❖特约专栏❖

基于双模全卷积网络的行人检测算法(特邀)

罗海波[1,2,3,4],何 森[1,2,3,4],惠 斌[1,3,4],常 铮[1,3,4]

(1. 中国科学院沈阳自动化研究所,辽宁 沈阳 110016; 2. 中国科学院大学,北京 100049;
3. 中国科学院光电信息处理重点实验室,辽宁 沈阳 110016;
4. 辽宁省图像理解与视觉计算重点实验室,辽宁 沈阳 110016)

摘 要:在近距离行人检测任务中,平衡算法的检测精度与检测速度对于检测算法的实际应用有着重要意义。为了快速并准确地检测出近景行人目标,提出了一种基于模型融合全卷积网络的行人检测算法。首先,通过全卷积检测网络对图像中的目标进行检测,得到一系列候选框;其次,通过弱监督训练的语义分割网络得到图像的像素级分类结果;最后,将候选框与像素级分类结果融合,完成检测。实验结果表明:算法在检测速度与精度方面都具有较高的性能。

关键词:深度学习; 弱监督训练; 行人检测; 语义分割

中图分类号:TP391.4 **文献标志码**:A **DOI**:10.3788/IRLA201847.0203001

Pedestrian detection algorithm based on dual-model fused fully convolutional networks(*Invited*)

Luo Haibo[1,2,3,4], He Miao[1,2,3,4], Hui Bin[1,3,4], Chang Zheng[1,3,4]

(1. Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. Key Laboratory of Opto-Electronic Information Processing, Chinese Academy of Sciences, Shenyang 110016, China;
4. The Key Lab of Image Understanding and Computer Vision, Liaoning Province, Shenyang 110016, China)

Abstract: In the task of close range pedestrian detection, the balance of the precision and speed were of great significance to the practical application of the detection algorithm. In order to detect the close range target quickly and accurately, a pedestrian detection algorithm based on fused fully convolutional network was proposed. Firstly, a fully convolutional detection network was used to detect the target in the image, and a series of candidate bounding boxes were obtained. Secondly, pixel level classification results of the image were obtained by using a semantic segmentation network with weakly supervised training. Finally, the candidate bounding boxes and the pixel level classification results were fused to complete the detection. The experimental results show that the algorithm has good performance in both the speed and the precision of detection.

Key words: deep learning; weakly supervised training; pedestrian detection; semantic segmentation

收稿日期:2017-08-10; **修订日期**:2017-10-28

作者简介:罗海波(1967-),男,研究员,博士生导师,博士,主要从事光电成像、图像处理、目标识别、目标跟踪等方面的研究。Email:luohb@sia.cn

通讯作者:何森(1992-),男,博士生,主要从事目标检测、目标跟踪、图像处理、模式识别等方面的研究。Email: hemiao@sia.cn

0 引 言

行人检测是目标检测技术的重要分支[1],是自动驾驶、机器人以及智能视频监控等研究领域的核心技术,行人检测有着重要的研究意义。近年来,这些技术引起了业界的广泛关注,吸引了大量学者加入其中。

为了应对行人检测任务,学者们主要提出了三类方法:基于决策森林的方法、基于 DPM(Deformable Part Model)的方法和基于卷积神经网络的方法[1]。其中,基于卷积神经网络的方法近年来得到了迅猛发展,由于深度卷积网络强大的特征提取能力,自 ImageNet 项目开始以来,越来越多的用于图像分类的网络结构被提出,不断提升了深度卷积网络在特征提取方面的优势,取得了相比于利用 HOG(Histogram of Oriented Gradient)等人工特征的传统方法[2]更高的检测精度。

由于深度学习方法需要大量的数据以及很长的训练时间才能收敛,因而迁移学习方法成为解决这一问题的一种有益的补充;Yosinski 等人发现,即便是微调从其他任务迁移来的权重也会比训练随机初始化的权重获得更好的特征提取能力[3]。在新的特征提取网络结构和迁移学习的帮助下,深度学习方法在更多的高级视觉任务中获得了成功,这些高级视觉任务包括目标跟踪[4]、目标检测、语义分割、骨架提取等。

