DOI:10.19651/j.cnki.emt.2518114

基于运动概率筛选和加权位姿估计的鲁棒动态 RGB-D SLAM*

于兴云[1,2] 程向红[1,2] 刘丰宇[1,2] 钟志伟[1,2]
(1.东南大学仪器科学与工程学院 南京 210096;2.微惯性仪表与先进导航技术教育部重点实验室 南京 210096)

摘 要:为减小动态物体对视觉 SLAM 的干扰,提出一种基于运动概率筛选和加权位姿估计的鲁棒动态 RGB-D SLAM。首先,利用实例分割网络 Yolact 获取场景的语义信息,结合语义信息和深度信息对动态掩膜边界修复,根据先验运动概率的大小计算语义动态概率。然后,采用基于语义引导的方法,计算特征点的几何动态概率,将语义动态概率和几何动态概率及其置信度,通过加权融合的方式构造特征点的运动概率模型,并设计具有自适应概率阈值的特征点筛选策略。最后,在系统的位姿跟踪、局部地图优化、全局优化过程中,设计基于特征点运动概率的加权代价函数,以区分不同特征点对位姿优化的贡献。此外,在移除动态物体之后,对静态场景建立全局点云地图。公开数据集的实验结果表明,相较于 ORB-SLAM2,所提算法在 TUM RGB-D 和 Bonn 数据集上的绝对轨迹误差的均方根误差分别平均降低 69.16%和 91.94%;与其他先进的动态 SLAM 算法相比,所提算法的位姿估计精度和鲁棒性均有一定程度的提升。在真实场景实验中,相较于 ORB-SLAM2、Dyna-SLAM,轨迹端点漂移误差分别平均降低 52.20%、19.15%。

关键词:RGB-D SLAM;动态物体;运动概率;加权位姿估计;全局点云地图
中图分类号:TP391.41;TN98 **文献标识码**:A **国家标准学科分类代码**:510.99

Robust dynamic RGB-D SLAM based on motion probability screening and weighted pose estimation

Yu Xingyun[1,2] Cheng Xianghong[1,2] Liu Fengyu[1,2] Zhong Zhiwei[1,2]
(1. School of Instrument Science & Engineering, Southeast University, Nanjing 210096, China; 2. Key Laboratory of Micro-inertial Instrument and Advanced Navigation Technology, Ministry of Education, Nanjing 210096, China)

Abstract: In order to reduce the interference of dynamic objects on visual SLAM, a robust dynamic RGB-D SLAM that combines the motion probability of feature point and weighted pose estimation is proposed. First, the instance segmentation network Yolact is used to obtain semantic information of scene, combine semantic information and depth information to restore the dynamic mask boundaries, and calculate the semantic dynamic probability according to the magnitude of the prior motion probability. Then, a semantically guided method is used to calculate the geometric dynamic probability of feature point, and the semantic dynamic probability, the geometric dynamic probability and their confidence are combined to construct the motion probability of the feature point, and a feature point screening strategy with adaptive probability threshold is designed. Finally, in the process of pose tracking, local map optimization, and global optimization of the system, a weighted cost function based on the motion probability of feature point is designed to distinguish the contribution of different feature points to pose optimization. In addition, after removing the dynamic objects, a global point cloud map is established for static scenes. Experimental results on the public datasets demonstrate that, compared with ORB-SLAM2, the Root Mean Square Error of Absolute Trajectory Error of the proposed algorithm on the TUM RGB-D and Bonn datasets is reduced on average by 69.16% and 91.94%, respectively. Moreover, compared with other state-of-the-art dynamic SLAM algorithms, the proposed method exhibits noticeable improvements in both pose estimation accuracy and robustness. In real-world experiments, compared with ORB-SLAM2 and Dyna-SLAM, the trajectory endpoint drift error is reduced by an average of 52.20% and 19.15% respectively.

Keywords: RGB-D SLAM; dynamic objects; motion probability; weighted pose estimation; global point cloud map

0 引 言

同步定位与建图(simultaneous localization and mapping, SLAM)是移动机器人实现自主定位和环境感知的核心技术,广泛应用于增强现实、自动驾驶、无人机等各种领域。随着实际应用对场景感知需求的快速增长,基于

收稿日期:2025-02-20
* 基金项目:国家自然科学基金(62273091)、国网江苏省电力有限公司省管产业单位科技项目(JC2024074)资助

相机的场景建图需求日益增加，这使得视觉 SLAM 的重要性愈发凸显。目前，已经涌现诸多经典的视觉 SLAM 框架，例如 ORB-SLAM2[1] 和 VINS-Mono[2]。然而，传统的视觉 SLAM 建立在场景静止的假设之上，这一假设在实际应用中具有较大的局限性[3]。动态对象会破坏特征匹配和位姿优化的过程，而这两个过程对 SLAM 系统的正常运行至关重要。

动态视觉 SLAM 系统通过设计有效的策略以减轻动态对象对系统的影响，从而确保其在动态场景的优异表现。目前，常见的解决方案分为 3 类：纯几何方法、纯深度学习方法、几何和深度学习相结合的方法。

一些研究人员采用纯几何方法来解决动态环境中的 SLAM 问题。这些方法主要依赖空间和时间一致性来应对动态物体带来的挑战。Zou 等[4]通过协方差矩阵来维护每个特征点的位置不确定性，并在新观测值出现时迭代细化地图点的位置。此外，利用地图点的三角测量一致性，在每一帧中，将地图点分类为动态点和静态点。Kim 等[5]提出了一种基于深度场景估计的非参数背景模型，采用结合能量驱动的方法来估计相机的自运动。Dai 等[6]提出采用 Delaunay 三角剖分生成稀疏地图，并根据地图点的相关性对其进行分组来区分动态点和静态点。与以往仅检测短时间跨度的连续帧之间的动态分量不同，LC-CRF-SLAM[7]通过图割(graph cut，GC)-随机采样一致性(random sample consensus，RANSAC)方法确定内点，初步估计相机位姿，并通过构建长期一致性的条件随机场模型来辅助 3D 动态路标检测，从而区分静态/动态路标点。但是在面临大面积动态特征点时，该方法通过人工设置阈值，无法处理新场景问题，导致处理动态目标的能力严重下降。

基于深度学习的方法通过卷积神经网络从图像中识别先验动态对象，例如人和车辆。Dynamic-DSO[8]通过 Mask R-CNN[9]模型来检测高动态性物体。为了保留静态背景，系统仅在特征点四邻域内的像素点不属于先验动态掩膜时，才将其标记为静态特征点。尽管深度学习方法在某些特定场景中取得不错的效果，但其性能受到模型精度限制。此外，这些方法只能识别预定义的动态对象，缺乏对未定义的潜在动态对象的抗干扰能力。

