基于单目视觉的同时定位与建图算法研究综述*

朱 凯, 刘华峰, 夏青元
(南京理工大学 计算机科学与工程学院, 南京 210094)

摘 要: 与传统基于激光传感器的同时定位与建图(SLAM)方法相比,基于图像视觉传感器 SLAM 方法能廉价地获得更多环境信息,帮助移动机器人提高智能性。不同于用带深度信息的 3D 传感器研究 SLAM 问题,单目视觉 SLAM 算法用二维图像序列在线构建三维环境地图并实现实时定位。针对多种单目视觉 SLAM 算法进行对比研究,分析了近 10 年来流行的单目视觉定位算法的主要思路及其分类,指出基于优化方法正取代滤波器方法成为主流方法。从初始化、位姿估计、地图创建、闭环检测等功能组件的角度分别总结了目前流行的各种单目视觉 SLAM 或 Odometry 系统的工作原理和关键技术,阐述它们的工作过程和性能特点;总结了近年最新单目视觉定位算法的设计思路,最后概括指出本领域的研究热点与发展趋势。

关键词: 单目相机; 视觉定位; 视觉里程计; 视觉同时定位与建图
中图分类号: TP391.41 **文献标志码**: A **文章编号**: 1001-3695(2018)01-0001-06
doi:10.3969/j.issn.1001-3695.2018.01.001

Survey on monocular visual SLAM algorithms

Zhu Kai, Liu Huafeng, Xia Qingyuan
(*School of Computer Science & Engineering*, *Nanjing University of Science & Technology*, *Nanjing* 210094, *China*)

Abstract: Compared with traditional laser scanner based SLAM method, camera based SLAM algorithms outperformed in cost as well as information, and could make mobile robot smarter. Instead of using 3D sensors, monocular visual SLAM utilized 2D image sequence to reconstruct 3D map and performed real-time localization. This paper aimed at providing a survey on monocular visual SLAM algorithms. It studied the most popular visual SLAM algorithms in the last decade and discussed the main principle and their classification. Then it pointed out that the optimization method would be the mainstream. It summarized the principle and practice of the state of the art visual SLAM or visual Odometry systems from the angle of initialization, pose estimation, map generation and loop closure. According to above, this paper also elaborated the design scheme and performance of the state of art monocular SLAM algorithms. At last, it concluded with the viewpoint for the research trend.

Key words: monocular camera; visual localization; visual odometry; visual SLAM(simultaneous localization and mapping)

0 引言

随着计算机视觉技术的发展和硬件性能的快速增长,基于视觉的同时定位与地图创建技术及其应用在过去 10 年里进展丰硕。民用硬件方面,高性能嵌入式计算设备的大量出现和廉价精确成像设备的小型化,已经可以支持在廉价小巧如手机和平板电脑等手持设备上实现虚拟现实(virtual reality)和增强现实(augment reality)应用。在工业界,随着小型无人机的普及和自动驾驶研究的兴起,视觉同时定位与地图创建和视觉里程计作为未来智能机器人发展的基础技术也越来越激起学术界与工程界的研究热情。这些新兴理论和技术的基础就是计算机视觉研究中的从运动中恢复结构(structure from motion)和视觉三维重建(visual 3D reconstruction)等课题。机器人研究领域的研究者称其为基于视觉传感器的同时定位与地图创建(visual simultaneous localization and mapping,visual SLAM)或者基于视觉传感器的里程计(visual odometry)[1]。本文所指的视觉定位与地图创建技术包括常见的 visual SLAM 和 visual Odometry 算法。前者研究同步定位与三维地图生成,而后者关注实时向机器人提供实时位姿估计,两者的目的均是帮助机器人实现在各种 GPS 难以工作的环境进行实时在线定位、运动位姿估计和三维地图创建。

机器人视觉定位有两种不同的实现思路:

a)基于图像匹配的定位方法(image-based localization)。此方法将采集到的图像和采集时刻的位姿信息预先处理成为图像数据库,而在定位阶段将实时采集到的图像放入数据库中搜索最匹配的图像从而估计当前相机的位姿。图像匹配定位方法中使用的图像数据库需要离线处理以便检索匹配,定位与地图创建是分开完成的。

b)摈弃了方法 a)依赖预先创建地图的弱点,将地图创建工作与定位工作分别实现于两个并发线程中,同时完成地图的创建与相机位姿的定位。该方法一般采用视觉里程计技术或 SLAM 过程降低位姿估计的累计漂移,同时在后端采用闭环检

收稿日期: 2016-12-27; **修回日期**: 2017-02-23 **基金项目**: 国家自然科学基金资助项目(61403202);中国博士后科学基金面上资助项目(2014M561654)

作者简介:朱凯(1977-),男,江苏南京人,助理实验师,硕士,主要研究方向为智能机器人系统、机器人自主导航(kevinsd@njust.edu.cn);刘华峰(1988-),男,湖北随州人,博士研究生,主要研究方向为智能机器人系统、机器人自主导航;夏青元(1980-),男,江苏盐城人,讲师,博士(后),主要研究方向为智能机器人环境理解与导航技术、无人飞行器仿真与控制.

