DOI: 10.19650/j.cnki.cjsi.J2108317

一种多焦距动态立体视觉 SLAM*

冯明驰[1],刘景林[1],李成南[1],汪静姝[2]
(1. 重庆邮电大学先进制造工程学院 重庆 400065; 2. 重庆理工大学机械工程学院 重庆 400054)

摘 要:现有的双目同步定位与建图(SLAM)都使用标准立体相机,所处环境为静态的假设会影响其在动态环境中的精度。提出了一种多焦距动态立体视觉 SLAM 方法,它克服了标准立体相机无法兼顾远距离和宽视场感知场景的缺点,并去除了动态物体对 SLAM 的影响。具体来说,对传统的立体校正方法进行了改进,并使用校正参数修正了特征点的位置,而不是整张图像,还提出了一种自适应特征提取和匹配方法以增加多焦距图像的特征匹配数量。综合使用多视图几何、区域特征流和相对距离检测动态对象,剔除动态对象上的特征点。在公开数据集 KITTI 上,该方法相对 ORB-SLAM3 和 DynaSLAM 的定位精度都提高了 6.97%,在自建数据集中,该方法的定位精度比 ORB-SLAM3 提高了 26.64%,比 DynaSLAM 提高了 32.09%。

关键词: 同时定位与建图;多焦距立体视觉;实例分割;动态对象检测
中图分类号: TH85 TP391 **文献标识码**: A **国家标准学科分类代码**: 510.40

A multi-focal length dynamic stereo vision SLAM

Feng Mingchi[1], Liu Jinglin[1], Li Chengnan[1], Wang Jingshu[2]
(*1. School of Advanced Manufacturing Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; 2. College of Mechanical Engineering, Chongqing University of Technology, Chongqing 400054, China*)

Abstract: The existing stereo simultaneous localization and mapping (SLAM) methods all use standard stereo cameras, and the assumption of the static environment has influence on their accuracy in dynamic environment. A multi-focal dynamic stereo vision SLAM is proposed. It could overcome the insufficiency of standard stereo cameras that cannot perceive the scene at long distance and wide field of view. The impact of dynamic objects is also removed. To be specific, the stereo calibration method is improved and the calibration parameters are utilized to rectify ORB features instead of rectifying stereo images. For multi-focal stereo images, a feature extraction and matching method is also proposed to increase the number of matched features. Multi-view geometry, regional feature flow and relative distance are used to detect dynamic objects. The feature points on the dynamic objects are eliminated. Compared with ORB-SLAM3 and DynaSLAM, the positioning accuracy of the proposed method on the public data set KITTI is increased by 6.97%, and the positioning accuracy on the self-made data set is increased by 26.64% and 32.09%, respectively.

Keywords: simultaneous localization and mapping; multi-focal length stereo visual; instance segmentation; dynamic object detection

0 引 言

同步定位与建图(simultaneous localization and mapping, SLAM)技术是移动载体在未知环境下实现自主定位与导航的关键技术[1]。视觉传感器因其分辨率高、采集环境图像信息丰富、简易便携、硬件成本低、定位精度高且无需环境先验信息等优势成为 SLAM 研究的热门领域。视觉 SLAM 技术是计算机视觉、机器人和增强现实等各种上层应用的基础和关键模块,在机器人定位和导航方面发挥着重大作用[2]。

近几十年来,众多研究者提出了许多优秀的视觉 SLAM 系统,如实时单目 SLAM(real-time single camera SLAM, MonoSLAM)[3]、并行跟踪与建图(parallel tracking

收稿日期:2021-07-26 Received Date: 2021-07-26
*基金项目:重庆市科技局(cstc2019jscx-zdztzxX0050, cstc2019jscx-mbdX0004)、国家自然科学基金(51505054)、重庆市教育委员会(KJZD-M201801101, KJQN201801147)项目资助

and mapping, PTAM)[4]、快速特征点提取和描述 SLAM (oriented fast and rotated BRIEF SLAM, ORBSLAM)系列[5-7]、大范围直接单目 SLAM (large-scale direct monocular SLAM, LSD-SLAM)[8],这些系统使用单目或者立体相机可获得令人满意的性能。但是单目 SLAM 无法获取地图和轨迹的尺度,立体 SLAM 系统相机一般为左右分布,根据左右图像视差计算场景真实深度,构建真实尺度地图,并准确定位自身位置。现有的立体视觉 SLAM 系统都是使用相同焦距相机,并且现有的可用于视觉 SLAM 的立体数据集都是采用较短焦距镜头录制,短焦距镜头的优点是能获取广视野场景信息,缺点是对远距离对象的可探测性比较差。

视觉 SLAM 方法分为间接法和直接法。间接法视觉 SLAM 系统首先提取图像序列中每一帧的特征点,然后匹配图像特征点,再进行相机轨迹的跟踪和估计。程序员角度 SLAM (graph SLAM from a programmer's perspective, ProSLAM)[9]、ORB-SLAM 和一个通用的 SLAM 框架和基准(a general SLAM framework and benchmark, GSLAM)[10]是典型的开源间接视觉 SLAM 框架,它们实现了不同的功能应用且具有卓越的性能。直接法视觉 SLAM 系统基于灰度不变假设,利用图像的光度信息和图像一致性以对准图像进行位姿估计,可以在较低纹理环境中正常工作,LSD-SLAM 是典型的直接视觉 SLAM 框架。

ProSLAM 是一个简单的立体 SLAM 框架,它使用立体图像作为唯一的输入。它拥有 4 个模块:三角化测量、增量运动估计、地图管理和重新定位。该框架在仅单个线程中执行,并且没有执行光束法平差(bundle adjustment, BA)优化。ORB-SLAM2 是能使用单目、立体和 RGBD 相机的完整视觉 SLAM 框架。它有 3 个主要线程:跟踪、局部建图和回环检测,全局 BA 线程作为整个系统的全局优化。ORB-SLAM3 在 ORB-SLAM2 上改进和升级,增加了多地图模块。跟踪丢失时,多地图模块重新生成一个新地图,当检测到与之前地图有重合时,新地图将与以前地图无缝合并。Zhao 等[10]提出了一个通用的、跨平台的 GSLAM。它被设计成兼容不同类型的传感器,包括但不限于单目、立体、RGBD 和多相机视觉惯性里程计与多传感器融合。它可以很好地支持基于特征法、直接法和基于深度学习的 SLAM。GSLAM 还提供了 Python 接口,可以很容易地部署对象检测。