近年来，学者们尝试将几何和深度学习方法相结合，以弥补各自的局限性。Dyna-SLAM[10]结合 Mask R-CNN[9]和多视图几何来处理人或其他物体上的动态点，同时修复被遮挡的背景以构建静态 3D 地图。然而，该方法的计算成本较高，难以满足实时性要求，并且存在剔除过多特征点导致跟踪失败的问题。DS-SLAM[11]采用轻量级分割网络 SegNet[12]来分割对象，通过对极几何约束检测动态特征点，但是该约束不能处理运动物体沿极线方向运动的情况。为了解决语义分割存在分割不准确的问题，Blitz-SLAM[13]使用原始掩膜和深度信息来修复掩膜以更准确地覆盖动态对象。OVD-SLAM[14]通过融合语义信息、深度信息和光流来区分前景和背景，并通过多帧之间的平均重投影误差来恢复运动物体上的静态点，从而适应非刚性运动和低动态环境。文献[15]通过大量实验研究特征点匹配距离与极线距离之间的关系，提出了一种自适应阈值方法来剔除动态特征点。文献[16]则结合稀疏场景流和马氏距离来剔除动态特征点，并对保留的特征点分配权重。

综上所述，基于几何方法的动态视觉 SLAM 依赖人工设置阈值，在物体大范围运动时，处理异常特征点的性能严重下降。深度学习方法需要预先定义物体的类别，但在实际场景中，一些静止的物体也可能运动。文献[7,10,15]在处理大面积的动态特征点时，性能较差。为了筛选动态特征点，可以采用特征点运动概率的方法。文献[14]采用统计学的方法将深度学习输出的检测框和深度信息结合来进行前景和背景区分。在计算特征点运动概率的时候，该方法只考虑检测框的先验运动性和深度信息，未考虑到物体的实际运动性，而且该方法严重依赖前景和背景的分割结果。文献[16]采用语义分割和光流结合的方式筛选动态特征点，但未考虑在处理物体交界处时，常常出现分割精度下降问题，进而导致筛选特征点出现错误。为了解决上述问题，本文提出一种基于运动概率筛选和加权位姿估计的鲁棒动态 RGB-D SLAM 方法，主要创新点如下：

1）利用实例分割网络 Yolact[17]获取语义先验信息。为了提高动态区域边界的分割精度，采用掩膜图像和深度图像结合对动态掩膜边界修复，然后根据先验运动概率的大小计算语义动态概率。

2）为了有效筛选动态特征点，结合语义动态概率、几何动态概率及其置信度来更新特征点的运动概率，并根据运动概率的大小进行自适应概率阈值筛选。

3）考虑不同特征点的动态特性来平衡特征点对代价函数的影响，改进后端优化的代价函数，以提供更准确的定位精度，同时，对静态环境建立稠密点云地图。

1　系统设计

所提算法基于 ORB-SLAM2 框架进行开发，该系统的框架如图 1 所示，包括五大线程：1）追踪线程；2）实例分割线程；3）局部建图线程；4）静态场景重建线程；5）回环线程。加粗的黑色虚线框内为主要工作。系统通过 RGB-D 相机获取 RGB 图像和深度图像，提取 ORB 特征点进行匹配。采用 Yolact 模型检测图像中潜在动态对象来获取检测框和语义分割掩膜。设计基于运动概率筛选和加权位姿估计的方案：首先，结合掩膜图像和深度图像对动态掩膜边界修复，通过运动概率计算模块评估特征点的运动概率；然后，通过自适应概率阈值筛选特征点；最后，将剩余特征点及其权重整合到后端优化模块，包括帧间加权位姿估计、局部加权捆绑调整(bundle adjustment，BA)优化、全局加权 BA 优化，联合优化相机的位姿以及权重。

静态场景重建线程解决动态物体对建立环境地图造成的 重影问题。

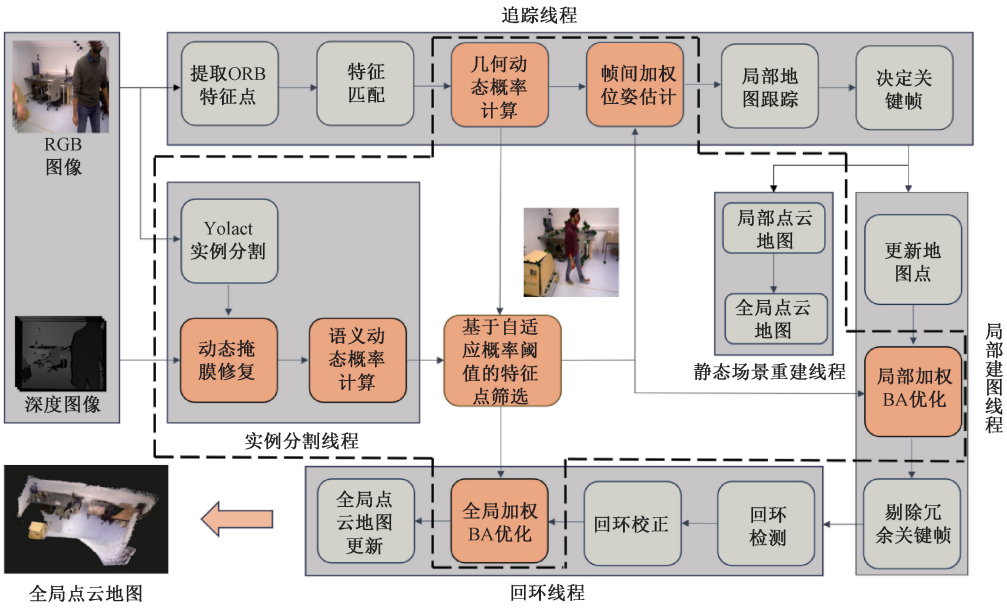


图 1 系统框架

Fig. 1 System framework

2 基于运动概率的动态特征点筛选

根据先验知识,SLAM 场景中的常见元素,如人和汽车,通常被视为潜在的动态目标。基于深度学习的方法将潜在动态对象边界框内的所有特征点标记为动态,这在低动态场景中表现较差。人体作为非刚性物体,常进行局部运动,将人体对象上的所有特征点删除会导致后端优化性能下降。

2.1 基于 Yolact 的语义动态概率估计

深度学习网络 Yolact 模型已经在 MS COCO 数据集上进行了训练,该数据集包含 80 个类别,涵盖多种常见的动态对象。该模型对输入的彩色图像进行处理,输出每个对象的检测框、掩膜、标签和置信度。选取室内场景下较为常见的 5 个类别:人、鼠标、椅子、显示器和背景。针对这 5 种对象,分别为它们分配不同的先验运动概率,如图 2 所示。输入图像的第 i 个特征点的先验运动概率 M_i^{obj} 由 Yolact 获取的对象类别决定,$M_i^{\text{obj}} \in [0,1]$。当 $M_i^{\text{obj}} > 0.5$ 时,认为该点为动态特征点;反之,则为静态特征点。Yolact 在分割动态对象的边界时,常出现像素级偏差,尤其在物体交界处,部分动态区域被误判为静态区域。

