doi:10.19306/j.cnki.2095-8110.2017.06.002

RGB-D SLAM 综述

王琉军,陈家斌,余 欢,朱汇申

(北京理工大学 自动化学院,北京 100081)

摘 要:RGB-D SLAM 是指使用 RGB-D 相机作为视觉传感器,进行同时定位与地图构建(SLAM)的技术。RGB-D 相机是近几年推出的能够同时采集环境 RGB 图像和深度图像的视觉传感器。首先对主流 RGB-D 相机,RGB-D SLAM 算法框架流程做了介绍,然后对 RGB-D SLAM 算法的国内外主要标志性成果,以及 RGB-D SLAM 的研究现状进行介绍,并对 RGB-D SLAM 方法前端视觉里程计中特征检测与匹配、后端位姿图优化、回环检测等关键技术进行介绍总结。最后,对 RGB-D SLAM 算法的优缺点进行了分析,并对 RGB-D SLAM 算法的研究热点及发展趋势进行了讨论。

关键词:RGB-D 相机;同时定位与地图构建;视觉里程计;位姿图优化;回环检测

中图分类号:TP242.6 **文献标志码**:A **文章编号**:2095-8110(2017)06-0009-10

An Overview of RGB-D SLAM

WANG Liu-jun, CHEN Jia-bin, YU Huan, ZHU Hui-shen

(School of Automation, Beijing Institute of Technology, Beijing 100081, China)

Abstract: RGB-D SLAM refers to Simultaneous Localization and Mapping (SLAM) using RGB-D camera as a visual sensor. RGB-D camera is a kind of vision sensor which can be used to capture RGB images and depth images of environment. Firstly, this paper introduced the RGB-D camera used frequently and the RGB-D SLAM algorithm framework. Then the main achievements of RGB-D SLAM method at home and abroad, research status of RGB-D SLAM and the key technologies of the RGB-D SLAM method, such as feature detection and matching, the pose graph optimization of the back end and the loop closure detection were introduced and summarized. Finally, the advantages and disadvantages of the RGB-D SLAM method were analyzed, and the research hotspot and development trend of the RGB-D SLAM method were discussed.

Key words: RGB-D camera; SLAM; Visual odometry; Pose graph optimization; Loop closure detection

0 引言

对于自主移动机器人,在导航定位方面涉及三个方面的问题:1)我在哪里(where am I);2)我要去哪里(where am I going);3)我要怎么去(How should I get there)[1]。其中第一个是定位问题,第二个是任务规划问题,第三个为路径规划问题。

同时定位与地图构建(Simultaneous Localization and Mapping,SLAM)问题描述为:一个机器人在一个陌生的环境中运动,通过自身携带的传感器对周围环境进行感知,然后绘制陌生环境的地图,同时定位自己在地图中的位置[2]。利用 SLAM

收稿日期:2017-04-15;**修订日期** 2017-09-06

基金项目:国家国防基金(9140A09050313BQ01127);国家自然科学基金(91120010)

作者简介:王琉军(1992-),男,硕士,主要从事智能导航、视觉 SLAM 方向的研究。E-mail:wangliujun_bit@163.com

通讯作者:陈家斌(1964-),男,教授,博导,从事惯性导航、智能导航方向研究。E-mail:chenjiabin@bit.edu.cn

方法可以解决自主移动机器人的定位与建图问题。使用相机作为传感器的 SLAM 方法被称为视觉 SLAM(VSLAM)。相比于传统的惯性器件(Inertial Measurement Unit,IMU)和激光雷达(Laser Scanner)等传感器,相机具有体积小、质量小和价格低等突出的优点,因此,VSLAM 成为近年来 SLAM 算法研究的热点。

RGB-D 相机是新兴的视觉传感器,它可以同时获取周围环境的 RGB 图像和每个像素的深度(Depth)信息。相比于单目相机和双目立体相机利用算法计算空间点的三维坐标,RGB-D 相机获取空间点的 3D 信息更加直接方便,深度信息通过红外结构光(Structured Light)或飞时(Time-of-flight,TOF)原理测得,和激光雷达有些相似。所以,有时候 RGB-D 相机又被称为伪激光雷达(Fake Laser)。由于 RGB-D 相机能相对容易地获取 RGB 图像上每一个像素的深度数据,并且 RGB-D 相机价格相对便宜,近年来 RGB-D SLAM 技术得到快速发展。

1 RGB-D SLAM 介绍

RGB-D SLAM 使用 RGB-D 深度相机作为传感器实现同时定位与地图构建。经过十多年的研究,虽然不同的研究团队使用的具体 VSLAM 算法有所区别,但是这些 VSLAM 算法都可以归为前端和后端两部分,RGB-D SLAM 作为 VSLAM 的一个分支,当然也不例外。本节首先介绍目前使用的主流 RGB-D 相机的种类并做出对比,然后简要介绍 RGB-D SLAM 算法流程,最后介绍 RGB-D SLAM 的标志性成果。

1.1 RGB-D 相机介绍

微软公司于 2010 年推出的 Kinect 相机是世界上首款 RGB-D 相机,Kinect 相机是微软公司针对 Microsoft Xbox 360 发布的一款由 PrimeSense 公司开发的体感设备。它由 RGB 相机、3D 深度传感器、麦克风阵列和机动倾斜马达等组成[3](图 1)。Kinect 相机中间部位是一款 RGB 彩色镜头,图像分辨率为 640×480,最大帧率为 30Hz。两边分别为红外发射和接收装置,组成 Kinect 的深度传感器。其深度传感器的分辨率为 320×240,帧率同样为 30Hz。RGB-D 相机采集的图像数据如图 2 所示[4]。

