视觉SLAM综述

zhuanlan.zhihu.com/p/53836358

《机器学习-原理、算法与应用》，清华大学出版社，**雷明著**，由SIGAI公众号作者倾力打造。

什么是视觉SLAM

SLAM是“Simultaneous Localization And Mapping”的缩写，可译为同步定位与建图。概率 SLAM 问题 (the probabilistic SLAM problem) 起源于 1986 年的IEEE Robotics and Automation Conference 大会上，研究人员希望能将估计理论方法 (estimation-theoretic methods) 应用在构图和定位问题中。 SLAM最早被被应用在机器人领域，其目标是在没有任何先验知识的情况下，根据传感器数据实时构建周围环境地图，同时根据这个地图推测自身的定位[1]。

假设机器人携带传感器 (相机) 在未知环境中运动，为方便起见，把一段连续时间的运动变成离散时刻 t=1，...k，而在这些时刻，用 x 表示机器人的自身位置，则各时刻的位置就记为 *x1*，*x2...xk*，它构成了机器人的轨迹。地图方面，假设地图由许多个路标组成，而每个时刻，传感器会测量到一部分路标点，得到它们的观测数据。设路标点共有 N 个，用 *y1*，*y2...yn* 表示。通过运动测量 u 和传感器读数 z 来求解定位问题 (估计 x)和建图问题 (估计 y)。

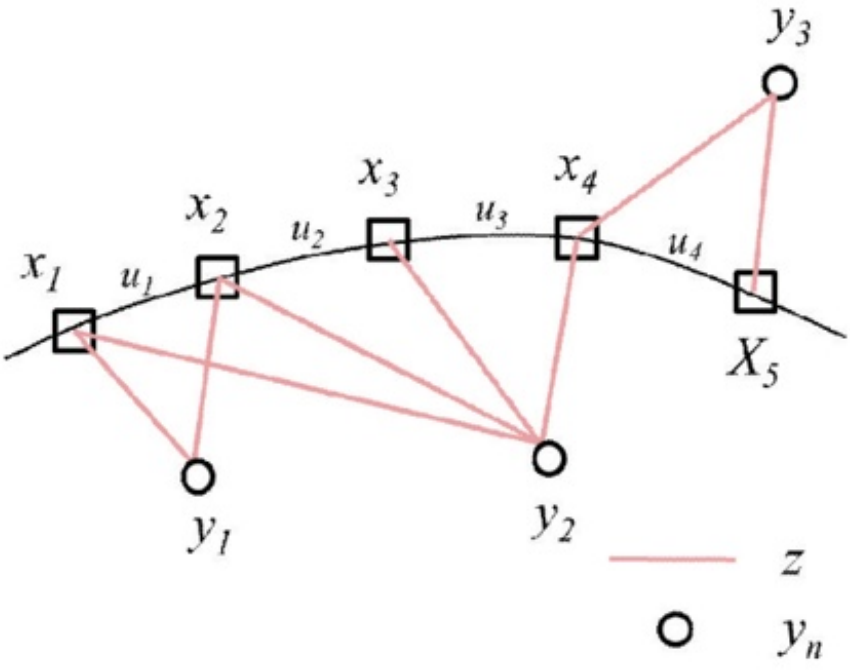

图 1 SLAM系统示意图[16]

只利用相机作为外部感知传感器的SLAM称为视觉SLAM (vSLAM [2])。相机具有视觉信息丰富、硬件成本低等优点，经典的vSLAM系统一般包含前端视觉里程计、 后端优化、闭环检测和构图四个主要部分[3]。

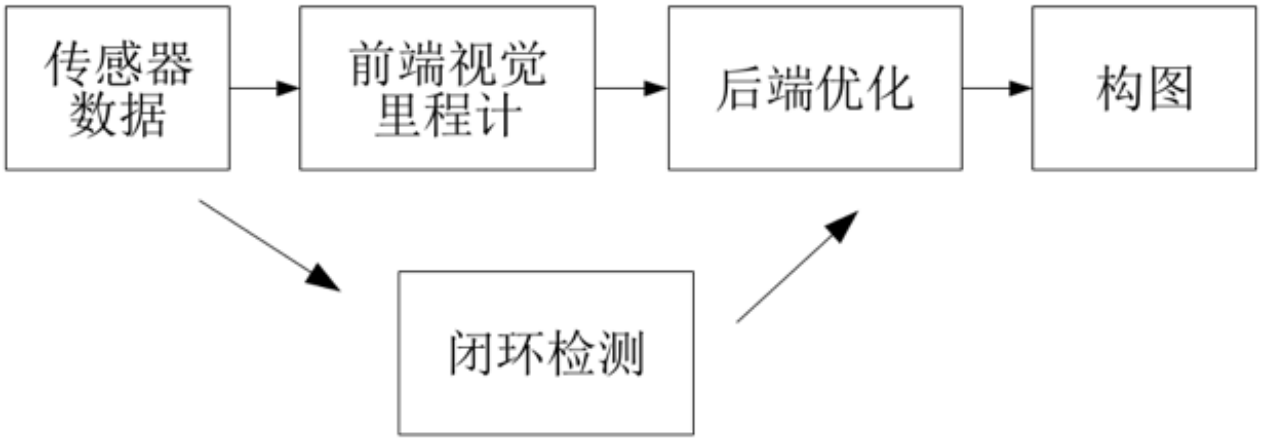

图 2 经典vSLAM系统流程图[5,16]

● 视觉里程计 (Visual Odometry)：仅有视觉输入的姿态估计[4]；

● 后端优化 (Optimization): 后端接受不同时刻视觉里程计测量的相机位姿，以及闭环检测的信息，对它们进行优化，得到全局一致的轨迹和地图[5]；

● 闭环检测 (Loop Closing): 指机器人在地图构建过程中, 通过视觉等传感器信息检测是否发生了轨迹闭环, 即判断自身是否进入历史同一地点[6];

● 建图 (Mapping): 根据估计的轨迹，建立与任务要求对应的地图[5]。

根据生成方法的不同，SLAM可以分成两大类：间接方法和直接方法。下面依次介绍这两种方法。

间接法及其典型系统

间接法首先对测量数据进行预处理来产生中间层，通过稀疏的特征点提取和匹配来实现的，也可以采用稠密规则的光流，或者提取直线或曲线特征来实现。然后计算出地图点坐标或光流向量等几何量，因此间接法优化的是几何误差：

