

3.1 定位技术简介

终端(手机、平板、AR 眼镜)作为增强现实的硬件平台,允许用户在真实世界中自由行动是一个自然的功能需求。但为了将虚拟内容稳定地放置在真实世界中,终端的实时位姿就是一个必须知道的信息。

终端的实时位姿,既包括该终端在整个世界(地球)的全局坐标,也包括它在一个小环境(如一个房间内)中移动的局部坐标。全局坐标可通过北斗等全球定位系统获取。卫星定位技术虽然应用广泛,但是在室内以及有遮挡的地方,接收不到卫星信号,就无法通过卫星进行定位。局部坐标可以依靠终端自身的传感器计算得到,也可以通过外部布置的传感器辅助得到。

仅依靠自身的传感器(例如摄像机、惯性传感单元等)计算终端的实时位姿是一件有挑战的事情。如果是在一个陌生环境中移动,则是一个更艰巨的任务,这意味着算法需要在“黑暗”中摸索,并记录自己移动的轨迹。这是机器人领域的一个典型问题:即时定位与地图构建(Simultaneous Localization And Mapping, SLAM)。本章首先介绍 SLAM 领域的主流方法,然后重点介绍基于移动终端的典型算法,最后介绍一些辅助定位技术。

3.2 SLAM 技术简介

SLAM 问题可以被描述为一个机器人在未知环境中的未知位姿,能否通过自己的连续移动,在构建地图的同时,估计出自身在地图中的位姿。

几年前扫地机器人的出现使得 SLAM 技术走入人们的视野。其实经过多年的发展,SLAM 技术已经愈发成熟,在无人驾驶、无人机、增强现实、服务机器人等工业和商用领域有着广泛的应用。SLAM 在近三十多年的发展历程可以分为三个阶段,第一阶段是 1986—2004 年,这一时期引入了 SLAM 概率论推导方法,包括基于扩展卡尔曼滤波、粒子滤波和最大似然估计。基于概率方法的激光 SLAM 起步较早,可靠性高,精度好,技术相对成熟,在无人驾驶、扫地机器人以及仓储物流方面有较好的应用。但是激光 SLAM 主要是受激光器的体积和成本约束,在无人机、增强现实、手术机器人等方面应用受限。视觉 SLAM 因为依赖的相机体积小、便于携带、易于安装等优点获得研究者的关注。最早基于概率的视觉

SLAM 算法是由 Davison 设计的 MonoSLAM。

第二阶段是 2004—2015 年,视觉 SLAM 算法产生了较大突破,基于优化方法产生了许多经典算法。这些经典算法可以分为两类:一类是基于特征点方法,一类是基于直接法。基于特征点的 SLAM 方法作为经典的定位算法,在室内和室外等自然环境中均取得较好的效果。它从图像中提取特征,并从特征之间的匹配关系估计环境中特征点的 3D 位置和相机位姿。比较常用的特征有 SIFT、SURF 和 ORB 等。相比于只能基于稀疏地图进行定位的特征点方法,直接法可以基于每个像素构建稠密地图,其在扩展性和鲁棒性方面更有优势。直接法不依赖于特征点提取和匹配,因此它对环境特征和图像质量的丰富性(例如模糊、图像噪声等)不敏感。

从 2015 年至今,SLAM 已经进入鲁棒感知阶段,主要从鲁棒性、对环境的高层次理解、资源敏感和对任务驱动的判断这四个方面进行发展。2015 年,Mur-Artal 根据先前工作,基于 ORB 特征提出了性能最好的 ORB-SLAM 框架,其中,跟踪线程以最小化重投影误差来优化相机的姿势,而构图线程通过三角测量来估计地图点的深度。这些技术使得 SLAM 的性能更加稳定,可以长时间进行精确定位。在此基础上,可以进行机器人定位导航、协同探索等应用。随着深度学习的发展,有许多研究者将深度学习方法融入传统的 SLAM 算法中,在相机轨迹的精确性和重建点云的准确性上有所提升。Tateno 等首次提出在关键帧上用训练好的 CNN 网络来预测单帧图深度值的 CNN-SLAM,同时进行了语义分割。Laidlow 等人提出了基于对数深度图像梯度的 DeepFusion 方法,与当前先进算法效果相当。2018 年,Magic Leap 提出了基于自监督训练的特征点检测和描述符提取方法 SuperPoint,最终的结果显示效果要明显高于 ORB 的特征提取性能。

在已有工作的基础上,一些项目基于 LSD-SLAM 进行街景重建,如图 3.1(a)所示。有的工作则关注于无人机上对建筑进行在线半稠密重建,如图 3.1(b)所示。还有的项目侧重于根据多个 RGB-D 相机协同进行室内场景恢复,如图 3.1(c)所示。

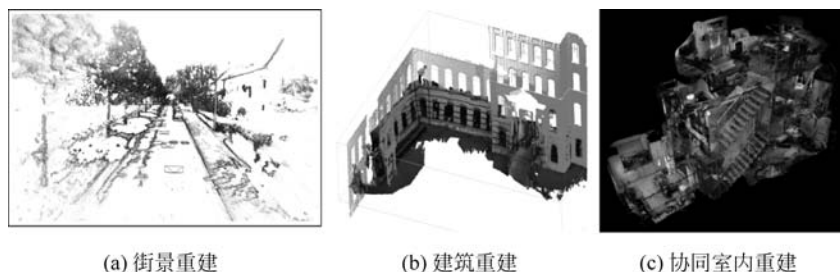


图 3.1 基于 SLAM 的实时三维重建

研究人员在智能手机上通过直接法对齐不同图像帧的像素,构建了半稠密深度地图,实现了在手机端的增强现实应用。发表于 2014 年的研究工作就可以超过 30 帧/秒,重建网格分辨率超过 320×240 ,为手机端的增强现实广泛应用奠定了基础。这些工作普遍是利用智能手机的前置摄像头,进行位姿估计和网格重建(见图 3.2)。基于单目摄像头的跟踪方法是富有挑战性的,因为缺少足够的深度信息。在后续工作中,结合单目摄像头和惯性传感单元将成为主流的技术路线。

