电子科技大学学术学位硕士研究生学位论文文献综述

|  |  |  |
| --- | --- | --- |
| 姓名：武博文 | 学号：202221090225 | 学科：软件工程 |
| 综述题目：室外动态场景下基于实例分割的视觉SLAM研究 | | |
| 导师意见：    导师签字：  日期： | | |

1. **选题依据和研究意义**

同步定位与地图构建（SLAM）是搭载激光雷达、IMU（Inertial Measurement Unit）、相机等传感器的移动载体在未知环境下同步进行定位与地图构建的过程[1]。SLAM一般可分为激光SLAM和视觉SLAM。激光SLAM利用激光雷达、编码器和惯性测量单元（IMU）等多种传感器相结合，已在理论和应用方面相对成熟。然而，激光雷达具有较高的价格使其难以普及到个人小型设备，并且雷达信息获取量有限。视觉SLAM利用视觉传感器，如单目、双目和RGB-D（带有深度信息的彩色图像）相机等，来构建环境地图。相机能够获取丰富的图像信息，并且视觉传感器具有低廉的价格，简单的结构和小巧便携的特点，因此成为近年来研究者们关注的热点，也成为SLAM技术中的主要研究方向。视觉SLAM能够广泛应用于无人驾驶，自主机器人，导盲避障等领域，对视觉SLAM的研究具有现实意义。

经过近二十年的发展，视觉同时定位与建图（Visual Simultaneous Localization And Mapping，V-SLAM）框架已趋于成熟，在机器人视觉感知领域中占有重要地位，最先进的V-SLAM算法提供了高精度定位和场景重建的能力[2]。现阶段，V-SLAM系统大多数建立在非动态环境的假设上，即假设移动载体在跟踪过程中不存在动态物体。然而，这种假设是一种强假设，在现实场景中很难成立。在室内场景中，常出现移动的人和桌椅等等；在室外场景中，常常出现移动的车和动物等等，这些动态物体的出现对V-SLAM系统的影响巨大，尤其是对V-SLAM中的前端模块的影响。SLAM前端求解存在两种方案，直接法和特征点法。直接法基于光度不变假设来描述像素随时间在图像之间的运动方式，每个像素在两帧之间的运动是一致的，通过此估计相机的运动。然而由于相机获得的图像受到光线，噪声等影响，光度不变假设往往不成立，如果再出现动态物体，直接使用此方法更会影响相机的运动估计。特征点法是一种间接的方法，它首先提取图像的特征点，然后通过两帧间特征点的匹配和位置变化求解相机运动。特征点的选择与使用大幅提高了V-SLAM系统定位的准确性，例如著名开源视觉SLAM框架ORB-SLAM2[3]、ORB-SLAM3[4]、VINS-Mono[5]都采用了特征点法。但是，一旦出现动态物体，这些特征点中会包含动态物体上的点，动态物体的移动造成了特征点移动的不一致性，从而对相机运动的估计造成了巨大影响。这种影响会导致后端失效，定位精度大幅减弱，不能忽视。随着视觉SLAM技术的发展，如何解决动态影响受到广泛关注，具有重要的研究价值。

1. **国内外研究现状和发展态势**

**2.1 视觉SLAM研究现状**

视觉SLAM问题最早可追溯到滤波技术的提出，Smith等人提出了采用状态估计理论的方法处理机器人在定位和建图等方面的问题[6]。随后出现各种基于滤波算法的SLAM系统，例如粒子滤波[7]和卡尔曼滤波[8]。2007年视觉SLAM取得重大突破，A. J. Davison等人提出第一个基于单目相机的视觉SLAM系统MonoSLAM[9]。该系统基于扩展卡尔曼滤波算法（Extended Kalman Filter, UKF），是首个达到实时效果的单目视觉SLAM系统，在此之前其他的算法都是对预先拍好的视频进行处理，无法做到同步。MonoSLAM的发布标志着视觉SLAM的研究从理论层面转到了实际应用，具有里程碑式意义。同年，Klein 等人提出了PTAM( Parallel Tracking And Mapping) [10]，创新地以并行的方式进行跟踪和建图线程，解决了MonoSLAM计算复杂度高的问题，这种并行的方式也是当下SLAM框架的主流。PTAM应用了关键帧和非线性化优化理论而非当时多数的滤波方案，为后续基于非线性化优化的视觉SLAM开辟了道路。

