计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(2):22−30 [doi: 10.15888/j.cnki.csa.008244]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

单目视觉下基于逆投影空间的车辆细粒度识别①

王　伟[1,2], 唐心瑶[1], 田尚伟[1], 梅占涛[3]

[1](长安大学 信息工程学院, 西安 710064)
[2](安徽科力信息产业有限责任公司, 合肥 230088)
[3](内蒙古第一机械集团股份有限公司, 包头 014030)
通信作者: 唐心瑶, E-mail: andy19966212@126.com

摘　要: 当前车辆识别大多采用深度学习方法, 直接输入图像数据进行训练以获得车辆分类的深度网络, 由于图像本身存在透视形变及尺度变化, 因此不得不采取大量不同类型数据进行训练, 同时也无法获取车辆相关的物理信息. 为了改进上述问题, 本文提出基于逆投影空间训练的车辆细粒度识别方法. 首先利用标定信息及几何约束, 对单目投影下的车辆构建精细化的三维包络框. 然后将车辆三维包络展开, 获得规范化及标准化的逆投影空间数据. 最后利用深度卷积网络对这些展开的规范数据进行训练分类及回归, 获得 5 种常见车辆细分类结果及对应的物理尺寸信息. 实验结果表明, 与传统端到端的深度学习车辆分类算法相比较, 本文算法在利用更少的训练数据的前提下, 能有效的提升车辆分类准确率, 同时可获取车辆三维物理尺寸信息.

关键词: 深度学习; 智能交通; 三维包络框; 三维空间标准化数据; 车辆细粒度识别

引用格式: 王伟,唐心瑶,田尚伟,梅占涛.单目视觉下基于逆投影空间的车辆细粒度识别.计算机系统应用,2022,31(2):22–30. http://www.c-s-a.org.cn/1003-3254/8244.html

Fine-grained Recognition of Vehicles Based on Inverse Projection Space in Monocular Vision

WANG Wei[1,2], TANG Xin-Yao[1], TIAN Shang-Wei[1], MEI Zhan-Tao[3]

[1](School of Information Engineering, Chang'an University, Xi'an 710064, China)
[2](Anhui Keli Information Industry Co. Ltd., Hefei 230088, China)
[3](Inner Mongolia First Machinery Group Co. Ltd., Baotou 014030, China)

Abstract: Most of the current vehicle recognition methods rely on deep learning to directly input image data for training, thus obtaining a deep network. Due to the perspective distortion and scale change of an image, a large number of different types of data have to be used for training, without obtaining the vehicle-related physical information. To address the above problems, we propose a method of vehicle fine-grained recognition based on inverse projection space. First, the three-dimensional bounding boxes are constructed for vehicles under projection of a monocular camera by calibration information and geometric constraints. Second, the bounding boxes are unfolded to obtain normalized and standardized three-dimensional data in the inverse projection space. Finally, a deep convolutional network is introduced to obtain vehicle recognition results and its corresponding physical sizes of five common types of vehicles by training these standardized data. Experimental results show that, compared with traditional end-to-end vehicle recognition methods based on deep learning, the proposed method can effectively improve the accuracy of recognition while using less training data, and the three-dimensional physical sizes of vehicles can also be obtained simultaneously.

Key words: deep learning; intelligent transportation; three-dimensional bounding box; three-dimensional standardized spatial data; fine-grained recognition of vehicles

① 基金项目: 陕西省社会发展领域项目 (2019SF-258); 内蒙古自治区交通运输发展研究中心开放基金 (2019KFJJ—003); 陕西省交通运输厅交通科技项目 (20—25K)
收稿时间: 2020-12-04; 修改时间: 2021-01-14; 采用时间: 2021-04-20; csa 在线出版时间: 2022-01-17

在无人驾驶领域及智能交通应用中, 车辆三维信息的准确获取在车辆行驶路径规划, 安全行驶及车辆违规判断上都有着重要的应用[1], 同时, 详细的车型信息对于精确检测与统计车流[2], 车辆违规处罚[3–6] 等应用上都提供了基础数据支撑. 因此在交通应用中, 十分关注车辆的三维尺寸及车型分类信息, 本文所定义的车辆细粒度即为这两类信息.

目前主流的车辆识别方法主要包括: (1) 基于目标二维局部特征的方法[7–11]. 这类方法利用车牌、车灯、车标或车脸等信息, 对输入的车辆局部特征进行传统的模型识别, 获取车辆的识别结果, 由于检测精度低、特征设计复杂, 这类方法在实际应用中逐渐淡出视线. (2) 基于深度学习的方法, 随着数据集的增加及深度学习目标检测技术的成熟, 出现了一批优秀的车辆检测及分类网络, 尤其以 YOLO 系列[12] 为典型. 但 (1) 和 (2) 类方法都属于二维目标检测方法, 仅能检测识别出车辆目标的存在性及粗略的型号分类, 并不能获取车辆物理尺寸等相关的其他辅助信息, 做不到精细化描述. (3) 基于三维目标检测的方法. 与二维目标检测相比, 三维目标检测能够消除图像成像的透视形变, 同时能够在三维目标检测能在物理尺度上描述车辆信息, 因此更加适合车辆的细粒度描述. 目前基于三维的方法主要基于深度相机 (RGB-D camera)[13,14], 激光相机 (laser camera)[15] 等, 而这些设备价格昂贵且数据量冗余. 相比之下, 单目相机 (monocular camera) 价格便宜维护简单, 同时具有视野范围大且数据量相对较小等优势, 一直是视频监控系统中的主流应用, 但是由于透视形变及投影造成的信息损失, 直接通过单目相机获取车辆目标的三维信息有一定的难度. 综上, 基于单目视觉下三维目标检测的车辆识别研究具有重要意义.