其中 *ui* 为 *Ik-1* 中任意像素点，它投影到空间点的坐标为 *Pi*， u′*i*是 *Pi* 投影到 *Ik* 上的坐标。之后利用中间层的数值来估计周围环境的三维模型和相机运动。

$$T_{k-1,k} = \underset{T}{argmin} \sum_{i} ||u_i' - \pi(p_i)||^2_{\sum}$$

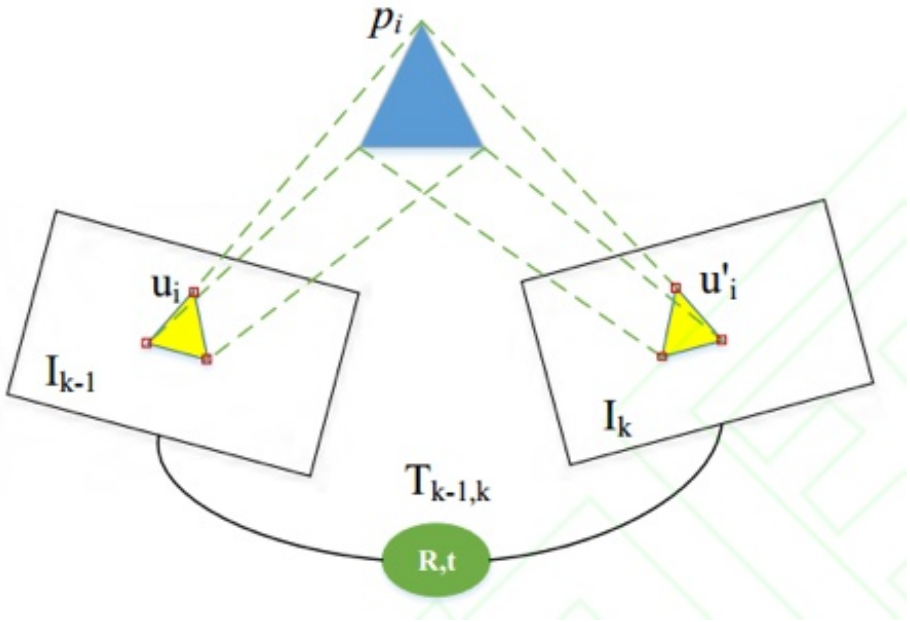

图 3 间接法示意图[3]

MonoSLAM

MonoSLAM 是第一个实时的单目视觉 SLAM 系统[7]。 MonoSLAM 以EKF(扩展卡尔曼滤波)为后端，追踪前端稀疏的特征点，以相机的当前状态和所有路标点为状态量，更新其均值和协方差。在 EKF 中，每个特征点的位置服从高斯分布，可以用一个椭球表示它的均值和不确定性，它们在某个方向上越长，说明在该方向上越不稳定。 该方法的缺点：场景窄、路标数有限、稀疏特征点易丢失等。

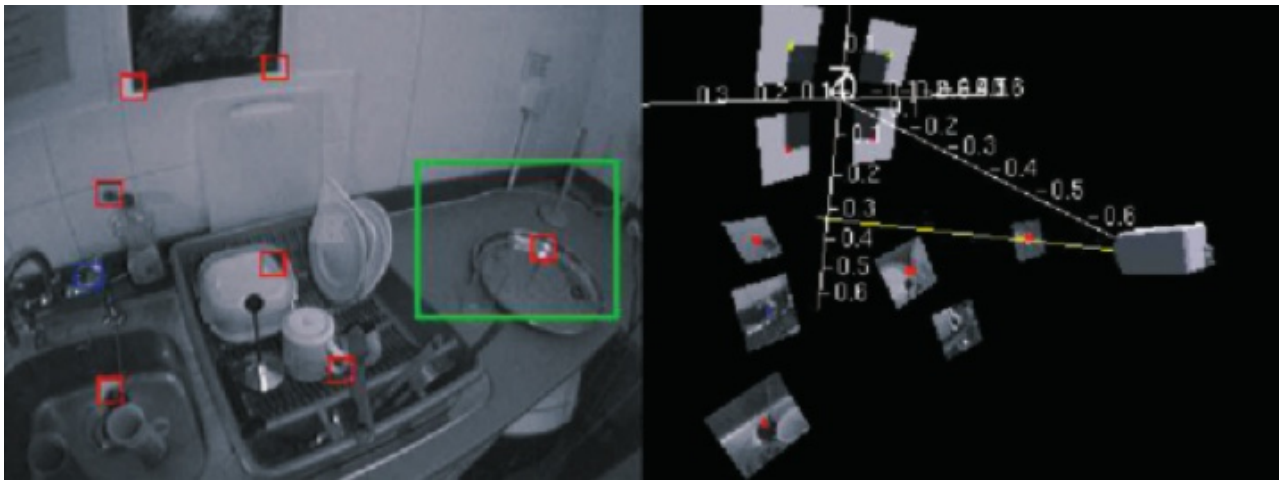

图 4 MonoSLAM效果图[7]

PTAM

PTAM[8]提出并实现了跟踪和建图的并行化，首次区分出前后端 (跟踪需要实时响应图像数据，地图优化放在后端进行)，后续许多视觉SLAM 系统设计也采取了类似的方法。PTAM 是第一个使用非线性优化作为后端的方案，而不是滤波器的后端方案。提出了关键帧 (keyframes)机制，即不用精细处理每一幅图像，而是把几个关键图像串起来优化其轨迹和地图。该方法的缺点是：场景小、跟踪容易丢失。