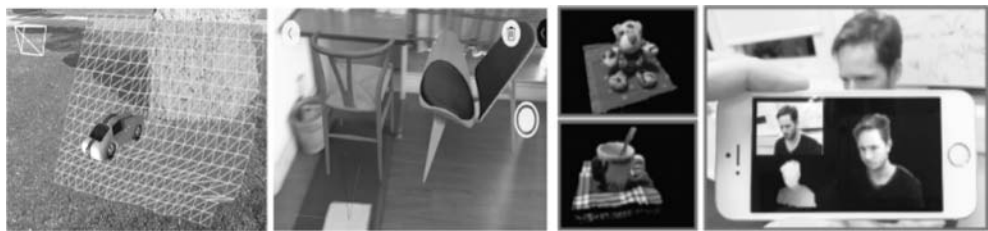


图 3.2 SLAM 在手持设备增强现实中的应用

3.3 常见 SLAM 系统介绍

下面将根据系统使用的硬件来进行分类,主要包括激光、RGB、RGBD 和惯性传感单元(IMU),并介绍各类中最经典的方法。

3.3.1 激光 SLAM

便携式激光测距仪(也称为 LIDAR)是获取平面图的有效方法。因此,激光 SLAM 是较早发展起来的 SLAM 技术,在扫地机器人、无人驾驶领域、城市搜索和救援、环境测量等方面有较为广泛的应用。激光 SLAM 方法可以直接获取从激光器到周围环境中障碍物的距离,以点云的形式表达周围环境。根据生成的点云数据,测算障碍物位置、大小、形状等。激光 SLAM 在构建地图的时候,优点是精度较高,且没有累计误差。即使是入门级别的激光雷达构建的距离测量精度也可达到 2cm 左右。但设备体积、重量都较大,无法集成于小型的增强现实硬件(例如智能手机中),同时设备购置费用较高,不适合成本敏感的应用场景。比较经典的激光 SLAM 系统有 Cartographer、Gmapping 和 Hector_SLAM。Cartographer 方法可实现 5cm 分辨率的实时建图和回环优化。Gmapping 依据机器人的运动和最近的观察结果,提出计算准确的分配方法,并选择性执行重采样操作,从而减少了滤波器预测步骤中有关机器人姿势的不确定性和粒子耗尽的问题。Hector_SLAM 结合使用 LIDAR 系统的鲁棒扫描匹配方法和基于惯性传感的 3D 姿态估计,提出了一种在线计算的系统,该系统需要较少的计算资源。通过使用地图梯度的快速逼近和多分辨率网格,可以在各种挑战性环境中实现可靠的定位和地图绘制功能。

3.3.2 视觉 SLAM 系统

MonoSLAM 是最早提出的实时单目 SLAM 系统,可以恢复单眼相机的 3D 轨迹,在未知的场景中快速移动。该方法的核心是在概率框架内在线创建稀疏的自然地标地图,使用通用运动模型来平滑相机运动,并对单眼特征初始化和特征方向的估计提出一种解决方法。PTAM 是由 Klein 和 Murray 提出的突破性 SLAM 工作,他们所提出的关键帧选择、特征匹配、三角化、每帧的摄像机定位以及跟踪失败后的重新定位对之后的工作有较大影响。但是,该方法限于小规模操作,缺乏闭环和对遮挡的适当处理,需要进行人工干预。SVO 是一种在 2014 年提出的半直接单目视觉测距算法。该算法适用于 GPS 受限的环境中的微型航空器状态估计,并且在嵌入式计算机上以每秒 55 帧的速度运行,在消费类笔记本电脑上以

每秒 300 帧的速度运行。SVO 不进行特征匹配而是直接对像素强度进行操作,从而在高帧速率下产生亚像素精度。同年,Jakob Engel 等人提出了一种在 CPU 上实时运行的直接(无特征的)SLAM 算法,即 LSD-SLAM。该算法能够获取高精度关键帧的姿势图,且允许实时构建大规模半密集的深度图。此外,该方法可以应对存在巨大变化的场景尺度。

ORB-SLAM 在 2015 年由西班牙学者 Raul Mur-Artal 提出并发表在 *IEEE Transactions on Robotics*。该系统在室内室外、大小型环境中基本都可以实时运行,对于剧烈运动鲁棒性较好,在开源数据集上有较好的表现。ORB-SLAM 是基于特征的典型方法之一,基本包含一个 SLAM 系统所需的步骤,包括跟踪、建图、重定位、回环检测等,且系统代码已经开源,代码风格简明规范,适合初学者。

近些年,由于深度学习方法的流行,将深度学习与 SLAM 方法进行结合出现了大量工作,如 DeepFusion 和 CNN-SLAM。CNN-SLAM 是基于卷积神经网络(CNN)在深度预测方面取得的最新进展,该方法优先考虑在单眼 SLAM 方法趋于失败的图像位置进行深度预测,例如沿低纹理区域。单目 SLAM 的最大局限是无法获取绝对尺度,因此使用基于 CNN 的深度预测可以估计重建后的绝对尺寸。此外,还可以有效地融合语义标签。

3.3.3 RGB-D SLAM

随着 Kinect 等硬件设备的发展,RGB-D SLAM 依据硬件可以获取场景中深度的优势,发展出了一系列的方案,如 RGBDSLAM、RTAB-MAP、ElasticFusion、KinectFusion 和 Kintinuous。RGBDSLAM 仅使用 RGB-D 相机可靠地生成高精度的 3D 地图,可以强大地应对挑战性的场景,例如,快速的摄像机运动和特征点较稀疏的环境,同时又足够快地进行在线操作,适用于小型家用机器人(如吸尘器)以及飞行机器人(如直升机)。RTAB-MAP 是一种基于外观的闭环检测方法,具有内存管理功能,可处理大规模且长期的在线操作。自 2013 年提出以来,它逐渐在各种机器人和移动平台上实现。ElasticFusion 是一种实时密集视觉 SLAM 法,能够以增量在线方式获得密集且全局一致的房间规模环境地图。该方法通过使用密集的图片帧到模型的相机跟踪以及通过非刚性表面变形进行模型细化。通过对场景的外观属性进行简单的假设,该方法以在线方式递增地估计环境中多个光源的数量和位置。KinectFusion 是一种在可变照明条件下能准确实时地绘制复杂和任意室内场景地图的系统。该系统将来自 Kinect 传感器的所有深度数据,通过使用由粗到细的迭代最近点(ICP)算法获得当前传感器的姿态,之后将观测的深度数据实时融合到全局隐式表面模型中。实验证明,它允许实时重构密集的表面,并具有一定程度的细节和强大的功能。

