视觉 SLAM 技术及应用

2018 年 9 月 16 日

目录

摘要

SLAM（Simultaneous Localization and Mapping）意为同时定位与建图，而视觉 SLAM 是以视觉传感器为主的 SLAM 技术。视觉 SLAM 技术被广泛应用于机器人、自动驾驶、增强现实等领域。本文先介绍了视觉 SLAM 技术及其面临的实际问题；为了解决这些问题，然后介绍了课题组的五项工作：非连续帧的特征匹配、大尺度场景下的优化策略、如何应对动态场景、移动端 SLAM 和增量式优化的 RGB-D SLAM；最后讨论了视觉 SLAM 技术的应用领域和发展趋势。

1 视觉 SLAM 简介

SLAM 属于机器人领域。SLAM 的定义是，在未知环境下，同时估计设备的位置姿态和三维环境结构。因为 SLAM 技术能估计自身位姿，所以被广泛应用于增强现实（AR）、虚拟现实（VR）、机器人、自动驾驶等领域。根据地图的不同形式，SLAM 可分为稀疏 SLAM 和稠密 SLAM。

SLAM 技术经过多年发展，目前它的框架趋于稳定（图1）。其框架主要包括了并行的跟踪（Tracking）和建图（Mapping）线程，前端根据传感器输入的数据，实时跟踪地图并估计

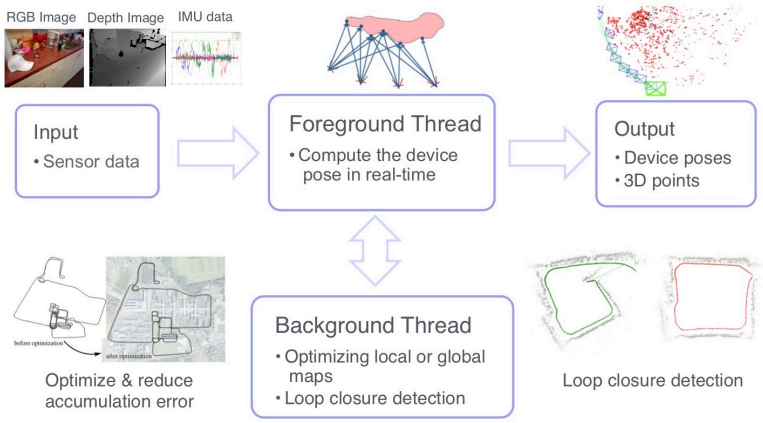

图 1: SLAM 技术框架

相机的位姿，后端则进行局部或全局地图优化和回路闭合，从而消除误差累积。最后实时输出设备的位姿和三维地图。

视觉 SLAM 的传感器主要是相机，分为单目相机、双目相机和多目相机。广义上的视觉 SLAM 是指，以视觉传感器为主、其他传感器为辅的技术，其他的传感器包括 IMU、GPS、深度相机等。视觉 SLAM 的主要优势在于它的成本低、小尺度场景下定位精度高、不需要预先的场景布置。

视觉 SLAM 技术的理论基础是多视图几何（Multi-View Geometry）。根据相机投影方程 $X_{ij} = \pi(P_i X_j)$ 我们知道三维空间点在像平面的观测投影点和理论计算得到的结果应当是相同的，但实际上却存在重投影误差，所以我们需要优化变量，最小化重投影误差。视觉 SLAM 技术经过多年发展，其理论已经成熟，但在实际复杂场景应用时却会遇到以下几个方面的挑战：

- 如何处理循环回路序列和多视频序列
- 如何在大尺度场景下实现高精度实时定位
- 如何在动态场景下稳定跟踪
- 如何应对快速运动和剧烈旋转的情况

围绕这些问题，我们课题组完成了以下五个工作。

2 非连续特征跟踪

运动恢复结构（SfM）的是指根据视频序列或多张图像，离线或在线地恢复出相机位姿。运动恢复结构首先需要解决好特征匹配，特征匹配是根本性的前提，如果特征匹配做不好，会导致后面的结果不稳定。

第一个需要解决的问题：如何匹配不同子视频序列的公共特征点。我们提出了非连续特征跟踪方法，分为连续特征跟踪和非连续轨迹匹配两个步骤。第一步针对连续特征跟踪，我们提出了 TWO-Pass 匹配方法：先用 SIFT 特征匹配方法得到若干匹配点，再假设场景可以用

