基于卷积神经网络和 RGB-D 图像的车辆检测算法

王得成[1]，陈向宁[2*]，赵峰[1,3]，孙浩燃[4]

[1]航天工程大学研究生院，北京 101416；

[2]航天工程大学航天信息学院，北京 101416；

[3] 61618 部队，北京 100094；

[4]酒泉卫星发射中心，甘肃 酒泉 730000

摘要 针对利用彩色图像进行车辆检测时会受到路面阴影、车辆反光和光线不足等复杂情况影响的问题，提出一种基于卷积神经网络并融合彩色与深度图像的车辆检测算法。设计单通道 RG-D 融合网络和双通道 RGB-D 融合网络两种改进模型，分别用于提高检测速度和准确度。实验使用 GTA(Grand Theft Auto)车辆数据集对该算法进行测试，并与基于 RGB 图像的其他流行算法进行对比和分析，结果表明：与基于彩色图像的 Yolo v2 算法相比，利用双通道 RGB-D 融合网络检测的准确率和召回率分别提升 5.69%和 6.31%，利用单通道 RG-D 融合网络对单一图像的最快检测速度达到 24 ms。实验证明，基于 RGB-D 图像的改进网络模型能够实现实时检测，并有效提高车辆检测精度。

关键词 图像处理；车辆检测；计算机视觉；卷积神经网络；RGB-D 图像

中图分类号 TP391.4 **文献标识码** A **doi**：10.3788/LOP56.181003

Vehicle Detection Algorithm Based on Convolutional Neural Network and RGB-D Images

Wang Decheng[1], Chen Xiangning[2*], Zhao Feng[1,3], Sun Haoran[4]

[1] *Graduate School, Space Engineering University, Beijing* 101416, *China*;

[2] *School of Space Information, Space Engineering University, Beijing* 101416, *China*;

[3] 61618 *Troops, Beijing* 100094, *China*;

[4] *Jiuquan Satellite Launch Centre, Jiuquan, Gansu* 730000, *China*

Abstract Aiming at the problem that using RGB images for vehicle detection are affected by complex conditions such as road shadow, vehicle reflection and insufficient light. The paper proposes a vehicle detection algorithm based on convolutional neural network and combination of RGB and depth images. Two improved models of single-channel RG-D and double-channel RGB-D fusion networks are designed to improve detection speed and accuracy respectively. The algorithm is tested with (Grand Theft Auto) vehicle dataset and compared with other popular algorithms based on RGB images. The results show that compared with Yolo v2 algorithm based on RGB images, detection accuracy and recall rates increase 5.69% and 6.31% respectively by double-channel RGB-D fusion network, and the fastest detection speed of single image reaches 24 ms with single-channel RG-D fusion network. Experiments show that the improved network model based on RGB-D images can achieve real-time detection and effectively improve vehicle detection accuracy.

Key words image processing; vehicle detection; computer vision; convolutional neural network; RGB-D images

OCIS codes 100.4996; 100.2980; 100.4999

收稿日期：2019-02-25；**修回日期**：2019-03-14；**录用日期**：2019-04-01

基金项目：国防科技创新特区专项(18-H863-01-ZT-002-055)

* **E-mail**：18810836867@163.com

1 引　　言

目标检测与识别是当前计算机视觉领域中一个重要的研究方向[1]。车辆检测属于目标识别的范畴，早期的图像识别系统主要采用尺度不变特征变换(SIFT)[2]和方向梯度直方图(HOG)[3]等传统的特征提取方法，将提取到的特征输入至支持向量机(SVM)[4]等分类器中进行分类识别。这些图像识别系统一般都是针对某个特定的识别任务，且数据规模不大，泛化能力较差，难以在实际应用中实现精准的识别效果[5]。

2006 年，Hinton 在文献[6]中首次提出深度学习的思想。到目前为止，基于深度学习的 RGB 图像目标识别技术已经取得巨大成果。但在实际应用过程中，RGB 图像在目标重叠、遮挡及光照不足等复杂场景下，存在目标识别率低、场景分类效果不佳及稳健性差等问题[7]。为解决这些困难，近几年国内外学者利用 RGB-D 图像对目标检测和场景分类进行相关研究。所谓 RGB-D 图像，就是在原来 RGB 图像的基础上增加一维深度(Depth)信息，由于深度不受亮度和颜色变化的影响，因此深度数据可以在复杂场景下提供有用的额外信息，并且提高光照变化情况下的目标识别准确率及稳定性。

