电子科技大学专业学位硕士研究生　　学位论文文献综述

姓名：陈佐玓	学号：202222090508	工程领域：电子信息
综述题目：基于 RGB-D 相机的物体级语义 SLAM 算法研究		
导师意见： 校内导师： 企业方导师：		

基于 RGB-D 相机的物体级语义 SLAM 算法研究

文献综述

摘要： 传统 SLAM 系统主要依赖于几何信息，然而，在实际应用中，对环境进行深度理解所需的信息不仅仅包括几何结构，还需要对环境中不同物体的语义信息有更精准的认知。深度学习技术，充分利用图像中的语义信息，不仅能够感知环境的形状，还能精准认知其中的不同物体及其语义含义。这种技术的引入为语义 SLAM 系统注入了更为高级的理解能力，使机器能够更全面地理解和应对复杂的现实世界场景，为智能机器人在导航、交互等任务中提供了更强大的感知和认知基础。这一演进推动了 SLAM 技术的发展，将其从传统的几何定位演变为对语义信息更敏感的感知系统。

关键词：目标检测、SLAM、语义、物体重建

一、研究背景

由于对机器人在未知环境中实时感知和导航的现实需求，SLAM（Simultaneous Localization and Mapping）开始被研究者提出。自 1986 年首次提出以来，SLAM 引起了众多研究人员的广泛关注，并在机器人、虚拟现实等领域迅速发展。 SLAM 是指基于位置和地图的自我定位，以及基于自我定位构建增量地图。主要用于解决机器人在未知环境中移动时的定位和地图构建问题[1]。SLAM 作为一项基础技术，早期已应用于移动机器人定位与导航。随着计算机技术（硬件）和人工智能（软件）的发展，机器人研究受到越来越多的关注和投入。许多研究人员致力于让机器人变得更加智能。 SLAM 被认为是促进移动机器人真正自主的关键。然而，传统 SLAM 主要关注几何信息，对于动态物体和语义理解的处理存在一定的局限性。随着对机器人系统功能的不断提升以及对更丰富环境理解的需求，语义 SLAM 在广泛的关注之下出现。语义 SLAM 引入语义信息，使机器不仅能够理解环境的几何结构，还能够对不同物体进行语义区分，从而更全面、深入地理解环境。

如图 1 所示，视觉 SLAM 和语义 SLAM 相关论文的引用次数正在迅速增加。尤其是在 2017 年左右，视觉 SLAM 和语义 SLAM 的引用量猛增。传统的 VSLAM 研究已经取得了许多进展。为了使机器人能够从更高的层面感知周围环境，语义 VSLAM 的研究受到了广泛的关注。近年来，语义 SLAM 受到越来越多的关注。

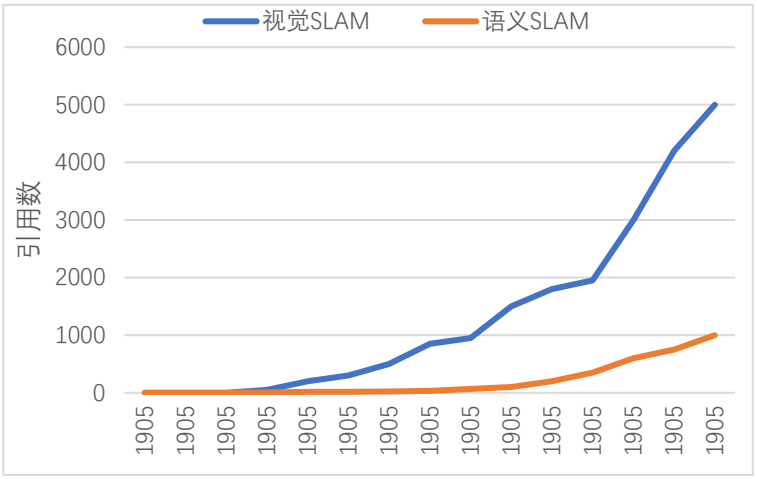

图 1 关于语义 SLAM 的论文数

视觉 SLAM（VSLAM）由于其传感器成本低廉、易于与其他传感器融合、环境信息更丰富等优点而得到迅速发展。传统的基于视觉的 SLAM 研究已经取得了许多成果，但在充满挑战的环境中可能无法达到预期的结果[2]。语义 SLAM 引入神经网络获取图像中的语义信息，辅助 SLAM 系统的进行。但是需要考虑目标检测的耗时问题、识别准确率以及物体的构建等问题。