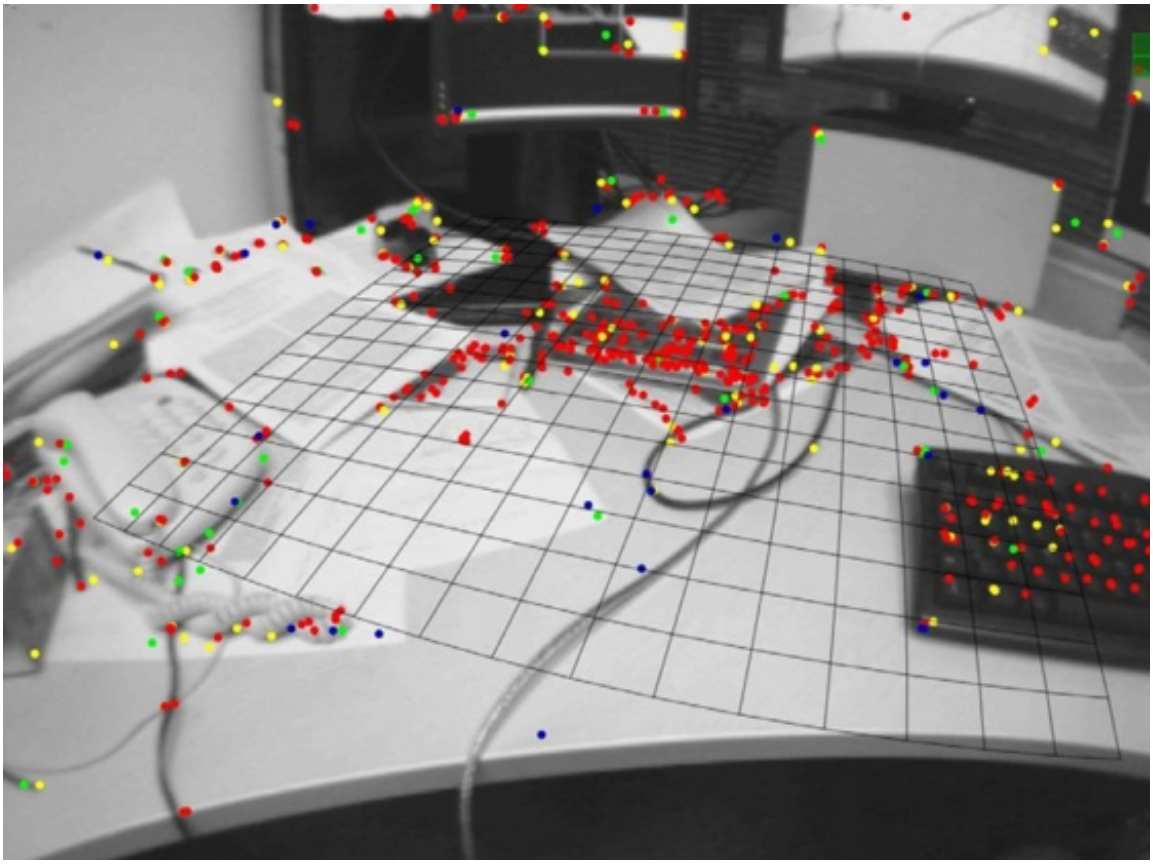

图 5 PTAM效果图（图片来源于文献[8]）

ORB-SLAM (ORB_SLAM2)

ORB-SLAM[9]围绕ORB 特征计算，包括视觉里程计与回环检测的 ORB字典。ORB 特征计算效率比 SIFT 或 SURF 高，又具有良好的旋转和缩放不变性。ORB-SLAM 创新地使用了三个线程完成 SLAM，三个线程是：实时跟踪特征点的Tracking线程，局部 Bundle Adjustment 的优化线程和全局 Pose Graph 的回环检测与优化线程。该方法的缺点：每幅图像都计算一遍 ORB 特征非常耗时，三线程结构给 CPU带来了较重负担。稀疏特征点地图只能满足定位需求，无法提供导航、避障等功能。

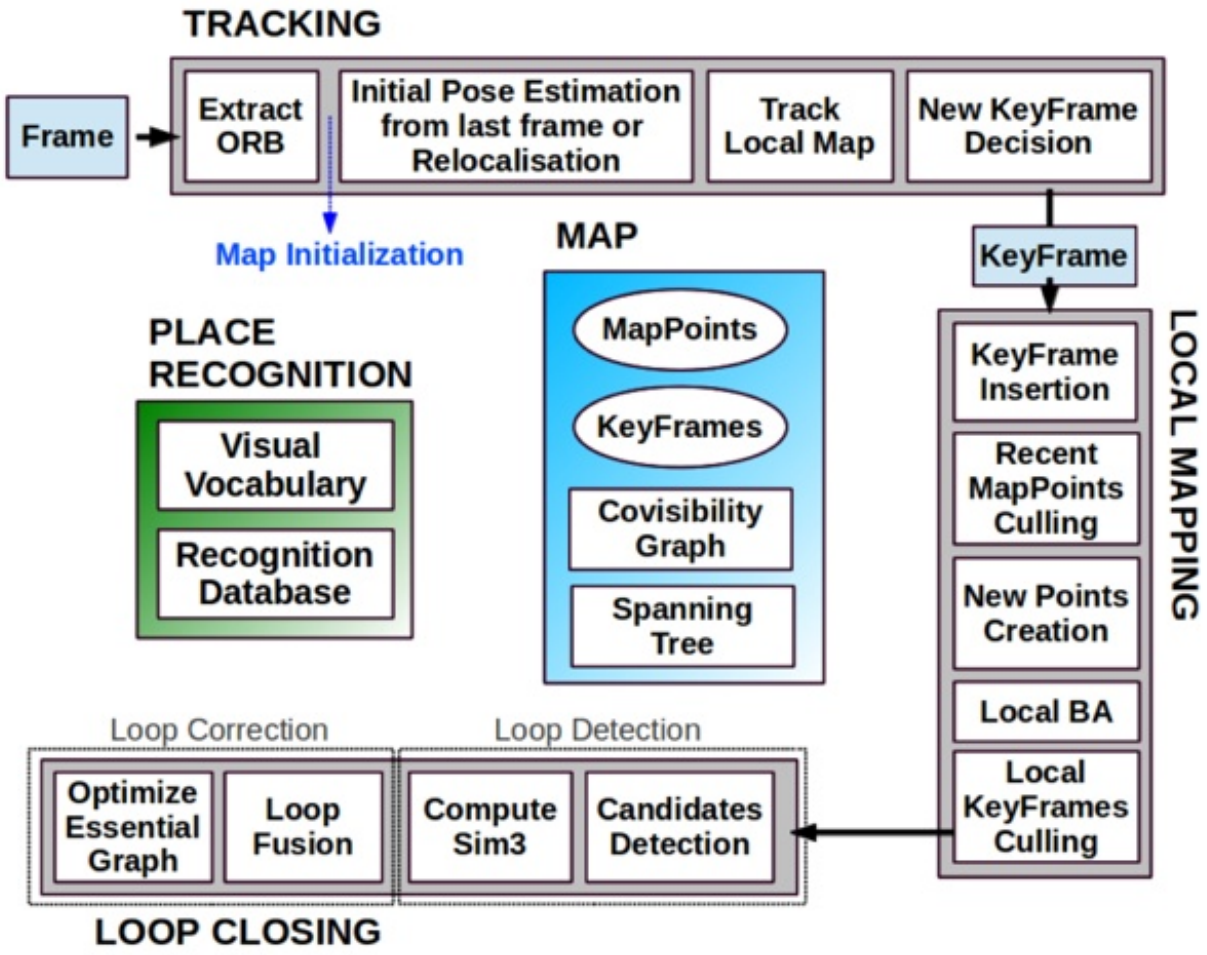

图 6 ORB-SLAM三线程流程图[9]