随后华硕也发布了其体感控制设备 Xtion Pro Live,它在外观上和 Kinect 相似,但比 Kinect 尺寸略小,RGB 和深度传感器的配置和 Kinect 相差无几。华硕 Xtion Pro Live 相机如图 3 所示。Xtion Pro Live 和 Kinect 参数对比如表 1 所示。

图 1 微软 Kinect 相机

Fig. 1 Microsoft Kinect

图 2 RGB 图像(左)和 Depth 图像(右)[4]

Fig. 2 RGB image (left) and Depth image (right)[4]

图 3 华硕 Xtion Pro Live

Fig. 3 Asus Xtion Pro Live

表 1 Xtion Pro Live 和 Kinect 参数对比

Tab. 1 Comparison of Xtion Pro Live and Kinect

属性	微软 Kinect v1	华硕 Xtion Pro Live
长/cm	28	18
宽/cm	6	3.6
高(带底座)/cm	7.5	5
深度感应有效距离/m	1.2～3.5	0.8～3.5
有效视角/(°)	水平:57 垂直:43	水平:58 垂直:45
电源/接口	外接电源+USB2.0	USB2.0
图像大小、帧率	彩色 640×480、32bit、30fps 深度 320×240、16bit、30fps	彩色 640×480、32bit、30fps 深度 320×240、16bit、30fps

后来，微软和英特尔（Intel）又相继推出了 Kinect V2 和 Realsense 体感设备，使用了更加先进的技术，并在硬件品质上有所提升。由于 RGB-D 相机功能强大并且价格低廉，引起了社会的极大关注，并逐渐在 SLAM 领域占有一席之地。

1.2　RGB-D SLAM 算法流程

RGB-D SLAM 算法大体上可分为前端视觉里程计、后端优化、回环检测（又称闭环检测）和建图几个部分。

算法前端根据输入 RGB 图像和 Depth 图像，对 RGB 图像进行特征点检测和特征描述子的计算（仅讨论基于特征的方法）；然后根据特征描述子进行相邻两帧图像的特征匹配，得到 2D-2D 特征匹配点集；然后根据 Depth 图像的深度信息，计算 2D-2D 特征匹配点对的空间三维坐标，得到 3D-3D 匹配点集。由匹配好的 3D-3D 点就可以计算出相邻两帧图像间的旋转和平移矩阵[4]，最后对运动估计误差进行优化，得到误差最小的位姿估计结果。这样就可以根据输入的视频流，不断地得到相机位姿的增量变化，所以算法前端构建了视觉里程计（VO）。

算法后端主要是为了优化 SLAM 过程中的噪声问题。实际应用当中，再精确的传感器获取的数据也会带有一定的噪声。所以，通过前端得到相邻两帧图像之间的运动估计之后，还要关心这个估计带有多大的噪声。后端优化就是从这些带有噪声的数据中，估计整个系统的状态，给出这个状态的最大后验概率（Maximum a Posteriori，MAP）。具体来说，后端接收不同时刻视觉里程计测量的相机位姿和回环检测的约束信息，采用非线性优化得到全局最优的位姿。在 VSLAM 中，前端和计算机视觉研究领域更为相关，例如图像的特征点检测与匹配，而后端则主要是滤波与非线性优化算法。

回环检测又称为闭环检测（Loop Closure Detection），主要解决机器人位置随时间漂移的问题。回环检测就是让机器人具有识别曾经到达过的场景的能力。视觉回环检测就是通过比较两幅图像数据的相似性，由于图像信息丰富，使得视觉回环检测比较容易实现。如果回环检测成功，则认为机器人曾经来过这个地点，把比对信息输送给后端优化算法，后端根据回环检测的信息，调整机器人轨迹和地图。通过回环检测，可以显著地减小累积误差[4]（图 4）。

建图（Mapping）是指构建地图的过程，只有构建环境的地图，才能实现机器人的定位或导航功能。根据实际应用的需要，SLAM 算法构建的地图大体可分为度量地图（Metric Map）和拓扑地图（Topological Map）两种。

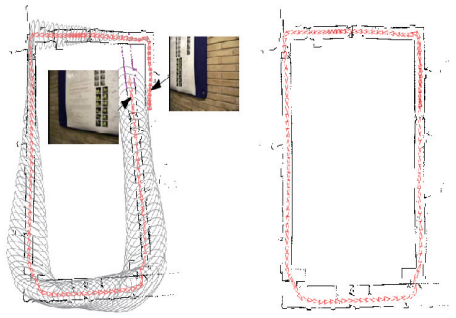

图 4　累积误差与回环检测校正结果[4]

Fig. 4　Cumulative error and the result after loop closing[4]