测等手段来提高定位与地图的精度和一致性。

本文综述的对象属于方法 b),这种方法具备在未知环境下工作的能力,定位精度不依赖预先创建的地图。

视觉 SLAM 算法有多种传感器方案,目前广泛使用的是 RGB-D 深度传感器、立体视觉相机和单目相机等。其中,新出现的 RGB-D 传感器如微软公司的 Kinect/Kinect v2、华硕公司的 Xtion 和 Intel 公司的 RealSense 传感器因廉价和体积小巧较为流行。RGB-D 相机的优势在于可以直接获得空间深度信息,但现有 RGB-D 传感器的视场小、深度测距范围一般限于 0.5 ~5 m 且室外抗干扰能力不足,这些局限导致它们往往工作在室内环境。双目立体视觉相机既可以在静止状态也可以在运动状态估计深度,室内室外都能工作,但是立体视觉的计算量大且要求精确而复杂的标定。本文采用的是单目视觉传感器方案的 visual SLAM 算法,但其他传感器方案可以与单目视觉方案相互借鉴,多数基于单目视觉的算法,最终可以实现出不同传感器方案的版本,并能发挥出不同传感器方案的长处[2~4]。

Visual SLAM 和 visual Odometry 近 10 年来研究成果众多,一些学者已经对这一领域的研究状况作出了综述。Strasdat 等人在文献[5]详细地总结了基于滤波器法、马尔可夫随机场法和最优化方法在立体视觉和单目视觉传感器方案上的应用,并指出最优化方法较滤波器方法和其他方法有一定优势。Fuentes-Pacheco 等人在 2012 年也对视觉定位与地图创建算法作出综述。Yousif 等人于 2015 在文献[6]中对 visual SLAM 和 visual Odometry 进行了较为全面的综述,内容涵盖了基于滤波器、非滤波器的各种方法及 RGB-D 传感器下的相关主题。本文主要跟踪了近 10 年来的单目视觉 SLAM/Odometry 的研究,分类阐述了不同技术方案的特点,同时讨论了影响较大算法的设计思路和实现。

1 单目视觉 SLAM 概述

1.1 单目视觉 SLAM 技术的分类

1.1.1 滤波器与非滤波器方法

以系统建模的方式分类,单目视觉定位与地图创建有两种主流类别,即基于滤波器的方法和基于优化的方法。其中前者出现得较早,是激光传感器 SLAM 的常用方法;后者出现较晚,并已经逐渐发展成为一种主流方法。

Dellaert 在文献[7]中指出 SLAM 问题可以用图状模型描述。定义每次观测的位置为 L_i,环境中的静态路标为 P_j,观测位置 L_i 上观测到的路标 P_j 表示为 u_{ij}。观测位置 $L \in \mathcal{L}$ 和观测 u 随着相机的移动不断增加,新探索到的环境也会引入新的 $P \in \mathcal{P}$。以 L 和 P 为顶点、u 为边,上述元素构成如图 1 所示的图模型。

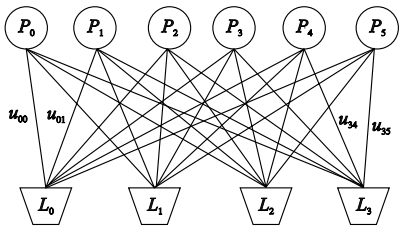

图1 定位与地图创建问题的图模型表示

在滤波器方法中,计算过程中一般包含未来会用到的路标 P 和当前的观测位置 L。例如,在图 1 中位姿节点只有 L_3 而路标节点是路标集合的一个子集。因为不再考虑每一个历史观测位置信息,滤波器算法必须保存和更新路标之间的关系。路标的数量随着运动路径的延伸不断增多,算法中更新联合概率分布的过程将消耗越来越多的计算和存储资源。所以,基于滤波器的算法必须严格控制路标的规模,防止计算量快速扩大而使效率急剧下降。

在优化方法中,算法采用合理的方式抽取并保留一些历史观测位置和路标。这些被保留的观测位置和观测路标合称为一个关键帧(key frame)。地图中的路标一般从关键帧提取和计算。这一策略的思路是:要尽可能地保证图模型维持稀疏,同时关键帧的高质量路标参与算法计算以消除采样观测位姿带来的损失,而低质量的路标或待选的关键帧可以被剔除以缩减计算量同时保证定位和地图精度。

