基于加速区域卷积神经网络的夜间行人检测研究

叶国林[1,2], 孙韶媛[1,2], 高凯珺[1,2], 赵海涛[3]

[1]东华大学信息科学与技术学院, 上海 201620;

[2]东华大学数字化纺织服装技术教育部工程研究中心, 上海 201620;

[3]华东理工大学信息科学与工程学院, 上海 200237

摘要 行人检测是机器人和无人车夜间工作应用中的重要任务之一,采用加速区域卷积神经网络框架实现夜间红外图像中的行人检测,用区域建议网络生成候选区域,无需单独从图像中生成候选区域。区域建议网络和用于分类以及位置精修的卷积网络中,采用卷积层参数共享机制,使得该框架具有端到端的优点,因此无需手动选取目标特征,实现了从输入图像直接到行人检测的功能。实验结果表明,与使用传统方法和快速区域卷积神经网络相比,使用加速区域卷积网络框架对红外图像进行行人检测的准确率从 68.2%和 73.4%提高到了 90.9%,检测时间从 3.6 s/frame和 2.3 s/frame 缩短到了 0.04 s/frame,达到了实际应用中的实时性要求。

关键词 图像处理; 红外图像; 行人检测; 加速区域卷积神经网络; 区域建议网络

中图分类号 TP391 **文献标识码** A

doi: 10.3788/LOP54.081003

Nighttime Pedestrian Detection Based on Faster Region Convolution Neural Network

Ye Guolin[1,2], Sun Shaoyuan[1,2], Gao Kaijun[1,2], Zhao Haitao[3]

[1] *College of Information Science and Technology, Donghua University, Shanghai* 201620, *China*;

[2] *Engineering Research Center of Digitized Textile and Fashion Technology, Ministry of Education, Donghua University, Shanghai* 201620, *China*;

[3] *School of Information Science and Engineering, East China University of Science and Technology, Shanghai* 200237, *China*

Abstract Pedestrian detection is one of the most important tasks of robots and unmanned vehicles at nighttime. Faster region convolution neural network framework is used to realize the pedestrian detection of infrared image at nighttime. This framework uses region proposal network to generate region proposals. Therefore, it is unnecessary to generate region proposals separately from the image. The parameter sharing mechanism is adopted in the convolutional layers in region proposal network and convolutional network for classification and bounding box regression, which makes the framework an end-to-end advantage. Thus, the pedestrian detection can be implemented from the input image to the detection result directly and it is unnecessary to manually select the features of the target. Experimental results show that the proposed method increases the recognition accuracy from 68.2% and 73.4% to 90.9% and shortens the recognition time from 3.6 s/frame and 2.3 s/frame to 0.04 s/frame compared with the traditional method and fast region convolution neural network, respectively, which reaches the required real-time level in practical applications.

Key words image processing; infrared image; pedestrian detection; faster regional convolution neural network; region proposal network

OCIS codes 100.4996; 100.5010; 040.3060; 110.4155

收稿日期: 2017-03-02; **收到修改稿日期**: 2017-04-01

基金项目: 国家自然科学基金(61375007)、上海市科委基础研究项目(15JC1400600)

作者简介: 叶国林(1992—),男,硕士研究生,主要从事红外图像处理方面的研究。E-mail: 863939325@qq.com

导师简介: 孙韶媛(1974—),女,博士,教授,主要从事红外图像处理方面的研究。

E-mail: shysun@dhu.edu.cn(通信联系人)

1 引 言

在红外图像处理中,行人检测技术应用广泛[1],但由于红外图像具有无色彩信息[2]、纹理细节少[3]、信噪比低[4]等特点,因此红外图像行人检测技术的实时性和准确性较差。传统的行人检测方法通常是结合行人特征提取和机器学习,用滑动窗口遍历一幅完整的图像,再利用训练好的分类器对窗口进行行人与非行人的分类判别,达到行人检测的目的。这类方法虽然能够得到较好的检测结果,但由于在检测时利用多尺度的滑动窗口对整幅图像进行遍历,产生了大量的检测窗口,并且依次对所有的检测窗口进行特征提取,因此计算量剧增,速度较慢。如 Dalal 等[5]先计算正负样本图像的方向梯度直方图(HOG)描述子,组成一个特征向量矩阵,同时对应产生一个指定每个特征向量的类标向量,并输入到支持向量机(SVM)中进行训练得出分类器。林成竹[6]分别提取 HOG 特征和局部二值模式(LBP)特征进行多特征融合,并使用级联的分类器来提高检测正确率。