根据以上的叙述，RGB-D SLAM 算法流程如图 5 所示。

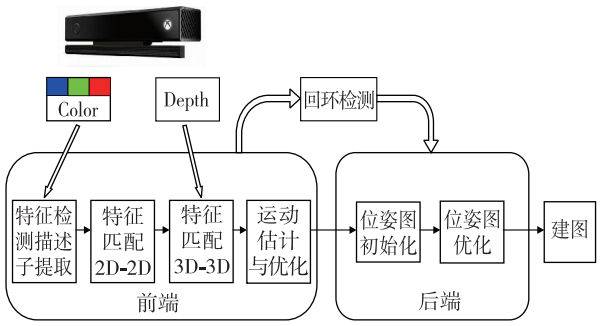

图 5　RGB-D SLAM 算法流程图

Fig. 5　Flow chart of RGB-D SLAM algorithm

1.3　RGB-D SLAM 标志性成果

华盛顿大学（University of Washington）的 Peter Henry、Michael Krainin、Evan Herbst 联合英特尔（Intel）实验室的 Xiaofeng Ren、Dieter Fox[4]最早提出基于 RGB-D 相机的 SLAM 算法。Peter Henry 等使用尺度不变特征转换（Scale Invariant Feature transform，SIFT）[6]方法对相邻两帧 RGB 图像进行特征检测与描述子提取，然后加入 Depth 图像的数据生成 3D-3D 特征点对信息，使用随机采样一致性（Random Sample Consensus，RANSAC）方法[8]对 3D-3D 匹配点对进行配准并求出对应的变换矩阵，采用迭代最近点（Iterative Closest Point，ICP）[10]方法对运动变换进行优化。后端采用基于树的网络优化（Tree-based network Optimizer，TORO）[11]算法并加入基于视觉的回环检测约束最终得到全

局最优的三维场景地图。

随后不久，Peter Henry 等又对他们之前提出的算法进行了改进[12]。针对 SIFT 方法速度较慢的缺陷，在特征提取时采用基于加速段的特征检测子(Features from Accelerated Segment Test，FAST)[13]方法，描述子的计算采用 Calonder 方法。并对 RANSAC 阶段预投影误差进行了优化，提出只有当特征匹配失败或者是匹配结果中包含的匹配点对数目较少时才进行 ICP 优化。后端使用性能更加优秀的稀疏光束平差法(Sparse Bundle Adjustment，SBA)[14]代替 TORO 算法进行全局优化，通过场景识别提高了闭环检测的效率。

德国弗莱堡大学(University of Freiburg)计算机科学系的 Nikolas Engelhard、Felix Endres 和 Jurgen Hess 开发出了一套基于手持 Kincet 相机的 RGB-D SLAM 系统[15]，在 SLAM 算法前端使用基于加速鲁棒性特征(Speeded Up Robust Features，SURF)[16]算法对输入的 RGB 图像进行特征检测与特征描述子提取，对相邻两帧 RGB 图像进行特征匹配。然后加入 Depth 图像中的深度信息对匹配的特征点计算出三维空间坐标。运动估计与优化方面使用 RANSAC 方法估计两帧图像之间的运动，使用改进的 ICP 方法优化相机的运动转移矩阵。算法后端使用 HOGMAN 位姿图求解方法对前端计算出的运动转移矩阵进行全局优化以得到全局最优位姿。最后输出彩色点云数据表示的三维环境地图。

之后，Felix Endres 等又开发出了一套新的 RGB-D SLAM 系统[17]。在新的系统中，他们提出在后端优化部分使用开源的图优化库(General Graph Optimization，g2o)[18]进行全局位姿图优化，得到一个全局的 3D 点云表示的三维环境地图。最后，使用基于八叉树(Octtree-based Mapping，OctoMap)的地图构建框架对得到的点云进行体素(Voxel)化表示，最终得到一个 3D 栅格地图。所得到的 3D 栅格地图可以直接被用于机器人定位、路径规划和导航。

Willow Garage 公司的 Nicola Fioraio 和 Kurt Konolige 提出了一种利用 RGB-D 相机构建稠密地图的 SLAM 算法[19]。该算法在 Depth 图像的两两配准中使用光束平差法(Bundle Adjustment，BA)[14]同时对 Depth 图像和 RGB 图像进行匹配，一方面根据所得 2D 匹配结果生成稀疏图(Sparse Graph)，另一方面根据所得的 2D 匹配结果使用 RANSAC 方法估算相机位姿，再利用 GICP(Generalized-ICP)[20]方法以所得相机位姿为初始值对采样后的两帧图像进行配准；后端全局配准中每当检测到闭环时则使用 g2o 方法对稀疏图进行优化，以得到全局最优配准结果。号称此算法对两帧图像进行 GICP 配准时只需要 10ms，具有实时性，比其他几个 RGB-D SLAM 算法性能更优秀。

为了对来自世界各个机构和高校研究者提出的 RGB-D SLAM 算法进行评估，德国 Munich 理工大学的 Jurgen Sturm 和 Daniel Cremers 联合 Freiburg 大学的 Nikolas Engelhard 和 Felix Endres 一起制作了标准的 RGBD SLAM 数据集[21]。他们采用高速运动摄像捕捉系统 VICON 捕捉手持的 Kinect 的位置和姿态数据。这些数据被作为真实的轨迹(Ground Truth)，可以和通过 RGBD SLAM 算法程序估算出来的位姿和姿态信息作比较。此外，标准数据集还包括以时间戳命名的匹配好的 RGB 图像和 Depth 图像。录制数据集时，选取了工厂、办公室等许多场景，具有代表性的数据集有 FR1 ROOM 、FR2 DESK 、FR3 LONG OFFICE 等 。