Julier 等人[8]证明,基于 EKF 的方法在长时间工作后会出现不一致。为克服不一致缺陷并改善传统滤波器方法在大地图场景中的计算复杂度过高的问题,Clemente 和 Piniés 等人[9,10]提出了一些基于子图(sub map)的改进算法。Lim 等人[11]将一策略引入到最优化方法中,取得了较好的成果;也有一些基于最优化理论的方法提出,部分路标可以在计算过程中不予考虑从而平衡计算量与精度;也有一些算法在运行的不同阶段采用不同方法以增加灵活性。

Strasdat 等人在文献[5]中指出,基于滤波器的方法与基于最优化的方法在精度上并没有明显差异。在定位和地图创建过程中,影响精度提高的因素是有效观测数量的多寡,单纯提高关键帧的数量对提高精度作用并不总是有效的。但是从计算量角度考虑,最优化方法的计算量随观测量呈线性阶增长,而滤波器方法随观测量呈立方阶或平方阶增长。所以在相同精度的要求下,采用最优化方法构架单目视觉定位系统理论上更为高效。未来单目视觉定位与地图创建技术的研究关注点将转到基于最优化方法及其改进。

1.1.2 直接方法与非直接方法

从使用数据的方式来看,单目视觉 SLAM 可以分为直接方法和非直接方法,两者区别在于如何利用传感器信息。

直接方法也被称做稠密方法,其思路是利用相邻帧图像中每一个像素的亮度信息差来估计相机的运动量。通常这一方法下的运动估计问题会被建模成一个图像对齐问题,进而使用 Lucas 等人[12]提出的 FAIA 方法求解。为了缩减计算量,FCIA、ICIA、IAIA 等改进方法也在各种研究中得以引入[13]。这种方法的优点是在环境纹理匮乏时算法还能够保持一定的鲁棒性;而弱点是该算法难以在光照变化剧烈的场景工作,同时计算量偏大。

非直接方法使用图像特征代替像素亮度以克服直接法的弱点。常用的特征包括但不限于 SIFT、SURF、BRIEF、FAST、ORB 等。Hartmann 等人[14]将常见的图像特征在视觉定位与地图创建算法中的应用作了详尽的对比分析。非直接方法将同时定位与地图创建问题建模成数据关联问题,计算效率较直接法有所提高。但是这类方法非常依赖特征抽取的结果,建模出的地图稀疏且缺乏直观性。

上述两种方法也可以混合使用。例如,SVO 中就使用直

接法进行特征的关联,同时使用特征来改善定位精度,这一思路使得 SVO 可以在计算能力受限的场景下获得良好的实时定位效果,实验证明可以在小型无人机上实时工作[15]。

1.2　单目视觉 SLAM 的发展近况和特点

自 2002 年起,视觉 SLAM 算法的研究开始引起关注,但重大的进展是从 2007 年 Klein 等人[16]提出的并行跟踪与建图(parallel tracking and mapping,PTAM)算法开始的。他们提出一个可以将地图创建和位姿估计过程放在两个并发线程中运行的非滤波器算法框架,并取得了实时运行效果。其后,本领域的大部分研究成果都延续了 PTAM 框架中的思想。表 1、2 按照滤波器法和优化方法分类,分别整理了自 2007 年起一些重要的视觉同步定位与地图创建算法和视觉里程计算法,算法采用的是直接法或是非直接法也反映在表中。

表 1　重要 visual SLAM/Odometry 方法文献(滤波器方法)

年份	文献名称	类别
2007	MonoSLAM	I
2008	Square root UKF for visual MonoSLAM	I
2008	Efficient view-based SLAM using visual loop closures	I
2009	Towards a robust visual SLAM approach: addressing the challenge of life-long operation	I
2010	On combining visual SLAM and visual Odometry	I
2010	Monocular SLAM with locally planar landmarks via geometric Rao-Blackwellized particle filtering on Lie groups	I
2013	Monocular SLAM for indoor aerial vehicles	I
2014	Real-time camera tracking using a particle filter combined with unscented Kalman filter	I
2015	StructSLAM: visual SLAM with building structure lines	I

表 1 中的算法均采用滤波器方法(如 EKF、UKF、PF 等),I 表示非直接法。

表 2　重要 visual SLAM/Odometry 方法文献(优化方法)

年份	文献名称	类别
2007	Parallel tracking and mapping	I
2008	An efficient direct approach to visual SLAM	D
2010	Live dense reconstruction with single moving camera	H
2010	Scale drift-aware large scale monocular SLAM	I
2011	Online environment mapping	I
2011	Omnidirectional dense large scale mapping and navigation based on meaningful triangulation	D
2011	Dense tracking and mapping	D
2013	Robust monocular SLAM in dynamic environments	I
2014	Semi-direct visual odometry	H
2014	DT-SLAM: deferred triangulation for robust SLAM	I
2014	Real-time 6-DOF monocular visual SLAM in a large scale environment	I
2014	LSD-SLAM: direct monocular SLAM	D
2015	Robust large scale monocular visual SLAM	I
2015	DPPTAM: dense piecewise planar tracking and mapping from a monocular sequence	D
2015	ORB SLAM:a versatile and accurate monocular SLAM system	I
2016	Direct sparse odometry	I