语义 SLAM 系统由两个基本组件构成：语义提取器和现代 V-SLAM 框架。具体来说，语义信息的提取和导出主要经过两个过程，目标检测和语义分割[3]。目标检测主要分为有两个阶段：(1) 获取对象的二维位置；(2)分类对象。目标检测用于确定相关目标的 2D 位置以及分类，语义分割是在像素级层面进行目标像素的获取。V-SLAM 框架整体分为五个部分，如图 2 所示。

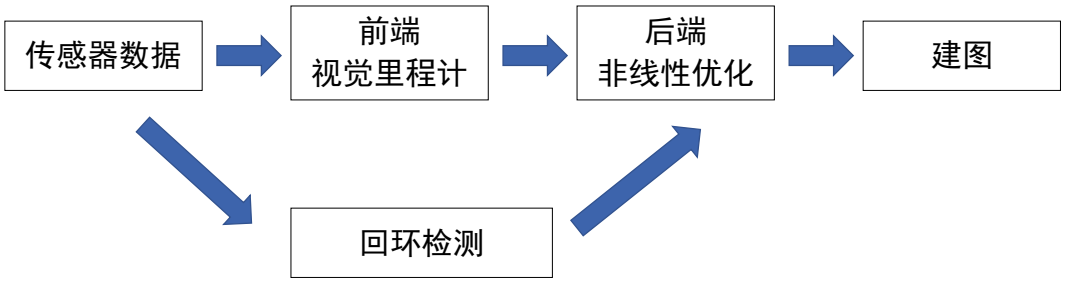

图 2 SLAM 流程框架

前端视觉里程计在相邻图像之间追踪特征点或特征描述符的运动，从而推断相机的位姿变化。后端非线性优化后端不同时刻视觉里程计测量的相机位姿，以及回环检测的信息，以最小化误差，提高对机器人轨迹和环境地图的估计精度，得到全局一致的轨迹和地图。回环检测判断机器人是否到达过先前的位置。如果检测到回环，它会把信息提供给后端进行处理。建图部分根据估计的轨迹，建立

与任务要求对应的地图。

二、研究现状

2.1 SLAM 发展阶段

SLAM 技术是机器人领域中一个重要的研究方向，其目标是使移动机器人在未知环境中能够实时进行自身位置估计和地图构建。Cesar Cadena 等人将 SLAM 发展历程分为了三个阶段，如图 3 所示，经典时代(1986-2004)、算法时代(2004-2015)以及鲁棒感知时代(2015-现在) [4]。在经典时代，引入了 SLAM 的主要概率公式，包括基于扩展卡尔曼滤波器、RaoBlackwellized 粒子滤波器和最大似然估计等方法；此外，这个时期还涵盖了与效率和强大的数据关联相关的基本挑战。Durrant-Whyte 和 TimBailey 的两项工作[5,6] 对经典时代的早期发展和主要公式结论进行了详细回顾，内容基本全面覆盖了整个经典时代的发展。接着是算法时代，Gamini Dissanayake 等人的工作[7]回顾内容涵盖这个时期的一些发展，并提出了一些 SLAM 面临的一些挑战。目前，我们正处于鲁棒感知时代，其中涉及到一些新的挑战如，鲁棒性能、高层次理解、资源感知和任务感知、驱动推理。

