第55卷第5期 南京航空航天大学学报 Vol. 55 No. 5
2023年10月 Journal of Nanjing University of Aeronautics & Astronautics Oct. 2023

DOI:10.16356/j.1005-2615.2023.05.005

顾及动态物体感知的增强型视觉SLAM系统

李 佳[1],李明磊[1],魏大洲[2],吴伯春[2],郭文骏[2]

(1.南京航空航天大学电子信息工程学院,南京 211106;2.中国航空无线电电子研究所,上海 200233)

摘要: 传统的同步定位与制图(Simultaneous localization and mapping,SLAM)系统在复杂环境下工作时,无法分辨环境中的物体是否存在运动状态,图像中运动的物体可能导致特征关联错误,引起定位的不准确和地图构建的偏差。为了提高SLAM系统在动态环境下的鲁棒性和可靠性,本文提出了一种顾及动态物体感知的增强型视觉SLAM系统。首先,使用深度学习网络对每一帧图像的动态物体进行初始检测,然后使用多视图几何方法更加精细地判断目标检测无法确定的动态物体区域。通过剔除属于动态物体上的特征跟踪点,提高系统的鲁棒性。本文方法在公共数据集TUM和KITTI上进行了测试,结果表明在动态场景中定位结果的准确度有了明显提升,尤其在高动态序列中相对于原始算法的精度提升在92%以上。与其他顾及动态场景的SLAM系统相比,本文方法在保持精度优势的同时,提高了运行结果的稳定性和时间效率。

关键词: 同步定位与制图;动态环境;目标检测;多视图几何

中图分类号: TP242 **文献标志码:** A **文章编号:** 1005-2615(2023)05-0789-09

Enhanced Visual SLAM System Considering Dynamic Objects

LI Jia[1], *LI Minglei*[1], *WEI Dazhou*[2], *WU Bochun*[2], *GUO Wenjun*[2]

(1. College of Electronic and Information Engineering, Nanjing University of Aeronautics & Astronautics, Nanjing 211106, China; 2. China Institute of Aeronautical Radio Electronics, Shanghai 200233, China)

Abstract: When working in complex scenarios, traditional simultaneous localization and mapping (SLAM) systems cannot distinguish whether the visible objects are moving. Moving objects in the images may lead to wrong feature association, resulting in the inaccuracy of positioning and the deviation of mapping. To improve the robustness and reliability of the SLAM system in dynamic scenarios, an enhanced visual SLAM system with dynamic object perception is proposed in this paper. Firstly, the object detector is used to initially detect the dynamic objects in each image, and then the multi-view geometry method is further used to extract the dynamic regions that cannot be determined by the object detection. The robustness of the system is improved by eliminating feature points belonging to dynamic objects. The proposed method is tested in public datasets TUM and KITTI. The results show that the localization accuracy of the proposed method in dynamic scenes has been significantly improved, especially in high dynamic sequences. Compared with the original algorithm, the accuracy is improved by more than 92%. Compared with other SLAM systems in dynamic scenarios, the proposed method not only maintains the accuracy advantage, but also improves the stability of running results and time efficiency.

Key words: simultaneous localization and mapping; dynamic environment; object detection; multi-view geometry

基金项目: 国家自然科学基金(42271343);核工业北京地质研究院国家级重点实验室基金(6142A010403)。

收稿日期: 2022-10-30;**修订日期:** 2023-01-03

通信作者: 李明磊,男,副教授,E-mail:minglei_li@nuaa.edu.cn。

引用格式: 李佳,李明磊,魏大洲,等.顾及动态物体感知的增强型视觉SLAM系统[J].南京航空航天大学学报,2023,55(5):789-797. LI Jia, LI Minglei, WEI Dazhou, et al. An enhanced visual SLAM system considering dynamic objects [J]. Journal of Nanjing University of Aeronautics & Astronautics, 2023, 55(5):789-797.

同步定位与制图(Simultaneous localization and mapping, SLAM)技术是很多机器人应用的前提条件,它为路径规划、无碰撞导航和环境感知等任务提供支持。与激光雷达传感器相比,视觉传感器能够获取图像的纹理信息,可以拓展其他基于视觉的研究。视觉SLAM的框架通常包括图像信息读取、视觉里程计、后端优化、回环检测与建图。其中,视觉里程计作为SLAM系统的前端,能够通过传感器读取的图像信息来估计相机运动,其实现方法根据处理技术的不同可以分为直接法和特征点法两类[1-2]。直接法基于灰度不变假设,使用每帧图像的全部像素信息,通过最小化光度误差来优化相机的位置姿态,对于图像的灰度值变化比较敏感。特征点法需要对图像进行特征的提取与匹配,通过匹配的特征构建并最小化重投影误差来优化相机的位置姿态,其相关研究的积累比较丰富,系统性能相对稳定。

对于大部分传统视觉SLAM系统[3-5],无论是直接法还是特征点法,均是以基于场景为静态且主要变化由相机运动造成的假设为前提。然而在实际环境中,动态物体的存在不可避免,例如运动的行人和车辆。从动态物体中提取的特征跟踪点会增加系统不确定性,降低对相机位置姿态估计的精度,甚至导致定位的失败。一些系统[6-8]会在特征匹配时采用随机抽样一致(Random sample consensus, RANSAC)算法[9]去除错误的匹配点对,提高在动态环境下的鲁棒性。但是这种方法具有随机性,无法从根本上针对性地剔除位于动态物体中的特征点。因此,对动态物体的感知和处理成为提高视觉SLAM系统的定位与制图精度的重要突破方向。由于相机自身的运动会给动态物体的检测带来很大的挑战,因此使用单一的方法不能完整地分割出动态区域,而由于检测方法的不同,能够检测出动态物体的类型会略有不同。