2014年慕尼黑工业大学计算机视觉组Jakob Engel等人[11]提出LSD-SLAM，该方案是一种基于直接法的单目视觉SLAM算法，不需要计算特征点，通过最小化光度误差进行图像像素信息的匹配，实现了效果不错的建图，可以生成半稠密的深度图。该方案的出现证明了基于直接法的视觉SLAM系统的有效性，为后续的研究奠定了基础。但该方案仍旧存在尺度不确定性问题，以及在相机快速移动时容易丢失目标的问题等等。同年SVO（semi-direct monocular visual odometry）被Forster等人提出[12]。这是一种基于稀疏直接法的视觉SLAM方案，结合了特征点和直接法，使用了特征点，但是不计算特征点的描述子，特征点的匹配使用特征点周围像素利用直接法匹配。SVO有着较快的计算速度，但是缺少了后端的功能，对相机的运动估计有较为明显的累计误差，应用场景受限。

2015年Mur-Artal等人参考PTAM关键帧和并行线程的方案，提出了ORB-SLAM框架[13]。该框架是一种完全基于特征点法的单目视觉SLAM系统，包括了跟踪，建图和回环检测三个并行线程。跟踪线程负责提取ORB[14]（oriented FAST and rotated BRIEF）特征点，这该系统最为经典的一部分，采用的ORB特征点具有良好的尺度不变性和旋转不变性，能实现提取速度和效果的平衡。跟踪线程还完成估计位姿的工作，并且适时选出新的关键帧来实现建图。建图线程接收跟踪线程选出的关键帧，删除冗余的关键帧和地图点，再进行全局优化。回环线程接收建图线程筛选后的关键帧，与其他关键图进行回环检测，然后更新相机位姿和地图。ORB-SLAM因为回环检测线程的加入，有限消除了累计误差的影响，提高了定位和建图的准确性。但是其系统只适用于单目相机，精度低且应用场景受限。随着相机的进步，2017年Mur-Artal 等人对ORB-SLAM进行了改进，扩展了对双目和RGB-D相机的支持，提出ORB-SLAM2[3]。相比于原版，该系统支持三种相机，同时新增重定位，全局优化和地图复用等功能，更具鲁棒性。

2017年，香港科技大学Qin Tong等人[15]提出VINS Mono系统，该系统在单目相机中融合IMU传感器，在视觉信息短暂失效时可利用IMU估计位姿，视觉信息在优化时可以修正IMU数据的漂移，两者的结合表现出了优良的性能。2019年提出改进版系统VINS-Fusion[16]，新增对双目相机和GPS传感器的支持，融合后的系统效果更优。

2020年Carlos Campos等提出了ORB-SLAM3[4]，该系统在ORB-SLAM2的基础上，加入了对视觉惯性传感器融合的支持，并在社区开源。系统对算法的多个环节进行改进优化，例如加入了多地图系统和新的重定位模块，能够适应更多的场景，同时精度相比上一版增加2-3倍。在2021年底，系统更新了V1.0版本，继承了ORB-SLAM2的优良性能，成为现阶段最有代表性的视觉SLAM系统之一。

**2.2 动态SLAM研究现状**

针对动态物体的影响，已经有许多研究人员开展了相关工作，尝试解决动态场景下的视觉SLAM问题。解决这一问题的主要挑战就是如何高效地检测到动态物体和其特征点，并将动态特征点剔除以恢复相机运动。

最早的解决思路是根据几何约束来筛除动态物体的特征点，如WANG等[17]首次使用K-Means将由RGB-D相机计算的3D点聚类，并使用连续图像之间的极线约束计算区域中内点关键点数量的变化，内点数量较少的区域被认定是动态的。利用极线约束是一种判断动态物体特征点的常见方法，但是如果相邻帧间存在高速移动物体或者运动物体沿着极线方向移动，这种方法效果会大大减弱。为了更好地利用几何信息，研究人员提出借助光流信息来提高动态物体的检测。Fang[18]使用光流法检测图像之间的动态物体所在位置，对其特征点进行滤除。该方法利用光流提高检测的精度，有效地降低了帧之间极线约束的误差。尽管基于几何约束的方法可以在一定程度消除动态特征点的影响，但随着深度学习的发展，图像中语义信息逐渐被重视和利用起来。