图 2 室内物体的先验运动概率划分

Fig. 2 Classification of prior movement probability of indoor objects

为了减小错误分割对后续特征匹配和位姿估计的影响,首先对动态掩膜边界进行修复,然后引入特征点的语义动态概率模型。由于背景和前景的深度差异显著,例如人和墙壁,因此利用检测框的 4 个角点和中心点的深度构造深度阈值,并根据此阈值判断当前动态掩膜边界像素点 5×5 邻域的像素点是否为动态点。定义检测框的 4 个角点的平均深度为 d_{avg},中心点的深度为 d_{c},深度阈值为 d_{th},如式(1)。通过对深度差异 $|d_{\text{avg}}-d_{\text{c}}|$ 判断,增加一层噪声过滤,以减少异常值的影响。增量 $\alpha\cdot|d_{\text{avg}}-d_{\text{c}}|$ 动态依赖深度差异,以适应不同的深度分布,τ 为深度阈值。

$$d_{\text{th}}=\begin{cases}\dfrac{1}{2}(d_{\text{avg}}+d_{\text{c}})+\alpha\cdot|d_{\text{avg}}-d_{\text{c}}|, & d_{\text{c}}>0\cap|d_{\text{avg}}-d_{\text{c}}|>\tau\\ d_{\text{c}}+\alpha\cdot|d_{\text{avg}}-d_{\text{c}}|, & d_{\text{c}}>0\cap|d_{\text{avg}}-d_{\text{c}}|\leqslant\tau\\ d_{\text{avg}}, & d_{\text{c}}=0\\ \infty, & \text{其他}\end{cases} \tag{1}$$

对于 5×5 像素块中的任意像素点,其深度为 d,如果 $d\leqslant d_{\text{th}}$,则将该像素点归入动态掩膜区域。为了计算语义动态概率,定义动态掩膜内的特征点集合和动态掩膜边界像素点集合分别为 $\mathbb{P}=\{\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_m\}$,$\mathbb{B}=\{\boldsymbol{b}_1,\boldsymbol{b}_2,\cdots,\boldsymbol{b}_n\}$。对于动态掩膜内的任一特征点 $\boldsymbol{p}_i$ 到边界像素点的欧几里得距离为 $dist_i$,如式(2)所示。

$$dist_i=\min_{\boldsymbol{b}_j\in\boldsymbol{B}}\|\boldsymbol{p}_i-\boldsymbol{b}_j\|_2 \tag{2}$$

对于该距离,采用指数映射的方式获取该特征点的语义动态概率 M_i^{obj},如式(3)所示。

$$M_i^{\text{obj}}=\frac{1}{1+\exp(-dist_i)} \tag{3}$$

2.2　基于重投影残差的几何动态概率估计

为了提高特征点运动概率估计的准确性，基于先验信息计算的语义动态概率需进一步细化。为此，引入一种新颖的模块来感知特征点的实际运动情况，从而使得运动概率估计地更加准确和高效。在系统的跟踪阶段，采用语义引导的重投影残差计算方法，并将其作为性能指标来量化特征点的运动状态。

对于第 k 帧的每个掩膜中的第 i 个特征点，利用式(4)计算重投影残差 e_i，其中 $\boldsymbol{p}_{k,i}$ 为第 k 帧的第 i 个特征点，$\boldsymbol{P}_{k-1,i}$ 为 $\boldsymbol{p}_{k-1,i}$ 对应的地图点，$\pi(\cdot)$ 为相机的投影函数，$\boldsymbol{T}_{c_k c_{k-1}}$ 为相邻帧之间的变换矩阵。统计重投影残差的最大值 $e_{\max}$、平均值 $\bar{e}$ 以及标准差 σ。利用式(5)估计基于重投影残差的几何动态概率 M_i^{geo}，m 为当前掩膜内满足 3σ 准则的特征点数量。

$$e_i = \| \boldsymbol{p}_{k,i} - \pi(\boldsymbol{T}_{c_k c_{k-1}} \boldsymbol{P}_{k-1,i}) \|_2 \tag{4}$$

$$M_i^{\text{geo}} = \begin{cases} \dfrac{1}{1+\exp(-\sum_{j=1}^{m} e_j/m \cdot e_{\max})}, & |e_i - \bar{e}| \leqslant 3\sigma \\ \dfrac{1}{1+\exp(-e_i/e_{\max})}, & \text{其他} \end{cases} \tag{5}$$

对于相邻两帧之间的一对匹配点，重投影残差的大小不仅与对应空间点是否符合静态环境假设有关，还与求解位姿矩阵时的约束数量的多少以及位姿矩阵是否满足约束条件相关。为了估计特征点的几何动态概率的置信度，引入位姿矩阵统计的置信度 C_s 和位姿矩阵计算的置信度 C_c，如式(6)和(7)所示。n 表示当前帧中参与位姿优化的地图点数量，e_{th} 为最大重投影残差的 0.75 倍。最终，基于重投影残差的几何动态概率的置信度 $C^{\text{geo}} = C_s C_c$。

$$C_s = \frac{1}{1+e^{-n}} \tag{6}$$

$$C_c = 1 - \frac{\sum e_i}{n \cdot e_{\text{th}}} \tag{7}$$

2.3　基于语义和几何信息加权融合的特征点筛选

将特征点的语义动态概率和几何动态概率进行加权融合，得到特征点最终的运动概率 M_i，如式(8)所示，C^{obj} 为语义分割获得的对象置信度。

$$M_i = \frac{M_i^{\text{geo}} C^{\text{geo}} + M_i^{\text{obj}} C^{\text{obj}}}{C^{\text{geo}} + C^{\text{obj}}} \tag{8}$$

对于当前帧中所有特征点的运动概率集合 $\boldsymbol{M} = \{M_1, M_2, M_3, \cdots, M_n\}$，将能够呈现数据集中分布趋势的中值 $\tilde{M}$ 作为自适应动态概率阈值。如果 $\tilde{M} < 0.5$，说明静态场景在当前图像中占主导地位，此时将阈值 $\tilde{M}$ 设置为 0.5；否则，动态场景占主导地位，将阈值 $\tilde{M}$ 设置为中值。如果第 i 个特征点的运动概率 $M_i \geqslant \tilde{M}$，将其视为动态点并剔除，否则保留该特征点，参与后续的位姿优化。

3　加权位姿估计

ORB-SLAM2 的位姿优化可分为粗优化和精优化。在跟踪线程的恒速模型跟踪、参考关键帧跟踪和重定位跟踪过程中，执行粗优化，仅优化相机的位姿。在局部建图和回环检测线程的局部 BA 优化和全局 BA 优化中，执行精优化，同时优化相机的位姿和路标点。当前大多数的动态 SLAM 方法，在特征点剔除之后，执行文献[1]、[7]、[10]、[15]的非线性优化方法完成位姿估计，对应的代价函数如式(9)所示。

$$\{\boldsymbol{R}_{cw}^*, \boldsymbol{t}_{cw}^*\} = \underset{\boldsymbol{R}_{cw}, \boldsymbol{t}_{cw}}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} \rho \| \boldsymbol{p}_i - \pi(\boldsymbol{R}_{cw} \boldsymbol{P}_i^w + \boldsymbol{t}_{cw}) \|_{\sum^{-1}}^2 \tag{9}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} (s^2)^m & 0 \\ 0 & (s^2)^m \end{bmatrix} \tag{10}$$