本文提出一种基于深度学习和几何约束的算法来处理视觉SLAM过程中的动态物体,能够适用于RGB-D、双目和单目等多种类型的影像数据。其中,深度学习方法将动态物体根据语义知识定义为车辆和人这类具有自主移动能力的潜在运动对象,而几何约束方法将不满足几何约束的点集标记为动态,检测的是场景中真实运动的动态物体。基于ORB-SLAM2[8]系统框架,添加一个前端的处理模块来实现动态物体的感知检测。在特征跟踪步骤中,基于区域掩膜剔除属于动态物体部分的特征跟踪点,从而提高特征关联的可靠性,使系统拥有更准确的输出。

1 相关研究

1.1 SLAM

近十几年来,视觉SLAM取得了快速的发展,它因为成本低、体积小等优点受到很多研究人员的关注。Davison等[5]提出了Mono-SLAM来实现通过单目相机进行实时定位与建图的目标,是较早期的一种视觉SLAM系统。随后,Klein等[6]提出的PTAM(Parallel tracking and mapping)创造性地将整个系统划分为两个线程:跟踪和建图,成为一个参考基准。Engel等[10]提出的LSD-SLAM将直接法应用到了半稠密的单目SLAM中,可以构建大规模、一致的环境地图。同时,Forster等[11-12]提出了一种将直接法和特征点法结合的视觉里程计(Semi-direct monocular visual odometry, SVO)。除此之外,直接稀疏里程计(Direct sparse odometry, DSO)[13]、单目视觉惯性状态估计器(Visual inertial navigation system-monocular, VINS-mono)[14]等框架也都使用了直接法。虽然直接法在跟踪和匹配方面能够节省计算资源,但其稳定性仍有待提高。基于特征点提取与匹配的方法能够保证在SLAM跟踪中位姿估计的准确性,Leutenegger等[15]提出的基于双目相机和惯性导航系统的OKVIS和Mur-Artal等[7-8]提出的ORB-SLAM和ORB-SLAM2都是基于特征跟踪的经典SLAM系统。

ORB-SLAM2的框架采纳了多线程机制,使用了ORB(Oriented FAST and rotated BRIEF)[16]特征点和3个主要的并行线程,使系统能够在大场景、大回环下长时间运行,从而保证了相机轨迹与地图的全局一致性。ORB-SLAM2具有良好的定位与建图性能,但在处理动态环境问题方面仍然有许多不足,其稳定性会随着动态物体在影像中的增加而显著下降,甚至直接引起定位失败。

1.2 动态环境下的SLAM

目前研究人员对于提升视觉SLAM在动态环境下的性能所采取的解决思路是基本一致的,即在前端视觉里程计之前,使用某种方法检测图像中的动态物体并进行筛除,然后仅使用环境中的静态特征关联点来参与计算[17]。得益于深度学习技术的发展,如今一些检测器已经能够很好地识别图像中一些特定的动态物体(如汽车、行人和动物等)。Zhong等[18]开发的Detect-SLAM在ORB-SLAM2的基础上结合目标检测网络SSD-NET对关键帧中的动态物体进行检测,将特征点属于动态物体的概率称为运动概率。通过特征匹配点和周围区域的点来更新普通帧中的特征点的运动概率,从而标

记出所有帧中特征点的运动情况。不过，通过深度学习的方法剔除动态目标往往会受其训练数据集的约束，因此也有一些方法是联合深度学习和几何约束一起进行检测[19]。

Yu 等[20]开发的 DS-SLAM 将语义分割网络 SegNet 设置为一个独立的线程，对于前后两帧图像通过极线几何方法联合语义分割结果检测动态特征点，然而文中仅将人设置为要分割的动态物体，并且网络的分割效果还有很大的提升空间。Bescos 等[21]开发的 DynaSLAM 使用 Mask R-CNN 网络进行语义分割识别出先验的动态物体，结合多视图几何方法增强动态范围感知能力。Li 等[22]开发的 DP-SLAM 将语义分割网络和极线几何方法的检测结果转换为观测概率，基于贝叶斯定理对特征点的移动概率进行更新，然后剔除移动概率较高的特征点。

除了使用深度学习和几何约束的方法，还可以采用光流法对动态区域进行检测。艾青林等[23]提出了一种在室内环境中检测动态物体的 RGB-D SLAM 算法，通过单应变换来补偿由相机运动带来的背景变化，使用双向的光流法对运动的前景物体进行判断，最后根据几何连通性与深度图像聚类结果对动态物体进行分割。

现有的 DynaSLAM[21]和 DP-SLAM[22]等系统主要使用两阶段检测网络模型提取像素级的目标分割来筛除动态物体，对动态物体的像素分割精度很高，避免了一阶段检测网络提取的包围框会损失大量有效静态场景像素区域的情况。然而，两阶段网络模型的计算复杂度更高，并且对于边界区域一般需要使用膨胀算法进行处理，使得每一次分割的边缘都具有了不确定性。本文的思想是利用一阶段的目标检测网络提高计算的时效性，同时通过多视图几何的方法，减少矩形包围框带来的静态场景特征区域的损失。

