SLAM | 视觉SLAM中的前端：视觉里程计与回环检测

标题以下，全是干货

什么是SLAM?

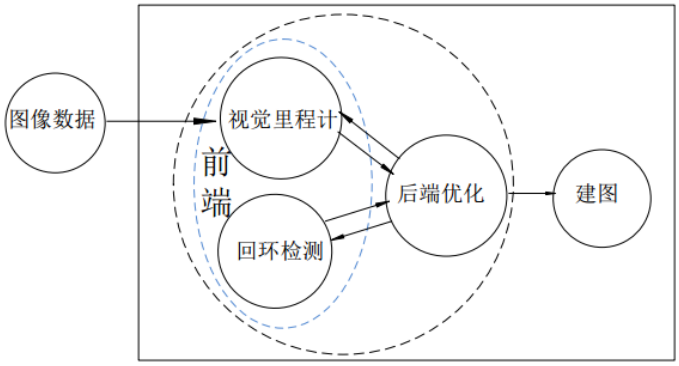
同时定位与地图构建 （simultaneous localization and mapping， SLAM）是**机器人进入未知环境遇到的第一个问题。**它是指机器人搭载特定传感器，在**没有环境先验信息的情况下**，于运动过程中**对周围环境建模并同时估计自身的位姿。**如果传感器主要为相机，那么就称为视觉SLAM（VSLAM）。SLAM 技术已经研究和发展了三十多年，研
究人员已经做了大量的工作，近十年来，随着计算机视觉的发展，VSLAM 以其硬件成本低廉、轻便、高精度等优势获得了学术界和工业界的青睐。

接下来，我们将有一系列的文章带大家来领略SLAM的魅力。

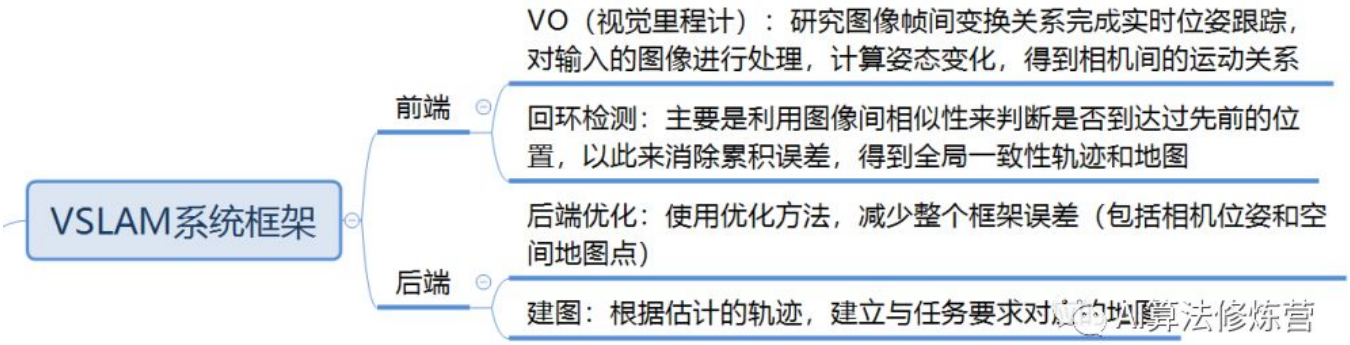
视觉SLAM系统框架

VSLAM 是利用多视图几何理论，根据相机拍摄的图像信息对相机进行定位并同时构建周围环境地图。按照相机的分类，有单目、双目、 RGBD、鱼眼、全景等。

VSLAM 主要包括**视觉里程计（visual odometry， VO）、后端优化、回环检测、建图。** 其中VO 研究图像帧间变换关系完成实时的位姿跟踪， 对输入的图像进行处理，计算姿态变化， 得到相机间的运动关系；但是随着时间的累计，误差会累积，这是由于仅仅估计两个图像间的运动造成的，后端主要是使用优化方法，减小整个框架误差（包括相机位姿和空间地图点）。回环检测，又称闭环检测，主要是利用图像间的相似性来判断是否到达过先前的位置，以此来消除累计误差，得到全局一致性轨迹和地图。建图，根据估计的轨迹，建立与任务要求对应的地图。

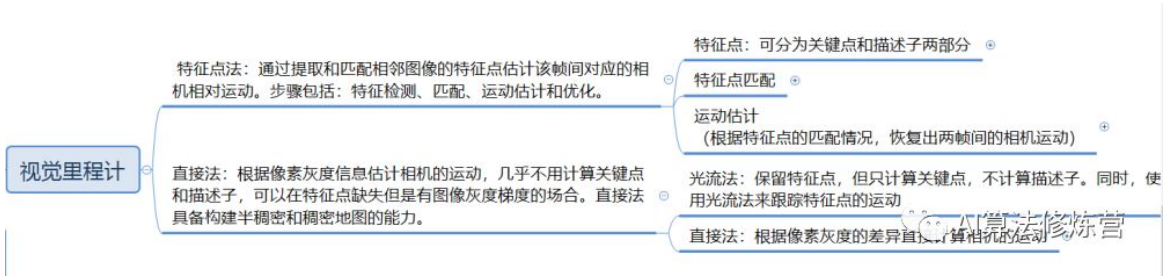


现在比较通常的惯例是把 VSLAM 分为**前端和后端。前端为视觉里程计和回环检测**，相当于是对图像数据进行关联；**后端是对前端输出的结果进行优化**，利用滤波或非线性优化理论，得到最优的位姿估计和全局一致性地图。