最早的实时传统行人检测算法是 VJ 算法[5],该算法最初用于人脸检测,在 2003 年时被作者用于行人检测,该算法利用积分图、Haar 特征以及 Adaboost 分类器,奠定了传统行人检测技术的发展基础。2005 年,Dalal 等提出了经典的利用 HOG 特征和 SVM 分类器的行人检测方法[6],成为行人检测研究领域的里程碑。2014 年,ACF 算法[7]利用多尺度多通道特征(HOG+LUV)、快速特征金字塔和 Adaboost 分类器,提升了检测的准确率,并在检测速度上达到了实时性要求。

最早的基于深度学习的目标检测方法是 Overfeat[8],该方法基于多尺度滑动窗口获得候选区域,并利用卷积网络进行分类和位置预测,精度上相比于传统方法有了很大提高,但速度只能达到 0.5 fps。在此基础上,出现了 JointDeep[9]等利用深度学习方法和传统方法相结合的算法。JointDeep 利用基于 HOG 特征的传统算法作为前端,并利用深度学习网络对前端的检测结果进行筛选,这种方法大大提高了行人检测的精度,但是检测速度仍然难以满足实时性要求。之后出现了更多基于候选区域的方法,如 R-CNN,fast R-CNN 以及 faster R-CNN[10]等。这些方法的检测精度和检测速度都得到了有效提高,但其中最快的 faster R-CNN 在以 VGG[11]网络为基础网络的条件下也只能达到 5 fps,检测速度依然是深度学习网络在实际应用中的重要制约因素。Angelova 等人提出了一种基于级联的深度学习行人检测算法 DeepCascade[12],首先利用一个小型的卷积网络对整幅图片中的大量图像块进行筛选,再利用一个更深的卷积网络对筛选出的信任度较高的图像块进行评价,该算法兼顾了深度神经网络的精度与级联分类器的效率。Redmon 等人于 2015 年提出的 YOLO[13]网络结构不再采用基于候选框的方式,而是采用回归的方式寻找目标边界框,极大提高了检测速度,但是在检测精度上有所降低。

语义分割是一种像素级的标注方法,较为经典的用于语义分割的网络结构是 Long 等人提出的 FCN (Fully Convolutional Networks)[14]网络。FCN 网络利用包括 VGG-16,ResNet[15]等经典网络的卷积结构提取不同尺度的特征图,而这些经典网络后端的全连接结构被替换为反卷积结构用于生成稠密的像素级标签,这些像素级的标签反映了输入图中各个像素的分类信息。将语义分割得到的像素级分类信息与目标检测得到的候选框融合,可以一定程度上提高目标检测的检测精度。

在行人检测任务中,相比于远景行人目标检测,近景目标的检测在自动驾驶,智能监控等方面有着更重要的意义,但检测速度也是制约算法推广到实际应用中的一个重要因素,因此平衡检测精度与检测速度是目前亟待解决的难题之一。

文中针对提高行人检测中近景目标的检测速度与精度的需求,提出了一种检测网络与语义分割网络相融合的网络结构用于行人检测。算法通过检测模型与语义分割模型的融合,在保证较高的检测精度的前提下,获得了优于绝大部分传统算法的检测速度。在加州理工行人检测数据集上对算法进行了评估,实验结果表明,论文提出的算法在对近景目标(高度在 100 pixel 以上)的检测精度与速度方面获得了较高的性能。

1 基于全卷积网络的行人检测算法

1.1 基于回归的全卷积网络

基于回归的全卷积网络的整体结构如图 1 所示，该网络是在 YOLO 网络结构的基础上改进而来的，旨在在保证检测速度不受影响的前提下提升网络的检测精度。网络将整幅图像划分为 15×15 的网格，在每一个格子中预测两个目标候选框并给出相应的置信度。每个格子预测的候选框的中心都落在对应的格子内，因此每张图像都会预测出 450 个候选框。

