文献标志码: A **中图分类号**:TP242.6

基于目标检测的视觉 SLAM 改进方法

王晓超[1]，王春林[1+]，袁成祥[2]

(1.杭州电子科技大学，自动化(人工智能)学院，浙江 杭州 310018；

2.浙江工商大学，计算机与信息工程学院，浙江 杭州 310018)

摘 要： 同时定位与地图构建（SLAM）系统大多假设场景是静态的，在动态场景中，系统性能会大大下降。传统的 SLAM 系统构建了几何信息地图，很难得到物体的语义信息。为了实现系统在动态场景中准确定位和构建包含语义信息的地图，提出了一种基于目标检测算法改进的视觉 SLAM 系统。该系统在 ORB-SLAM2 的跟踪线程中，采用目标检测算法 YOLOv4 去除动态特征提高系统位姿估计精度。建图线程中，对点云地图进行超体素聚类，与 YOLOv4 获取的物体标签融合构建语义地图。实验结果表明该视觉 SLAM 系统有效地去除了动态特征，减少了位姿估计误差，构建了分层清晰的语义地图。

关键词： 同时定位与地图构建；动态场景；深度学习；目标检测；语义地图

Improved Visual SLAM Method Based on Object Detection

WANG Xiaochao[1],WANG Chunlin[1],YUAN Chengxiang[2]

(1.College of Automation(Artificial Intelligence),Hangzhou Dianzi University,Hangzhou,Zhejiang 310018,China;

2.College of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou,Zhejiang 310018,China)

【Abstract】 Simultaneous Localization and Mapping(SLAM) systems are mostly assumed to be static, and the system performance will be greatly reduced in dynamic scene.The traditional SLAM system constructs the geometric information map, and it is difficult to get the object's semantic information.In order to accurately locate in the dynamic scene and construct the map with semantic information, an improved visual SLAM system based on object detectionis proposed. In the tracking thread of ORB-SLAM2,the system uses the object detection algorithm YOLOv4 to remove dynamic features and improve the accuracy of pose estimation. In the mapping thread, The point cloud map is segmented by Supervoxel, and the semantic map is constructed by merging the object labels obtained by Yolov4.The experimental results show that the system can effectively eliminate the dynamic features, reduce the error of pose estimation, and construct a clear and hierarchical semantic map.

【Key words】 SLAM; Dynamic scene;Deep learning;Object detection;Semantic map

0 概述

同时定位与地图构建（Simultaneous Localization and Mapping,SLAM）技术是机器人通过传感器对未知环境构建地图的同时实现自定位的过程。大多数视觉 SLAM 系统是将环境假设为静态场景，而在实际场景中往往会有像人类这样的活动对象，这些动态特征被提取后会严重影响相机的位姿估计，造成轨迹漂移严重，甚至导致系统崩溃。另外，这些 SLAM 系统主要构建的是几何信息地图，缺少对物体具体语义层次的理解，不能提供带语义信息的地图，制约了移动机器人交互能力和导航能力。

近年来，深度学习在语义信息获取方面的进展及应用为解决这些问题提供了一个可行的方向。深度学习在图像分类[1]、识别、图像分割[2]等几大领域的表现都远远高于传统人工设计的算法。深度学习与 SLAM 结合可以使机器人从几何和语义

基金项目: 浙江省自然科学基金(No.ZL20f030013)。

作者简介: 王晓超(1995-)，男，硕士研究生，研究方向为视觉 SLAM 与深度学习；王春林+，副教授，博士；袁成祥，讲师，博士。

E-mail: wchl@hdu.Edu.cn

两个层次对场景进行抽象理解，获得高层次的感知，提高机器人对周围环境的理解。Yu 等[3]提出了 DS-SLAM 系统，该系统是在 ORB-SLAM2[4]的跟踪线程中加入语义分割 SegNet 网络[5]，去除每一帧图像中的动态特征，从而降低位姿估计的误差。Bescos 等[6]提出 DynaSLAM 系统，采用实例分割网络 MASK-RCNN[7]对当前帧中的动态物体进行分割，并根据前 20 个关键帧对去除的地方进行背景修复。由于 MASK-RCNN 复杂的网络结构，DynaSLAM 系统实时性较差。Sünderhauf 等[8]同样以 ORB-SLAM2 系统为基础，在系统中加入 SSD[9]目标检测算法对构建的三维点云地图物体识别和分割，最终构建了带有语义信息的三维点云语义地图。Mccormac 等[10]提出 SemanticFusion 系统，该系统使用卷积神经网络（CNN）进行语义分割，再与 ElasticFusion 系统结合，构建了稠密的三维语义地图。

