硕士学位论文

基于学习方法的高精度SLAM算法研究

RESEARCH ON LEARNING BASED HIGH PRECISION SLAM ALGORITHM

冯爱迪

哈尔滨工业大学

2018 年 6 月

国内图书分类号：TP391.4　　学校代码：10213
国际图书分类号：621.3　　密级：公开

工学硕士学位论文

基于学习方法的高精度SLAM算法研究

硕士研究生：冯爱迪

导　　师：张大鹏教授

申请学位：工学硕士

学　　科：计算机科学与技术

所在单位：计算机科学与技术学院

答辩日期：2018 年 6 月

授予学位单位：哈尔滨工业大学

Classified Index: TP391.4
U.D.C: 621.3

Dissertation for the Master's Degree in Engineering

RESEARCH ON LEARNING BASED HIGH PRECISION SLAM ALGORITHM

Candidate: Aidi Feng
Supervisor: Prof. David Zhang
Academic Degree Applied for: Master of Engineering
Specialty: Computer Science and Technology
Affiliation: School of Computer Science and Technology
Date of Defence: June, 2018
Degree-Conferring-Institution: Harbin Institute of Technology

摘 要

同时定位和地图构建（Simultaneous Localization and Mapping，SLAM）是一种三维重建的方法。它是一种在未知环境中自主定位并进行地图构建的方法。SLAM在自动驾驶、虚拟现实等多个领域都有广泛的应用。

在众多SLAM算法当中，LSD-SLAM算法是一种可以实时地对大规模场景进行重建的方法。但是LSD-SLAM算法还存在提升的空间。近些年，基于学习的方法活跃在各个领域中，并取得了不错的成绩。因此，本文尝试利用基于学习的方法提升LSD-SLAM算法的重建鲁棒性。

SLAM系统由传感器、前端视觉里程计、后端非线性优化、闭环检测和地图构建五部分组成。其中，视觉里程计部分和闭环检测部分可以加入基于学习的方法，本文从这两个方面对LSD-SLAM算法做了改进。首先本文提出了一个基于可信控制点的深度置信度估计算法，对LSD-SLAM的视觉里程计部分做了提升和改进。置信度估计算法使用随机森林算法训练可信控制点预测模型，模型使用在立体匹配过程中可以方便快捷的计算出的特征，使模型在保证准确度的情况下，时间的消耗尽量小。在跟踪估计的过程中，通过模型得到一个深度置信度的估计，将置信度作为权值融入到深度估计和相机运动估计当中，得到更为准确的估计结果，提升LSD-SLAM算法在前端视觉里程计部分的精度，进而提升整个系统的重建精度。

本文还提出了一种基于二阶特征的闭环检测网络模型。模型采用深度卷积网络，基于二阶的特征信息，得到高精度的闭环检测结果。模型采用的损失函数为三元组损失函数。通过这种弱监督学习，不断缩小同一个地点的特征之间的距离，不断加大不同地点的特征距离，使得相同地点的特征聚为一类。将提出的闭环检测网络模型放入LSD-SLAM当中，提升模型在闭环检测环节的准确性，使重建的结果更加的准确。

利用上述两个模型，本文对LSD-SLAM算法的重建精度和鲁棒性做了提升，使算法可以得到更好的重建效果。

关键词：三维重建；可信控制点；立体匹配；闭环检测；二阶特征

Abstract

Simultaneous Localization and Mapping (SLAM) is a 3D reconstruction method. It is a method of self-localization and mapping in an unknown environment. SLAM has a wide range of applications in areas such as automatic driving and virtual reality.

Among various SLAM algorithms, the LSD-SLAM algorithm is a method that can reconstruct large-scale scenes in real time. However, there is room for improvement in the LSD-SLAM algorithm. In recent years, learning-based methods have been active in various fields and have achieved good results. Therefore, this project attempts to use learning-based methods to improve the robustness of LSD-SLAM.

The SLAM system consists of a sensor, a front-stage visual odometry, back-stage nonlinear optimization, loop closure, and mapping. Among them, the visual odometry part and the loop closure part can join the learning-based method. This subject has improved the LSD-SLAM algorithm from these two aspects. Firstly, this topic proposes a depth confidence estimation algorithm based on ground control points, which improves the visual odometry part of the LSD-SLAM. The confidence estimation algorithm uses a random forest algorithm to train the gorund control points prediction model. The model can use the features that can be calculated conveniently and quickly in the process of stereo matching, so that the model costs as little as possible while prediction. In the process of tracking estimation, an estimate of the depth confidence is obtained through the model, and the confidence is used as a weight into the depth estimation and the camera motion estimation to obtain a more accurate estimation result and improve the LSD-SLAM algorithm in the front-stage visual odometry part accuracy, and then improve the reconstruction accuracy of the entire system.

