智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

视觉SLAM研究进展

王霞， 左一凡

引用本文:
王霞, 左一凡. 视觉SLAM研究进展[J]. 智能系统学报, 2020, 15(5): 825–834.
WANG Xia, ZUO Yifan. Advances in visual SLAM research[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(5): 825–834.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202004023

DOI: 10.11992/tis.202004023
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20200904.1708.002.html

视觉 SLAM 研究进展

王霞，左一凡
（北京理工大学 光电成像技术与系统教育部重点实验室，北京 100081）

摘　要：视觉 SLAM 是指相机作为传感器进行自身定位同步创建环境地图。SLAM 在机器人、无人机和无人车导航中具有重要作用，定位精度会影响避障精度，地图构建质量直接影响后续路径规划等算法的性能，是智能移动体应用的核心算法。本文介绍主流的视觉 SLAM 系统架构，包括几种最常见的视觉传感器，以及前端的功能和基于优化的后端。并根据视觉 SLAM 系统的度量地图的种类不同将视觉 SLAM 分为稀疏视觉 SLAM、半稠密视觉 SLAM 和稠密视觉 SLAM 3 种，分别介绍其标志性成果和研究进展，提出视觉 SLAM 目前存在的问题以及未来可能的发展。

关键词：视觉同步定位与创建地图；稀疏视觉 SLAM；半稠密视觉 SLAM；稠密视觉 SLAM；视觉传感器；优化；视觉 SLAM 系统；度量地图
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1673−4785(2020)05−0825−10

中文引用格式：王霞，左一凡. 视觉 SLAM 研究进展 [J]. 智能系统学报, 2020, 15(5): 825–834.
英文引用格式: WANG Xia, ZUO Yifan. Advances in visual SLAM research[J]. CAAI transactions on intelligent systems, 2020, 15(5): 825–834.

Advances in visual SLAM research

WANG Xia, ZUO Yifan
(Key Laboratory of Photo-electronic Imaging Technology and System, Ministry of Education of China, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Visual SLAM, i.e., simultaneous localization and mapping with cameras, plays an important role in the navigation of robots, unmanned aerial vehicles, and unmanned vehicles. As the location accuracy affects the obstacle avoidance accuracy and the mapping quality directly affects the path planning performance, the visual SLAM algorithm is the core aspect of intelligent mobile applications. This paper introduces the architecture of the mainstream visual SLAM system, including several common visual sensors, the function of the front end, and the optimized back end. According to the type of the metric map model created by the visual SLAM system, visual SLAM can be classified into three types: sparse visual SLAM, semi-dense visual SLAM, and dense visual SLAM. The landmark achievements and research progress of visual SLAM are reviewed in this paper, and its current problems and possible future developments are discussed.

Keywords: visual simultaneous localization and mapping; sparse visual SLAM; SemiDense visual SLAM; dense visual SLAM; visual sensors; optimization; visual SLAM system; metric map

同步地图构建和定位 (simultaneous localization and mapping, SLAM) 包含定位和建图两方面，是移动机器人领域的一个重要的开放问题：移动机器人要想精确移动，必须有精确的环境地图；然而，为了建立一个精确的地图，必须知道移动机器人精确的位置[1]，所以，这是一个相辅相成的过程。1990 年，有学者首次提出利用拓展卡尔曼滤波器对机器人姿态的后验分布进行增量估计[2]。事实上，在未知的位置、未知的环境中，机器人通过在运动过程中反复观察环境特征确定自身位

收稿日期：2020−04−23.　**网络出版日期**：2020−09−08.
基金项目：装备预先研究项目 (41417070401).
通信作者：左一凡. E-mail: zuoyifan_bit@outlook.com.

置，然后根据自己的位置构建一个增量的周边环境地图，从而达到同时定位和地图构建的目的。随着中央处理器 (central processing unit, CPU) 和图形处理器 (graphic processing unit, GPU) 的发展，图形处理能力越来越强大。相机传感器变得更价廉、更轻便，同时具有更多功能。在过去的十几年中，视觉 SLAM 发展迅速。该系统可在微 PC 和嵌入式设备上运行，甚至可在智能手机等移动设备上运行[3-7]。视觉 SLAM 可使用单目相机、立体相机等视觉传感器进行数据采集、前端视觉里程计、后端优化、回环检测和地图构建等[8]。有些 SLAM 系统包含重定位模块，作为更稳定和准确的视觉 SLAM 附加模块[9]。本文主要介绍几种重要的视觉传感器，并根据数据的稀疏程度分类综述视觉 SLAM 的现阶段成果。

1　视觉 SLAM 系统

视觉 SLAM(visual SLAM, VSLAM) 系统的架构包括两个主要部分:前端和后端。前端抓取传感器数据，并进行状态估计，后端对前端产生的数据进行优化。后端可为前端提供反馈，并进行回环检测[10]。该体系结构如图 1 所示。

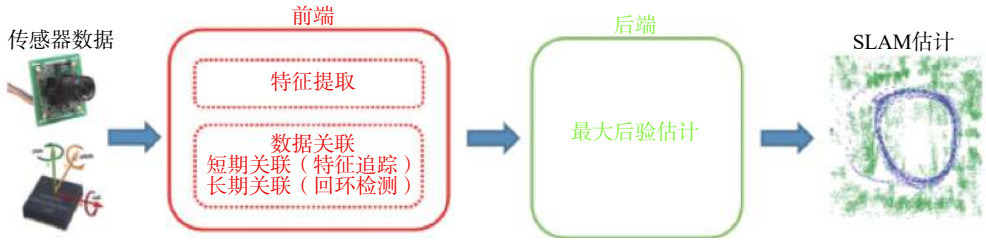

图 1　典型的 SLAM 系统图

Fig. 1　A typical SLAM system

1.1　视觉传感器

大多数视觉传感器是基于相机，相机可以分为单目相机、立体相机、RGB-D 相机和事件相机等[11]。

1.1.1　单目相机

单目相机定位和建图具有和现实世界的真实比例关系，但没有真实的深度信息和绝对尺度，这称为尺度模糊[12]。基于单目相机的 SLAM 必须进行初始化来确定尺度，而且面临漂移问题，但单目相机价格低廉、计算速度快，在 SLAM 领域受到广泛应用。