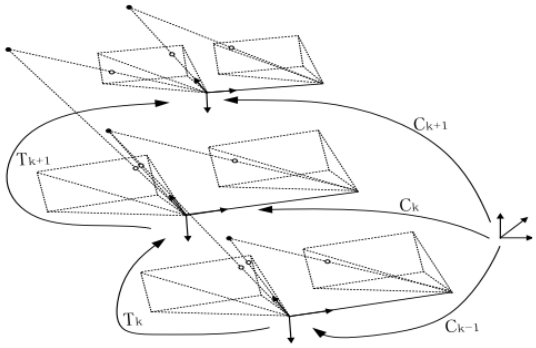


1 前端：图像数据的关联

1.1 视觉里程计



前端中的视觉里程计， 为通过采集的图像得到相机间的运动估计，视觉里程计问题可由下图描述（双目立体视觉里程计）。



视觉系统在运动过程中，在不同时刻获取了环境的图像，而且**相邻时刻的图像必须有足够的重叠区域，**则视觉系统的**相对旋转和平移运动**可被估算出来，然后将每两个相邻时刻之间视觉系统的运动

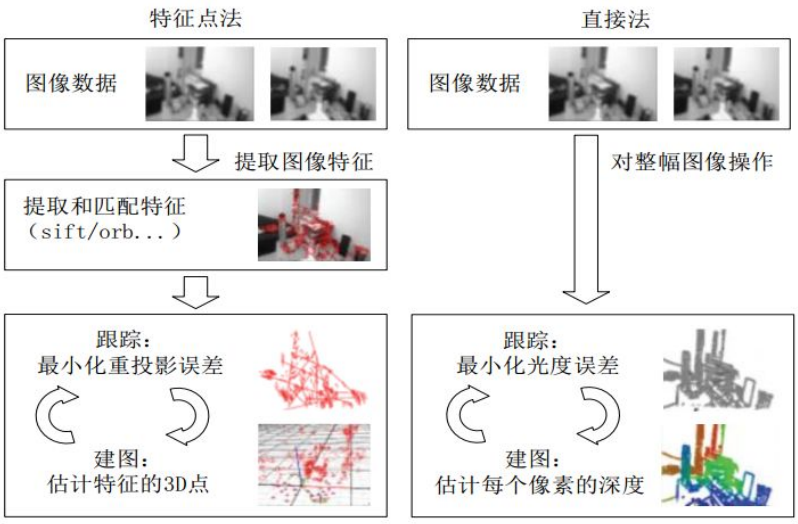
串联起来,可以得到**累计的视觉系统相对于参考坐标系的旋转和平移。**

如上图所示， 视觉里程计的任务就是已知 k = 0 的初始位置 C0 （这可以根据情况自己定义），求相机的运动轨迹当前的位置 Ck 通过 Tk 和上一时刻的位置Ck-1 来计算，Tk 为 K 和 K+1 时刻的相机相对位置变化，可根据相应时刻采集的图像计算出来，从而恢复相机的运动轨迹。

视觉里程计可分为**特征点法和直接法。**

特征点法主要是根据图像上的特征匹配关系得到相邻帧间的相机运动
估计，它**需要对特征进行提取和匹配**，然后根据匹配特征**构建重投影误差函数**，并将其最小化从而得到相机的相对运动；

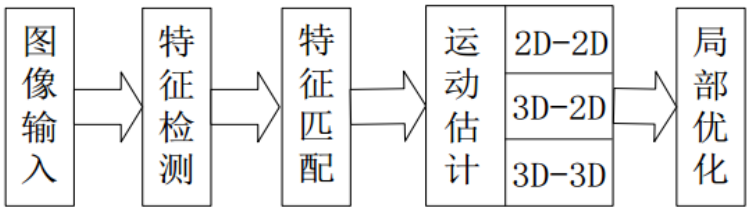
直接法是假设两帧图像中的**匹配像素的灰度值不变**，构建**光度误差函数**， 也将其最小化求解帧间的相机运动。



