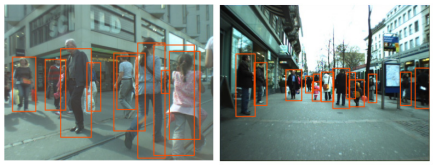
Opto-Electronic Engineering
光 电 工 程

Article

2019 年，第 46 卷，第 9 期

DOI: 10.12086/oee.2019.180606

基于改进 R-FCN 的多遮挡行人实时检测算法



刘　辉，彭　力，闻继伟*

江南大学物联网工程学院物联网应用技术教育部工程中心，江苏 无锡 214122

摘要：当前车辆辅助驾驶系统的一个主要挑战就是在复杂场景下实时检测出多遮挡的行人，以减少交通事故的发生。为了提高系统的检测精度和速度，提出了一种基于改进区域全卷积网络(R-FCN)的多遮挡行人实时检测算法。在 R-FCN 网络基础上，引进感兴趣区域(RoI)对齐层，解决特征图与原始图像上的 RoI 不对准问题；改进可分离卷积层，降低 R-FCN 的位置敏感分数图维度，提高检测速度。针对行人遮挡问题，提出多尺度上下文算法，采用局部竞争机制进行自适应上下文尺度选择；针对遮挡部位可见度低，引进可形变 RoI 池化层，扩大对身体部位的池化面积。最后为了减少视频序列中行人的冗余信息，使用序列非极大值抑制算法代替传统的非极大值抑制算法。检测算法在基准数据集 Caltech 训练检测和 ETH 上产生较低的检测误差，优于当前数据集中检测算法的精度，且适用于检测遮挡的行人。

关键词：多遮挡行人；可分离卷积层；多尺度上下文；可形变 RoI 池化层

中图分类号：TP391.4　　**文献标志码：**A

引用格式：刘辉，彭力，闻继伟. 基于改进 R-FCN 的多遮挡行人实时检测算法[J]. 光电工程，2019，**46**(9): 180606

Multi-occluded pedestrian real-time detection algorithm based on preprocessing R-FCN

Liu Hui, Peng Li, Wen Jiwei*

Engineering Research Center of Internet of Things Technology Applications of the Ministry of Education, School of Internet of Things Engineering, Jiangnan University, Wuxi, Jiangsu 214122, China

Abstract: One of main challenges of driver assistance systems is to detect multi-occluded pedestrians in real-time in complicated scenes, to reduce the number of traffic accidents. In order to improve the accuracy and speed of detection system, we proposed a real-time multi-occluded pedestrian detection algorithm based on R-FCN. RoI Align layer was introduced to solve misalignments between the feature map and RoI of original images. A separable convolution was optimized to reduce the dimensions of position-sensitive score maps, to improve the detection speed. For occluded pedestrians, a multi-scale context algorithm is proposed, which adopt a local competition mechanism for adaptive context scale selection. For low visibility of the body occlusion, deformable RoI pooling layers were introduced to expand the pooled area of the body model. Finally, in order to reduce redundant information in the video sequence, Seq-NMS algorithm is used to replace traditional NMS algorithm. The experiments have shown that there is low detection error on the datasets Caltech and ETH, the accuracy of our algorithm is better than that of the detection algorithms in the sets, works particularly well with occluded pedestrians.

收稿日期：2018-11-21；　收到修改稿日期：2019-01-10

基金项目：教育部中国移动创新基金资助项目(MCM20182019)

作者简介：刘辉(1992-)，男，硕士研究生，主要从事模式识别的研究。E-mail：1391570995@qq.com

通信作者：闻继伟(1981-)，男，博士，副教授，主要从事控制理论的研究。E-mail：wjw8143@aliyun.com

Keywords: multi-occluded pedestrian; separable convolution layer; multi-scale context; deformable RoI pooling layer

Citation: Liu H, Peng L, Wen J W. Multi-occluded pedestrian real-time detection algorithm based on preprocessing R-FCN[J]. *Opto-Electronic Engineering*, 2019, **46**(9): 180606

1 引 言

在车辆辅助驾驶系统中，系统的主要检测对象就是出现在视野范围内不同位置的行人和车辆，特别在遮挡情况下准确地检测出行人的位置并及时做出反应尤为重要。2018 年，Uber 的自动驾驶汽车由于目标检测系统未能成功识别出行人位置，导致发生了致命的交通事故。从事故的原因可以看出，车辆辅助驾驶的目标检测系统还需要进一步的完善。当前行人遮挡检测问题已经成为行人检测领域内公认的难点之一[1-2]，如何提高遮挡条件下行人的检测精度，一直是学者们研究的方向。因此，本文在基于改进区域全卷积网络(region-based fully convolutional networks，R-FCN)[3]的基础上，提出了一种用于在复杂场景下，可以实时检测遮挡行人的算法。

文献[4]详细分析了目前深度学习在无人驾驶汽车的应用，提到的关键问题就是如何提高检测系统的精度和算法的运行速度。针对运动目标的检测问题，文献[5]将深度卷积神经网络引入到运动目标光流检测中，将前后帧图像及目标光流场图像作为网络的输入，自适应地学习运动目标光流，用于运动目标的检测。考虑到传统方法的精度问题，本文采用基于卷积神经网络(convolution neural network，CNN)的方式训练出高精度的检测模型，而当前基于 CNN 的目标检测算法主要分为单步检测和两步检测两种方式。单步检测的算法主要是基于单次多盒检测器(single shot multibox detector，SSD)[6]，两步检测的主要是基于 Faster R-CNN[7]。单步目标检测的算法虽然在速度上可以达到实时的效果，但对于遮挡的目标检测精度远没有达到工业的要求。基于两步的目标检测算法虽然精度很高，但是由于存在区域建议网络(region proposal networks，RPN)[6]用于提取区域建议，导致检测速度无法达到实时的需求。文献[3]在 Faster-RCNN 的基础上提出了 R-FCN 目标检测算法，通过引进位置敏感分数图将目标的位置信息融合进全局平均池化层中，替代了 Faster R-CNN 检测网络中的全连接层部分，检测精度高于 Faster R-CNN，但由于生成了高维度的位置敏感分数图，达不到实时检测的速度。SSD[5]则是直接在不同尺度的特征图上进行预测，输出目标边界框的坐标和类别，没有 RPN 网络提取区域建议的过程。由于 SSD 算法是通过低层的卷积层来输出小目标的特征图，高层的卷积输出大目标的特征图，而低层卷积含有的相关信息较少，导致遮挡的小目标的检测效果不理想。