2　算法设计

2.1　系统框架设计

本文的算法在 ORB-SLAM2 的框架基础上结合深度学习和几何约束的方法来提高系统在动态环境下的鲁棒性，同时考虑了单目、双目以及 RGB-D 相机 3 种情况，系统的总体框架如图 1 所示。

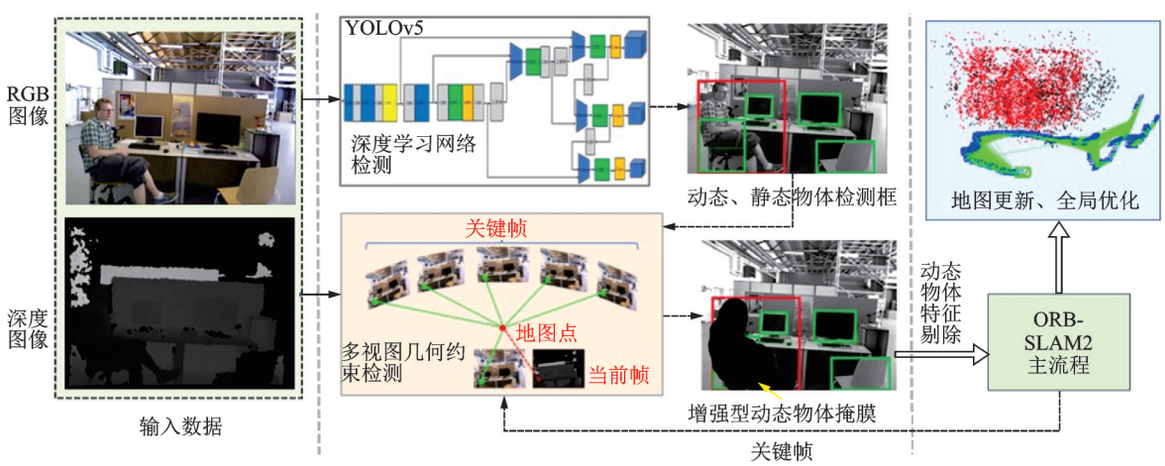


图 1　本文方法的框图

Fig.1　Diagram of the proposed method

由图 1 可以看出，对于使用单目和双目相机的情况，系统通过 YOLOv5 网络将人和车辆视作潜在的运动物体进行检测，利用剩余的静态区域进行后续的地图更新与全局优化。对于使用 RGB-D 相机的情况，由于 RGB-D 相机能够直接采集图像中每一个像素点的深度信息，系统利用提取的深度信息增加了一个基于深度变化的多视图几何方法[21]的运动一致性判断模块，该模块针对目标检测遗漏的区域以及先验知识无法确认但可能被移动的静态物体（例如被人拿起的书）进行判断。本文将考虑了 YOLOv5 网络和多视图几何综合判别的方法简称为 YG-SLAM 算法。

2.2　基于 YOLOv5 的动态物体检测

目前，一阶段目标检测网络的检测精度和检测速度已经具有良好的实时应用优势[24]，常见的模型有 YOLO 系列和 SSD。与早期的 YOLO 相比，SSD 采用了和 Faster R-CNN 相似的先验框概念，并删除了 bounding box proposal 以及后续的重采样步骤，检测结果更精确，检测速度也相近。但是 SSD 的先验框需要人工设置参数，导致调试过程依赖经验。YOLOv5 采用了自动锚框计算，可以在不同训练集中自适应地计算出最佳的锚框值，它

在YOLOv4[25]算法上做了进一步的改进，增加了Focus等模块，在保证模型识别精度的同时，进一步提高了运算速度，并且模型的权重相对YOLOv4而言更小。YOLOv5一共有4个版本，分别为YOLOv5x、YOLOv5l、YOLOv5m和YOLOv5s。本文采用网络深度和特征图宽度均最小的网络YOLOv5s作为动态物体检测的基准网络。

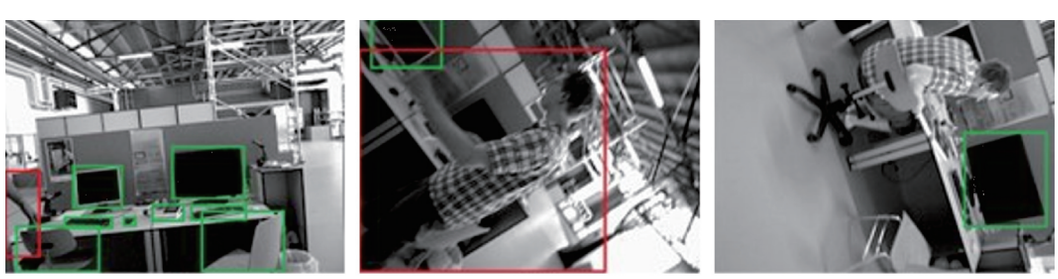
首先，利用YOLOv5的语义知识检测先验的动态物体，将检测网络输出的检测框分为高动态和低动态两类。高动态框内检测的是能够自主移动的物体，即人与车辆，低动态框内检测的是不会自主移动的物体，例如椅子、书本等。由于使用的是矩形包围框的形式来框选检测的目标，因此低动态框与高动态框相互之间会存在相交区域，需要设计一套处理机制分别分析各个部分的像素类型。