表 2 中的算法均采用最优化方法(光束平差),I 表示非直接法,D 表示直接法,H 表示混合方法。

综合表 1 和 2 可以发现,过去 10 年间 visual SLAM 的研究开始从滤波器方法主导逐渐向优化方法发展,直接法和非直接法在优化方法框架内都受到研究者的重视。一些最近极受关注的算法中,采用优化方法逐渐成为主流方法。

2　单目视觉 SLAM 算法设计思路

当前,单目视觉 SLAM 算法一般分为初始化、位姿估计、地图生成、闭环检测等基本功能组件。其中,初始定位组件负责创建并初始化初始位姿和三维初始地图;位姿估计是单目视觉 SLAM 系统的核心步骤,其工作内容是已知前一帧位姿并利用相邻帧之间的关系估计当前帧的位姿;地图生成模块负责生成和维护全局地图;闭环检测负责通过检测“闭环”而减少全局地图“漂移”现象。本章通过分解单目视觉 SLAM 算法的基本结构,从初始定位、位姿估计、地图生成和闭环检测四个环节对比分析近年来重要单目视觉 SLAM 算法的设计思路。

2.1　初始化

早期算法的初始化过程需要人工辅助完成[17],随着新方法的提出,该步骤开始逐渐降低对手工输入的依赖。从策略角度分类,初始化方法分为本征矩阵分解、单应矩阵分解和随机深度初始化等策略。

PTAM 使用了 Faugeras 等人[18]提出的单应矩阵分解方法。其初始化无须输入深度值,但要求人工指定第一和第二帧关键帧。该算法在初始化阶段从第一个关键帧中提取 FAST 特征点[19],并持续跟踪匹配到第二个关键帧。其间,相机要求平行于假设平面运动。采用单应矩阵分解初始化时,被观测场景假设是一个平面,当第二个关键帧确定后,初始化算法采用随机采样极大似然估计(MLESAC)迭代并求出两个关键帧之间的位姿变化[20]。PTAM 的初始化过程需要满足若干基本假设,这导致初始化过程容易受到手动步骤的影响而失败。SVO 采用了与 PTAM 同样的单应矩阵分解方法作为初始化策略,SVO 改进了 PTAM 的初始化过程,不再要求手工输入。其第一个关键帧是算法读入的第一个有效帧,而第二个关键帧由算法自动识别。与 PTAM 一样,SVO 在第一个与第二个关键帧之间跟踪匹配 FAST 特征,但跟踪匹配算法采用 Lucas-Kanade tracking 完成[21]。

DT-SLAM 没有在算法中设计一个专门的模块用于初始化,而是将此过程集成在位姿估计组件使其估计单应性矩阵。DT-SLAM 不要求 PTAM 与 SVO 一样的假设。

ORB-SLAM 则混合使用了上述两种方法来完成初始化,该方法无须手工选择关键帧,也不需要预先假设场景是否为平面。ORB-SLAM 的初始化过程是:a)找到两帧图像 ORB 特征点的初始对应关系;b)同时计算单应矩阵 $\boldsymbol{H}_{cr}$ 和基础矩阵 $\boldsymbol{F}_{cr}$ 并为其分别评分,通过比较评分确定选择单应矩阵还是基础矩阵作为模型;c)采用 SFM 方法获得最终结果。由于 ORB-SLAM 采用了混合方法,其环境适应性得到很好的平衡,无论在平面场景、低视差场景还是在非平面场景、高视差场景,它都能自动选择较优的模型来完成初始化工作。

LSD-SLAM 中使用随机深度初始化策略以类似于滤波器方法的思路来完成初始化。Engel 等人在 LSD-SLAM 中将图像中的像素以随机的深度初始化,并随后利用新产生数据不断迭

代优化直至收敛。这种方法在一定程度上克服了上述两种方法对环境和相机运动敏感的缺点,但其计算量偏大且耗时可能会较长。Concha 等人在 DP-PTAM 也采用了与 Engel 等人在 LSD-SLAM 中使用的同一初始化策略。

2.2 位姿估计

位姿估计是视觉定位系统的核心环节。位姿估计的任务是利用先验位姿估计和帧间关系计算出相机的位姿或位姿变化量。位姿估计过程的计算量比初始的数据关联小,但对系统实时性影响大,这一过程也称为跟踪(tracking)。