3.3.4 结合惯性传感单元的 SLAM 方法

近年来,结合视觉和惯性传感单元来完成同时定位与地图构建任务的相关技术经过了快速发展,涌现出了大量的优秀工作,比较有代表性的有 MSCKF_VIO、VINS-Mono 和 BASALT。MSCKF_VIO 是一种基于滤波器的视觉惯性里程计方法(Visual-Inertial Odometry, VIO),使用了多状态约束卡尔曼滤波器(MSCKF)。实验表明,该方法可与最新的单眼解决方案相提并论,同时具有更好的鲁棒性。VINS-Mono 是融合视觉和惯性传感器,用于六自由度状态估计的 SLAM 方法。但是,每个传感器的时间戳通常会因不同原因导致不同传感器之间的时间偏移,极大地影响了传感器融合的性能。VINS-Mono 中使用了

一种在线方法,用于校准视觉和惯性测量之间的时间偏移。BASALT 也是一种相机和惯性测量单元结合使用的视觉惯性里程计系统。BASALT 主要针对全局一致的映射问题,提出基于非线性因子方法,对 VIO 系统累积的轨迹信息进行了最佳近似,并使用集束调整将这些因素与闭环约束结合起来。

3.4 视觉里程计

视觉里程计(Visual Odometry, VO)是利用单个或多个相机的输入信息估计运动信息的方法。它只利用相机完成,相比于激光雷达,大大降低了成本。此外,基于最新的深度学习技术,视觉信息还可以提供丰富的环境理解功能,完成物体识别等任务。视觉里程计不仅在增强现实,也同时在无人车、无人机、水下机器人等领域有广泛应用,其中最成功的应用当属美国 NASA 开发的火星探测器“勇气号”和“机遇号”。

ORB-SLAM 方法成功的原因之一是它集合并继承了许多优秀工作,在此基础上进行了进一步的开发,比如整体架构来源于 PTAM,回环检测任务中用到的词袋模型等。总结来说,ORB-SLAM 系统有以下几个创新点。一是将整体流程分成处理速度不同的三个线程:跟踪、建图以及回环检测线程。三个线程相互配合既保证了系统的速度,又提供了较高的精度。二是在三个线程中都使用的是 ORB 特征,这使得系统整体有较高的效率和准确性。三是提出了新颖稳定的地图初始化方法,在全自动初始化功能和长时间地图构建方面有了较大优化。四是基于 DBOW2 中的词袋技术进行了回环检测和重定位,提高系统的准确性和鲁棒性。下面将针对这些创新点,对 ORB-SLAM 系统进行详细介绍。首先,从系统框架、关键数据结构以及回环检测方法等方面对系统进行整体回顾。其次,根据图片处理顺序,按照地图初始化、跟踪、建图、回环检测的顺序对每一部分进行展开,介绍其中包含的原理、基本方法及实验效果。

3.4.1 系统整体回顾

系统包含三个并行且分工明确的跟踪、建图和回环检测线程,如图 3.3 所示。跟踪线程负责在每帧中对相机进行定位、确定何时插入新的关键帧、对于跟踪丢失进行重定位;局部建图线程处理新的关键帧并执行局部束调整(Bundle Adjustment),根据匹配关系通过三角测量创建新的地图点,删除冗余关键帧和旧地图点,以实现相机姿势周围环境的重建;回环检测线程对每个新关键帧进行回环检测及优化,降低累积漂移并保证全局一致性。跟踪线程会新建关键帧并传递给局部建图线程。如果在局部建图线程不处于空闲时插入了关键帧,则会发送一个信号以停止局部束调整,以便它可以尽快处理新的关键帧。

如图 3.3 所示,三个线程之间主要通过关键帧(KeyFrames)和地图点(MapPoints)传递信息。每一个地图点中存储的是该点在世界坐标系中的 3D 位姿、观察方向(即将该点与观察该点的关键帧光学中心相连的光线平均单位向量)、代表性的 ORB 描述符以及可以观察到该点的最大和最小距离。每一个关键帧包含相机姿态、相机固有特性(包括焦距和主点)以及框架中提取的所有 ORB 要素。

此外,还需要重点维护的结构是共视图(Covisibility Graph)和生成树(Spanning Tree),这是保证系统长时间稳定运行、节省内存的关键。关键帧之间的共视信息表示为无