随着深度学习的发展,动态视觉 SLAM 与深度学习更加紧密地结合,可以通过语义分割和对象检测消除动态特征对 SLAM 的影响。Cui 等[11]提出一种面向动态环境的语义视觉 SLAM(a semantic visual SLAM for dynamic environments, SOF-SLAM)将语义信息和几何信息紧耦合,该方法通过增加语义分割线程,为映射点提供语义标签。利用语义分割信息辅助多视图几何方法,有效地去除动态三维地图点,生成可靠的地图点。在后续的改进中提出用于动态环境的语义深度过滤 SLAM(semantic depth filter SLAM for dynamic environments, SDF-SLAM)[12],在基于语义光流的地图初始化后,加入基于逆深度滤波的动态场景地图更新策略,该方法克服了从二维图像水平检测动态特征点的缺陷,能够更可靠地检测出物体的动态特征。与 SOF-SLAM 相比较,极大的增加了在动态场景中相机定位和跟踪的精度。Zhang 等[13]提出一种解决动态环境下的动态目标检测方法,该方法使用卷积神经网络模型从 RGB 图像中检测目标对象,通过融合深度图像信息,并使用 k-means 算法对目标特征进行聚类。最后,使用多视图几何方法识别在动态环境下的运动对象。Bescós 等[14]提出一种动态场景中的追踪、建图和修复(tracking, mapping and inpainting in dynamic scenes, DynaSLAM),该方法在 ORBSLAM2 的基础上使用实例分割网络(mask region-based convolution neural networks, Mask R-CNN)[15]增加了对动态对象的检测和背景修复功能,该系统适应于单目相机、立体相机和 RGB-D 相机,能够通过多视图几何,深度学习或者两者结合的方式检测运动对象。系统在检测运动对象后,移除动态对象区域,再根据静态区域修复补全动态区域图像,得到一个全图静态图像。最近 Bescós 等[16]升级了 DynaSLAM 为紧密耦合的多目标跟踪 SLAM(tightly-coupled multi-object tracking and SLAM, DynaSLAM II),该可视化 SLAM 系统可配置在立体相机或者 RGB-D 相机上,它紧耦合多对象跟踪功能。DynaSLAM II 系统利用实例分割和 ORB 特征跟踪动态对象,提出一种新的 BA 优化方法优化 SLAM 系统中的静态对象、动态对象和自身姿态,并且跟踪对象的 3D 边界框也可以在固定的时间内进行计算和优化。

目前还没有立体 SLAM 能够组合多焦距相机,多焦距相机可以检测宽视野和远距离的物体。标准的视觉 SLAM 系统基于环境静态假设,贾松敏等[17]使用随机抽样一致方法(random sample consensus, RANSAC)等鲁棒性的系统模块消除静态或低动态环境中的异常地图点,鲁棒的系统模块可以获得良好的性能。然而,真实环境中的动态对象无处不在,假设场景静态的 SLAM 系统在高度动态的环境中可能完全不可靠。

本文提出一种多焦距动态立体视觉 SLAM,它可以感知更远视野,并且剔除了动态物体对 SLAM 的影响。智能驾驶数据集(the karlsruhe institute of technology and toyota technological institute dataset, KITTI)[18]和自建数据集验证了本文方法的性能。实验结果表明,与传统相同焦距立体 SLAM 系统相比,本文方法具有更好的鲁棒性和精度。在自建多焦距数据集中,本文方法比 DynaSLAM 和 ORB-SLAM3 有更高的精度,时间性能与

DynaSLAM 一致。本文的后续安排如下:第 1 节描述多焦距动态立体视觉 SLAM;第 2 节对多焦距动态立体视觉 SLAM 进行实验验证;第 3 节对本文进行总结。

1 多焦距动态立体视觉 SLAM

本文基于 ORB-SLAM3 提出了一种多焦距动态立体视觉 SLAM。多焦距立体视觉由两个(或多个)焦距差异较大的相机构成,利用成像设备从不同位置获取被测物体的图像,通过图像匹配和三角化得到物体三维信息,多焦距立体视觉可以兼顾较大的检测视野和较远的检测距离。由于焦距较为接近,传统的双目立体视觉 SLAM 在图像金字塔的邻近层中匹配特征点,而多焦距立体视觉不要求两个相机具有相近的焦距,不能只在图像金字塔的临近层匹配特征点,需要增加更多尺度的特征匹配来提高精度。

本文系统的结构如图 1 所示,首先将改进的立体标定方法用于多焦距立体相机的校正,并将相机校正参数用于特征点的立体校正。还提出了一种自适应特征提取和匹配方法增加特征匹配数量,提高了 SLAM 精度。使用实时实例分割网络(you only look at CoefficienTs, YOLACT)[19-20]分割先验动态对象,例如人或车辆,同时还使用语义分割网络(searching for MobileNetV3, MobileNetV3)[21]分割车道区域,用于计算先验动态对象的距离。然后将检测所有先验动态对象是否是动态的,如果它们是动态的,则将剔除动态对象上的特征点。

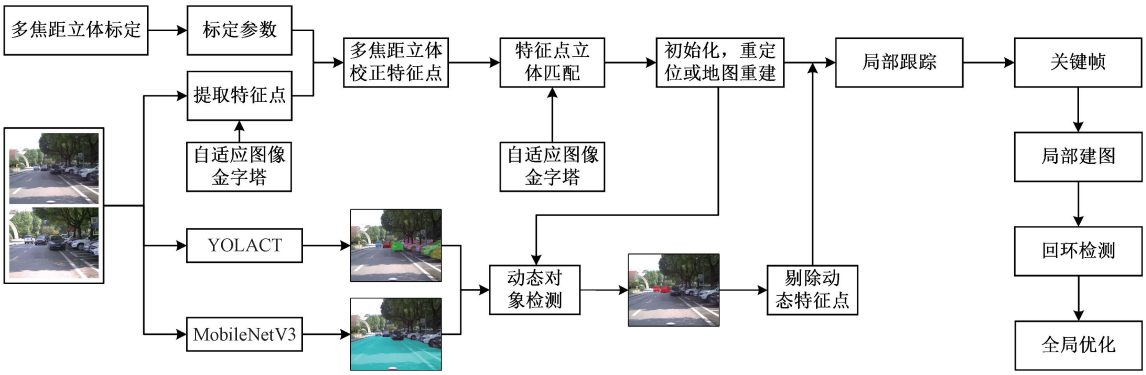


图 1 多焦距动态立体视觉 SLAM 框架

Fig. 1 The architecture of multi-focal length dynamic stereo vision SLAM

1.1 改进的多焦距立体相机校正

标准立体相机图像如图 2 所示,左图像和右图像的焦距相等。假设三维点 P 在标准立体相机左图像的像素坐标为 (u_l, v_l),右图像的像素坐标为 (u_r, v_r),因为水平上极线对齐,所以 $v_l = v_r$,因此可根据视差 $d = u_l - u_r$ 求出三维点 P 深度 $z = f \cdot b/d$,其中 f 为相机的焦距,b 为立体相机的基线距离。