恒速运动模型是生成先验位姿估计的常规方法,该模型假设相机的运动轨迹平滑且运动速度均匀。PTAM、DT-SLAM、ORB-SLAM 和 DP-PTAM 均假设恒速运动模型,通过上两帧的运动量来估计先验位姿;而 SVO 和 LSD-SLAM 则使用上一帧的位置直接作为先验位姿以适应相机运动轨迹不光滑的情况,代价是图像采集帧率和系统运行速度需足够高以满足连续两帧之间运动量足够小的假设。因此在不同的机器人本体上,采用合适的运动模型假设可以优化先验位姿估计。获得先验位姿后,位姿估计组件通过自身策略优化重投影误差函数以确定相机位姿。

PTAM 用 $SE(3)$ 表示位姿变换,应用李代数使 $SE(3)$ 可以映射成一个 6 参数的 $S\xi(3)$ 以提高随后最优化过程的工作效率。PTAM 采用恒速运动模型预生成一个先验的位姿估计,并采用一种以 Tukey-biweight 为目标函数的鲁棒的重投影残差最小化算法获得位姿结果[22]。具体过程是:首先系统利用运动模型为新获得的图像确定一个先验相机位姿,然后利用先验相机位姿将地图上的点投影到新获得的图像中,随后在最粗尺度下选取少量特征点优化相机位姿,最后将较多数量的特征点重投影到当前图像计算最终相机位姿。在整个特征点跟踪的过程中,算法可能出现跟踪无效的情况。PTAM 算法在每一帧中都会计算并监视跟踪性能指标,若跟踪性能指标较低,算法照常运行并停止创建关键帧;若指标过低,则启动故障恢复。

SVO 与 PTAM 表示位姿变换的方式相同。SVO 没有假设恒速运动模型,其先验位姿估计直接设置为上一帧的位姿估计结果。SVO 位姿估计的整体流程是:利用直接法计算位姿变化值的初始估计,再通过最优化特征点非线性重投影误差函数来改善该初始估计精确度。

a)受到相机微小运动的影响,特征点投影到图像上会有一定位置变化,算法采用稀疏的图像对齐过程来最小化特征块的光度差,以获得一个位姿变化的初步估计。由于 SVO 系统处理速度较高,相邻帧采集到的图像特征块形变很小,所以该步骤中忽略了形变。

b)在步骤 a)得到的位姿基础上将更多地图中的点投影到图像,对每个当前帧可见的地图点,算法找到观测到该点角度最小的关键帧,分别优化得到新的位姿。优化过程需要考虑到仿射变换,但不考虑极线约束。

c)利用步骤 b)得到的地图点与投影点的关系,算法运行光束平差优化出位姿信息。

SVO 允许在算法中设置忽略步骤 c)的光束平差最优化步骤以换取更高的运行效率。

与 PTAM 不同,ORB-SLAM 抽取 FAST 特征并生成 ORB 描述子。ORB-SLAM 采用恒速运动模型生成先验位姿估计,并利用该先验估计跟踪匹配上一帧的地图点。若上述先验位姿与实际运动不符而导致匹配效果不佳,算法会扩大搜索范围以达到匹配目的。最后,算法利用上面得到的数据关联信息通过光束平差确定本帧对应的相机位姿估计。ORB-SLAM 的特别之处在于其采用了局部地图的方法使得它能在大范围环境工作。在全局地图中的路标需要投影到当前帧,而构建局部地图可以精选出有效的路标,最终实现减少计算复杂度的同时保持位姿估计的性能。

LSD-SLAM 采用 $SE(3)$ 表示位姿变换,并在位姿估计阶段使用上一帧相机位姿作为先验位姿估计,最终求得本次输入与活动关键帧之间的位姿关系。由于 LSD-SLAM 采用纯直接法,算法当前输入帧的相机位姿采用 Gauss-Newton 法迭代优化当前帧与活动关键帧之间的光度残差得到[23]。DP-PTAM 采用与 LSD-SLAM 相似的策略,但用于估计先验位姿的运动模型改为恒速运动模型,并在该模型失效时应用 SVO 中使用的策略,即直接使用上一帧位姿作为先验位姿估计。同时,DP-PTAM 在优化最终位姿时使用了与 PTAM 基本相同的对应算法。

2.3 地图生成

地图生成是与位姿估计同时完成的工作,即所谓的同时定位与地图创建机制。算法创建出的地图可分为使用非直接法创建的稀疏地图和使用直接法创建的稠密地图。一些工作在大场景的算法中也使用拓扑地图(topological maps)。拓扑图不记录尺度、距离等度量信息,但可以通过给定对应的度量信息来完成拓扑地图到度量地图的转换。拓扑图常与其他类型的地图同时使用。Konolige[24] 提出的混合地图即结合了拓扑地图与度量地图的优势。本节内容主要讨论度量地图的生成。