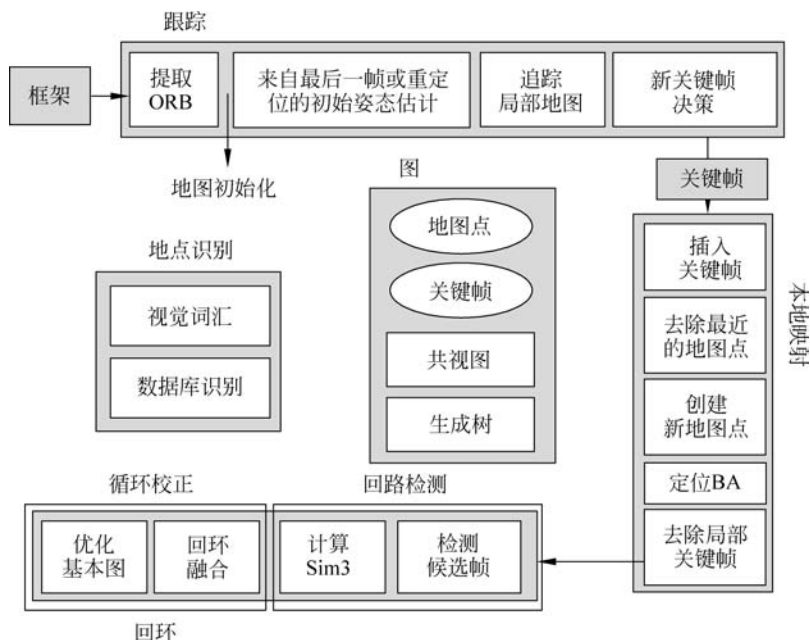


图 3.3 ORB-SLAM 系统架构图

向加权图,每个节点都是一个关键帧,如果两个关键帧之间共享部分地图点(至少 15 个)的观测值,则存在两个关键帧之间的一条边,即边缘的权重为公共地图点数。如图 3.4(b)所示,共视图中包含所有边,使得连接非常密集。为了更有效地进行用于回环检测的位姿优化,ORB-SLAM 系统维护了一个基本图(Essential Graph),保留所有节点(关键帧),但保留较少的边缘,同时仍保留可产生准确结果的强大网络。系统从初始关键帧开始逐步构建生成树,该生成树是具有最小边数共视图的连接子图,如图 3.4(c)所示。在生成和淘汰关键帧时,会更新关键帧之间的连接。在闭环优化实验过程中,束调整存在收敛问题,即使经过 100 次迭代,误差仍然很高。实验结果表明,基本图优化实现了快速收敛并取得了更准确的结果,说明减少边缘的数量可以大大减少时间。

3.4.2 地图初始化

因为单目相机不能获取深度信息,所以单目 SLAM 需要通过初始化确定空间尺度并且得到一个初始地图。单目 SLAM 的初始化,需要手持摄像头平移一小段距离。在这个过程中,算法一直在跟踪每一帧的特征点。如果跟踪得好的话,就可以得到两个关键帧。通过计算两个帧之间的相对姿态,以对一组初始的地图点进行三角测量。束调整优化之后就可以得到初始化的地图,然后进入跟踪状态。

ORB-SLAM 系统并行计算两个几何模型,一个单应矩阵(Homography Matrix)模型假设平面场景,一个基础矩阵(Fundamental Matrix)模型假设非平面场景。然后,使用启发式方法选择模型,并尝试使用针对所选模型的特定方法来恢复相对姿势。地图初始化的步骤是,首先提取当前帧中的 ORB 特征,并在参考帧中搜索匹配项;其次,在并行线程中分别计算单应和基础矩阵模型。在每次迭代过程中,为每个单应矩阵模型和基础矩阵模型分别计

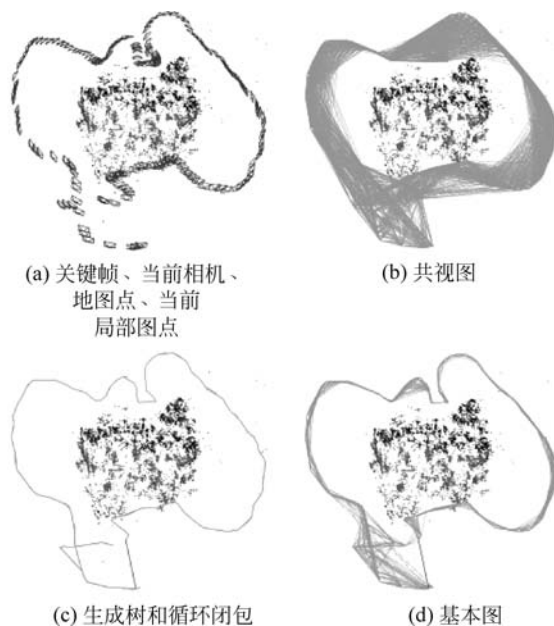


图 3.4 共视图和基本图结构示例

算分数。如果场景是平面的、接近平面的或视差较低可以用单应性表达,具有足够视差的非平面场景用基础矩阵来表达。

3.4.3 跟踪

跟踪线程会对从相机获得的每一帧图像提取 FAST 角点。算法针对不同分辨率的图像,将采取不同的提取策略,比如对于从 $512 \times 384\text{px}$ 到 $752 \times 480\text{px}$ 的图像分辨率适合提取 1000 个角点,对于更高的分辨率 $1241 \times 376\text{px}$ 适于提取 2000 个角点。如果某些单元格不包含角点(无纹理或低对比度),则调整每个单元格保留的角点数量。然后在保留的 FAST 角上计算方向和 ORB 描述符。

跟踪线程根据跟踪状态获取相机位姿。如果跟踪成功,将使用恒速运动模型来预测相机的姿态,并对在最后一帧中观察到的地图点进行指导搜索;如果当前运动不满足恒速运动模型,将在最后一帧中围绕地图点的位置进行更广泛的搜索。如果跟踪失败,则通过使用词袋模型,为当前帧在已有数据库中查找获取关键帧,计算与每个关键帧中的地图点关联的 ORB 的对应关系。然后,该方法对每个关键帧执行替代性的 RANSAC 迭代,并尝试找到相机姿势。如果找到具有足够内点支持的相机姿态,则会优化该姿态并在候选关键帧的映射点上进行更多匹配的引导搜索。一旦该方法估计了摄像机的姿势并获得了一组初始的特征匹配,就可以将地图投影到框架中并搜索更多的地图点对应关系。最后,使用框架中找到的所有地图点来优化相机姿态。

跟踪的最后一步是确定是否将当前帧作为新的关键帧。跟踪的原则是提高对于具有挑战性的摄像机运动(通常是旋转)的鲁棒性,因此会尽快插入关键帧。但是为了确保良好的重定位、跟踪效果,需要满足以下 4 个条件才能插入新的关键帧:①自上次全局重新定位以来,必须已经传递了二十多个帧;②局部建图线程处于空闲状态,或者从上一次插入关键帧