本文针对机器人视觉 SLAM 系统中存在的对动态物体难以处理和无法构建带有语义信息地图的问题，以 ORB-SLAM2 为基础，采用基于深度学习的目标检测算法 YOLOv4 对系统进行改进，减少系统位姿估计误差，提高系统的鲁棒性，并添加一个点云语义地图构建的线程，构建稠密的三维语义地图，以提高视觉 SLAM 系统的感知能力。

1 系统构成

本文提出的基于目标检测的视觉 SLAM 系统改进方法的框架如图 1 所示。本文采用 YOLOv4 对传统 ORB-SLAM2 系统进行部分改进。在跟踪线程中，采用 ORB 算法提取出图像帧中的特征点，然后判断这些特征点是否在由 YOLOv4 定位出的动态物体上，如果是，则去除这些特征点，避免这些动态特征点对系统的干扰，以提高系统位姿估计的准确率。在建图线程中，构建 3D 点云地图，再对点云地图采用基于图结构的超体素聚类算法生成初步语义地图，与 YOLOv4 提供的语义标签构建最后的语义地图。

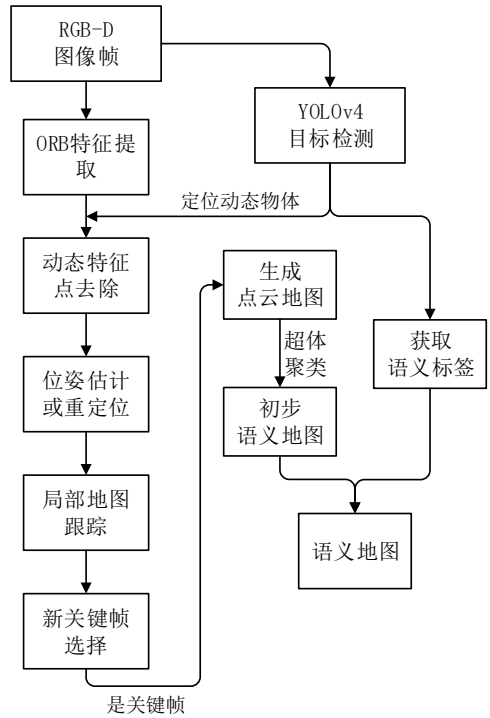


图 1 基于 YOLOv4 改进的 SLAM 系统框架图

2 基于 YOLOv4 的视觉 SLAM 改进方法

2.1 YOLOv4

YOLOv4[11]在以残差块结构的 Darknet-53 为骨干网络的 YOLOv3 基础上做出了全面的提升，在骨干网络 Darknet-53 上加入 CSPNet(Cross StagePartial Network)网络结构，减少计算量的同时提高了推理速度和准确性。另外，骨干网络还加入了 SPP(Spatial Pyramid Pooling)模块，可以提升模型的感受野，分离更重要的上下文信息。与此同时，还采用 PANet(Path Aggregation Network)改进骨干网络结构，加强了特征金字塔的结构，缩短了高低层特征融合的路径。

本文利用 MS COCO 数据集来训练 YOLOv4 的网络模型，数据集里包含人、茶杯、键盘、显示器、鼠标、玩具熊等 80 个类别。由于网络结构的优势，YOLOv4 可以在 MS COCO 数据集上 AP(Average Precision)为 43.5%的同时可以达到 65FPS(Frames Per Second)，是目前最新的快速而高效的目标检测器。