图 2 标准立体视觉相机图像

Fig. 2 The images of the standard stereo camera

多焦距立体相机图像如图 3 所示,多焦距立体相机的左相机为短焦距,右相机为长焦距,右相机整个图像位于左相机图像中的虚线区域。三维点 P 在多焦距立体相机左图像的像素坐标为 (u_l, v_l),在右图像的像素坐标为 (u_r, v_r),其中 $v_l \neq v_r$,因此不能根据标准焦距的视差求出三维点 P 的深度。

图 3 多焦距立体相机图像

Fig. 3 The images of the multi-focal length stereo camera

针对多焦距动态立体视觉 SLAM,本文提出了一种改进的多焦距立体相机的校正方法,它可以校正多焦距立体相机。首先使用单目标定获取短焦距和长焦距相机的焦距 f_{lx}, f_{ly} 和 f_{rx}, f_{ry},改进投影矩阵 $\boldsymbol{P}_l$ 和 $\boldsymbol{P}_r$ 中的 f_x 和 f_y,令 $f_x = f_y = (f_{lx} + f_{ly})/2$,光心 c_x 和 c_y 分别为 $(c_{lx} + c_{rx})/2$、$(c_{ly} + c_{ry})/2$。

在本文方法中没有立体校正图像,而是使用改进标定方法得到的投影矩阵 $\boldsymbol{P}_l$ 和 $\boldsymbol{P}_r$ 对特征点立体校正。该方法预先进行立体标定,得到的立体校正参数作为本文

系统的输入,这种方法避免了短焦距图像的裁剪和长焦距图像的压缩。

1.2　基于自适应图像金字塔的特征提取与匹配

在 ORB-SLAM3 中,图像金字塔主要分为 8 层,每层提取不同数量的特征点,式(1)表示每层提取特征的数量, N 是特征点的总数,s 是金字塔的缩放因子,N_α 是在 α 层中提取的特征点的数量,n 是图像金字塔的层数。特征点匹配将特征点的尺度限制在同一图像金字塔层次或者相差一个图像金字塔层次,超出这个层次范围的匹配将被视为误匹配。

$$N_\alpha = \frac{N(1-s)}{(1-s^n)}s^\alpha \tag{1}$$

多焦距立体图像如图 3 所示。短焦距图像由 16 mm 焦距相机拍摄,而长焦距图像由 25 mm 焦相机拍摄,25 mm 焦距相机的图像是短焦距图像的一部分。本文将长焦距图像在短焦距图像的相同图像内容区域定义为感兴趣区域(region of interest, ROI)。为了获得 ROI 的边界,使用归一化互相关算法[22]计算 ROI 的角点。相关系数最大值对应位置作为 ROI 的角点。在 ROI 区域内的图像,短焦距与长焦距图像组成多焦距立体视觉,在 ROI 区域外的图像被视为单目图像,不能获取其中对象的深度信息。

长焦距和短焦距图像具有不同的尺度,若使用 ORB-SLAM3 的特征提取和匹配方法将增加特征点误匹配数量,因此本文改进了特征提取和匹配算法。

本文提出了一种基于图像金字塔自适应图层提取特征的方法,在图像金字塔的部分图层中提取特征点,这样能支持不同焦距立体相机,同时也支持标准焦距立体相机。图像金字塔的缩放因子由长短焦距比例决定,如式(2)所示。在本文的方法中,λ 的值为 1.7, $th1$ 的值为 1.15,$th2$ 的值为 1.28。提取特征点的图层由相机焦距 f_l, f_r 和比例因子 s 确定。L_{dis} 定义为相同特征在左右图像金字塔中的层次差,如式(3)。多焦距立体特征的匹配策略如图 4 所示,根据图层差 L_{dis},可以确定左图像金字塔的图像层 $Level\_i$ 与右图像金字塔的图像层 $Level\_i + L_{dis}$ 相同。式(4)展示了改进的基于图像金字塔自适应图层提取特征方法,对左图像金字塔,不提取 $n - L_{dis} + 1$ 图层以上的特征点,对右侧图像金字塔,不提取 $L_{dis} - 1$ 图层以下的特征点。

$$s = \begin{cases} th1, & \dfrac{f_r}{f_l} < \lambda \\ th2, & \dfrac{f_r}{f_l} \geqslant \lambda \end{cases} \tag{2}$$

$$L_{dis} = \log_s \frac{f_r}{f_l} \tag{3}$$

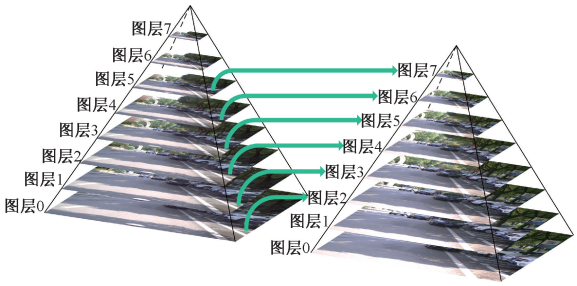


图 4　多焦距立体视觉图像的特征匹配策略

Fig. 4　Features matching strategy of the multi-focal length stereo image

$$N_\alpha = \frac{N(1-s)}{(1-s^{n-L_{dis}+1})}s^\alpha \tag{4}$$

本文还提出一种增加 ROI 区域特征数量的方法,如式(5)所示。该方法应用在图像金字塔的特征提取层中,使用基于四叉树的方法均匀增加左图像 ROI 区域中特征提取的数量,减少 ROI 外提取特征的数量,而不改变特征点的总数。s_{img}^2 是整个图像的面积,s_{roi}^2 是 ROI 的面积, ρ 是 ROI 区域提取特征的增加百分比。在本文的方法中, ρ 的值为 0.4。

$$N_a = \left(N_\alpha \frac{s_{roi}^2}{s_{img}^2}(1+\rho)\right)_{roi} + \left(N_\alpha\left(1-\frac{s_{roi}^2}{s_{img}^2}(1+\rho)\right)\right)_{outer} \tag{5}$$