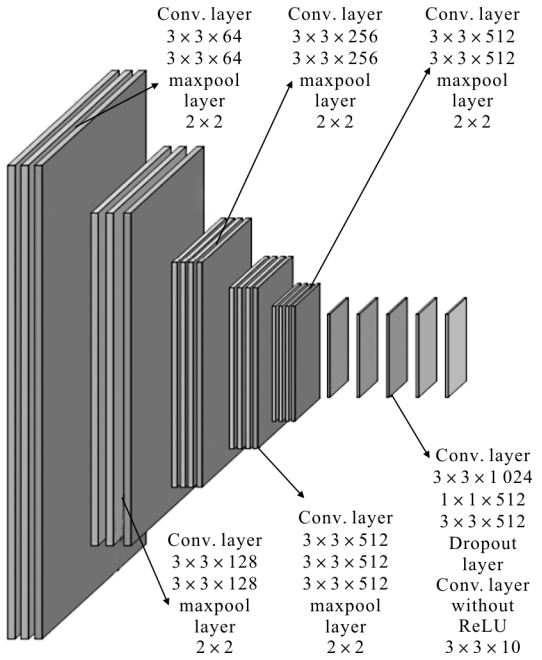

图 1 检测网络整体结构

Fig.1 Structure of the detection network

网络的前半部分利用 VGG 网络进行特征提取，输入图像的大小为 480×480。完整的 VGG 网络要求输入图像尺寸为 224×224，故笔者使用全卷积结构，去除 VGG 网络 pool5 之后的所有全连接层，藉此去除对输入图像的尺寸限制，这种修改并不会影响 VGG 网络的特征提取能力。图像经过特征提取后，得到了 pool5 层输出的大小为 15×15 的特征图。在 pool5 之后加入一层 1 024 通道的 3×3 的卷积层，进而加入一层 512 通道的 1×1 卷积层进行降维和进一步融合。经过一层 dropout 层抑制过拟合之后，利用 10 通道的 3×3 卷积核输出一个 10 通道 15×15 的输出。输出的特征图上每一个点的位置 $\mathrm{loc} \in \Omega(\Omega \subset Z^2)$ 都对应图像划分的 15×15 的一个网格区域，10 个通道分别对应网络预测的两个候选框的置信度 c、中心位置(x,y)，宽度 w 和高度 h。

1.2 Loss 函数

计算代价前，先将数据集的标签转换为更适合于网络的格式，由于网络输出结果以每张图的网格为单位，每张图 450 个小格，输出结果以(15,15,10)的张量表示。将数据集标签同样转换成张量形式，张量形状为(15,15,5)，5 个通道分别对应网格位置上是否有边界框，边界框的中心位置(x,y)，宽度 w 和高度 h。数据集标签的格式为“图像序号，目标边界框左上角 x 坐标，目标边界框左上角 y 坐标，目标边界框宽度(pixel)，目标边界框高度(pixel)”。将目标边界框的左上角坐标转换为目标中心坐标，并将这一边界框标注到标签中目标的中心坐标对应的网格位置$(\mathrm{loc}_x, \mathrm{loc}_y)$上。首先将这一位置标签的第一维置$(\mathrm{loc}_x,\mathrm{loc}_y, 0)$为 1，表示有目标边界框中心落在此位置对应的网格区域。进而将中心坐标映射为中心坐标相对于这一网格左上角坐标的偏移量，并利用网格的像素宽度与高度进行归一化。最后将边界框的宽和高利用整张图像的宽和高进行归一化。如果标签网格位置$(\mathrm{loc}_x, \mathrm{loc}_y)$上并没有标注到任何边界框，将其第一维$(\mathrm{loc}_x, \mathrm{loc}_y, 0)$置为 0，$(\mathrm{loc}_x,\mathrm{loc}_y, 1{:}4)$置为 1。此时得到了这一张图像的标签张量，张量上所有值都已经被归一化。

