; 然后采用视频帧间上下文融合算法来弥补单帧图像检测算法存在的漏检和误检. 实验结果表明, 在 Caltech 行人检测数据库上, 在每幅图像虚警率 (FPPI) 为 10% 的条件下, 该算法丢失率仅为 14.04%, 比 Faster R-CNN 单模型丢失率 (16.09%) 降低 2.05%; 利用多模型和帧间信息融合对行人检测结果进行校正, 能提高行人检测性能.

**关键词:** 行人检测; 卷积神经网络; 模型融合; 帧间信息融合  
**中图法分类号:** TP391.41

## Pedestrian Detection with Fusion of Multi-models and Intra-frame Information

Wang Bin<sup>1,3)</sup>, Liu Yang<sup>2)\*</sup>, Tang Sheng<sup>1)</sup>, and Guo Junbo<sup>1)</sup>

<sup>1)</sup> (Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

<sup>2)</sup> (National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029)

<sup>3)</sup> (University of Chinese Academy of Sciences, Beijing 100049)

**Abstract:** Pedestrian detection has significant applications in the field of driver assistance, video surveillance, intelligent robot and so on. To address the problem of many false detections and missing windows produced by existing pedestrian detection algorithm when the videos are of low resolution with complicated background, this paper propose a fusion method with multi-models and intra-frame information for pedestrian detection. It combines complementary detection results generated from Fast R-CNN and Faster R-CNN which can obtain more precise detection windows at first. Then a frame-context information fusion method is proposed to further remove the false positives and false negatives resulting from single frame information. In the Caltech pedestrian detection dataset, under the condition of false positive per image (FPPI) equaling 10%, the missing rate can be reduced to 14.04% which is 2.05% lower than that (16.09%) of Faster R-CNN model. It shows that fusion of multi-models and intra-frame information can correct the results of previous pedestrian detection and improve the detection performance accordingly.

**Key words:** pedestrian detection; convolutional neural networks; model fusion; intra-frame information fusion

收稿日期: 2016-03-16; 修回日期: 2016-07-22. 基金项目: 国家自然科学基金(61572472, 61525206); 国家“八六三”高技术研究发展计划(2014AA015202); 北京市自然科学基金(4152050). 王 斌(1990—), 男, 博士研究生, 主要研究方向为基于深度学习的多媒体内容分析; 刘 洋(1983—), 女, 硕士, 工程师, 论文通讯作者, 主要研究方向为多媒体处理与信息检索; 唐 胜(1972—), 男, 博士, 副研究员, 硕士生导师, CCF 高级会员, 主要研究方向为媒体大数据内容分析与检索、计算机视觉与模式识别; 郭俊波(1975—), 男, 博士, 高级工程师, 主要研究方向为多媒体检索.

近年来,随着交通安全、视频监控、智能机器人等领域的智能化需求的提高,行人检测技术成为支持这些领域的关键技术之一。然而由于行人的衣着多样性、姿态多样性、场景多样性、光照变化多样性使得设计一种实时性好、准确率高、鲁棒性强的行人检测算法成为一个难点。

传统的行人检测算法大体分为候选窗口选择、特征提取和分类器设计 3 个阶段。为了获取更好的行人检测效果,研究者们提出了用候选区域(region proposal)算法<sup>[1]</sup>代替传统的滑动窗口策略,使得候选窗口定位更加准确,而且时间复杂度更低。为了获取更好的特征,研究者使用更加鲁棒的卷积神经网络(convolutional neural networks, CNNs)特征代替经典的梯度方向直方图(histogram of oriented gradients, HOG)等特征。本文重点考虑如何使用多模型融合并结合帧间上下文信息对检测窗口进行修正,使得最终行人检测结果得到提升。

## 1 相关工作

2005 年, Dalal 等<sup>[2]</sup>提出 HOG 特征,可在一定程度上抑制平移旋转带来的问题,并且对光照变化有一定的适应性, HOG 的提出大大加快了行人检测的发展,大量改进算法被相继提出。Zhu 等<sup>[3]</sup>使用积分直方图技术快速计算 HOG 特征大大提高了行人检测速度; Tuzel 等<sup>[4]</sup>提出协方差描述子,在黎曼流形上训练 LogitBoost 分类器从而提高检测性能; Wu 等<sup>[5]</sup>提出 CENTRIST 特征并结合级联分类器应用于行人检测; Felzenszwalb 等<sup>[6]</sup>提出形变部件模型(deformable part mode, DPM),进一步推动了目标检测的发展; Yan 等<sup>[7]</sup>结合 DPM 提出了鲁棒的多分辨率行人检测算法; 陈锐等<sup>[8]</sup>提出基于稳定区域的梯度方向直方图。