行人检测的主要工作就是检测系统可以在不同场景下准确的框出行人的位置坐标，提醒系统做出相应的措施。但是由于周围环境的复杂性(例如遮挡，光照弱等)，导致目标检测系统的精度受到了很大的挑战。相比于非遮挡的行人，遮挡的行人容易出现由于遮挡导致检测信息的丢失，检测得分值下降到低于阈值而出现漏检问题。为了克服噪声和光照的影响，文献[8]提出了一种基于完整的局部二值数(completed local binary count，CLBC)和方向梯度直方图(histogram of oriented gradient，HOG)特征融合的行人检测算法，通过计算原始图像的 CLBC 特征，将图像的三种特征融合来描述图像，并使用主成分分析(principal component analysis，PCA)方法降低特征维度，最后使用直方图交叉核 SVM(histogram intersection Kernel support vector machine，HIKSVM)分类器实现最终对行人的检测。但传统方法容易受到外部因素的影响，而且检测精度不高。因此为了克服复杂背景下行人相互遮挡的影响，学者们采用了不同的方法提高遮挡行人的检测精度。文献[9]将行人检测中特征提取、形变处理、遮挡处理和分类联合在一起，最大化其能力，检测遮挡行人；文献[10]用深度学习结合部分部位模型得到一个深度部位(DeepParts)模型来解决行人检测中的遮挡问题，构建一个部分池化层，然后对每个部分训练一个检测器，综合所有检测器的分数，得到整个行人检测结果；文献[11]为了处理部分检测器的不完善性，提出了一种概率行人检测框架。使用基于部分可变形的模型来获得整个部分检测器的分数，并且将部分的可见性建模为隐变量。使用多模型融合算法综合每个部分的检测效果，但由于存在多模型融合过程，导致速度过慢。虽然最新的算法在解决遮挡行人的问题取得了显著的进步，但系统的检测速度和精度依然需要提高。

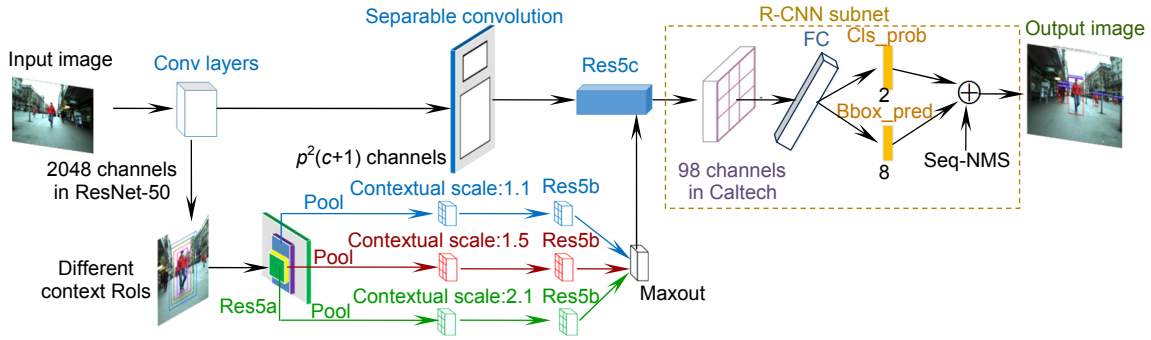


图 1 整体网络结构图

Fig. 1 Schematic of the network structure

为了提高复杂场景下遮挡行人的检测精度与速度，本文提出了快速可形变全卷积行人检测网络(fast deformable fully convolutional networks，Fast D-FCN)，其结构如图 1 所示，1) 在特征提取层中引入可分离卷积(separable convolution)[12]，提高检测速度；2) 在 res5a 层引入多尺度上下文，提取不同上下文信息的行人感兴趣区域；在 res5b 层使用可形变池化代替传统的池化,增加不同池化部位的感受面积；3) 通过 res5c 层输出固定维度的通道特征向量，在分类层输出分类概率，在回归层输出边界框信息，最后通过序列非极大值抑制(sequence non-maximum suppression，Seq-NMS)[13]算法输出检测结果，提高了 R-FCN 算法在复杂场景下多行人遮挡检测精度和速度。

2 R-FCN 网络

最初的 R-FCN 网络如图 2 所示，主要由基础的卷积网络如 ResNet-50[14]、RPN、位置敏感池化层和全局平均池化层构成。R-FCN 主要的计算量集中在位置敏感分数图和全局平均池化层。位置敏感分数图块大小为 $w\times h$ 的感兴趣区域(region of interest，RoI)，其中每块(bin)的大小为 $\frac{w}{k}\times\frac{h}{k}$，为每类产生 k^2 维分数图。对第 (i,j) 块($0\le i,j\le k-1$)位置敏感 RoI 池化操作为

$$r_c(i,j\,|\,\Theta)=\frac{1}{n}\sum_{(x,y)\in bin(i,j)} z_{i,j,c}(x+x_0,y+y_0\,|\,\Theta)\quad,\quad(1)$$