ORB-SLAM2[10]基于单目的 ORB-SLAM 做了如下贡献：第一个用于单目、双目和 RGB-D 的开源 SLAM 系统，包括闭环，重定位和地图重用; RGB-D 结果显示，通过使用 bundle adjustment，比基于迭代最近点(ICP) 或者光度和深度误差最小化的最先进方法获得更高的精度; 通过使用近距离和远距离的立体点和单目观察结果，立体效果比最先进的直接立体 SLAM 更准确; 轻量级的本地化模式，当建图不可用时，可以有效地重新使用地图。

直接法及其典型系统

直接法跳过预处理步骤直接使用实际传感器测量值，例如在特定时间内从某个方向接收的光，如下图所示。在被动视觉的情况下，由于相机提供光度测量，因此直接法优化的是光度误差：

$$T_{k-1,k} = \underset{T}{argmin} \sum_{i} \| I_K(u_i') - I_{k-1}(u_i) \|_\sigma^2$$

$$u_i' = \pi(T \cdot \ (\pi^{-1} \ (u_i) \cdot d))$$

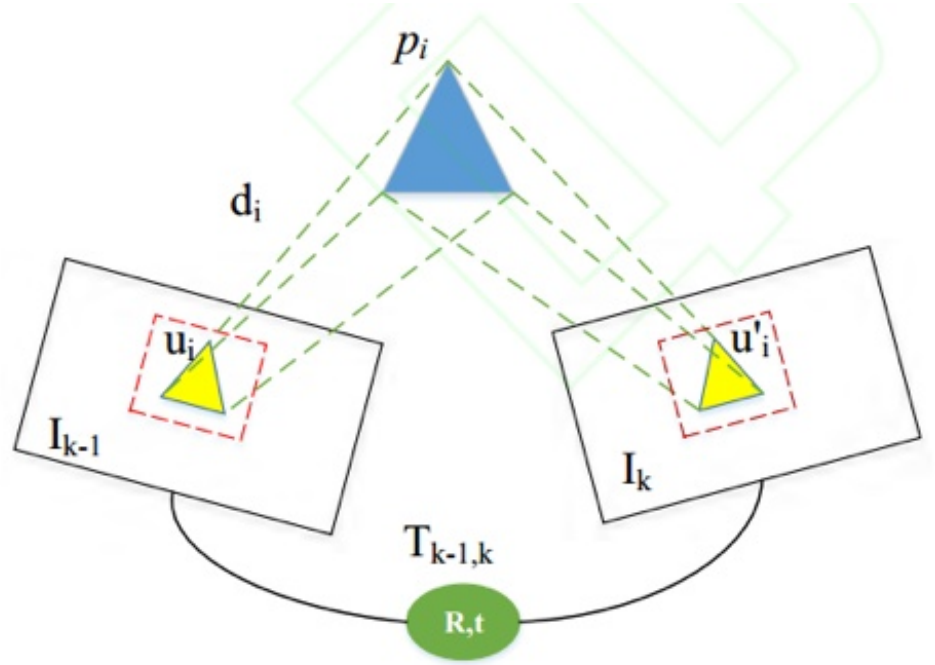

图 7 直接法示意图[3]

DTAM

DTAM[11]是单目 VSLAM 系统, 是一种直接稠密的方法，通过最小化全局空间规范能量函数来计算关键帧构建稠密深度图，而相机的位姿则使用深度地图通过直接图像匹配来计算得到。对特征缺失、 图像模糊有很好的鲁棒性。该方法的缺点是：计算量非常大，需要 GPU 并行计算。 DTAM 假设光度恒定，对全局照明处理不够鲁棒。

LSD-SLAM

LSD-SLAM[12]建了一个大尺度直接单目 SLAM 的框架，提出了一种用来直接估计关键帧之间相似变换、尺度感知的图像匹配算法，在CPU上实现了半稠密场景的重建。该方法的缺点：对相机内参敏感和曝光敏感，相机快速运动时容易丢失，依然需要特征点进行回环检测。

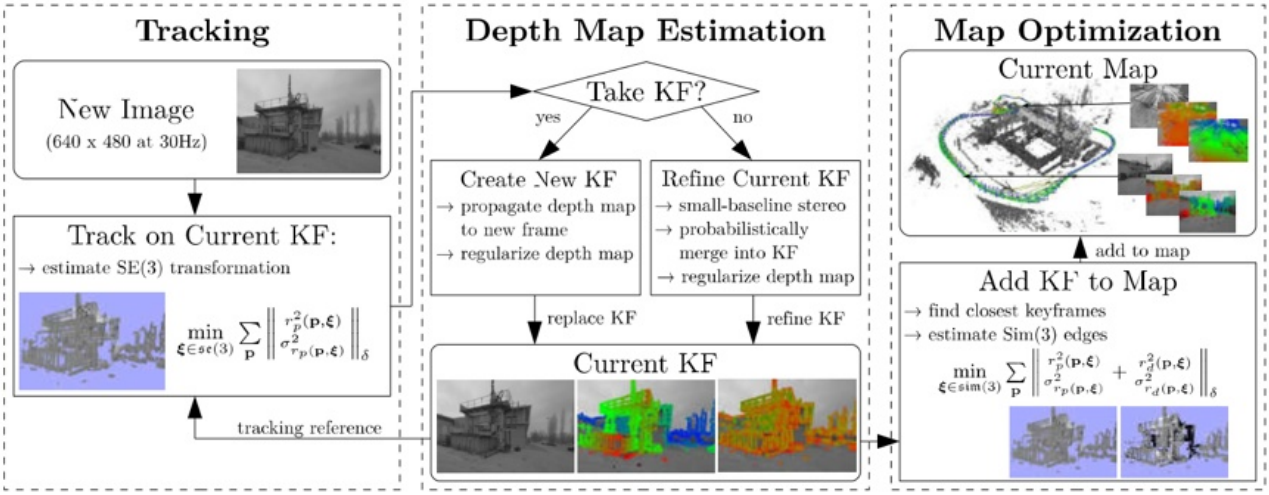

图 8 LSD-SLAM流程图[12]