2006 年, Hinton 等<sup>[9]</sup>在 Science 上发表文章,开启了深度学习在学术界的浪潮。Krizhevsky 等<sup>[10]</sup>使用深度 CNNs 在 2012 年国际大规模视觉识别大赛(ImageNet large scale visual recognition challenge, ILSVRC)中将分类任务的 Top-5 错误率降低到 15.3%。之后 CNNs 在图像处理领域更是取得了突飞猛进的发展<sup>[11-13]</sup>。深度学习提取的特征相对于传统特征鲁棒性更强,用于行人检测可以得到更好的检测性能。Ouyang 等<sup>[14]</sup>结合 CNNs 特征和 DPM 思想,提出 JointDeep 算法用于行人检测。Sermanet 等<sup>[15]</sup>根据卷积稀疏编码提出了非监督学习算法 ConvNet-U。Luo 等<sup>[16]</sup>提出的可切换深度网络模型

(switchable deep network, SDN)结合行人各个部件的特征进行行人检测。Wang 等<sup>[17]</sup>使用扩充数据训练基于 R-CNN<sup>[18]</sup>的行人检测器,并使用多窗口融合算法提升行人检测准确率。Tian 等<sup>[19]</sup>结合行人属性和场景属性,提出 TA-CNN 模型用于行人检测。

本文使用 Fast R-CNN<sup>[20]</sup>以及 Faster R-CNN<sup>[21]</sup>框架作为基本检测模型。Faster R-CNN 的端到端模型不需要先提取候选窗口,可以直接产生最终的窗口,大大提高了检测的速度;同时由于 Faster R-CNN 网络结构的共享性,使得网络能够得到更好的行人检测结果。本文利用这 2 个模型各自优点进行模型融合;同时考虑到大多数算法并没有使用视频之间关联信息,提出一种视频帧间信息融合算法,进一步提高了行人检测性能。

## 2 基于多模型和帧间信息融合的行人检测

在目标检测任务中, Fast R-CNN 和 Faster R-CNN 都取得了不错的效果。Faster R-CNN 发表在 ICCV2015 上的版本实际上是候选区域网络(region proposal networks, RPN)和 Fast R-CNN 的结合。后期的改进版本则使得 Faster R-CNN 成为独立的一个网络结构,其获取的检测窗口本质上不同于 Fast R-CNN 的两阶段算法:首先获取图像的候区域,然后提取 CNN 特征并进行分类。2 种算法融合可以得到更好的检测结果。本文算法整体结构如图 1 所示。本文算法中视频帧分别输入到 Fast R-CNN 和 Faster R-CNN 检测框架,然后对检测得到结果进行融合,最后利用帧间上下文信息对当前视频帧检测结果进行校正。

### 2.1 基于 Fast R-CNN 的行人检测

1) 利用边缘框(Edge Boxes)算法对输入图像帧提取候选区域。Fast R-CNN 算法在目标检测领域取得了显著效果,主要得益于鲁棒的 CNNs 特征和更加精准的候选窗口。候选区域算法相比于滑动窗口算法可以获得数量少且召回率高的待检测窗口,对于后继的特征提取和分类任务既节省时间,又大大减少了假阳例的数目。

候选区域算法分为分组算法和窗口得分两大类<sup>[1]</sup>。选择搜索(Selective Search)算法<sup>[22-23]</sup>是一种优秀的分组算法,其中综合考虑了目标的颜色、纹理、包含关系等信息;边缘框(Edge Boxes)算法<sup>[24]</sup>则是优秀的窗口得分算法,考虑了目标的边缘信息;两者检测性能相似,但 Edge Boxes 算法速度更快。为此,本文使用 Edge Boxes 算法代替