多个平面近似表达，那么可以用多个单应矩阵引导特征匹配，最终找到更多的匹配点，这样可以有效延长特征点跟踪的轨迹。第二步针对非连续轨迹匹配，此时暴力匹配无法处理长视频序列。为了解决非连续帧的匹配问题，我们先快速估计匹配矩阵：对连续帧跟踪得到的特征轨迹，采用对每帧描述量求和取平均值的方法得到轨迹的描述量，同时用 hierarchical K-means 方法快速得到匹配矩阵的初始值。然后同时迭代优化特征匹配和匹配矩阵。实验证明该方法对初始匹配矩阵的要求非常低，能在线性时间复杂度内完成非连续特征轨迹匹配。这部分工作属于 ENFT-SfM[1]。

3 大尺度场景下的优化策略

第二个需要解决的问题：如何在特征匹配已知的情况下高效地用全局优化消除累积误差。Bundle Adjustment 会把所有相机参数和三维点一起优化，所以它的变量数目和内存的需求都非常大，迭代优化速度非常慢；如果每次只优化一部分，固定其他部分，则容易陷入局部最优解，无法得到全局最优解；如果先消去三维点，再进行 Pose Graph 优化，虽然减少了变量的数目，但是在消元过程中使用线性化带入的误差会导致结果并非最优。

针对大尺度场景，我们提出了 Segment-based BA 优化方法，我们把长视频序列分解为若干短视频序列，再对每个短视频序列独立进行 SfM，然后把它们对齐。每个短视频序列有 7 个自由度（空间 6 自由度和 Scale 自由度），且内部的三维结构是固定的，每次对短视频做整体的变换。我们根据公共特征点对齐每段视频，如果对齐过程中误差较大，那么我们需要找到使能量下降最快的分裂点，把视频一分为二，增加视频的自由度，增强优化能力，反复分裂，直到重投影误差足够小为止。Segment-based BA 是一个从粗到细（Coarse-to-Fine）的优化过程，最主要的优点是不容易陷入局部最优解，容易得到全局最优解。实验说明了我们算法的高效性，针对 6 段的长视频序列（约 10 万帧），后端单线程优化耗时 16 分钟，平均每秒 17.7 帧。作为对比，VisualSFM 在有 GPU 加速的情况下，后端优化耗时达 57 分钟，故我们算法的速度有 1 到 2 个数量级的提升。我们进一步比较结构，我们的优化结果可以完整的把子图拼在一起，VisualSFM 未能完整重现三维结构，ORB-SALM 的回路闭合效果较差（图2）。测试 KITTI 和 TUM 数据集发现，在复杂回路情况下的非连续特征跟踪和 Segment-based BA 方法的优势会体现出来。这部分工作属于 ENFT-SfM[2]。

图 2: 三维结构比较

上面的过程是一个离线的运动恢复结构（SfM），为了满足实时性，我们采用了并行的 Tracking 和 Mapping 框架处理。特征跟踪采用 ENFT（Efficient Non-Consecutive Feature Tracking），主要修改了非连续轨迹匹配，改为不再计算匹配矩阵。通过实验与 ORB-SLAM 比较，我们的回路闭合效果更好，一方面因为非连续特征轨迹匹配算法比基于 Bag-of-words 场景识别更有效，另一方面 Segment-based BA 优化方法比 Pose Graph 优化结合传统 BA 的方法更好，在复杂回路的场景下不容易陷入局部最优解。这部分工作我们叫 ENFT-SLAM。

4 动态场景下的 SLAM

动态场景下视觉 SLAM 存在一些问题，比如 Outlier 和遮挡，包括视角变化产生的遮挡和动态物体造成的遮挡。为了解决这个问题，我们提出了 RDSLAM 框架 [3]（图3）。