Socher 等[8]在 2012 年首先提出卷积递归神经网络(CNN-RNN)融合模型来学习 RGB-D 特征，最后采用 softmax 分类器获得较好的分类性能。但该方法对深度图像缺失的信息未进行恰当的处理，影响后续目标识别的准确度。Eitel 等[9]将基于监督学习的深度卷积神经网络用于 RGB-D 物体识别任务中，对 Depth 图像进行 jet 彩色映射处理，将训练好的 CaffeNet 模型作为网络的预训练参数，然后使用 RGB-D 数据集进行调优训练。这种方法在当时的 RGB-D 目标检测算法中准确率最高达到91.3%，但是在 Depth 图像的 jet 彩色映射中，彩色化的效果依赖于距离信息，并不能完全反映目标物体的外观形状。李威[10]提出的基于混合结构的 RGB-D 目标识别算法中，分别利用深层卷积神经网络(VGG-16)和改进定向法向量直方图(HONV)描述子的 Fisher Vector 编码对每一帧的 RGB 图像与深度图像提取特征表达，虽然取得了非常优异的识别性能，但 SVM 在分类过程中计算量过大，导致识别速度缓慢。Xu 等[11]在 2017 年提出一种 RGB-D 场景下目标识别的新方法，利用共享权重策略和无参数相关层进行 RGB-D 对象检测和区域对象识别，通过后期融合，多模态 RGB-D 物体识别效果得到有效提升。2018 年刘帆等[12]设计了双流卷积神经网络，在卷积层根据最优权值对两个网络进行融合，实现 RGB-D 图像的联合检测，其准确率和成功率相比早期和后期融合分别提高 4.1%和 3.5%。

目前利用深度学习方法进行车辆检测存在两个问题，1)彩色图像的车辆检测在路面阴影、车面反光及光照不足等情况下存在误检与漏检[13]；2)基于 RGB-D 图像的车辆检测过程中彩色与深度信息融合不充分导致网络模型检测效果不佳。故本文提出双通道 RGB-D 融合网络(D-RGBD)和单通道 RG-D 融合网络(S-RGD)两种结构，分别用于提升检测精度和速度。使用 GTA(Grand Theft Auto)汽车数据集大约 1 万张 RGB-D 图像对网络进行训练与测试，最后对实验结果进行比较评估和定量评价。

2 算法原理

以当前流行的检测算法 Yolo v2[14]为基础，改进其网络模型，实现对 RGB-D 图像的目标检测。Yolo 算法对目标检测的原理如图 1 所示，首先将输入的图像分割成 $S\times S(S=7)$网格，然后每个单元格负责检测那些中心点落在该格子内的目标。每个单元格需预测多个边界框及边界框的置信度 C。边界框信息包含 5 个数据值，分别是 x，y，w，h 和 C，及 1 个类别概率，类别概率为当前网格预测得到物体的边界框所包含物体的车辆概率。(x,y)是指边界框的中心位置坐标；(w,h)是边界框的宽度和高度；C 反映当前边界框是否包含车辆及其位置的准确性，即检测边界框对其检测到车辆的置信度。

Yolov2 使用均方和误差作为损失函数(x_{loss})来优化模型参数，损失函数定义为预测数据与标定数据之间的坐标定位误差、交并比 R_{IOU} 误差和分类误差三项的和，即

$$x_{\text{loss}}=\lambda_{\text{coord}}\sum_{i=0}^{S^2}\sum_{j=1}^{B}I_{ij}^{\text{obj}}[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2]+\lambda_{\text{coord}}\sum_{i=0}^{S^2}\sum_{j=1}^{B}I_{ij}^{\text{obj}}[(\sqrt{w_i}-\sqrt{\hat{w}_i})^2+(\sqrt{h_i}-\sqrt{\hat{h}_i})^2]+$$
$$\sum_{i=0}^{S^2}\sum_{j=1}^{B}I_{ij}^{\text{obj}}(C_i-\hat{C}_i)^2+\lambda_{\text{noobj}}\sum_{i=0}^{S^2}\sum_{j=1}^{B}I_{ij}^{\text{noobj}}(C_i-\hat{C}_i)^2+\sum_{i=0}^{S^2}I_i^{\text{obj}}\sum_{c\in c_{\text{classes}}}[p_i(c)-\hat{p}_i(c)]^2,\tag{1}$$