特征点法和直接法的直观对比

1.1.1 特征点法

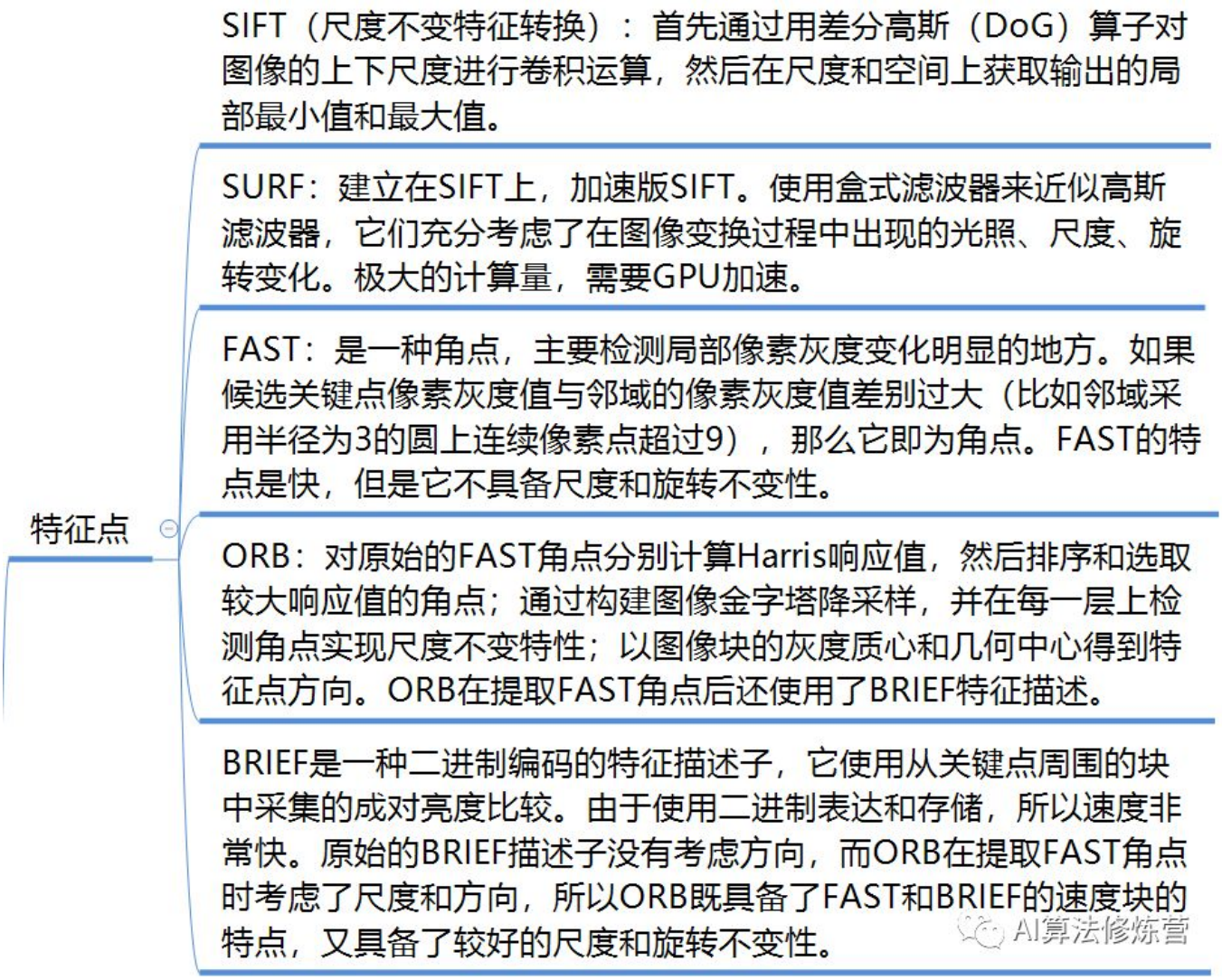
特征点法的原理为：通过**提取和匹配相邻图像的特征点**估计该帧间对应的相机相对运动。特征点法的步骤包括**特征检测、匹配、运动估计和优化。**



1．特征检测与匹配

特征点可以称为兴趣点、显著点、关键点等。以点的位置来表示的点特征是一种最简单的图像特征。 **特征点可以分为关键点和描述子两部分。**事实上，特征点是一个具有一定特征的局部区域的位置标志，称其为点，是将其抽象为一个位置概念，以便于确定两幅图像中同一个位置点的对应关系，所以在特征匹配过程中是以该特征点为中心，将邻域的局部特征进行匹配。

也就是说**在进行特征匹配时首先要为这些特征点建立特征描述，这种特征描述通常称之为描述子。**



特征点匹配

早期多采用跟踪方式，比如采用光流跟踪得到关键点的匹配。通常为了排除误跟踪，可以采用一致性检测。这种适合相邻帧之间的运动量和外观变化较小的情况。

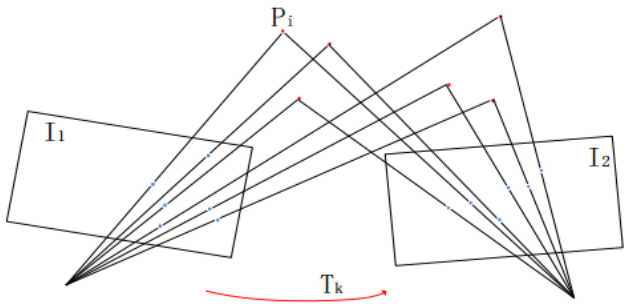
如果两帧之间的运动量和外观变化较大，需要计算两帧之间的特征点和描述子，比较描述子之间的距离（如汉明距离）。由于计算量关系，多采用恒速模型在预期区域中搜索潜在的对应关系。如果是双目匹配或者深度滤波器中计算每个像素的深度，通常采用极线搜索和采用归一化互相关（NCC）或绝对误差和（SSD）找到匹配点。