其中，$\boldsymbol{\Sigma}$ 为协方差矩阵，s 和 m 分别为图像金字塔的缩放因子和层数，在实验中，s 和 m 分别设置为 1.2 和 8；$\boldsymbol{p}_i$ 和 $\boldsymbol{P}_i^w$ 为第 i 个特征点的归一化坐标和对应地图点的世界坐标；$\boldsymbol{R}_{cw}$ 和 $\boldsymbol{t}_{cw}$ 分别表示世界坐标系到相机坐标系的旋转矩阵和平移向量；$\rho(\cdot)$ 为鲁棒核函数；n 为图优化的残差边数量。为了提升系统在不同动态场景下位姿估计的准确性和鲁棒性，在经过概率筛选之后，将剩余特征点的运动概率作为区分不同特征点对位姿估计的贡献程度，从而实现所提算法对动态环境的高度适应性。引入概率加权因子 ω_i，如式(11)所示。特征点的运动概率越小，ω_i 的值越大，则对应的特征点对位姿估计的贡献越大。构造的概率加权代价函数如式(12)所示，位姿优化使用 g2o 库，采取列文伯格-马尔夸特方法进行加权位姿估计。图 3 为加权位姿优化的可视化示意图，其中红色圆圈为被剔除的动态点，绿色圆圈为保留的特征点，圆圈的大小表示该点的运动可能性的大小。

$$\omega_i = (1 - M_i) \tag{11}$$

$$\{\boldsymbol{R}_{cw}^*, \boldsymbol{t}_{cw}^*, \boldsymbol{\omega}_i^*\} = \underset{\boldsymbol{R}_{cw}, \boldsymbol{t}_{cw}, \boldsymbol{\omega}_i}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} \rho \| \omega_i (\boldsymbol{p}_i - \pi(\boldsymbol{R}_{cw} \boldsymbol{P}_i^w + \boldsymbol{t}_{cw})) \|_{\sum^{-1}}^2 \tag{12}$$

4　静态场景重建

为了生成高质量的三维点云地图，引入静态场景重建线程。为了解决动态物体对三维点云地图造成的重影问题，仅传入与静态场景相关的信息。考虑到室内环境中的人是最常见的动态对象，本文将删除深度信息中属于人的部分，并滤除深度值<0.01 m 或>10 m 的异常像素点。当传入新的关键帧后，将关键帧的位姿、深度信息、彩色信息以及相机内参传递给静态场景重建线程。利用每个像素点对应的世界坐标构建局部点云地图，并将其合并到全局点云地图。为了有效降低点云的密度，采用体素滤波的方法，设置滤波器的分辨率为 0.005 m。该方法在保留点

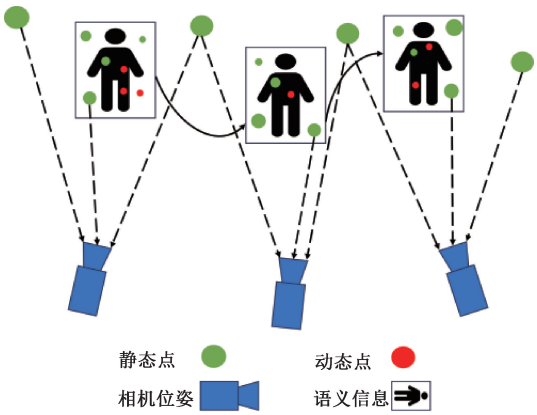


图 3 加权位姿优化的模型

Fig. 3 Model for weighted pose optimization

云细节的同时,有效去除噪声与冗余数据。如果检测到回环关键帧,则通过全局优化调整关键帧位姿,并更新全局点云地图。

5 实验与分析

为验证所提算法的有效性,在公开数据集 TUM RGB-D 和 Bonn 进行一系列实验,包括特征点筛选实验、导航定位精度实验和静态场景重建实验。为验证所提算法的实用性,在真实场景中进行相关实验。与之对比的先进基准算法包括 ORB-SLAM2[1]、Dyna-SLAM[10]、LC-CRF-SLAM[7]和 OVD-SLAM[14]。在评估算法的导航定位精度时,采用绝对轨迹误差(absolute trajectory error,ATE)的均方根误差(root mean squared error,RMSE)和方差(standard deviation,SD)作为评价指标。所有算法均在配备 Intel Core i5-8400 @ 2.80 GHz CPU 和 16 GB RAM 的电脑上运行。为了减小随机误差,每个算法运行 5 次,然后取平均值。表中加粗的数据表示所有算法在该序列的 ATE 的 RMSE 和 SD 的最小值。

TUM RGB-D 数据集由 Microsoft Kinect 传感器以 30 Hz 频率在室内不同环境中捕获,包含 39 个低动态和高动态序列。在低动态序列中,场景通常涉及两个人坐在椅子上进行语言和手势交流;而在高动态序列中,场景则涉及大范围的活动。选择 4 个高动态序列(fr3_walking_xyz、fr3_walking_rpy、fr3_walking_half、fr3_walking_static)和一个低动态序列(fr3_sitting_static)进行实验。

Bonn 数据集由伯恩大学摄影测量与机器人实验室发布,包含 24 个序列,记录了不同的室内活动,如人群行走、搬运箱子、人物追踪等。选取 6 个具有代表性的序列进行实验,包括 crowd1、crowd2、move_no_box1、move_no_box2、person_tracking1、person_tracking2。

