视觉定位方法

本项目所使用到的视觉定位技术主要分为两部分：

1. 点云重建。在这一步骤中，利用已有的图像进行点云重建，可以离线进行；
2. 视觉定位。给定（1）中得到的点云模型以及需要进行定位的图像，恢复图像相对于已有模型的位姿，包括旋转矩阵和平移向量。

接下来将分别对以上两个步骤进行详细介绍。

1. 点云重建

现有的点云重建方法有很多，Bundler，VisualSfM，COLMAP。考虑到COLMAP效果较好，速度也较快，此项目拟采用COLMAP软件来进行点云重建。

由于COLMAP本身重建是基于SIFT描述子，而现在的一些基于学习的描述子（例如HardNet、SOSNet）在一些标准数据集（例如Brown数据集、HPtches数据集）展现出了明显优于SIFT描述子的性能。本项目拟替换掉COLMAP软件中的特征匹配模块，改为采用SOSNet[2]描述子。

进行点云重建的具体内容如下：

1. 关键点提取：对所有图像采用高斯差分（Difference of Gaussian, DoG）检测器进行关键点提取，在关键点处提取图像块，将图像块输入SOSNet中，得到相应的描述子。对于DoG检测器，可以调用OpenCV库的接口；对于SOSNet，作者公开了代码以及训练好的模型，可以直接调用。
2. 描述子匹配：采用相互最近邻（Mutual Nearest Neighbor, MNN）和ratio test方法对描述子进行匹配；
3. 丛运动恢复结构：得到的匹配对将作为COLMAP软件重建模块的输入，得到最终的点云模型。具体而言，每一个点的属性包含(x,y,z)坐标以及对应的128维描述子。
4. 相机定位

相机定位的目的是恢复给定图像相对于(1)中所获点云模型的位姿，主要包含二维-三维点匹配和PnP算法求解位姿两个步骤：

* 1. 二维-三维匹配

现有文献中的方法主要包含二维-三维(2D-3D)匹配以及三维-二维(3D-2D)匹配。其中2D-3D法相较于3D-2D法性能更好，但速度更慢。为了在保证匹配精度的同时提升匹配的速度，本项目主要参考了文献[1]中的快速匹配方法。

具体而言，本算法首先构建一颗词汇树(vocabulary tree)，将所有三维点分配到其对应的视觉词汇上。这一步骤只需要离线进行一次。接下来，对于每一图像，匹配的过程如下：

1. 利用DoG +SOSNet对图像进行特征提取；
2. 找到图像描述子对应的视觉词汇，以及该视觉词汇包含的三维点（定义为**潜在匹配点**）。三点的数目可以衡量匹配这一描述子所需要的开销，我们需要以尽可能小的开销来得到一定数量的描述子以供后续的PnP算法使用；
3. 得到每个描述子的潜在匹配点之后，根据潜在匹配点的数量对描述子进行排序，并挑选出K个潜在匹配点数量最少的描述子；
4. 采用MNN+ratio test方法进行特征匹配。
   1. 位姿求解

得到前一步骤中获得的二维-三维点对后，即可利用PnP算法来进行相机位姿的求解。由于匹配点对中仍然可能存在一些外点，可以采用一些基于RANSAC的方法进行剔除，例如MAGSAC[3]，MAGSAC++[4]等等。其中MAGSAC算法的改进版MAGSAC++方法在OpenCV库中有实现，可以直接调用。

[1] Fast Image-Based Localization using Direct 2D-to-3D Matching, ICCV2011

[2] SOSNet: Second Order Similarity Regularization for Local Descriptor Learning

[3] MAGSAC: Marginalizing Sample Consensus

[4] MAGSAC++, a fast, reliable and accurate robust estimator