2．运动估计

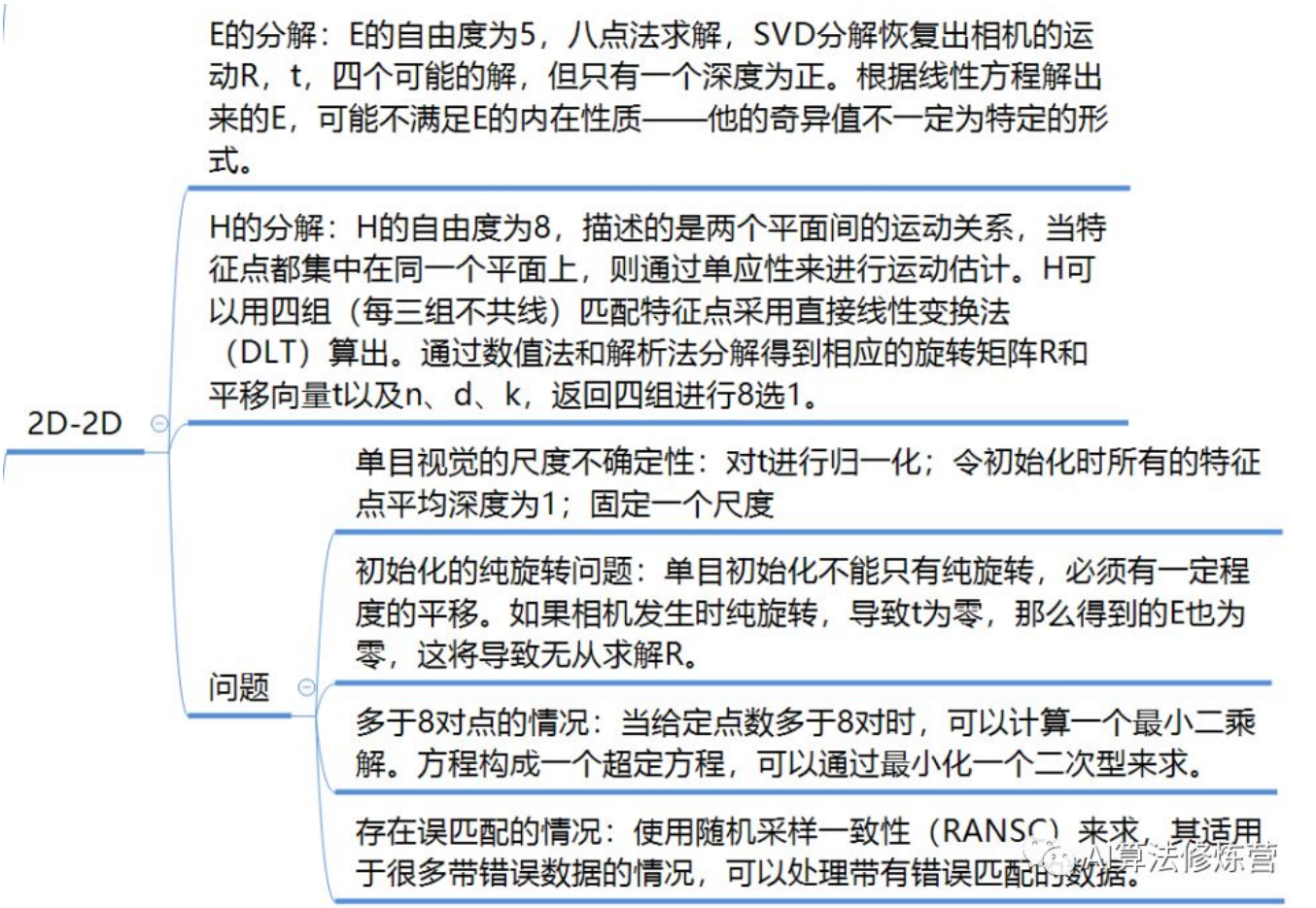
运动估计就是**根据特征点的匹配情况，恢复出两帧间的相机运动。**针对特征点匹配的情况，运动估计分为 2D-2D、3D-2D、 3D-3D。其求解方法可以分为几何方法和优化方法。**几何方法主要是根据对极几何理论得到两帧间的对应关系；而优化方法主要是构建两帧间的重投影误差并使其最小，从而得到帧间变换。**

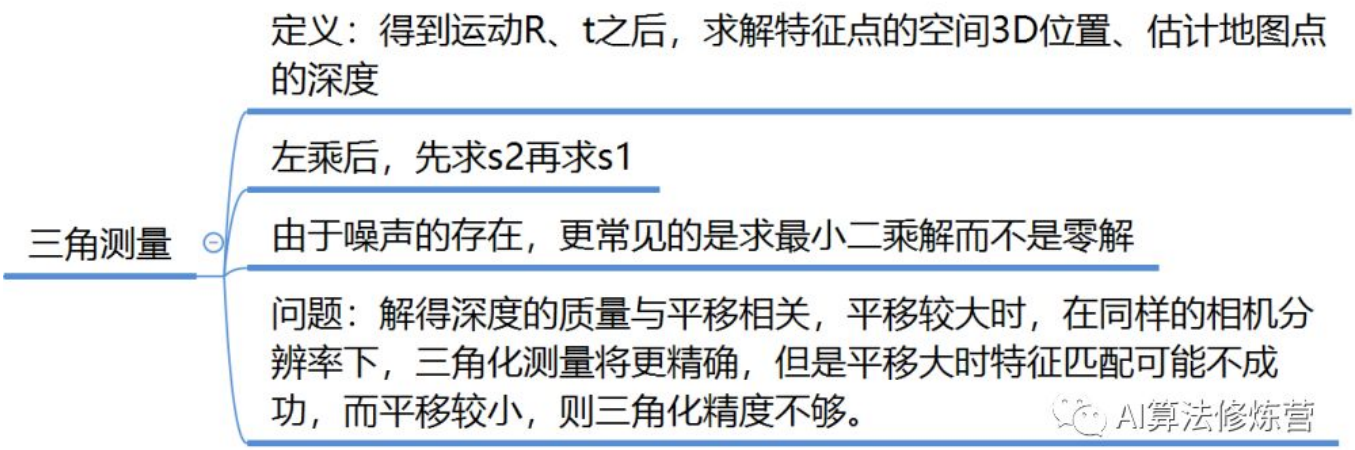
2D-2D：

2D-2D 主要是针对**单目相机的初始化过程**，在不知道空间中 3D 点的情况下（如未进行初始化）通过两帧间匹配的特征点进行帧间相机运动估计。这涉及到**对极几何中本质矩阵（E）或单应性矩阵（H）的相关理论及其分解，**通常在图像的特征匹配中难免会有“外点”，可以**采用随机采样一致（RANSAC）得到最大“内点”子集的 E 或 H。**