其中：$r_c(i,j\,|\,\Theta)$ 为第 c 类第 (i,j) 块的池化响应。$z_{i,j,c}$ 为 $p^2(c+1)$ 维分数图中的输出，(x_0,y_0) 为 RoI 的左上角坐标，n 为图块的像素总数，且 Θ 为网络的参数。对于训练 c 类目标，位置敏感分数图输出 $p^2(c+1)$ 维分数图(p 表示 RoI 的尺寸)。对该 RoI 每类的所有相对空间位置的分数平均池化：

$$r_c(\Theta)=\sum_{i,j} r_c(i,j\,|\,\Theta)\quad,\quad(2)$$

$r_c(\Theta)$ 表示每个类别的平均得分，最后通过 Softmax 函数分类出每个类别的概率。在回归层输出每个类别的位置坐标。R-FCN 的主要计算量是生成了 $p^2(c+1)$ 维数的位置敏感分数图，导致平均池化层的计算量过大。随着训练类别的增加，分数图的维度也在增加，导致算法的检测速度无法达到实时。

3 快速可形变全卷积网络(Fast D-FCN)

本文算法针对 R-FCN 计算量的问题和行人遮挡问题做了以下的改进：在 R-FCN 网络的基础上，引进可分离卷积层，使用卷积组代替单个卷积实验，减少 R-FCN 所产生的位置敏感分数图的维数，加速 R-FCN 的检测速度；引进 RoI 对齐层[15]解决特征图与原始图像上 RoI 的不对齐问题。针对行人遮挡问题，提出多尺度上下文算法，提取不同尺度的行人 RoIs；引进可形变 RoI 池化层[16]，扩大模型对身体部位的池化面积。最后为了降低检测视频中行人的冗余信息，引进 Seq-NMS 算法，代替传统的非极大值抑制(sequence non-maximum suppression，NMS)算法。相比于 R-FCN，

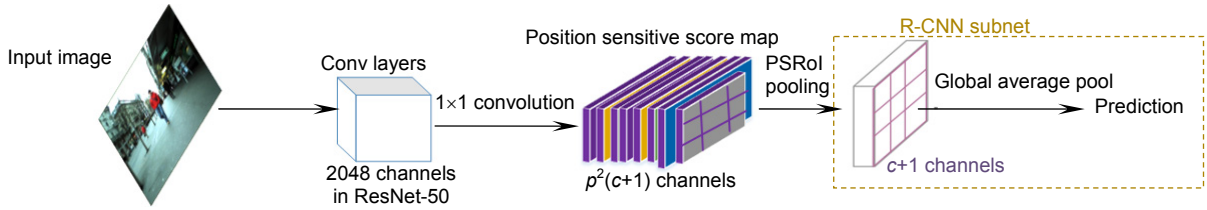


图 2 R-FCN 结构图

Fig. 2 Schematic of the R-FCN structure

本文的 Fast D-FCN 算法在检测遮挡行人方面取得了显著的效果，而且达到了实时检测。

3.1 改进的 R-FCN 网络

3.1.1 可分离卷积

针对 R-FCN 中位置敏感池化层输出的高维数的分数图，本文使用了可分离卷积，使用少的通道数量提取特征图，将 R-FCN 的位置敏感分数图变“薄”，减少位置敏感分数图的维数，加快 R-FCN 的检测速度。选择 ResNet-50 作为基础网络，提取特征图。图 1 中“Conv layers”表示 ResNet-50 的结构，网络参数如表 1 所示。由于 R-FCN 的 conv5_x 的 2048 维输出接一个 1024 维(用 $p\times p\times(c+1)$表示)的 1×1 的过滤器去卷积生成位置敏感分数图(p 表示池化核大小，c 表示类别)，因此本文将可分离卷积设置在 conv5_x 层，工作流程如图 3 所示。本文设置可分离卷积核的大小 K=15，代替 1×1 卷积，相比于 R-FCN 中 K=7，可以获取 $15^2/7^2$=4.6 倍的特征图感受野。为了减少位置敏感分数图的计算量，本文使用 1×15 和 15×1 代替 15×15 的卷积核操作，其中输入通道数 $C_{\rm in}$=2048，$C_{\rm mid}$=256。由于只有行人和背景两类检测目标，所以 $C_{\rm out}$ 为 $2\times p^2$=98，其中 p 表示池化大小为 7，最后进行总和输出。本文输入大小为 $w\times h\times$2048 特征图，输出为 $w\times h\times$98 的特征图。根据模型参数量计算公式可得：

原始参数：$2048\times15\times15\times98=4.5\times10^7$。

加入可分离卷积后的参数量：

$(15\times1\times2048\times256+1\times15\times256\times98)+(1\times15\times2048\times256+15\times1\times256\times98)=1.6\times10^7$。

使用大的卷积核，再加入分离卷积之后，生成的位置敏感分数图的参数量是之前的三分之一，扩大了感受野，提高了模型的训练和检测速度。

表 1 ResNet-50 网络参数表

Table 1 ResNet-50 network parameter

Layer	Output size	K	Output channels
Image	224×224		
Conv1	112×112	3×3	256
maxPool	56×56	3×3	256
Stage2	28×28		512
	28×28		
Stage3	14×14		1024
	14×14		
Stage4	7×7		2048
	7×7		
FC	1×1		1000
Comp*			98 M

(Comp*表示模型的复杂度，K 表示卷积内核大小)

3.1.2 RoI 对齐层

在 RPN 中，获取到的预选框的位置主要是通过模型回归得到的，通常为浮点数，而池化后的特征图要求尺寸固定为整数。所以在将候选框边界量化为整数坐标值和将量化后的边界区域平均分割成 $k\times k$ 单元时，候选框与回归位置出现了偏差(即 RoI 不对齐)。为了解决偏差，本文引进了 RoI 对齐层。RoI 对齐首先遍历了每一个候选区域，保持浮点数边界不变。然后将候选区域分割成 $k\times k$ 单元，每个单元的边界不做量化。最后在每个单元中计算固定四个坐标位置，用双线性插值的方法计算出这四个位置的值，然后进行最大池化操作。本文引进了 RoI 对齐层，减少了池化操作的计算量，提高了检测精度。