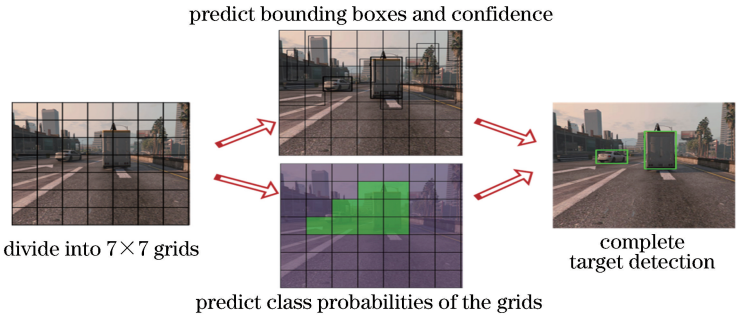

图 1 Yolo 目标检测模型

Fig. 1 Object detection model of Yolo

式中:λ 为权重,位置误差权重 $\lambda_{\text{coord}}=5$,置信误差权重 $\lambda_{\text{noobj}}=0.5$;$x$,$y$,$w$,$h$,$c$,$p$ 为网络预测值,其中 c 为预测的类别,p 为预测框类别正确的概率,$\hat{x}$,$\hat{y}$,$\hat{z}$,$\hat{h}$,$\hat{c}$,$\hat{p}$ 为标注值;I_i^{obj}、I_{ij}^{obj} 和 I_{ij}^{noobj} 分别为 x_{loss} 函数表达式的权值系数,表示若物体落入检测网格 i 中,则 I_i^{obj} 取值为 1,若物体落入格子 i 的第 j 个边界框内,则 I_{ij}^{obj} 取值为 1,同理若物体未落入格子 i 的第 j 个边界框内,则 I_{ij}^{noobj} 取值为 1;S^2 为网格总数,B 为边界框总数,c_{classes} 为目标的类别集合。

2.1 S-RGD 结构

S-RGD 是指数据输入端将原来 RGB 图像的最后一个通道换作深度信息,融合成为 RGD 数据。利用这种方法给原始图像补充额外的深度信息,可压缩卷积网络复杂度,在不改变检测速度的前提下,增加复杂环境中车辆检测的准确率,适合应用于实时检测系统。

Yolov2 算法使用的是只包含卷积层和池化层的 DarkNet-19 卷积网络[14],在网络中最后用卷积层替代全连接层,这主要是为了在图像中有多个类别物体时,卷积层可以利用滑动窗口,减少重复卷积的计算,使得检测网络更加高效[15]。但由于该网络层数较多,对于大场景中车辆的检测速度不佳,并且本文所检测的目标单一,因此针对数据集特点设计一种用于改善中小目标检测的卷积神经网络。如图 2所示,整个网络共有 11 层卷积层、5 层最大池化层及 2 层全连接层,最后通过 softmax 分类函数输出。所有卷积层和第一个全连接层都配有修正线性单元(ReLU)激活函数[16]。

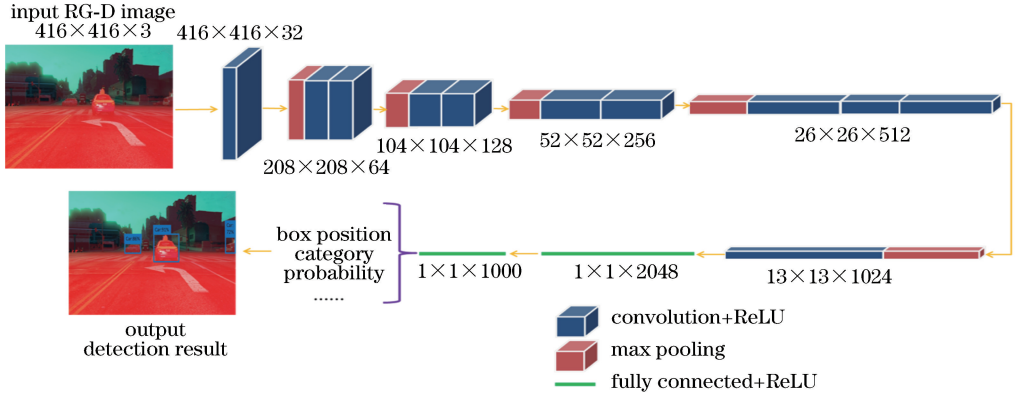

图 2 S-RGD 结构图

Fig. 2 S-RGD structure

2.2 D-RGBD 结构

D-RGBD 是指在卷积部分设计两个通道,分别为 RGB 通道和 Depth 通道,由于深度图像的像素值仅表示距离信息,表达内容较为单一,若使用与彩色图像一样的网络提取特征,必然会造成参数冗余。因此针对深度图像简化卷积层数,设计了由 7 层卷积层和 5 层池化层组成的卷积网络,使结构更加灵活紧凑。

D-RGBD 网络结构如图 3 所示,RGB 图像和 Depth 图像分别在 11 层和 7 层两个卷积神经网络中提取特征,将两类特征在第一个全连接层中融合,然后进行分类,输出检测结果图。这种双通道网络可以拓宽网络结构,结合深度图像检测车辆,有利于提升检测性能,但同时由于卷积层数的增加和两种特征的融合,使得检测速度略有下降。