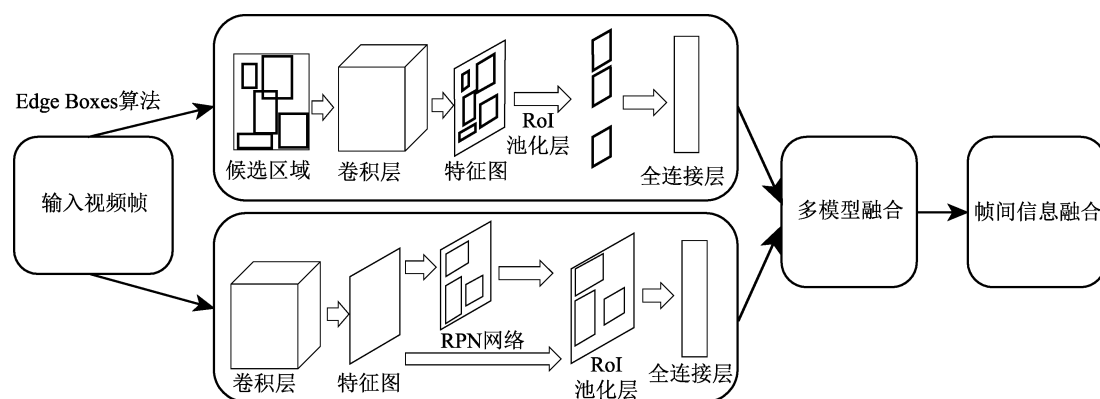


图 1 行人检测框架图

Fast R-CNN 中使用的 Selective Search 算法对行人进行候选区域的提取。

2) 提取完候选区域后, 将视频帧和对应的候选区域位置信息输入到 Fast R-CNN 的网络, 获取每个候选区域特征并进行打分。网络中的 ROI 池化层使 CNNs 只需要对全图提取一次特征, 然后直接将每个候选区域的特征做一次映射, 即可得到候选区域对应的特征, 大大加快了检测速度。网络中的多任务损失函数可以让网络在获取候选窗口类别的同时对候选窗口进行位置校正, 使最终检测结果定位更加准确。

经过 Fast R-CNN 获取的检测窗口有很多冗余, 所以使用非极大抑制算法(non-maximum suppression, NMS)去除重叠的冗余窗口。

## 2.2 基于 Faster R-CNN 的行人检测

不同于 Fast R-CNN 算法, Faster R-CNN 在进行目标检测时不需要提前获取候选区域, 而是直接利用 CNNs 实现候选窗口的提取; 之后网络继续对候选窗口进行打分筛选和回归, 得到最终的检测结果。

Faster R-CNN 中 RPN 网络结构实现了候选窗口的提取, 本文使用的 RPN 网络如图 2 所示。考虑到 Caltech 行人检测数据库<sup>[25]</sup>视频帧中行人分辨率较低, 本文在原始 RPN 网络锚点(anchor)的尺度上增加一个 64 的尺度, 设置为 {64, 128, 256, 512} 4 类, 长宽比设置为 {1:1, 1:2, 2:1} 3 类, 最终锚点数为 12。输入的视频帧经过 VGG16 网络的卷积层提取到了 512 个特征图, RPN 网络设置了 3×3 的滑动窗口, 在特征图上提取锚点对应的特征。每个滑动窗口对应输入的 12 种窗口, 12 种窗口对应着 512 维的特征向量。特征向量后接 2 个损失层: 一个负责分类损失计算(由于行人检测是二值分类, 12 种窗口总共对应 24 种划分), 另一个负责窗口位置损失计算(每个窗口 4 个坐标, 总共对应 48 个坐标)。

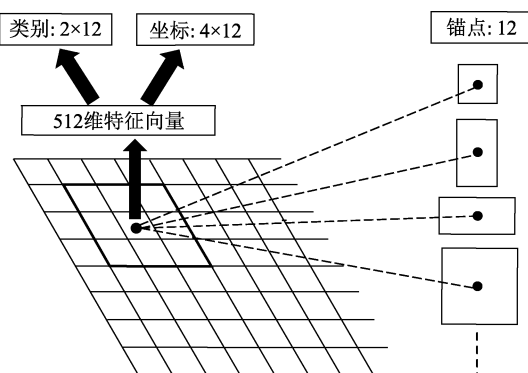


图 2 行人检测 RPN 网络结构

接下来对 RPN 网络所有窗口按得分进行排序, 取前 300 个窗口作为最终的候选窗口; 然后输入到后继的网络进行类别划分; 最后利用 NMS 算法对冗余窗口进行过滤。