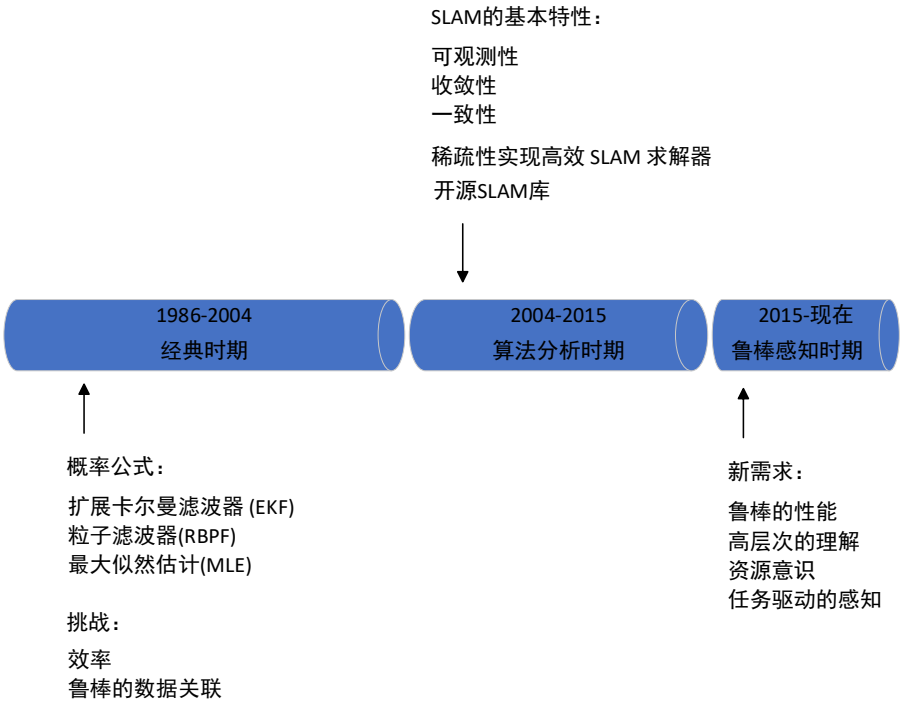

图 3 SLAM 的发展时期

2.2 SLAM 分类

1.视觉 SLAM

传统的 SLAM 主要依赖于昂贵的激光雷达、IMU 等传感器进行环境感知。通过几何特征点的提取和匹配，传统 SLAM 系统能够实现机器人的定位和环境地图构建。然而，这些传感器通常较昂贵，限制了 SLAM 系统的广泛应用。随着技术的发展，视觉 SLAM(Visual SLAM)开始出现，利用相机等视觉传感器进行环境信息的获取，推动了 SLAM 技术的更广泛应用，并催生了多种经典算法。

2015 年，Mur-Artal 等人提出的 ORB-SLAM[8] 是一种基于特征的单目 SLAM 系统，可以在小型和大型、室内和室外环境中实时运行，成为模块化 SLAM 领域的一项重要工作，很多后续出现的基于特征匹配的 SLAM 系统都是由 ORB-SLAM 发展而来。Mur-Artal 等人接下来提出的 ORB-SLAM2[9]，在保持框架整体性的基础上，对一些细节进行了改进，使其能够适用于更多种类的相机，包括深度相机和双目相机，此外，跟踪线程中引入了预处理模块，最后有一个全局 BA(Bundle adjustment) 提高系统的鲁棒性；ORB-SLAM3[10]在此基础之上耦合了惯导 IMU、加入了融合估计以及子地图功能。另一方面，Tong Qin 等人提出的 VINS[11,12]系列是不同于 ORB-SLAM 的又一经典框架，支持多种视觉惯性传感器类型包括 IMU、GPS 等。

2.动态 SLAM

SLAM 框架早期是建立在静态假设成立的基础上，即认为环境中的所有物体都是静态不动的，唯一移动的物体是传感器本身。这种假设导致在存在动态物体的环境中，位姿估计容易变得不准确，甚至在高度动态的环境中可能完全失效。为了解决这一问题，提出了动态场景下的 SLAM，即动态 SLAM(Dynamic SLAM)。动态 SLAM 将环境中的物体分为了动态和静态两类来进行区分。在一些动态 SLAM 中，动态物体被剔除，不纳入位姿估计计算当中。例如，Seungwon Song 等人提出的 DynaVINS[13]同时估计相机姿势并丢弃与运动前验明显偏离的动态对象的特征；Berta Bescos 等人提出的 DynaSLAM[14]通过物体掩码和角度差、深度信息来判断物体的状态，并只使用静态区域且非动态物体掩膜的 ORB 特征点进行相机位姿估计；Daniela Esparza 等人提出的 STDyn-SLAM[15]采用对极几何的方法，通过建立当前帧和上一帧光流, 根据对极几何约束判断是否为动态点，从而将动态点剔除。对于连续两帧匹配上的特征点，计算点与极线的距离，如公式(1)所示。

$$d(X', l') = \frac{X'^{T} F X}{\sqrt{(FX)_1^2 + (FX)_2^2}} \tag{1}$$

其中$d(X', l')$就是计算的距离，$(FX)_1$和$(FX)_2$表示极线元素。由于噪声的存在，静态的特征点有一定的概率不在极线上，所以设定距离阈值，大于阈值的认为是动态点。