现阶段有许多优秀的深度学习网络，如YOLO[19]，SegNet[20]，Mask R-CNN[21]等等。这些神经网络有着强大的特征提取能力和语义信息提取能力，可以帮助SLAM系统更轻松地辨别出动态物体的存在，提供语义先验信息，从而消除其影响。Fangwei Zhong等人提出的Detect-SLAM[22]，利用目标检测网络获取环境中的动态的人和车等，为了实时性，只在关键帧中进行目标检测，最后去除所有检测到的动态点来恢复相机位姿。LIU和MIURA[23]提出了 RDS-SLAM。基于ORB-SLAM3[4]的RDS-SLAM框架使用模型的分割结果初始化移动对象的移动概率，将概率传播到随后的帧，以此来区分动静点。这种只基于深度学习的方法仅能提供图像中的语义信息，但无法判断图像中的物体是否真的在运动，比如静止的人或者路边停靠的汽车。若根据语义信息将其标记为动态物体后直接去除其特征点，这种方法会导致系统丢失有用的特征点，对相机的运动估计有所影响。因此仅利用深度学习不能很好解决动态物体对SLAM系统的影响。

许多研究开始探索语义信息和几何信息的结合。清华大学Chao Yun等提出的DS-SLAM[24]，该系统首先利用SegNet网络进行语义分割，再利用极线约束过滤移动的物体，达到了不错的效果。Berta Bescos等人首次利用Mask R-CNN网络进行实例分割，提出了DynaSLAM[25]。该系统结合基于多视几何深度的动态物体分割和区域生长算法，大幅降低了位姿估计的误差。Runz等人提出了MaskFusion，一种考虑物体的语义和动态RGD-D SLAM系统[26]。这个系统基于MASK-RCNN语义分割和几何分割，将语义分割和SALM线程放在两个线程以保证整个SLAM系统的实时性。但是该系统物体边界分割常包含背景，仍有改善空间。等人提出RS-SLAM，一种使用RGB-D相机解决动态环境不良影响的SLAM[27]。该系统采用语义分割识别动态对象，通过动态对象和可移动对象的几何关系来判断可移动对象是否移动。动态内容随后被剔除，跟踪模块对剔除过的静态背景图像帧进行ORB特征提取并估计相机位姿。

利用深度学习得来的语义信息和几何信息结合来解决SLAM中的动态场景问题渐渐成了一种主流，但是上述大多系统只是为了恢复相机的位姿而剔除动态物体的特征点，而没有估计动态物体的位姿。同时估计相机运动和跟踪动态物体运动，将动态物体的点加入优化步骤正在发展为一种趋势。Henein等人提出一种新的基于特征的，无模型的动态SLAM算法Dynamic SLAM（Dynamic SLAM: The Need For Speed）[28]。该方法利用语义分割场景中的刚体物体的运动，并提取运动物体的速度，有效性在各种虚拟和真实数据集上得到了验证。Javier Civera等人提出的DOT SLAM（Dynamic Object Tracking for Visual SLAM）[29]主要工作在前端，结合实例分割为对态对象生成掩码，通过最小化光度重投影误差跟踪物体。AirDOS被卡内基梅隆大学Yuheng Qiu等人提出[30]，将刚性和运动约束引入模型铰接对象，通过联合优化相机位姿、物体运动和物体三维结构，来纠正相机位姿估计。VDO SLAM[31]利用Mask R-CNN掩码和光流区分动静点，将动态环境下的SLAM表示为整体的图优化，同时估计相机位姿和物体位姿。

总体来说，目前动态场景下的视觉SLAM问题的解决需要借助几何信息和深度学习的语义信息，语义信息提供更准确的物体，几何信息提供物体真实的运动状态，两者结合来估计相机运动和跟踪物体。