多焦距立体图像使用 ORB-SLAM3 的立体校正方法会导致短焦距图像的裁剪,长焦距图像的压缩,因此 ORB-SLAM3 的立体特征匹配方法不能直接用于本文系统。本文的立体匹配利用多焦距校正参数对特征点立体校正,优点是不矫正图像,而矫正所有特征点。

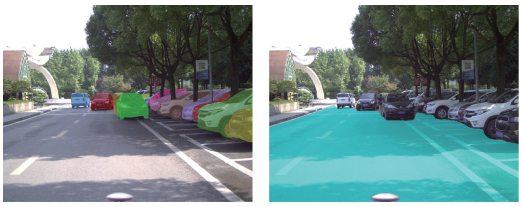
1.3　分割先验动态对象和道路

为了检测动态对象,使用 YOLACT 实例分割先验动态对象,它可以获得像素级的实例分割和标签,本文的分割类别有行人,小汽车,公交车,卡车,自行车,摩托车等,分割的先验动态对象如图 5(a)所示。YOLACT 在 NVIDIA RTX 2060 上的速度超过 25 FPS。

本文方法中检测先验动态对象是否是动态需要计算对象的相对位置,短焦距图像 ROI(如图 4 所示)中的对象可使用立体匹配计算相对位置,但是短焦距图像 ROI 外的对象被视为单目对象,无法计算其精确距离。本文使用逆透视变换(inverse perspective mapping, IPM)方法测量 ROI 区域外物体的位置。为了使用 IPM 计算短焦距图像 ROI 外先验动态对象的相对位置,使用 MobileNetV3 分割车道区域,如图 5(b)所示。

1.4　动态特征点检测和剔除

动态特征点影响着 SLAM 的精度,动态特征点的剔除成为提高精度的关键。YOLACT 用于分割先验动态对象,



(a) YOLACT分割 (b) MobileNetV3分割
(a) YOLACT segmentation (b) MobileNetV3 segmentation

图 5 分割先验动态对象和道路平面

Fig. 5 Segmentation of priori dynamic objects and lane plane

并结合多视图几何、区域特征流和相对距离检测先验对象是否为动态对象。检测到动态对象后,剔除动态对象上的所有特征点,还剔除位于动态对象轮廓边缘的特征点,以消除高梯度区域的不确定性对 SLAM 精度的影响。

1)多视图几何识别动态对象

为计算前一帧和当前帧之间特征点到极线距离,需计算前一帧和当前帧之间的基本矩阵 $\boldsymbol{F}$。由于当前帧的基本矩阵 $\boldsymbol{F}_{cur}$ 未知,因此本文使用前一帧的基本矩阵 $\boldsymbol{F}_{last}$ 计算当前帧的极线距离。在 KITTI 数据集和自建数据集中,图像均以 10FPS 的速度捕获。本文假设当前帧的运动与前一帧的运动相同,则位姿变换矩阵 $\boldsymbol{T}$ 几乎没有变化,或者位姿变换影响特征点到极线距离误差较小。

$$l=\begin{bmatrix}X\\Y\\Z\end{bmatrix}=\boldsymbol{F}_{last}\boldsymbol{P}_1=\boldsymbol{F}_{last}\begin{bmatrix}u_1\\v_1\\1\end{bmatrix} \tag{6}$$

其中,$[X,Y,Z]^{\mathrm{T}}$ 表示极线的向量,$\boldsymbol{F}_{last}$ 是前两帧到前一帧的基础矩阵。由于运动具有连续性,在连续两帧图像中,当前帧图像动态特征点到上一帧特征点极线的距离有一定偏差,因此前一帧和当前帧特征点到极线 l 的距离为 d,如式(7)所示:

$$d=\frac{|\boldsymbol{P}_2^{\mathrm{T}}\boldsymbol{F}_{last}\boldsymbol{P}_1|}{\sqrt{\|X\|^2+\|Y\|^2}} \tag{7}$$

并且计算所有背景特征点的平均距离 d_{mean}^{back} 和对象 k 中所有特征点的平均距离 $d_{mean}^{k\in obj}$:

$$d_{mean}^{back}=\frac{1}{n}\sum_{back}^{n}{}_{i=0}d_i^{back}=\frac{1}{n}\sum_{back}^{n}{}_{i=0}\frac{|\boldsymbol{P}_{i2}^{\mathrm{T}}\boldsymbol{F}_{last}\boldsymbol{P}_{i1}|}{\sqrt{\|X\|^2+\|Y\|^2}} \tag{8}$$

$$d_{mean}^{k\in obj}=\frac{1}{n}\sum_{k\in obj}^{n}{}_{i=0}d_i^{k\in obj}=\frac{1}{n}\sum_{k\in obj}^{n}{}_{i=0}\frac{|\boldsymbol{P}_{i2}^{\mathrm{T}}\boldsymbol{F}_{last}\boldsymbol{P}_{i1}|}{\sqrt{\|X\|^2+\|Y\|^2}} \tag{9}$$

如果 $d_{mean}^{k\in obj}>d_{mean}^{back}\cdot th$ 时,认为这个对象在多视图几何方法中是动态,则将 m_1^k 置为 1,否则为 0,本文的 th 值为 1.2。

2)区域特征流识别动态对象

特征流类似于光流,但是特征流基于 ORB 特征点,ORB 特征点由 FAST 特征点和 BRIEF 描述子构成,而光流法仅由 FAST 特征点构成,BRIEF 描述子使特征流具有更高的精度,并且不需要花费额外时间提取光流点。前一帧和当前帧的特征点像素坐标用于计算特征流的运动距离和方向,特征流的运动距离和运动方向计算方法为:

$$\rho=\sqrt{(u_1-u_2)^2+(v_1-v_2)^2} \tag{10}$$

$$\theta=\arctan((u_1-u_2)/(v_1-v_2)) \tag{11}$$

每个对象的平均运动距离 $\rho_{mean}^{k\in obj}$ 和运动方向 $\theta_{mean}^{k\in mean}$ 为:

$$\rho_{mean}^{k\in obj}=\frac{1}{n}\sum_{k\in obj}^{n}{}_{i=0}\rho_i \tag{12}$$

$$\theta_{mean}^{k\in mean}=\frac{1}{n}\sum_{k\in mean}^{n}{}_{i=0}\theta_i \tag{13}$$