起已经传递了二十多个帧；③当前帧至少跟踪 50 个点；④前一个关键帧的特征点在此帧里已经有 90% 观测不到。

3.4.4 局部建图

首先,对于每个新生成的关键帧,在共视图和生成树中添加相应节点并连接边。然后,对局部地图中的地图点进行管理,一方面进行严格的筛选策略,比如至少从三个关键帧中可观察的地图点才能被保留;另一方面进行可靠的生成策略,通过从共视图中的已连接关键帧中三角测量 ORB 特征点来创建新的地图点。此外,局部建图线程也会根据条件对冗余的关键帧进行删除,比如 90% 的地图点在至少其他三个相同或更高数量中可以看到的关键帧。局部建图对于地图点和关键帧的严格操作,为 ORB-SLAM 长时间鲁棒稳定运行提供了保证。否则,关键帧的数量可能会无限制地增长,而集束调整的复杂度会随着关键帧的数量而增加。最后,在局部建图线程中的局部束调整,会优化当前处理的关键帧、共视图中与其连接的所有关键帧以及这些关键帧看到的所有映射点。其他看到这些地图点但与当前处理帧未连接的所有关键帧都包含在优化中,但保持不变。在优化的中间和结束时,将标记为异常的观察值丢弃。

3.4.5 回环检测

回环检测线程对局部建图线程处理的最后一个关键帧进行回环检测和回环优化。在回环检测和重定位模块,ORB-SLAM 主要基于 DBoW2 项目。在 DBoW2 项目中,主要使用了词袋技术。词袋模型(见图 3.5)原本是在信息检索技术中,用于对文本进行特征建模的方法。在计算机视觉中,视觉单词是描述符空间的离散化,可以从大量图像中提取的 ORB 描述符离线创建词汇表。如果图像足够通用,则可以将相同的词汇表用于不同的环境以获得良好的性能。该系统以增量方式构建一个数据库,该数据库包含一个反向索引,该索引为词汇表中的每个视觉词汇存储了已看到的关键帧,因此可以非常有效地查询数据库获取共视关键帧。在删除或者新建关键帧时,也会更新数据库。由于关键帧之间存在视觉重叠,因此在查询数据库时,将不存在评分较高的唯一关键帧。该系统不仅考虑在时间上接近的图像,还包括在同一位置但在不同时间插入的关键帧,最后返回所有得分高于最佳得分 75% 的关键帧匹配项。

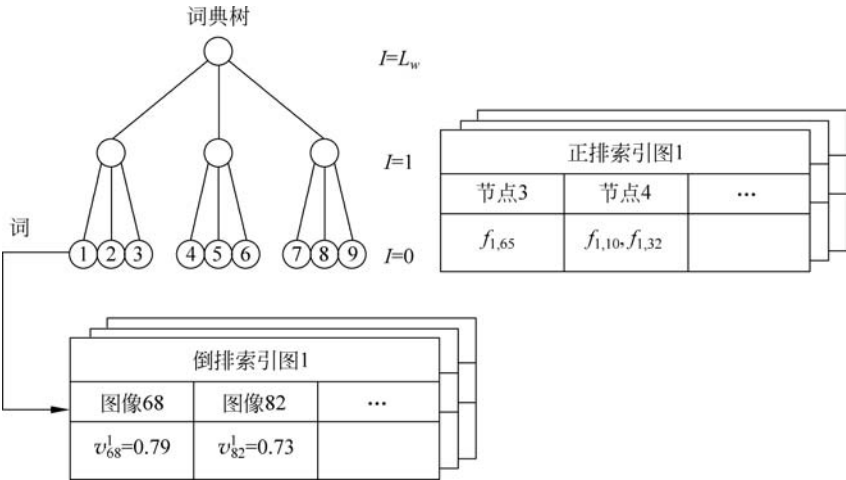


图 3.5 词袋模型

首先,需要根据词袋模型在共视图中计算相似度,获取回环候选帧。必须连续检测三个一致的回环候选(共视图中连接的关键帧)才能认定存在回环。如果有多个地方相似,则可以有多回环候选。在进行回环检测后,该方法需要计算从当前关键帧到回环关键帧的相似性变换。这是因为在单眼 SLAM 中,除了三个平移和三个旋转这六个自由度之外,还存在一个比例因子。计算相似变换的第一步是计算与当前关键帧中的映射点关联的 ORB 与回环候选关键帧之间的对应关系。可以对每个候选对象执行 RANSAC 迭代,尝试使用 Horn 的方法找到相似性变换。回环检测最关键的一步是根据相似变换将重复的地图点融合在一起,对共视图进行更新。首先,使用相似度变换对当前关键帧姿态进行校正,并将此校正传播到最后关键帧的所有邻居,并连接变换,以使回环的两头对齐。回环关键帧及其邻居看到的所有地图点都投影到最后关键帧中,并且在投影周围的狭窄区域中搜索其邻居和匹配项。匹配所有相关地图点,并融合那些在计算相似度时不合理的点。融合中涉及的所有关键帧都将在共视图中更新其边缘,从而有效地创建连接回环闭合的边缘。图 3.6 显示了闭环之前和之后的重构。以红色显示局部映射,该局部映射在回环闭合之后沿闭合处的两侧延伸。

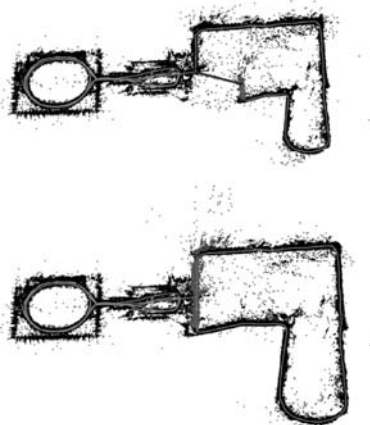


图 3.6 回环检测效果

3.5 视觉惯性里程计

惯性测量单元(Inertial Measurement Unit, IMU)是一种电子设备,可以结合使用加速度计、陀螺仪,有时还包括磁力计,可以测量速度、方向和重力。IMU 通常被合并到惯性导航系统中,该系统利用原始 IMU 测量值来计算相对于全局参考系的姿态、角速度、线速度和位姿。

视觉惯性里程计(Visual-Inertial Odometry, VIO)利用视觉和惯性两种传感器实现智能设备的三维空间轨迹跟踪。这种融合充分利用了两种类型的传感器在不同方面的优势互补。两种类型传感器信号的融合主要基于 IMU 预积分技术。传感信号的融合在 VIO 项目中主要体现在初始化、基于窗口融合、重定位以及优化过程中。视觉传感器在大多数纹理丰富的场景中效果很好,但是如果遇到玻璃、白墙等特征较少的场景,基本上无法工作;快速