2.2 跟踪线程

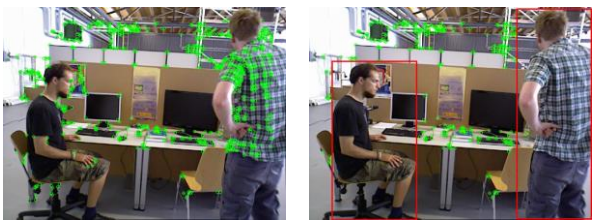
2.2.1 特征提取与特征匹配

ORB（Oriented FAST and Rotated BRIEF）[12]是一种快速特征点提取和特征描述的算法，由关键点和描述子两部分组成，ORB 特征提取主要分为以下两个步骤：一是方向 FAST 特征点检测；二是 BRIEF 特征描述。

特征点提取后，接下来就是特征匹配。特征匹配是 SLAM 系统中的重要部分，为保证位姿估计的精度，特征点间必须正确地、有效地相互匹配。以上的特征描述子均是二进制描述子，为后续计算特征相似度减少了计算压力。这里比较特征相似度采用计算汉明距离的方式，汉明距离是指两字符串之间对应位置字符不同的总数量。当两特征描述子的汉明距离低于设定的阈值时，便认为这两个特征是同一个点。

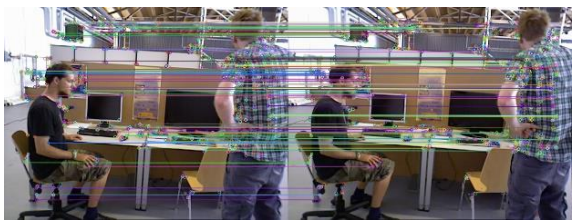
2.2.2 YOLOv4 去除动态特征

将当前图像帧输入 YOLOv4 网络模型，经过目标检测算法定位出动态物体的位置，本文将人类视为动态物体类别，将人类用矩形框框出来，并将框内的特征点视为动态特征点，全部去除，如图 2 所示。当动态特征剔除后，进行特征匹配时，避免了很多在人身上的特征点的匹配，以提高后续位姿估计的精度，如图 3 所示。

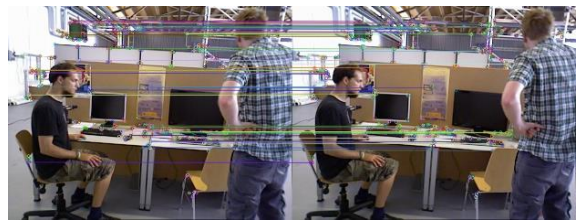


(a) 传统特征提取后　　(b) YOLOv4 改进后特征提取

图 2　特征提取结果图



(a) 传统 ORB-SLAM2 特征匹配结果



(b) YOLOv4 改进后特征匹配结果

图 3　特征匹配结果图

2.2.3 位姿估计

本文采用 PnP(Perspective-n-Point)[13]算法来计算相机的位姿估计。PnP 又叫重投影误差，是一种将匹配点从三维空间投影到像平面并计算误差来估计相机运动的方法，如图 4 所示。

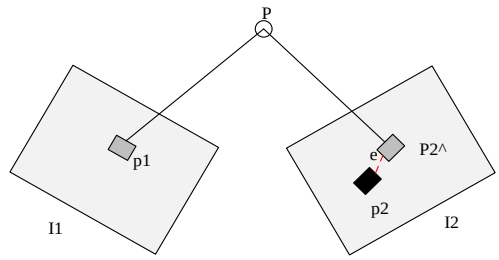


图 4　重投影误差示意图

其中，空间点 P 坐标 $P=[X,Y,Z,1]^T$，其在图像 I_1 中投影的像素坐标为 $p_1=[u_1,v_1]^T$，在图像 I_2 中投影的像素坐标为 $p_2^{'}=[u_2^{'},v_2^{'}]^T$，而点 P 在图像 I_2 中的观测值为 $p_2=[u_2,v_2]^T$，则重投影误差为 $e_2=p_2-p_2^{'}$。

图中的理想重投影过程表达如下：

$$d_2u_2=K\exp\left(\xi^{\wedge}\right)P \tag{1}$$

其中，d_2 表示空间点 P 在图像 I_2 所在相机坐标系的深度，K 表示相机内参，$\exp\left(\xi^{\wedge}\right)$ 表示相机从图像 I_1 到图像 I_2 的姿态变换阵，ξ 表示其对应李代数。