5.1 特征点筛选实验

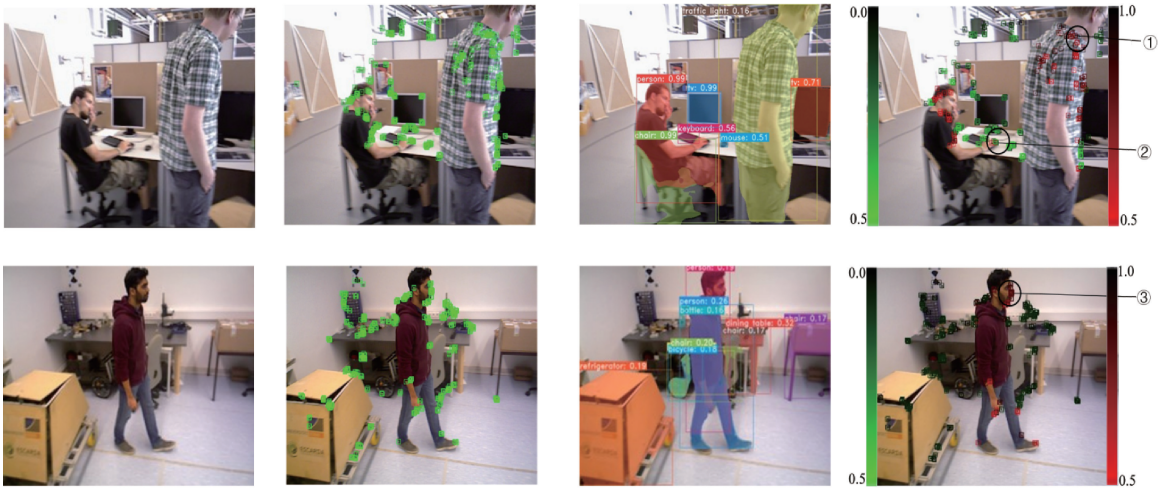
为验证基于运动概率的特征点筛选算法的效果,选取典型序列 fr3_walking_rpy 和 person_tracking1 的某时刻进行特征点筛选实验。图 4(a)~(d)的上下两行分别展示序列 fr3_walking_rpy 和 person_tracking1 某时刻的特征点筛选过程,包括原始图像、ORB 特征点分布图、Yolact 分割后的图像以及特征点经过运动概率筛选后的结果图。图 4(d)左右两侧的两个矩形的颜色的深浅表示特征点运动概率的大小。在序列 fr3_walking_rpy 的某时刻,右侧的人正在运动,而左侧的人在移动鼠标。在图 4(d)中的②处,尽管鼠标的语义动态概率较低,但由于人手部的移动,其几何动态概率较高;在图 4(d)中的①和③处,人体作为非刚体,尽管语义动态概率较高,但这些区域未参与整体运动,因此几何运动概率较低。从特征点筛选结果可以看到,结合语义动态概率和几何动态概率的方法能够筛选出高动态特征点,并保留低动态特征点。

5.2 TUM RGB-D 数据集上的导航定位精度实验

为验证所提算法在动态环境中的优势,选取 TUM RGB-D 数据集中的 4 个高动态序列和一个低动态序列进行实验。在图表中,使用 Ours 表示所提算法。图 5 为所提算法和 ORB-SLAM2 的导航定位曲线对比,在每个子图中,实线表示估计的相机轨迹、虚线表示地面真实轨迹、粗实线表示二者之间的误差。

图 5(a)~(d)分别对应序列 fr3_walking_xyz、fr3_walking_rpy、fr3_walking_half、fr3_ sitting _static 的轨迹,每个子图的第 1 行为 ORB-SLAM2 估计的轨迹,第 2 行为所提算法估计的轨迹。实验结果表明,在低动态序列 fr3_sitting_static 中,所提算法和 ORB-SLAM2 均能实现较高精度的定位。然而,在其他高动态序列中,ORB-SLAM2 由于缺乏处理动态对象对位姿估计干扰的能力,定位精度明显下降,而所提算法能够保持较高的定位准确性。为定量评估所提算法的性能,表 1 给出了所提算法和 ORB-SLAM2 在 ATE 的 RMSE、SD 的对比结果。其中,Ours(w/o 概率加权)和 Ours(w/o 动态掩膜边界修复)分别表示未使用概率加权和未使用动态掩膜边界修复模块。从表 1 中可以看出,与 ORB-SLAM2 相比,所提算法在 ATE 的 RMSE 和 SD 上分别平均降低 69.16%、67.72%。

为验证各模块的有效性,进一步分析了仅使用动态掩膜边界修复模块、仅使用概率加权模块以及两者组合对定位精度的影响。从表 1 中可以看到,与 ORB-SLAM2 相比,Ours(w/o 概率加权)和 Ours(w/o 动态掩膜边界修复)均能显著提升定位精度;与 Ours(w/o 概率加权)和 Ours(w/o 动态掩膜边界修复)相比,所提算法在高动态序列的定位精度均有进一步提升。这充分证明了所提算法的准确性和鲁棒性,以及概率加权和动态掩膜边界修复模块的有效性。概率加权模块通过对保留的特征点赋予不同的权重,实现自适应位姿估计;动态掩膜边界修复模块通过区分前景和背景,解决语义分割方法未能正确分割边界点的问题。这两个模块相辅相成,在系统中协同作用,实现



(a) 原始图像
(a) Original images

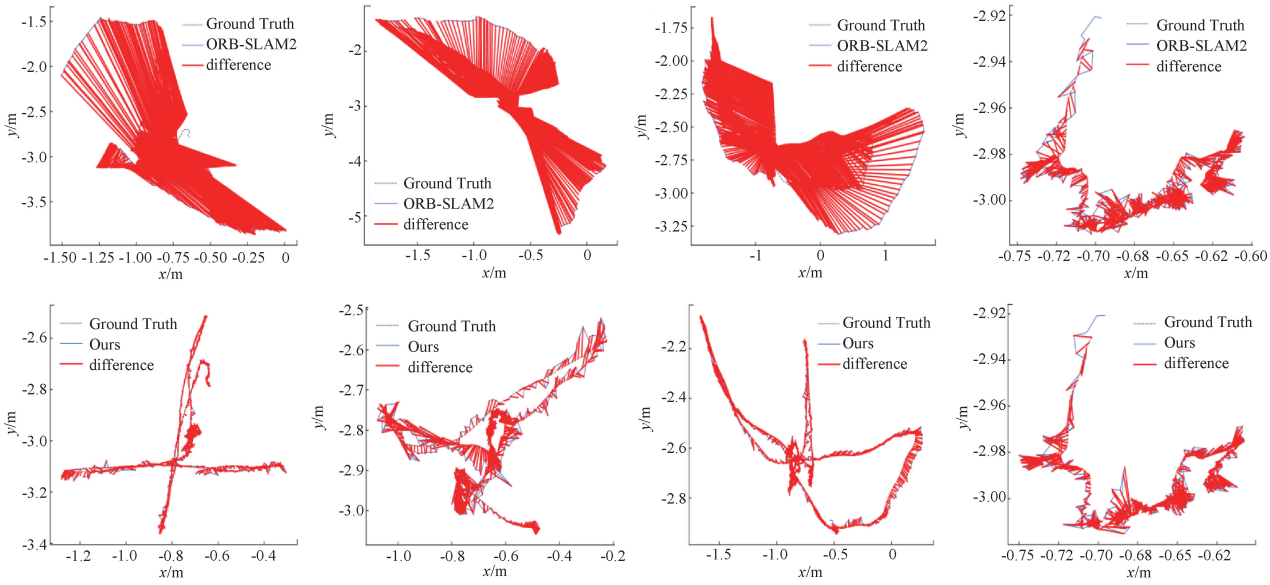
(b) 特征点分布
(b) Distributions of feature points