本方法中不计算所有先验静区域的平均特征流距离 ρ_{mean}^{back} 和平均特征流方向 θ_{mean}^{back},仅计算每个对象所在图像列区域内所有先验静态特征点的平均特征流距离 $\rho_{mean}^{kcol\in back}$ 和平均特征流方向 $\theta_{mean}^{kcol\in back}$。

对象 k 的特征流如果满足 $\theta_{mean}^{k\in mean}-\theta_{mean}^{kcol\in back}>45°$ 和 $\rho_{mean}^{k\in obj}-\rho_{mean}^{kcol\in back}>0.2\rho_{mean}^{kcol\in back}$,则认为对象 k 在区域特征流方法中被视为动态对象,则 m_2^r 置为 1,否则为 0。

3)相对距离识别动态对象

当对象位于 ROI 区域内时使用多焦距立体视觉测量距离,当对象位于 ROI 边界上或超出 ROI 时,使用自适应 IPM 测量距离。

逆透视矩阵 $\boldsymbol{H}$ 由 MobileNetV3 分割区域和该区域的特征点求解,逆透视矩阵 $\boldsymbol{H}$ 用于计算 ROI 外对象的相对距离。位于道路掩码区域的特征点在相机坐标系下的坐标为 $\boldsymbol{P}_i=[X_i,Y_i,Z_i]^{\mathrm{T}}$,对应的像素坐标系坐标为 $p_i=(u_i,v_i)$,则逆透视变换方程为:

$$\begin{pmatrix}X_i\\Z_i\\1\end{pmatrix}=\boldsymbol{H}\begin{pmatrix}u_i\\v_i\\1\end{pmatrix} \tag{14}$$

将分割的所有先验动态对象使用逆透视矩阵 $\boldsymbol{H}$ 变换,然后计算所有先验动态对象相对于相机的距离。每个先验动态对象掩码逆透视变换后在相机坐标系中离 Z 轴最近的点为距离点,最后得到 ROI 外对象 k 的距离 D^k。

利用前一帧和当前帧之间成功匹配的特征点关联每个掩码,将它们视为同一个对象。对象 r 在前一帧图像相机坐标系的距离为 D_{last}^r,对象 k 在当前帧图像相机坐标系的距离为 D_{cur}^k,对象 r 和 k 为同一对象。相对距离判断动态对象如式(15)所示,其中 t_z 为前一帧到当前帧在相机坐标系 Z 轴方向的运动距离。如果 $|D|>0.5|t_z|$,在相对距离识别动态对象方法中这个对象将被视为动态对象,则 m_3^k 置为 1,否则为 0。

$$D = D^{r}_{last} - D^{k}_{cur} - t_z \tag{15}$$

4)剔除动态特征点

当动态对象和相机向不同的方向移动时,多视图几何可以很好地判断先验动态对象是否是动态的。当对象和相机在同一方向移动时,使用距离方法确定先验动态对象是否动态。因此,多视图几何与相对距离方法之间有很好的互补性。

本文提出的实现算法如下:

算法 1　动态对象检测算法

输入:图像 I_i,所有特征点 p^{all}_{ni},基本矩阵 $\boldsymbol{F}_{last}$

输出:静态特征点 p^{static}_{ni}

1)采用 YOLACT 分割先验动态对象和 MobileNetV3 分割车道区域;

2)采用多视图几何、区域特征流和相对距离检测先验对象,得到 m^k_1, m^k_2, m^k_3 的值;

3)如果($m^k_1 \&\& m^k_3 \,||\, m^k_2$) 为真则将第 k 对象判定为动态对象,否则为静态对象;

4)重复步骤 2)~3),直到所有对象都判定完毕,位于动态对象上和边缘的特征点都剔除。

2　实验验证

据集上与 ORB-SLAM3 和 DynaSLAM 进行比较。在实验中,从所有图像中提取了 2 000 个特征点,并且所有实验均在 Intel i7-8700 CPU、NVIDIA RTX 2060 GPU、16 G 内存的计算机上进行。因为现有公共数据库没有多焦距立体图像,因此本文使用多焦立体相机采集自建数据集。自建数据集由 3 个摄像头同时收集,相机 0 和相机 1 组合是标准焦距立体视觉,相机 1 和相机 2 组合是多焦距立体视觉,它们的基线距离都为 0.45 m。同时利用实时动态定位(real time kinematic, RTK)卫星定位系统以 10 Hz 记录经度、纬度和海拔信息作为真实轨迹评估 SLAM 的精度。相机的安装位置如图 6 所示,左侧相机设置为相机 0,中间相机为相机 1,右侧相机为相机 2,立体相机的基线为 0.45 m,采样频率设置为 10 Hz。作为对比,相机 0 和相机 1 镜头的焦距相同,相机 2 镜头的焦距大于相机 0 和相机 1。自建数据集的图片像素大小为 1 218×962。图 7 为对应的多焦距图像,第 1 列为相机 0 的图像,第 2 列为相机 1 的图像,第 3 列为相机 2 的图像。第 1 行为 6 mm 镜头和 12.5 mm 镜头拍摄图像,第 2 行为 16 mm 镜头和 25 mm 镜头拍摄图像。

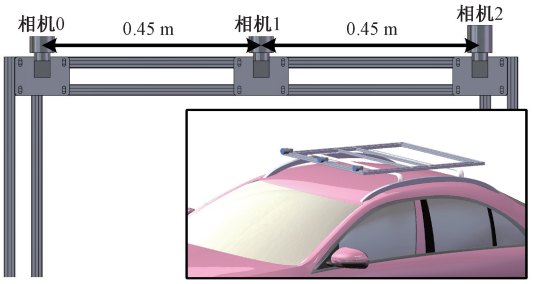


图 6　实验相机布局

Fig. 6　Experimental camera layout

图 7　自建数据集图像

Fig. 7　The images of our dataset

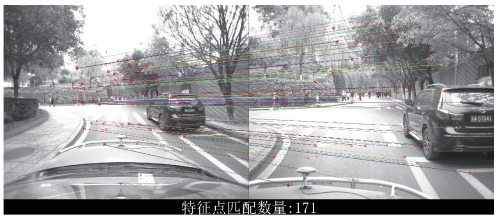
2.1　特征提取和匹配

为了验证基于自适应图像金字塔的特征提取和匹配方法,实验在室外环境中进行,相机安装在实验车上。使用两种不同焦距的镜头组合获得不同焦距组合数据集,分别是 6 和 12.5 mm、16 和 25 mm,真实轨迹由 RTK 同时采集。如表 1 所示为 3 种特征提取与匹配算法的实验结果。首先是 ORB-SLAM3 中的原始方法直接应用于立体图像,其特征立体匹配如图 8(a)所示,二是在 ORB-SLAM3 中的原始方法中增加 ROI 区域特征点数量,其特征立体匹配如立体匹配如图 8(b)所示,三是自适应图层提取特征与 ROI 区域增加提取特征数量两种方法结合,其特征立体匹配如图 8(c)所示。