1.1.2　立体相机

立体相机是两个单目相机的组合，其中两个相机之间的距离是已知的，称之为基线。使用立体相机，可以通过定标、较正、匹配和计算 4 个步骤获取深度信息，进而确定尺度信息，但这个过程会消耗很大的计算资源。

1.1.3　RGB-D 相机

RGB-D 相机也称为深度相机，因为这种相机可以直接以像素形式输出深度信息。深度相机可以通过立体视觉、结构光和飞行时间 (time of flight, TOF) 技术来实现。结构光理论是指红外激光向物体表面发射具有结构特征的图案，红外相机收集不同深度的表面图案的变化信息。TOF 通过测量激光飞行的时间计算距离。

1.1.4　事件相机

事件相机不是以固定的速率捕获即时消息，而是异步地测量每个像素的亮度变化[13]。事件相机具有非常高的动态范围、高时间分辨率、低功耗，并且不会出现运动模糊。因此，事件相机在高速、高动态范围的情况下性能优于传统相机。事件相机包含动态视觉传感器[14-17]、动线传感器[18]、动态和主动像素视觉传感器[19] 和异步基于时间的图像传感器[20]。

以上 4 种视觉传感器各有其优缺点，如表 1 所示。

表 1　4 种视觉传感器的优缺点

Table 1　Advantages and disadvantages of 4 kinds of visual sensor

传感器类型	优点	缺点
单目相机	成本低、结构简单、速度快	没有深度信息和尺度信息
立体相机	可通过基线估计深度	计算量巨大
RGB-D相机	可估计像素级深度信息	测量范围窄、噪声大
事件相机	高动态范围、高时间分辨率、低延时、低功耗	噪声大、特征点难以提取

1.2　VSLAM 前端

在实际的机器人应用中，可能很难将传感器的测量值直接写成传感器状态量的解析函数。例如，原始传感器数据是一个图像，那么可能很难

将每个像素的强度表示为 SLAM 状态的函数;这是由于无法设计一个足够普遍、但又易于处理的函数来表示环境与传感器状态的关系;即使存在这样一种普遍的表示, 也很难写出一个将测量值与传感器状态联系起来的解析函数。因此, 在数据进入 SLAM 后端之前, 通常需要一个前端模块提取传感器原始图像的相关特征。例如, 在 VSLAM 中, 前端提取特征点位置, 后端可根据这些特征点的位置进行优化处理。同时, 前端模块负责初始化, 例如单目 SLAM 中的初始化, 利用多角度观测图像三角化将尺度信息固定。

前端数据关联模块包括短期数据关联和长期数据关联, 短期数据关联负责联系传感器数据的帧间特征点及追踪特征点, 常用的方法有特征匹配和光流法等。长期关联是关联新的信息是否和过去的所有信息有关联, 即回环检测, 常用的方法有词袋法和深度学习[10]。

1.3　基于优化的 SLAM 后端

早期 SLAM 后端主要是基于滤波的方法, 但由于优化方法的精度明显优于滤波的方法而逐渐成为主流。文献 [21-22] 综述了滤波的方法, 本文主要介绍基于优化的 SLAM 后端。

在 SLAM 问题中, 需要估计的未知变量 X 包括机器人的位姿和路标点的物理坐标。给定观测数据 $Z=\{z_k:k=1,2,\cdots,m\}$,观测方程可表示为未知变量 X 的函数。例如 $z_k=h_k(X_k)+\varepsilon_k$, 其中 $X_k\subseteq X$, $h_k(\cdot)$ 是已知的测量或观测函数, ε_k 是随机测量误差。根据贝叶斯理论, 通过 X 的测量值 X^*估计最大后验概率 $p(X|Z)$:

$$X^*=\arg\max_X p(X|Z)=\arg\max_X p(Z|X)p(X) \tag{1}$$

式中: $p(Z|X)$ 是给定 X 的观测量 Z 的似然; $p(X)$ 称为 X 的先验概率。先验概率包含 X 所有的先验信息; 在没有先验信息的情况下, 先验概率为常量, 在优化中不起作用, 最大后验估计简化为似然估计。不同于卡尔曼滤波的方法, 最大后验概率估计不需要区分运动模型和观测模型, 这两个模型都被视为因子整合到估计的过程中。

假设测量值 Z 是独立的, 即噪声不相关, 将式 (1) 因式分解为

$$X^*=\arg\max_X p(X)\prod_{k=1}^m p(z_k|X)=\arg\max_X p(X)\prod_{k=1}^m p(z_k|X_k) \tag{2}$$

式 (2) 可以用因子图来解释, 因子图是一种图模型, 可以建立第 k 个因子与相应变量 X_k 的依赖关系。$p(z_k|X_k)$ 项和 $p(X)$ 项都称之为因子, 通过节点建立约束。因子图可直观表示约束, 如图 2 展示了一个用因子图表示的简单的 SLAM 问题。其中蓝色代表机器人位姿, 绿色代表地标点坐标, 红色为相机标定参数, 黑色方块是因子, 代表变量间的约束。通过因子图的方式可以表示复杂的多优化变量模型。

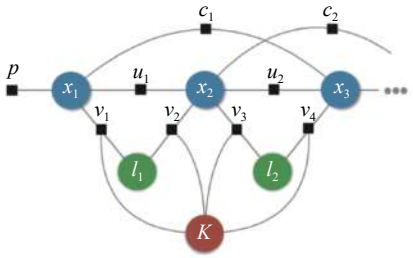

图 2　因子图表示的 SLAM
Fig. 2　SLAM as a factor graph

假设测量噪声 ε_k 是零均值的高斯噪声, 其信息矩阵表示为 Ω_k, 则式 (2) 中的似然可表示为

$$p(z_k|X_k)\propto\exp\left(-\frac{1}{2}\|h_k(X_k)-z_k\|^2_{\Omega_k}\right) \tag{3}$$