根据矩形框的相对位置关系，本文将位于高动态框内、低动态框外的区域，划分为动态区域，认为其中提取的特征点具有高不可靠性，故将其去除；将位于高动态框和低动态框相交的区域，划分为待定区域，其中特征点的动态特性需要等候做进一步判断；将其余没有产生检测框的区域划分为静态区域，该区域内检测到的跟踪匹配的特征点被视为是可靠性较高的特征关联。如图2所示，其中红色区域表示动态区域，黄色区域表示待定区域。可以看出，由于检测框的特性，黄色区域不仅包含低动态对象，也包含高动态对象的局部，三幅图中的黄色区域都包含了人的部分身体。

图2 目标检测网络确定区域

Fig.2 Object detector determines the region

本文使用的YOLOv5目标检测网络的模型是通过在COCO数据集[26]上预先训练得到参数，能够确定有限类别的目标对象。由于训练样本有限，在复杂情况下使用目标检测网络检测动态区域有可能会出现漏检和错检的情况，如图3所示。图3(a)表示人移动椅子的情况，由于使用先验的语义知识将椅子判断为了低动态类，故目标检测对于此时移动的椅子无法进行有效地判断；图3(b)表示检测框在整幅图像占比过大的情况，由于目标检测网络的结果无法做到像素级语义分割网络那样精确，因此在相机旋转角度过大或者要检测的目标动态物体离相机距离过近时，得到的动态区域占整幅图像的比例会过大，此时如果将其中的特征点全部去除会出现特征匹配过少导致的相机跟踪失败的情况；图3(c)表示目标检测网络漏检的情况，当输入图像比较模糊或者旋转角度过大时可能会出现检测失败的情况。对于上述场景，本文采用几何约束的方法进行联合检测。



(a) False detection (b) Oversized detection (c) Missing detection

图3 目标检测无法满足的情况

Fig.3 Situations of object detection not to met

2.3 基于多视图几何方法的动态物体检测

对于划分的待定区域以及目标检测网络的错检和漏检情况，使用多视图几何方法从像素级层面进行进一步的判断。

首先，对于输入的每一帧图像，通过目标检测网络确定动态区域后，使用位于静态区域的特征点进行轻量级的相机跟踪，得到当前帧的一个估计位姿。选择跟踪的地图点超过50个并且和上一关键

帧的地图点重叠度小于 90% 的帧作为关键帧，这样做的目的是为了使插入的关键帧之间保持一定距离，减少信息的冗余。计算当前帧与每个关键帧之间的旋转和距离来衡量重叠度，令 d 作为两帧之间的重叠度，计算公式为

$$d=0.7\frac{t}{t_{\max}}+0.3\frac{r}{r_{\max}} \tag{1}$$

$$r=\sqrt[2]{\boldsymbol{e}_{\mathrm{KF}}\cdot\boldsymbol{e}_{\mathrm{CF}}} \tag{2}$$

$$t=\sqrt[2]{\boldsymbol{t}_{\mathrm{KF}}\cdot\boldsymbol{t}_{\mathrm{CF}}} \tag{3}$$

式中：$\boldsymbol{e}_{\mathrm{KF}}$ 和 $\boldsymbol{e}_{\mathrm{CF}}$ 分别为关键帧和当前帧位姿的欧拉角；$\boldsymbol{t}_{\mathrm{KF}}$ 和 $\boldsymbol{t}_{\mathrm{CF}}$ 分别为关键帧和当前帧位姿的平移向量；r 为两帧之间位姿的欧拉角的模长；t 为两帧之间平移的距离；$t_{\max}$ 和 $r_{\max}$ 分别为所有 t 和 r 中的最大值。选择与其重叠度最高的至多 5 个关键帧，根据三角测量原理，将其中的二维像素特征点 $\boldsymbol{x}=[u,\ v,\ 1]^{\mathrm{T}}$ 转换到世界坐标系中得到三维地图点 $\boldsymbol{M}=[X,\ Y,\ Z,\ 1]^{\mathrm{T}}$，简化的计算方程表示为

$$\begin{bmatrix}X\\Y\\Z\\1\end{bmatrix}=\boldsymbol{Z}\cdot\boldsymbol{T}_{\mathrm{cw}}^{-1}\boldsymbol{K}^{-1}\begin{bmatrix}u\\v\\1\end{bmatrix} \tag{4}$$

式中：$\boldsymbol{T}_{\mathrm{cw}}$ 为从世界坐标系到相机坐标系的变换矩阵；$\boldsymbol{K}$ 为相机的内参矩阵。利用轻量级相机跟踪得到的当前帧位姿，将世界坐标系下的地图点投影到当前帧中，得到特征点 $\boldsymbol{x}'$ 和投影深度 Z_{proj}。

计算特征点 $\boldsymbol{x}$ 与 $\boldsymbol{x}'$ 对应的地图点之间的夹角 φ，即视差角

$$\varphi=\frac{180}{\pi}\arccos\frac{(\boldsymbol{M}-\boldsymbol{t}_{\mathrm{KF}})\cdot(\boldsymbol{M}-\boldsymbol{t}_{\mathrm{CF}})}{\|\boldsymbol{M}-\boldsymbol{t}_{\mathrm{KF}}\|_2\|\boldsymbol{M}-\boldsymbol{t}_{\mathrm{KF}}\|_2} \tag{5}$$

如果这个角度大于 30°，那么该点可能存在被遮挡的情况，此时将其划分到动态区域。此外，将关键帧的特征点投影到当前帧中得到的投影深度 Z_{proj} 与在当前帧的深度图中直接获得的深度 Z' 进行比较，如果差值超过了某个阈值，也认为该特征点落在了动态区域中。