## 2.3 模型融合

Fast R-CNN 中候选区域是 Edge Boxes 算法根据边缘信息提取得到的, 而端到端的 Faster R-CNN 模型是 RPN 网络直接利用 CNNs 特征进行提取的。前者适合提取任意尺度和长宽比的候选区域, 但对边缘信息有一定要求; 后者由于 RPN 网络滑动窗口的遍历性, 会对所有位置兼顾, 但由于锚点只有 12 类, 并不能很好地兼顾任意尺度和长宽比(虽然窗口位置回归可以在一定程度上缓解这个问题)。本文结合两者的优点, 对检测结果进行融合并设计了 3 种模型融合方式。下文中  $A$  和  $B$  分别表示使用 Fast R-CNN 和 Faster R-CNN 模型得到的行人检测窗口。

### 1) $A+B$ 型

$A+B$  表示取两者检测结果的并集, 合并时分为 3 种情形: 窗口在  $A, B$  中同时出现则使用加权平均合并窗口并提高此窗口得分; 如果只在效果好的模型中出现, 窗口得分保持不变; 如果只在效果差的模型中出现, 对窗口得分进行抑制(降低窗口得分)。

## 2) A-B型

A-B表示以 Fast R-CNN 检测结果作为基准,使用 Faster R-CNN 得到的结果对其进行过滤。如果 Fast R-CNN 中检测到的行人同时出现在 Faster R-CNN 的检测的行人集合中,即检测窗口交集比并集(intersection-over-union, IoU)的数值大于 0.3,则认为这个检测到的行人是正确的;如果检测的行人得分较低并且没有出现在 Faster R-CNN 中检测到的行人集合中( $\text{IoU} \leq 0.3$ ),则该检测结果是假阳例的可能性更大,直接过滤掉。

## 3) B-A型

B-A表示以 Faster R-CNN 检测结果作为基准,使用 Fast R-CNN 检测结果进行过滤。

## 2.4 帧间上下文融合

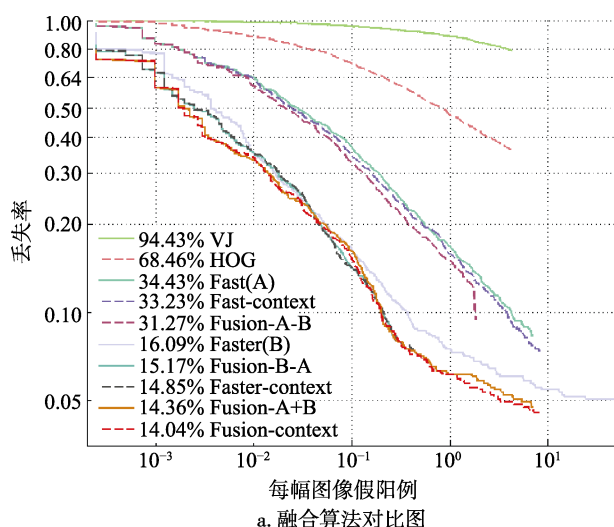
对于视频的行人检测,视频当前帧与上下帧之间有着极大的相关性,本文根据帧间信息对检测结果进一步融合,进一步提高检测结果。利用帧间信息,本文从2个方面进行检测窗口的融合。由于视频的当前帧和前后帧的变化很小,假定当前帧检测的结果应该和前后帧检测结果非常接近。

1) 如果在视频当前帧某个位置存在行人并且前后帧此位置都不存在行人,那么当前帧的检测结果是误检可能性大,对此窗口得分进行抑制。

2) 如果在视频当前帧某个位置不存在行人但前后帧在这个位置都检测出行人,认定当前帧漏检。所以结合前后帧信息得到新的检测窗口

$$(x_1, y_1, x_2, y_2)_{\text{Cur}} = \frac{(x_1, y_1, x_2, y_2)_{\text{Prev}} + (x_1, y_1, x_2, y_2)_{\text{Next}}}{2}$$

其中,  $(x_1, y_1, x_2, y_2)$  表示包围盒左上点和右下点坐标, Cur 表示当前帧, Prev 表示前一帧, Next 表示下一帧。



## 3 实验结果与分析

本文使用 Caltech 行人数据库进行实验,该数据库采用车载摄像头拍摄,视频总长度大约为 10 h,帧率为 30 帧/s;其中标注了 25 万帧(约 137 min),总共包含 2 300 个独立的行人,35 万个包围盒;视频分辨率为  $640 \times 480$ ,质量相对较差。