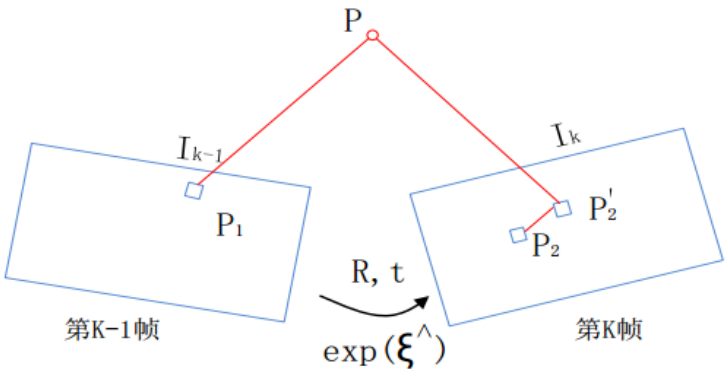
2D-2D示意图





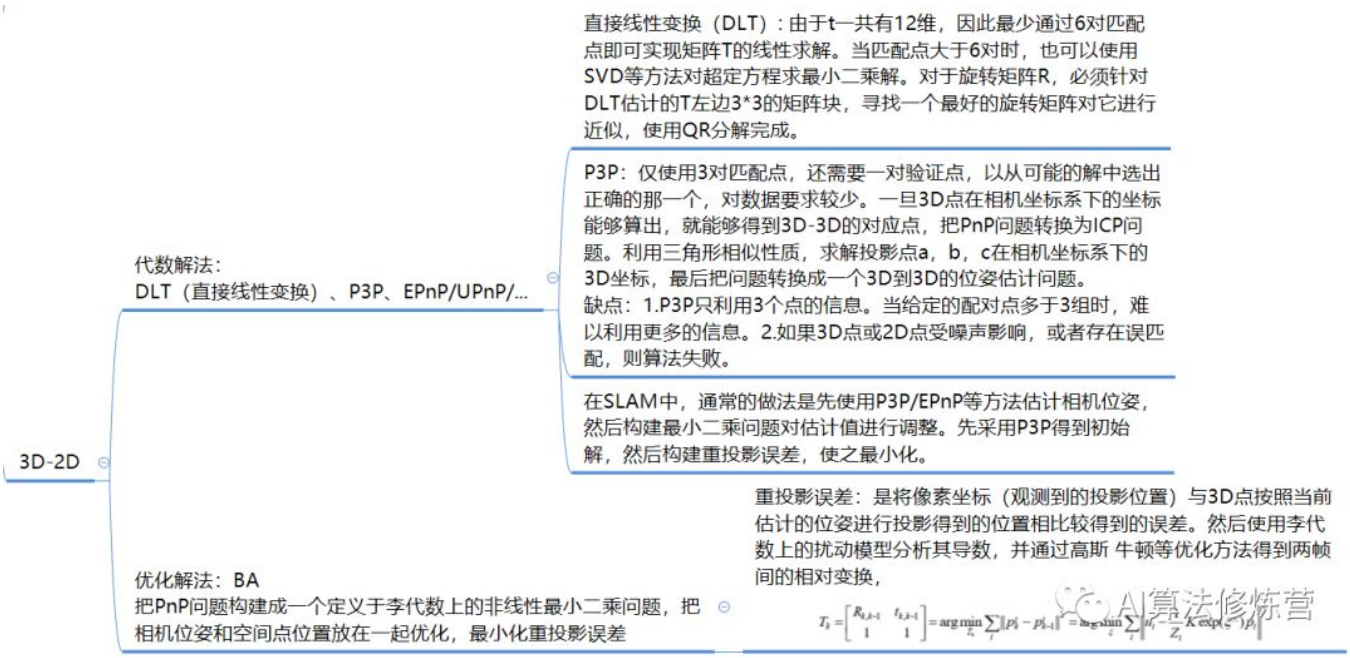
3D-2D：

3D-2D 就是 PnP（perspective-n-point）求解 3D 到 2D 点对运动的方法，描述的是当**知道 N 个 3D 空间点及其投影位置时**（例如单目，已经初始化完毕，知道特征点的 3D 位置），如何估计相机位姿。当然双目或者深度相机可以直接使用PnP。对它的求解有 DLT、 P3P、 EPnP、UPnP。**现在常用的做法是先采用 P3P 得到初始解，然后构建重投影误差，使之最小化。**



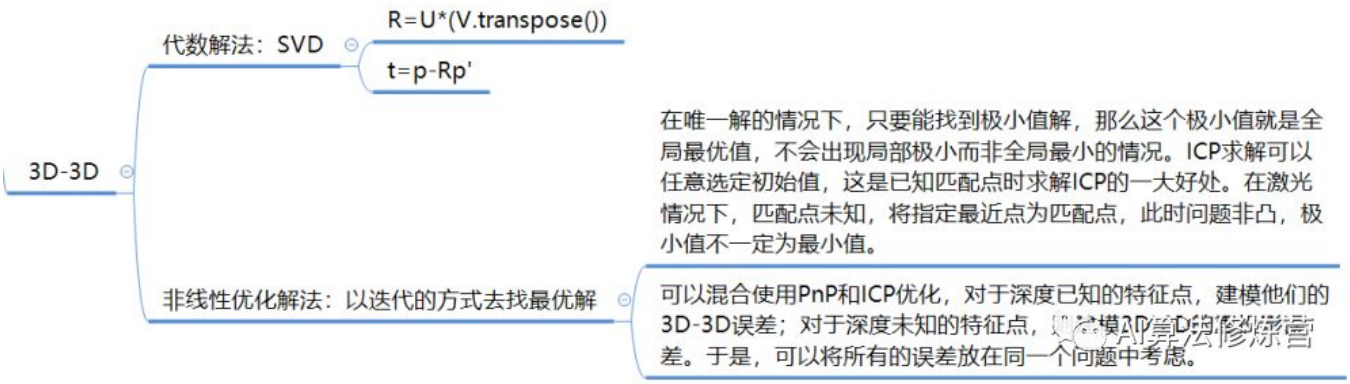
$$T_k = \begin{bmatrix} R_{k,k-1} & t_{k,k-1} \\ 1 & 1 \end{bmatrix} = \arg\min_{T_k} \sum_i \left\| p_k^i - p_{k-1}^i \right\|^2 = \arg\min_{\xi} \sum_i \left\| u_i - \frac{1}{Z_i} K \exp(\xi^\wedge) p_i \right\|^2$$