检测网络的输出结果共分为三部分，分别为置信度，中心位置以及候选框大小。对于每一个标签边界框，选择与其中心落在同一预测网格中且 IOU (Intersection Over Union)更大的预测边界框，对其置信度、中心位置和宽高进行回归预测，训练过程中也只对这些预测边界框的中心位置和宽高进行训练。而对于其他预测框，只训练其置信度，置信度的代价函数选择 sigmoid 交叉熵函数。

由于将数据映射到网格后，大部分预测框没有对应的标签框，正负样本极不平衡，引入 focal loss[16] 的方法平衡代价函数。利用 focal loss 可以使得已经被正确分类的样本对整体代价函数的贡献更小，让网络集中于训练未能正确分类的样本，抑制正负样本不平衡带来的影响。对于中心位置直接选用差的平方进行回归，且只训练有对应标签框的预测框参与代价函数计算。

对于候选框大小 (即候选框的宽和高)，选择与中心位置相似的方式进行回归，然而由于相同的预测误差对于大目标的代价应该小于小目标的代价，如图 2(a)所示，所以将代价中的宽和高替换为宽和高的二次方根来达到这样的目的，其效果如图 2(b)所示。

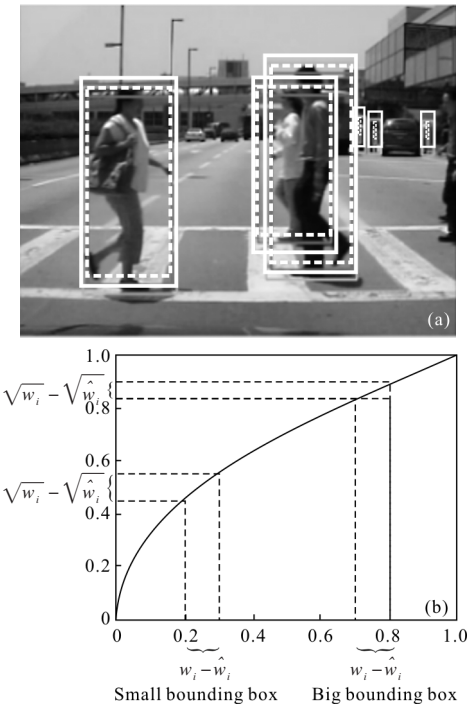

图 2 不同尺度目标在同预测误差下代价示意图

Fig.2 Schematic diagram of the loss of targets in different cales with the same prediction error

整体的代价函数可表示为：

$$\sum_{I=0}^{S^2}\sum_{j=0}^{B} if_{ij}^{\text{obj}}[-(1-c)^2\log(c)]+$$

$$\sum_{I=0}^{S^2}\sum_{j=0}^{B} if_{ij}^{\text{noobj}}[-c^2\log(1-c)]+$$

$$\sum_{I=0}^{S^2}\sum_{j=0}^{B} if_{ij}^{\text{obj}}[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2]+$$

$$\sum_{I=0}^{S^2}\sum_{j=0}^{B} if_{ij}^{\text{obj}}[(\sqrt{w_i}-\sqrt{\hat{w}_i})^2+(\sqrt{h_i}-\sqrt{\hat{h}_i})^2] \quad (1)$$

1.3 非极大值抑制

由于在检测网络中，每个网格中预测两个候选框可能造成两个候选框同时预测同一目标，如在图 3(a)所示的结果中，最左和最右两个目标上各有两个几乎重合的预测框。在对检测结果进行评价时，这两个几乎重合的预测框中会有一个被评价为未检测到目标，这将导致预测的精度降低。考虑到由两个候选框同时预测同一目标造成的候选框重叠度一般会大于由于相互遮挡造成的候选框重叠度，所以采用阈值较高的非极大值抑制方法来解决这一问题。如图 3(b)所示为选择 IOU 阈值为 0.75 的非极大值抑制方法的检测结果。实验结果表明，在此阈值下，非极大值抑制对检测结果的提升效果达到最佳，该方法较好地避免了对同一个目标的重复检测。