近年来, 基于单目视觉的车辆三维检测算法呈上升趋势. 该类算法主要基于以下两种思路: (1) 基于 CAD/可变模型+局部特征设计[16–18], 如 Zhang 等[19] 提出一种基于可变模型的车辆识别方法, 主要使用 Hog 特征生成初始三维车辆模型, 能够识别轿车, 掀背车, 公交车等常见的 8 种车辆. Corral-Soto 等[20] 也提出一种基于三维可变模型的车辆识别方法, 对于拥挤高速公路下的车辆遮挡有一定的鲁棒性, 该方法对车辆前景根据蒙特卡洛方法和马尔科夫链方法 (MCMC) 沿着车道线方向滑动模型来获取最贴合的三维模型, 从而解决解决道路中的车辆相互遮挡造成的识别失效. Prokaj 等[21] 采用三维 CAD 模型结合 DPM 分类检测器的思路, 通过车辆的局部特征数据训练出一个能够将二维图像和三维模型在几何与视角上进行对齐的 DPM 分类检测器. 该类方法在 CAD 模型库过大时, 存在检索速度慢等缺点, 同时手工设计目标特征在深度学习流行的今天, 也远远达不到理想的准确率. (2) 三维包络盒+机器学习. 该类方法摒弃使用 CAD 模型贴合车辆目标, 而采用更灵活的三维包络盒的方式进行三维检测. Zapletal 等[22] 提出将车辆三维包络盒在逆投影空间中展开, 继而进行训练实现精细化识别的思路, 具体方法为, 对于展开的逆投影空间包络面, 首先利用 HOG 特征对于逆投影空间进行描述, 然后用 SVM 算法进行训练识别, 获取车型识别结果. 由于采用的是传统的手工特征设计, 在较复杂的数据集下, 该算法的识别准确率并不高, 仅能达到 60%. Sochor 等[23] 对车辆前面, 侧面和顶面的二维平面图像进行标准化展开, 然后进行标注, 通过深度学习训练网络的方式学习车辆类型, 在一般场景中精确度高达 83.2%, 但该方法采用 3 个互相正交的消失点的方式对车辆进行三维包络, 而消失点在某些方向存在不稳定现象, 因此在一定视角下该类方法对于车辆的三维检测并不稳定.

基于上述对当前算法的综述分析, 采用三维包络盒 + 深度学习的思路开展本文研究, 与当前已有的方法相比, 本文的创新与贡献有:

(1) 前期工作中对于交通场景构建了自标定模型, 本文将基于相机自标定参数, 单灭点与车辆二维投影的几何约束构建车辆精细化的三维包络.

(2) 对二维车辆目标进行逆投影空间标准化展开, 构建联合物理尺寸标签的损失函数, 训练出更具区分性的车辆细粒度识别网络.

1 前期工作

1.1 相机标定模型建立

相机标定是获得三维世界空间与二维图像空间映射关系的必要步骤, 可为后续第 1.2 节中的车辆 3D 包络框的构建提供依据.

如图 1 所示, 为道路场景相机空间模型示意图, 前期工作对该场景下的相机自动标定和优化问题进行了相关的研究[24], 是本文的基础.

如图 1 所示, 在此空间模型中, 世界坐标系包含 xyz 轴, 相机坐标系包含 $x_cy_cz_c$ 轴, 相机焦距为 f, 相机

距离地面的高度为h, 相机俯仰角和偏转角分别为ϕ和θ. 将世界坐标表示为齐次形式: $\boldsymbol{x}=[x, y, z, 1]^{\mathrm{T}}$, 则在图像坐标中对应为: $\boldsymbol{p}=[\alpha u, \alpha v, \alpha]^{\mathrm{T}}$, $\alpha\neq 0$表示尺度因子. 由文献 [24] 推导可知, 世界坐标到图像坐标的投影方程为:

$$\begin{bmatrix}\alpha u\\ \alpha v\\ \alpha\end{bmatrix}=\begin{bmatrix}f & 0 & 0 & 0\\ 0 & -f\sin\phi & -f\cos\phi & fh\cos\phi\\ 0 & \cos\phi & -\sin\phi & h\sin\phi\end{bmatrix}\begin{bmatrix}x\\ y\\ z\\ 1\end{bmatrix} \tag{1}$$

将式 (1) 展开可得直观的世界坐标至图像坐标的表示形式:

$$\begin{cases}u=\dfrac{\alpha u}{\alpha}=\dfrac{fx}{y\cos\phi-z\sin\phi+h\sin\phi}\\ v=\dfrac{\alpha v}{\alpha}=\dfrac{fh\cos\phi-fy\sin\phi-fz\cos\phi}{y\cos\phi-z\sin\phi+h\sin\phi}\end{cases} \tag{2}$$

由式 (2) 可知, 当给定目标高度为 z_0 时, 即可计算得图像坐标在世界坐标的逆投影. 标定参数(f, h, ϕ, θ)可通过道路标识[25](如道路虚线, 道路宽度等) 间接求取. 在多标识的约束下, 还可在参数空间对于标定参数进行迭代优化. 在文献 [24] 中有详尽的描述, 此处不再赘述. 由此, 通过建立的相机空间标定模型, 可得道路场景下世界坐标与图像坐标的投影与逆投影变换, 从而获得后续构建车辆 3D 包络的基础.