近几年,卷积神经网络(CNN)在目标检测上的应用取得了突破,Girshick 等[7]提出了区域卷积神经网络(R-CNN)框架将图像的目标检测问题转化为分类问题,该方法先在图像中生成若干个候选区域,再用卷积网络对每个候选区域提取目标特征,并用 SVM 训练一个分类器,对候选区域进行分类,最后根据每个区域分类得分利用非极大值抑制算法优化得出最终的目标边界。但是 R-CNN 提取特征的卷积网络和用于分类的分类器要分开训练,导致了训练过程要耗费大量的时间和存储空间;而且分类器的训练与特征提取网络不相关,这种不合理影响了目标检测的准确率。因此 Girshick[8]又提出了快速区域卷积神经网络(Fast R-CNN)模型,将特征提取和分类融合进一个分类框架,提高了训练模型的速度和目标检测的准确率。但 Fast R-CNN 使用选择性搜索算法[9]单独生成候选区域非常耗时,使该算法不具有实时性。因此,Ren 等[10]在 Fast R-CNN 上增加区域建议网络(RPN)来生成候选区域,构成一种端到端的加速区域卷积神经网络(Faster R-CNN)模型,大大提高了运算速度。

本文利用 Faster R-CNN 实现夜间红外图像行人检测,利用该网络框架中的 RPN 来生成候选区域,利用 Fast R-CNN 来提取特征、分类以及位置精修,由于该框架中的 RPN 和 Fast R-CNN 的卷积层采用参数共享机制,因此整个框架具有端到端的特点,从而提高了行人检测的速度,实现了夜间模式下的实时行人检测。

2 R-CNN 基本原理

在 CNN 出现之前,一直使用全连接的深度神经网络(DNN),该神经网络在图像处理中会导致待训练的参数数目和计算量很大。而 CNN 使用滑窗遍历图像的方式,基于局部关联和权值共享的原则作卷积计算,大大减少了训练的参数数目和计算量,而且对平移、比例缩放、倾斜或其他形式的变形具有高度不变性。CNN 典型结构如图 1 所示。

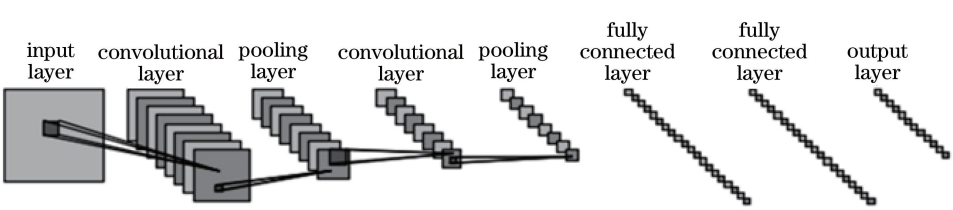

图 1 典型的 CNN 结构图

Fig. 1 Structure of a typical CNN

CNN 直接将整幅图像作为输入,在卷积层中通过滑动窗口对输入图像作卷积,生成特征图,池化层也通过滑窗对特征图中的局部块作最大值或平均值计算,用降采样特征图来压缩数据量和防止过拟合。一般经过若干个卷积层和池化层,再经过激励层,把卷积结果做非线性映射,最终,采用和 DNN 相同的方法,连接几个全连接层来生成全局的特征。

目标检测需要先生成若干个候选区域,且输入到卷积网络中用于检测的也是该图像中的这些候选区域图像,因此称为 R-CNN,其结构如图 2 所示。

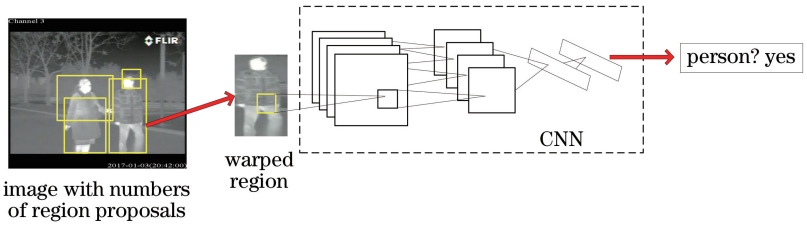

图 2 R-CNN 的结构图

Fig. 2 Structure of R-CNN

3 基于 Faster R-CNN 红外行人检测算法

3.1 Faster R-CNN 整体结构

本文算法的整体网络框架结构如图 3 所示，该网络输入为一张图片，输出为目标的边界框以及概率得分。RPN 生成 300 个候选区域输入给目标识别网络 Fast R-CNN，由于 RPN 和 Fast R-CNN 前面部分都有若干卷积层来计算特征图，因此该框架将这两个网络结合成一个网络，卷积层参数共享，卷积层同时输出给 RPN 和 Fast R-CNN，构成端到端的网络结构。