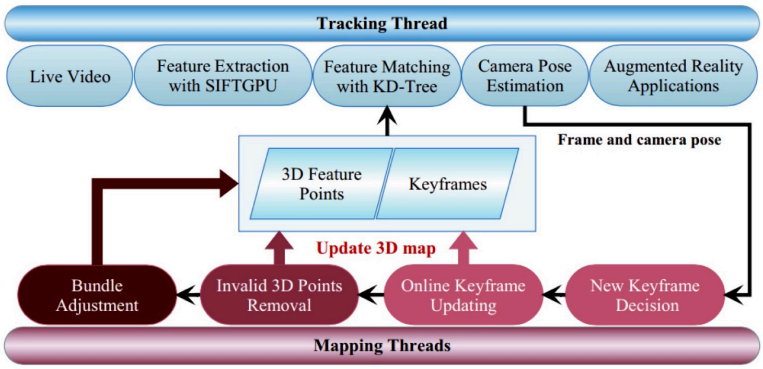

图 3: 针对动态场景的框架

我们的主要贡献是，通过 Mapping 线程在线检测三维点云的变化，如果某三维点云发生变化，算法会及时把它剔除掉，同样如果某关键帧的点云变化很多，那么将替换掉整个关键帧。新的策略可以及时检测点云的变化并更新，实现纯视觉的稳定跟踪。实验比较发现，在动态场景下，PTAM 的结果很差，因为传统 SLAM 方法假设场景是静止的。

5 用于移动端的 SLAM

SLAM 一般分为两类：基于 Keyframe-based SLAM 和基于 Filtering-based SLAM（图4），基于 Keyframe-based SLAM 采用 Keyframe BA 优化方法。SLAM 问题可以抽象为 Markov Random Field 问题，点之间的连线表示约束。基于卡尔曼滤波的方法每次只保存最新的相机状态，会增加复杂度。Keyframe BA 把非关键帧的约束删除掉，故它的求解速度快，稳定性高。但 Keyframe-based SLAM 要求关键帧之间的 baseline 比较大，这样的话会导致发生纯旋转时无法加入新的关键帧。基于 Keyframe-based SLAM 和基于 Filtering-based SLAM 都很难处理快速运动、运动模糊和弱纹理的情况。而 Visual-Inertial SLAM 可以用 IMU 传感器提高鲁棒性，比如有基于 Filtering-based 的框架有 MSCKF、Project Tango、ARCore、ARKit 等，采用非线性优化的框架有 OKVIS、VINS 等。

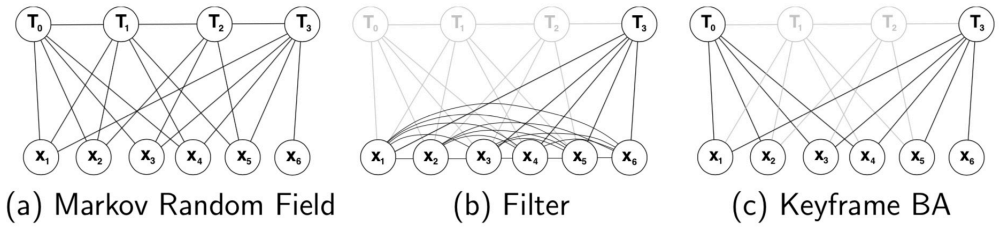

图 4: 优化求解的模型

我们考虑如何只用单目相机实现类似 Visual-Inertial SLAM 的鲁棒效果，提出了 RK-SLAM 框架 [4]。框架的第一部分是 Multi-Homography based 跟踪，我们采用了 FAST 角点，用 Global homography、Specific homography 和 Lcoal homography 提高跟踪的稳定性。我们借鉴 Visual-Inertial 方法，提出了基于 Sliding-window 的位姿优化。针对手机等移动设备，引入运动约束，即假设位移为零，通过相邻帧匹配点解算出旋转矩阵。当遭遇快速运动而导致特征点全部丢失时，我们用缩略图去对齐（类似于直接法跟踪），提高系统的稳定性。

实验比较发现，我们的方法比 ORB-SLAM、LSD-SLAM、PTAM 跟踪的成功率更高、初始化更快。我们还用 RKSLAM 框架设计了一款移动端的家具展示应用。RKSLAM 的另外一个优点是运行速度非常快，在 PC 上可以做到每秒 1 到 200 帧，在移动端能做到实时运行，比如在 iPhone 7 上接近每秒 100 帧。