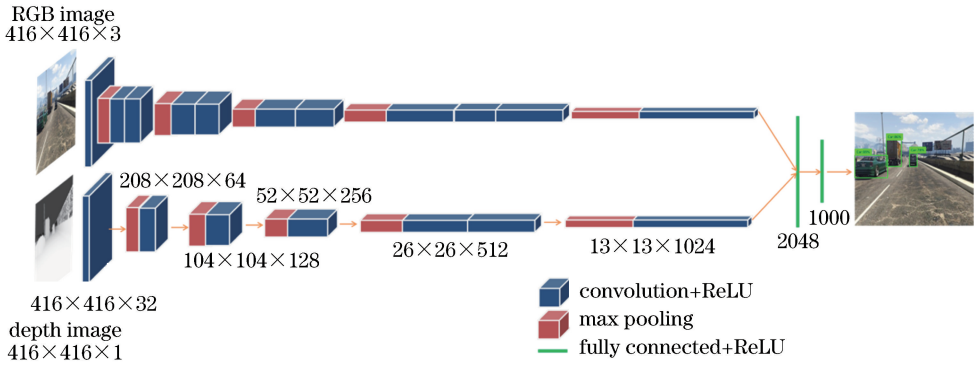

图 3 D-RGBD 结构图

Fig. 3 D-RGBD structure

3 实验过程与结果

3.1 RGB-D 数据预处理

实验使用 GTA 汽车数据集 1 万张 RGB-D 图片，其中训练图片 8800 张，测试图片 1200 张。预处理阶段将所有图片的尺度都缩放为 416 pixel×416 pixel，降低图像分辨率。另外原始深度图中的像素代表深度值，如图 4(b)所示，其对比度不强。为了增强可视化效果，预处理阶段用直方图均衡法简单调整深度图像的对比度，使目标信息更为突出，如图 4(c)所示。图 4(a)和(d)分别为原始 RGB 图像和改变通道的 RG-D 融合图像。

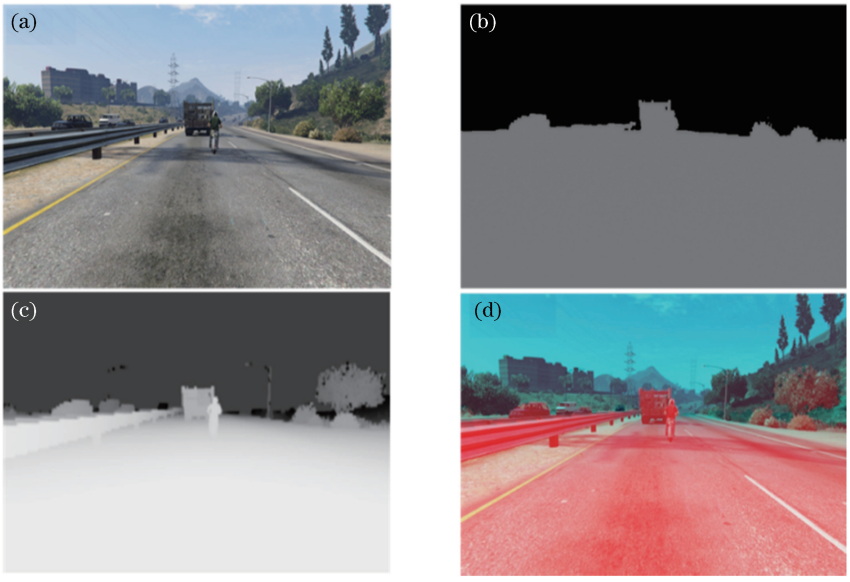

图 4 实验预处理图像。(a)原始 RGB 图像；(b)原始深度图像；(c)对比度增强后的深度图像；(d)改变通道的 RG-D 融合图像

Fig. 4 Experimental preprocessed images. (a) Original RGB image; (b) original depth image; (c) contrast enhanced depth image; (d) channel changed RG-D fused image

3.2 训练及测试环境

实验采用批量标准化操作(BN)，为节省内存空间，每次迭代 128 张图片进行训练，共迭代 30000 次。训练阶段采用动量项为 0.9 的随机梯度下降，权值的初始学习率为 0.001，衰减系数设为 0.0005。实验在 Win10 系统下搭载 6 GB NVIDIA GeForce GTX 1060 的图形处理器中进行，采用了 CUDA9.0 版本图形处理器(GPU)加速驱动，在 Darknet 深度学习框架中完成样本训练与测试。训练过程中对指标 R_{IOU}，R_{recall}，x_{loss} 进行可视化，以评价模型的性能，其中 R_{IOU} 为预测边界框与物体真实区域面积的交并比，R_{recall} 为样本中的正样本被预测正确的概率，计算公式分别为

$$R_{\mathrm{IOU}}=\frac{a_{\mathrm{rea}}(B_{\mathrm{Bdt}}\cap B_{\mathrm{Bgt}})}{a_{\mathrm{rea}}(B_{\mathrm{Bdt}}\cup B_{\mathrm{Bgt}})},\tag{2}$$

$$R_{\mathrm{recall}}=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}},\tag{3}$$