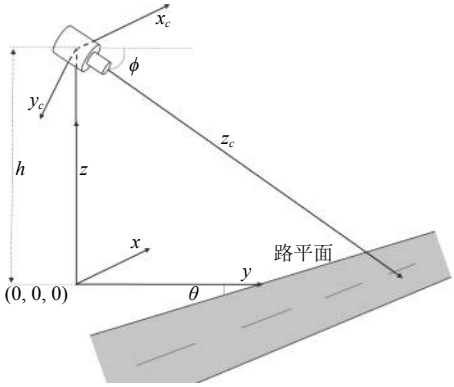


图 1　道路场景中相机空间模型示意图

1.2　车辆 3D 包络框粗构建

基于第 1.1 节中的相机标定, 可进一步构建车辆的 3D 包络框, 为第 2 节中的包络框展开及车辆细粒度识别奠定基础.

如图 2 所示, 为本文车辆 3D 包络框粗构建的示意图. 设车辆 3D 包络框 8 个顶点的世界坐标为 $P_i=(x_i, y_i, z_i)$, i=1, 2,⋯, 8, 图像坐标为 $p_i=(u_i, v_i)$, i=1, 2,⋯, 8, 车辆的初始尺寸为 (l_v, w_v, h_v), 分别表示车辆的长度、宽度和高度, 单位为 m. 由图 1 的标定模型可推导出车辆在长度、宽度和高度的方向向量分别为$\boldsymbol{d}_l=(-\sin\theta, \cos\theta, 0)$, $\boldsymbol{d}_w=(\cos\theta, \sin\theta, 0)$, $\boldsymbol{d}_h=(0, 0, 1)$. 将 P_2 作为车辆基准点，通过式 (3) 可得其余 7 点坐标.

$$\begin{bmatrix}P_1 & P_2\\ P_3 & P_4\\ P_5 & P_6\\ P_7 & P_8\end{bmatrix}=\begin{bmatrix}(x_1, y_1, 0) & P_1+w_v\boldsymbol{d}_w\\ P_2+l_v\boldsymbol{d}_l & P_1+l_v\boldsymbol{d}_l\\ P_1+h_v\boldsymbol{d}_h & P_2+h_v\boldsymbol{d}_h\\ P_3+h_v\boldsymbol{d}_h & P_4+h_v\boldsymbol{d}_h\end{bmatrix} \tag{3}$$

在车辆 3D 框粗包络的过程中, 并不能保证所有参数均准确, 因此需要进一步对粗包络进行调整, 得到更准确的车辆 3D 包络框.

参考前期工作对于车辆空间形态优化的思路[26], 将调整过程看作包络框参数的优化过程, 优化参数包括(l_v, w_v, h_v), 和车辆的偏转角 θ_v 构成的车辆空间形态向量 $\boldsymbol{V}$, $\boldsymbol{V}_1$ 为其初始值. 构造车辆投影凸包与车辆轮廓的约束算法如下, 其中车辆轮廓使用 Mask-RCNN[27] 进行提取.

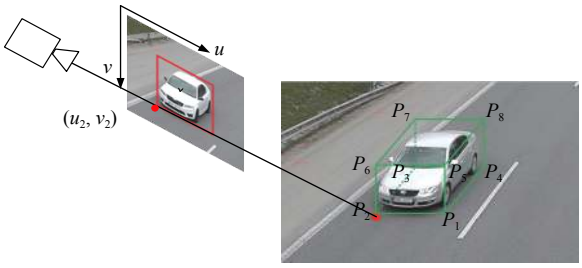


图 2　车辆三维包络框的粗构建

算法 1. 车辆 3D 包络框构建优化算法

1) 通过式 (3) 构建车辆 3D 粗包络, 车辆初始尺寸 (l_v, w_v, h_v) 可通过 Mask-RCNN 得到的车辆类型查阅车辆外廓尺寸获取.

2) 将 3D 包络的 8 点物理坐标代入式 (2) 可反求出 8 个投影点并求凸包, 获得某组已知 $\boldsymbol{V}_1$ 对应的 3D 包络投影凸包, 将式 (3) 得到的世界坐标点通过式 (2) 投影至图像坐标中, 获得车辆投影凸包顶点, 记为$\{s_i|1\leqslant i\leqslant m\}$, m 为投影凸包的顶点数量.

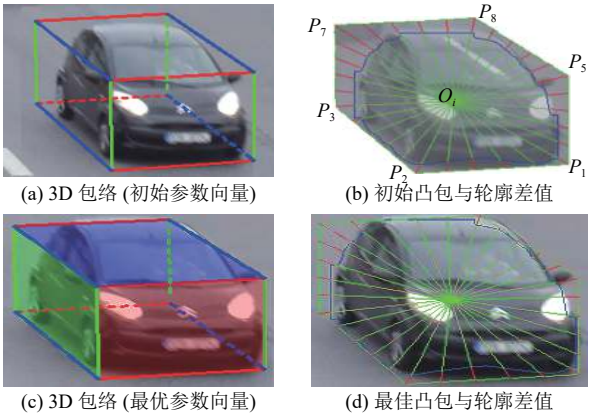
3) 为了更好地构建约束, 在相邻的投影凸包顶点等间隔插入 v 个新顶点, 则稠密投影凸包可表示为 $\{s_i|1\leqslant i\leqslant m(v+1)\}$.

4) 求车辆轮廓 C 的重心 O, 连接 Os_i 获得与 C 的交点 q_i, 得到约束误差为 $\sum_{i=1}^{m(v+1)} s_iq_i$.