**参考文献**

1. 孔德磊, 方正. 基于事件的视觉传感器及其应用综述[J]. 信息与控制, 2021, 50(1): 1-19. KONG D L, FANG Z. A review of event-based vision sensors and their applications[J]. Information and Control, 2021, 50(1): 1-19.
2. J. Engel, V. Koltun, and D. Cremers, “Direct sparse odometry,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 3, pp. 611–625, Mar. 2016.
3. Mur-Artal R , JD Tardós. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras[J]. IEEE Transactions on Robotics, 2017.
4. Campos C, Elvira R, Rodriguez J, et al. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM[J]. IEEE Transactions on Robotics: A publication of the IEEE Robotics and Automation Society, 2021, 37(6): 1874-1890.
5. Tong, Qin, Peiliang, et al. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator[J]. IEEE Transactions on Robotics, 2018.
6. Smith R, Self M, Cheeseman P. Estimating Uncertain Spatial Relationships in Robotics [J]. Machine Intelligence & Pattern Recognition, 1988, 5(5):435-461.
7. Grisetti G, Stachniss C, Burgard W. Improved Techniques for Grid Mapping With Rao-Blackwellized Particle Filters [J]. IEEE Transactions on Robotics, 2007, 23(1):34-46.
8. Kalman R E. A New Approach To Linear Filtering and Prediction Problems[J]. Journal of Basic Engineering, 1960, 82D:35-45.DOI:10.1115/1.3662552.
9. 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.
10. Klein G, Murray D. Parallel Tracking and Mapping for Small AR Workspaces[C]. IEEE and ACM International Symposium on Mixed and Augmented Reality. IEEE, 2007:1-10.
11. ENGEL J, SCHOPS T, CREMERS D, LSD-SLAM: Large-scale direct monocular SLAM[C]. European Conference on Computer Vision(ECCV), 2014:834–849.
12. FORSTER C, PIZZOLI M, SCARAMUZZA D. SVO: Fast semi-direct monocular visual odometry[C]. Hong Kong, China: IEEE International Conference on Robotics and Automation (ICRA), 2014: 15-22.
13. MURARTAL R, MONTIEL J M, TARDOS J D. ORB-SLAM: A Versatile and Accurate Monocular SLAM System[J]. IEEE Transactions on Robotics, 2015, 31(5):1147-1163.
14. Rublee E,Rabaud V,Konolige K,et al.ORB:An efficient alternative to SIFT or SURF[C].2011 International conference on computer vision. IEEE, 2011:2564-2571.
15. TONG Q, PEILIANG L, SHAOJIE S. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator[J]. IEEE Transactions on Robotics, 2017,99:1-17.
16. QIN T, PAN J, CAO S, et al. A general optimization-based framework for local odometry estimation with multiple sensors[J]. ArXiv, 2019:1901.03638.
17. WANG R, WAN W, WANG Y, et al. A new RGB-D SLAM method with moving object detection for dynamic indoor scenes[J]. Remote Sensing, 2019, 11(10): 1143.
18. Fang Y, Dai B. An improved moving target detecting and tracking based on Optical Flow technique and Kalman filter[J]. IEEE, 2009.DOI:10.1109/ICCSE.2009.5228464.
19. Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]//Computer Vision & Pattern Recognition. IEEE, 2016.DOI:10.1109/CVPR.2016.91.
20. 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.
21. Gkioxari G, He K, Piotr Dollár, et al. Mask R-CNN [J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 42(2): 386-397.
22. Zhong F, Wang S, Zhang Z, et al. Detect-SLAM: Making Object Detection and SLAM Mutually Beneficial[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018.DOI:10.1109/WACV.2018.00115.
23. LIU Y, MIURA J. RDS-SLAM: Real-time dynamic SLAM using semantic segmentation methods[J]. IEEE Access, 2021, 9: 23772-23785.
24. C. Yu, et al. DS-SLAM: A Semantic Visual SLAM towards Dynamic Environments[A]. //2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018, pp. 1168-1174.
25. B. Bescos, J. M. Fácil, J. Civera and J. Neira, DynaSLAM: Tracking, Mapping, and Inpainting in Dynamic Scenes[J]. IEEE Robotics and Automation Letters, 2018, 3(4)4076-4083.
26. Runz M, Buffier M, Agapito L. MaskFusion: real-time recognition, tracking and reconstruction of multiple moving objects[J]. 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), 2018, pp. 10-20.
27. T. Ran, L. Yuan, J. Zhang, D. Tang and L. He. RS-SLAM: A Robust Semantic SLAM in Dynamic Environments Based on RGB-D Sensor[J]. IEEE Sensors Journal, 2021, vol. 21, no. 18, pp. 20657-20664.
28. M. Henein, J. Zhang, R. Mahony and V. Ila. Dynamic SLAM: The Need For Speed[C]. 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020: 2123-2129.
29. Ballester I, Fontan A, Civera J, et al. DOT: Dynamic Object Tracking for Visual SLAM[J]. 2020.DOI:10.48550/arXiv.2010.00052.
30. Qiu Y, Wang C, Wang W, et al. AirDOS: Dynamic SLAM benefits from articulated objects[C]//2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022: 8047-8053.
31. Zhang J, Henein M, Mahony R, et al. VDO-SLAM: A Visual Dynamic Object-aware SLAM System[J]. 2020.DOI:10.48550/arXiv.2005.11052.