则重投影误差表示为：

$$e_2=u_2-\frac{1}{d_2}K\exp\left(\xi^{\wedge}\right)P \tag{2}$$

假设有 N 个特征点，则构成求相机位姿 ξ 的最小二乘问题：

$$\xi^{*}=\arg\min_{\xi}\frac{1}{2}\sum_{i=1}^{N}\|u_i-\frac{1}{d_i}K\exp\left(\xi^{\wedge}\right)P\|_2^2 \quad (3)$$

通过 G2O(General Graph Optimization)求解公式(3)的最小二乘问题，得到优化后的相机的位姿。

2.3 建图线程

本文在 ORB-SLAM2 建图线程后添加一个构建带有语义信息的点云地图线程。以 ORB-SLAM2 的点云地图为输入，采用基于超体素聚类(Supervoxel Clustering）的算法对点云地图进行初步分割，再与 YOLOv4 获得的语义标签融合构建最终的语义地图。

2.3.1 基于超体素聚类的物体分割算法

超体(Supervoxel)[14]是一种集合，集合的元素是“体”，其本质是一个个小方块。超体素聚类并不是分割出某个物体，而是对点云过分割(over segmentation)，将场景点云化成许多的小方块，根据颜色，法向量方向等进行局部分割。

超体素聚类算法是在空间中选择一定数量的种子点作为超体的初始化。首先将空间分割成半径为 R_{seed} 分辨率的体素网格，R_{seed} 的大小要比各像素间的范围 R_{voxel} 大得多，然后将种子作为每个体素网格的中心。如图 5 所示，其中 R_{search} 表示放置种子空间距离。

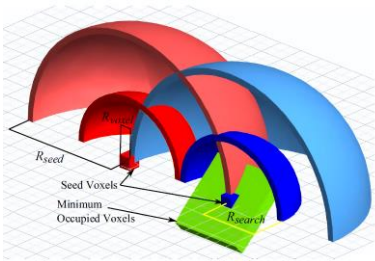


图 5 影响超体素聚类的不同半径

从种子点开始通过计算特征空间如空间范围，法向量，颜色等进行扩展，通过公式(4)计算两个素体之间的距离：

$$D=\sqrt{w_c D_c^2+\frac{w_s D_s^2}{3R_{seed}^2}+w_n D_n^2} \quad (4)$$

其中 D_c 为 RGB 空间中的归一化欧氏距离，D_s 为归一化的空间的欧氏距离，D_n 为归一化的法向量角度距离。

超体是使用迭代的方式增长的。首先从种子最近邻点开始，如果计算出体素是距离当前种子最近的点，则将该体素加入当前超体中，接着使用近邻图继续将该体素所有近邻点加入到搜索队列中，然后再处理下个种子，一直迭代到超体的边界。这种方式能保证在处理过程中每个像素里中心点的层次水平是相同的。在超体搜索结束后，接着更新每个超体的中心为其组成成分的重心点，这样经过几次迭代，直达超体中心稳定停止。

2.3.2 基于 YOLOv4 获取语义标签

以上将点云进行超体素聚类属于几何结构的分割，虽有不错的分割效果，但是不能得到物体的语义信息。为获取图像中的语义信息，本文采用 YOLOv4 进行目标检测获取图像帧的语义标签，目标检测效果如图 6 所示。由于点云地图属于三维，所以需将二维的语义标签投影到三维的点云地图中，实现一个合理的语义标签地图。再与之前基于超体素聚类获得的点云分割地图融合，构建成最终的语义地图。

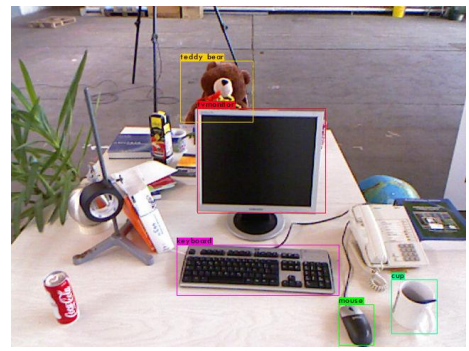


图 6 YOLOv4 目标检测

3 实验与分析

本节中，我们使用开源数据集对本文构建的系统进行评估，将本文系统、传统 ORB-SLAM2 和基于 SegNet 改进的 DS-SLAM 三系统进行比较，分析系统在动态环境中的性能是否提高，并构建带有语义信息的稠密地图。所有实验均在处