运动时定位跟踪容易丢失；单目视觉无法测量尺度。但是由于视觉不产生漂移，可以直接测量旋转平移。相反，IMU 有输出频率高、能输出 6DoF 测量信息等优点，在短时间内，其相对位移数据有很高的精度；但由于零偏和噪声的存在，导致其长时间使用有非常大的累积误差。VIO 方法获得广泛应用的另外一个主要原因是以手持智能设备为主的增强现实平台普遍都会搭载这两款传感器。

视觉和 IMU 融合可以分为基于滤波和基于优化两种方法。按照是否把图像特征信息加入状态向量来进行分类，可以分为松耦合和紧耦合两大类。

松耦合将视觉传感器和 IMU 分别计算得到的位姿直接进行融合，融合过程对二者本身不产生影响，作为后处理方式输出，一般通过 EKF 进行融合。

紧耦合将视觉传感器和 IMU 的状态通过一个优化滤波器合并在一起，紧耦合需要把图像特征加入到特征向量中，共同构建运动方程和观测方程，然后进行状态估计，最终得到位姿信息的过程，其融合过程本身会影响视觉和 IMU 中的参数（如 IMU 的零偏和视觉的尺度）。比较典型的是基于优化思路的 VINS-Mono。在后续章节中，也将以这个方法为代表，重点介绍 VIO 方法。

总体来说，视觉惯性里程表是一个研究的热点和难点，本节仅简要介绍。读者若对其中涉及的具体流程感兴趣，可以参考发表的论文或者其他资料。

3.5.1 单目 VIO 初始化

单眼紧密耦合的视觉惯性里程表是一个高度非线性的系统。由于无法从单眼相机直接观察到尺度，因此很难在没有良好初始值的情况下直接融合这两个测量。同时，IMU 传感器必然存在偏置值，在后续步骤中还需要获得精确的重力加速度和速度等信息。因此，需要在紧耦合系统处理之前计算出这些量。初始化的目的就是计算出绝对尺度、陀螺仪偏置、加速度偏置、重力加速度和每个 IMU 时刻的速度。VIO 采用一种松耦合的传感器融合方法来获取初始值。纯视觉 SLAM 方法具有良好的初始化属性，如前所述，可以通过从相对运动方法中得出的初始值来进行初始化。然后，通过将 IMU 预积分与仅视觉的初始化信息对齐。如图 3.7 所示为非线性单目 VIO 的初始化对齐过程。

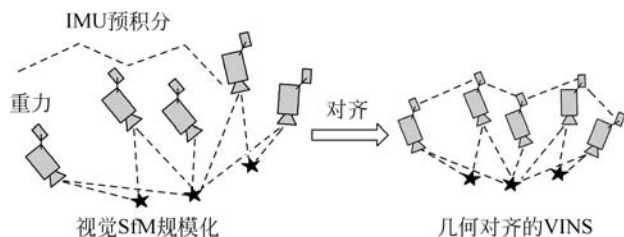


图 3.7 初始化阶段 VIO 对齐过程说明

相机的旋转位姿可以通过两种方式求得，一种是陀螺仪测量值，一种就是视觉观测值。按照正常的理解两者的大小一定是相等的（假设没有误差），但实际情况肯定有误差。陀螺仪的误差有两部分：测量噪声和陀螺仪偏置。噪声暂时可以忽略，而视觉的误差就只有观测噪声，因此两者差值的绝对值就是陀螺仪偏置，将整个滑动窗口的所有的旋转做差构成了一个最小化误差模型，便可初始化旋转值。

3.5.2 紧耦合重定位模块

为了消除漂移,研究人员提出了与单眼 VIO 无缝集成的紧密耦合的重新定位模块。与 ORBSLAM 相同,重新定位过程从一个识别已经访问过的回环检测模块开始,然后在回环闭合候选者和当前帧之间建立连接,从而以最小的计算开销实现了无漂移状态估计。图 3.8 图示了重新定位和姿势图优化过程。图 3.8(a)显示了重新定位过程,它从只有 VIO 的姿势估计开始,过去的状态用另一种颜色记录。姿态图优化如图 3.8(b)所示。

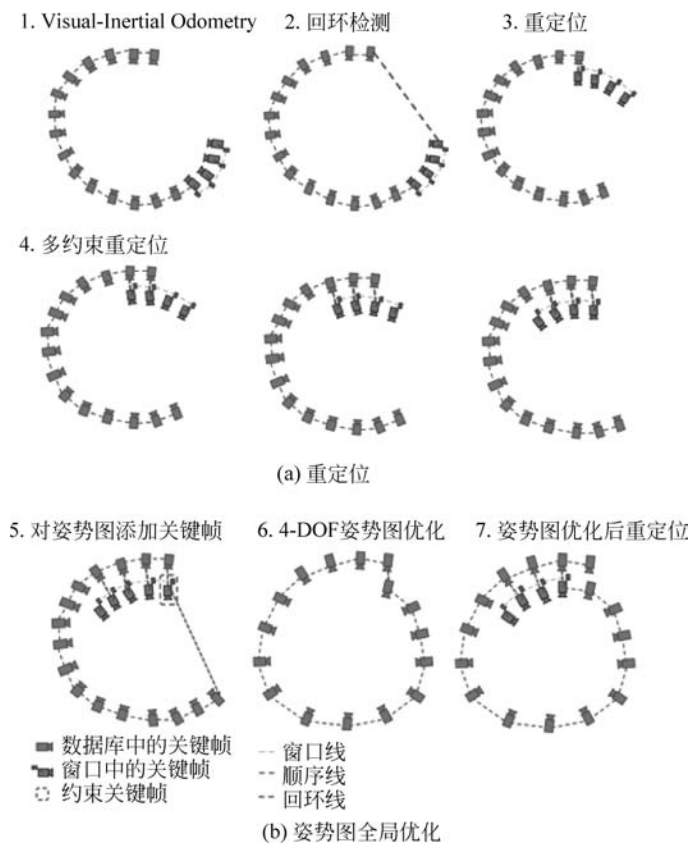


图 3.8 重新定位和姿势图优化过程示意图

VIO 与 ORBSLAM 一样利用 DBoW2 进行回环检测。当检测到回路时,通过检索特征对应关系在本地滑动窗口和闭合候选之间建立连接。但是,仅依靠简要描述符匹配找到的对应关系可能会导致大量异常值。因此,使用两种方法将几何离群值剔除。一方面,使用对当前图像和回环闭合候选图像中检索到的特征的 2D 观察来执行基础矩阵测试。另一方面,基于局部滑动窗口中特征的已知 3D 位姿以及闭环候选图像中的 2D 观察进行 PnP 算法,并从候选对象中选择合适的回环关键帧,执行重新定位。