(c) 分割结果
(c) Results of segmentation

(d) 筛选结果
(d) Results after screening

图 4　特征点筛选过程

Fig. 4　Process of feature points screening



(a) fr3_walking_xyz

(b) fr3_walking_rpy

(c) fr3_walking_half

(d) fr3_sitting_static

图 5　ORB-SLAM2 与所提算法的绝对轨迹误差曲线对比

Fig. 5　Comparison of Absolute Trajectory Error curves between ORB-SLAM2 and the proposed algorithm

表 1　TUM RGB-D 数据集上的 ATE 对比

Table 1　Comparison of ATE in the TUM RGB-D dataset

m

序列	ORB-SLAM2[1]		Ours(w/o 概率加权)		Ours(w/o 动态掩膜边界修复)		Ours		性能提升/%	
	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
fr3_walking_xyz	0.291 9	0.138 7	0.015 5	0.009 2	0.014 8	0.007 5	0.012 7	0.006 5	95.65	95.31
fr3_walking_rpy	0.157 9	0.079 4	0.031 3	0.019 5	0.028 7	0.021 6	0.027 9	0.015 7	82.33	80.22
fr3_walking_half	0.348 9	0.146 1	0.023 0	0.012 5	0.024 5	0.014 3	0.020 9	0.010 9	94.01	92.54
fr3_walking_static	0.015 6	0.009 8	0.006 8	0.002 9	0.008 3	0.003 6	0.006 5	0.003 5	58.33	64.29
fr3_sitting_static	0.007 1	0.003 2	0.006 1	0.002 9	0.005 8	0.002 7	0.006 0	0.003 0	15.49	6.25

了更好的定位效果。

为进一步验证所提算法的性能优势,将其与当前先进的动态 SLAM 方法:Dyna-SLAM、LC-CRF-SLAM、OVD-SLAM 进行 ATE 的 RMSE 对比,结果如表 2 所示。在极具挑战性的序列 fr3_walking_rpy 中,场景不仅包含行走的人物,还存在绕三轴旋转的相机。这种运动模式会导致图像模糊,从而大幅减少特征点数量。Dyna-SLAM 直接剔除运动目标上的特征点,因此在人物占据较大区域时,特征点数量不足导致跟踪失败,其定位效果较差。在序列 fr3_walking_static 中,相机固定且人物在场景中行走。

由于 LC-CRF-SLAM 的标准参数固定,一些静态路标被错误地标记为动态,从而降低了初始位姿估计的准确性,其定位精度在该序列中最低。所提算法仍能保持与另外两种算法的定位精度相当。在前 4 个序列中,与 Dyna-SLAM、LC-CRF-SLAM、OVD-SLAM 和文献[16]相比,所提算法在 ATE 的 RMSE 上分别平均降低 18.83%、29.81%、8.96%、10.20%。

表 2 不同算法在 TUM RGB-D 数据集上的 ATE 的 RMSE 对比

Table 2 Comparison of ATE RMSE for different algorithms in the TUM RGB-D dataset m

序列	Dyna-SLAM[10]	LC-CRF SLAM*[7]	OVD-SLAM*[14]	文献[15]*	文献[16]*	Ours
fr3_walking_xyz	0.016	0.016	0.014	0.152	0.014	0.013
fr3_walking_rpy	0.037	0.046	0.035	—	0.034	0.028
fr3_walking_half	0.031	0.028	0.023	—	0.025	0.021
fr3_walking_static	0.007	0.011	0.007	—	0.007	0.007
fr3_sitting_static	0.008	—	—	—	—	0.006

注:由于未跑通 LC-CRF-SLAM 和 OVD-SLAM,表中带“*”标记的算法使用原始论文的数据,“—”表示参考文献未提供相关结果。

5.3 Bonn 数据集上的导航定位精度实验

为进一步验证所提算法的优势,在 Bonn 数据集的部分高动态序列进行相关实验,结果如表 3 所示。

从表 3 可以看到,所提算法在 2/3 的序列中表现出持平/优于其他算法的导航定位精度,在高动态人群序列 crowd1 中实现了 0.017 m 的最高定位精度。在 crowd2 序列中,即使存在多个运动目标重叠以及遮挡导致的边界分割不准确问题,所提算法通过动态掩膜边界修复模块对边界点修复,达到了与 OVD-SLAM 相同的定位精度。相比 ORB-SLAM2、Dyna-SLAM、LC-CRF-SLAM、OVD-SLAM 和文献[16],所提算法的 ATE 的 RMSE 分别平均降低 91.94%、16.71%、4.49%、0.12%、2.52%。这进一步验证了所提算法在高动态环境中的准确性和鲁棒性。

表 3 不同算法在 Bonn 数据集上的 ATE 的 RMSE 对比

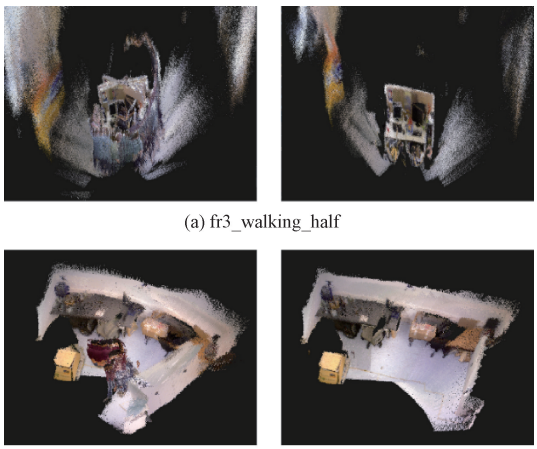
Table 3 Comparison of ATE RMSE for different algorithms in the Bonn dataset m

序列	ORB-SLAM2[1]	Dyna-SLAM[10]	LC-CRF SLAM*[7]	文献[15]*	文献[16]*	OVD-SLAM*[14]	Ours
crowd1	1.085	0.023	0.019	—	0.020	0.018	0.017
crowd2	2.109	0.029	0.031	—	0.024	0.023	0.023
move_no_box1	0.207	0.027	0.018	—	0.020	0.018	0.021
move_no_box2	0.102	0.034	0.038	—	0.029	0.033	0.026
person_tracking1	0.800	0.045	0.035	—	0.041	0.041	0.047
person_tracking2	0.856	0.041	0.040	—	0.038	0.038	0.036

注:由于未跑通 LC-CRF-SLAM 和 OVD-SLAM,表中带“*”标记的算法使用原始论文的数据,“—”表示参考文献未提供相关结果。

5.4 建图质量测试

为进一步验证所提算法的优势并展示相机位姿估计的准确性,在 TUM 数据集的 fr3_walking_half 和 Bonn 数据集的 person_tracking1 序列,进行所提算法与 ORB-SLAM2 的建图质量对比实验,结果如图 6(a)、(b)所示。由于动态物体的影响,ORB-SLAM2 在定位精度和鲁棒性方面表现较差,并且在建图过程中出现了明显的重影现象,如图 6(a)、(b)的第 1 列所示。相比之下,所提算法通过从当前图像中剔除运动中的人,有效消除了重影问题,生成较为干净的静态地图,如图 6(a)、(b)的第 2 列所示。