This topic also proposes a loop closure detection network model based on second-order features. The model uses a deep convolutional network, based on the second-order information, to obtain high-precision loop closure detection results. The loss function used by the model is the triplet loss function. Through this kind of weakly supervised learning, the distance between the features of the same place is continuously reduced, and the feature distances of different locations are constantly increased, so that the features of the same place keep clustering. The proposed loop closure detection network model is put

into the LSD-SLAM to improve the accuracy of the model in the loop closure detection make the reconstruction result more accurate.

Using the above two models, the subject has improved the reconstruction accuracy and robustness of the LSD-SLAM algorithm, so that the algorithm can obtain better reconstruction results.

Keywords: 3D Reconstruction, Ground Control Points, Stereo Matching, Loop Closure, Second-order

目 录

第 1 章 绪论

1.1 课题背景及研究意义

SLAM是Simultaneous Localization and Mapping的缩写，即同时定位与地图构建。它是三维重建多种应用中的一种。它主要是指传感器（如手机、机器人等设备）在自身位置不确定并且没有周围环境的先验信息的情况下，创建地图，同时自主定位。在虚拟现实、增强现实等领域SLAM都有着广泛的应用。近些年，基于学习的方法被应用到越来越多的研究方向中，并且都取得了不错的成绩。将基于学习的方法加入到SLAM中，也成为了一种趋势。

SLAM系统主要由以下五个部分组成：传感器采集数据，前端视觉里程计，后端非线性优化，闭环检测以及地图构建。SLAM系统的整体框架图如图1-1所示。

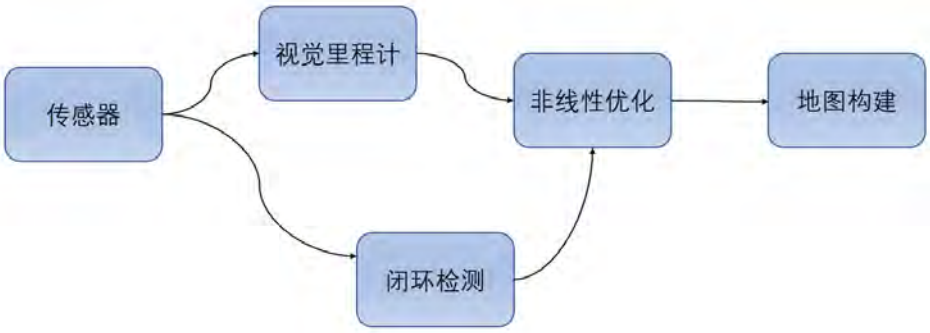


图 1-1 SLAM系统框架示意图

传感器负责采集数据，如RGB图像、深度信息等。它是整个系统的输入模块，负责处理输入数据，并将数据传递给前端视觉里程计。

视觉里程计负责估计相邻帧之间相机的运动。知道了相机的运动信息，就可以得到被重建景物的深度信息，从而进行三维重建。但是由于视觉里程计对于相机运动的估算会存在一定的误差并且它只考虑相邻两帧之间的运动，这样就不可避免的会将之前产生的误差累积到下一帧的计算当中。从而使整

个SLAM系统出现累积误差，进而导致重建结果不准确。由于计算误差的产生是不可避免的，从提高视觉里程计计算精度、改善相机运动估计的算法的角度上，是无法解决上述问题的。因此，研究者们希望找到一种技术，它能及时更正误差累积带来的漂移，从而尽可能地减少误差。

闭环检测就是这样的一种技术。它不断地检测相机是否到过当前地点，如果相机曾经来过相同的地方，它就将这个信息传送给SLAM的后端，后端非线性优化模块就可以将之前在相同地点得到的计算结果融入现在的计算中，从而减少误差累积带来的漂移，使得最后构建的地图更为准确。

后端的非线性优化是对一段时间内的运动状态进行估计。由于前端视觉里程计部分一定会产生误差，造成运动估计结果的漂移，需要及时更正这种误差。上文提到的闭环检测只能在再次访问相同地点的时候才能起作用，并不能及时最小化误差。后端通过获得的带有噪声的数据，利用卡尔曼滤波、集束调整（bundle adjustment，BA）[1]或者图优化[2]等方法，对当前的最新状态进行估计。并获得闭环检测模块得到的结果融入进去，得到更加鲁棒的相机运动状态估计。

地图构建模块是对传感器经过的地方进行地图构建的过程，它是SLAM系统的输出模块。通常是将估计出的结果以可视化形式展现出来，例如以点云的形式展现在三维坐标系中。

以上的五个环节中，视觉里程计部分和闭环检测部分可以使用近些年在多个领域都取得了不错效果的基于学习的方法，进一步提高估计的精度。本课题将在前端视觉里程计和闭环检测部分加入基于学习的方法，提高算法的精度。如果可以提高上述两个环节的精度，就能使SLAM系统的整体精度得以提高，从而使整个重建结果更加精确。