理器为 Intel i5 8400，显卡 GTX1060 6G 显存和内存 16GB 的计算机上进行。

3.1 数据集

为评价系统的综合能力，实验采用的数据集是德国慕尼黑工业大学开源的 TUM RGB-D 数据集[15]，该数据集由 39 个序列组成，这些序列是由 Microsoft Kinect 传感器以 30Hz 速率记录在不同的室内场景中，包含 RGB 图片，深度图片和地面实况数据。

本文实验主要采用TUM RGB-D数据集中的 5 个序列。freiburg3_walking_xyz 序列相机沿三个方向(x,y,z)移动，freiburg3_walking_static 序列相机保持在适当位置静止，freiburg3_walking_rpy 序列中相机沿主轴(滚转-俯仰-偏航)在相同的位置旋转，freiburg3_walking_halfsphere 序列中相机移动范围为直径一米的小半球。该 4 个序列都是高动态序列。最后一个 freiburg2_xyz 序列是相机沿着(x,y,z)方向缓慢移动，确保了数据足够清晰。

此外，TUM RGB-D 数据集还提供了用于系统评估的两种方法：一是绝对轨迹误差(Absolute Trajectory Error,ATE)，代表运动轨迹的全局一致性。二是相对位姿误差(Relative Pose Error,RPE)，测量平移和旋转漂移。

3.2 实验结果

3.2.1 定量结果

本小节为对比本文系统、ORB-SLAM2 和 DS-SLAM 在 TUM RGB-D 数据集中的 4 个高动态序列的实验结果，评价指标为均方根误差(RMSE)、平均误差(MEAN)、标准偏差(S.D.)。均方根误差(RMSE)计算估计值与真实值之间的偏差；平均误差(MEAN)描述所有估计误差的平均水平；标准偏差(S.D.)反映系统轨迹估计的离散程度。三种评价指标可以很好的体现 SLAM 系统的稳定性和可靠性。

从表 1-3 的比较结果可以看出，本文系统相比于传统的 ORB-SLAM2 系统和 DS-SLAM 系统在三种评价误差中均有不同程度的减少。图 7 显示了在 freiburg3_walking_halfsphere 序列中传统 ORB-SLAM2、DS-SLAM 和本文的 SLAM 系统的 ATE 和 RPE 图。由图可明显看出本文系统的绝对轨迹误差和相对位姿误差均降低了很多。

表 1 绝对路径误差(ATE)对比

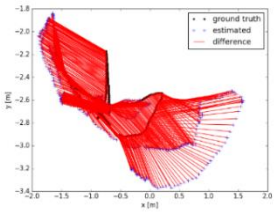
序列	ORB_SLAM2(m)			DS_SLAM(m)			本文系统(m)		
	RMSE	MEAN	S.D.	RMSE	MEAN	S.D.	RMSE	MEAN	S.D.
fr3_walking_xyz	0.8510	0.6839	0.4207	0.0237	0.0220	0.0108	**0.0161**	**0.0139**	**0.0080**
fr3_walking_static	0.3713	0.3527	0.1160	0.0103	0.0094	0.0041	**0.0069**	**0.0060**	**0.0033**
fr3_walking_rpy	0.5429	0.4954	0.2220	0.4560	0.3942	0.2291	**0.0271**	**0.0198**	**0.0187**
fr3_walking_halfsphere	0.7598	0.6812	0.3365	0.0580	0.0375	0.0443	**0.0322**	**0.0274**	**0.0169**