帝国理工学院(Imperial College London)的 Richard A. Newcombe、Andrew J. Davison 等联合微软研究院(Microsoft Research)的 Shahram Izadi、Otmar Hilliges 等提出的 KinectFusion[22]是第一个基于 Kinect 深度相机的，能在 GPU 上实时构建稠密三维环境地图的 SLAM 算法。该算法仅使用 Kinect 相机一个传感器计算相机的位姿并构建环境的三维地图。

西班牙萨拉戈萨大学(University of Zaragoza)的 Raul Mur-Artal 和 Juan Domingo Tardos 继 2015 年提出比较完整的单目 ORB-SLAM 算法[23]后，2016 年他们又加入了对双目相机和 RGB-D 相机的支持[24]。ORB-SLAM 算法采用多线程处理，分为追踪(Tracking)、地图构建(Local Mapping)、闭环检测(Loop Closing)几个线程进行处理。特征提取与匹配、稀疏地图创建和回环位置识别都是基于 ORB(Oriented Robust Brief)[25]特征，在标准 CPU 上运行就可以实时地进行 SLAM 运算，而且精度很高。

2 RGB-D SLAM 关键技术

2.1 特征点提取

VSLAM 算法根据利用图像信息的不同可以分

为基于特征的 SLAM 算法和直接 SLAM(Direct SLAM)[26]算法。基于特征点法的前端,长久以来被认为是 SLAM 的主流方法,它运行稳定,对光照、动态物体不敏感,是目前比较成熟的解决方案。

特征点是由关键点(Key Point)和描述子(Descriptor)两部分组成。关键点是指该特征点在图像里的位置,描述子通常是一个向量,按照人们的需求描述关键点周围像素的信息。常用的特征提取算法有 SIFT[6]、SURF[16]和 ORB[25]等。

1999 年 British Columbia 大学的 David G. Lowe 教授总结了基于不变量技术的特征检测方法,并正式提出了一种基于尺度空间的、对图像缩放、旋转甚至仿射变换保持不变性的图像局部特征描述算子 SIFT,这种算法在 2004 年被加以完善[6]。SIFT 算法经过十多年的发展,已经取得了巨大的成功[31-34]。在文献[4]中,Peter Henry 等使用 SIFT 算法对相邻两帧 RGB 图像进行特征提取与描述子计算,在室内大型场景环境下测试取得了较好的结果。

SIFT 算法充分考虑了图像变换过程中出现的光照、尺度和旋转等变化,但是计算量也随之增大。到目前为止,普通的 CPU 还无法实时地计算 SIFT 特征,为此,Herbert Bay 等在 2006 年提出了 SURF[16]算法。SURF 算法中采用积分图、近似的 Hessian 矩阵和 Haar 小波变换运算来提高时间效率,采用 Haar 小波变换增加鲁棒性。与 SIFT 特征相比,SURF 算法时间复杂度有所降低,同样具有尺度和旋转不变性,且相对于 SIFT 特征的算法速度提高了 3～7 倍[35-37]。文献[15]使用速度更快的 SURF 算法对 Kinect 相机采集的 RGB 图像进行特征检测与描述子提取,最终实现了实时地构建环境的 3D 模型。

ORB 算法是由 Ethan Rublee 在 ICCV 2011 上提出的[25]。该算法采用改进的具有方向性的 FAST[38]和速度极快的二进制稳健基元独立特征(Binary Robust Independent Elementary Features, BRIEF)[39]描述子。在文献[25]中,作者测试了 SIFT、SURF 和 ORB 算法的性能:对同一幅图像提取约 1000 个特征点的情况下,SIFT 耗时约 5228.7ms,SURF 花费约 217.3ms,而 ORB 则用了 15.3ms。由此可以看出,ORB 算法在兼有 SIFT 和 SURF 算法旋转、尺度不变性的基础上速度方面大为提升。文献[23]中,University of Zaragoza 的 Raul Mur-Artal 和 Juan Domingo Tardos 构建了基于单目相机的 ORB 特征检测 SLAM 框架,并在小型、大型室内环境和常用的数据集上测试 ORB-SLAM 算法,实验结果表明 ORB-SLAM 算法相比其他顶尖的 SLAM 算法性能大为提升。文献[24]中,Raul Mur-Artal 和 Juan Domingo Tardos 改进了原有的算法提出 ORB-SLAM2 算法,并增加了对双目立体相机和 RGB-D 相机的支持。

2.2　后端优化

后端的优化方法一般分为两大类:基于滤波器的方法和基于图优化(Graph Optimization)[42]的方法。基于滤波器的方法理论基础是概率论里的贝叶斯公式,利用控制信息对机器人的位姿进行先验估计,然后利用观测信息对机器人位姿和地图进行后验估计。早期的 SLAM 方法多采用滤波器优化,通常使用扩展卡尔曼滤波(Extended Kalman Filter,EKF)和粒子滤波(Particle Filter, PF)滤波器。直至 21 世纪早期,基于 EKF 滤波器的方法仍然占据了 SLAM 的主导地位。文献[40]提出的最早的实时 SLAM 系统既是基于 EKF 滤波器开发的。

为了克服 EKF 滤波器的缺点:线性化误差和噪声高斯分布假设,研究者们提出了粒子滤波和非线性优化[41]等方法。基于图优化的非线性优化方法将机器人的位姿转化成图论中的顶点(Vertex),将机器人位姿之间的约束,以及位姿与观测量间的约束则构成了边(Edge),从而将 SLAM 问题转化为一个优化问题,并利用最小二乘法进行求解(图 6)。由于 SLAM 问题中雅可比矩阵具有稀疏结构,并得益于计算机技术和算法的优化,使得图优化方法成为现实。