另一方面，一些 SLAM 框架采用不同的策略，将动态物体位姿进行估计并纳入到优化中，一同用来估计相机位姿。如，Yuheng Qiu 等人提出的 AirDOS[16]将刚性和运动约束引入模型铰接对象，引入简单而有效的刚性和运动约束一般动态铰接物体。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；Shichao Yang 等人提出的 Cubeslam[17] 联合优化摄像机、物体和点的姿态。物体可以提供远距离的几何和比例约束，以改进摄像机的姿态估计。待优化的最小二乘问题如式(2)所示。

$$C^*, O^*, P^* = \lim_{\{C,O,P\}} \sum_{C_i,O_j,P_k} \| \mathbf{e}(c_i, o_j) \|^2_{\Sigma_{ij}} + \| \mathbf{e}(c_i, p_k) \|^2_{\Sigma_{ik}} + \| \mathbf{e}(o_j, p_k) \|^2_{\Sigma_{jk}} \quad (2)$$

其中 C 表示相机、O 表示物体、P 表示特征点，三个信息共同进行优化。

3.语义 SLAM

传统 SLAM 系统主要依赖于几何信息，在某些场景下可能限制了对环境的深度理解。随着深度学习技术的兴起，V-SLAM 系统得到了显著的改进。深度学习方法被广泛用于图像特征提取、深度图的生成、对抗性训练，可以提高 SLAM 系统的鲁棒性等性能。深度学习为 V-SLAM 系统获取更多的环境语义信息，增强对环境的高层次理解能力，从而更好的感知环境。在 SLAM 系统中加入语义信息，可以形成语义 SLAM(Semantic SLAM)。语义 SLAM 利用深度学习网络对物体进行分割，能更好的识别可能的动态物体，同时构建出包含语义信息的地图，在导航和环境交互等方面有更好的效果。2017 年 Martin R¨unz 等人提出的 Co-Fusion[18]利用 SharpMask[19]将场景分割成不同的对象（使用运动或语义线索），同时跟踪和重建真实的 3D 形状，并随时间推移改进物体在地图上的模型。2018 年 Martin R¨unz 等人进行改进，提出了 Mask-Fusion[20]，使用 MASK-RCNN[21]网络对场景中的不同对象进行识别，并在 SLAM 线程之外添加了一个用于分割的语义线程，以提高系统的实时性。2022 年 Shuhong Cheng 等人提出的 SG-SLAM[22]在 ORB-SLAM2 的基础框架上添加了两个新的并行线程：一个用于获取 2D 语义信息的对象检测线程和一个语义地图线程，然后利用语义信息和几何信息快速剔除动态点，使用静态特征点进行位姿估计，将静态物体构建到语义地图中。

语义 SLAM 的崛起标志着 SLAM 技术在智能感知领域的巨大飞跃。传统 SLAM 侧重于通过传感器获得环境的几何信息，而语义 SLAM 则将其推向一个更为智能化的阶段，引入了对环境中物体语义信息的理解。这不仅使机器能够感知到物体的存在，更能够理解它们的含义和相互关系。在自主导航方面，语义 SLAM 不仅关注机器人在空间中的位置，还注重周围环境中物体的语义理解，使机器能够更智能地避障和与环境进行交互。总体而言，语义 SLAM 为各种机器人和智能系统赋予了更深层次的感知能力，使它们能够更加智能地与真实世界互动。这种技术的进步不仅为科技发展注入新的活力，更为未来智能化社会的建设带来了更为广阔的前景。

参考文献

[1]G. Deng, J. Li, W. Li and H. Wang, "SLAM: Depth image information for mapping and inertial navigation system for localization," 2016 Asia-Pacific Conference on Intelligent Robot Systems (ACIRS), Tokyo, Japan, 2016, pp. 187-191, doi: 10.1109/ACIRS.2016.7556210.

[2] W. Chen et al, "An Overview on Visual SLAM: From Tradition to Semantic," Remote Sensing, vol. 14, (13), pp. 3010, 2022. Available: https://www.proquest.com/scholarly-journals/overview-on-visual-slam-tradition-semantic/docview/2686170995/se-2. DOI: https://doi.org/10.3390/rs14133010.

[3] Xia L, Cui J, Shen R, Xu X, Gao Y, Li X. A survey of image semantics-based visual simultaneous localization and mapping: Application-oriented solutions to autonomous navigation of mobile robots. International Journal of Advanced Robotic Systems. 2020;17(3). doi:10.1177/1729881420919185.

[4]C. Cadena et al., "Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age," in IEEE Transactions on Robotics, vol. 32, no. 6, pp. 1309-1332, Dec. 2016, doi: 10.1109/TRO.2016.2624754.