式中：a_{rea}为面积；B_{Bgt}为训练的参考标准框；B_{Bdt}为检测边界框；∩和∪为相交(重合)与相并；T_P表示被判定为正样本，事实上也是正样本的个数；F_P表示被判定为正样本，但事实上是负样本的个数；F_N表示被判定为负样本，但事实上是正样本的个数。以S-RGD网络为例，平均R_{IOU}、平均R_{recall}和x_{loss}随Batch变化曲线如图5所示。

由于在训练过程中迭代次数较多，因此经过降采样后能清楚观察出模型训练情况。其中图5(a)和(b)分别表示采样率为0.01%的整个训练过程R_{IOU}与R_{recall}变化曲线，可以看出训练过程中两条曲线都呈螺旋式上升，R_{IOU}曲线最终稳定在0.73，R_{recall}曲线最终稳定在0.89。图5(c)为迭代前500个Batch的x_{loss}变化曲线，可以看出训练之前的x_{loss}值约为24，在前100个Batch内迅速下降，之后变化极为缓慢，最终x_{loss}值稳定在0.28。

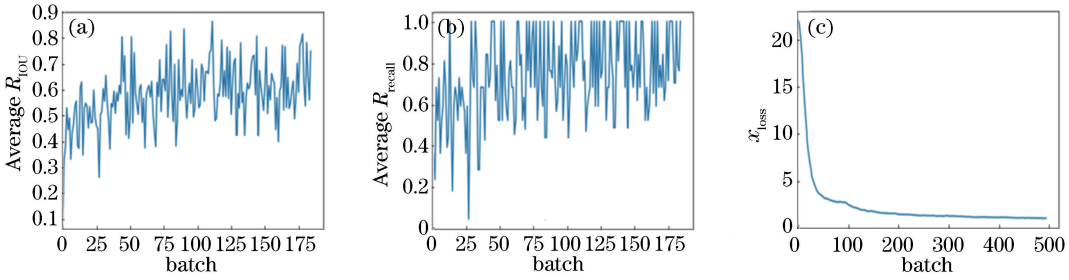

图5 S-RGD训练过程各项指标可视化。(a)采样率为0.01%的平均R_{IOU}变化曲线；(b)采样率为0.01%的平均R_{recall}变化曲线；(c)迭代前500次的x_{loss}变化曲线

Fig. 5 Visualization of indicators in the S-RGD training processing. (a) Average R_{IOU} curves for sampling rates of 0.01%; (b) average R_{recall} curves for sampling rates of 0.01%; (c) top 500 times' iteration of the x_{loss} curve

3.3 实验结果及性能评价

基于D-RGBD和S-RGD两种改进方案，分别进行车辆检测。图6所示为两种网络的检测结果，其中图6(a)为S-RGD和D-RGBD在正常环境下对车辆的检测结果；图6(b)为S-RGD和D-RGBD在隧道、反光、夜晚环境下的检测结果。

训练过程中还使用$P_{precision}$指标来定量评价模型的性能，$P_{precision}$为预测为正的样本中真正的正样本所占比重，表达式为

$$P_{precision}=\frac{T_P}{T_P+F_P}。\tag{4}$$

本文提出的两种方法性能指标对比如表1所示。从表中可知，就平均准确率来讲两种网络相差无几，D-RGBD网络的平均召回率较高，这主要得益于网络容量的增大，能够学习到更深的特征，从而检测到更多目标。但是S-RGD网络的检测速度优于D-RGBD网络，平均检测一张图片所需时间是D-RGBD网络的1/8，对于环境单一条件下的车辆实时检测来讲，S-RGD网络更加方便灵活。

表1 S-RGD和D-RGBD性能指标对比

Table 1 Comparison of S-RGD and D-RGBD performance indicators

Network	Time /ms	R_{IOU}/%	$P_{precision}$/%	R_{recall}/%
S-RGD	24	73.52	85.26	89.27
D-RGBD	198	78.91	88.43	91.63