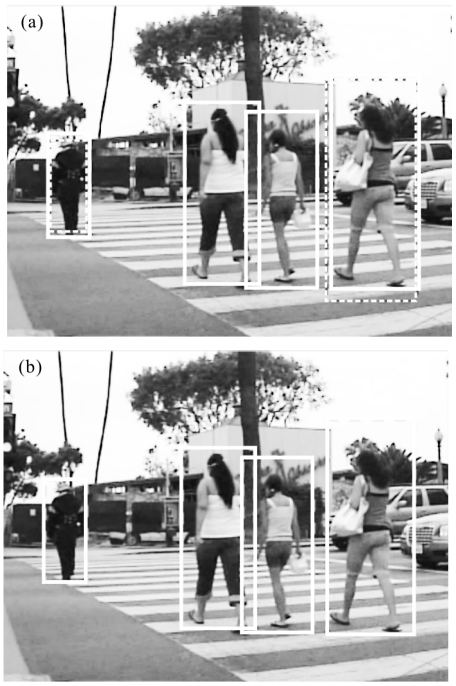

图 3 非极大值抑制

Fig.3 Non-maximum suppression

1.4 弱监督学习语义分割

一般的语义分割方法使用强监督学习，利用像素级标定的标签对神经网络进行训练，使得网络获得语义分割的能力。像素级标定需要付出大量的时间成本，而弱监督学习可以利用对数据集不完全可靠的标记(例如不充分的标记、不完全正确的标记或者局部标记)对网络进行训练，使得网络同样获得一定的语义分割能力[17]。而对于语义分割任务而言，边界框级标注就属于一种弱监督标注。文中利用目标边界框对 FCN 网络进行训练，使得网络获得一定的语义分割能力。

FCN 网络利用 VGG 网络的 pool5 层之前的卷积部分提取特征，之后通过 3 层卷积得到分辨率为原图 32 倍下采样的语义分割结果。由于高层特征包含的语义信息更强，而低层特征的定位信息更准，FCN 网络通过将低层的特征与高层特征不断融合可得到更准确的语义分割结果。首先将之前的 32 倍下

采样结果进行一次转置卷积,使得特征图放大 2 倍,并与 pool4 得到的特征图经过卷积后求和,融合到一起。进而再次将融合得到的特征图进行 2 倍上采样后与 pool3 得到的特征图经过卷积后求和。最后利用转置卷积进行 8 倍上采样得到最终的语义分割结果,网络结构如图 4 所示。

网络训练之前,需要对边界框标注信息进行处理,将其处理为可以用于语义分割训练的标注。将所有边界框内的像素点在标签中的对应位置标注为 1,边界框外的像素点对应位置标注为 0,如图 5 所示。在该任务中,笔者选择的输入图像大小为 480×480,故将输入图像与像素级标签的尺寸均调整为 480×480。与上文中提到的一样,全卷积网络由于没有全连接层,所以对输入图像的大小没有限制。