数据集总共分为 11 个集合: set00~set05 为训练集, set06~set10 为测试集。对于训练集视频,本文每隔 3 帧取 1 帧作为训练数据,这样既能保证训练 CNNs 的数据量充足,也避免了取所有帧带来的数据冗余。本文实验在合理子集<sup>[25]</sup>上进行评估,只检测那些行人高度大于 49 个像素、遮挡率小于 35% 的行人。

为了证明算法的通用性,本文在 ETH 行人检测数据库<sup>[26]</sup>上进行进一步验证。

### 3.1 Fast R-CNN 行人检测实验

本文使用 Edge Boxes 算法对训练视频帧提取候选区域,统计发现,每幅图像平均提取到约 7 000 个候选区域。考虑到行人的宽高比特性,将宽高比大于 1 的候选区域全部去掉(一半左右),去除无用的候选区域可以加快网络的收敛速度。

行人检测是二分类问题,将 Fast R-CNN 网络输出数目设置为 2,窗口回归数目设置为 8。使用训练完的网络进行行人检测,最终检测精度(每 10 幅图像出现一个假阳例的情况下对应的丢失率)为 34.43%,如图 3a 中的 Fast(A)。

### 3.2 Faster R-CNN 行人检测实验

考虑到 Caltech 行人检测数据库中图像的分辨率较低,本文在 RPN 原始网络锚点的 3 种尺度的

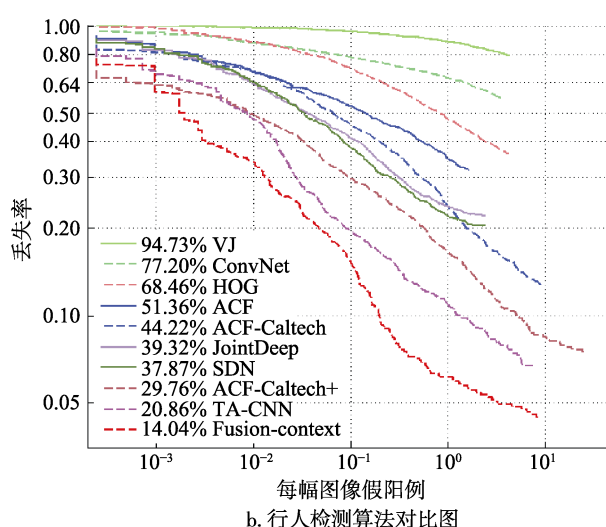


图3 Caltech 行人检测数据库检测性能图

基础上 {128, 256, 512} 添加了 64 像素的尺度, 并且将图像保持长宽比放大到短边 800 个像素. 使用训练的模型进行行人检测的丢失率仅为 16.09%, 如图 3a 中的 Faster(B).

### 3.3 多模型融合实验

使用  $A+B$  模型, 从图 3a 中 Fusion- $A+B$  可以看出:  $A+B$  模型融合得到的检测结果为 14.36%, 丢失率较最好的单模型 Faster(B)低了 1.73%. 这是因为 Fast R-CNN 和 Faster R-CNN 模型融合后窗口的召回率更高, 而且对窗口得分进行了校正, 所以性能得到了很好的提升.

另外, 单个行人检测模型容易将复杂的背景误检测为行人. 本文考虑用  $A-B$  和  $B-A$  模型算法剔除误判为行人的检测窗.

如图 3a 中的 Fusion- $A-B$ ,  $A-B$  模型(以 Fast R-CNN 为基准)融合得到的检测结果为 31.27%, 丢失率较最好的单模型 Faster(A)增加了 15.18%; 但相对于 Fast(B)丢失率降低了 3.16%.

如图 3a 中的 Fusion- $B-A$ ,  $B-A$  模型(以 Faster R-CNN 为基准)融合得到的检测结果为 15.17%, 丢失率较最好的单模型 Faster(A)降低了 0.92%.

从  $A-B$  和  $B-A$  模型可以看出, 这种模型融合可以减少误判从而提高检测性能, 虽然 Fast R-CNN 检测结果比较差, 但是用其对 Faster R-CNN 的假阳例进行剔除仍然可以提升行人检测性能.

本文最终使用性能最好的  $A+B$  算法进行融合.