(a) fr3_walking_half

(b) person_tracking1

图 6　ORB-SLAM2 与所提算法的静态场景重建结果对比

Fig. 6　Comparison of static scene reconstruction results between ORB-SLAM2 and the proposed algorith

5.5　真实场景实验

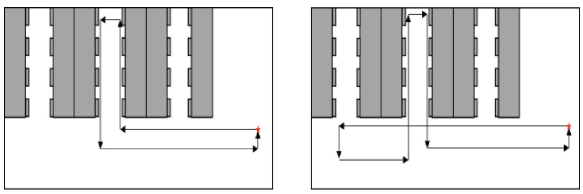
为评估所提算法的实用性，在室内实验室进行实证测试。实验中，搭载 Intel Realsense D435i 相机的四旋翼飞行器平台和实验场景如图 7(a)、(b)所示。数据集由四旋翼飞行器在室内动态场景中捕获，总共采集两个序列，每个序列的行驶路线如图 8(a)、(b)所示。红色五角星表示每个序列的起点，飞行器按照箭头所示的方向行驶。在实验中，实验者不仅自身移动，而且通过随机移动箱子、书本等物体引入额外的复杂性。图像的分辨率为 1 280×720，相机频率为 30 Hz。

(a) 四旋翼飞行器
(a) Quadrotor

(b) 室内环境
(b) Indoor environment

图 7　实验平台以及实验环境

Fig. 7　Experimental platform and environment

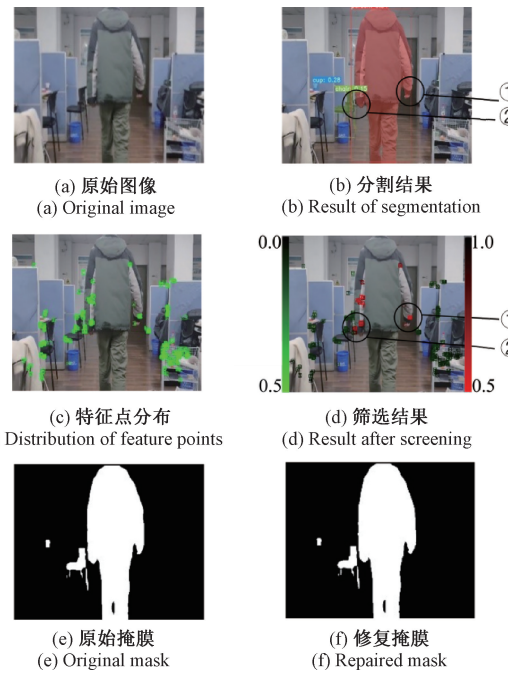


(a) 01　　(b) 02

图 8　四旋翼飞行器运动轨迹布局

Fig. 8　Layouts of the motion trajectory of the Quadrotor

选取序列 01 中某时刻人行走的图像，进行动态掩膜边界修复和特征点运动概率筛选实验，结果如图 9 所示。图 9(a)～(d)分别展示原始图像、Yolact 分割后的图像、ORB 特征点分布图以及经过运动概率筛选后的图像。图 9 (e)、(f)分别展示 Yolact 分割后的二值掩膜图像以及通过动态掩膜边界修复模块修复的图像。Yolact 分割方法存在边界精度较低的问题，而所提算法结合动态掩膜与深度信息，有效修复了掩膜边界。如图 9(f)所示，所提算法更精确地还原人体掩膜，为运动概率筛选提供坚实的基础条件。从图 9(d)中可以看到，即使在边界①和②处，所提算法仍能为不同的特征点分配不同的权重，确保特征点筛选的准确性。



(a) 原始图像
(a) Original image

(b) 分割结果
(b) Result of segmentation

(c) 特征点分布
(c) Distribution of feature points

(d) 筛选结果
(d) Result after screening

(e) 原始掩膜
(e) Original mask

(f) 修复掩膜
(f) Repaired mask

图 9　所提算法在真实场景上的动态掩膜边界修复和特征点筛选结果

Fig. 9　The dynamic mask boundary restoration and feature points selection results of the proposed algorithm in real-world scenarios

由于缺乏地面真值作为参考，因此保持四旋翼飞行器的起点和终点位置一致，计算轨迹端点漂移来评估定位精度。在测试中，将所提算法与 ORB-SLAM2、Dyna-SLAM 对比。如图 10 所示，矩形虚线框内为轨迹端点位置的放大图，轨迹端点漂移误差如表 4 所示。

从实验结果中可以看出，所提算法在两个序列均表现出最低的轨迹端点漂移误差。由于动态物体占据了较大区域，若直接删除动态物体上的特征点，将导致特征点数量不足，从而影响 Dyna-SLAM 的性能，表现不及所提算法。由于缺乏对动态物体的有效抗干扰能力，ORB-SLAM2 的误差显著增加。相比 ORB-SLAM2，所提算法在序列 01 和 02 的轨迹端点漂移误差分别平均降低 54.38%、50.02%，结果表明所提算法在动态环境下具有

更高的准确性和鲁棒性。

表 4 不同算法的轨迹端点漂移误差对比

Table 4 Comparison of trajectory end drift errors of different algorithms m

序列	ORB-SLAM2[1]	Dyna-SLAM[10]	Ours	性能提升
01	0.238 5	0.150 8	0.108 8	54.38%
02	0.307 0	0.169 3	0.151 6	50.02%

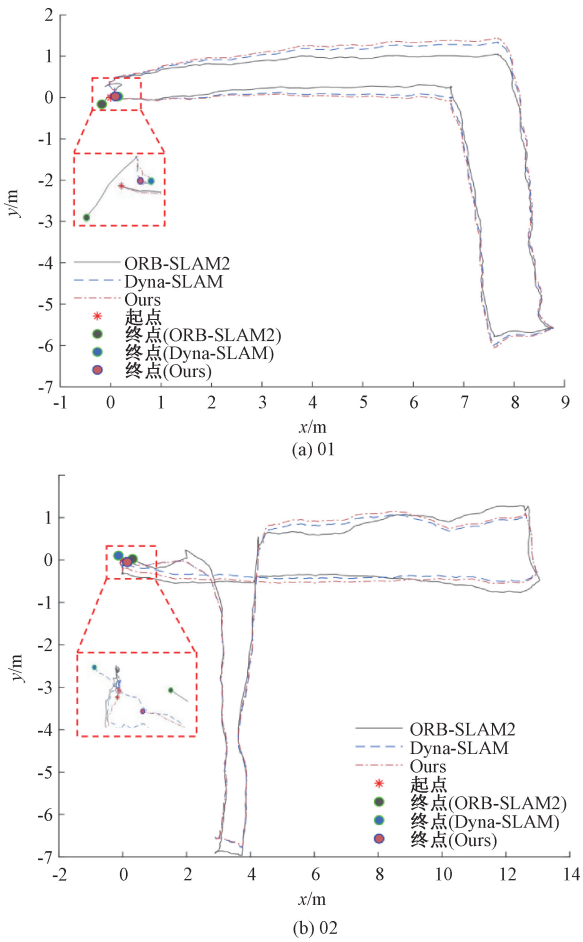


图 10 不同算法估计的轨迹对比

Fig. 10 Comparison of trajectories estimated by different algorithms

为了评估算法的实时性,采用真实场景中每帧图像的平均跟踪时间作为指标。表 5 给出所提算法与 ORB-SLAM2、Dyna-SLAM 在相同平台下的运行结果。尽管 ORB-SLAM2 的运行速度最快,但是其定位精度严重下降。Dyna-SLAM 采用的语义分割网络 Mask-RCNN 本身速度较慢,而且该算法涉及耗时较长的采用历史多帧图像对原始图像进行修复的方法,其平均跟踪时间为 450.56 ms。与 Dyna-SLAM 相比,所提算法的速度更快,能够以接近 10 Hz 的频率运行,基本上满足实时性的要求。