3.2 多尺度上下文

遮挡处理是行人检测中最具挑战性问题之一，使用多区域、多上下文特征是比较有效的处理方法[17-19]。考虑到 Caltech 数据集中行人有多种尺度的变化和复杂的遮挡情况，不同的行人需要不同的上下文信息。尤其是被严重遮挡的行人，在视频中可分辨度较低，比没有被遮挡的行人需要更多的上下文信息，因此本文提出了多尺度上下文算法，从多个离散尺度中提取不同区域的上下文信息，如图 1 所示。假设每个行人 RoI 具有宽度 w，高度 h，S 为 RoI 的上下文比例因子。因此，多尺度上下文区域具有的宽度为 $w\times S$ 和高度为 $h\times S$，与原始 RoI 有相同的中心，比例因子 S 为 s_1,s_2,s_3(实验中 s_1=1.1，s_2=1.5，s_3=2.1)。为了提高上下文特征对多尺度 RoI 的自适应选择能力，本文通过引进 Maxout[20]作为该层的激活函数，融合多个行人 RoIs 输出到池化层，其中 Maxout 公式如下：

$$o_{\rm out}=\max_{k=1.2,3}(s_k(w_k*x+b_k)) \quad , \tag{3}$$

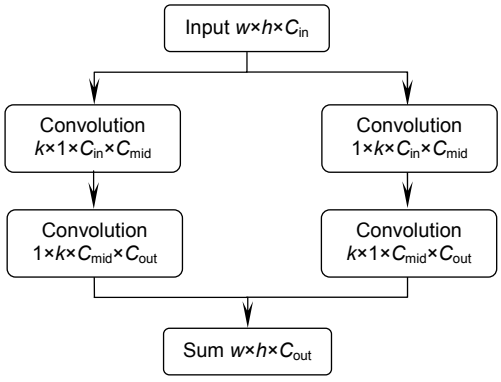


图 3 可分离卷积执行 $k\times1$ 和 $1\times k$ 卷积顺序

Fig. 3 Separable convolution performs $k\times1$ and $1\times k$ convolution

其中：w_k，b_k 表示第 k 个神经元的学习参数，w_k 表示 k 个神经元的权重，b_k 表示偏移量。当使用 Maxout 时，三个特征图融合成具有相同维度的单个特征图。这些不同的特征图通过 Maxout 以数据驱动的方式进行选择，在 RoI 池化层之前共享每层的权重。最后每个不同尺度的 RoI 前向传播到可形变位置敏感 RoI 池化层，获得固定分辨率的特征图。

3.3 可形变池化层

本文在 res5a_branch2a 层、res5a_bran-ch2b 层和 res5a_branch2c 层分别引入可形变 RoI 池化层，池化多尺度上下文信息提取行人 RoIs。可形变 RoI 池化层中可形变 RoI 组的大小为 7，池化大小为 7，每个部位采样 4 次，空间尺度大小为 0.0625，在加入偏移之后，改变了原池化层长方形的结构，变成指向不同方向的偏移点，增大了池化层的空间采样位置的额外偏移量(如图 4 所示)。池化层中每个 RoI 大小为 $w\times h$，划分为 $k\times k$ 块。最初的 RoI 池化层从输入特征图 $\boldsymbol{x}$ 获得 $k\times k$ 维的池化特征图 $\boldsymbol{y}$，第 (i,j) 个$(0\le i,j<k)$网格的特征图表示为

$$\boldsymbol{y}(i,j)=\sum_{p\in bin(i,j)}\boldsymbol{x}(p_0+p)/n_{i,j}\quad , \tag{4}$$

其中：$\boldsymbol{y}(i,j)$ 表示池化后的特征图，p_0 表示 RoI 的左上角，p 表示池化大小，$n_{i,j}$ 表示网格的像素值。为了扩大对形变部位的池化面积，可形变池化层增加偏移量 $(\Delta p_{i,j},0\le i,j<k)$ 到空间网格的位置中。等式(4)为

$$\boldsymbol{y}(i,j)=\sum_{p\in bin(i,j)}\boldsymbol{x}(p_0+p+\Delta p_{i,j})/n_{i,j}\quad 。 \tag{5}$$

由于 Δp_n 为分数，所以等式(5)通过双线性插值为

$$\boldsymbol{x}(p)=\sum_q G(q,p)\cdot\boldsymbol{x}(q)\quad , \tag{6}$$

其中：p 表示任意位置 $(p=p_0+p_n+\Delta p_n)$，q 枚举特征图 $\boldsymbol{x}$ 中的所有整体空间位置，$G(\cdot,\cdot)$ 为双线性插值内核，G 函数可以分为两个一维内核的乘积：

$$G(q,p)=g(q_x,p_x)\cdot g(q_y,p_y)\quad , \tag{7}$$

其中：$g(x,y)=\max(0,1-|x-y|)$，可以快速计算 $G(q,p)$ 非零时的 q 值。由于 R-FCN 的位置敏感 RoI 池化是可变的，在位置敏感分数图上用于目标分类和边界框的回归，但没有可学习的加权图层。所以本文将位置敏感 RoI 池化扩展为可形变的位置敏感 RoI 池化，应用到位置敏感分数图，使偏移学习对 RoI 大小具有不变性。相比于 R-FCN 网络中传统的位置敏感 RoI 池化层，通过引进了可形变位置敏感 RoI 池化层，扩大了池化层的池化面积。

