基于单目视觉的同时定位与地图构建方法

1 Induction

1.1 增强现实技术

增强现实技术 (**augmented reality**) 是一种将真实世界信息和虚拟世界信息“无缝”集成的新技术，是把原本在现实世界的一定时间空间范围内很难体验到的实体信息 (视觉信息, 声音, 味道, 触觉等), 通过电脑等科学技术，模拟仿真后再叠加，将虚拟的信息应用到真实世界，被人类感官所感知，从而达到超越现实的感官体验。比起传统方式来说，它更加的直观，更加的高效，因此也有着更加广阔的应用前景。近年来，增强现实技术已在军事，生活，游戏等众多领域运营并取得了成功。例如，宜家家居公司已经开发了一个 APP 使得用户可以使用智能手机观察不同的家具在自己房间的摆放效果；而任天堂公司也开发了 Pokemon-Go 游戏，使得玩家可以通过智能手机在现实世界里发现精灵。

1.2 增强现实的定位方式

增强现实需要实时定位设备在环境中的方位，定位的方案虽然有许多种，但多数方案都存在局限或者代价太高难以普及，例如 GPS 无法在室内及遮挡严重的环境里使用，且精度较低，而基于无线信号的定位方案则需要事先布置场景。基于视觉的同时定位与地图构建技术（*visual simultaneous localization and mapping* **V-SLAM**）以其成本低廉、小场景精度较高、无需预先布置场景等优势成为比较常采用的定位方案。

1.3 V-SLAM 技术

V-SLAM 技术指的是使用图像作为外部信息的唯一来源，来定位一个机器人、一辆车或者一个移动的相机在整个场景中的位置，同时，重建环境的三维结构。

1.3.1 V-SLAM 的基本原理

V-SLAM 技术根据拍摄的视频、图像信息推断摄像头在环境的方位，同时构建环境地图，其原理为多视图几何原理 (**Multiple view geometry theroy**) V-SLAM 的目标为同时恢复出每帧图像对应的相机运动参数 C_1,C_2 $\cdots C_m$ 以及场景三维结构 X_1,X_2 $\cdots X_n$, 每个相机运动参数 C_i 包含了相机的位置和朝向信

息，通常表达为一个 3×3 的旋转矩阵 R_i 和一个三维位置变量 p_i。R_i 与 p_i 将一个世界坐标系下的三维点 X_j 变换至 C_i 的局部坐标系

$$(X_{ij}, Y_{ij}, Z_{ij})^{\mathrm{T}} = R_i(X_j - p_i) \tag{1}$$

进而投影至图像中

$$h_{ij} = (f_x X_{ij}/Z_{ij} + c_x . f_y Y_{ij}/Z_{ij} + c_y)^{\mathrm{T}} \tag{2}$$

其中，f_x,f_y 分别为沿图像 x,y 轴的图像焦距,(c_x, c_y) 为镜头光心在图像中的位置，通常假设这些参数已实现标定且保持不变，由式 (1) (2)，三维点在图像中的投影位置 h_{ij} 可表示为一个关于 C_i 和 X_j 的函数，记为

$$h_{ij} = h(C_i, X_j) \tag{3}$$

V-SLAM 算法需要将、对不同图像中对应于相同场景的图像点进行匹配，而这个过程是通过求解如下目标函数

$$\underset{C_1,\cdots C_m, X_1,\cdots X_n}{\arg\min} \sum_{i=1}^{m}\sum_{j=1}^{n} \|h(C_i, X_j) - \tilde{x_{ij}}\|_{\Sigma_{ij}} \tag{4}$$

得到一组最优的 C_1，$C_2 \cdots C_m$,X_1,X_2 $\cdots X_n$，使得所有 X_j 在 C_i 图像中的投影位置 h_{ij} 与观测到的图像点位置 x_{ij} 尽可能靠近，这里假设图像观测点符合高斯分布 $x_{ij} \sim N(\tilde{x_{ij}}, \Sigma_{ij})$，$\| e \| = e^{\mathrm{T}} \Sigma^{-1} e$ 求解目标函数(4)的过程也成为集束调整（**bundle adjustment,BA**），该最优化问题可利用线性方程的稀疏结构高效求解。

1.3.2 基于关键帧 BA 的单目 V-SLAM 系统

由于现阶段大多数 AR 产品都以智能手机以及平板电脑作为载体，而智能手机的摄像头大多以单目为主，双目、三目摄像头甚至深度摄像头都未得到普及，因此本文主要讨论基于单目视觉的同时定位与地图构建方法。目前，主流的 V-SLAM 方法主要为：基于滤波器、基于关键帧 BA 和基于直接跟踪，我们先来看看这三种方法。比较并分析其优劣，而后详细介绍基于关键帧 BA 的 V-SLAM 方法。其中比较具有代表性的有 MonoSLAM 以及 MSCKF

基于滤波器的 V-SLAM 的方法将系统每一时刻的状态 t 用一个高斯概率模型表达，$x_t \sim N(\tilde{x_t}, P_{ij})$, 其中 $\tilde{x}_t$ 为当前时刻系统状态估计值，P_t 为该估计值误差的协方差矩阵，系统状态由滤波器不断更新。

而基于关键帧 BA 的 V-SLAM 方法是近年来最流行的方法之一，他的主要思想是将相机跟踪（Tracking）和地图构建（Mapping）作为两个独立的任务在两个

线程并行执行，而 Mapping 线程仅维护视频流中抽取的关键帧。PTAM 是最著名的基于关键帧 BA 的方法之一，也是我们介绍的重点

基于直接跟踪的 V-SLAM 方法则是直接通过比较像素颜色来求解相机运动，具有代表性的算法有 DTAM 以及 LSD-SLAM。