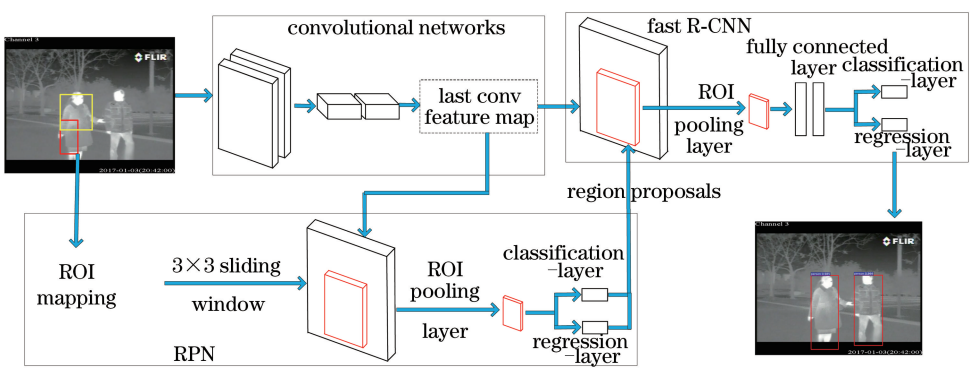

图 3 Faster R-CNN 的结构图

Fig. 3 Structure of Faster R-CNN

3.2 感兴趣区域(ROI)采样层

传统的 R-CNN 框架是将单独提前生成的每个候选区域当成一张图像(可能需要先经过缩放或拉伸等操作使其达到固定大小)输入到网络中进行后续操作，即 CNN 提取特征及 SVM 分类。假设一幅图像有 2000 个候选区域，则就要进行 2000 次操作，其中大部分是进行卷积计算特征图，非常耗时。但由于这 2000 个候选区域都是一张图片的一部分，因此，采用对整幅图像提取一次卷积层特征，然后将候选区域在原图像的位置映射到卷积层特征图上，得出各个候选区域特征图的方法，对于一幅图像只需要提取一次卷积层特征，大大节省了运算时间。

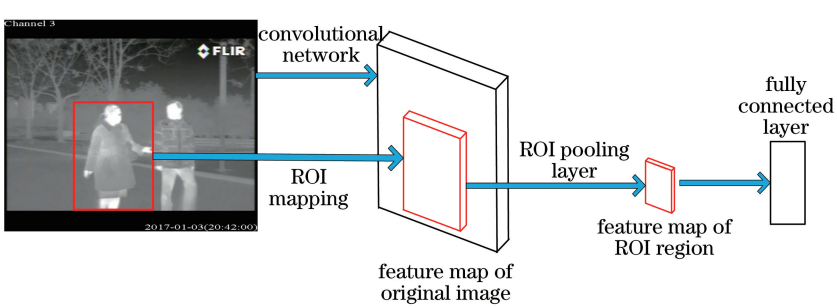

图 4 ROI 采样层

Fig. 4 ROI sampling layer

上述由候选区域即 ROI 在原图的位置到特征图映射任务的层称为 ROI 采样层，如图 4 所示。在映射出各个候选区域的特征图后，需要将这些特征图输入给全连接层，而全连接层需要大小一样的特征输入，但候选区域的大小却是不相同的，因此，该框架对文献[11]中空间金字塔池化层(SPP-Layer)进行修改，采用单尺度输出 7 pixel×7 pixel 的特征图，假设输入的候选区域为 (r,c,h,w)，ROI 采样层首先产生 7×7 个 $r\times c\times(h/7)\times(w/7)$ 的块，然后用 Max-Pooling 方式求出每一个块的最大值，这样输出都是 7 pixel×7 pixel的特征图。