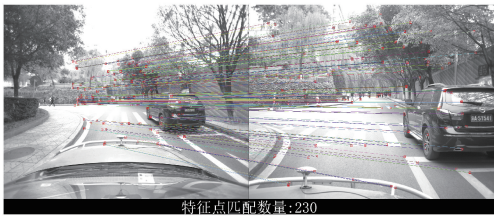
表 1　立体匹配结果

Table 1　Stereo matching results　　%

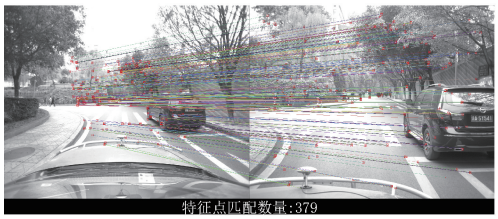
焦距组合	相机 1 和相机 2			相机 0 和相机 2		
	ORB-SLAM3	ROI	本文方法	ORB-SLAM3	ROI	本文方法
6 和 12.5 mm	188	27.19	79.62	141	34.95	92.42
16 和 25 mm	309	8.87	32.23	233	8.61	31.61



(a) ORB-SLAM3的特征提取与匹配
(a) ORB-SLAM3's feature extraction andmatching



(b) ROI中的特征提取和匹配
(b) Feature extraction and matching in ROI



(c) 本文特征提取和匹配
(c) Our feature extraction and matching

图 8 立体匹配结果比较

Fig. 8 Comparison of stereo matching results

实验证明自适应图层提取特征与 ROI 区域增加提取特征数量结合可以大大提高立体匹配的数量。与 ORB-SLAM3 的立体匹配数量相比,6 和 12.5 mm 镜头组合的平均增幅为 86.01%,16 和 25 mm 镜头的平均增幅为 31.97%。6 和 12.5 mm 镜头效果好于 16 和 25 mm 的原因在于一方面 6 和 12.5 mm 的 ORB-SLAM3 原始平均匹配数量低于 16 和 25 mm,另一方面 6 和 12.5 mm 的图像视差比高于 16 和 25 mm。

2.2 KITTI 数据集实验

KITTI 数据集常用于评估基于视觉和激光的里程计或 SLAM 的数据集。本文选择 KITTI 的 00-07 序列数据集进行对比实验,其具有真实姿态,可与 SLAM 姿态对比评估。本文使用绝对轨迹误差(absolute trajectory error, ATE)被用作评估指标,ATE 是一种评估 SLAM 的流行方法,它测量了相应时间的真实姿态和估计姿态之间的欧几里德距离,此外本文还使用平均跟踪时间评估本文系统的速度。

ORB-SLAM3、DynaSLAM 和本文方法在 KITTI 数据集 00~07 序列中的平均跟踪时间和 ATE 的均方根误差(root mean squared error of ATE, ATE RMSE)比较结果如表 2 所示。与 ORB-SLAM3 相比,本文方法在 00、01、02、03、07 序列的 ATE RMSE 较低,平均 ATE RMSE 降低了 6.97%。与 DynaSLAM 相比,本文方法在 00、01、02、07 的 ATE RMSE 较低,平均 ATE RMSE 降低了 6.97%。