SVO

SVO[13](Semi-direct Visual Odoemtry) 是一种半直接法的视觉里程计，它是特征点和直接法的混合使用：跟踪了一些角点，然后像直接法那样，根据关键点周围信息估计相机运动及位置。由于不需要计算大量描述子，因此速度极快，在消费级笔记本电脑上可以达到每秒 300 帧，在无人机上可以达到每秒 55 帧。该方法的缺点是：舍弃了后端优化和回环检测，位姿估计存在累积误差，丢失后重定位困难。

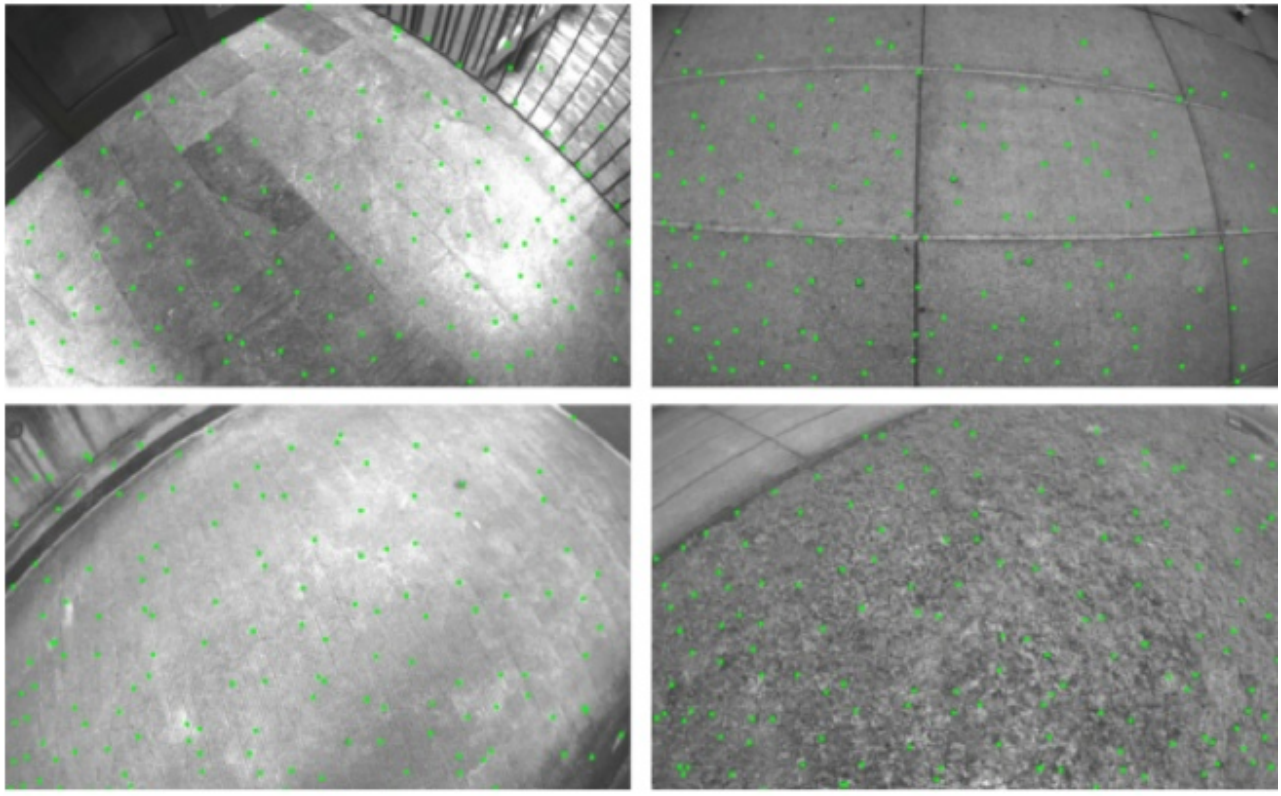

图 9 SVO效果示意图[13]

DSO

DSO[14](Direct Sparse Odometry) 是基于高度精确的稀疏直接结构和运动公式的视觉里程计的方法。不考虑几何先验信息，能够直接优化光度误差。并且考虑了光度标定模型，其优化范围不是所有帧，而是由最近帧及其前几帧形成的滑动窗口，并且保持这个窗口有 7 个关键帧。DSO 中除了完善直接法位姿估计的误差模型外，还加入了仿射亮度变换、 光度标定、 深度优化等。该方法没有回环检测。

图 10 DSO效果示意图[14]

基于深度学习的SLAM

传统的视觉SLAM在环境的适应性方面依然存在瓶颈，深度学习有望在这方面发挥较大的作用。目前，深度学习已经在语义地图、重定位、回环检测、特征点提取与匹配以及端到端的视觉里程计等问题上有了相关工作，下面列举一些典型成果：

●CNN-SLAM[17]在LSD-SLAM[12]基础上将深度估计以及图像匹配改为基于卷积神经网络的方法，并且可以融合语义信息，得到了较鲁棒的效果；

● 剑桥大学开发的PoseNet[18]，是在GoogleNet[19]的基础上将6自由度位姿作为回归问题进行的网络改进，可以利用单张图片得到对应的相机位姿；

● 《视觉SLAM十四讲》[5]一书的作者高翔，利用深度神经网络而不是常见的视觉特征来学习原始数据的特征，实现了基于深度网络的回环检测[20]；

● LIFT[21]利用深度神经网络学习图像中的特征点，相比于SIFT[22]匹配度更高，其流程图如下图所示：

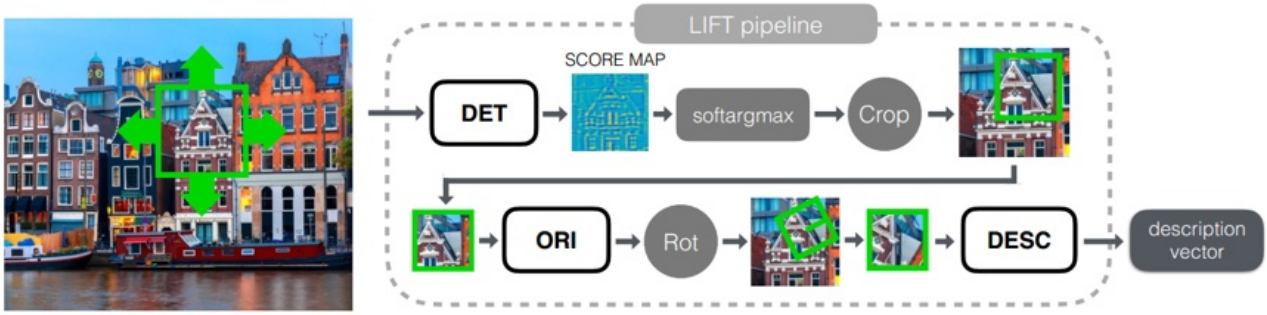

图 11 LIFT流程图[21]