其中, $\|\boldsymbol{e}\|^2_\Omega=\boldsymbol{e}^{\mathrm{T}}\Omega\boldsymbol{e}$, 同理, 假设先验可以写为 $p(X)\propto\exp\left(-\frac{1}{2}\|h_0(X)-z_0\|^2_{\Omega_k}\right)$, $h_0(\cdot)$ 为给定的函数, z_0 为先验均值, Ω_0 为信息矩阵。因为最大后验相当于最小后验的负对数, 所以最大后验估计可写为

$$X^*=\arg\min_X -\log(p(X)\prod_{k=1}^m p(z_k|X_k))=\arg\min_X\sum_{k=0}^m\|h_k(X_k)-z_k\|^2_{\Omega_k} \tag{4}$$

式 (4) 是一个最小二乘问题, 在 SLAM 问题中, $h_k(\cdot)$ 是一个非线性函数。求解相机位姿即求解此最小二乘问题。通常使用 Gauss-Newton 法或者 Levenberg-Marquardt 法求解, 得到优化变量 X。

2　度量地图模型的 VSLAM 系统

度量地图是对环境结构的一种表示方法。选择合适的 SLAM 度量表示方法十分重要, 会影响很多研究领域, 例如长时间导航、环境交互和人机交互等领域。根据利用图像信息的方法不同, 可分为直接法和特征点法, 直接法会产生半稠密和稠密的结构, 特征点法会产生稀疏的结构。本文根据 SLAM 系统产生的不同稀疏程度结构的特点, 将 VSLAM 分为稀疏 VSLAM、半稠密 VSLAM 和稠密 VSLAM。

2.1　稀疏 VSLAM

稀疏 VSLAM 的前端算法以特征点匹配为主，光流追踪以及直接法等方法也在不断发展，但特征匹配仍为稀疏 VSLAM 的主流前端算法。后端算法主要分为基于滤波的算法和基于优化的算法，早期由于算力的限制，主要以基于滤波的后端算法为主，随着 CPU 和 GPU 的发展，基于优化的后端由于其具有更好的精度而逐渐成为主流。

MonoSLAM 是在 2007 年提出的一种可以通过单目相机实现实时场景三维重建的算法[23]，该算法首次实现单目 SLAM 系统，可实现实时且无漂移的运动恢复结构，后端使用拓展的卡尔曼滤波算法，前端使用稀疏的特征点匹配，实现在线稀疏地图的持续构建。虽然 MonoSLAM 应用场景很窄，特征点也容易丢失，但作为第一个 SLAM 系统，具有里程碑意义。2007 年，一种专门为小型 AR 工作空间中手持摄像机设计的系统——PTAM 系统出现[24]，此系统首次将特征点追踪和地图构建分为两个独立的任务，使用并行处理的方式，并首次使用重投影误差进行后端优化，因此 PTAM 的出现对于 SLAM 的发展具有重要意义。同样，PTAM 也存在场景小、跟踪容易丢失等特点。

2009 年，Klein 等[5] 提出关键帧的概念和重定位的方法。同年，此团队又提出应用于照相手机的基于关键帧的 SLAM 系统[25]。2015 年，建立了 ORB-SLAM[3]，一种基于 ORB 特征匹配[26] 的单目实时 SLAM 系统，此系统在稀疏 VSLAM 领域具有里程碑意义，系统十分完善，可应用于多种场景，对于运动杂波具有较强的鲁棒性。具有追踪、建图、重定位和回环检测功能，其标志性地使用 3 个线程 (如图 3 所示)，分别为特征点追踪线程、局部重投影误差优化线程和基于位姿图的全局优化线程。对于选择重建点和关键帧具有良好鲁棒性，可生成增量地图，这使得基于特征点的 SLAM 成为当时的主流。目前此项目源代码已开源。2017 年的 ORB-SLAM2[27] 支持单目相机、立体相机和 RGB-D 相机的 SLAM 系统，可在 CPU 上实时工作。

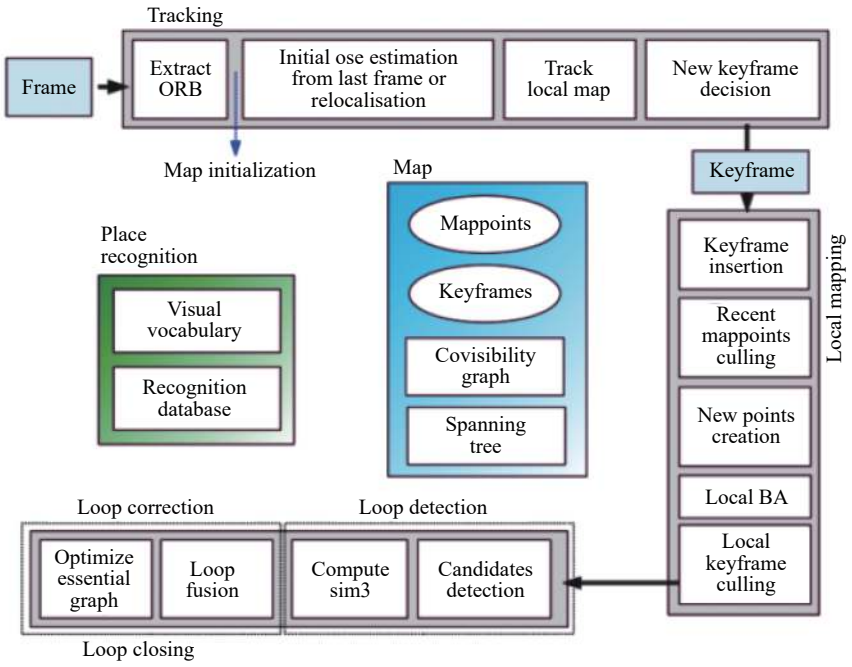

图 3　ORB-SLAM 的三线程结构

Fig. 3　Three-thread structure of ORB-SLAM

与其他基于特征点提取的稀疏 VSLAM 不同，Forster 等[28] 在 2016 年提出了一种半直接 VO (semi-direct visual odometry, SVO)，是一种直接法和特征点混合的方法，它使用直接的方法跟踪和三角化像素，这些像素具有较高的图像梯度，但依赖于基于特征方法的联合优化。半直接 VO 加上鲁棒的概率深度估计算法，能够有效地跟踪像素的角点和边。该算法可以很容易地扩展到多个相机跟踪，包括运动先验，并可适用于大视场相机，如鱼眼和反折射相机。相对于其他 VSLAM，SVO 的优点是速度快、计算要求低。但只适用于平面运动，而且没有后端优化和回环检测，不是