3.3 RPN

Fast R-CNN 不具有实时性的原因，是输入到网络中的候选区域由基于图论和最小生成树的选择性搜索方法获得，该方法计算量很大。因此，为了提高识别速率，需要采用其他思路解决候选框生成问题。使用 CNN 直接生成候选区域，该网络称为 RPN，如图 5 所示。在 RPN 中，最后一个卷积层有 256 个卷积核，所以特征图有 256 个，特征维度为 256 维，每个特征图大小约为 40 pixel×60 pixel。用3 pixel×3 pixel 的滑窗滑动特征图，当滑窗滑到每个位置时，预测输入图像为三种尺度(128 pixel，256 pixel，512 pixel)和三种长宽比(1∶1，1∶2，2∶1)的候选区域，所以每一个滑动位置有 9 个候选区域，一幅图像会生成约 2000(40 pixel×60 pixel×9)个候选区域。在卷积层后面接两个分支的全连接层，一个是分类层输出 2 个得分，用于判定候选区域是目标还是背景，另一个是边界回归层输出 4 个得分，用于对候选区域的边界进行微调(与 Fast R-CNN 类似)。虽然由 RPN 选取的候选区域约有 2000 个，但是该框架根据候选区域的得分高低筛选了前 300 个输入到目标识别网络，用于提高运算速度。

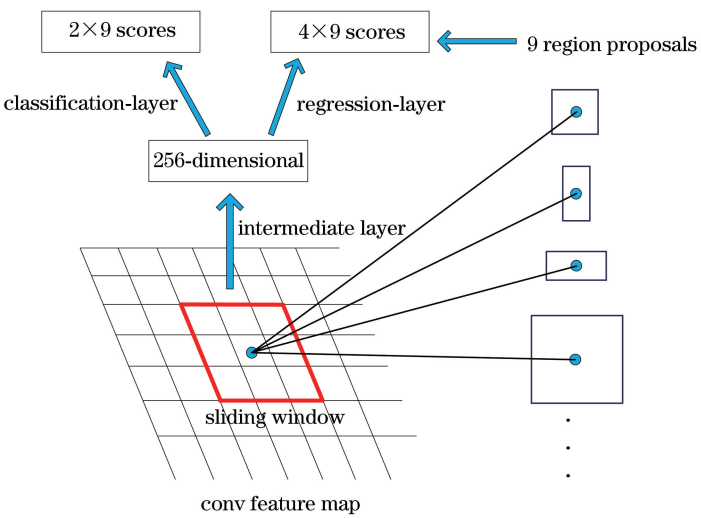

图 5 RPN 的结构图

Fig. 5 Structure of RPN

3.4 优化算法

由 Faster R-CNN 结构的特殊性，该框架采用交替训练的方式。训练过程为：

1) 用 ImageNet 预训练的模型初始化，端到端地微调 RPN 参数用于候选框提取；

2) 用步骤 1)生成的候选区域，由 Fast R-CNN 训练一个单独的检测网络，该网络同样是由 ImageNet 预训练的模型初始化；

3) 用步骤 2)训练好的 Fast R-CNN 再训练 RPN，但共享的卷积层保持不变，仅微调 RPN 独有的层，再次重新提取候选框；

4) 保持共享的卷积层不变，用步骤 3)新生成的候选区域再微调 Fast R-CNN 的其他层。

Fast R-CNN 的训练过程与文献[8]相同。为了训练 RPN，给每个候选区域分配一个二进制标签，正标签可以分配给两类候选区域：1) 与某个真实目标(GT)边界框有最高的交集并集之比(IoU)交叠的候选区域，IoU 比率可小于 0.7；2)与任意 GT 边界框有大于 0.7 的 IoU 交叠的候选区域。一个 GT 边界框可能分配正标签给多个候选区域，而负标签则只分配给与所有 GT 边界框的 IoU 比率都低于 0.3 的候选区域，非正

非负的候选区域对训练目标没有任何作用。

与 Fast R-CNN 一样，RPN 训练时也遵循多任务损失，最小化目标函数的原则。一张图像的 RPN 损失函数定义为

$$L(p_i,\boldsymbol{t}_i)=\frac{1}{N_{\mathrm{cls}}}\sum_i L_{\mathrm{cls}}(p_i,p_i^*)+\lambda\frac{1}{N_{\mathrm{reg}}}\sum_i p_i^* L_{\mathrm{reg}}(\boldsymbol{t}_i,\boldsymbol{t}_i^*),\tag{1}$$