6 基于 RGB-D 相机的稠密 SLAM

第五个工作是基于 RGB-D 相机和 RKSLAM 技术提出了 RKD-SLAM[5] 框架（图5）。

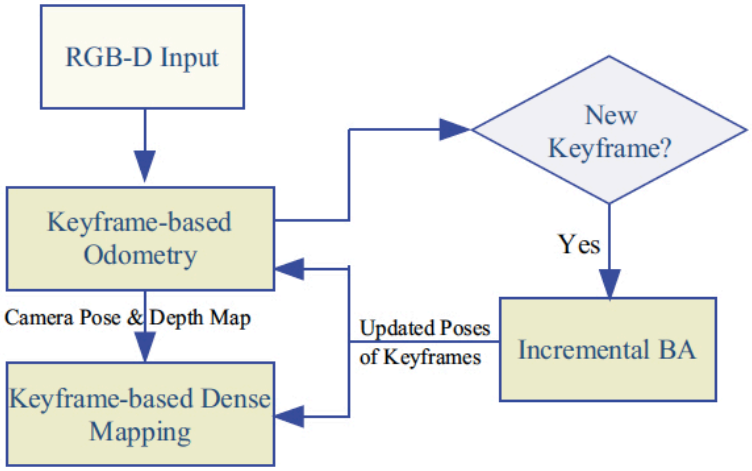

图 5: RKD-SLAM 框架

我们还提出了高效增量式 BA（EIBA）算法。BA 一般分为 Local BA、Global BA 和增量式 BA。Local BA 优化能力有限，Global BA 优化速度缓慢，而增量式 BA 利用了之前的计算结果，每次不需要完全重新计算，只需要更新矩阵，这样可以大大提高优化速度。实验比较发现，我们 EIBA 的计算速度比 iSAM2 有数量级的提升。我们还提出了基于 Keyframe-based 的深度表达和融合方法，在检测到回环后能对三维结构进行在线动态调整。

7 应用与趋势

视觉 SLAM 技术的应用领域包括：

- 三维重建
- 视频分割与编辑
- 增强现实
- 自动驾驶

例如我们可以用 SLAM 技术实现视频分割和编辑，我们先在三维场景上面进行编辑，编辑完后重新投影到二维视频，生成修改后的视频。我们还可以把一个视频序列抽取出来放到另一个视频序列中，但需要在三维场景上面实现，否则无法做到无缝融合。在自动驾驶领域，把视觉 SLAM 和道路线检查结合起来，能实现更加鲁棒的定位。

最后谈谈视觉 SLAM 的发展趋势，第一个趋势是如何减少特征跟踪的问题，目前有结合 Edge Tracking、Direct Tracking 和 Learning based 的方法。第二个趋势是如何实现稠密的三维重建，分为基于视觉和基于深度相机的方法，但都需要考虑简化表达。第三个趋势是多传感器融合，所有传感器都有自身的优点和缺点，我们最好有效地融合各传感器数据，进行优势互补。

我们课题组目前和未来的工作包括：无人机和机器人的协同、稠密 SLAM、场景分析和理解、SLAM 在 AR/VR/机器人/自动驾驶领域的应用。

参考文献

[1] G. Zhang, Z. Dong, J. Jia, T.-T. Wong, and H. Bao, "Efficient non-consecutive feature tracking for structure-from-motion," in *European Conference on Computer Vision*, pp. 422–435, Springer, 2010.

[2] G. Zhang, H. Liu, Z. Dong, J. Jia, T.-T. Wong, and H. Bao, "Efficient non-consecutive feature tracking for robust structure-from-motion," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5957–5970, 2016.

[3] W. Tan, H. Liu, Z. Dong, G. Zhang, and H. Bao, "Robust monocular slam in dynamic environments," in *Mixed and Augmented Reality (ISMAR), 2013 IEEE International Symposium on*, pp. 209–218, IEEE, 2013.

[4] H. Liu, G. Zhang, and H. Bao, "Robust keyframe-based monocular slam for augmented reality," in *Mixed and Augmented Reality (ISMAR), 2016 IEEE International Symposium on*, pp. 1–10, IEEE, 2016.

[5] H. Liu, C. Li, G. Chen, G. Zhang, M. Kaess, and H. Bao, "Robust keyframe-based dense slam with an rgb-d camera," *arXiv preprint arXiv:1711.05166*, 2017.