在当前帧的深度图中，利用获得的动态像素点进行区域增长，得到动态区域的像素级掩膜。图 4 给出了示例图像，其中第 2 行图像用黑色像素展示了利用上述方法得到的掩膜区域。可以看到，针对 2.2 节中提到的目标检测网络不能满足的情况，多视图几何方法都能给出补充的检测结果。

图 4　基于多视图几何方法得到的像素级掩膜

Fig.4　Mask (black pixels) obtained by multi-view geometry method

但是由于本方法依赖于 RGB-D 相机提供的深度信息，对于超出深度量程的物体，RGB-D 相机并不能提供准确的测量结果，这就导致多视图几何方法不能检测出距离相机过远的动态物体。同时，为了提高计算的准确度，对动态物体的每一次判断都至少需要 5 帧关键帧参与，这会给结果带来一定的滞后性。而目标检测网络直接对单帧图像进行检测，不受深度信息的限制，不依赖已有的关键帧，可以和多视图几何方法实现互补，达到更加精确的检测效果。

3　实验与分析

实验分别采用 TUM 数据集[27]和 KITTI 数据集[28]来评价系统的综合能力。本文实验环节采用文献[21]的设置，系统对于每个测试图像序列都运行了 10 次。所有实验均在 Intel CoreTM i5-114000F CPU、12 核主频 2.60 GHz、内存 16.0 GB 配置的 Windows 操作系统的台式机上完成。实验中使用文献[27]提出的绝对轨迹误差（Absolute trajectory error，ATE）来评估算法在定位上的性能。

为更好地反映系统的鲁棒性和稳定性，采用均方根误差(Root mean square error，RMSE)作为评价指标。

3.1 TUM数据集

RGB-D TUM数据集[27]由Microsoft Kinect传感器在不同室内场景下以30 Hz频率记录的39个序列组成，每个序列包括RGB图像、深度图像和使用高精度光学捕捉系统获得的标准轨迹。在名为sitting的低动态序列中，有两个人坐在桌子前一边说话一边做手势。在名为walking的高动态序列中，有两个人同时在背景和前景中持续行走，这对于标准的SLAM系统具有挑战性。对于sitting(s)和walking(w)这2种类型的序列，有xyz、rpy、半球面、静止这4种类型的相机运动，例如，xyz表示相机沿着x、y、z三个轴的方向运动。

表1展示了本文算法在TUM数据集上进行消融实验的结果。其中，算法YOLOv5-SLAM表示在SLAM系统中只使用YOLOv5对动态目标进行检测；算法Geo-SLAM表示在SLAM系统中只使用多视图几何方法对动态区域进行检测；算法YG-SLAM表示联合了YOLOv5和多视图几何的方法识别动态区域的增强型系统。从表1可以看出，在大多数序列中，使用改进算法YG-SLAM系统是最精确的。与只使用深度学习的YOLOv5-SLAM相比，增加了几何约束检测的改进算法对动态对象的检测更加细化。同时，由于弥补了目标检测网络的局限性，其在10次运行中的结果也更加稳定。

表1 改进算法进行消融实验的ATE

Table 1 ATE of ablation experiment with proposed method

图像序列	ORB-SLAM2[8]	YOLOv5-SLAM	Geo-SLAM	YG-SLAM	准确度提升/%
w_半球面/m	0.351	0.020	0.035	**0.018**	94.87
w_xyz/m	0.459	**0.013**	0.312	**0.013**	97.17
w_rpy/m	0.662	0.040	0.251	**0.032**	95.17
w_静止/m	0.090	0.008	0.009	**0.007**	92.22
s_半球面/m	0.020	**0.016**	0.018	0.017	15.00
s_xyz/m	**0.009**	0.010	**0.009**	0.010	−11.1

与原始的ORB-SLAM2系统相比较，在高动态的walking序列中，改进算法YG-SLAM对于原系统定位准确度的提升在92%以上；在低动态的sitting序列中，人的运动幅度小，进行筛除操作后留下的特征点距离相机较远，因此YG-SLAM的结果与原始ORB-SLAM2系统的结果接近。图5直观地展示了改进算法YG-SLAM与ORB-SLAM2的序列轨迹与真值的差距。