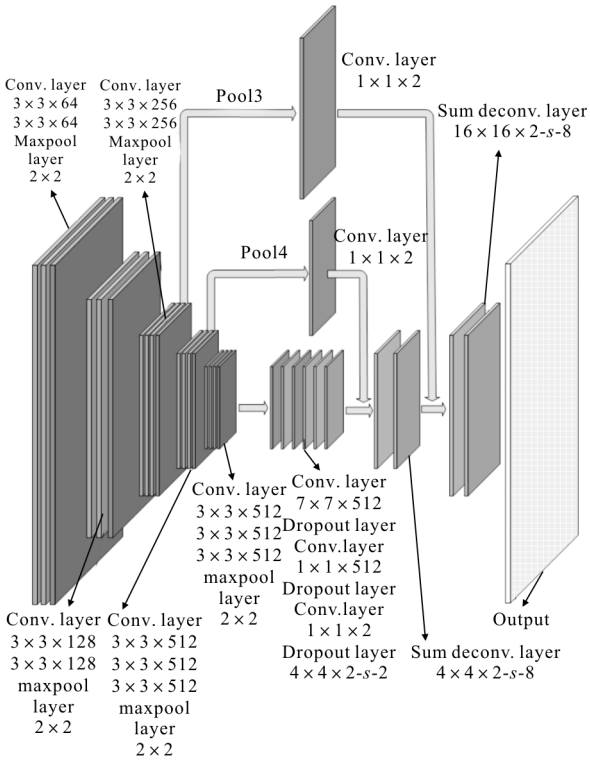

图 4 语义分割网络结构

Fig.4 Structure of semantic segmentation network

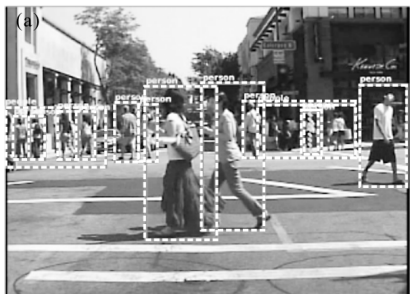

图 5 标签处理

Fig.5 Label processing

训练时的代价函数选择Softmax 交叉熵函数,Softmax 回归函数表示为:

$$p_k(x)=\frac{e^{\alpha_k(x)}}{\sum_{k'=1}^{K} e^{\alpha_{k'}(x)}} \tag{2}$$

式中:$\alpha_k(x)$特征图第 k 个通道在位置 $x\in\Omega(\Omega\subset Z^2)$ 的激活值。

Softmax 交叉熵可表示为:

$$L=\sum_{x\in\Omega}\log(p_{I_x}(x)) \tag{3}$$

式中:I_x为在位置 $x\in\Omega(\Omega\subset Z^2)$的真实标签。

1.5 模型融合

在获得检测与语义分割结果之后,通过模型融合的方式提高检测精度。通过基于回归的全卷积网络得到的检测结果中仍有部分边界框处于背景部分,增大了检测结果的错误率。而语义分割方式具有将前景与背景分割开来的作用,所以利用语义分割结果修正从检测网络获得的 450 个边界框的置信度,可以提升网络的整体检测精度。

笔者将语义分割的结果作为一层二进制语义掩模,1 和 0 分别对应前景部分和背景部分,进而将检测结果的位置映射到语义掩模上,通过语义掩模与检测得到的候选区域边界框的重叠度对边界框置信度进行调整,如图 6 所示。

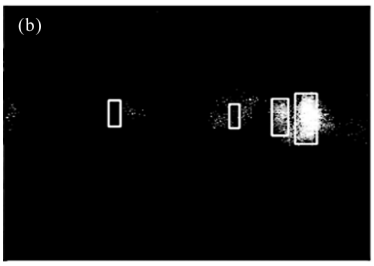

图 6 检测结果与掩模的融合方法

Fig.6 Fusion method of detection result and the mask

笔者使用以下策略进行调整：

(1) 当语义掩模与预测候选框的重叠度大于10%时，新的置信度为检测网络预测候选框置信度和重叠度的乘积；

(2) 当语义掩模与候选框重叠度小于 10%时，新的置信度为检测网络预测候选框置信度的 10%。

上述调整策略可表示为：

$$S_{\text{final}}=\begin{cases} s_{\text{det}}\times\dfrac{s_{\text{mark}}}{s_{\text{bounding}}} & \text{if } \dfrac{s_{\text{mark}}}{s_{\text{bounding}}}>0.1 \\ s_{\text{det}}\times 0.1 & \text{otherwise} \end{cases} \tag{4}$$