4 结果分析与比较

4.1 与基于Yolo v2的RGB图像车辆检测对比实验

为更加直观评价本文提出的两种车辆检测方法的性能，图7给出利用Yolo v2对RGB和RGB-D两种图像进行目标检测的结果对比，由于在实验过程中S-RGD和D-RGBD两种方法对于目标检测的定性结果相差无几，所以只给出S-RGD网络的检测结果。当车辆在夜晚和隧道行驶(见图中第一列和第五列)，由于光线较暗，RGB图像中无法识别车辆，但RGB-D图像中的识别结果比较准确；第二列图中由于深度传感器作用距离的限制，在远处的小目标车辆未能检测到；第三列中左侧的建筑物由于形状类似车辆，网络也将其误检为车辆目标；第四列图中图像中仅露出车辆头部，网络无法准确识别，主要源于在训练样本中类似目标较少，没有训练充分。虽然可以人为增加类似样本进行训练，但是耗时过长不宜采用，合理的方法是增强网络的泛化能力，减少过拟合。第六列中两辆车距离较近几乎重叠，网络无法将分离，仅识别为一个目标。为对比在正常环境中RGB和RGB-D网络对于同一目标的检测结果，如图8所示，通过模拟增加图像曝光量和清晰度来降低光照、能见度等因素影响。在光照充足、环境单一的情况下，RGB图像的检测结果与RGB-D

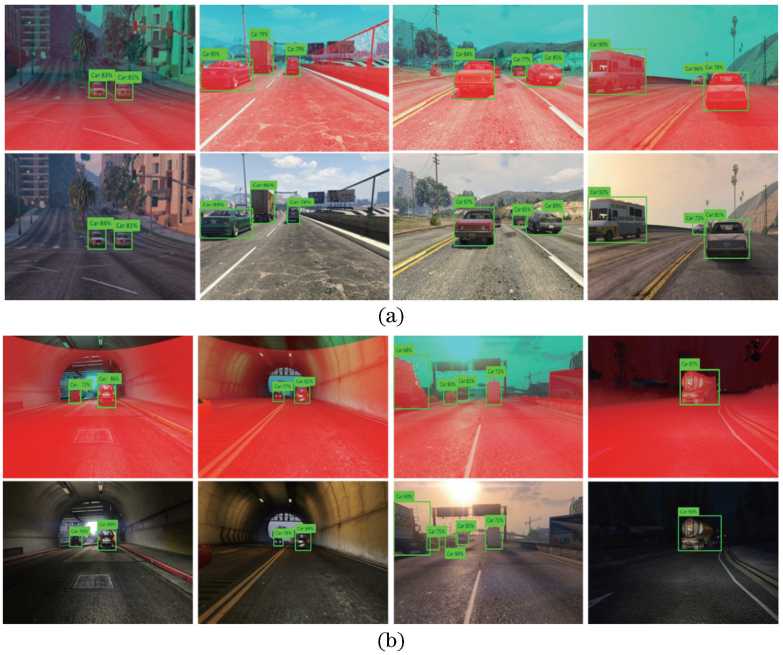

图 6 分别利用 S-RGD 和 D-RGBD 在不同环境中检测车辆的结果。(a)正常环境;(b)隧道、反光、夜晚环境

Fig. 6 Vehicle detection results of using S-RGD and D-RGBD in different environments. (a) Normal environment; (b) tunnel, reflect light, night

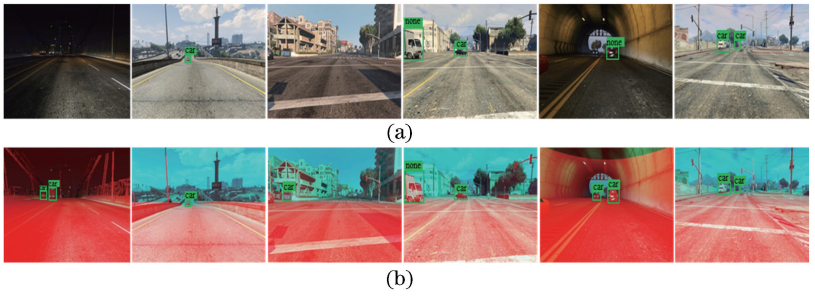

图 7 利用 Yolo v2 对 RGB 和 RGB-D 目标检测结果对比。(a) RGB 检测结果;(b) RGB-D 检测结果

Fig. 7 Contrast results of RGB and RGB-D target detection by Yolo v2. (a) RGB detection results; (b) RGB-D detection results

图像基本一致,但是在夜晚、隧道等光线昏暗、能见度差及环境复杂的情况下,在增强图像对比度时,基于 RGB-D 图像的检测结果表现更加优秀,但同时也有漏检、误检等错误发生。

深度图像可以提供目标轮廓和目标表面凹陷特征,而彩色图像主要提供目标表面的纹理特征,将两者结合可以获得更加充分的信息来有效提升检测准确率。虽然基于 RGB-D 图像的目标检测在某些情况下仍然存在不足,但是相比于同一算法下基于 RGB 图像的目标检测来讲,准确率有了大幅提升,利用 S-RGD 进行检测所需时间与 RGB 图像基本相同。