完整的 SLAM 系统。2018 年, Loo 等[29] 提出利用神经网络预测单目图像深度的 SVO 版本, 可根据单图像深度来预测网络的深度预测结果, 通过初始化特征点处深度的均值和方差改进 SVO 建图。

此外, 还有大量优秀的稀疏 VSLAM 系统。2016 年, Zhang 等[30] 提出 ENFT-sfm 系统, 它是一种特征跟踪方法, 能够有效地匹配一个或多个视频序列之间的特征点对应。升级版的 ENFT-SLAM 可大规模运行。DSO 是 2017 年 Engel 等[31] 提出的基于单目相机在不需要检测和描述特征点的情况下, 采用直接法和稀疏法建立的一个可视化导航系统。2018 年, Schlegel 等[32] 提出一种简单的轻量级立体 VSLAM, 此方法重点在数据结构和算法提升方面, 可达到目前最优秀算法级别的准确性, 同时大大减少计算资源。OpenVSLAM 是 Sumikura 等[33]2019 年提出的一个具有高可用性和可扩展性的可视化 SLAM 框架, 基于具有稀疏特征的间接 SLAM 算法。传统的开源 VSLAM 框架并不适合作为第三方程序调用, 此框架易于扩展和使用, 该系统支持透视、鱼眼等相机, 甚至支持自己设计的相机模型。2019 年, 通过使用 AprilTag 基准标记实现 SLAM 的 TagSLAM 系统出现[34]。该系统提供了一个前端的 GT-SAM 因子图优化器, 此优化器可设计大量的实验,包括完整 SLAM 系统、相机标定、视觉定位、回环检测以及位姿估计等。UcoSLAM 是 2019 年提出的一种融合自然地标和人工地标的同步定位方法[35]。多数 SLAM 方法使用自然地标 (如关键点)。但是, 自然地标随着时间的推移是不稳定的, 在许多情况下是重复的, 或者不足以进行鲁棒的跟踪 (例如在室内建筑物中)。另一方面, 基于人工地标的其他方法 (例如平方基准标记) 可通过放置在环境中的地标帮助跟踪和重新定位。UcoSLAM 提出了一种将这两种方法相结合的方法, 以实现在许多场景下的长期鲁棒跟踪, 且具有更好的准确性。

综上所述, 稀疏 VSLAM 由于其计算量小、速度快, 一度成为 VSLAM 的主流方法, 但稀疏 VSLAM 无法构建稠密地图, 对于路径规划以及场景理解等高层任务无法很好地实现。

2.2　半稠密 VSLAM

由于特征点法只能产生稀疏的结构, 半稠密 VSLAM 主要以直接法和半直接法为主, 直接法不需要提取特征点, 直接根据像素变化估计相机运动, 因此计算量远高于特征点法, 其起步也晚于基于特征点法的 VSLAM。本文介绍几种经典的半稠密 VSLAM 系统。

LSD-SLAM 是 2014 年 Engel 等[36] 提出的一种直接 (无特征) 单目 SLAM 算法, 与目前最先进的直接方法相比, 可构建大规模、一致的环境地图, 标志着半稠密 VSLAM 的成功应用, 其运行结果如图 4 所示。2015 年立体相机直接 SLAM 算法在标准 CPU 上以高帧速率实时运行[37], 此前很少见到使用 CPU 实时建立半稠密地图的算法。此方法的创新性在于直接基于所有高对比度像素 (包括角、边和高纹理区域) 的亮度一致性对图像进行对齐, 同时, 通过固定基线的立体摄像机设置的静态立体以及利用摄像机运动的临时多视点立体估计这些像素的深度。2015 年, Caruso 等[38] 将鱼眼相机引入直接单目 SLAM 方法, 使直接法 SLAM 支持广角相机。但 LSD-SLAM 同时有着直接法 SLAM 的缺点, 对相机内参和曝光敏感, 相机快速运动时容易丢失。

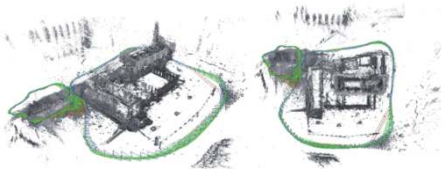

图 4　LSD-SLAM 运行图

Fig. 4　LSD-SLAM running figure

基于事件相机的 SLAM 系统的早期代表为 2013 年 Weikersdorfer 等[39] 提出的一种对单个像素事件进行操作的算法, 可生成具有精确机器人定位的高质量 2D 环境地图。2014 年又提出了一种基于事件的动态视觉传感器与一个基于帧的 RGB-D 传感器融合的 SLAM 系统, 以产生一个深度增强的 3D 点图[40]。EVO 是 2016 年提出的一种基于事件的视觉里程计算法[41], 此算法成功地利用了事件相机的特性来跟踪快速的相机运动, 同时恢复了半密集的 3D 环境地图。由于事件相机的性质, 算法不受运动模糊的影响, 并在具有挑战性的高动态范围条件下运行良好, 光照变化强烈。2018 年, Zhou 等[42] 提出了一种基于立体事件相机的 SLAM 系统, 可进行半稠密的三维重建。