### 3.4 帧间信息融合实验

从图 3a 中可以看出, 对 Fast R-CNN 进行帧间信息融合后, 检测丢失率从 34.43% 降到 33.23% (Fast-context); 同样地, 对 Faster R-CNN 检测结果使用帧间信息融合, 使得丢失率由原来的 16.09% 降低到 14.85% (Faster-context); 最后, 在 2 个单模型

帧间信息融合的基础上进行模型融合(Fusion-context, Fusion 是以 Faster-context 为基准), 最终模型检测丢失率仅为 14.04%. 因此本文算法优于很多已有模型, 如图 3b 所示.

为了验证本文算法的通用性, 利用 Caltech 上训练的 Fast R-CNN 和 Faster R-CNN 模型对 ETH 数据库进行测试并融合, 结果如图 4 所示. 可以看出, 相比于最好的单模型无上下文模型 Fast(48.8%), 使用本文算法(Fusion-context)使得丢失率降低 1.12%.

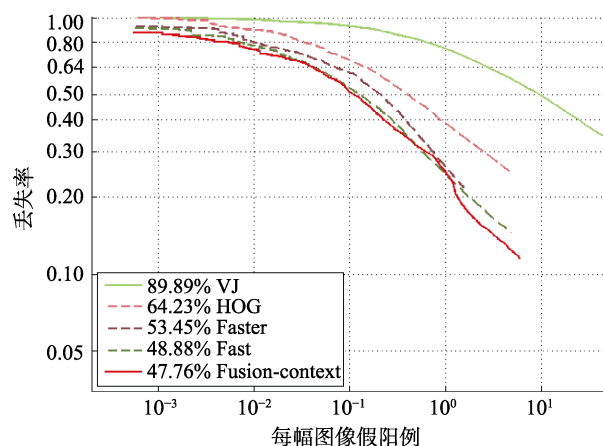


图 4 ETH 行人检测数据库检测性能图

Caltech 数据库行人检测效果图如图 5 所示(其中真实的行人只有 3 个). 图 5a 所示为 Fast R-CNN 检测结果, 有 4 个误检窗口; 图 5b 所示为对 Fast R-CNN 进行帧间信息融合, 去掉了 1 个误检窗口; 图 5c 所示为 Faster R-CNN 检测结果, 有 3 个误检; 图 5d 所示为对 Faster R-CNN 进行帧间融合, 去掉了 1 个误检窗口; 图 5e 所示为在 Faster R-CNN 基础上融合 Fast R-CNN 结果, 检测正确; 图 5f 表示在 Faster-context 基础上融合 Fast-context, 检测正确.

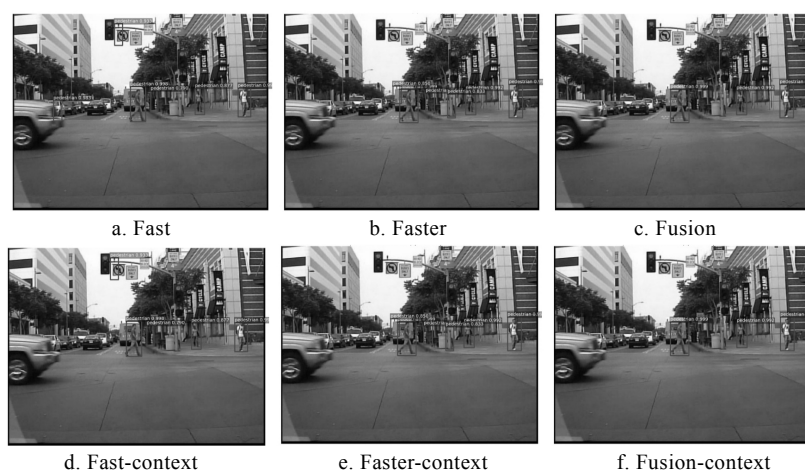


图 5 行人检测效果图

## 4 结 语

本文基于现有的目标检测框架,结合行人检测任务的特性进行改进,得到了较好的行人检测效果。本文提出的多模型融合可以对多模型检测窗口进行融合并对得分重新校正,同时对于视频行人检测,本文利用帧间上下文信息对当前帧检测结果进行更正。实验结果证明,本文的模型融合及帧间信息融合算法是有效的。尽管如此,本文使用的帧间信息只考虑了帧间行人的空间位置关系,所以今后我们将考虑融合更多动态信息到检测模型,进一步提升行人检测性能。

