第12周汇报

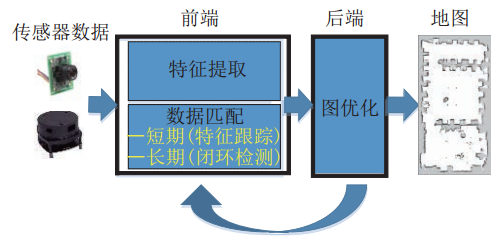
——张溢炉

1. 《基于深度学习的视觉 SLAM 综述》总结
2. 内容摘要

本文介绍了同时定位与地图构建（SLAM），介绍了一些传统的方法和深度学习方法

1. 总结及理解

* 典型SLAM系统前后端如下图，传统的方法包括运动方程和观测方程，运动方程利用两帧图像计算相机位姿，观测方程用于后端优化。单目摄像头较难计算像素点的深度，需要通过两帧图像计算；双目可以通过一帧图像计算深度；RGB-D摄像头直接得到深度
* 帧间估计：也称视觉里程计，通过前后两帧图像计算摄像头位姿
* 后端优化：优化相机运动轨迹，减小噪声的干扰和累积误差
* 闭环检测：判断自身是否进入历史同一地点．闭环检测发生时可触发 SLAM 后端全局一致性算法进行地图优化，消除累积轨迹误差和地图误差．闭环检测问题本质上是场景识别问题
* 地图构建：利用传感器得到的数据建立三维地图

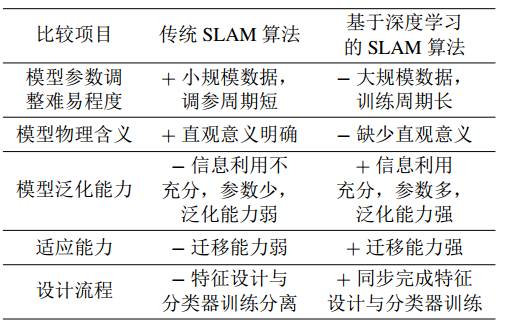


* 人工设计的稀疏图像特征当前有很多局限性，一方面如何设计稀疏图像特征最优地表示图像信息依然是计算机视觉领域未解决的重要问题，另一方面稀疏图像特征在应对光照变化、动态目标运动、摄像机参数改变以及缺少纹理或纹理单一的环境等方面依然有较多挑战
* 目前深度学习结合SLAM有3个方向
* 用深度学习方法替换传统SLAM一个或几个模块特征提取，如特征匹配，提高特征点稳定性，提取点线面等不同层级的特征点，深度估计，位姿估计，重定位；
* 在传统SLAM之上加入语义信息，如图像语义分割，语义地图构建；
* 端到端的SLAM，端到端是输入image输出action，没有定位和建图。 如机器人自主导航（深度强化学习）等。

本文介绍了基于深度学习的帧间估计、闭环检测和语义地图生成，目前深度学习多应用于 SLAM 局部的子模块，如定位模块或闭环检测模块

* SLAM如果能和一些CV、nlp任务结合起来，会产生一些比较好玩的任务，比如：
* 建图时利用图像描述生成一段自然语言描述，并保存当前帧，直到下一帧与当前帧的相似度低于某个阈值时，再生成一段描述，当语句到达一定数量或者描述完一个场景，再利用nlp中生成结尾的任务生成一段总的描述
* 建图时识别图中物体，然后从外部知识获取物体的属性（如颜色、是否可移动等）、物体间的关系，建立某空间的语义图谱
* 与EQA任务结合，SLAM可以实时建图，可以更新某个空间被移动过的物体，但是实时建的图会有很多噪声，不像仿真环境中那么干净

以上任务在小范围比较合适，如家里、仓库等

* 传统方法与深度学习方法比较如下图

1. 《Ongoing Evolution of Visual SLAM from Geometry to Deep Learning:Challenges and Opportunities》总结
2. 内容摘要