式中 i 为一个训练批量中候选区域的索引，p_i 是第 i 个候选区域为目标的预测概率。如果候选区域为正，GT 标签 p_i^* 则为 1，反之，p_i^* 为 0。$\boldsymbol{t}_i$ 是一个向量，即 $\boldsymbol{t}_i=(t_x,t_y,t_w,t_h)$，表示预测的边界框的 4 个参数化坐标，$\boldsymbol{t}_i^*$ 是与正候选区域对应的 GT 边界框的坐标向量，即 $\boldsymbol{t}_i^*=(t_x^*,t_y^*,t_w^*,t_h^*)$。$\lambda$ 为平衡权重，此处取为 10，N_{cls} 为训练批量的大小，即 256，N_{reg} 为候选区域的数量，约为 2000。分类损失 L_{cls} 是两个类别（目标和非目标）的对数损失，即：$L_{\mathrm{cls}}(p_i,p_i^*)=-\ln[p_i^*p_i+(1-p_i)(1-p_i^*)]$。对于回归损失 L_{reg}，由 $L_{\mathrm{reg}}(\boldsymbol{t}_i,\boldsymbol{t}_i^*)=R(\boldsymbol{t}_i-\boldsymbol{t}_i^*)$ 来计算，R 为具有稳健性的损失函数。$p_i^*L_{\mathrm{reg}}$ 意味着只有正候选区域，即 $p_i^*=1$ 时才有回归损失，其他情况下，即 $p_i^*=0$ 时无回归损失。

上述的参数化坐标向量 $\boldsymbol{t}_i$ 和 $\boldsymbol{t}_i^*$ 定义为

$$\begin{cases}t_x=(x-x_a)/w_a,\ t_y=(y-y_a)/h_a,\ t_w=\lg(w/w_a),\ t_h=\lg(h/h_a)\\ t_x^*=(x^*-x_a)/w_a,\ t_y^*=(y^*-y_a)/h_a,\ t_w^*=\lg(w^*/w_a),\ t_h^*=\lg(h^*/h_a)\end{cases},\tag{2}$$

式中 x、y、w、h 分别为预测边界框的中心坐标(x,y)、宽和高；x_a、y_a、w_a、h_a 分别为候选区域边界框的中心坐标(x_a,y_a)、宽和高；x^*、y^*、w^*、h^* 分别为 GT 边界框的中心坐标(x^*,y^*)、宽和高。$\boldsymbol{t}_i$ 和 $\boldsymbol{t}_i^*$ 用于计算回归损失，可以理解为从候选区域边界框到附近的 GT 边界框的边界框回归。

根据定义，多任务损失函数同 Fast R-CNN 等深度学习算法一样，采用随机梯度下降（SGD）优化算法，求得最优的权重参数。

4 实验及结果分析

4.1 实验配置

机器软硬件配置为：中央处理器（CPU）Intel i5-6600，内存 16 GB，图形处理器（GPU）NVIDIA GTX 1070，操作系统 Ubuntu 14.04，统一计算设备架构 CUDA8.0。使用深度学习中的 Caffe 框架，并且参考了该开源项目上一些层次结构的具体实现。

4.2 实验数据

使用的实验图片由实验室载有红外摄像机的机器人在夜晚拍摄所得。其中训练集为 2000 张，测试集为 500 张，大小为 720 pixel×576 pixel，按照格式命名图片文件并制作图片名称列表，通过 Python 编写程序，标出每张图片中真实目标的边界，并将坐标位置等信息记录到.xml 文件中。

4.3 实验步骤

基于 Faster R-CNN 框架的开源网络结构有 VGG_CNN_M_1024、ZF 和 VGG16 这三种，其网络结构的层次依次加深。VGG_CNN_M_1024 层次偏浅，虽然训练速度和测试速度都很快，但是不能完全地学习图像中的特征，影响了检测的正确率。VGG16 层次偏深，导致训练和测试的速度很慢，且对 GPU 的要求很高，导致检测无法实现实时性；此外，虽然 VGG16 能很好地学习图像特征，但是其深层次也容易使得训练达到过拟合。因此，采用层次居上述两者之中的 ZF 网络结构，能很好地权衡速度和正确率。