单目视觉定位系统创建地图的基本原理是将图像中的 2D 信息通过一定方法(三角测量法、粒子滤波器等)转换成为 3D 的地标(land mark),然后通过一定方法维护地图内容,最终生成一致性较好的全局地图。使用滤波器方法时,算法假设 3D 地标点位姿满足一个分布假设,随着新的视图不断加入到 3D 地标的更新过程,地标位姿分布的方差将会收敛到一定阈值,这时就可以确定地标的位姿信息。使用优化方法时,在地图创建阶段相机位姿变化量已知,对同一个地标在不同位置的观测也是已知的,将地标重投影到至少两帧视图并将重投影残差最小化即可获得该地标的 3D 信息。上述两种方法都依赖于相机的运动存在有效的平移量以构成视差基线,否则就难以估计出地标的 3D 信息。为处理视差基线不足的情况,一些如 DT-SLAM 的方法会在运行上述方法前先检测并单独处理。

随着地图中的路标信息不断增加,地图需要运行维护算法不断地自我维护以剔除异常值并抑制漂移。局部和全局的光束平差是常用方法,它们利用非线性最优化过程提高定位与地图质量。许多算法仅仅在关键帧上运行使用光束平差计算以节省计算资源。另外一种常用方法是图优化技术(pose graph optimization)。图优化技术优化对象是位姿图,在位姿图较稀疏的情况下其速度明显优于在每帧数据上作光束平差。

PTAM 算法中,地图生成过程是由关键帧的选择驱动的。关键帧的选取要满足特征跟踪效果良好、具备一定间隔、离最近关键点距离大于一定值(满足基线条件)等要求。每个关键

帧采用位姿估计阶段的特征点及其位姿为初始值。当前关键帧可见的已知点被重投影到当前关键帧上以建立数据关联。随后,通过三角测量法从最近关键帧与当前关键帧得到新加入地图地标点的深度信息。PTAM 的地图生成模块也负责维护地图,它使用局部光束平差(local bundle adjustment)来优化局部地图,同时使用全局光束平差(global bundle adjustment)来促进全局地图收敛。

SVO 采用基于概率的方法估计三维地标点的深度信息。已知一段序列的图像及其对应观测位姿,SVO 使用 Vogiatzis 等人[25]的方法引入一个基于贝叶斯理论的框架更新深度的概率分布。当深度的高斯分布方差足够小时,算法认为深度值计算已经收敛,随即利用逆投影函数恢复出三维地标坐标。为了控制计算量同时提高算法效率,SVO 的关键帧数量是固定的,当达到关键帧数量上限时,系统会选择移除距离现在位置远的关键帧。SVO 选择关键帧的依据为:当某一帧场景相对于所有关键帧的场景平均深度超出 12% 时,则被选为关键帧。

ORB-SLAM 的地图生成模块运行关键帧抽取、地图点三角测量、关键帧与地图点的维护。ORB-SLAM 使用局部光束平差优化局部地图,并使用 Co-visibility 和 Essential 图维护数据关联信息。其中 Co-visibility 图是一个无向带权图[26],主要用来存储所有关键帧之间的位姿关系。图中各节点表示关键帧,而图中的边则表示两个关键帧之间有效的共同地标点(数量大于 15),其权重表示有效共同点数量。Co-visibility 图是稠密的,这在优化地图时引入巨大计算量。算法使用 Essential 图加速计算。ORB-SLAM 利用生成树算法从 Co-visibility 图中构建了一个包含了所有的关键帧的子图,子图中保留了权重强的边。这样,ORB-SLAM 就可以使用上述两种图精确并高效地创建与维护地图。为了防止地图在生成过程中受到异常值和冗余关键帧的影响,ORB-SLAM 在生成地图的同时不断地进行地标点和关键帧优选,这样可以剔除不稳定地标点和减少冗余关键帧。

LSD-SLAM 生成稠密地图,算法监视当前相机位姿角度和距离是否满足抽取关键帧的条件。由于 LSD-SLAM 属于纯粹的直接法,该算法的关键帧抽取强度较一般非直接法大。如果当前输入帧不是关键帧,这帧数据就用来优化最新选取的关键帧[23];如果当前场景被确定为关键帧,则从前关键帧中投影点以初始化本关键帧深度图,并执行空间归一化。LSD-SLAM 通过在位姿估计阶段记录所有成功匹配的像素以调整该像素是否是异常值的概率,最终剔除异常值,算法也在后台持续通过位姿图优化技术进行优化[27]。

2.4　闭环检测

视觉定位与地图创建算法总体上是一个滤波或优化过程。虽然算法中有很多机制用于提高位姿估计和地图创建环节中的精确性,但是系统在长期工作后依然无法避免出现漂移。为此,SLAM 算法中引入闭环检测功能来帮助减少累计漂移和误差。