LIFT(Learned Invariant Feature Transform)由三个部分组成：Detector，Orientation Estimator和Descriptor。每一个部分都基于CNN实现，作者用Spatial Transformers将它们联系起来，并用soft argmax函数替代了传统的非局部最大值抑制，保证了端到端的可微性。

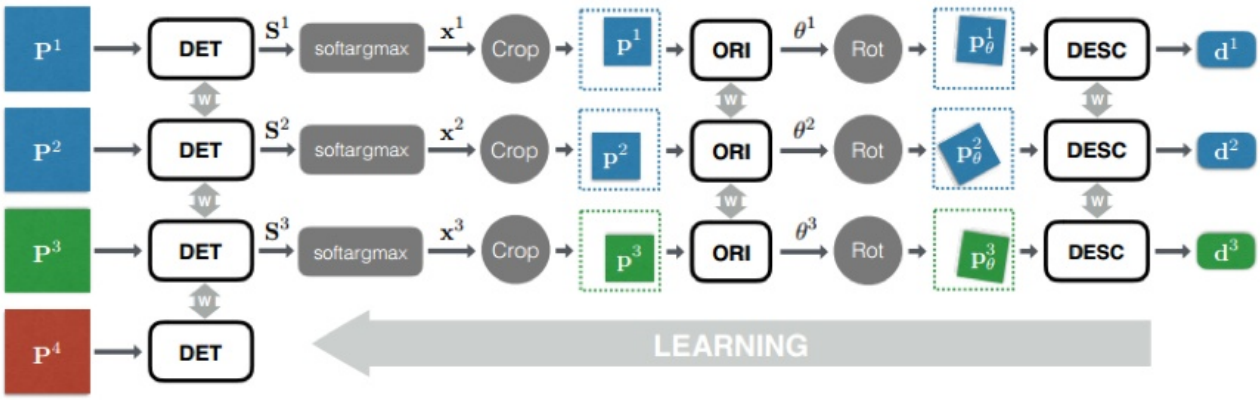

图 12 四分支孪生训练网络架构[21]

孪生训练架构（Siamese Network）包含四个分支，其中P1和P2对应同一个点的不同视角，作为训练Descriptor的正样本，P3代表不同的3D点，作为Descriptor的负样本；P4不包含特征点，仅作为训练Detector的负样本。由大P，Detector，softargmax和Spatial Transformer layer Crop共同得到的小p反馈到Orientation Estimator，Orientation Estimator和Spatial Transformer layer Rot提供pθ 给Descriptor，得到最终的描述向量d。作者给出了LIFT与SIFT特征匹配的效果对比。

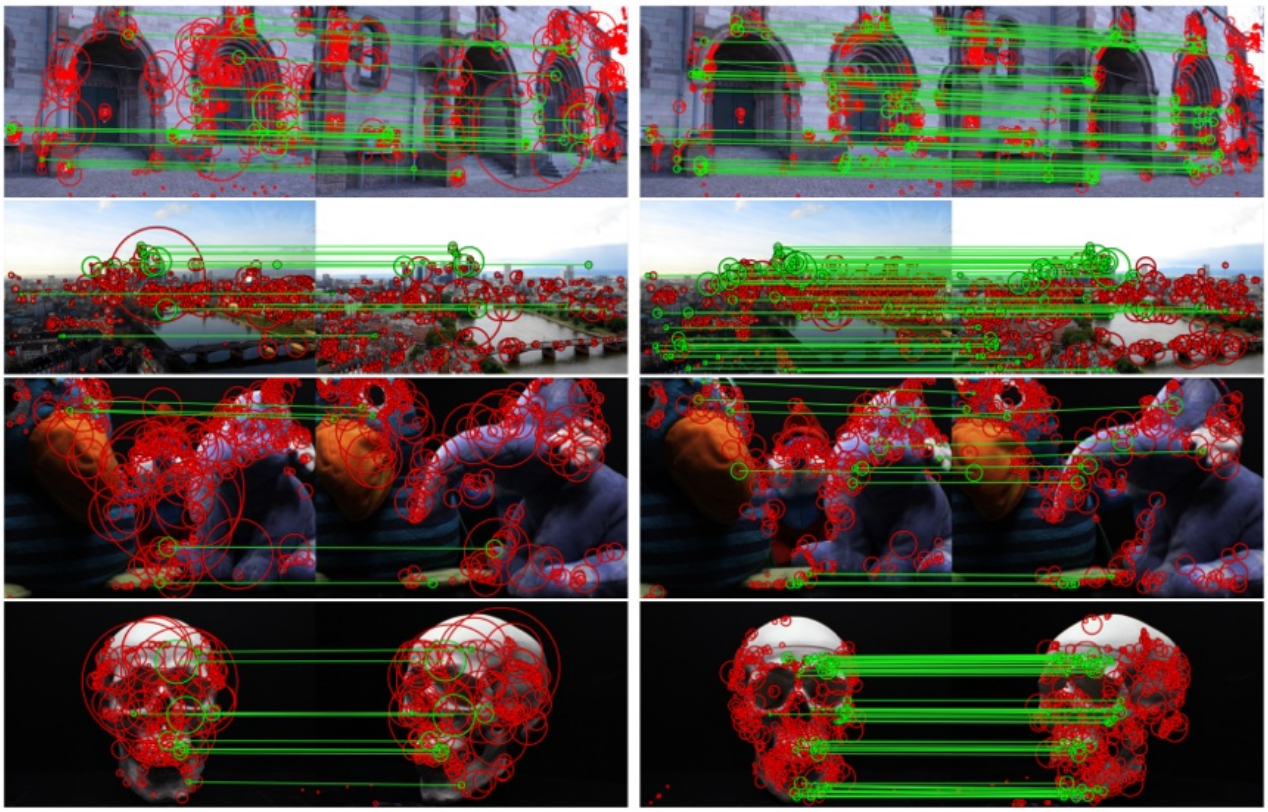

图 13 LIFT与SIFT特征匹配对比[21]