表 5 不同算法的平均跟踪时间对比

Table 5 Comparison of average tracking time of different algorithms ms

算法	平均跟踪时间
ORB-SLAM2[1]	32.66
Dyna-SLAM[10]	450.56
Ours	102.34

6 结 论

为减小动态物体对视觉 SLAM 的干扰,本文提出一种新颖的鲁棒动态 RGB-D SLAM 系统。在 ORB-SLAM2 的基础上,引入语义分割线程和静态场景重建线程。首先,结合语义信息和深度信息对动态掩膜进行修复;其次,在动态掩膜边界修复的基础上,构建结合基于语义分割的语义动态概率和基于语义信息引导的特征点几何动态概率的特征点运动概率模型,实现特征点的自适应概率阈值筛选;最后,将剩余特征点的运动概率作为系统的追踪、局部建图、回环线程的 BA 优化环节的权重以区分不同特征点对位姿优化的贡献,从而保证所提算法对动态环境的高度适应性。相较于 ORB-SLAM2、Dyna-SLAM、LC-CRF-SLAM、OVD-SLAM 和文献[16],所提算法在 TUM RGB-D 数据集上的 ATE 的 RMSE 分别平均降低 69.16%、18.83%、29.81%、8.96%、10.20%。相较于 ORB-SLAM2、Dyna-SLAM、LC-CRF-SLAM、OVD-SLAM 和文献[16],所提算法在 Bonn 数据集上的 ATE 的 RMSE 分别平均降低 91.94%、16.71%、4.49%、0.12%、2.52%。在真实场景实验中,相比于 ORB-SLAM2、Dyna-SLAM,轨迹端点漂移误差分别平均降低 52.20%、19.15%。与 Dyna-SLAM 相比,所提算法的实时性有所改善。

未来,将会探索动态对象跟踪和预测,并将动态对象加入到位姿估计的联合优化函数中。同时,考虑加入惯性测量单元,利用其不受动态物体影响的特点实现特征点的更准确的运动概率估计。

参考文献

[1] MUR-ARTAL R, TARDÓS J D. ORB-SLAM2: An open-source slam system for monocular, stereo, and RGB-D cameras[J]. IEEE Transactions on Robotics, 2017, 33(5):1255-1262.

[2] QIN T, LI P, SHEN SH J. VINS-Mono: A robust and versatile monocular visual-inertial state estimator[J]. IEEE Transactions on Robotics, 2018, 34(4): 1004-1020.

[3] 王爽,刘云平,张柄棋,等. 基于全景分割与多视图几何的动态 SLAM 方法[J]. 电子测量技术,2024,47(24):149-159.

WANG SH, LIU Y P, ZHANG B Q, et al. Dynamic SLAM approach based on panoptic segmentation and multi-view geometry [J]. Electronic Measurement Technology, 2024,47(24):149-159.

[4] ZOU D P, TAN P. CoSLAM: Collaborative visual SLAM in dynamic environments [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(2): 354-366.

[5] KIM D, KIM J. Effective background model-based RGB-D dense visual odometry in a dynamic environment[J]. IEEE Transactions on Robotics, 2016, 32(6): 1565-1573.

[6] DAI W CH, ZHANG Y, LI P, et al. RGB-D SLAM in dynamic environments using point correlations[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022,44(1):373-389.

[7] DU ZH J, HUANG SH SH, MU T J, et al. Accurate dynamic SLAM using CRF-based long-term consistency[J]. IEEE Transactions on Visualization and Computer Graphics, 2022,28(4): 1745-1757.

[8] SHENG CH, PAN SH G, GAO W, et al. Dynamic-DSO: Direct sparse odometry using objects semantic information for dynamic environments [J]. Applied Sciences-Basel, 2020,10, DOI:10.3390/app10041467.

[9] HE K M, GKIOXARI G, PIOTR D, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 2(42): 386-397.

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

[11] YU CH, LIU Z X, LIU X J, et al. DS-SLAM: A semantic visual SLAM towards dynamic environments[C]. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS), 2018: 1168-1174.

[12] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[13] FAN Y CH, ZHANG Q CH, TANG Y L, et al. Blitz-SLAM: A semantic SLAM in dynamic environments[J]. Pattern Recognition, 2022, 121, DOI:10.1016/j.patcog.2021.108225.

[14] HE J, LI M, WANG Y, et al. OVD-SLAM: An online visual slam for dynamic environments[J]. IEEE Sensors Journal, 2023,23(12): 13210-13219.

[15] 崔岸,张新颖,马耀辉.复杂环境下基于自适应极线约束的 AGV 视觉 SLAM 算法[J].中国惯性技术学报, 2024,32(3):234-241.
CUI AN, ZHANG X Y, MA Y H. Adaptive polar constraint-based visual SLAM algorithm for AGV in complex environment[J]. Journal of Chinese Inertial Technology, 2024,32(3): 234-241.

[16] 闫河,王旭,雷秋霞.动态场景结合稀疏场景流和加权特征的视觉 SLAM 方法[J].中国惯性技术学报, 2024,32(9):891-897.
YAN H, WANG X, LEI Q X. Visual SLAM method combining sparse scene flow and weighted features in dynamic environment[J]. Journal of Chinese Inertial Technology, 2024, 32(9): 891-897.

[17] BOLYA D, ZHOU C, XIAO F, et al.. YOLACT Real-time instance segmentation [C]. IEEE/CVF International Conference On Computer Vision(ICCV). 2019:9156-9165.

作者简介

于兴云,硕士研究生,主要研究方向为视觉 SLAM、组合导航。

E-mail:220223262@seu.edu.cn

程向红(通信作者),教授,博士生导师,主要从事导航、制导与控制方面的研究。

E-mail:101005578@seu.edu.cn

刘丰宇,博士研究生,主要研究方向为激光/视觉/惯性多传感器融合定位。

E-mail:liufengyu@seu.edu.cn

钟志伟,硕士研究生,主要研究方向为视觉 SLAM、组合导航。

E-mail:220223246@seu.edu.cn