使用李代数上的扰动模型分析其导数，并通过高斯牛顿等优化方法得到两帧间的相对变换，具体做法又叫作捆集优化（bundle adjustment， BA），在编程上一般采用General Graph Optimization（G2O）等优化库实现。



3D-3D：

3D-3D 主要是激光 SLAM 采用**迭代最近点（ICP）求解。**在 VSLAM 中，可以在 RGB-DSLAM 中使用，但由于 RGB-D相机的限制，**仅仅适用室内，而且适用小的场景。**这是由于深度的估计不准，导致误差比 3D-2D 大。直观的感觉是， 相机得到的 3D 位置误差较大（相机方向性好， 距离信息误差大）， 3D-2D 只使用一次深度信息，但是 3D-3D 采用两次深度信息，导致计算的精确度降低，所以**在普通相机中一般回避 3D-3D 的方式。**



特征点法的问题

特征点法有几个问题：

a) **关键点的提取和描述子的计算非常耗时**，如果保证SLAM 实时运行，需要 30 Frame/s，也就是每帧图像的处理时间约 30 ms，而实时性最好的 ORB也需要近 20 ms/Frame；

b) 特征点法仅仅使用了图像中几百个特征点，占整个图像几十万个像素的很小部分， **丢弃了大量可以利用的图像信息**；

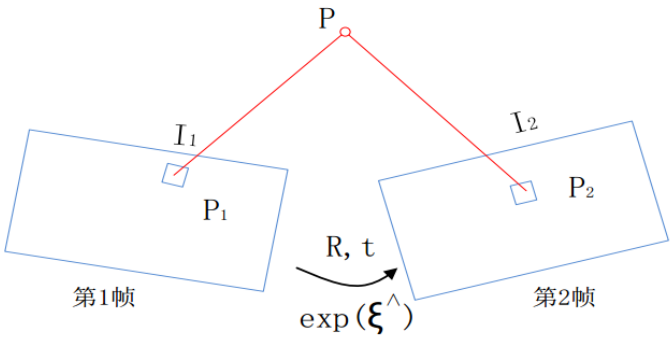
c) 特征点的寻找是根据人类自己设计的检测算法，并不完善，**有些图像没有明显的纹理，有些图像的纹理比较相似**，这种情况下特征点法的 VSLAM 就很难运行；

d) 特征点法**只能得到空间的稀疏三维点云。**离稠密地图尚有一定的距离，与用于机器人导航的地图差距就更大了。

1.1.2 直接法

直接法**根据像素灰度信息估计相机的运动**，几乎不用计算关键点和描述子，省去了计算关键点和描述子的时间，**可以在特征点缺失但是有图像灰度梯度的场合（当然对于一张白墙， 它也无能为力）。**相比于特征点法只能构建稀疏点云地图（构建半稠密或稠密需要采取其他技巧），直接法具备**构建半稠密和稠密地图的能力。**

直接法的不变量是对应像素点的灰度值。首先假设两个像素点在第一帧与第二帧之间灰度值保持不变，如下图所示， P1 和 P2 的灰度值是一样的，直接法的思路是**根据当前相机的位姿估计来寻找 P2 的位置**，如果相机位姿不好， P2 和 P1 的外观会有明显差别。



为了减少这个差别， 直接法优化相机位姿，**寻找与 P1更相似的 P2。**这就是**在灰度不变假设下，直接采用两帧图像中的匹配像素的灰度值，构建光度误差的优化函数，改变相机位姿使之最小化。**根据图像像素 P 的情况，直接法分为稀疏、半稠密和稠密直接法。P 如果是稀疏关键点， 称之为稀疏直接法；P 如果是图像中梯度明显的点， 称之为半稠密法；P 如果是图像中的所有像素， 称之为稠密法。