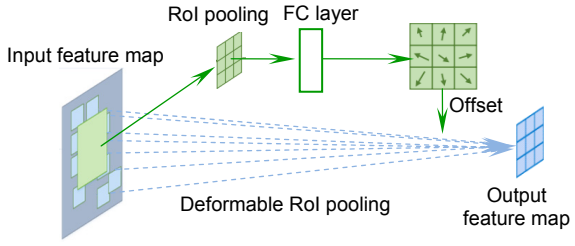


图 4 3×3 可形变 RoI 池化示例

Fig. 4 Illustration of 3×3 deformable RoI pooling

3.4 检测网络

3.4.1 Seq-NMS 算法

如图 1 所示，本文的目标检测网络部分主要由一个全连接层，输出 1024 维特征向量，和两个输出层(分类层，回归层)组成。由于只有行人和背景两类检测目标，所以最终的分类输出值为边界框为行人的概率和位置信息。最初的 R-FCN 主要用于在图片序列上检测目标，但在检测视频中的目标时会忽略了时间分量的显著信息。实验发现 NMS 算法经常会选择重合面积过大的区域建议框，导致检测出错误的边界框。为了解决这个问题，本文引进了 Seq-NMS 利用时间信息重新对视频序列中边界框的大小进行排序。Seq-NMS 主要分为三个步骤：1) 序列选择，2) 序列重新评分，3) 抑制。实验重复这三个步骤直到没有序列被留下。

1) 序列选择：视频序列 V 由 t 帧组成，$\{v_0,...,v_t\}$。其中每帧 v_t 都有一组区域建议框 b_t 和得分值 s_t。在给定一组区域边界框 $\boldsymbol{B}=\{b_0,...,b_t\}$ 和检测分数 $S=\{s_0,...,s_t\}$ 作为输入，如果它们的重合面积的交并比(intersection over union，IoU)大于某个阈值，则第 t 帧的边界框可以与第 t+1 帧的边界框相连接。然后在连接序列中找出最大得分序列：

$$i'=\arg\max_{i_s,...,i_e}\sum_{t=t_s}^{t_e}s_t[i_t]\quad , \tag{8}$$

其中 s.t $\mathrm{IoU}(b_t[i_t],b_{t+1}[i_{t+1}])>0.5,\forall t\in[t_s,t_e]$。

Seq-NMS 可以使每个区域建议框保持最高的得分顺序，返回一组索引 i' 用于提取最高得分序列 $\boldsymbol{B}^{\mathrm{seq}}=\{b_{t_s}[i_{t_s}],...,b_{t_e}[i_{t_e}]\}$ 和得分值 $\hat{\boldsymbol{S}}^{\mathrm{seq}}=\{s_{t_s}[i_{t_s}],...,s_{t_e}[i_{t_e}]\}$。

2) 序列重新评分：在序列被选择之后，通过平均值函数 F 产生 $\boldsymbol{S}^{\mathrm{seq}}=F(\hat{\boldsymbol{S}}^{\mathrm{seq}})$。

3) 抑制：如果视频帧 v_t 的候选框与真实框的重合面积比大于设定的阈值(0.5)，则将该边界框从候选框集合中移除。

相比于最初的 NMS 算法，本文可以使用动态索引按照得分顺序排序选择边界框，然后选择大于某个阈值的边界框作为最终的行人框，有效地降低了视频序列中的冗余信息。

3.4.2 多任务损失函数

由于每个训练的行人建议都有一个真实的类别 g 和一个真实边界框回归目标值 t^* 的标签。多任务损失函数 $\boldsymbol{L}$ 在每个目标建议 i 输出检测网络的参数：

$$\boldsymbol{L}=\boldsymbol{L}_{\rm cls}(s_i,g)+1[g\geq 1]\boldsymbol{L}_{\rm loc}(t_i^g,t^*)\ , \tag{9}$$

其中：$\boldsymbol{L}_{\rm cls}$ 和 $\boldsymbol{L}_{\rm loc}$ 分别是分类和边界框回归损失函数。$\boldsymbol{L}_{\rm cls}$ 是 Softmax 损失函数，输出分类的准确值；$\boldsymbol{L}_{\rm loc}$ 为 Smooth L_1 损失函数，输出边界框坐标和宽高。$1[g\geq 1]$ 表示当分类值 $g\geq 1$ 为 1，否则为 0。本文通过引进 Seq-NMS 算法输出检测结果，可以在视频序列中不同范围的遮挡下，准确定位出行人位置。

3.5 算法训练检测步骤

由于多尺度上下文在特征提取和表达与 R-FCN 存在一定的差异，这可能导致学习过程中的不稳定性，所以本文通过多步来训练和检测模型。

步骤 1：输入数据集 $V=\{v_1,v_2,...,v_n\}$，其中图片序列大小统一为 640×480，输入到网络中开始训练。

步骤 2：为了获取不同维度的特征图大小，设置特征提取网络 ResNet-50 依次输出 128，256，512，1024，2048 维的特征图。

步骤 3：训练 RPN 网络，本文使用了三个比例{1:2，1:1，2:1}，五个不同尺度{32^2，64^2，128^2，256^2，512^2}，采用了 15 种不用类型滑动窗口。设置锚点的重合面积大于阈值 0.5 为正样本，小于 0.3 为负样本进行训练。

步骤 4：对于 RPN 提取到的不同上下文的 RoIs，本文设置了具有不同比例因子的 Maxout 激活函数，自适应的选择最大区域的上下文边界框的 RoIs。

步骤 5：对于特征提取网络获取到的 2048 维特征图和多尺度上下文算法获取的 RoIs，本文设置了 7×7 大小的可形变池化操作，输出 1024 维特征向量到分类和回归层。