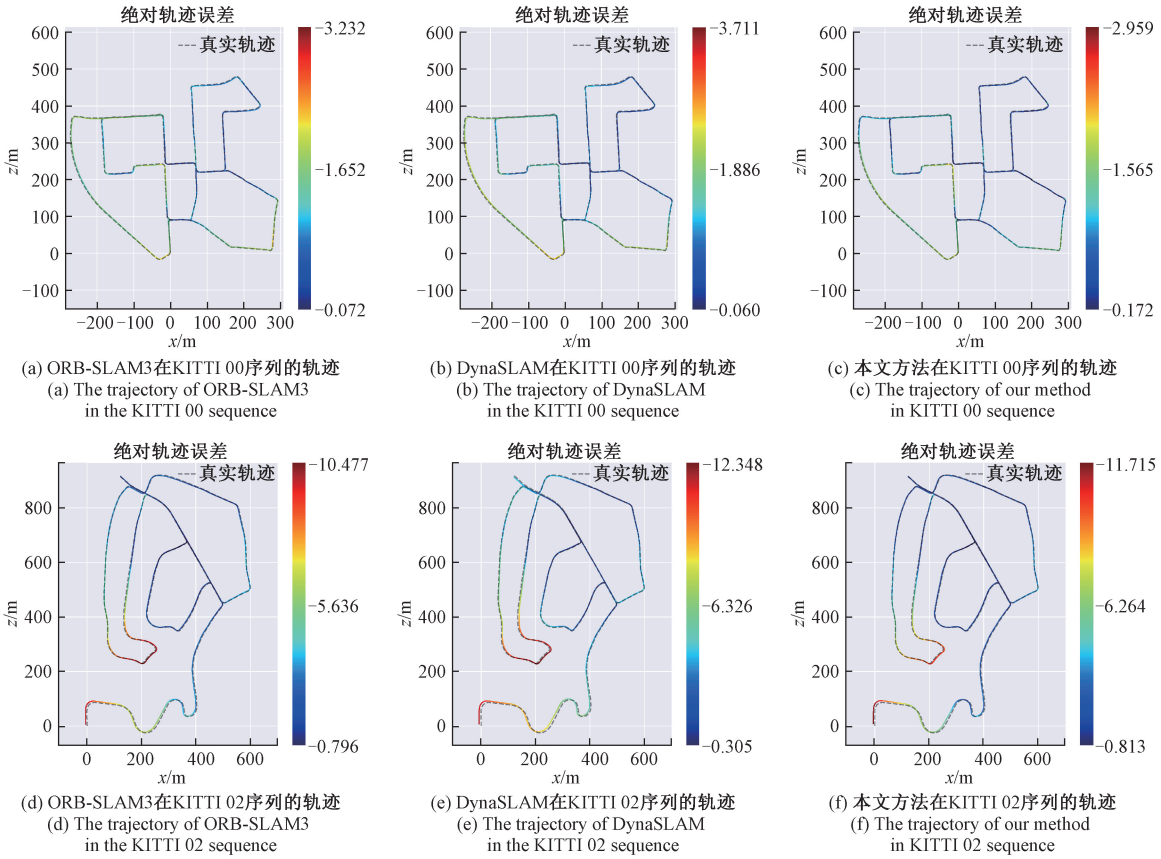
表 2 在 KITTI 数据集中实验

Table 2 Experiments in the KITTI dataset

序列	长度/m	ORB-SLAM3		DynaSLAM		本文方法	
		时间/ms	误差/m	时间/ms	误差/m	时间/ms	误差/m
00	3 724	65.15	1.25	92.60	1.37	111.69	**1.02**
01	2 453	86.78	10.15	117.58	9.61	129.14	**8.00**
02	5 067	65.32	4.58	94.38	5.59	107.61	**4.55**
03	561	67.73	1.33	87.13	**0.77**	103.26	1.24
04	394	67.51	**0.21**	87.11	0.23	107.03	0.60
05	2 206	66.09	0.85	92.51	**0.73**	107.87	1.11
06	1 233	73.28	**0.67**	97.22	0.69	111.83	1.18
07	695	62.96	0.50	80.01	0.53	105.12	**0.45**
均值	2 042	69.35	2.44	93.57	2.44	110.44	**2.27**

2.3 自制数据集实验

在图 9 中显示了 KITTI 序列 00 和 02 数据集的实验结果,图 9(a)和图 9(d)为 ORB-SLAM3 跟踪轨迹与真实轨迹的对比,图 9(b)和图 9(e)为 DynaSLAM 跟踪轨迹与真实轨迹的对比,图 9(c)和图 9(f)为本文方法的跟踪轨迹与真实轨迹的对比。在标准焦距立体视觉数据集的 6 个序列的实验中,如表 3 所示,本文方法的平均 ATE RMSE 比 ORB-SLAM3 降低了 17.64%,比 DynaSLAM 降低了 23.77%。在多焦距立体视觉数据集的 6 个序列的实验中,如表 3 所示,多焦距动态立体视觉 SLAM 的平均 ATE RMSE 比 ORB-SLAM3 降低了 26.64%,比 DynaSLAM 降低了 32.09%。



(a) ORB-SLAM3在KITTI 00序列的轨迹
(a) The trajectory of ORB-SLAM3 in the KITTI 00 sequence

(b) DynaSLAM在KITTI 00序列的轨迹
(b) The trajectory of DynaSLAM in the KITTI 00 sequence

(c) 本文方法在KITTI 00序列的轨迹
(c) The trajectory of our method in KITTI 00 sequence

(d) ORB-SLAM3在KITTI 02序列的轨迹
(d) The trajectory of ORB-SLAM3 in the KITTI 02 sequence

(e) DynaSLAM在KITTI 02序列的轨迹
(e) The trajectory of DynaSLAM in the KITTI 02 sequence

(f) 本文方法在KITTI 02序列的轨迹
(f) The trajectory of our method in KITTI 02 sequence

图 9　SLAM 和真实 GPS 轨迹比较

Fig. 9　Comparison with SLAM and real GPS trajectories

表 3　自建数据集的 ATE RMSE

Table 3　ATE RMSE in our dataset　m

序列	长度	同焦距			多焦距
		ORB-SLAM3	DynaSLAM	本文方法	本文方法
00	510	**1. 53**	1. 75	1. 68	1. 65
01	1 591	5. 97	8. 47	**4. 41**	**3. 31**
02	1 326	**2. 95**	3. 18	3. 09	3. 04
03	540	1. 52	**1. 27**	1. 82	1. 67
04	1 551	4. 72	4. 27	**3. 51**	**3. 15**
05	1 317	4. 11	3. 53	**2. 62**	**2. 44**
均值	1 139	3. 47	3. 74	**2. 85**	**2. 54**

表 4　自建数据集的平均跟踪时间

Table 4　The average computation time of our dataset　ms

序列	同焦距			多焦距
	ORB-SLAM3	DynaSLAM	本文方法	本文方法
00	86. 24	148. 87	124. 88	135. 15
01	83. 28	139. 33	128. 76	144. 51
02	82. 59	144. 20	135. 57	142. 41
06	90. 14	146. 74	145. 98	153. 41
07	88. 61	145. 75	143. 34	148. 88
08	92. 92	148. 16	145. 74	144. 29
均值	87. 30	145. 51	137. 38	144. 76

本文方法的速率约为 6. 7 FPS,与 DynaSLAM 的速度相同,自建数据集的平均跟踪时间如表 4 所示。值得注意的是,多焦距动态立体视觉 SLAM 在 KITTI 数据集中的速度可以达到 10 FPS。实验证明,本文的定位精度高于 ORB-SLAM3 和 DynaSLAM,时间性能与 DynaSLAM 一致。

3　结　　论

本文提出了一种可以应用于标准立体相机和多焦距立体相机的多焦距动态立体视觉 SLAM,它克服了标准立体相机无法兼顾宽视场和远距离感知场景的缺点,并去除了动态物体对 SLAM 的影响。

首先对传统的立体校正方法进行了改进,校正参数用于特征点立体校正。其次改进的自适应图像金字塔特征提取和匹配方法大大增加了立体匹配的特征数量,提高定位精度和稳定性。同时,YOLACT 分割先验动态对象,MobileNetV3 分割车道区域得到车道平面,使用 IPM 方法计算的得到 ROI 区域外对象的距离。最后使用多视图几何、区域特征流和相对距离对先验动态对象检测,剔除动态对象上和其边缘的特征点。

在 KITTI 数据集实验中,本文方法比 ORB-SLAM3 和 DynaSLAM 具有更好的鲁棒性和准确性,定位精度都提高 6.97%。在自建数据集中,本文方法同样比 ORB-SLAM3 和 DynaSLAM 具有更高的定位精度,时间性能与 DynaSLAM 几乎一致。更重要的是,因为短焦距和长焦距相机的结合,本文方法可以感知宽视野和更远距离的场景。在今后的工作中,我们将充分利用该方法检测广视野、范围远的优势,实现大范围、远距离区域内的更精准 SLAM 和动态目标跟踪。

参考文献

[1] 李帅鑫,李广云,周阳林,等. 改进的单目视觉实时定位与测图方法[J]. 仪器仪表学报,2017,38(11):2849-2857.
LI SH X, LI G Y, ZHOU Y L, et al. Improved monocular simultaneous localization and mapping solution[J]. Chinese Journal of Scientific Instrument, 2017, 38(11): 2849-2857.