图 3(a) 为$v=4$时的一组初始参数向量对应的 3D 包络, 图 3(b) 为初始投影凸包与车辆轮廓的差值, 使用红色线段表示, 投影凸包顶点为P_1, P_2, P_3, P_5, P_7, P_8. 约束函数可表示为:

$$\begin{cases} \min\limits_{v} \sum\limits_{i=1}^{m(v+1)} \dfrac{s_i q_i}{m(v+1)} \\ \text{s.t. } l_0 \leqslant l_v \leqslant l_1, w_0 \leqslant w_v \leq w_1,\ \ h_0 \leqslant h_v \leqslant h_1, \\ \quad (u_2, v_2) \in R,\ \ \theta - \varepsilon \leqslant \theta \leqslant \theta + \varepsilon \end{cases} \tag{4}$$

其中, (l_0, l_1), (w_0, w_1), (h_0, h_1)为车辆长度、宽度和高度的约束范围.基准点(u_2, v_2)的取值范围为 R, 使用矩形区域表示. ε为θ_v的取值范围的阈值. 限定参数的取值, 可进一步缩小参数优化的空间, 提升优化效率. 如图 3(c) 为最优参数向量对应的 3D 包络, 图 3(d) 为最优投影凸包与车辆轮廓的差值.



(a) 3D 包络 (初始参数向量) (b) 初始凸包与轮廓差值

(c) 3D 包络 (最优参数向量) (d) 最佳凸包与轮廓差值

图 3 车辆三维包络精细化过程实例图

2 车辆目标展开标准化及深度网络训练

2.1 基于车辆三维包络框的标准化展开

由于透视投影可知, 单目视觉下的目标会发生不同程度的透视畸变及尺度变化, 事实上这对于目标的识别有一定的影响, 传统的方法大多采用大量不同视角下及不同尺度下的目标数据进行训练, 继而弥补透视畸变及尺度变化对于目标识别产生的影响, 而本文可利用车辆三维包络框的标准化展开, 在数据输入端即可做透视畸变及尺度变化的校正. 通过这种方式, 在达到相同精度的情况下, 需要更少的数据集. 如图 4 所示, 通常视角下车辆的可视面有 3 面, 车辆目标正面 (F), 侧面 (S), 顶部 (V) (当然还有其他一些可视面的组合方式, 本文中暂时只考虑 F-S-V 的可视面组合方式), 可利用透视变换的原理对 3 个可视面进行矫正.

透视变换的公式如下, 其功能为投影图像至新的可视化平面.

$$\begin{bmatrix} x' \\ y' \\ w' \end{bmatrix} = \begin{bmatrix} T_1 & T_2 \\ T_3 & T_4 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \tag{5}$$

其中, 图像像素坐标变换前表示为 (u, v), 变换后为 (x', y'), T_1 表示线性变换, 为 2×2 的矩阵, T_2 表示透视变换, 为 2×1 的矩阵, T_3 表示图像平移, 为 1×2 的矩阵, T_4 为不为零的常数. 本文对 3D 包络框的每个可视面, 利用其四边形的每个顶点与变换之后的标准矩形顶点建立映射等式关系, 即可求取对应的透视变换矩阵, 继而可将 3 个可视面进行逆透视展开. 其展开的顺序和规范如图 5(a) 所示, 每个面的透视形变校正结果示例图如图 5(b) 所示. 设规范化的展开图像宽为 w_{sd}, 高为 h_{sd}, 如图 6 所示, 为一组车辆 3D 包络及标准化展开的示例图. 事实上除了展开的具有逆透视效果的规范化图像之外, 车辆的物理尺寸也可作为后续分类识别的有效信息.

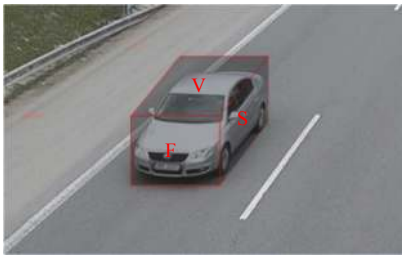
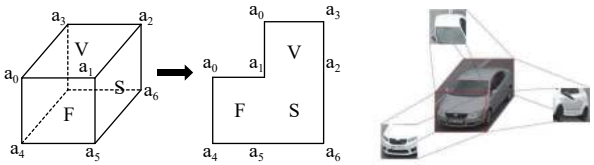


图 4 车辆三维包络可视面示意图



(a) 展开顺序和规范 (b) 各面的透视形变校正结果

图 5 车辆三维包络视变换示意图

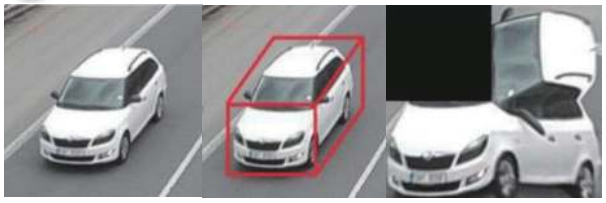


图 6 车辆三维包络及标准化展开示例图

2.2 基于标准化展开输入的深度网络训练分类模型

本文设计了一种可以同时预测车辆和车辆三维尺寸的细粒度识别网络, 如图 7 所示, 网络输入为由三维目标检测结果展开所得标准化展开图 (224×224), 网络输出为车辆分类v_{type} (Hatch-back, Sedan, SUV, Bus, Truck 共 5 类) 和车辆三维尺寸(l_v, w_v, h_v).

为了提升网络整体泛化性能, 防止过拟合, 本文采

用 ResNet[28] 作为 backbone, 网络共包含两个分支: 主分支和辅助分支, 这两个分支都可以完成车辆分类和车辆三维尺寸预测. 其中, 主分支用于训练和预测, 辅助分支借鉴了 GoogleNet 网络[29] 中辅助分类器的结构, 只在网络训练过程中使用, 能够防止一定程度的梯度消失.