特征匹配对比图中左列为SIFT匹配结果，右列为LIFT。绿色线条代表正确匹配，红色圆圈代表描述子区域，可以看到，LIFT得到了比SIFT更稠密的匹配效果。

UnDeepVO[23]能够通过使用深度神经网络估计单目相机的6自由度位姿及其视野内的深度，整体系统框架概图见下图。

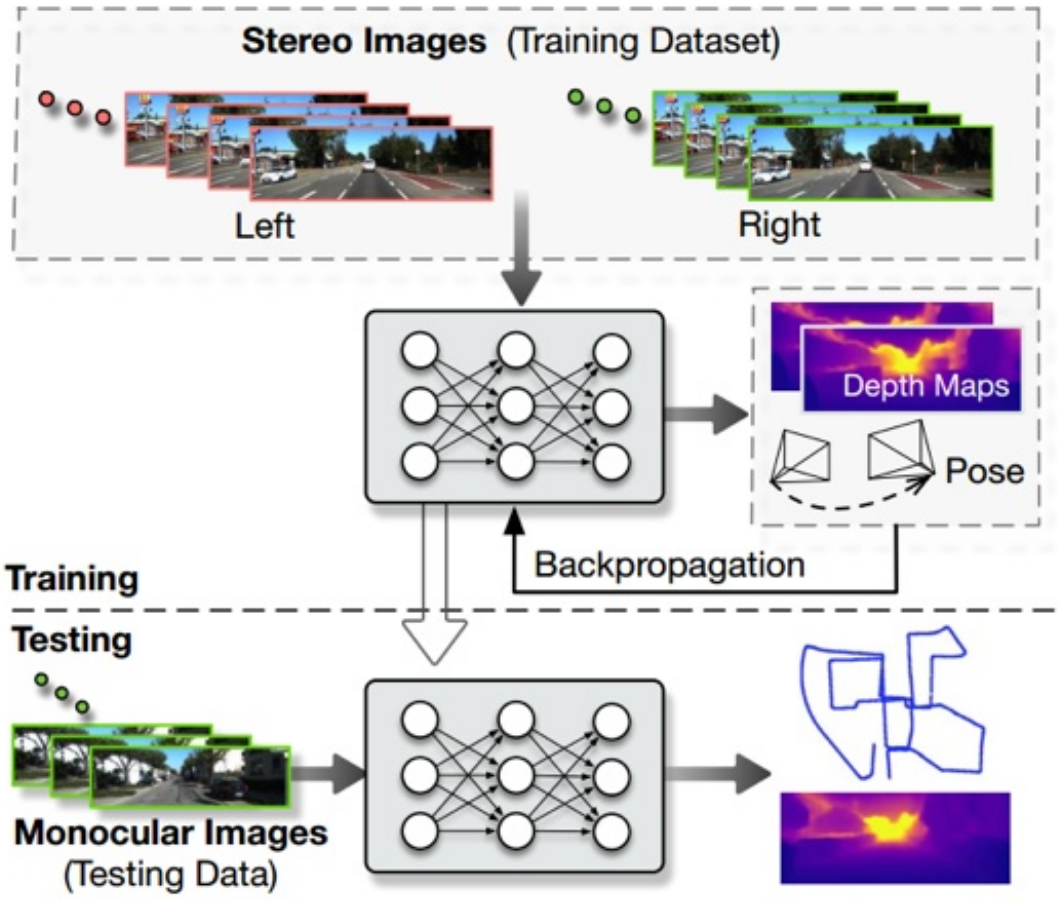

图 14 UnDeepVO系统框架概图[23]

UnDeepVO有两个显著的特点：一个是采用了无监督深度学习机制，另一个是能够恢复绝对尺度。UnDeepVO在训练过程中使用双目图像恢复尺度，但是在测试过程中只使用连续的单目图像。该文的主要贡献包括以下几点：

1.通过空间和时间几何约束，用无监督的方式恢复了单目视觉里程计的绝对尺度；

2.利用训练过程中的双目图像对，不仅估计了姿态还估计了稠密的带有绝对尺度的深度图；

3.在KITTI数据集上评价了该系统， UnDeepVO对于单目相机有良好的位姿估计结果。

UnDeepVO由位姿估计和深度估计构成，两个估计系统均把单目连续图像作为输入，分别以带有尺度的6自由度位姿和深度作为输出。对于位姿估计器，它是基于VGG的 CNN架构。 它把两个连续的单目图像作为输入，并预测它们之间的6自由度变换。 由于旋转（由欧拉角表示）具有高非线性，因此与平移相比通常难以训练。 在有监督训练中，一种常用的方法是将旋转损失作为一种归一化方式给予更大的权重。 为了使用无监督学习更好地训练旋转，作者在最后一个卷积层之后用两组独立的全连接层解耦平移和旋转。这使得作者能够引入一个权重来标准化旋转和平移的预测，以获得更好的性能。对于深度估计器，它基于encoder-decoder结构来生成稠密的深度图。 与其他深度估计方法不同的是，该方法从网络中产生视差图像（深度的倒数），UnDeepVO的深度估计器可以直接预测深度图，以这种方式训练整个系统更容易收敛。系统结构图如下图所示。