2.3　稠密 VSLAM

稠密 VSLAM 由于可构建三维稠密地图并应用于路径规划中, 使其具有前者不具备的优势, 在近些年得到了广泛关注。与半稠密 VSLAM 类似, 稠密 VSLAM 也是以直接法和半直接法为主。

DTAM 是 2011 年 Newcombe 等[43] 提出的一种不依赖于特征提取而是依赖于稠密的逐像素方法, 使用 RGB-D 相机和一种非凸优化框架中最小

化全局空间正则化的能量泛函实现实时地追踪与重建系统。这是直接法 SLAM 系统的典型例子，具有里程碑意义。

Newcombe 等[44] 在 2011 年提出了 Kinect Fusion，一个使用 Kinect 传感器的实时建图和追踪系统，通过 ICP 算法跟踪深度相机的数据，并构建稠密的地图。同年，Izadi 等[45] 也提出了一个使用 Kinect 传感器进行三维重建的系统，这是第一个基于深度相机的三维重建系统。2012 年，Whelan 等[46] 提出了 KinectFusion 算法的一个扩展算法——Kintinuous，它允许扩展尺度环境的实时稠密建图。相比于 KinectFusion，此算法的区域空间可以动态变化，通过三角网格代替点云创建地图，该系统实现了一组能够实时操作的分层多线程组件，地图绘制能力大大超出了原始 KinectFusion 算法的范围，其运行结果图如图 5 所示。2013 年，Whelan 等[47] 又对 Kintinuous 算法进行了扩展，提出 3 点补充：1) 融合多种 6 自由度相机进行稳健跟踪；2) 实现基于 GPU 的新型稠密 RGB-D 视觉里程计算法；3) 采用先进的融合实时表面着色技术。这些扩展可为机器人和虚拟现实应用提升构建密集的全彩色空间扩展环境模型的能力，同时在具有挑战性的几何和视觉特征的场景中保持鲁棒。2015 年，Whelan 等[48] 又提出了一种新的 SLAM 系统，能够在数百米范围内实时生成高质量的全局一致性地表重建，且只需要一个低成本的商品级 RGB-D 传感器，实现了比使用原始 RGB-D 点云更高质量的地图。此方案创新性地使用一个基于 GPU 的 3D 循环缓冲区技巧，高效地扩展稠密图融合方法，并克服了相机位姿估计在各种环境中的局限性。

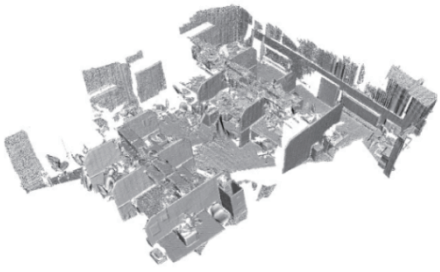

图 5　Kintinuous 运行结果图
Fig. 5　Kintinuous running figure

Labbé等[49] 在 2011 年提出的 RTAB-MAP 系统，是利用 RGB-D 传感器的经典 SLAM 系统，可利用 RGB-D 相机进行同步定位和局部建图、可在大规模和长时间 SLAM 中实时回环检测的方法，克服了回环检测随着时间推移影响实时处理的不足。两年后此团队又提出了基于外观的回环检测方法，该方法基于一种内存管理方法，它限制了用于回环检测的位置数量，从而使计算时间保持在实时约束下。包括将最近观察到的位置保存在工作内存 (WM) 中，用于回环检测，并将其他位置转移到长期内存 (LTM) 中。当在当前位置和存储在 WM 中的位置之间找到匹配时，可以更新并记住存储在 LTM 中的相关位置，以便进行回环检测[50]。1 年后，该团队又将 SLAM 系统与全局回环检测结合，解决多机器人初始值定位问题[51]。2019 年，该团队继续发展了 RTSB-MAP 系统，使其能同时支持视觉和激光雷达 SLAM[52]。

除上述系统外，稠密 VSLAM 系统还包括基于光度和深度误差的 DVO[53-54]，能够实时重建非刚性变形场景的密集 SLAM 系统的 DynamicFusion、VolumeDeform 和 Fusion4D[55-57]。实现在线增量地图的 ElasticFusion[58-59]。基于 CPU 的体素表示的三维重建系统 InfiniTAM v3 系统和统一的框架，即 InfiniTAM[6, 60-61]。此外，2014 年，ENDRES 等[62] 提出一种新型的建图系统 RGBD-SLAM-V2，只使用 RGB-D 传感器生成高精度的 3D 地图。2016 年，Greene 等[63] 提出 MLM SLAM 系统，一种基于单目相机、无需 GPU 即可在线重建稠密的三维模型，从而解决了多分辨率深度估计和空间平滑处理问题。

2.4　发展趋势

当前 SLAM 地图表达主要使用点云图和截断符号距离函数 (truncated signed-distance function, TSDF) 进行三维建模，这些表示方法有两个主要缺陷。第一，浪费大量内存，点和体素这两种表示都需要大量参数去编码一个简单的环境。第二，这两种表示都不能完整表达环境信息。比如机器人无法确定在房间内移动还是走廊中移动。因此，通过增加语义信息数据关联、人机交互等方式为 SLAM 提供更强有力的数据支撑是地图表示的趋势。

现在已有大量关于 SLAM 建图的研究，但是很少有人研究可以指导研究人员进行地图选择的标准和地图的评价指标。例如简单的室内环境、简单的参数变化可以满足三维环境表达，但网格表示对复杂的室外环境更适用。因此，制定一套评价不同地图表示以及指导研究者选择地图的标准也是亟待解决的问题。

设计环境表示方法是一个困难的问题，而且设计的表示方法往往不够灵活，缺乏适应性，如何让机器人根据所处环境的变化自动设计地图表示形式也是重要的发展方向，尤其对于长时间导