如果检测到最新关键帧的回环,如第二幅图所示,则会发生重新定位。请注意,由于使用了特征级别对应关系进行重新定位,因此该方法能够合并来自多个过去关键帧的回环闭合约束,如最后三个图所示。当关键帧从滑动窗口边缘化时,会将关键帧添加到姿势图中。如果在此关键帧和任何其他过去的的关键帧之间存在回环,则回环闭合约束(公式化为 4-

DOF 相对刚体变换)也将添加到姿势图中。使用单独线程中的所有相对姿势约束优化姿势图,并且重新定位模块始终相对于最新姿势图配置运行。

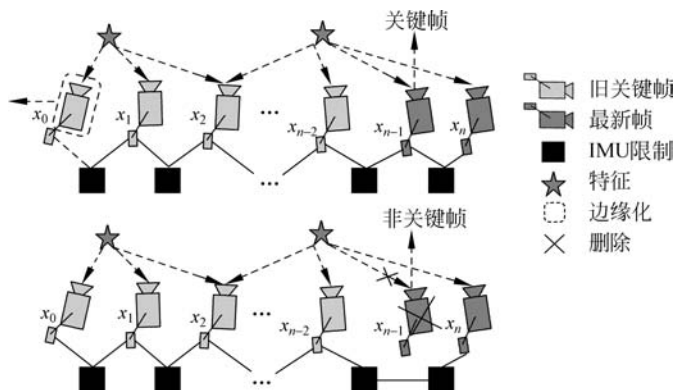
重新定位将单眼 VIO 保持的当前滑动窗口与过去的姿态图对齐。在重新定位期间,VIO 将所有闭环帧的位姿都视为常量。该方法使用所有 IMU 测量值,局部视觉测量值以及从闭环中检索到的特征对应关系共同优化滑动窗口。视觉测量的模型与前面相同,唯一的区别是从姿势图或直接从过去的里程表获取的闭环框架的姿势被视为不变。

3.5.3 全局位姿图优化

重新定位后,局部滑动窗口与过去的姿势对齐。利用重新定位的结果,结合姿态图优化以确保位姿的全局一致性。由于惯性测量单元可以完全观察到侧倾角和俯仰角,因此累积的漂移仅发生在四个自由度(x, y, z 和偏航角)中。为此,VIO 忽略了估计无漂移侧倾和俯仰状态,仅执行四自由度姿态图优化。当关键帧从滑动窗口进行边缘化时,它将添加到姿势图。该关键帧用作姿势图中的顶点,并且通过两种类型的边与其他顶点连接,一种是顺序边,顺序边表示局部滑动窗口中两个关键帧之间的相对转换,该值直接从 VIO 获取;另一种闭环边,如果新边缘化的关键帧具有回路连接,则它将通过姿势图中的回路闭合边与回路闭合帧连接。

3.5.4 如何在手机平台取得优异的实时性能

为了限制基于优化的 VIO 的计算复杂性,引入了边缘化。VIO 系统从滑动窗口中选择性地边缘化 IMU 状态和特征,同时将与边缘化状态相对应的度量转换为先验。目的是为特征三角测量提供足够的视差,最大化保持加速计测量的可能性。具体过程如图 3.9 所示,如果第二个最新帧是关键帧,该方法将其保留在窗口中,并边缘化最旧的帧及其相应的视觉和惯性测量值。如果第二个最新帧不是关键帧,该方法将简单地删除该帧及其所有相应的视觉度量。但是对于非关键帧,会保留预集成的惯性测量值,并且预积分过程将继续进行到下一帧。



对于计算能力较低的设备(例如手机),由于非线性优化的计算量很大,因此紧密耦合的单目视觉里程计无法实现摄像机速率的输出。为此,可以采用轻量级的仅考虑运动视觉惯

性束调整策略,以将状态估计提升至摄像机速率(30Hz)。

在视觉惯性里程计中,这个束调整策略的成本函数与单目视觉里程计的成本函数相同。但是该方法没有优化滑动窗口中的所有状态,而是仅优化了一定数量的最新 IMU 状态的姿态和速度。图 3.10 显示了所建议策略的示意图。与完全紧密耦合的单眼 VIO 形成对比的是,在先进的嵌入式计算机上,VIO 可能会产生 50ms 以上的时间消耗,而这个束调整策略的计算仅需 5ms。这样可以实现低延迟相机速率的姿势估计,对于无人机和增强现实应用特别有用。

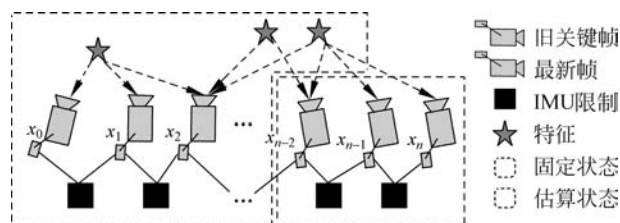


图 3.10 仅考虑运动视觉惯性束优化示意图

与视觉测量相比,IMU 测量的速率要高得多。尽管 VIO 方法的频率受到图像捕获频率的限制,但该方法仍可以使用最新的 IMU 测量值直接传播最新的 VIO 估计值,以最大程度地利用 IMU 高频性能。此外,当行进距离增加时,姿势图的大小可能会无限增长,从而限制系统的实时性能。为此,VIO 实施了下采样过程以将姿势图数据库保持在有限的大小。将保留所有具有回环闭合约束的关键帧,而其他过于接近或与相邻对象的方向非常相似的关键帧则可以删除。