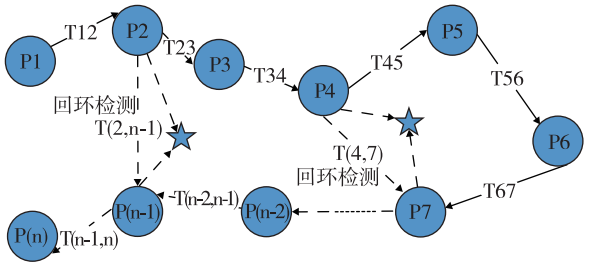

图 6　图优化示意图

Fig. 6　Graph Optimization

g2o[18]是一个开源的图优化库,是目前最常用的后端优化器,可供选择的梯度下降方法有 GN(Gauss-Newton)、LM(Levernberg-Marquardt)[43]

和 DogLeg(Powell's dogleg)等。文献[17]中 Felix Endres 等利用 g2o 优化器对位姿进行优化,得到全局一致的三维地图。针对 FR1 数据包[21],该文详细地叙述了其算法的运行速度和定位精度。因此该方法可以成为利用 RGB-D 数据完成 SLAM 的比较基准。

2.3 回环检测

回环检测的关键,就是如何能正确有效地判断相机经过同一个地方,从而为后端提供更加有效的位姿约束,进而消除累积误差,得到全局一致(Global Consistent)的位姿估计(图 7)。

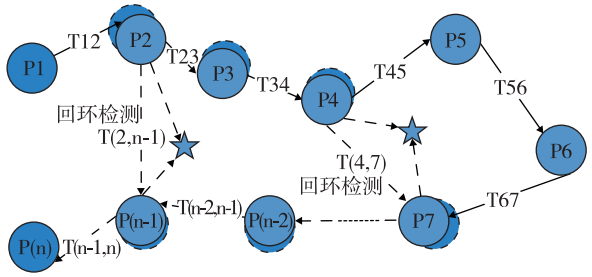

图 7 回环检测与优化

Fig. 7 Loop Closing and Optimization

回环检测一般有两种方案:基于里程计(Odometry Based)[44]和基于外观(Appearance Based)[47]的。基于里程计的回环检测通过当前相机的位置判断是否曾经到达过先前的某个位置附近,检测有没有回环关系。但是由于累积误差的存在,检测相机回到之前位置附近在逻辑上存在悖论,因此,累积误差较大时无法给出准确的结果[45]。基于外观的回环检测仅根据图像相似关系来检测回环,与之前的位姿没有直接关系,因而使得回环检测相对独立于前端和后端,能给出更加真实有效的检测结果,成为 VSLAM 回环检测的主流方法[23,46-47]。

基于外观的回环检测多使用图像特征的方法,词袋(Bag of Words,BoW)[48]方法因其有效性得到了广泛应用[23,49-51]。词袋方法首先通过对大量的训练图像提取特征(SIFT、SURF、ORB 等),并对这些特征(Word)进行分类(经典的是 K-means 聚类算法[52])获得叶子节点即为词典(Dictionary)。这样,一幅图像就可以根据是否出现相应的单词(Word)描述为词典下的一个向量。

文献[51]中,使用一种 K-means 扩展的 K 叉树来表达词典。在根节点,使用 K-means++[53]方法把所有样本聚成 k 类;对每一层的每个节点,把属于该节点的样本再聚成 k 类,得到下一层;以此类推,最终得到深度为 L,每层分叉为 k 的树,可以容纳 k^L 个单词。树的中间结构仅供快速查找时使用,在查找给定特征对应的单词时,只需将它与每个中间节点的聚类中心比较(一共比较 L 次),即可找到最后的单词,保证了对数级别查找的高效率。文献[6]使用 SIFT 特征进行全局定位,用 KD 树来排列地图点。文献[54-55]使用基于 SURF 特征描述子的词典方法去进行闭环检测特征,提取约耗时 400ms。文献[23-24]使用效率和精度折中的 ORB 特征进行回环检测,即使用基于 ORB 特征的词典筛选出闭环,再通过相似性[56-57]计算进行闭环的验证。

3 RGB-D SLAM 优缺点与发展趋势

3.1 RGB-D SLAM 优缺点

单目 SLAM 系统无法通过单张图像获取像素的深度信息,一般通过三角测量的方法估计像素的深度。并且由于单目视觉的尺度不确定性,单目 SLAM 系统必须进行初始对准。双目 SLAM 系统通过视差原理测得深度信息,利用左右相机的图像的特征匹配获取像素点对,消除了单目 VSLAM 系统的初始化问题。相比单目、双目等 SLAM 系统,RGB-D SLAM 系统能够通过传感器在硬件层面上测得图像点的深度,无需考虑单目 SLAM 系统的初始对准问题,也不必像双目 SLAM 系统消耗大量的资源计算深度。利用 RGB-D SLAM 进行稠密地图的构建相对容易,并且 RGB-D 相机使用红外结构光或飞时原理测量深度,保证了深度数据对纹理的无关性,这样即使面对纯色的物体,只要它能够反光,RGB-D SLAM 也能够比较准确地获取深度数据。