本文介绍了视觉SLAM的一些传统的方法和与深度学习结合的方法，讲述了深度学习应用到视觉SLAM上面临的一些机遇和挑战

1. 总结及理解

* 传统的方法如下：
* 基于特征（Feature-Based）：首先通过特征匹配找出两帧图像相同的像素点，可以通过关键帧的位姿计算这些像素点的世界坐标，然后利用几何算法计算相机位姿。因为相机可以看成一个小孔成像的模型，对于单目相机具有尺度不确定性，需要通过两帧图像来计算像素点的深度。整个计算过程中默认前一帧图像的结果是对的，所以会有累积误差，这就需要利用后端和闭环检测来减小累积误差。后端采用图优化的方法，将一些路标点投影到相机成像画面，通过一些算法（如BA）使投影点和实际点尽可能接近。比较有代表性的特征有SIFT、SURF和ORB，目前ORB用的比较多。得到位姿和深度后，建图还是比较容易，单目使用这种方法只能建立稀疏的三维图
* 直接法：也称光流法，有稠密和半稠密的方法。假设相邻两帧图像的像素值不变，通过梯度下降的方法调整位姿，使两幅图像对应像素点的方差最小，这就要用到李代数，因为旋转公式有约束条件，不好做梯度下降。可以进行稠密和半稠密建图。目前比较好的方法是Direct Sparse Odometry (DSO)
* CNN、RNN、LSTM、encoder-decoder、无监督学习在视觉SLAM方面都有应用，引入了CV和NLP方面的大多数模型。文中主要提到深度学习在SLAM的深度估计、位姿估计、语义建图（如下图）三个方面结合比较多。
* 数据集
* 深度估计：KITTI [35], TUM [40], NYU [41]
* 重定位：7-scenes dataset [46]，Cambridge landmarks [33]
* 视觉里程计（VO）：KITTI [35], Robotcar [38],M’alaga [47], EuRoc MAV [39], NYU [41],TUM [40]
* 场景分割：PASCAL VOC [42], NYU [41], Synthia [43],Cityscapes [37], KITTI [35],ADK20 [45]
* 传统的方法在定位和建图已经具有比较好的准确性和实时性，但鲁棒性不太好，会面临特征稀少、光强变化、尺度不确定性、相机内参漂移等问题，将深度学习应用到SLAM可提高其鲁棒性、语义信息以及学习能力，可以使robots更好地与环境交互。CNN可以提取大量的隐性图像特征，具有尺度不变性、旋转不变性。
* 监督学习需要数据集来驱动，但不同robots面临不一样的未知环境，很难建立通用的数据集，建立数据集也需要花费大量的人力和时间，这是当前面临的瓶颈。无监督学习、传统方法与深度学习的结合、多传感器（雷达、超声波、激光、陀螺仪）的融合有利于解决目前面临的一些问题。

1. 《Visual SLAM for Automated Driving:Exploring the Applications of Deep Learning》总结
2. 内容摘要

本文介绍了视觉SLAM在自动驾驶上的应用，很多内容和之前两篇相似，在这就不再重复

1. 总结及理解

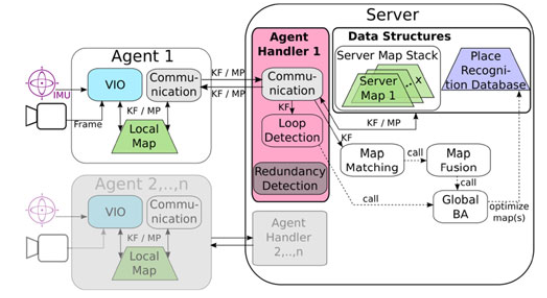
* 视觉SLAM比较适合自动驾驶的以下三个情景
* 停车：停车时要识别周围物体，选择车位，避开障碍物。在外面停车每次都得重新建图，对于生活中常去的停车的地方，如家里，可以将图存储下来，方便下次使用，虽然每次物体放置都不一样，但停车场整体结构一般不会变，有助于重定位
* 高速路：高速路需要比较高的实时性，每秒需要处理30张图片，传统方法实时性较好
* 城市驾驶：城市环境比较复杂，移动物体较多，DSO-SLAM表现较好
* 自动驾驶对安全性要求很高，因为一个失误可能就是人命，似乎很难找到一个置信度百分之百的深度学习模型，需要综合各种方法来防止意外的发生。因为世界各地环境都不相同，也很难找到一个通用的数据集
* 文后还提到了一些CNN Based Pipelines，如Tateno et al. [54]提出联合学习语义分割和建立深度地图

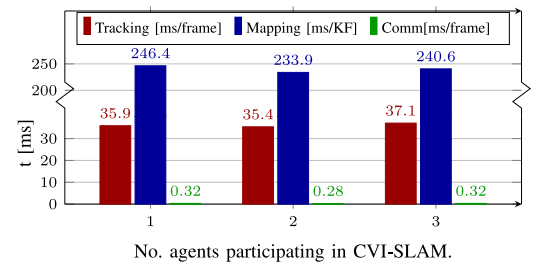
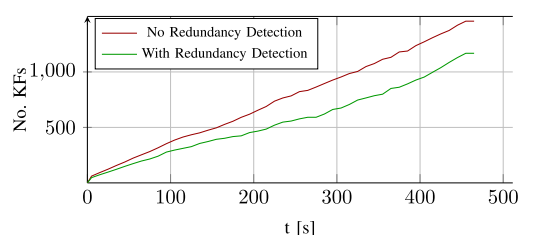
1. 《CVI-SLAM—Collaborative Visual-Inertial SLAM》总结
2. 内容摘要

本文介绍了多个智能体协作的Visual-Inertial SLAM，每个智能体上搭载了摄像头和Inertial Measurement Unit（IMU），与中央处理器协同定位和建图

1. 总结及理解

* 整体架构如下图所示，agent负责处理实时性较高的计算，如定位。Server负责处理计算量大且实时性不高的计算，如冗余检测、闭环检测、地图融合、BA。这样可以利用server强大计算力，也可以保持agent相对独立，但是这对两者之间的通信速度和质量要求比较高，图像的数据量比较大。从本文实验结果来看，该方法在agent数量上具有较好的可拓展性，对网速的需求可以接受。



* 下图为不同agent在frame tracking, local mapping of the VIO, communication所用时间，通信时间相对处理时间还是比较少的，但随着agent的增加，agent的处理量不变，但通信压力会变大，server的处理数据量会线性增大
* 下图是在4个agent情况下有无冗余检测模块随着时间增长待处理关键帧的数量变化图，由图可见冗余检测可以有效减少待处理关键帧的数量，减少server的处理压力
* 从本文大致了解了CVI-SLAM的方法，里面涉及的很多公式还需以后继续弄明白