[5] T. Bailey and H. Durrant-Whyte, "Simultaneous localization and mapping (SLAM): part II," in IEEE Robotics & Automation Magazine, vol. 13, no. 3, pp. 108-117, Sept. 2006, doi: 10.1109/MRA.2006.1678144.

[6] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part I," in IEEE Robotics & Automation Magazine, vol. 13, no. 2, pp. 99-110, June 2006, doi: 10.1109/MRA.2006.1638022.

[7]G. Dissanayake, S. Huang, Z. Wang and R. Ranasinghe, "A review of recent developments in Simultaneous Localization and Mapping," 2011 6th International Conference on Industrial and Information Systems, Kandy, Sri Lanka, 2011, pp. 477-482, doi: 10.1109/ICIINFS.2011.6038117.

[8] R. Mur-Artal, J. M. M. Montiel and J. D. Tardós, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," in IEEE Transactions on Robotics, vol. 31, no. 5, pp. 1147-1163, Oct. 2015, doi: 10.1109/TRO.2015.2463671.

[9] Y. Diao, R. Cen, F. Xue and X. Su, "ORB-SLAM2S: A Fast ORB-SLAM2 System with Sparse Optical Flow Tracking," 2021 13th International Conference on Advanced Computational Intelligence (ICACI), Wanzhou, China, 2021, pp. 160-165, doi: 10.1109/ICACI52617.2021.9435915.

[10] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel and J. D. Tardós, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM," in IEEE Transactions on Robotics, vol. 37, no. 6, pp. 1874-1890, Dec. 2021, doi: 10.1109/TRO.2021.3075644.

[11] T. Qin, P. Li and S. Shen, "VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator," in IEEE Transactions on Robotics, vol. 34, no. 4, pp. 1004-1020, Aug. 2018, doi: 10.1109/TRO.2018.2853729.

[12] Qin, T., Pan, J., Cao, S., and Shen, S., “A General Optimization-based Framework for Local Odometry Estimation with Multiple Sensors”, arXiv e-prints, 2019. doi:10.48550/arXiv.1901.03638.

[13] S. Song, H. Lim, A. J. Lee and H. Myung, "DynaVINS: A Visual-Inertial SLAM for Dynamic

Environments," in IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 11523-11530, Oct. 2022, doi: 10.1109/LRA.2022.3203231.
[14] B. Bescos, J. M. Fácil, J. Civera and J. Neira, "DynaSLAM: Tracking, Mapping, and Inpainting in Dynamic Scenes," in IEEE Robotics and Automation Letters, vol. 3, no. 4, pp. 4076-4083, Oct. 2018, doi: 10.1109/LRA.2018.2860039.
[15] D. Esparza and G. Flores, "The STDyn-SLAM: A Stereo Vision and Semantic Segmentation Approach for VSLAM in Dynamic Outdoor Environments," in IEEE Access, vol. 10, pp. 18201-18209, 2022, doi: 10.1109/ACCESS.2022.3149885.
[16] Qiu, Y., Wang, C., Wang, W., Henein, M., and Scherer, S., “AirDOS: Dynamic SLAM benefits from Articulated Objects”, arXiv e-prints, 2021. doi:10.48550/arXiv.2109.09903.
[17] S. Yang and S. Scherer, "CubeSLAM: Monocular 3-D Object SLAM," in IEEE Transactions on Robotics, vol. 35, no. 4, pp. 925-938, Aug. 2019, doi: 10.1109/TRO.2019.2909168.
[18] M. Rünz and L. Agapito, "Co-fusion: Real-time segmentation, tracking and fusion of multiple objects," 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 2017, pp. 4471-4478, doi: 10.1109/ICRA.2017.7989518.
[19] Lin, TY. et al. (2014). Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham. https://doi.org/10.1007/978-3-319-10602-1_48
[20] M. Runz, M. Buffier and L. Agapito, "MaskFusion: Real-Time Recognition, Tracking and Reconstruction of Multiple Moving Objects," 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Munich, Germany, 2018, pp. 10-20, doi: 10.1109/ISMAR.2018.00024.
[21] He, K., Gkioxari, G., Dollár, P., and Girshick, R., “Mask R-CNN”, arXiv e-prints, 2017. doi:10.48550/arXiv.1703.06870.
[22] S. Cheng, C. Sun, S. Zhang and D. Zhang, "SG-SLAM: A Real-Time RGB-D Visual SLAM Toward Dynamic Scenes With Semantic and Geometric Information," in IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1-12, 2023, Art no. 7501012, doi: 10.1109/TIM.2022.3228006.