RGB-D SLAM 也有其相应的缺点。由于 RGB-D 相机测量深度的原理,使得 RGB-D 相机容易受到日光或者其他传感器发射的红外光的干扰,因此使用多个 RGB-D 相机时会相互干扰,在室外使用效果也不好。对于透明材质的物体,因为反射光较少,也无法很好地测量其深度。而且由于主流 RGB-D 相机深度有效测量距离在 0.5~4m 的区间内,使得 RGB-D SLAM 无法应用在室外大场景下,应用环境受限。此外,RGB-D SLAM 算法实时运行对计算平台要求较高,还不能应用到轻量级的嵌入式平台上。

3.2　RGB-D SLAM 发展趋势

RGB-D 的一个研究热点方向就是和深度学习相结合。到目前为止，SLAM 的方案都处于特征点或者像素级别，利用特征点或像素的方法和我们日常生活实践中的方式很不一样。

很久之前，研究者就试图将物体信息结合到 SLAM 中。文献[58-61]中把物体识别和 VSLAM 结合起来，构建带标签的地图。文献[62]将标签信息引入到优化端的目标函数中进行优化。以上的工作都称为语义 SLAM(Semantic SLAM)(图 8[70])。综合来说，SLAM 与语义的结合点有以下两个方面[63]：一方面语义辅助 SLAM。传统的语义分割和物体识别往往只考虑一幅图片，而在 SLAM 中利用一台移动的相机，如果把语义分割应用到 SLAM 中，将得到一个带有标签的语义地图。另外，语义信息也可以为回环检测和 BA 优化提供更多的信息。另一方面 SLAM 辅助语义。物体识别和语义分割都需要大量的训练数据，并且需要人工从不同视角采集该物体的图片，输入分类器进行识别。利用 SLAM，可以自动地计算物体在图像中的位置，节省人力成本，并且能加快分类器的训练。

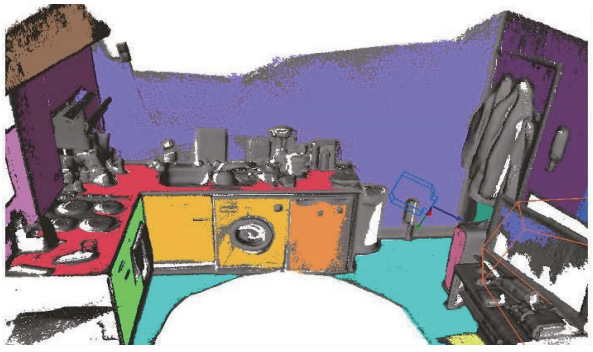

图 8　语义 SLAM[70]

Fig. 8　Semantic SLAM[70]

此外，基于词袋的回环检测算法完全依赖于外观而没有利用任何其他信息，这导致外观相似的图像容易被当成回环。从词袋的模型来说，它本身是一个非监督的机器学习过程：构建词典相当于对特征描述子进行聚类。因此，基于深度学习的框架完全可以应用到回环检测当中[68-69]。

文献[64-65]使用深度学习网络对图像进行识别、检测和分割。文献[66]利用 RGB-D 相机采集的室内 RGB 图像和 Depth 图像，进行了轮廓检测，自底向上分组，目标检测和语义分割(图 9)。文献[67]创新性的将卷积神经网络(Convolutional Neural Networks，CNN)应用到 SLAM 本身的位姿估计与回环检测当中。虽然这些方法还没有成为主流，但将 SLAM 与深度学习结合处理图像将是一个很有前景的研究方向。

图 9　RGB-D 图像语义分割[66]

Fig. 9　Semantic segmentation of RGB-D images[66]

4　结束语

自从微软 2010 年针对 Microsoft Xbox 360 发布第一款 RGB-D 相机 Kinect 以来，世界各地的研究者们就竞相将 RGB-D 相机应用到 VSLAM 技术当中。经过近些年的发展，RGB-D SLAM 算法框架趋于成熟，也取得了比较好的实验效果。但是，RGB-D SLAM 算法远没有达到完善的地步：除了比较常用的特征点法，基于像素的直接法近年来也崭露头角；实际应用当中，RGB-D 相机与惯性器件(IMU)、激光雷达等传感器的融合会得到更好的效果；将深度学习应用到 RGB-D SLAM 中将是今后的一个发展方向；RGB-D SLAM 应用到 AR、VR 领域将会改变人们的生活。相信随着硬件技术的进步、算法的日趋完备，RGB-D SLAM 最终能扬长避短，得到广泛的应用。

参考文献

[1] Leonard J J, Durrant-Whyte H F, Cox I J. Dynamic map building for an autonomous mobile robot[M]. Sage Publications, Inc. 1992.

[2] 祝继华，郑南宁，袁泽剑，等. 基于 ICP 算法和粒子滤波的未知环境地图创建[J]. 自动化学报，2009，35(8)：1107-1113.

[3] 丁洁琼. 基于 RGB-D 的 SLAM 算法研究[D]. 西安：西安电子科技大学，2014.

[4] Henry P, Krainin M, Herbst E, et al. RGB-D mapping: Using depth cameras for dense 3D modeling of

indoor environments[J]. International Journal of Robotics Research, 2010, 31(5):647-663.

[5] Newman P, Ho K. SLAM-loop closing with visually salient features[C]//Proceedings of the 2005 IEEE International Conference on Robotics and Automation. IEEE, 2006:635-642.