式中：s_{det} 为检测网络预测候选框置信度；s_{mark} 为语义掩模与预测候选框的重叠面积；s_{bounding} 为候选框面积。

通过实验，笔者将分界值取为 10%。利用这样的融合方法，既可以保证候选框的置信度范围仍然保持在区间，并按照比例自适应地对置信度进行调整，防止两个模型判定结果出现巨大差异时对整体结果造成影响，又可以避免由于掩模的漏检(候选框对应掩模区域全为 0)使得部分正确候选框置信度变为 0 而彻底失效的问题。

1.6 整体结构与训练

网络整体结构如图 7 所示，网络通过迁移学习的方式，载入利用 ImageNet 预训练过的 VGG 网络 pool5 以及之前卷积层的权重用于特征提取。首先采取对整个网络进行微调的方式对检测网络进行训练，检测网络训练完成后，冻结 pool5 之前的所有权重，只对 pool5 之后的语义分割网络各层权重通过随机梯度下降法进行微调，完成语义分割网络的训练。语义分割网络与检测网络共用相同的特征提取结构，最后将两者的结果进行模型融合。

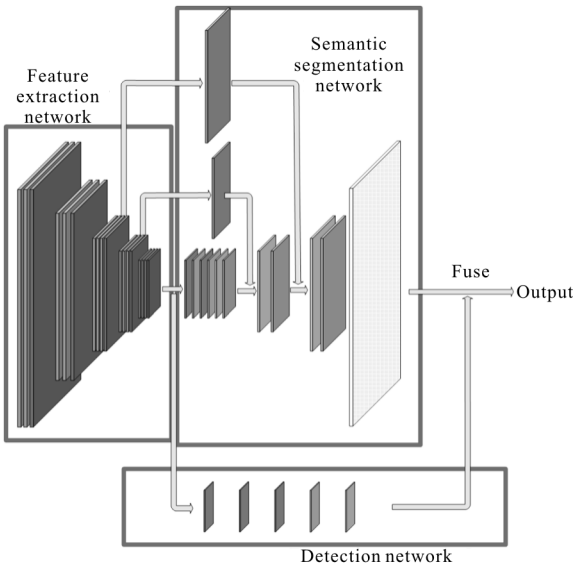

图 7 整体网络结构

Fig.7 Structure of the whole network

2 实验结果与分析

文中选用知名的加州理工行人检测数据集对算法进行训练与测试，该数据集共有 250 000 余张图像，350 000 余个目标框，图像尺寸均为 640 pixel×480 pixel。数据集通过装于机动车上的摄像头在城市环境正常交通行驶中采集，是自动驾驶研究领域最具影响力的数据集之一。训练数据包括 6 个数据集，每个数据集内包括 6~13 个 1 min 长的视频图像序列；测试数据共包括 5 个数据集，规模与训练数据集相似。

2.1 对于近景目标的检测效果对比

加州理工行人检测数据集的评价指标将数据集分为 18 个子集，针对该实验的实验目的，选择近景目标子集进行实验效果对比。其 ROC(Receiver Operating Characteristic)曲线对比结果如图 8 所示。

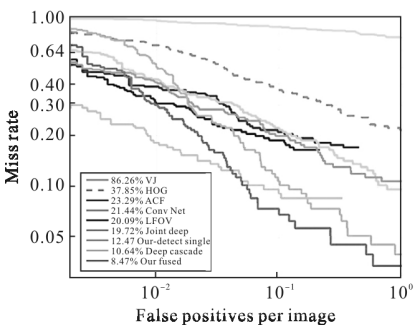

图 8 算法 ROC 曲线对比图

Fig.8 Comparison of ROC curves of algorithms

通过图 8 所示的曲线可以看出，文中提出算法的漏检率指标优于较流行的 ACF、HOG 等经典算法以及利用深度学习的 JointDeep、DeepCascade 等算法，在近景目标检测实验中，漏检率低于 10%。这些算法在具体场景中的检测结果如图 9 所示。可以看出，笔者提出的算法漏检的目标和误检的预测框均少于对比算法。与此同时，利用融合模型可以使得算法的漏检率从 12.47%(our-detect single)降低到8.47%(our-fused)，可以看出，融合模型能够大幅提高算法的检测准确率。