航有着巨大的促进作用。

3　机遇和挑战

SLAM 的概念早在 1986 年由 Smith 等[64] 提出, 由于没有发现海森矩阵的稀疏性导致长期未被实际应用, 经历了几十年的发展, VSLAM 已经被广泛应用于机器人、无人机、无人车和增强现实等领域, 但 SLAM 对环境光照、高速运动、运动干扰等问题较为敏感, 如何提升系统的鲁棒性以及长时间构建大规模地图等问题都是值得挑战的领域。在 SLAM 主要应用的两大场景是基于智能手机或无人机等嵌入式平台和 3D 重建、场景理解和深度学习。如何平衡实时性和准确性是一个重要的开放性问题。针对动态、非结构化、复杂、不确定和大规模等诸多环境的解决方案仍有待探索[65]。此外, VSLAM 和语义信息以及与其他类型传感器的结合也给 SLAM 带来了新的机遇和挑战。

3.1　鲁棒性

VSLAM 仍然面临着光照条件、高动态环境、快速运动、强烈旋转和低纹理环境等问题。首先, 利用新型传感器可以解决高动态和光照条件等问题。例如, 动态视觉传感器这样的事件相机每秒可产生 100 万个事件, 这对于高速、高动态范围内的快速运动已经足够。其次, 结合语义特征, 如边缘、平面、表面特征可减少对特征的依赖, 可以解决低纹理环境等问题。语义 SLAM 也是一个重要的研究方向。第三, 每一次运动信息都会减少一次定位不确定性, 但同时也增大一次计算量, 如何平衡精度与计算量之间的关系, 进行大尺度地图构建仍是一个重要问题。

3.2　多传感器融合

实际的机器人和硬件设备通常携带不止一种传感器, 往往是多个传感器的融合。新的传感器的诞生往往是 SLAM 的一大驱动因素。例如, 将视觉信息与 IMU 信息相结合, 实现了两个传感器的互补优势, 为 SLAM 的小型化和低成本提供了非常有效的解决方案。事件相机有可能解决高动态环境等问题。目前的传感器包括激光雷达、声纳、IMU、红外、相机、GPS、雷达等。传感器的选择取决于环境和所需的地图类型。

3.3　基于深度学习的 SLAM

深度学习在机器视觉的众多领域获得了成功, 也不断有学者将深度学习引入 SLAM 的各个模块, 例如回环检测、特征识别等, 甚至深度学习理论上可以代替整个 SLAM 系统, 但基于精确的数学公式推导出的导航函数仍优于学习得到的导航函数。所以基于最大后验概率的后端 SLAM 依旧是目前的主流。事实上, 人类识别物体的运动是基于感知, 而不是图像的特征。SLAM 中的深度学习可实现目标识别和分割, 帮助 SLAM 系统更好地感知周围环境。语义 SLAM 还可以在全局优化、循环关闭和重定位等方面发挥作用。传统的 SLAM 依赖于点、线 (PL-SLAM[66]、Struct-SLAM[67])、面等几何特征来推断环境结构。在大规模场景中, 高精度实时定位的目标可以通过语义 SLAM 来实现[68]。

4　结束语

本文介绍了常见的 VSLAM 系统并根据数据的稀疏性将 VSLAM 分为 3 类, 分别介绍了 3 种类型的 VSLAM 发展历程, 最后提出了 VSLAM 的机遇与挑战。从近些年的发展来看, VSLAM 在向鲁棒性、实时性更强的方向发展, 也有越来越多的新型技术不断涌现, 使得 VSLAM 已经应用于实际生活中, 尤其是视觉惯导融合领域, 已在无人车和手持设备中实现应用, 但对于许多应用环境, 许多重大的挑战仍亟需解决, 如实现长时间鲁棒的感知和导航、挑战光环境下导航等。随着新的系统、新型传感器和新的计算工具等的开发, 相信未来 VSLAM 技术一定会在导航定位领域发挥重要作用。

参考文献:

[1] LEONARD J J, DURRANT-WHYTE H F. Simultaneous map building and localization for an autonomous mobile robot[C]//Proceedings IROS'91: IEEE/RSJ International Workshop on Intelligent Robots and Systems' 91. Osaka, Japan, 1991: 1442−1447.

[2] SMITH R, SELF M, CHEESEMAN P. Estimating uncertain spatial relationships in robotics[M]//COX I J, WILFONG G Y. Autonomous Robot Vehicles. New York, USA: Springer, 1990: 167−193.

[3] 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.

[4] QIN Tong, LI Peiliang, SHEN Shaojie. VINS-MONO: a robust and versatile monocular visual-inertial state estimator[J]. IEEE transactions on robotics, 2018, 34(4): 1004–1020.

[5] KLEIN G, MURRAY D. Parallel tracking and mapping on a camera phone[C]//Proceedings of the 2009 8th IEEE In-

ternational Symposium on Mixed and Augmented Reality. Orlando, USA, 2009: 83−86.

[6] KÄHLER O, PRISACARIU V A, REN C Y, et al. Very high frame rate volumetric integration of depth images on mobile devices[J]. IEEE transactions on visualization and computer graphics, 2015, 21(11): 1241–1250.

[7] LYNEN S, SATTLER T, BOSSE M, et al. Get out of my lab: large-scale, real-time visual-inertial localization[C]// Proceedings of Robotics: Science and Systems. Rome, Italy, 2015.

[8] 高翔, 张涛, 刘毅, 等. 视觉 SLAM 十四讲 [M]. 北京: 电子工业出版社, 2017: 13−19.

[9] TAKETOMI T, UCHIYAMA H, IKEDA S. Visual slam algorithms: a survey from 2010 to 2016[J]. IPSJ transactions on computer vision and applications, 2017, 9(1): 16.

[10] CADENA C, CARLONE L, CARRILLO H, et al. Past, present, and future of simultaneous localization and mapping: toward the robust-perception age[J]. IEEE transactions on robotics, 2016, 32(6): 1309–1332.

[11] HUANG Baichuan, ZHAO Jun, LIU Jingbin. A survey of simultaneous localization and mapping with an envision in 6G wireless networks[EB/OL]. (2020-02-14)[2020-03-20]. https://arxiv.org/pdf/1909.05214.pdf.