4.2 与其他目标检测方法的对比实验

本文在相同的 RGB 数据集上分别训练当前流行的 Faster R-CNN [17] 和 SSD [18] 两种目标检测算法模型,并在测试集上与基于 Yolo v2 改进的 S-RGD 和 D-RGBD 结果进行对比,如表 2 所示。同时,为验证两种检测网络的通用性和稳健性,增加了在纽约大学 RGBD 室内数据集 NYU Depth v2[19] 上检测结果的对比实验,为快速有效地达到验证效果,其中标记并训练了 2 类样本进行测试,结果如图 9 所示。

从表 2 和图 9 的对比结果可以看出,在基于 RGB 图像的检测方法中 Faster R-CNN 对目标检测的稳健性最好,这是由于算法的主网络是 VGG16,其网络结构复杂,卷积层数多,且每个类别都要训练一个回归器,但同时导致在训练模型和检测物体时速度缓慢。当输入数据类型变为RGB-D后,利用

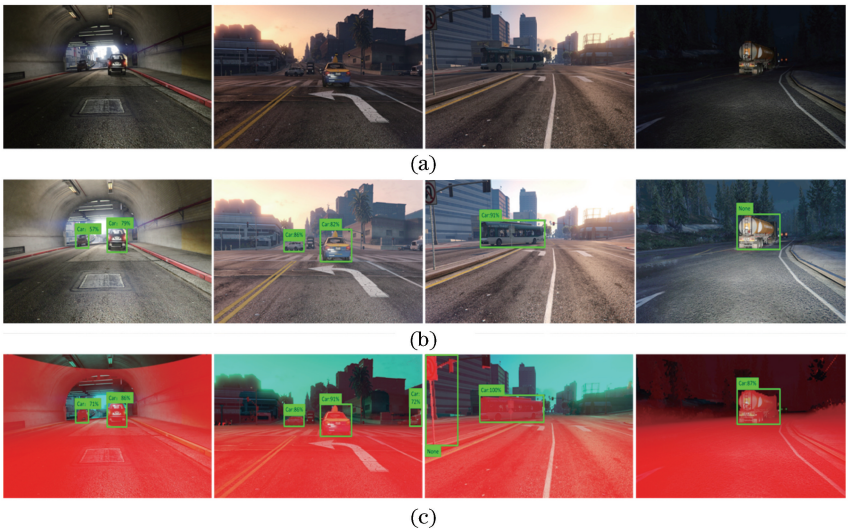

图 8 增强后的 RGB 图像与 RGB-D 图像目标检测结果对比。(a)原始图像;(b) RGB 图像增强后的检测结果;(c) RGB-D 检测结果

Fig. 8 Comparison of enhanced RGB images and the RGB-D images. (a) Original images; (b) RGB detection results after image enhancement; (c) RGB-D detection results

表 2 本文算法与其他目标检测方法结果对比

Table 2 Comparison of the algorithm in this paper with other target detection methods

Data type	Algorithm category	Time / ms	$P_{\text{precision}}$ / %	R_{recall} / %
RGB images	SSD	23	79.65	84.90
	Faster R-CNN	161	84.30	87.13
	Yolov2	19	82.74	85.32
RGB-D images	S-RGD	24	85.26	89.27
	D-RGBD	198	88.43	91.63

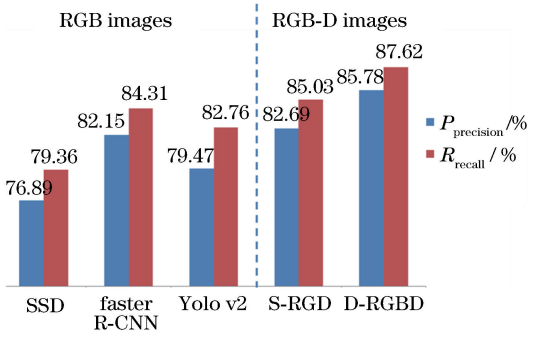

图 9 本文算法在数据集 NYU Depth v2 上与其他方法的对比结果

Fig. 9 Comparison between the proposed algorithm and other methods in the dataset NYU Depth v2

本文提出的两种网络模型对车辆的平均检测精度均有明显提升,与 Faster R-CNN、SSD 相比,S-RGD 检测速度接近于 SSD,比 Faster R-CNN 约提升 6 倍。由于 D-RGBD 复杂的网络结构及彩色与深度图像的充分融合,该网络模型的召回率提升最为明显。

5 结 论

实现了基于卷积神经网络和 RGB-D 图像的实时车辆检测。利用彩色信息和深度信息在 Yolo v2 的目标检测框架下设计并测试 S-RGD 和 D-RGBD 两种网络结构,前者压缩了网络复杂度,提高了检测速度,方便灵活,适用于移动检测设备;后者双通道有利于提升网络容量,提高检测精度,避免了两种信息相互之间的跨通道影响。与 Yolo v2 算法下基于 RGB 图像的目标检测相比,D-RGBD 对车辆检测的准确率和召回率分别提升 5.69%和 6.31%,但检测时间较长,利用 S-RGD 可做到实时检测,但对于准确率和召回率的提升不够明显。