3.6 辅助定位技术

3.6.1 全局位置跟踪和导航

这部分将主要介绍可以实现全局位置跟踪和导航的相关技术。卫星定位是获取地球上任意一个终端位置的成熟技术。全球导航卫星系统(GNSS)指的是一个卫星群,提供来自太空的信号,这些信号将定位和定时数据传输到 GNSS 接收器。然后,接收器使用此数据来确定位置。GNSS 的例子包括欧洲的伽利略系统(GALILEO)、美国的全球定位系统(GPS)、俄罗斯的格洛纳斯系统(GLONASS)和中国的北斗导航卫星系统(BDS)。

北斗卫星导航系统是中国自主研发、独立运行的全球卫星导航系统。该系统分为三代,即北斗一代系统、北斗二代系统和北斗三号全球卫星导航系统。我国 20 世纪 80 年代决定建设北斗系统,2003 年,北斗卫星导航验证系统建成。该系统由 4 颗地球同步轨道卫星、地面控制部分和用户终端三部分组成。2020 年,北斗三号全球卫星导航系统正式开通,在全球范围内提供定位与导航服务。

RTK(Real-Time Kinematic)是一种利用 GPS 载波相位观测值进行实时动态相对定位的技术。RTK 的工作原理是基准站通过数据链将其观测值和测站坐标信息一起传送给终端。终端不仅通过数据链接收来自基准站的数据,还要采集 GPS 观测数据,并在系统内组成差分观测值进行实时处理,同时给出厘米级定位结果。此外,华为、Google 目前都有采用

基于视觉的全球位姿定位技术(Visual Positioning Service,VPS)。VPS在使用GPS定位时会同时采用几十个甚至几百个视觉参考点来做定位,可以更准确地获取终端实时的位姿朝向。VPS技术的理论精度可以达到厘米级别。除了使用视觉参考点,VPS一般还会使用惯性传感器来进一步提升精度。通过城市级的3D重建和视觉定位技术可以获取更精准的全球位姿定位,并可以解决室内全球位姿定位的难题。

已构建的全球网络大多数依靠陆地和海底光纤电缆进行数据传输。建设这些光缆需要耗费大量的人力物力,也只能服务在光缆沿线的用户。美国公司Starlink通过将小型卫星网络置入低地球轨道,建立一个由12000颗卫星组成的卫星网,其中大约三分之二的运行轨道在500km以上,剩下的则为1200km。该网络可以连接到地球上的任何地方,比目前的卫星服务快40倍。Starlink卫星互联网服务的延迟为20ms,目标价格约为每月80美元(约合人民币566元),可以为地球上的大部分用户提供便捷的网络服务。

3.6.2 局部位置跟踪与导航

局部位置跟踪与导航技术有很多的实现方案,本节简单介绍以下四种主流技术,包括超宽带、Wi-Fi、射频和蓝牙。

超宽带(Ultra-Wide Bandwidth,UWB)是一种无线电技术,可以在很大一部分无线电频谱上使用非常低的能量水平进行短距离、高带宽通信。最近的应用目标是传感器数据收集,精确定位和跟踪应用。它的优点是功耗低,精度高。定位精度可以在30cm内。苹果手机iPhone 11就集成了UWB技术。这个精度非常适合在室内场景下做绝对位置的定位,再结合视觉-惯导技术进行实时跟踪,则可以进一步提升精度。这为室内场景的增强现实应用提供了理想的定位方案。

Wi-Fi是基于IEEE 802.11系列标准的无线网络技术家族,通常用于设备的局域网和Internet访问。一个接入点(或热点)在室内的射程通常约为20m。热点覆盖范围可以小到单独的一个房间,它的墙壁可以挡住无线电波,也可以使用多个重叠的接入点在它们之间允许漫游的情况下,将其覆盖到最大的平方千米。主流算法通常利用信号从热点到终端的传输时间、信号衰减模型估算出移动设备距离各个热点的距离。再同时计算终端到超过三个热点之间距离,形成三角形结构,确定终端的三维位置。

射频识别(RFID)使用电磁场来自动识别和跟踪附着在物体上的标签。RFID标签由一个微小的无线电应答器构成。当RFID标签被附近的RFID读取器设备发出的电磁询问脉冲触发时,会将识别库存编号的数字数据发送回读取器。无源标签由RFID阅读器询问无线电波的能量供电,范围在10m内。有源标签由电池供电,因此可以在更大范围(可达数百米)内从RFID读取器读取。与条形码不同,该标签不需要位于阅读器的视线范围内,因此可以将其嵌入被跟踪的对象中。一个常见的使用场景就是在商场里的重要商品,通常会有一个RFID标签,用于防止物品盗窃。在定位的应用场景下,利用RFID技术去探测接收器与三个发射器之间的距离,同样形成一个三角形结构,进而确定该接收器在空间内的三维位置。

蓝牙也是在AR里面常用的辅助定位技术之一。现在的智能手机普遍搭载基于蓝牙4.0的低功耗蓝牙技术(Bluetooth Low Energy,BLE),可以检测到由固定位置的信号发射器发出的蓝牙信号,通过软件和硬件的结合,从而大大提高室内精度,甚至达到1m以内的

定位精度。由苹果公司推出的室内定位技术 iBeacon, 就可以让附近的手持电子设备检测到 50~80m 的发射器信号。

小 结

本章首先介绍 SLAM 领域的主流方法, 然后重点介绍基于移动终端的典型算法, 最后介绍一些辅助定位技术。

习 题

1. 假设要实现一个室外景区的增强现实应用, 可能需要哪些定位技术? 为什么?
2. 假设要实现一个室内博物馆的增强现实应用, 可能需要哪些定位技术? 为什么?
3. 基于激光、RGB、RGBD、惯性等不同传感器的 SLAM 方法, 各自的优缺点分别是什么?
4. SLAM 问题和图像感知是否有联系的地方?
5. SLAM 问题中的回环检测主要是解决什么问题?
6. 在视觉惯性里程计的方法中, 视觉传感器和惯性传感器是如何相互补充的?