表 2 相对位移误差(RPE)对比

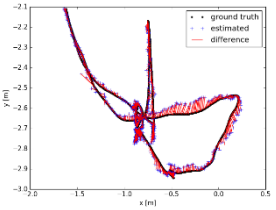
序列	ORB_SLAM2(m)			DS_SLAM(m)			本文系统(m)		
	RMSE	MEAN	S.D.	RMSE	MEAN	S.D.	RMSE	MEAN	S.D.
fr3_walking_xyz	1.3025	1.0707	0.7416	0.03721	0.0276	0.0141	**0.0230**	**0.0204**	**0.0106**
fr3_walking_static	0.5323	0.3885	0.3638	0.0168	0.0149	0.0076	**0.0098**	**0.0087**	**0.0045**
fr3_walking_rpy	0.7743	0.6634	0.4992	0.6517	0.5227	0.3892	**0.0385**	**0.0297**	**0.0230**
fr3_walking_halfsphere	0.3151	0.1864	0.2634	0.0837	0.0572	0.0611	**0.0460**	**0.0405**	**0.0218**

表 3 相对旋转误差(RPE)对比

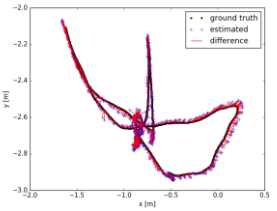
序列	ORB_SLAM2(°)			DS_SLAM(°)			本文系统(°)		
	RMSE	MEAN	S.D.	RMSE	MEAN	S.D.	RMSE	MEAN	S.D.
fr3_walking_xyz	7.7524	5.6683	5.1286	1.0577	0.8795	0.6837	**0.6509**	**0.5159**	**0.3969**
fr3_walking_static	4.0235	1.8643	3.6151	0.2994	0.2723	0.1244	**0.2809**	**0.2534**	**0.1212**
fr3_walking_rpy	7.6284	4.8835	5.2516	3.3458	1.9425	0.9743	**0.9256**	**0.7462**	**0.5477**
fr3_walking_halfsphere	7.0663	4.3241	4.9516	1.9046	1.3072	1.3851	**0.9202**	**0.8197**	**0.4181**



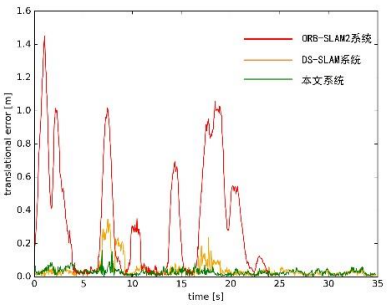
(a) ORB-SLAM2 绝对轨迹误差



(b) DS-SLAM 绝对轨迹误差



(c) 本文系统绝对轨迹误差



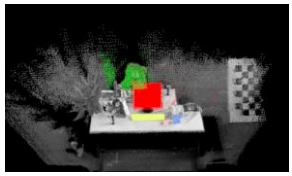
(d) 相对位姿误差对比

图 7　fr3_walking_halfsphere 序列中 ORB-SLAM2、DS-SLAM 以及本文系统的绝对轨迹误差和相对位姿误差对比图

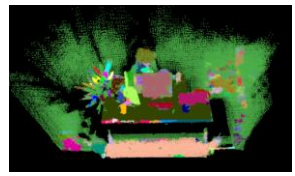
3.2.2 语义地图构建

本节实验是构建 freiburg2_xyz 序列数据集的稠密语义点云地图，其效果如图 8 所示。图 a 是传统 SLAM 系统所构建的点云地图，图中虽然可以看出物体的轮廓，但是各物体间分层不够清晰。图 b 是经过目标检测算法 YOLOv4 后，给几类物体的点云贴上了不同颜色的标签。图 c 是将点云进行超体素聚类后得到的地图，从图中可以看出物体成功的从周围环境中分割出来，但是超体素聚类是基于几何结构的分割，地图中并没有物体的语义信息。图 d 是将带语义标签的点云地图与超体素聚类的点云地图相融合后的结果，图中可以看出实验桌上物体分层清晰，由 YOLOv4 算法识别出的物体用不同颜色标出。构建带有语义信息的点云地图提高了 SLAM 系统的交互能力和感知能力。