本文算法可以稳定、快速、有效地在不同环境下检测车辆目标,但也存在小目标和重叠目标容易被漏检、误检等问题,并且由于深度信息的距离限制,并不适用于图像中较远车辆的检测。改善网络对微小目标的检测效果,获取更远距离目标的深度信息,仍需要进一步研究。

参 考 文 献

[1] Chen X. Three-dimensional plane target based on neural network recognition [D]. Changchun: Changchun University of Science and Technology, 2011: 7-9.
陈曦. 基于神经网络的三维飞机目标识别研究[D]. 长春: 长春理工大学, 2011: 7-9.

[2] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[3] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C] // 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), June 20-25, 2005, San Diego, CA, USA. New York: IEEE, 2005: 886-893.

[4] Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach [C] // Proceedings of the 17th International Conference on Pattern Recognition, August 26, 2004, Cambridge, UK. New York: IEEE, 2004, 3: 32-36.

[5] Lu H T, Zhang Q C. Applications of deep convolutional neural network in computer vision [J]. Journal of Data Acquisition & Processing, 2016, 31(1): 1-17.
卢宏涛, 张秦川. 深度卷积神经网络在计算机视觉中的应用研究综述 [J]. 数据采集与处理, 2016, 31(1): 1-17.

[6] Hinton G E. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.

[7] Tu S Q, Xue Y J, Liang Y, *et al*. Review on RGB-D image classification [J]. Laser & Optoelectronics Progress, 2016, 53(6): 060003.
涂淑琴, 薛月菊, 梁云, 等. RGB-D 图像分类方法研究综述 [J]. 激光与光电子学进展, 2016, 53(6): 060003.

[8] Socher R, Huval B, Bath B, *et al*. Convolutional-recursive deep learning for 3D object classification [C] // Proceedings of the 25th International Conference on Neural Information Processing Systems, December 3-6, 2012, Lake Tahoe, Nevada. USA: Curran Associated Inc., 2012, 1: 656-664.

[9] Eitel A, Springenberg J T, Spinello L, *et al*. Multimodal deep learning for robust RGB-D object recognition [C] // 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), September 28-October 2, 2015, Hamburg, Germany. New York: IEEE, 2015: 681-687.

[10] Li W. Research on RGB-D object recognition via feature learning [D]. Wuhan: Huazhong University of Science and Technology, 2016: 24-28.
李威. 基于特征学习的 RGB-D 目标识别算法研究 [D]. 武汉: 华中科技大学, 2016: 24-28.

[11] Xu X Y, Li Y C, Wu G S, *et al*. Multi-modal deep feature learning for RGB-D object detection [J]. Pattern Recognition, 2017, 72: 300-313.

[12] Liu F, Liu P Y, Zhang J N, *et al*. Joint detection of RGB-D images based on double flow convolutional neural network [J]. Laser & Optoelectronics Progress, 2018, 55(2): 021503.
刘帆, 刘鹏远, 张峻宁, 等. 基于双流卷积神经网络的 RGB-D 图像联合检测 [J]. 激光与光电子学进展, 2018, 55(2): 021503.

[13] Qu L, Wang K R, Chen L L, *et al*. Fast road detection based on RGBD images and convolutional neural network [J]. Acta Optica Sinica, 2017, 37(10): 1010003.
曲磊, 王康如, 陈利利, 等. 基于 RGBD 图像和卷积神经网络的快速道路检测 [J]. 光学学报, 2017, 37(10): 1010003.

[14] Redmon J, Farhadi A. YOLO9000: better, faster, stronger [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI. New York: IEEE, 2017: 6517-6525.

[15] 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.

[16] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C] // Proceedings of the 25th International Conference on Neural Information Processing Systems, December 3-6, 2012, Lake Tahoe, Nevada. New York: ACM, 2012, 1: 1097-1105.

[17] Ren S Q, He K M, 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.

[18] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot MultiBox detector [M] // Leibe B, Matas J, Sebe N, *et al*. Computer vision-ECCV 2016. Lecture notes in computer science. Berlin, Germany: Springer, 2016, 9905: 21-37.

[19] Silberman N, Hoiem D, Kohli P, *et al*. Indoor segmentation and support inference from RGBD images [M] // Fitzgibbon A, Lazebnik S, Perona P Berlin, *et al*. Computer vision-ECCV 2012. Lecture notes in computer science. Berlin, Heidelberg: Springer, 2012, 7576: 746-760.