步骤 6：在输出行人的位置坐标和分类概率后，针对视频中的冗余信息，本文实验了不同阈值的 Seq-NMS，在阈值为 0.5 时，得到了最佳的检测精度。

4 实验结果与分析

实验平台是基于 64 位的 Ubuntu16.04 操作系统和 NVIDIA GTX 1070 GPU，软件有 Matlab2014a、Python2.7，使用的深度学习框架为 caffe。改进 ResNet-50 在 ImageNet 预先训练的模型，使用梯度下降算法进行训练，学习率为 0.001，20k 批次的动量为 0.9，权重衰减设置为 0.0005。同时当训练次数达到 1 万次时，学习率降低 10%。一张随机选择的图片，每个 mini-batch 包括 128 个随机采样行人建议，包括 32 个正样本和 96 个负样本。为了提高训练样本的多样性，本文使用 Caltech 数据集作为训练集和检测集。针对遮挡的行人，本文对数据集进行了筛选，剔除了非行人的图片序列，并进行了水平翻转。为了验证算法的实用性，本文还测试了 ETH 行人数据集。在成绩评估过程中，如果检测框与真实标注框重合面积的交并比值大于某个阈值，则认为该检测框与真实标注框相匹配。本文实验表明交并比(intersection over union, IoU)大于 0.5，行人边界框预测为正值，IoU 小于 0.3，则为负值。IoU 公式为

$$IoU=\frac{area(B_{\rm dt}\bigcap B_{\rm gt})}{area(B_{\rm dt}\bigcup B_{\rm gt})}>0.5\ , \tag{10}$$

其中：$B_{\rm dt}$ 为最终检测框；$B_{\rm gt}$ 为真实标注框。若多个 $B_{\rm dt}$ 与 $B_{\rm gt}$ 匹配，则决策得分高的检测框将被选择，而没有被匹配的 $B_{\rm dt}$ 记为误检，未被匹配的 $B_{\rm gt}$ 记为漏检。

实验结果的评价标准根据 Dollar[1]等提出的工具箱，用 FPPI-miss rate(false positives per image against miss rate)来衡量滑动窗口行人检测算法的性能。把测试图片中包含行人的窗口切割下来，然后从不包含行人的测试集采样非行人的样本；最后将窗口作为测试集来评估算法的性能。

$$FPPI=\frac{FP}{TN+FP}\times 100\%\ , \tag{11}$$

$$MR=\frac{FN}{FN+TP}\times 100\%\ , \tag{12}$$

其中：TP、FP、FN 和 TN 分别表示将行人样本分类为行人样本数、将非行人样本分类为行人样本数、将行人样本分类为非行人样本数、将非行人样本分类为非行人样本数。

4.1 Caltech 实验结果比较

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，总共包含 2300 个独立的行人。数据集总共分为 11 个集合：set00~set05 为训练集，set06~set10 为测试集，图片分辨率为 640×480。对于训练集，每隔 5 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。测试集每隔 30 帧取 1 帧作为测试数据，共 4024 张图片，其中标记遮挡的行人样本是测试集的 30%左右，主要分为部分遮挡行人(遮挡度为 0~35%)和严重遮挡的行人(遮挡度为 35%~80%)。采用漏检率(MR)来总结探测器的性

能，其中 *MR* 越低表示算法性能越好。为了对比实验结果，本文将网络在训练集上迭代 20 万次生成最终的检测模型，选择了遮挡行人检测最新的 CNN 算法 DeepParts[10]、MS-CNN[18]、RPN+BF[20]、R-FCN[3]、TA-CNN[21]、F-DNN[22]等与本文的算法 Fast D-FCN 作对比。图 5(a)显示了部分遮挡的检测结果，算法 Fast D-FCN 的漏检率为 14.86%，比 R-FCN 低了 1.23%，比 F-DNN 低了 0.55%。此外，对于严重遮挡的行人 5(b)，本文的方法达到了最低漏检率 42.36%，比 R-FCN 的 55.81%和 F-DNN 的 55.13%，低了 13.45%和 12.77%。

为了测试算法的检测速度，对比了最新的目标检测算法 R-FCN、SSD 和本文的算法 Fast D-FCN 的检测速度，都使用相同的基础模型 ResNet-50，保证模型的参数量相同，结果如表 2 所示。本文算法在 Batch_size 为 1 时，图片分辨率为 640×480，检测数据集共花费了 1 分 30 秒，平均 48.71 f/s。SSD 在图片分辨率为 512×512 时，平均检测速度为 35.42 f/s。本文算法相比于 R-FCN 的 11.24 f/s，提高了 4 倍的检测速度，达到了实时的检测速度。

综合图表可知，在 Caltech 数据集中，本文的算法 Fast D-FCN 在 R-FCN 的基础上提高了算法的精度和速度，使得在部分行人遮挡的漏检率比 R-FCN 低了 1.23%，对于严重遮挡的行人，比 R-FCN 低了 13.45%，而且检测速度比 R-FCN 提高了四倍，达到了实时的检测效果。

4.2 ETH 数据实验比较

为了验证本文算法的适用性，在 ETH 数据集进行了测试，ETH 数据集共三个视频序列：set00，set01，set02，共 1804 张图片，分辨率大小为 640×480，主要拍摄的是城市街道的行人序列，环境比较复杂，而且互相遮挡严重。本文将所有图片作为测试集，使用在 Caltech 数据集上训练好的模型，选择了 ETH 数据集中经典的检测算法 ACF[23]、LDCF[24]、RPN+BF[20]、TA-CNN[21]作为对比算法，如图 6。算法评估了数据集中遮挡比例在 35%以上的行人的检测结果，比数据集中最新的算法 TA-CNN 的漏检率低了 4.75%，而且检测每张图片的花费时间为 21 ms，速度为 48 f/s，达到了实时的检测速度。

Caltech 数据检测效果如图 7(a)，7(b)所示，7(a)表示部分遮挡，7(b)表示严重遮挡。ETH 数据检测效果如图 7(c)，7(d)所示，7(c)表示部分遮挡，7(d)表示严重遮挡。