[6] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[7] Bay H, Ess A, Tuytelaars T, et al. Speeded-up robust features (SURF)[J]. Computer Vision and Image Understanding, 2008, 110(3): 346-359.

[8] Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF[C]//IEEE International Conference on Computer Vision. IEEE Computer Society, 2011:2564-2571.

[9] Fischler M A, Bolles R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography[J]. Communications of the ACM, 1981, 24(6):381-395.

[10] Zhang Z. Iterative point matching for registration of free-form curves and surfaces [J]. International Journal of Computer Vision, 1994, 13(2):119-152.

[11] Burgard W, Brock O, Stachniss C. A tree parameterization for efficiently computing maximum likelihood maps using gradient descent [C]//Robotics: Science and System, 2007:65-72.

[12] Henry P, Krainin M, Herbst E, et al. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments[M]//Experimental Robotics. Springer Berlin Heidelberg, 2014:647-663.

[13] Rosten E, Drummond T. Machine learning for high-speed corner detection[C]//Proceedings of the European Conference on Computer Vision. 2006:430-443.

[14] Triggs B, Mclauchlan P F, Hartley R I, et al. Bundle adjustment - a modern synthesis[C]//International Workshop on Vision Algorithms: Theory and Practice. Greece, 1999:298-372.

[15] Engelhard N, Endres F, Hess J, et al. Real-time 3D visual SLAM with a hand-held RGB-D camera[C]//Proceedings of the RGB-D Workshop on 3D Perception in Robotics at the European Robotics Forum. Västerås, Sweden, 2011:1-15.

[16] Bay H, Ess A, Tuytelaars T, et al. Speeded-up robust features[J]. Computer Vision & Image Understanding, 2008, 110(3):404-417.

[17] Endres F, Hess J, Engelhard N, et al. An evaluation of the RGB-D SLAM system[C]//IEEE International Conference on Robotics and Automation. IEEE, 2012: 1691-1696.

[18] Kümmerle R, Grisetti G, Strasdat H, et al. g2o: A general framework for graph optimization[C]//2011 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2011: 3607-3613.

[19] Fioraio N, Konolige K. Realtime visual and point cloud SLAM[C]//Proceedings of the RGB-D Workshop on Advanced Reasoning with Depth Cameras at Robotics: Science and Systems Conference (RSS). 2011: 27.

[20] Segal A, Haehnel D, Thrun S. Generalized-ICP[C]//Proceedings of the Robotics: Science and Systems Conference(RSS). 2009: 435.

[21] Sturm J, Engelhard N, Endres F, et al. A benchmark for the evaluation of RGB-D SLAM systems[C]//2012 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). IEEE, 2012:573-580.

[22] Newcombe R A, Izadi S, Hilliges O, et al. Kinect-Fusion: Real-time dense surface mapping and tracking [C]//2010 10th IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE Computer Society, 2011:127-136.

[23] Mur-Artal R, Montiel J M M, Tardós J D. ORB-SLAM: a versatile and accurate monocular SLAM system[J]. IEEE Transactions on Robotics, 2015, 31(5):1147-1163.

[24] Murartal R, Tardos 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.

[25] Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF [C]//2011 IEEE International Conference on Computer Vision (ICCV). IEEE Computer Society, 2011:2564-2571.

[26] Hmer J, Gumhold S, Cremers D. Real-time dense geometry from a handheld camera[C]//Joint Pattern Recognition Symposium. Springer, Berlin, Heidelberg, 2010: 11-20.

[27] Engel J, Sturm J, Cremers D. Semi-dense visual odometry for a monocular camera[C]//2013 IEEE International Conference on Computer Vision(ICCV). IEEE Computer Society, 2013:1449-1456.

[28] Newcombe R A, Lovegrove S J, Davison A J. DTAM: Dense tracking and mapping in real-time [C]//2011 IEEE International Conference on Computer Vision(ICCV). IEEE Computer Society, 2011:2320-2327.

[29] Forster C, Pizzoli M, Scaramuzza D. SVO: Fast

semi-direct monocular visual odometry[C]//2014 IEEE International Conference on Robotics and Automation(ICRA). IEEE, 2014:15-22.

[30] Engel J, Schöps T, Cremers D. LSD-SLAM: Large-scale direct monocular SLAM[C]//Proceedings of the European Conference on Computer Vision. Springer, Cham, 2014: 834-849.

[31] Ali A M, Nordin M J. SIFT based monocular SLAM with multi-clouds features for indoor navigation[C]//TENCON 2010 - 2010 IEEE Region 10 Conference. IEEE, 2010:2326-2331.

[32] Wu E Y, Zhao L K, Guo Y P, et al. Monocular vision SLAM based on key feature points selection [C]//2010 IEEE International Conference on Information and Automation(ICIA). Harbin, China, 2010: 1741-1745.

[33] Chen C H, Chan Y P. SIFT-based monocluar SLAM with inverse depth parameterization for robot localization [C]//2007 IEEE Workshop on Advanced Robotics and Its Social Impacts (ARSO). IEEE, 2007:1-6.

[34] Zhu D X. Binocular vision-SLAM using improved SIFT algorithm[C]//2010 2nd International Workshop on Intelligent Systems and Applications (ISA). IEEE, 2010:1-4.

[35] Zhang Z, Huang Y, Li C, et al. Monocular vision simultaneous localization and mapping using SURF [C]//2008 7th World Congress on Intelligent Control and Automation(WCICA). IEEE, 2008:1651-1656.