光流法

- 光流法**保留特征点**，但**只计算关键点，不计算描述子。**同时，使用光流法来跟踪特征点的运动。

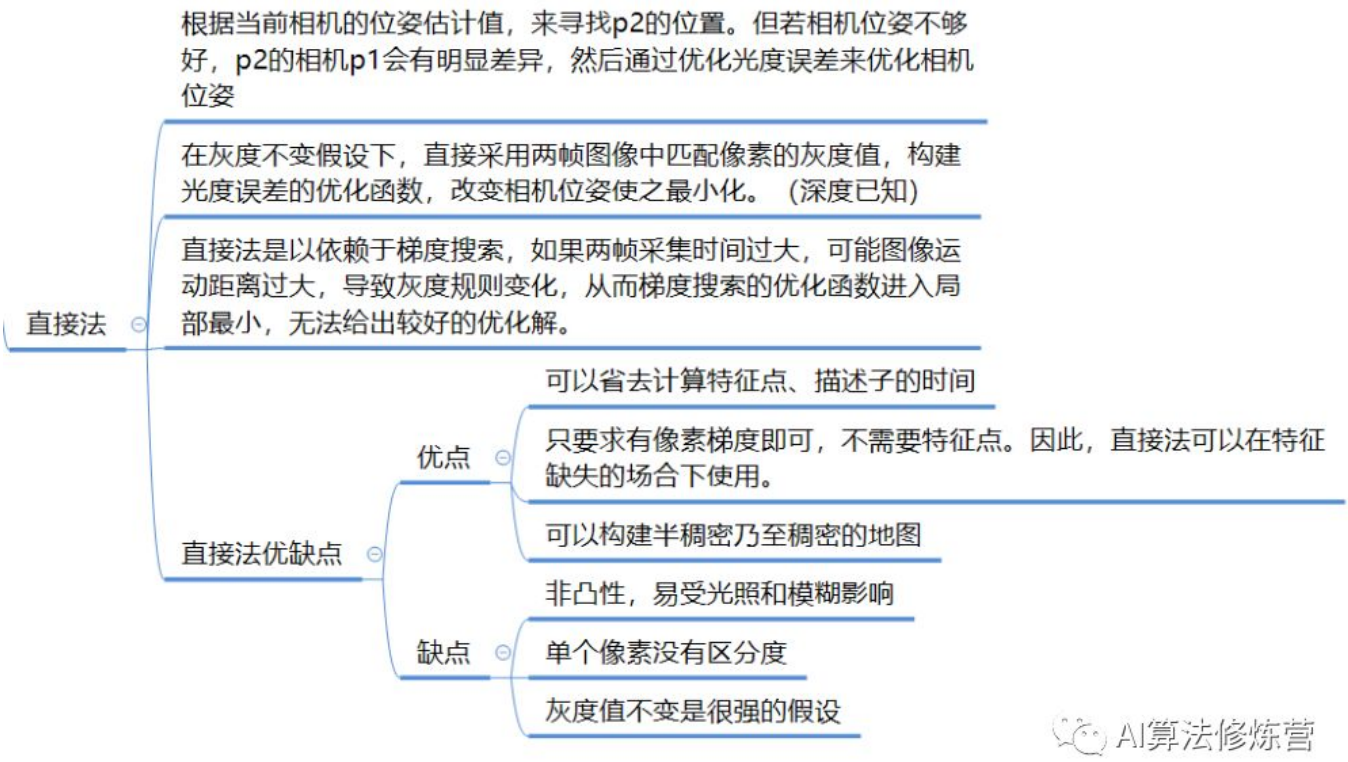
- 光流描述了**像素在图像中的运动**，跟踪源图像某个点在其他图像中的运动。遵循**灰度不变假设**：同一个空间点的像素灰度值，在各个图像中是固定不变的（强假设，不一定成立）。

- 计算部分像素运动的称为**稀疏光流（LK）**，计算所有像素的称为**稠密光流（HS）。**LK光流中，假设某一个窗口内的像素具有相同的运动。

- 可以看成**最小化像素误差的非线性优化**；每次使用泰勒一阶近似。在离优化点较远时效果不佳，往往需要迭代多次，**运动较大时需要使用金字塔**；LK光流可以**用于跟踪图像中稀疏关键点的运动轨迹**，得到匹配点后，后续计算与特征点法VO相同；按方法可以分为正向/反向+平移/组合的方式。

- 光流法可以**加速**基于特征点的视觉里程计算法，避免计算和匹配描述子的过程，但**要求相机运动较慢（或采集频率较高）。**

直接法



直接法也有自己的局限，首先它**需要满足光度不变性假设**，这对相机提出了很高的要求， 而且稠密法因为需要计算图像的所有像素（640*480 就是 30 万个像素），**很难在现有 CPU 上实时运行。**

在前端，特征点法和直接法最大的区别在于，**直接法是依赖于梯度搜索，如果两帧采集时间过大，可能图像运动距离过大，导致灰度不规则变化，**从而梯度搜索的优化函数进入局部最小，无法给出较好的优化解；而**特征点法对运动和光照有一定的鲁棒性**， 是根据特征点对距离和光照的鲁棒性来决定的，这也是未来 SLAM 发展的决定因素之一。