## 参考文献(References):

- [1] Hosang J, Benenson R, Dollar P, *et al.* What makes for effective detection proposals?[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(4): 814-830
- [2] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C] // *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2005, 1: 886-893
- [3] Zhu Q, Yeh M C, Cheng K T, *et al.* Fast human detection using a cascade of histograms of oriented gradients[C] // *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2006, 2: 1491-1498
- [4] Tuzel O, Porikli F, Meer P. Human detection via classification on riemannian manifolds[C] // *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2007: 1-8
- [5] Wu J X, Geyer C, Rehg J M. Real-time human detection using contour cues[C] // *Proceedings of the IEEE International Conference on Robotics and Automation*. Los Alamitos: IEEE Computer Society Press, 2011: 860-867
- [6] Felzenszwalb P F, Girshick R B, McAllester D, *et al.* Object detection with discriminatively trained part-based models[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, 32(9): 1627-1645
- [7] Yan J J, Zhang X C, Lei Z, *et al.* Robust multi-resolution pedestrian detection in traffic scenes[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2013: 3033-3040
- [8] Chen Rui, Peng Qimin. Pedestrian detection based on HOG of stable area[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2012, 24(3): 372-377(in Chinese)  
(陈 锐, 彭启民. 基于稳定区域梯度方向直方图的行人检测方法[J]. *计算机辅助设计与图形学学报*, 2012, 24(3): 372-377)
- [9] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. *Science*, 2006, 313(5786): 504-507
- [10] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[OL]. [2016-03-16]. <https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks>
- [11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. [2016-03-16]. <http://arxiv.org/abs/1409.1556>
- [12] Szegedy C, Liu W, Jia Y Q, *et al.* Going deeper with convolutions[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2015: 1-9
- [13] He K M, Zhang X Y, Ren S Q, *et al.* Deep residual learning for image recognition[OL]. [2016-03-16]. <http://arxiv.org/abs/1512.03385>
- [14] Ouyang W L, Wang X G. Joint deep learning for pedestrian detection[C] // *Proceedings of the IEEE International Conference on Computer Vision*. Los Alamitos: IEEE Computer Society Press, 2013: 2056-2063
- [15] Sermanet P, Kavukcuoglu K, Chintala S, *et al.* Pedestrian detection with unsupervised multi-stage feature learning[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2013: 3626-3633
- [16] Luo P, Tian Y L, Wang X G, *et al.* Switchable deep network for pedestrian detection[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2014: 899-906
- [17] Wang B, Tang S, Zhao R Z, *et al.* Pedestrian detection based on region proposal fusion[C] // *Proceedings of the 17th IEEE International Workshop on Multimedia Signal Processing*. Los Alamitos: IEEE Computer Society Press, 2015: 1-6
- [18] Girshick R, Donahue J, Darrell T, *et al.* Rich feature hierarchies for accurate object detection and semantic segmentation[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2014: 580-587
- [19] Tian Y L, Luo P, Wang X G, *et al.* Pedestrian detection aided by deep learning semantic tasks[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2015: 5079-5087
- [20] Girshick R. Fast r-cnn[C] // *Proceedings of the IEEE Conference on Computer Vision*. Los Alamitos: IEEE Computer Society Press, 2015: 1440-1448
- [21] 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, PP(99): Page 1
- [22] Uijlings J R R, van de Sande K E A, Gevers T, *et al.* Selective search for object recognition[J]. *International Journal of Computer Vision*, 2013, 104(2): 154-171
- [23] van de Sande K E A, Uijlings J R R, Gevers T, *et al.* Segmentation as selective search for object recognition[C] // *Proceedings of the IEEE International Conference on Computer Vision*. Los Alamitos: IEEE Computer Society Press, 2011: 1879-1886
- [24] Zitnick C L, Dollar P. Edge boxes: locating object proposals from edges[M] // *Lecture Notes in Computer Science*. Heidelberg: Springer, 2014, 8693: 391-405
- [25] Dollar P, Wojek C, Schiele B, *et al.* Pedestrian detection: an evaluation of the state of the art[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 34(4): 743-761
- [26] Ess A, Leibe B, Schindler K, *et al.* A mobile vision system for robust multi-person tracking[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2008: 1-8