根据上文所述，训练过程分为 4 步，4 个步骤设置的迭代训练次数分别为 40000、20000、40000 和 20000。此外，4 个步骤都采用固定的学习率 0.001，其他参数对训练和测试几乎没有影响，因此均采用默认数值。将图片以及记录有真实目标位置的.xml 文件作为训练集输入给配置好上述参数的 ZF Faster R-CNN 进行训练，训练时间约为 2 h，训练结束后保存模型参数，并进行测试验证，实验流程如图 6 所示。

为了理解该网络框架的迭代学习，在训练结束后，绘制若干卷积层的特征图。图 7(a)、(b)分别为训练后参数共享的第 2 和第 5 卷积层的特征图，图 7(c)为 RPN 特有的卷积层特征图。由于红外图像没有色彩信息，若特征图中场景比较完整，则该卷积核学习的是图像的灰度信息；若特征图中内容多为线条，则该卷积核学习的是图像纹理信息。这些特征均由卷积网络自动学习得到。

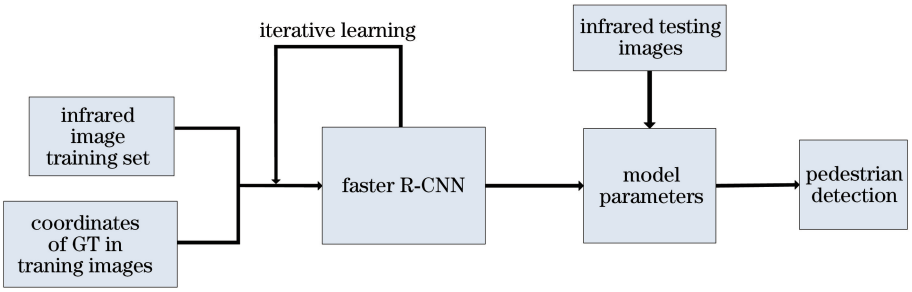

图 6 实验流程图

Fig. 6 Flow chart of experiment

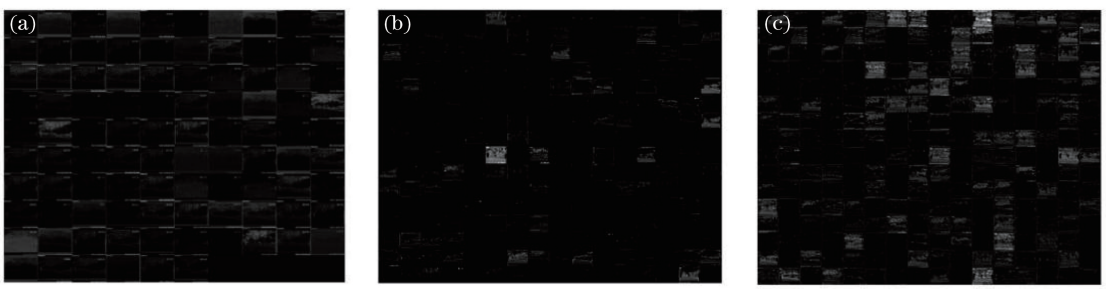

图 7 卷积层特征图。(a)共享卷积层的第 2 层特征图;(b)共享的第 5 层特征图;(c)RPN 特有卷积层的特征图

Fig. 7 Feature maps of convolution layer. (a) Feature map of the 2nd shared convolution layer; (b) feature map of the 5th shared convolution layer; (c) feature map of the peculiar convolution layer of RPN

4.4 结果与分析

对训练得出的模型输入一段视频进行测试,其中部分帧的检测结果如图 8 所示。可见,所有帧中的行人目标都能够完整地检测出来,且概率很大,目标边界恰好框出目标,此外,由于 RPN 用了三种不同的尺度,因此,图像中不同距离的行人目标也能用不同大小的边界框标出。然而,由图 8(d)可以看出,在红外图像上采用 Faster R-CNN 算法也存在一些误检测,这是因为深度学习在自动提取特征时,会将一个亮度很大的圆形区域视为人头,并将其下连接的区域视为身体和腿,因此出现了图中的误检测。