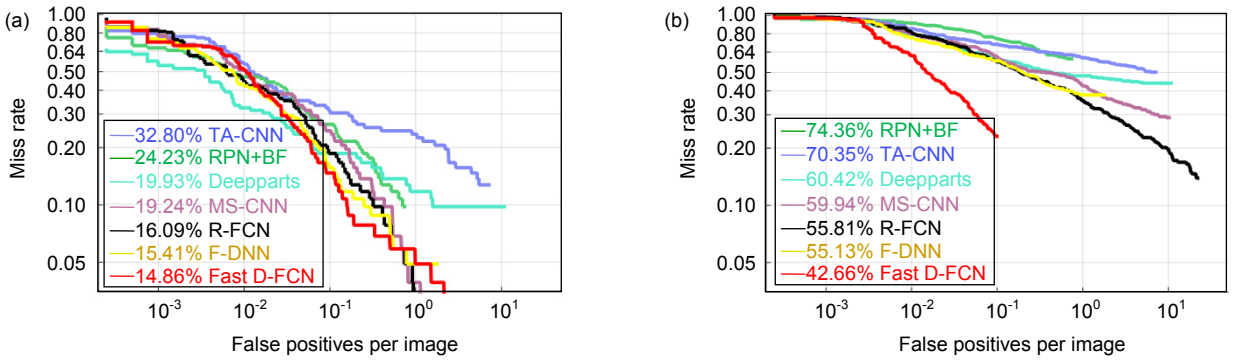


图 5 Caltech 数据集的结果比较。(a) 部分遮挡；(b) 严重遮挡

Fig. 5 Comparison results on the Caltech bench-mark. (a) Part-occlusion; (b) Heavy-occlusion

表 2 漏检率与检测速度比较

Table 2 Comparison of miss and detect rate

Algorithm	Fast D-FCN	SSD	R-FCN
Test size	640×480	512×512	640×480
Base-model	ResNet-50	ResNet-50	ResNet-50
Part-occlusion(MR)/%	14.86	20.49	16.09
Heavy-occlusion(MR)/%	42.36	57.64	55.81
Speed/(f/s)	48.71	35.42	11.24

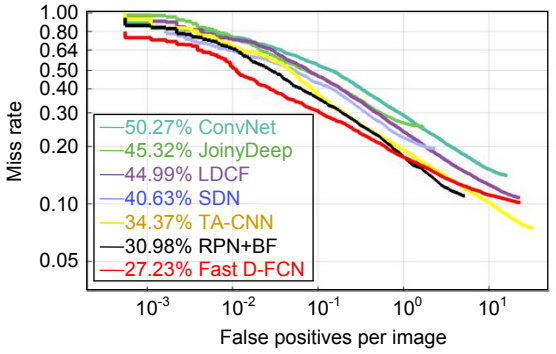


图 6 ETH 数据集检测结果

Fig. 6 Results on the ETH benchmark

图 7 算法检测效果

Fig. 7 Test result carried out by the algorithm

5 结 论

本文提出的遮挡行人实时检测算法的目的是为了解决 R-FCN 算法在复杂场景下，检测遮挡行人的精度较低和无法实时检测的问题。本文的算法在两个广泛应用的行人检测数据集 Caltech 和 ETH 上取得了很好的检测精度，而且速度达到了实时性，但由于生成的模型的参数量较大，无法用于嵌入式端处理。如何将生成模型压缩到可用于嵌入式端实时检测是我们下一步研究的主要方向。

参考文献

[1] 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.

[2] Wang X Y, Han T X, Yan S C. An HOG-LBP human detector with partial occlusion handling[C]//*Proceedings of the 12th IEEE International Conference on Computer Vision*, 2009: 32–39.

[3] Dai J F, Li Y, He K M, *et al*. R-FCN: object detection via region-based fully convolutional networks[C]//*Proceedings of the 30th Conference on Neural Information Processing Systems*, 2016: 379–387.

[4] Wang K J, Zhao Y D, Xing X L. Deep learning in driverless vehicles[J]. *CAAI Transactions on Intelligent Systems*, 2018, **13**(1): 55–69.
王科俊, 赵彦东, 邢向磊. 深度学习在无人驾驶汽车领域应用的研究进展[J]. 智能系统学报, 2018, **13**(1): 55–69.

[5] Wang Z L, Huang M, Zhu Q B, *et al*. The optical flow detection method of moving target using deep convolution neural network[J]. *Opto-Electronic Engineering*, 2018, **45**(8): 180027.
王正来, 黄敏, 朱启兵, 等. 基于深度卷积神经网络的运动目标光流检测方法[J]. 光电工程, 2018, **45**(8): 180027.

[6] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot MultiBox detector[C]//*Proceedings of the 14th European Conference on Computer Vision*, 2016: 21–37.

[7] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[C]//*Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015: 91–99.

[8] Cheng D Q, Tang S X, Feng C C, *et al*. Extended HOG-CLBC for pedstrain detection[J]. *Opto-Electronic Engineering*, 2018, **45**(8): 180111.
程德强, 唐世轩, 冯晨晨, 等. 改进的 HOG-CLBC 的行人检测方法[J]. 光电工程, 2018, **45**(8): 180111.

[9] Ouyang W L, Wang X G. Joint deep learning for pedestrian detection[C]//*Proceedings of 2013 IEEE International Conference on Computer Vision*, 2014: 2056–2063.

[10] Tian Y L, Luo P, Wang X G, *et al*. Deep learning strong parts for pedestrian detection?[C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*, 2015: 1904–1912.

[11] Ouyang W L, Zeng X Y, Wang X G. Partial occlusion handling in pedestrian detection with a deep model[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016, **26**(11): 2123–2137.

[12] Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision[J]. arXiv:1512.00567v3[cs.CV], 2015.

[13] Han W, Khorrami P, Le Paine P, *et al*. Seq-NMS for video object detection[J]. arXiv:1602.08465[cs.CV], 2016.

[14] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770–778.

[15] He K M, Gkioxari G, Dollár P, *et al*. Mask R-CNN[C]// *Proceedings of 2017 IEEE International Conference on Computer Vision*, 2017: 2980–2988.