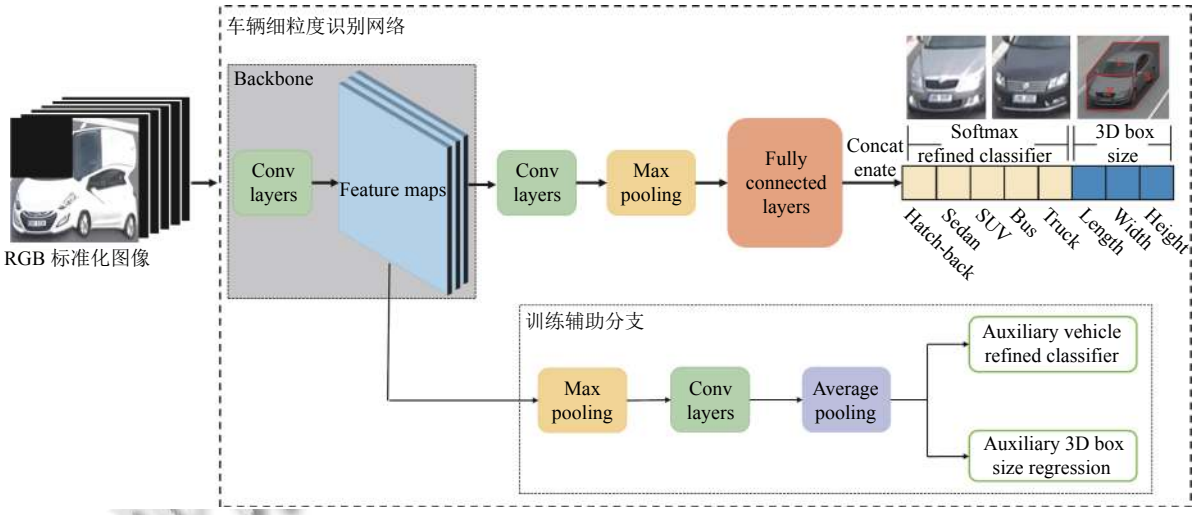


图 7 深度网络结构图

由于分类标签是一概率分布向量, 其网络输出值比车辆实际三维尺寸标签小很多, 因此, 为了使得模型更稳定, 本文对车辆三维尺寸标签做了归一化处理, 具体为将标准化展开图中车辆尺寸像素与实际物理尺寸相比, 作为最终的三维尺度因子, 其值范围在 0–1 之间. 如图 8 所示, 车辆像素尺寸标签大小为(l_{pix}, w_{pix}, h_{pix}), 物理尺寸标签为(l_v, w_v, h_v), 则新的标签设计为尺度因子: $s_l = l_{\text{pix}}/l_v$, $s_w = w_{\text{pix}}/w_v$, $s_h = h_{\text{pix}}/h_v$.

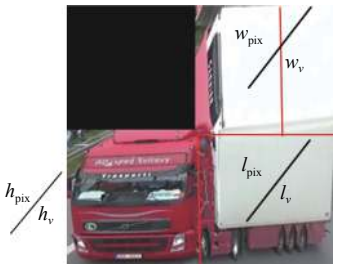


图 8 车辆三维物理尺寸标签尺度因子设计示例图

损失函数得设计共包含 3 个部分, 车辆分类损失, 车辆三维尺寸回归损失和辅助训练损失, 如式 (6) 所示, 辅助训练损失也由分类和回归损失组成. 具体形式如公式组 (7) 所示.

$$L_{\text{total}} = L_{\text{classifier}} + L_{\text{size}} + L_{\text{auxiliary}} \tag{6}$$

$$\begin{cases} L_{\text{classifier}} = -\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{K} y_{\text{gt}}^{(j)} \ln y_{\text{pre}}^{(j)} \\ L_{\text{size}} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left(\left|l_v^{\text{pre}} - l_v^{\text{gt}}\right| + \left|w_v^{\text{pre}} - w_v^{\text{gt}}\right| + \left|h_v^{\text{pre}} - h_v^{\text{gt}}\right|\right) \\ L_{\text{auxiliary}} = \lambda_c L_{\text{aux\_classifier}} + \lambda_s L_{\text{aux\_size}} \end{cases} \tag{7}$$

车辆分类损失$L_{\text{classifier}}$形式为多分类交叉熵损失, 如式 (7) 所示, N为网络训练时每批次输入的标准化展开图数量, K 为分类数, 本文中分别取 32 和 5, $y_{\text{gt}}^{(j)}$表示第j类的车辆分类标签, 如果车辆属于第j类, 则$y_{\text{gt}}^{(j)} = 1$, 否则$y_{\text{gt}}^{(j)} = 0$, $y_{\text{pre}}^{(j)}$表示经过全连接层及 Softmax 处理后车辆属于第j类的概率. 车辆三维尺寸回归损失L_{size}为L_1范数损失, 如式 (7) 所示, l_v^{pre}, w_v^{pre}, h_v^{pre}分别表示网络预测所得车辆三维尺寸尺度因子, l_v^{gt}, w_v^{gt}, h_v^{gt}分别表示车辆三维尺寸尺度因子真实归一化标签值. 辅助训练损失$L_{\text{auxiliary}}$如式 (7) 所示, $L_{\text{aux\_classifier}}$, $L_{\text{aux_size}}$分别与$L_{\text{classifier}}$, L_{size}具有相同的形式, λ_c和λ_s分别表示分类和回归损失在辅助训练损失中的权重系数, 本文中选取$\lambda_c = \lambda_s = 0.5$.

3 实验结果

3.1 实验过程及实例

本文所应用的场景为道路交通视频监控, 因此使

用针对道路交通监控场景下的 BrnoCompSpeed 数据集[30], 该数据集包含 6 个交通场景, 如图 9 所示, 其中, 单车道宽度为 3.5 m, 道路虚线长度为 3 m, 虚线间隔为 6 m. 同时该数据集对于经过的每辆车都有明确的车辆记录. 表 1 为本文对 6 个交通场景自动标定的结果.