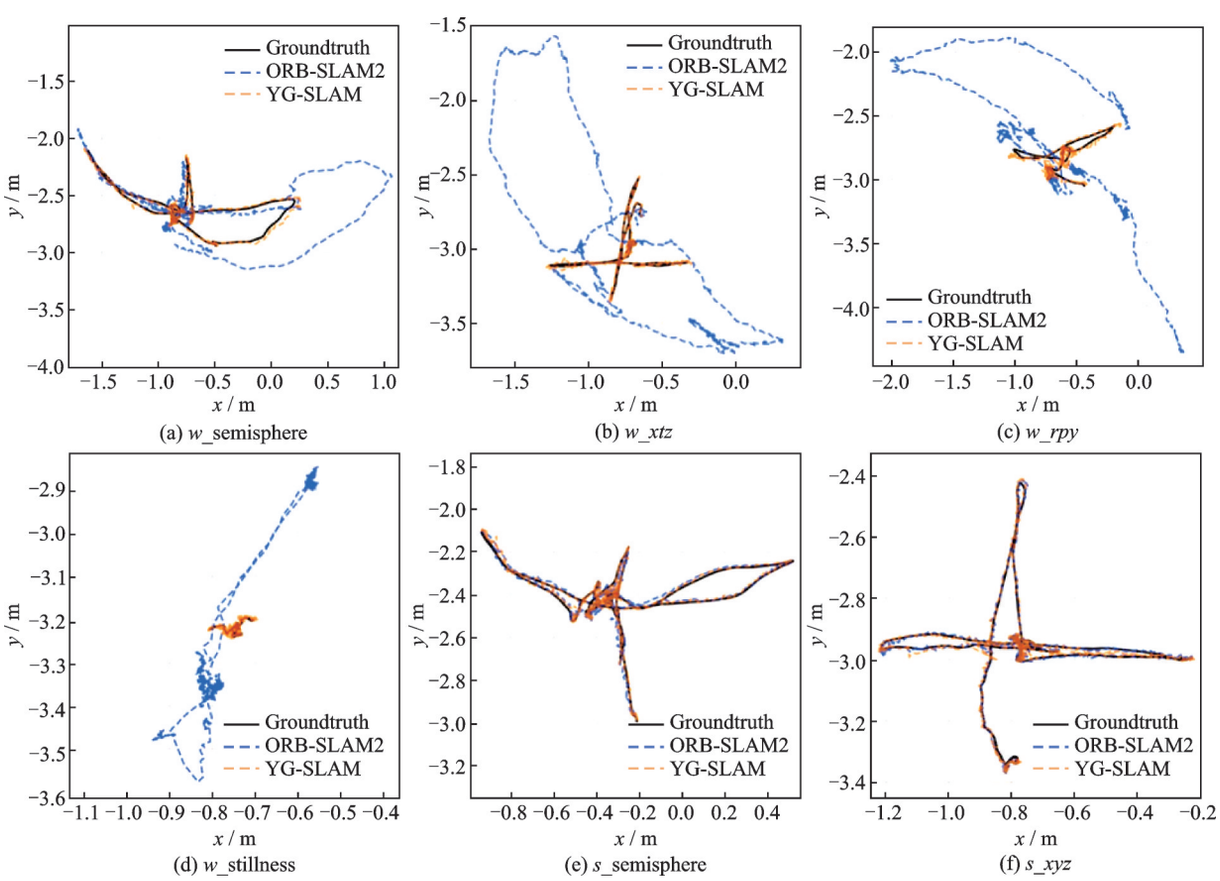


图5 TUM数据集序列轨迹

Fig.5 Trajectories of sequences from TUM dataset

为验证算法的先进性，实验内容将本文所提的 YG-SLAM 算法与目前先进的动态环境系统 DynaSLAM 进行了比较。表 2 给出了 YG-SLAM 和 DynaSLAM 在 TUM 数据集上的运行 10 次结果中的中值、最小值与最大值。可以看出，改进算法除了在 walking-静止的序列中的运行结果比 DynaSLAM 的运行结果略差，在其他序列中的运行结果都具有优势。DynaSLAM 使用 MaskR-CNN 和膨胀算法会增加物体边缘的检测结果的不确定性，从 10 次结果的最大值、中值和最小值的比较得出，本文 YG-SLAM 在大部分序列中的不确定性更小，结果更稳定。

表 2　基于 RGB-D 数据的 YG-SLAM 和 DynaSLAM 的 ATE

Table 2　ATE of YG-SLAM and DynaSLAM based on RGB-D data　m

图像序列	DynaSLAM[21]			本文 YG-SLAM		
	中值	最小值	最大值	中值	最小值	最大值
*w*_半球面	0.025	0.024	0.031	**0.018**	**0.018**	**0.021**
w_xyz	0.015	0.014	0.016	**0.013**	**0.013**	**0.014**
w_rpy	0.035	0.032	0.038	**0.032**	**0.028**	**0.037**
*w*_静止	**0.006**	**0.006**	**0.008**	0.007	0.007	0.009
*s*_半球面	0.017	0.016	0.020	0.017	**0.015**	0.020
s_xyz	0.015	0.013	0.015	**0.010**	**0.010**	**0.011**

表 3 给出了 YG-SLAM 与 DS-SLAM、Detect-SLAM、DP-SLAM 和 RGB-D SLAM 这几种目前先进的算法的运行结果比较，其中“—”表示计算失败情况。可以看出，YG-SLAM 的定位精度要优于其他方法。

表 3　基于 RGB-D 数据的 DS-SLAM、Detect-SLAM、DP-SLAM、RGB-D SLAM 和本文 YG-SLAM 的 ATE

Table 3　ATE of DS-SLAM, Detect-SLAM, DP-SLAM, RGB-D SLAM and YG-SLAM based on RGB-D data　m

图像序列	DS-SLAM[20]	Detect-SLAM[18]	DP-SLAM[22]	RGB-D SLAM[23]	本文 YG-SLAM
*w*_半球面	0.025 8	0.051 4	0.025 4	0.031 6	**0.018 0**
w_xyz	0.024 7	0.024 1	0.014 1	0.017 1	**0.013 0**
w_rpy	0.444 2	0.295 9	0.035 6	0.194 4	**0.032 0**
*w*_静止	0.008 1	—	0.007 9	0.009 1	**0.007 0**
*s*_半球面	—	0.023 1	0.018 2	**0.015 2**	0.017 0
s_xyz	—	0.020 1	—	0.012 3	**0.010 0**

3.2　KITTI 数据集

KITTI 数据集[28]包含了在城市和高速公路环境中行驶的汽车记录的双目视频序列，可以用来评估 SLAM 系统在户外动态环境下的定位性能。表 4 展示了本文提出的 YG-SLAM 算法在 11 个序列中运行的结果，与 ORB-SLAM2 和 DynaSLAM 的运行结果进行了对比，使用文献[27]中提出的 ATE 和文献[29]提出的相对平移误差（Relative pose error，RPE）与相对旋转误差（Relative translation and rotation errors，RRE）进行性能评估。表 5 给出了单目影像下的结果比较。