[36] Ye Y. The research of SLAM monocular vision based on the improved SURF feather[C]//2014 International Conference on Computational Intelligence and Communication Networks(CICN). IEEE, 2014:344-348.

[37] Wang Y T, Feng Y C. Data association and map management for robot SLAM using local invariant features[C]//2013 IEEE International Conference on Mechatronics and Automation(ICMA). IEEE, 2013: 1102-1107.

[38] Rosten E, Drummond T. Machine learning for high-speed corner detection[C]//Proceedings of the Computer Vision. 2006:430-443.

[39] Calonder M, Lepetit V, Strecha C, et al. BRIEF: Binary robust independent elementary features[C]//Proceedings of the European Conference on Computer Vision. Springer-Verlag, 2010:778-792.

[40] Davison A J, Reid I D, Molton N D, et al. MonoSLAM: Real-time single camera SLAM[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(6):1052-1067.

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

[42] Lu F, Milios E. Globally consistent range scan alignment for environment mapping[J]. Autonomous Robots, 1997, 4(4):333-349.

[43] Moré J J. The Levenberg-Marquardt algorithm: Implementation and theory[M]//Numerical Analysis. Springer, Berlin, Heidelberg, 1978: 105-116.

[44] Hahnel D, Burgard W, Fox D, et al. An efficient fast SLAM algorithm for generating maps of large-scale cyclic environments from raw laser range measurements[C]//2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE Xplore, 2003:206-211.

[45] Beeson P, Modayil J, Kuipers B. Factoring the mapping problem: Mobile robot map-building in the hybrid spatial semantic hierarchy[J]. The International Journal of Robotics Research, 2010, 29(4):428-459.

[46] Latif Y, Cadena C, Neira J. Robust loop closing over time for pose graph SLAM[J]. The International Journal of Robotics Research, 2013, 32 (14): 1611-1626.

[47] Ulrich I, Nourbakhsh I. Appearance-based place recognition for topological localization[C]//2000 IEEE International Conference on Robotics and Automation (ICRA). IEEE Xplore, 2000:1023-1029.

[48] Nister D, Stewenius H. Scalable recognition with a vocabulary tree[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2006:2161-2168.

[49] Mur-Artal R, Tardós J D. Fast relocalisation and loop closing in keyframe-based SLAM [C]//2014 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2014: 846-853.

[50] Angeli A, Filliat D, Doncieux S, et al. A fast and incremental method for loop-closure detection using bags of visual words[J]. IEEE Transactions on Robotics, 2008, 24(5):1027-1037.

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

[52] Lloyd S. Least squares quantization in PCM[J]. IEEE Transactions on Information Theory, 1982, 28 (2):129-137.

[53] Arthur D, Vassilvitskii S. k-means++: The advan-

tages of careful seeding[C]//Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, 2007: 1027-1035.

[54] Gálvezlópez D, Tardós J D. Real-time loop detection with bags of binary words[C]//2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2011: 51-58.

[55] Cummins M, Newman P. Appearance-only SLAM at large scale with FAB-MAP 2.0[J]. The International Journal of Robotics Research, 2011, 30(9):1100-1123.

[56] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos [C]//2003 IEEE International Conference on Computer Vision. IEEE Xplore, 2003:1470-1477.

[57] Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF[J]. Journal of Documentation, 2004, 60(5):503-520.

[58] Nüchter A, Hertzberg J. Towards semantic maps for mobile robots[J]. Robotics & Autonomous Systems, 2008, 56(11):915-926.

[59] Civera J, Gálvez-López D, Riazuelo L, et al. Towards semantic SLAM using a monocular camera [C]//2011 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). IEEE, 2011:1277-1284.

[60] Koppula H S, Anand A, Joachims T, et al. Semantic labeling of 3D point clouds for indoor scenes[C]// Advances in Neural Information Processing Systems. 2011: 244-252.

[61] Anand A, Koppula H S, Joachims T, et al. Contextually guided semantic labeling and search for three-dimensional point clouds [J]. The International Journal of Robotics Research, 2013, 32(1):19-34.

[62] Fioraio N, Stefano L D. Joint detection, tracking and mapping by semantic bundle adjustment [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2013:1538-1545.

[63] 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.

[64] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database [C]//2009 IEEE Conference onComputer Vision and Pattern Recognition(CVPR). IEEE, 2009:248-255.

[65] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016:770-778.

[66] Gupta S, Arbeláez P, Girshick R, et al. Indoor scene understanding with RGB-D images: Bottom-up segmentation, object detection and semantic segmentation [J]. International Journal of Computer Vision, 2015, 112(2): 133-149.

[67] Hou Y, Zhang H, Zhou S. Convolutional neural network-based image representation for visual loop closure detection[C]//2015 IEEE International Conference on Information and Automation. IEEE, 2015:2238-2245.

[68] Gao X, Zhang T. Loop closure detection for visual SLAM systems using deep neural networks [C]// 2015 34th Chinese Control Conference(CCC). IEEE, 2015:5851-5856.

[69] Gao X, Zhang T. Unsupervised learning to detect loops using deep neural networks for visual SLAM system[J]. Autonomous Robots, 2017, 41(1):1-18.

[70] Salas-Moreno R F, Glocken B, Kelly P H J, et al. Dense planar SLAM [C]//2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, 2014:157-164.