对于数据集的处理, 首先将视频数据集处理为图像数据集, 由于交通场景中车流量较小, 因此本文对数据集的处理方式为, 每隔 10 s 截取 1 帧, 去除车辆目标过小以及无车辆目标的图像帧, 整理得到图像数据集. 对图像数据集进行分类, 本文的分类标准是对轿车类 (Car) 中的两厢车 (Hatch-back) 和三厢车 (Sedan) 进行再分类, 总体车辆分为 Hatch-back, Sedan, SUV, Bus, Truck 共 5 类, 同时根据数据集提供的详细车辆信息查取其对应的三维尺寸(l_v, w_v, h_v).

图 9 BrnoCompSpeed 数据集下的交通场景

表 1 交通场景相机自标定结果

场景	f (pixel)	h (mm)	ϕ (rad)	θ (rad)
1	2 878.13	10 119.08	0.178 74	0.266 04
2	3 994.17	8 071.00	0.157 17	0.035 35
3	3 384.25	8 126.49	0.262 95	−0.248 69
4	3 435.40	8 058.36	0.176 30	0.170 00
5	7 443.22	6 389.58	0.953 24	0.152 07
6	1 240.86	6 471.77	0.683 18	−0.531 65

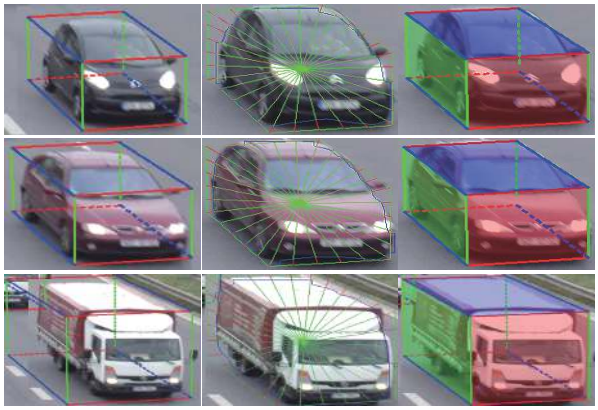
Mask-RCNN 网络集目标检测分类与分割于一体, 因此本文采用该网络对数据集中的车辆目标进行预处理, 获取车辆的预分类及边界分割结果. 该网络可以识别 80 种不同类别目标, 但对于车辆只能粗略分为 Car, Bus, Truck 3 类. 根据初始识别的车辆可根据统计给出物理尺寸取值范围, 各车辆的外轮廓尺寸取值范围实例如表 2 所示. 最后根据轮廓约束构建精细化的车辆 3D 包络. 实例图如图 10 所示.

表 2 各类型车辆外轮廓尺寸范围 (m)

车辆类型	宽w_v	长l_v	高h_v
小型轿车	1.5–1.7	3.6–4.4	1.3–1.5
中大型轿车	1.7–2.0	4.3–5.2	1.3–1.6
货车(重载)	2.4–2.5	12.5–18.5	2.4–2.7

对车辆三维包络 3 个可视面进行透视变换, 并对透视变换后的可视面进行标准化展开, 更多展开实例如图 11 所示.

本次实验从 BrnoCompSpeed 视频数据集中截取 3 000 张图片, 其中包含车辆 6 000 辆, 训练集 (4 000 辆) 和测试集 (2 000 辆), 训练集和测试集均包含 5 个类别的车辆, 并对每辆车都标注了对应的车辆及三维尺寸信息, 训练集中对较难区分的两厢车 (Hatch-back), 三厢车 (Sedan) 和 SUV 各 1 000 辆, 公交车类 (Bus) 和卡车类 (Truck) 各 500 辆. 为了便于网络的训练, 将展开的标准化图像分辨率调整为 224×224 大小. 实验在配置有 Intel i7-6800K CPU 和 GeForce GTX 1080Ti GPU 的 PC 机上运行.



(a) 粗 3D 包络 (b) 轮廓约束优化 (c) 精细化包络结果

图 10 车辆精细化三维包络构建实例图

图 11 车辆三维包络标准化展开实例

3.2 实验结果分析

本文选取 ResNet-101 作为主干特征提取网络, 为了提高检测精度, 采用其在 ImageNet[31] 上的预训练参数, 在训练的过程中进行微调 (fine-tune). 网络的输入

的是展开的标准化图像, 批次大小设置为 32, 分类输出 5 类车辆, 以及回归输出的是车辆物理尺寸长宽高.

由于本网络为多任务输出网络, 包括车辆分类及车辆物理尺寸输出, 因此, 实验结果可以从分类的精度及物理尺寸回归的结果两方面进行分析. 图 12 为细粒度识别结果在测试集上的车辆 P-R 曲线 (Precision-Recall curves) 图, 可看出对于特征区分度较高的 Bus 和 Truck, 本文方法的分类精度均超过 90%, Sedan 和 Hatch-back 车辆, 识别率也超过 80%, 由于 SUV 的特征区分度和二厢车及三厢车不大, 因此识别率稍低.

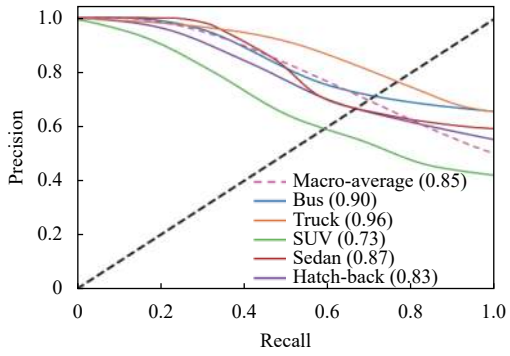


图 12 车辆分类 Precision-Recall 图

利用同样的网络结构及数据集, 本文分别用标准化展开数据及原始图像数据做为输入进行训练, 得到不同的识别结果, 以此证明本文方法对于识别结果的有益性, 如表 3 所示, 为两类方法识别的结果对比, 数值均为四舍五入的整数.