图 8 训练所得模型测试结果

Fig. 8 Testing results of trained model

使用传统的 HOG+LBP+AdaBoost 算法和 Fast R-CNN 算法在同样的视频上进行测试,并对三种算法结果进行了对比,如表 1 所示。由表 1 的对比结果可见,传统的机器学习方法在检测速度(3.60 s/frame)和正确率(0.682)上都低于采用深度学习的方法,这是由于传统方法在计算各种行人特征以及窗口遍历图像时需要进行大量计算,尽管如此,各种特征提取器学习的特征却依然没有深度学习的完全。Fast R-CNN 算法相比传统方法虽然有较大改善,但是其单独使用选择性搜索的方式生成候选区域,计算量较大,速度相对较慢(2.30 s/frame),造成视频的行人检测不具有实时性。而通过使用 Faster R-CNN 框架中的卷积网络生成候选区域,是一种端到端的处理方式,速度上有了极大的提高(0.04 s/frame),保证了视频中行人检测的实时性,此外,所使用的框架将候选区域生成、区域分类检测以及位置精修综合到一起,增强了它们之间的关联性,从而很好地将目标检测的正确率从 Fast R-CNN 的 0.734 提高到 0.909。因此,在红外图像上使用 Faster

R-CNN 框架检测行人,其检测速度和检测精度都达到了预期的要求。

表 1 三种算法结果对比

Table 1 Results contrast of three algorithms

Method	Time /(s·frame^{-1})	Mean average precision
HOG+LBP+AdaBoost	3.60	0.682
Fast R-CNN	2.30	0.734
Faster R-CNN	0.04	0.909

5 结 论

使用 Faster R-CNN 框架实现了夜间实时行人检测。该框架使用 RPN 代替传统的选择性搜索方法来生成候选区域,再将这些候选区域输入给 Fast R-CNN 网络进行分类检测,在分类检测网络中依旧采用 ROI 采样层从原图特征图中映射出这些候选区域的特征图。框架中的 RPN 与 Fast R-CNN 均采用多任务损失更新权重,通过卷积层参数共享机制形成一种端到端的框架结构,减少了训练的参数。训练时,采用交替的方式进行,使得网络权重能够以最快的学习速度达到最优。结果表明,通过该网络框架能够很好地实现夜间行人检测的高正确率(0.909)和实时性(0.04 s/frame),与其他方法相比,检测的精度和速度都有了极大的提高。将该框架引入红外图像目标检测的应用中,对推动机器人和无人车领域发展有极大的作用。

参 考 文 献

[1] Qin Jian, Wang Meihua. Fast pedestrian proposal generation algorithm using online Gaussian model[J]. Acta Optica Sinica, 2016, 36(11): 1115001.
覃 剑, 王美华. 采用在线高斯模型的行人检测候选框快速生成方法[J]. 光学学报, 2016, 36(11): 1115001.

[2] Xu Lu, Zhao Haitao, Sun Shaoyuan. Monocular infrared image depth estimation based on deep convolutional neural networks[J]. Acta Optica Sinica, 2016, 36(7): 0715002.
许 路, 赵海涛, 孙韶媛. 基于深层卷积神经网络的单目红外图像深度估计[J]. 光学学报, 2016, 36(7): 0715002.

[3] Xu Xin, Sun Shaoyuan, Sha Yujie, *et al*. A method of infrared image mosaic based on improved RANSAC[J]. Laser & Optoelectronics Progress, 2014, 51(11): 111001.
徐 鑫, 孙韶媛, 沙钰杰, 等. 一种基于改进 RANSAC 的红外图像拼接方法[J]. 激光与光电子学进展, 2014, 51(11): 111001.

[4] Zou Fangyu, Sun Shaoyuan, Xi Lin, *et al*. Color stereo vision method of vehicular infrared images with depth perception[J]. Laser & Optoelectronics Progress, 2013, 50(1): 011101.
邹芳喻, 孙韶媛, 席 林, 等. 具有深度视觉感的车载红外图像彩色化方法[J]. 激光与光电子学进展, 2013, 50(1): 011101.

[5] Dalal N, Triggs B. Histograms of oriented gradients for human detection[J]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, 1(1): 886-893.

[6] Lin Chengzhu. Pedestrian detection based on improved boosted cascade and fusion of multiple features[D]. Xiamen: Xiamen University, 2010: 27-37.
林成竹. 基于改进的级联分类器和多特征融合的行人检测方法研究[D]. 厦门: 厦门大学, 2010: 27-37.

[7] Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[J]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580-587.

[8] Girshick R. Fast-RCNN[J]. Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440-1448.

[9] Uijlings J R R, Sande K E A, Gevers T, *et al*. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154-171.

[10] Ren S, He K, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[C]. Advances in Neural Information Processing Systems, 2015: 91-99.

[11] He K, Zhang X, Ren S, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition[C]. European Conference on Computer Vision, 2014: 346-361.