1.2 回环检测

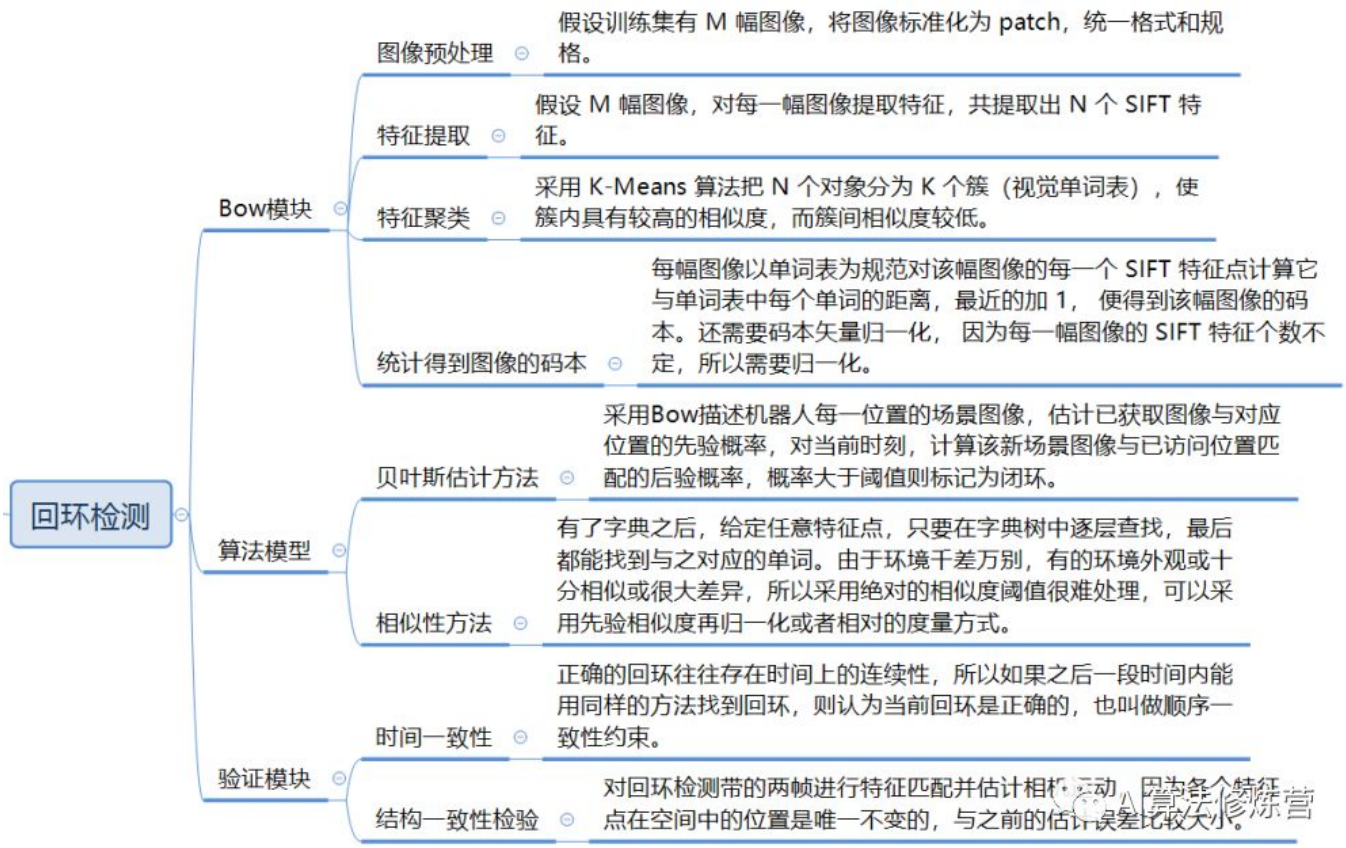
回环检测就是**利用传感器有效地检测出以前经过这里**，它对于 SLAM 系统意义非常重要，因为无论你的数据多么的精确，模型多么的优秀，**系统的累积误差始终存在。**如果能正确地检测到回环，对构建全局一致性地图是非常有帮助的；从另一方面，可以利用回环检测对跟踪失败后的情况进行**重定位。**

在 VLSAM 中回环检测大多数做法是**基于外观，比较图像间的相似性。**如果用特征点的方式，比如采用 SIFT 特征描述一幅图像，首先每个 SIFT 矢量都是 128 维的，假设每幅图像通常都包含 1000 个 SIFT 特征，在进行图像相似度计算时，这个计算量非常大，所以**通常不会直接采用特征点，而是采用词袋模型。**

词袋模型通过提取图像特征，再**将特征进行分类构建视觉字典**，然后采用视觉字典中的**单词集合可以表征任一幅图像。**换句话说，通过 BoW 可以**把一张图片表示成一个向量。**这对判断图像间的关联很有帮助，所以目前比较流行的回环解决方案都是采用的 BoW 及其基础上衍生的算

法 IAB-MAP。

回环检测主要由 **BoW 模块、 算法模块、 验证模块**三个部分组成。



目前还没有专门针对直接法的回环检测方法，主流的回环检测都是利用特征点采取 BOW 方式。换句话说**回环检测还是依赖于特征点，从这个角度来看，特征点法有很大的优势：特征点法已经提取了特征，直接用这些特征去做回环检测**；而直接法没有提取特征，如果想做回环检测，必须要另外提取特征。这也是 ORBSLAM 和 LSDSLAM 中的回环检测采取的不同的方式。

ORBSLAM 中的回环检测与整个系统结合得比较紧密，整个系统都是采用的 **ORB 特征**，首先**离线训练**得到 ORB 词典，在搜索时因为ORBSLAM 本身就已经计算了特征点和描述，可以**直接用特征来搜索**，而且 ORBSLAM 采用**正向和反向两种辅助指标**：反向指标在节点（单词）上储存到达这个节点的图像特征的**权重信息和图像编号**，因此可用于快速寻找相似图像。正向指标则储存**每幅图像上的特征以及其对应的节点在词典树上的某一层父节点的位置**，因此可用于快速特征点匹配（只需要匹配该父节点下面的单词）。

LSDSLAM 是采用 OpenFABMAP（OpenCV 上实现的FAB-MAP）来完成回环功能。FAB-MAP 在贝叶斯框架下，采用 **Chou-Liu tree估计单词的概率分布**，能够完成**大规模环境**下的闭环检测问题，但是它通过连续的当前帧数据与历史帧数据比较，效率较低，**不能满足实时地回环检测。**个人感觉 LSDSLAM 中的回环检测是为了完成这个大的系统，额外添加的模块，其实与系统契合度不是很高。

结语

该篇作为SLAM系列的第一篇，主要介绍了视觉SLAM系统主要结构，并详细介绍了前端中的视觉里程计和回环检测两大模板。接下来，SLAM系列文章会持续更新，会继续和大家介绍**后端优化算法，视觉SLAM的主流框架，VIO多传感器融合以及激光SLAM**等内容，欢迎大家持续关注~

参考：

1.视觉十四讲 高翔

2.https://www.cnblogs.com/gaoxiang12/p/3695962.html

3.https://zhuanlan.zhihu.com/p/64720052

-END-