表 3 车辆分类平均精确度对比 (%)

输入方式	Bus	Truck	SUV	Sedan	Hatch-back
原始图像输入	88	92	45	74	64
展开标准化图像输入	90	96	73	87	83

通过表 3 可以看出, 对于 Bus 及 Truck 等本身特征区分度很大车辆, 本文算法的准确率提高并不大, 而对于 SUV, Sedan 及 Hatch-back 等特征区分度较小的车辆, 本文方法的精确度有了明显的提高. 可证明本文采用的车辆目标三维展开规范化的输入方法, 可以有效的提高网络分类的性能.

对于车辆物理尺寸的回归输出, 本文对于预测输出的物理尺寸 $\boldsymbol{X}^{\text{pre}}=(l^{\text{pre}}, w^{\text{pre}}, h^{\text{pre}})$ 与标签物理尺寸 $\boldsymbol{X}^{\text{label}}=(l^{\text{label}}, w^{\text{label}}, h^{\text{label}})$, 利用式 (8) 计算准确率 P_{size}, 其中, $\|\cdot\|_2$ 表示欧氏距离的二范数:

$$P_{\text{size}}=1-\frac{\left\|\boldsymbol{X}^{\text{label}}-\boldsymbol{X}^{\text{pre}}\right\|_2}{\left\|\boldsymbol{X}^{\text{label}}\right\|_2}\times 100\% \tag{8}$$

车辆三维尺寸的识别受视角影响比较大, 如图 13 所示, 为测试数据集上不同偏转视角下网络预测车辆物理尺寸的平均精度.

从图 13 中可以看出, 当相机偏转角接近±45°左右时, 由于图像中车辆的 3 个可视面均可充分的展现, 因此在做三维包络展开时, 可以保留较多的特征信息, 也有助于最终车辆三维尺寸的回归输出. 而当相机偏转视角接近于 0°附近时, 图像中车辆目标纵向信息消失殆尽, 因此尤其对于车辆纵向长度的识别影响很大, 因此输出的车辆三维尺寸精确度较低.

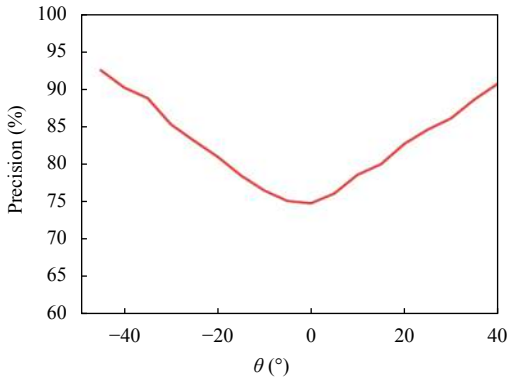


图 13 不同相机视角下车辆三维尺寸预测平均精度图

表 4 为车辆细粒度识别方法对比, 其中精度由车辆单个识别精度和追踪过程中综合识别精度组成.

表 4 不同车辆识别方法对比

网络	单车识别精度 (%)	综合识别精度 (%)	是否输出车辆尺寸
AlexNet[32]	66.65	77.75	False
VGG-16[33]	77.26	86.71	False
VGG-19[33]	76.74	86.06	False
ResNet-50[28]	75.48	84.61	False
ResNet-101[28]	76.46	85.31	False
ResNet-152[28]	77.68	86.20	False
BoxCars[23]	80.4	92.9	False
本文	81.5	93.1	True

表 4 中 BoxCars 也是采用消除透视畸变及 3D 展开输入的方法, 从中可知, 本文方法与 BoxCars 方法在识别精度上均有较大的提高. 本文方法相比于 BoxCars 还可回归输出车辆物理尺寸信息.

4 结论与展望

本文提出一种基于车辆三维包络展开的车辆识别方法, 该方法采用三维包络展开的规范化数据作为输入进行训练, 不仅可以提高车辆分类的精度, 而且可输出获得车辆物理尺寸信息. 通过在 BrnoCompSpeed 视频数据集中的实验表明, 相比于传统的原始图像数据直接输入训练, 基于三维包络展开规范化图像数据方法, 由于很大程度上消除了透视畸变及尺度因素的影响, 使得目标的特征更加突出及规范, 从而较大程度提升了细粒度识别的精度. 同时, 本文方法还可以回归输出车辆物理尺寸信息, 更加丰富了分类车辆描述的维度.

然而, 本文方法仍存在可优化的余地, 譬如车辆三维包络展开数据的规范程度依赖于车辆的 3D 包络准确程度, 而相机接近 0°视角下, 车辆的 3D 包络将会有较大误差. 同时, 与传统图像目标识别一样, 本文对于小目标的识别也存在较大误差, 主要原因就是小目标本身具有的图像特征较少, 数据规范化之后有可能造成较大变形, 影响分类识别结果. 后续工作将会着重探索和研究车辆在不同视角下的精确包络难题, 及小目标车辆的精确分类问题, 以进一步提高车辆分类的准确率以及稳定性.