[12] 刘浩敏, 章国锋, 鲍虎军. 基于单目视觉的同时定位与地图构建方法综述 [J]. 计算机辅助设计与图形学学报, 2016, 28(6): 855–868.
LIU Haomin, ZHANG Guofeng, BAO Hujun. A survey of monocular simultaneous localization and mapping[J]. Journal of computer-aided design & computer graphics, 2016, 28(6): 855–868.

[13] GALLEGO G, DELBRUCK T, ORCHARD G, et al. Event-based vision: a survey[J]. arXiv: 1904.08405, 2019.

[14] LICHTSTEINER P, POSCH C, DELBRUCK T. A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor[J]. IEEE journal of solid-state circuits, 2008, 43(2): 566–576.

[15] SON B, SUH Y, KIM S, et al. 4.1 A 640×480 dynamic vision sensor with a 9μm pixel and 300meps address-event representation[C]//Proceedings of 2017 IEEE International Solid-State Circuits Conference. San Francisco, USA, 2017: 66−67.

[16] POSCH C, MATOLIN D, WOHLGENANNT R, et al. A microbolometer asynchronous dynamic vision sensor for LWIR[J]. IEEE sensors journal, 2009, 9(6): 654–664.

[17] HOFSTÄTTER M, SCHÖN P, POSCH C. A SPARC-compatible general purpose address-event processor with 20-bit l0ns-resolution asynchronous sensor data interface in 0.18 μm CMOS[C]//Proceedings of 2010 IEEE International Symposium on Circuits and Systems. Paris, France, 2010: 4229−4232.

[18] POSCH C, HOFSTATTER M, MATOLIN D, et al. A dual-line optical transient sensor with on-chip precision time-stamp generation[C]//Proceedings of 2007 IEEE International Solid-State Circuits Conference. Digest of Technical Papers. San Francisco, USA, 2007: 500−618.

[19] BRANDLI C, BERNER R, YANG Minhao, et al. A 240×180 130 dB 3 μs latency global shutter spatiotemporal vision sensor[J]. IEEE journal of solid-state circuits, 2014, 49(10): 2333–2341.

[20] POSCH C, MATOLIN D, WOHLGENANNT R. A QVGA 143 dB dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS[J]. IEEE journal of solid-state circuits, 2011, 46(1): 259–275.

[21] BAILEY T, DURRANT-WHYTE H. Simultaneous Localization and Mapping (SLAM): Part II[J]. IEEE robotics & automation magazine, 2006, 13(3): 108–117.

[22] DURRANT-WHYTE H, BAILEY T. Simultaneous localization and mapping: Part I[J]. IEEE robotics & automation magazine, 2006, 13(2): 99–110.

[23] 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.

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

[25] KLEIN G, MURRAY D. Improving the agility of keyframe-based SLAM[C]//Proceedings of the 10th European Conference on Computer Vision. Marseille, France, 2008: 802−815.

[26] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: an efficient alternative to SIFT or SURF[C]//Proceedings of 2011 International Conference on Computer Vision. Barcelona, Spain, 2011: 2564−2571.

[27] 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.

[28] FORSTER C, ZHANG Zichao, GASSNER M, et al. SVO: semidirect visual odometry for monocular and multicamera systems[J]. IEEE transactions on robotics, 2017, 33(2): 249–265.

[29] LOO S Y, AMIRI A J, MASHOHOR S, et al. CNN-SVO: improving the mapping in semi-direct visual odometry using single-image depth prediction[EB/OL]. (2018-10-01)[2020-02-03]. https://arxiv.org/abs/1810.01011.

[30] ZHANG Guofeng, LIU Haomin, DONG Zilong, et al. Efficient non-consecutive feature tracking for robust structure-from-motion[J]. IEEE transactions on image processing, 2016, 25(12): 5957–5970.

[31] ENGEL J, KOLTUN V, CREMERS D. Direct sparse odometry[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 40(3): 611–625.

[32] SCHLEGEL D, COLOSI M, GRISETTI G. ProSLAM: graph SLAM from a programmer's perspective[EB/OL]. (2017-09-13)[2020-02-04]. https://arxiv.org/abs/1709.04377.

[33] SUMIKURA S, SHIBUYA M, SAKURADA K. Openvslam: a versatile visual slam framework[C]//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France, 2019.

[34] PFROMMER B, DANIILIDIS K. TagSLAM: robust slam with fiducial markers[EB/OL]. (2019-10-01)[2020-02-05]. https://arxiv.org/abs/1910.00679.

[35] MUÑOZ-SALINAS R, MEDINA-CARNICER R. UcoSLAM: simultaneous localization and mapping by fusion of keypoints and squared planar markers[J]. Pattern recognition, 2020, 101: 107193.

[36] ENGEL J, SCHÖPS T, CREMERS D. LSD-SLAM: large-scale direct monocular SLAM[C]//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland, 2014.

[37] ENGEL J, STÜCKLER J, CREMERS D. Large-scale direct SLAM with stereo cameras[C]//Proceedings of 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems. Hamburg, Germany, 2015: 1935−1942.

[38] CARUSO D, ENGEL J, CREMERS D. Large-scale direct SLAM for omnidirectional cameras[C]//Proceedings of 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems. Hamburg, Germany, 2015: 141−148.

[39] WEIKERSDORFER D, HOFFMANN R, CONRADT J. Simultaneous localization and mapping for event-based vision systems[C]//Proceedings of the 9th International Conference on Computer Vision Systems. Petersburg, Russia, 2013: 133−142.

[40] WEIKERSDORFER D, ADRIAN D B, CREMERS D, et al. Event-based 3D SLAM with a depth-augmented dynamic vision sensor[C]//Proceedings of 2014 IEEE International Conference on Robotics and Automation. Hong Kong, China, 2014: 359−364.

[41] REBECQ H, HORSTSCHAEFER T, GALLEGO G, et al. EVO: a geometric approach to event-based 6-DOF parallel tracking and mapping in real time[J]. IEEE robotics and automation letters, 2017, 2(2): 593–600.