LSD-SLAM 使用了 Glover 等人[28]在 2012 年提出的基于 FABMAP 的闭环检测算法。LSD-SLAM 在处理关键帧时,在最近 10 个关键帧中搜索得到一个闭环过程并得到两个端,随后利用位姿图优化来最小化相似度差异。ORB-SLAM 是通过“词袋模型”完成全局位置识别和闭环检测的[29]。算法首先在 Co-visibility 图上计算词袋向量并计算相似性提取候选闭环场景。由于单目视觉 SLAM 有三个平移、三个旋转和一个尺度共七个自由度,所以算法必须计算出从当前关键帧到闭环关键帧之间的相似变换(similarity transform)以得出两帧之间出现的累计误差,最后运行闭环修正算法优化地图。

3　总结与发展展望

上文将视觉定位与地图创建算法主要分为四大模块进行阐述,但是具体的算法中还可能包括故障恢复、参数调整、传感器标定等小功能组件。总结此类算法的设计思路可以发现,算法设计的重点在于如何在保证精度和鲁棒性的基础上尽可能地缩小计算开销、尽可能地利用场景中的各种信息防止定位和地图劣化。

当前,视觉 SLAM 发展趋势主要涵盖以下三方面:a)大规模复杂环境的适应性和鲁棒性的提高;b)提高计算效率并同时优化精度;c)研究多机器人和多传感器协同 SLAM。大规模复杂环境的适应性和鲁棒性是视觉 SLAM 系统实用化的必然要求,其趋势在于从当前的室内或简单室外环境转向大范围的、动态的、复杂的室外环境研究。如文献[30]开展了适应动态环境的相关工作,文献[31]利用双目视觉实现了原来运行在单目相机上的算法,文献[32,33]详细论述了抗光照变化和适应室外环境的算法,一些低纹理环境下的视觉 SLAM 技术、无须初始化的视觉 SLAM 技术也正在发展。此外,模式识别和机器学习理论的发展也正在使得闭环检测更加强大实用。在提高计算效率方面,一些基于嵌入式、低功耗设备上 SLAM 技术的研究也逐渐受到技术企业的关注,例如 VR 和 AR 头盔乃至手机设备都能提供较好的使用体验,同时摆脱对笨重计算设备的依赖。在多机器人和多传感器系统 SLAM 方面,多机器人协同建图、基于多传感器融合的 SLAM 系统、visual-inertial SLAM 等研究也逐渐兴起。KIT 和 CMU 等大学将自有地面无人车辆采集的数据打包发布供研究者测试和对比各自算法在真实环境中的效果[34],这些数据不仅包括单目相机的数据,还有如惯性导航、GPS、双目视觉、激光雷达数据等数据。因此,未来基于多传感器融合的 visual SLAM 算法将是一个热门的研究方向。

上述发展趋势将把视觉 SLAM 系统从实验室引入生活和工业应用场景中,使得机器人越来越智能化。

4　结束语

本文分析了典型的视觉定位与地图创建算法的各个基本组件,并以近年来重要的算法为例对比了不同算法在不同模块上的设计思路。综合近年来的重要成果不难发现,视觉定位与地图创建算法的发展正在朝着越来越注重准确性、实时运行和具备较强的故障恢复能力等方向发展,这一趋势也为此类算法在各种机器人系统实用上奠定了基础。

参考文献:

[1] Nistér D, Naroditsky O, Bergen J. Visual odometry[C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2004: 652-659.

[2] Engel J, Schops T, Cremers D, *et al*. LSD-SLAM: large-scale direct monocular SLAM[C]//Proc of European Conference on Computer

Vision. 2014:834-849.

[3] Engel J, Stückler J, Cremers D. Large-scale direct SLAM with stereo cameras[C]//Proc of IEEE/RSJ International Conference on Intelligent Robots and Systems. 2015.

[4] Caruso D, Engel J, Cremers D. Large-scale direct SLAM for omnidirectional cameras[C]//Proc of IEEE/RSJ International Conference on Intelligent Robots and Systems. 2015:141-148.

[5] Strasdat H, Montiel J M M, Davison A J. Visual SLAM: why filter? [J]. Image and Vision Computing, 2012, 30(2):65-77.

[6] Yousif K, Bab-Hadiashar A, Hoseinnezhad R. An overview to visual Odometry and visual SLAM: applications to mobile robotics[J]. Intelligent Industrial Systems, 2015, 1(4):289-311.

[7] Dellaert F. Square root SAM[C]//Robotics: Science and Systems. Cambridge: Massachusetts Institute of Technology, 2005:177-184.

[8] Julier S J, Uhlmann J K. A counter example to the theory of simultaneous localization and map building[C]//Proc of IEEE International Conference on Robotics & Automation. 2001:4238-4243.

[9] Clemente L A, Davison A J, Reid I D, *et al*. Mapping large loops with a single hand-held camera[C]//Robotics: Science and Systems III. Georgia: Georgia Institute of Technology, 2007:297-304.