图 9 算法实际效果对比图

Fig.9 Comparison of the effect of the algorithms

2.2 检测速度分析

笔者比较了几种经典算法和文中算法在近景目标(高度大于 100 pixel 的目标)的检测精度与检测速度，并绘制成散点图，如图 10 所示。

从图 10 中可以看出，文中算法更好地平衡了算法的检测精度与速度，在漏检率低于 10%的前提下，检测速度达到 17.07 fps，检测准确率最优；检测速度优于大部分算法，虽然不及 ACF 算法，但漏检率比 ACF 算法下降了近 15 个百分点，综合性能优于目前国际上的先进水平。

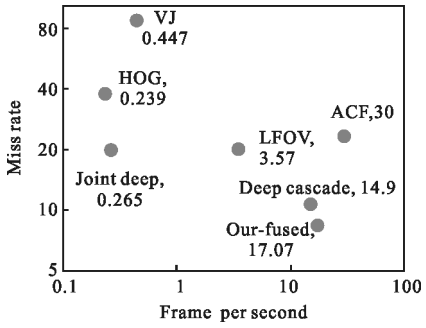

图 10 各算法检测速度对比

Fig.10 Comparison of detection speed of each algorithm

3 结 论

文中提出了一种近景行人检测算法，该算法在同一全卷积深度学习网络中，分别利用全卷积目标检测网络和弱监督反卷积语义分割网络对待检测图像进行处理；与此同时，文中还提出了一种模型融合方法，通过对网络检测和分割两方面的输出结果进行融合，提升了算法的精度。文中算法在加州理工行人检测数据集上进行的实验结果表明，在近实时的检测速度下，漏检率低至 8.47%，近景目标检测的综合性能优于当前国际上的先进水平。

在文中算法的基础上，还可以在检测网络中通过利用语义信息丰富的高层特征图与感受野较小的低层特征图融合进行检测，提高算法的多尺度适应能力，使得算法同时适用于近景行人检测和远景行人检测。

参考文献：

[1] Benenson R, Omran M, Hosang J, et al. Ten years of pedestrian detection, what have we learned [C]//European

Conference on Computer Vision, 2014: 613-627.

[2] Zhang Difei, Zhang Jinsuo, Yao Keming, et al. Infrared ship-target recognition based on SVM classification [J]. ***Infrared and Laser Engineering***, 2016, 45(1): 0104004. (in Chinese)

[3] Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks [C]//Advances in Neural Information Processing Systems, 2014: 3320-3328.

[4] Luo Haibo, Xu Lingyun, Hui Bin, et al. Status and prospect of target tracking based on deep learning [J]. ***Infrared and Laser Engineering***, 2017, 46(5): 0502002. (in Chinese)

[5] Viola P, Jones M J. Robust real-time face detection [J]. ***International Journal of Computer Vision,*** 2004, 57 (2): 137-154.

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

[7] 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.

[8] Sermanet P, Eigen D, Zhang X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks [C]//International Conference on Learning Representations, 2014, arXiv preprint arXiv: 1312.6229v4.

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

[10] 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***, 2017, 39(6): 1137-1149.

[11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [C]//Computer Vision and Pattern Recognition, 2014, arXiv preprint arXiv: 1409.1556.

[12] Angelova A, Krizhevsky A, Vanhoucke V, et al. Real-time pedestrian detection with deep network cascades [C]//BMVC, 2015, 32: 1-12.

[13] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[14] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation [J]. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 2017, 39(4): 640-651.

[15] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[16] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection [C]//Computer Vision and Pattern Recognition, 2017, arXiv preprint arXiv: 1708.02002.

[17] Khoreva A, Benenson R, Hosang J, et al. Simple does it: Weakly supervised instance and semantic segmentation [C]// Computer Vision and Pattern Recognition, 2016, arXiv preprint arXiv: 1603.07485.