[2] 孙曼晖,杨绍武,易晓东,等. 基于 GIS 和 SLAM 的机器人大范围环境自主导航[J]. 仪器仪表学报,2017,38(3):586-592.
SUN M H, YANG SH W, YI X D, et al. Autonomous navigation of robot in large-scale environments based on GIS and SLAM [J]. Chinese Journal of Scientific Instrument, 2017, 38(3): 586-592.

[3] DAVISON A J, REID I D, MOLTON N D, et al. MonoSLAM: Real-time single camera SLAM[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6): 1052-1067.

[4] KAMEDA Y. Parallel tracking and mapping for small AR workspaces (PTAM)[J]. 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, 2012, 66(1): 45-51.

[5] MUR-ARTAL R, MONTIEL J M M, TARDOS J, et al. ORB-SLAM: A versatile and accurate monocular SLAM system [J]. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.

[6] MUR-ARTAL R, TARDOS, JUAN D. ORB-SLAM2: An open-source SLAM system for monocular, stereo and RGB-D cameras [J]. IEEE Transactions on Robotics, 2016, 33(5): 1255-1262.

[7] CAMPOS C, ELVIRA R, RODRÍGUEZ J J G, et al. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM [J]. IEEE Transactions on Robotics, 2021, DOI: 10.1109/TRO.2021.3075644

[8] ENGEL J, SCHPS T, CREMERS D. LSD-SLAM: Large-scale direct monocular SLAM[J]. European Conference on Computer Vision Springer, 2014, 8690: 834-849.

[9] SCHLEGEL D, COLOSI M, GRISETTI G, et al. ProSLAM: Graph SLAM from a programmer's perspective[C]. 2018 IEEE International Conference on Robotics and Automation, 2018: 3833-3840.

[10] ZHAO Y, XU S B, BU S H, et al. GSLAM: A general SLAM framework and benchmark[C]. 2019 IEEE/CVF International Conference on Computer Vision. Los Alamitos. 2019: 1110-1120.

[11] CUI L Y, MA C W. SOF-SLAM: A semantic visual SLAM for dynamic environments [J]. IEEE Access, 2019, 7: 166528-166539.

[12] CUI L Y, MA C W. SDF-SLAM: Semantic depth filter SLAM for dynamic environments [J]. IEEE Access, 2020, 8: 95301-95311.

[13] ZHANG X, PENG Y, YANG M, et al. Moving object detection for camera pose estimation in dynamic environments[C]. 2020 10th Institute of Electrical and Electronics Engineers International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER). 2020: 93-98.

[14] BESCÓS B, FÁCIL J M, CIVERA J, et al. DynaSLAM: Tracking, mapping and inpainting in dynamic scenes[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 4076-4083.

[15] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]. 16th IEEE International Conference on Computer Vision (ICCV), Venice, ITALY. 2017: 2980-2988.

[16] BESCÓS B, CAMPOS C, TARDÓS J D, et al. DynaSLAM II: Tightly-coupled multi-object tracking and SLAM [J]. IEEE Robotics and Automation Letters, 2020, 6(3): 5191-5198.

[17] 贾松敏,郑泽玲,张国梁,等. 基于混合特征的机器人定位与地图创建[J]. 仪器仪表学报, 2018, 39(12): 201-209.
JIA S M, ZHENG Z L, ZHANG G L, et al. Robot localization and map building based on hybrid features[J]. Chinese Journal of Scientific Instrument, 2018, 39(12): 201-209.

[18] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: The KITTI dataset [J]. Int J Robot Res, 2013, 32(11): 1231-1237.

[19] BOLYA D, ZHOU C, XIAO F Y, et al. YOLACT real-time instance segmentation [C]. 2019 IEEE/CVF International Conference on Computer Vision IEEE. 2019: 9156-9165.

[20] BOLYA D, ZHOU C, XIAO F, et al. YOLACT++: Better real-time instance segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, DOI: 10.1109/TPAMI.2020.3014297

[21] HOWARD A, SANDLER M, CHU G, et al. Searching for MobileNetV3 [C]. 2019 IEEE/CVF International Conference on Computer Vision. Los Alamitos, IEEE Computer Soc. 2019: 1314-1324.

[22] WU P, LI W, SONG W L. Fast, accurate normalized cross-correlation image matching [J]. Journal of Intelligent & Fuzzy Systems, 2019, 37(4): 4431-4436.

作者简介

冯明驰(通信作者),分别 2008 年和 2014 年于中国科学技术大学获得学士学位和博士学位,现为重庆邮电大学副教授,主要研究方向为视觉测量、智能汽车环境感知。

E-mail: fengmc@cqupt.edu.cn

Feng Mingchi (Corresponding author) received his B. Sc. degree and Ph. D. degree both from University of Science and Technology of China in 2008 and 2014, respectively. He is currently an associate professor at Chongqing University of Posts and Telecommunications. His research interests include vision measurement and environmental perception of intelligent vehicles.

刘景林,2019 年于重庆邮电大学获得学士学位,现为重庆邮电大学硕士研究生,主要研究方向为计算机视觉。

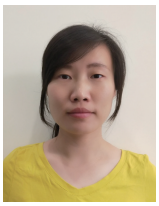
E-mail: jinlnl@163.com

Liu Jinglin received his B. Sc. degree from Chongqing University of Posts and Telecommunications in 2019. He is currently a master student at Chongqing University of Posts and Telecommunications. His research interest is Computer Vision.

李成南,2020 年于重庆邮电大学获得学士学位,现为重庆邮电大学硕士研究生,主要研究方向为计算机视觉。

E-mail: lichengnan1998@163.com

Li Chengnan received his B. Sc. degree from Chongqing University of Posts and Telecommunications in 2020. He is currently a master student at Chongqing University of Posts and Telecommunications. His research interest is Computer Vision.



汪静姝,分别 2008 年和 2013 年于中国科学技术大学获得学士学位和博士学位,现为重庆理工大学副教授,主要研究方向为视觉测量、智能制造。

E-mail: donotbreeze@163.com

Wang Jingshu received her B. Sc. degree and Ph. D. degree both from University of Science and Technology of China in 2008 and 2013, respectively. She is currently an associate professor at Chongqing University of Technology. Her research interests include vision measurement and intelligent manufacturing.