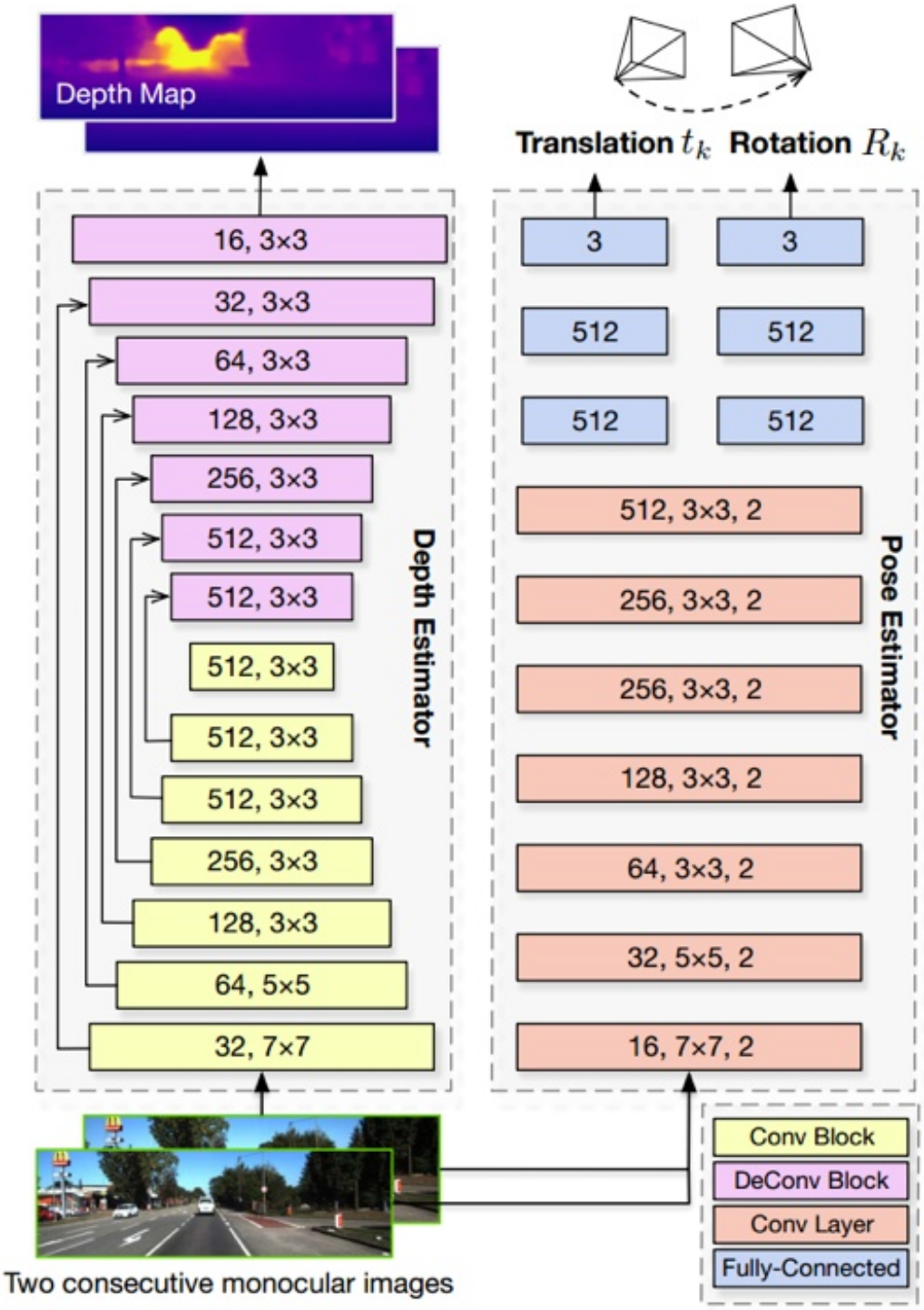

图 15 网络结构图[23]

总结

本文介绍了基于传统算法和深度学习的代表性SLAM方法。当前的 SLAM 算法在复杂的机器人运动和环境中很容易失效 (例如：机器人的快速运动， 高度动态性的环境)，通常不能面对严格的性能要求，例如，用于快速闭环控制的高速率估计。大多数的SLAM没有自由主动地收集数据，行动方案不够高效，并且，目前 vSLAM 方案中所采用的图像特征的语义级别太低，造成特征的可区别性太弱[15]。因此，今后的视觉SLAM将向着主动SLAM、语义SLAM以及与其它传感器（例如IMU）融合的方向发展。

参考文献

[1] Durrant-Whyte, H, and Bailey, Tim. ”Simultaneous Localization and Mapping: Part I.” IEEE Robotics & Amp Amp Automation Magazine 13.2(2006):99 - 110.

[2] Fuentes-Pacheco, Jorge, J. Ruiz-Ascencio, and J. M. Rendón-Mancha. ”Visual simultaneous localization and mapping: a survey.” Artifcial Intelligence Review 43.1(2015):55-81.

[3] 陈常, 朱华, 由韶泽. 基于视觉的同时定位与地图构建的研究进展 [J/OL]. 计算机应用研究,2018,(03):1-9(2017-08-18).

[4] Nister, D, O. Naroditsky, and J. Bergen. ”Visual odometry.” Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on IEEE, 2004:I-652-I-659 Vol.1.

[5] 高翔. 视觉 SLAM 十四讲 [M]. 北京: 电子工业出版社, 2017.

[6] 赵洋等. ” 基于深度学习的视觉 SLAM 综述.” 机器人 39.6(2017):889-896.

[7] Davison, Andrew J., et al. ”MonoSLAM: Real-time single camera SLAM.” IEEE transactions on pattern analysis and machine intelligence 29.6 (2007): 1052-1067.

[8] Klein, Georg, and David Murray. ”Parallel tracking and mapping for small AR workspaces.” Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on. IEEE, 2007.

[9] Mur-Artal, Raúl, J. M. M. Montiel, and J. D. Tardós. ”ORB-SLAM: A Versatile and Accurate Monocular SLAM System.” IEEE Transactions on Robotics 31.5(2015):1147-1163.

[10] Mur-Artal, Raul, and Juan D. Tardós. ”Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras.” IEEE Transactions on Robotics 33.5 (2017): 1255-1262.

[11] Newcombe, Richard A, S. J. Lovegrove, and A. J. Davison. ”DTAM: Dense tracking and mapping in real-time.” International Conference on Computer Vision IEEE Computer Society, 2011:2320-2327.

[12] Engel, Jakob, T. Schöps, and D. Cremers. ”LSD-SLAM: Large-Scale Direct Monocular SLAM.” 8690(2014):834-849.

[13] Forster, Christian, M. Pizzoli, and D. Scaramuzza. ”SVO: Fast semi-direct monocular visual odometry.” IEEE International Conference on Robotics and Automation IEEE, 2014:15-22.

[14] Engel, Jakob, V. Koltun, and D. Cremers. ”Direct Sparse Odometry.” IEEE Transactions on Pattern Analysis & Machine Intelligence PP.99(2016):1-1.

[15] Cadena, Cesar, et al. ”Past, Present, and Future of Simultaneous Localization and Mapping:Toward the Robust-Perception Age.” IEEE Transactions on Robotics 32.6(2016):1309-1332.

[16] 吕霖华. 基于视觉的即时定位与地图重建(V-SLAM)综述[J]. 中国战略新兴产业, 2017(4).

[17] Tateno K, Tombari F, Laina I, et al. CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, 2.

[18] Kendall A, Grimes M, Cipolla R. Posenet: A convolutional network for real-time 6-dof camera relocalization[C]//Proceedings of the IEEE international conference on computer vision. 2015: 2938-2946.

[19] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

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

[21] Yi K M, Trulls E, Lepetit V, et al. Lift: Learned invariant feature transform[C]//European Conference on Computer Vision. Springer, Cham, 2016: 467-483.

[22] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.

[23] Li R, Wang S, Long Z, et al. Undeepvo: Monocular visual odometry through unsupervised deep learning[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018: 7286-7291.

编辑于 2019-10-21