表 4　基于双目数据的 ORB-SLAM2、DynaSLAM 与 YG-SLAM 的 RPE、RRE 和 ATE

Table 4　RPE, RRE and ATE of ORB-SLAM2, DynaSLAM and YG-SLAM based on stereo data

图像序列	ORB-SLAM2[8]			DynaSLAM[21]			本文 YG-SLAM		
	RPE/%	RRE/((°)·100 m^{-1})	ATE/m	RPE/%	RRE/((°)·100 m^{-1})	ATE/m	RPE/%	RRE/((°)·100 m^{-1})	ATE/m
KITTI 00	**0.70**	0.25	1.3	0.74	0.26	1.4	0.71	0.25	1.3
KITTI 01	**1.39**	**0.21**	10.4	1.57	0.22	**9.4**	1.49	0.23	10.55
KITTI 02	0.76	0.23	**5.7**	0.80	0.24	6.7	**0.75**	0.23	5.9
KITTI 03	0.71	0.18	0.6	**0.69**	0.18	0.6	0.70	0.18	0.6
KITTI 04	0.48	0.13	0.2	0.45	**0.09**	0.2	**0.42**	0.11	0.2
KITTI 05	0.40	0.16	0.8	0.40	0.16	0.8	0.41	0.16	0.8
KITTI 06	0.51	**0.15**	0.8	0.50	0.17	0.8	**0.45**	0.19	**0.7**
KITTI 07	0.50	0.28	0.5	0.52	0.29	0.5	**0.48**	**0.26**	0.5
KITTI 08	1.05	0.32	3.6	1.05	0.32	**3.5**	1.07	0.32	3.7
KITTI 09	**0.87**	0.27	3.2	0.93	0.29	**1.6**	0.90	**0.26**	3.15
KITTI 10	0.60	0.27	**1.0**	0.67	0.32	1.2	**0.59**	**0.21**	1.1

表5 基于单目数据的ORB-SLAM2、DynaSLAM与YG-SLAM的ATE

Table 5 ATE of ORB-SLAM2, DynaSLAM and YG-SLAM based on monocular data m

图像序列	ORB-SLAM2[8]	DynaSLAM[21]	本文YG-SLAM
KITTI 00	**5.33**	7.55	6.37
KITTI 02	**21.28**	26.29	23.46
KITTI 03	1.51	1.81	**1.01**
KITTI 04	1.62	0.97	**0.74**
KITTI 05	4.85	**4.60**	5.19
KITTI 06	12.34	14.74	**12.17**
KITTI 07	2.26	2.36	**2.23**
KITTI 08	46.68	**40.28**	42.91
KITTI 09	6.62	**3.32**	7.97
KITTI 10	8.80	**6.78**	8.05

改进算法在单目和双目影像下的运行结果较为类似。可以看到对于某些序列，如KITTI 04，其中所有出现的车辆都在移动，因此使用改进算法的SLAM系统在其中的运行轨迹精度得到了提高。但在大部分序列中，出现的车辆大都停放在路边，处于静止状态，这会增加动态判别模块对动态区域的误判率，使提取的特征点数量减少，所以运行结果的绝对轨迹误差更大。不过，由于去除了具有移动潜力的对象，由静态环境的特征点生成的地图能够长期重复使用，这使得回环检测和重定位算法更加稳健。

3.3 运行时间分析

根据官方给出的数据，目前YOLOv5s模型[30]对一帧图像的处理时间最快可以达到0.009 s，而Mask R-CNN[31]的处理时间为0.195 s。在本文实验所使用的运行环境下实际运行时，YOLOv5s模型处理一帧图像的平均时间为0.068 s，而Mask R-CNN的平均处理时间为0.95 s，ORB-SLAM2对一帧图像进行跟踪的平均时间为0.027 s，而本文的增强型YG-SLAM每一帧的处理时间为0.139 s。其中由于几何约束中使用的区域增长算法是一种迭代的方法，其时间开销较大。在低动态序列中，由于运动幅度小，YG-SLAM使用几何约束方法的迭代次数会减少，计算时间会显著回升。尽管相比于ORB-SLAM2计算效率有所下降，但相比于DynaSLAM等使用了两阶段检测网络的系统而言，YG-SLAM可以保持较好的实时性。

4 结 论

本文提出了一个在传统多线程框架上进行增强型的视觉SLAM系统，联合深度学习与几何约束的方法检测图像中的动态区域，剔除不稳定的特征跟踪点，使SLAM系统能够在动态环境中对单目、双目和RGB-D影像数据具有更好的鲁棒性。本系统提高了相机跟踪的精度，并创建一个基于静态环境的可重复使用的场景地图。实验结果表明，与其他方法相比，本文方法在多数情况下都达到了良好的精度表现，运行结果也更加稳定，减少了由于动态对象检测给系统带来的不确定性。由于系统使用了深度学习的方法检测运动对象，当发生检测到的潜在动态物体并未进行运动的情况，例如静止的汽车，则会减少跟踪的特征点的数量，影响跟踪结果。因此，未来的研究工作将增加针对SLAM系统视频流中的动态物体运动模型估计以及运动参数计算的内容，进一步优化SLAM系统的鲁棒性。

参考文献：

[1] 丁文东，徐德，刘希龙，等. 移动机器人视觉里程计综述[J]. 自动化学报，2018，44(3)：385-400.
DING Wendong，XU De，LIU Xilong，et al. Review on visual odometry for mobile robots[J]. Acta Automatica Sinica，2018，44(3)：385-400.