[16] Dai J F, Qi H Z, Xiong Y W, *et al*. Deformable convolutional networks[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*, 2017: 764–773.

[17] Bell S, Zitnick C L, Bala K, *et al*. Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2874–2883.

[18] Cai Z W, Fan Q F, Feris R S, *et al*. A unified multi-scale deep convolutional neural network for fast object detection[C]//*Proceedings of the 14th European Conference on Computer Vision*, 2016: 354–370.

[19] Goodfellow I J, Warde-Farley D, Mirza M, *et al*. Maxout networks[J]. *JMLR WCP*, 2013, **28**(3): 1319–1327.

[20] Zhang L L, Lin L, Liang X D, *et al*. Is faster R-CNN doing well for pedestrian detection?[C]//*Proceedings of the 14th European Conference on Computer Vision*, 2016: 443–457.

[21] Tian Y L, Luo P, Wang X G, *et al*. Pedestrian detection aided by deep learning semantic tasks[C]//*Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015: 5079–5087.

[22] Du X Z, El-Khamy M, Lee J, *et al*. Fused DNN: a deep neural network fusion approach to fast and robust pedestrian detection[C]//*Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision*, 2017.

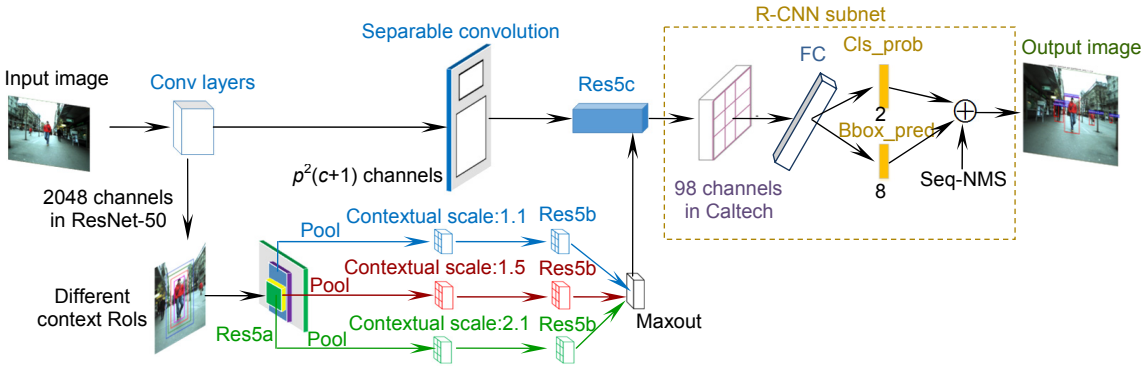
[23] Dollár P, Appel R, Belongie S, *et al*. Fast feature pyramids for object detection[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, **36**(8): 1532–1545.

[24] Nam W, Dollár P, Han J H. Local decorrelation for improved pedestrian detection[C]//*Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014: 424–432.

Multi-occluded pedestrian real-time detection algorithm based on preprocessing R-FCN

Liu Hui, Peng Li, Wen Jiwei*

Engineering Research Center of Internet of Things Technology Applications of the Ministry of Education,
School of Internet of Things Engineering, Jiangnan University, Wuxi, Jiangsu 214122, China



Multi-pedestrians occlusion detection network structure

Overview: Pedestrian detection is a research hot in the fields of pattern recognition and machine learning. It is widely used in areas such as video surveillance, intelligent driving and robot navigation. Computer realizes pedestrian detection automatically, which can reduce the burden of people in a certain extent. With the development of deep learning theory, the convolutional neural network has made remarkable achievements in the field of pedestrian detection by improving the generation strategy of candidate regions and optimizing the network structure and training methods. Different from the usual object detection, pedestrian is a moving target and not a rigidity instance because of the change of occlusion and height. The methods base on feature extraction cannot meet the industrial requirements. So we choose a method base on convolutional neural network to achieve higher accuracy and real-time detection for multi-occluded pedestrians. The main work of pedestrian detection is to accurately draw the position coordinates of pedestrians in different scenarios and output the detection accuracy of the system. However, due to the complexity of the surrounding environment (such as multiple occlusion, weak illumination, etc.), the accuracy of the pedestrian detection system is greatly challenged. Compared with non-occluded pedestrians, multi-occluded pedestrians are easier to lose the detection information, and cause the decrease of pedestrian detection score below the threshold, thus missed the detection. In order to improve the detection accuracy and speed of multi-occlusion pedestrians in complex scenes, we propose a fast deformable full convolutional pedestrian detection network (called Fast D-FCN). Based on R-FCN, we introduced RoI Align layer to solve misalignments between the feature map and RoI of original images. To improve detection speed, we improved a separable convolution to reduce dimensions of position-sensitive score maps, put it on feature extraction layers of ResNet-50. For multi-occluded pedestrians, we proposed a multi-scale context in res5a of ResNet-50, which adopt a local competition mechanism for adaptive context scale selection. In the case of low visibility of the body occlusion, we introduced deformable RoI pooling layers to expand the pooled area of the body model in res5b of ResNet-50. Through the res5c layer, the channel feature vector of the fixed dimension, classification probability in the classification layer, and bounding box information in the regression layer are outputted. Finally, in order to reduce redundant information in the video sequence, we used Seq-NMS algorithm to replace traditional NMS algorithm. The experiments have shown that on the datasets Caltech, the detection error about part occlusion and heavy occlusion decrease 0.55% and 12.77% respectively compared to F-DNN. On the ETH dataset, our algorithm is better than the accuracy of other detection algorithms, and works particularly well with multi-occluded pedestrians.

Citation: Liu H, Peng L, Wen J W. Multi-occluded pedestrian real-time detection algorithm based on preprocessing R-FCN[J]. *Opto-Electronic Engineering*, 2019, **46**(9): 180606

Supported by Education Ministry and China Mobile Science Research Foundation (MCM20182019)

* E-mail: wjw8143@aliyun.com