[42] ZHOU Yi, GALLEGO G, REBECQ H, et al. Semi-dense 3D reconstruction with a stereo event camera[C]//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany, 2018: 242−258.

[43] NEWCOMBE R A, LOVEGROVE S J, DAVISON A J. DTAM: dense tracking and mapping in real-time[C]//Proceedings of 2011 International Conference on Computer Vision. Barcelona, Spain, 2011: 2320−2327.

[44] NEWCOMBE R A, IZADI S, HILLIGES O, et al. KinectFusion: real-time dense surface mapping and tracking[C]//Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality. Basel, Switzerland, 2011: 127−136.

[45] IZADI S, KIM D, HILLIGES O, et al. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera[C]//Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology. Santa Barbara, USA, 2011: 559−568.

[46] WHELAN T, KAESS M, FALLON M, et al. Kintinuous: spatially extended kinectfusion[C]//Proceedings of RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras. Sydney, Australia, 2012.

[47] WHELAN T, JOHANNSSON H, KAESS M, et al. Robust real-time visual odometry for dense RGB-D mapping[C]//Proceedings of 2013 IEEE International Conference on Robotics and Automation. Karlsruhe, Germany, 2013: 5724−5731.

[48] WHELAN T, KAESS M, JOHANNSSON H, et al. Real-time large-scale dense RGB-D SLAM with volumetric fusion[J]. The international journal of robotics research, 2015, 34(4/5): 598–626.

[49] LABBÉ M, MICHAUD F. Memory management for real-time appearance-based loop closure detection[C]//Proceedings of 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. San Francisco, USA, 2011: 1271−1276.

[50] LABBÉ M M, MICHAUD F. Appearance-based loop closure detection for online large-scale and long-term operation[J]. IEEE transactions on robotics, 2013, 29(3): 734–745.

[51] LABBÉ M, MICHAUD F. Online global loop closure detection for large-scale multi-session graph-based slam[C]//Proceedings of 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. Chicago, USA, 2014: 2661−2666.

[52] LABBÉ M, MICHAUD F. RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation[J]. Journal of field robotics, 2019, 36(2): 416–446.

[53] KERL C, STURM J, CREMERS D. Dense visual SLAM for RGB-D cameras[C]//Proceedings of 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. Tokyo, Japan, 2013: 2100−2106.

[54] KERL C, STURM J, CREMERS D. Robust odometry estimation for RGB-D cameras[C]//Proceedings of 2013 IEEE International Conference on Robotics and Automation. Karlsruhe, Germany, 2013: 3748−3754.

[55] NEWCOMBE R A, FOX D, SEITZ S M. Dynamicfusion: reconstruction and tracking of non-rigid scenes in real-time[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 343−352.

[56] INNMANN M, ZOLLHÖFER M, NIEßNER M, et al. Volumedeform: real-time volumetric non-rigid reconstruction[C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 362−379.

[57] DOU Mingsong, KHAMIS S, DEGTYAREV Y, et al. Fusion4D: real-time performance capture of challenging scenes[J]. ACM transactions on graphics, 2016, 35(4): 114.

[58] WHELAN T, LEUTENEGGER S, SALAS MORENO R, et al. Elasticfusion: dense SLAM without a pose graph[C]//Proceedings of Robotics: Science and Systems. Rome, Italy, 2015.

[59] WHELAN T, SALAS-MORENO R F, GLOCKER B, et al. ElasticFusion: real-time dense SLAM and light source estimation[J]. The international journal of robotics research, 2016, 35(14): 1697–1716.

[60] KÄHLER O, PRISACARIU V A, MURRAY D W. Real-time large-scale dense 3D reconstruction with loop closure[C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 500−516.

[61] PRISACARIU V A, KÄHLER O, GOLODETZ S, et al. InfiniTAM v3: a framework for large-scale 3D reconstruction with loop closure[EB/OL]. (2017-08-02)[2020-02-25]. http://arxiv.org/abs/1708.00783.

[62] ENDRES F, HESS J, STURM J, et al. 3-D mapping with an RGB-D camera[J]. IEEE transactions on robotics, 2014, 30(1): 177–187.

[63] GREENE W N, OK K, LOMMEL P, et al. Multi-level mapping: real-time dense monocular SLAM[C]//Proceedings of 2016 IEEE International Conference on Robotics and Automation. Stockholm, Sweden, 2016: 833−840.

[64] SMITH R C, CHEESEMAN P. On the representation and estimation of spatial uncertainty[J]. The international journal of robotics research, 1986, 5(4): 56–68.

[65] SUALEH M, KIM G W. Simultaneous localization and mapping in the epoch of semantics: a survey[J]. International journal of control, automation and systems, 2019, 17(3): 729–742.

[66] GOMEZ-OJEDA R, MORENO F A, ZUÑIGA-NOËL D, et al. PL-SLAM: a stereo SLAM system through the combination of points and line segments[J]. IEEE transactions on robotics, 2019, 35(3): 734–746.

[67] ZHOU Huizhong, ZOU Danping, PEI Ling, et al. StructSLAM: visual SLAM with building structure lines[J]. IEEE transactions on vehicular technology, 2015, 64(4): 1364–1375.

[68] ATANASOV N, BOWMAN S L, DANIILIDIS K, et al. A unifying view of geometry, semantics, and data association in SLAM[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 5204−5208.

作者简介：

王霞，副教授，博士生导师，光电成像与信息工程研究所副所长，主要研究方向为光电成像技术和光电检测技术。主持省部级以上项目和横向合作项目多项。获授权国家/国防发明专利 10 余项，研究成果获省级技术发明二等奖 1 项、科技进步三等奖 3 项、中国电子科技集团公司科技进步三等奖 1 项。编辑出版教材 2 部，发表学术论文 70 余篇。

左一凡，博士研究生，主要研究方向为视觉 SLAM、多传感器融合导航。