参考文献

1 Chen XZ, Kundu K, Zhang ZY, *et al*. Monocular 3D object detection for autonomous driving. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2147–2156. [doi: 10.1109/CVPR.2016.236]

2 Taghvaeeyan S, Rajamani R. Portable roadside sensors for vehicle counting, classification, and speed measurement. IEEE Transactions on Intelligent Transportation Systems, 2014, 15(1): 73–83. [doi: 10.1109/TITS.2013.2273876]

3 Sochor J, Juránek R, Herout A. Traffic surveillance camera calibration by 3D model bounding box alignment for accurate vehicle speed measurement. Computer Vision and Image Understanding, 2017, 161: 87–98. [doi: 10.1016/j.cviu.2017.05.015]

4 武非凡, 宋焕生, 戴喆, 等. 交通监控场景下的相机标定与车辆速度测量. 计算机应用研究, 2020, 37(8): 2417–2421. [doi: 10.19734/j.issn.1001-3695.2019.03.0089]

5 Liu XC, Liu W, Mei T, *et al*. PROVID: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. IEEE Transactions on Multimedia, 2018, 20(3): 645–658. [doi: 10.1109/TMM.2017.2751966]

6 关济民. 高速公路联网收费稽查管理系统设计与实现 [硕士学位论文]. 南京: 南京理工大学, 2018.

7 Duan K, Parikh D, Crandall D, *et al*. Discovering localized attributes for fine-grained recognition. Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 3474–3481. [doi: 10.1109/CVPR.2012.6248089]

8 Berg T, Belhumeur PN. POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 955–962. [doi: 10.1109/CVPR.2013.128]

9 Deng J, Krause J, Li FF. Fine-grained crowdsourcing for fine-grained recognition. Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 580–587. [doi: 10.1109/CVPR.2013.81]

10 Zhang BL. Reliable classification of vehicle types based on cascade classifier ensembles. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(1): 322–332. [doi: 10.1109/TITS.2012.2213814]

11 Llorca DF, Colás D, Daza IG, *et al*. Vehicle model recognition using geometry and appearance of car emblems from rear view images. Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems. Qingdao: IEEE, 2014. 3094–3099. [doi: 10.1109/ITSC.2014.6958187]

12 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788. [doi: 10.1109/CVPR.2016.91]

13 Qi CR, Liu W, Wu CX, *et al*. Frustum PointNets for 3D object detection from RGB-D data. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 918–927. [doi: 10.1109/CVPR.2018.00102]

14 Zhou Y, Tuzel O. VoxelNet: End-to-end learning for point cloud based 3D object detection. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4490–4499. [doi: 10.1109/CVPR.2018.00472]

15 Fei ZG, Guo JJ, Wang JD, *et al*. The application of laser and CCD compound measuring method on 3D object detection. Proceedings of 2010 IEEE International Conference on

Mechatronics and Automation. Xi'an: IEEE, 2010. 1199–1202. [doi: 10.1109/ICMA.2010.5588049]

16 Zia MZ, Stark M, Schiele B, *et al*. Detailed 3D representations for object recognition and modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(11): 2608–2623. [doi: 10.1109/TPAMI.2013.87]

17 Krause J, Jin HL, Yang JC, *et al*. Fine-grained recognition without part annotations. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 5546–5555. [doi: 10.1109/CVPR.2015.7299194]

18 Chabot F, Chaouch M, Rabarisoa J, *et al*. Deep MANTA: A coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1827–1836. [doi: 10.1109/CVPR.2017.198]

19 Zhang ZX, Tan TN, Huang KQ, *et al*. Three-dimensional deformable-model-based localization and recognition of road vehicles. IEEE Transactions on Image Processing, 2012, 21(1): 1–13. [doi: 10.1109/TIP.2011.2160954]

20 Corral-Soto ER, Elder JH. Slot cars: 3D modelling for improved visual traffic analytics. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 889–897. [doi: 10.1109/CVPRW.2017.123]

21 Prokaj J, Medioni G. 3-D model based vehicle recognition. Proceedings of 2009 Workshop on Applications of Computer Vision. Snowbird: IEEE, 2009. 1–7. [doi: 10.1109/WACV.2009.5403032]

22 Zapletal D, Herout A. Vehicle re-identification for automatic video traffic surveillance. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas: IEEE, 2016. 25–31. [doi: 10.1109/CVPRW.2016.195]

23 Sochor J, Špaňhel J, Herout A. BoxCars: Improving fine-grained recognition of vehicles using 3-D bounding boxes in traffic surveillance. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(1): 97–108. [doi: 10.1109/TITS.2018.2799228]

24 王伟, 张朝阳, 唐心瑶, 等. 道路场景下相机自动标定及优化算法. 计算机辅助设计与图形学学报, 2019, 31(11): 1955–1962. [doi: 10.3724/SP.J.1089.2019.17737]

25 Kanhere NK, Birchfield ST. A taxonomy and analysis of camera calibration methods for traffic monitoring applications. IEEE Transactions on Intelligent Transportation Systems, 2010, 11(2): 441–452. [doi: 10.1109/TITS.2010.2045500]

26 王伟, 唐心瑶, 张朝阳, 等. 跨相机交通场景下的车辆空间定位方法. 计算机辅助设计与图形学学报, 2021, 33(6): 873–882. [doi: 10.3724/SP.J.1089.2021.18612]

27 He KM, Gkioxari G, Dollár P, *et al*. Mask R-CNN. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2980–2988. [doi: 10.1109/ICCV.2017.322]

28 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

29 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. Proceedings of 2015 IEEE International Conference on Computer Vision. Boston: IEEE, 2015. 1–9. [doi: 10.1109/CVPR.2015.7298594]

30 Sochor J, Juránek R, Špaňhel J, *et al*. Comprehensive data set for automatic single camera visual speed measurement. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(5): 1633–1643. [doi: 10.1109/TITS.2018.2825609]

31 Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

32 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 1097–1105.

33 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2015.