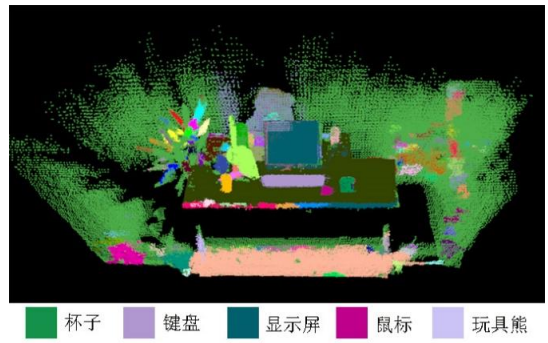
(a) 传统 SLAM 系统点云



(b) 语义标签点云



(c) 超体素聚类后的点云地图



(d) 融合后的语义点云地图

图 8 稠密语义点云构建过程效果图

4 结束语

为减少动态环境对 SLAM 系统的位姿和轨迹估计的影响和无法建立带有语义信息的地图，构建了一种基于目标检测的视觉 SLAM 系统。在 ORB-SLAM2 的跟踪线程中采用 YOLOv4 定位出动态物体，从而去除动态特征。与传统的 ORB-

SLAM2 相比，系统的性能有了明显的提升。在建图线程中加入构建稠密点云语义地图线程，用超体素聚类对点云地图初步分割，再与 YOLOv4 获取的标签融合得到最终的语义地图，地图构建效果良好。在接下来的工作中，重点研究如何在效果和本文差不多或更好的情况下提高系统的实时性。

参考文献:

[1] Dubnicki C, Ungureanu C, Kilian W. FPN: A distributed hash table for commercial applications[C]// Proceedings. 13th IEEE International Symposium on High performance Distributed Computing, 2004. IEEE, 2004: 120-128.

[2] Choudhury A R, Vanguri R, Jambawalikar S R, et al. Segmentation of Brain Tumors Using DeepLabv3+ [C]//International Miccai Brainlesion Workshop. Springer, Cham, 2018.

[3] Yu C, Liu Z, 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). IEEE, 2018.

[4] Mur-Artal R, Tardos J D.ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras[J]. IEEE Transactions on Robotics, 2017:1-8.

[5] Badrinarayanan, Kendall, Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12):2481-2495.

[6] Bescos, Fácil, Civera, et al. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 4076-4083.

[7] He, Gkioxari, Dollár, et al. Mask r-cnn[C]. Proceedings of the IEEE international conference on computer vision, 2017: 2961-2969.

[8] Niko Sünderhauf, Pham T, Latif Y, et al. Meaningful maps with object-oriented semantic mapping[C] //IEEE/RSJ International Conference on Intelligent Robots & Systems. IEEE, 2017.

[9] Liu, Anguelov, Erhan, Szegedy, Berg. SSD: Single Shot MultiBox Detector[C]. European Conference on Computer Vision, 2016

[10] Mccormac J , Handa A , Davison A , et al. SemanticFusion: Dense 3D Semantic Mapping with Convolutional Neural Networks[J]. 2016.

[11] Bochkovskiy, Alexey, Chien Y W, et al. "YOLO-v4: Optimal Speed and Accuracy of Object Detection." arXiv preprint arXiv:2004.10934, 2020.

[12] Rublee E, Rabaud V, Konolige K, et al. ORB:An efficient alternative to SIFT or SURF[C]//2011 International Conference on Computer Vision. IEEE, 2012.

[13] 高翔,刘毅,张涛.视觉 SLAM 十四讲——从理论到实践[M].电子工业出版社,2017.

[14] Papon J, Abramov A, Schoeler M, et al. Voxel Cloud Connectivity Segmentation-Supervoxels for Point Clouds[C]//Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013.

[15] Jürgen Sturm, Engelhard N , Endres F, et al. A benchmark for the evaluation of RGB-D SLAM systems[C]// IEEE/RSJ International Conference on Intelligent Robots & Systems. IEEE, 2012

联系方式：

邮编：310018

地址：浙江省杭州市江干区白杨街道杭州电子科技大学

联系人：王晓超

电话：19858190416

邮箱：630572121@qq.com

联系人：王春林

电话：13221011256

邮箱：wchl@hdu.Edu.cn

联系人：袁成祥

电话：13588013730

邮箱：yuancx_zjgsu@126.com