[2] 邹雄，肖长诗，文元桥，等. 基于特征点法和直接法VSLAM的研究[J]. 计算机应用研究，2020，37(5)：1281-1291.
ZOU Xiong，XIAO Changshi，WEN Yuanqiao，et al. Research of feature-based and direct methods VSLAM[J]. Application Research of Computers，2020，37(5)：1281-1291.

[3] WHELAN T，LEUTENEGGER S，SALAS-MORENO R，et al. ElasticFusion：Dense SLAM without a pose graph[C]//Proceedings of the 11th Conference on Robotics-Science and Systems. Cambridge，USA：MIT Press，2015.

[4] KERL C，STURM J，CREMERS D. Dense visual SLAM for RGB-D cameras[C]//Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. New York，USA：IEEE，2013：2100-2106.

[5] 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.

[6] KLEIN G，MURRAY D. Parallel tracking and mapping for small AR workspaces[C]//Proceedings of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality. New York，USA：IEEE，2007：225-234.

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

[8] 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.

[9] FISCHLER M A, BOLLES R C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography[J]. Communications of the ACM, 1981, 24(6): 381-395.

[10] ENGEL J, SCHÖPS T, CREMERS D. LSD-SLAM: Large-scale direct monocular SLAM[C]// Proceedings of the 13th European Conference on Computer Vision. Berlin, German: Springer, 2014: 834-849.

[11] FORSTER C, PIZZOLI M, SCARAMUZZA D. SVO: Fast semi-direct monocular visual odometry [C]//Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA). New York, USA: IEEE, 2014: 15-22.

[12] FORSTER C, ZHANG Z, GASSNER M, et al. SVO: Semidirect visual odometry for monocular and multicamera systems[J]. IEEE Transactions on Robotics, 2016, 33(2): 249-265.

[13] ENGEL J, KOLTUN V, CREMERS D. Direct sparse odometry[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(3): 611-625.

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

[15] LEUTENEGGER S, LYNEN S, BOSSE M, et al. Keyframe-based visual-inertial odometry using nonlinear optimization[J]. The International Journal of Robotics Research, 2014, 34(3): 314-334.

[16] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: An efficient alternative to SIFT or SURF [C]//Proceedings of the 2011 IEEE International Conference on Computer Vision. New York, USA: IEEE, 2011: 2564-2571.

[17] 王柯赛, 姚锡凡, 黄宇, 等. 动态环境下的视觉SLAM研究评述[J]. 机器人, 2021, 43(6): 715-732.
WANG Kesai, YAO Xifan, HUANG Yu, et al. Review of visual SLAM in dynamic environment[J]. Robot, 2021, 43(6): 715-732.

[18] ZHONG F W, WANG S, ZHANG Z Q, et al. Detect-SLAM: Making object detection and SLAM mutually beneficial[C]//Proceedings of the 18th IEEE Winter Conference on Applications of Computer Vision. New York, USA: IEEE, 2018: 1001-1010.

[19] 高兴波, 史旭华, 葛群峰, 等. 面向动态物体场景的视觉SLAM综述[J]. 机器人, 2021, 43(6): 733-750.
GAO Xingbo, SHI Xuhua, GE Qunfeng, et al. A survey of visual SLAM for scenes with dynamic objects [J]. Robot, 2021, 43(6): 733-750.

[20] YU C, LIU Z X, LIU X J, et al. DS-SLAM: A semantic visual SLAM towards dynamic environments [C]//Proceedings of the 25th IEEE/RSJ International Conference on Intelligent Robots and Systems. New York, USA: IEEE, 2018: 1168-1174.

[21] BESCOS 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.

[22] LI A, WANG J K, XU M, et al. DP-SLAM: A visual SLAM with moving probability towards dynamic environments[J]. Information Sciences, 2021, 556: 128-142.

[23] 艾青林, 王威, 刘刚江. 室内动态环境下基于网格分割与双地图耦合的RGB-D SLAM算法[J]. 机器人, 2022, 44(4): 431-442.
AI Qinglin, WANG wei, LIU Gangjiang. RGB-D SLAM algorithm in indoor dynamic environments based on gridding segmentation and dual map coupling [J]. Robot, 2022, 44(4): 431-442.

[24] ZAIDI S S A, ANSARI M S, ASLAM A, et al. A survey of modern deep learning based object detection models[J]. Digital Signal Processing, 2022. DOI: 1048550/arXiv.2104.11892.

[25] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[EB/OL]. [2022-05-30]. https://doi.org/10.48550/arXiv.2004.10934.

[26] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//Proceedings of the 13th European Conference on Computer Vision. Berlin, German: Springer, 2014: 740-755.

[27] STURM J, ENGELHARD N, ENDRES F, et al. A benchmark for the evaluation of RGB-D SLAM systems[C]//Proceedings of the 25th IEEE/RSJ International Conference on Intelligent Robots and Systems. New York, USA: IEEE, 2012: 573-580.

[28] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: The KITTI dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[29] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2012: 3354-3361.

[30] Ultralytics. YOLOv5[EB/OL]. (2021-10-12). https://www.mstx.cn/ultralytics/yolov5.

[31] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of the 16th IEEE International Conference on Computer Vision. New York, USA: IEEE, 2017: 2961-2969.

(编辑:刘彦东)