[10] Piniés P, Tardós J D. Large-scale SLAM building conditionally independent local maps: application to monocular vision[J]. IEEE Trans on Robotics, 2008, 24(5):1094-1106.

[11] Lim J, Frahm J M, Pollefeys M. Online environment mapping[C]// Proc of IEEE Conference on Computer Vision & Pattern Recognition. 2011:3489-3496.

[12] Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision[C]//Proc of International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc, 1981:285-289.

[13] Baker S, Matthews I. Lucas-kanade 20 years on: a unifying framework[J]. International Journal of Computer Vision, 2004, 56(3): 221-255.

[14] Hartmann J, Klüssendorff J H, Maehle E. A comparison of feature descriptors for visual SLAM[C]//Proc of European Conference on Mobile Robots. 2013:56-61.

[15] Forster C, Pizzoli M, Scaramuzza D. SVO: fast semi-direct monocular visual odometry[C]//Proc of IEEE International Conference on Robotics and Automation. 2014:15-22.

[16] Klein G, Murray D. Parallel tracking and mapping for small AR workspaces[C]//Proc of IEEE and ACM International Symposium on Mixed and Augmented Reality. Washington DC: IEEE Computer Society, 2007:1-10.

[17] Davison A J, Reid I D, Molton N D, *et al*. MonoSLAM: real-time single camera SLAM[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(6):1052-1067.

[18] Faugeras O D, Lustman F. Motion and structure from motion in a piecewise planar environment[J]. International Journal of Pattern Recognition and Artificial Intelligence, 1988, 2(3):485-508.

[19] Rosten E, Drummond T. Machine learning for high-speed corner detection[C]//Proc of European Conference on Computer Vision. [S. l.]: Springer-Verlag, 2006:430-443.

[20] Torr P H S, Zisserman A. MLESAC: a new robust estimator with application to estimating image geometry[J]. Computer Vision and Image Understanding, 2000, 78(1):138-156.

[21] Tomasi C, Kanade T. Detection and tracking of point features[D]. Pittsburgh: School of Computer Science, Carnegie Mellon University, 1991.

[22] Maronna R A, Martin D R, Yohai V J. Robust statistics: theory and methods[M]. Hoboken: Wiley, 2006.

[23] Engel J, Sturm J, Cremers D. Semi-dense visual odometry for a monocular camera[C]//Proc of IEEE International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2013: 1449-1456.

[24] Konolige K. Sparse sparse bundle adjustment[C]//Proc of British Machine Vision Conference. 2010:1-11.

[25] Vogiatzis G, Hernández C. Video-based, real-time multi-view stereo [J]. Image and Vision Computing, 2011, 29(7):434-441.

[26] Strasdat H, Davison A J, Montiel J M M, *et al*. Double window optimisation for constant time visual SLAM[C]//Proc of International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2011:2352-2359.

[27] Kümmerle R, Grisetti G, Strasdat H, *et al*. g2o: a general framework for graph optimization[C]//Proc of IEEE International Conference on Robotics and Automation. 2010.

[28] Glover A, Maddern W, Warren M, *et al*. OpenFABMAP: an open source toolbox for appearance-based loop closure detection[C]//Proc of IEEE International Conference on Robotics and Automation. 2012: 4730-4735.

[29] Galvez-López D, Tardos J D. Bags of binary words for fast place recognition in image sequences[J]. IEEE Trans on Robotics, 2012, 28(5):1188-1197.

[30] Tan Wei, Liu Haomin, Dong Zilong, *et al*. Robust monocular SLAM in dynamic environments[C]//Proc of IEEE International Symposium on Mixed and Augmented Reality. 2013:209-218.

[31] Mei C, Sibley G, Cummins M, *et al*. RSLAM: a system for large-scale mapping in constant-time using stereo[J]. International Journal of Computer Vision, 2011, 94(2):198-214.

[32] McManus C, Churchill W, Maddern W, *et al*. Shady dealings: robust, long-term visual localisation using illumination invariance [C]//Proc of IEEE International Conference on Robotics and Automation. 2014:901-906.

[33] Pascoe G, Maddern W, Newman P. Direct visual localisation and calibration for road vehicles in changing city environments[C]//Proc of IEEE International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2015:98-105.

[34] Urtasun R, Lenz P, Geiger A. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2012:3354-3361.

[35] Konolige K, Agrawal M. FrameSLAM: from bundle adjustment to real-time visual mapping[J]. IEEE Trans on Robotics, 2008, 24 (5):1066-1077.

[36] Nist D. An efficient solution to the five-point relative pose problem [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2003, 26(6):756-777.

[37] Benhimane S, Malis E. Homography-based 2D visual tracking and servoing[J]. International Journal of Robotics Research, 2007, 26(7):661-676.