1.2 国内外研究现状

1.2.1 视觉里程计研究现状

传统的前端方法分为基于特征点法和直接法。特征点法顾名思义是利用特征点之间匹配进行进行运动估计的方法。特征点是在图片中具有代表性的点，这些点可以在视角发生一定范围内的偏移时，保持不变。特征点由关键点和描述子两部分组成。关键点利用图像的像素信息，找到一些有特点的点。描述子就是关键点周围的像素信息组成的特征。找到较小的视角变化下两张图片中有

特点易识别的关键点，再提取关键点周围的特征信息，就可以对两张图片中的特征点进行匹配，进而估计出相机的运动。

ORB-SLAM[3]（Oriented Brief Simultaneous Localization and Mapping）就是一种使用特征点法进行前端运动估计的SLAM算法。它的整体框架如图1-2所示。

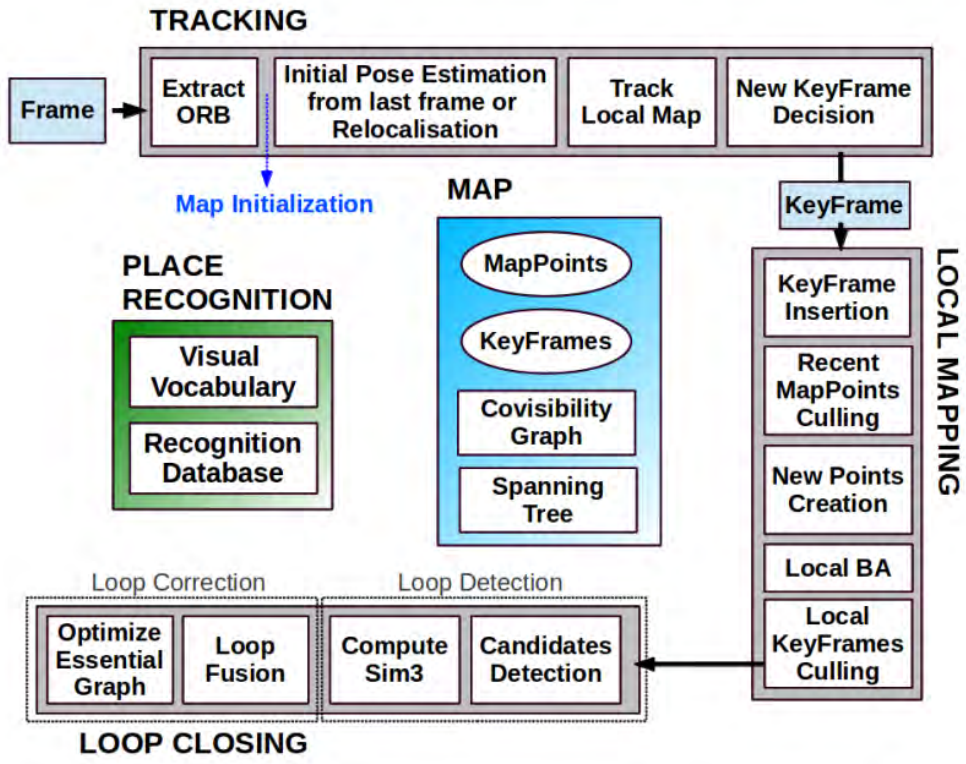


图 1-2 ORB-SLAM系统框架示意图[3]

它的视觉里程计部分使用的特征是ORB特征。ORB 特征[4]是一种简单的二值特征，它是对FAST特征点与BREIF 特征描述子的一种结合与改进，BRIEF是计算机视觉中比较常用的二值特征，通过比较一个经过高斯平滑的块内两个像素得到二值结果，每个块可以得到固定维数的向量，作为其特征描述。但BRIEF特征没有旋转不变性、尺度不变性，同时还对噪声敏感。ORB 利用一种灰度质心法来解决这个问题。

特征点法虽然在大多数SLAM的视觉里程计模块得到了应用，但是它也存在着一些缺点：首先，特征点的提取和计算十分耗时。上文提到的ORB特征在CPU上每帧需要20毫秒的时间进行计算，严重影响了SLAM系统的效率。其次，通过像素信息提取到的关键点不一定能涵盖图片的所有有用信息，只使用关键点，可能会损失有用的图像信息。第三，在纹理信息不强的时候，例如墙

面，能提取到的关键点很少，可能会造成两帧的匹配失败。

直接法是直接根据图像的亮度信息进行运动估计的视觉里程计方法。它不必要计算关键点和描述子，是一种使用所有像素估计运动的方法，避免了关键点带来的有用信息的遗失。在时间方面，它节省了计算特征点的时间。同时在一定程度上也可以克服特征缺失的问题。

直接法和特征点法相同，都是求解一个优化问题。不同的是，特征点法最小化的是重投影误差，而直接法需要优化测量误差，即最小化两个像素之间的亮度误差。

LSD-SLAM[5](Large-Scale Direct Monocular Simultaneous Localization and Mapping)使用的就是直接法作为前端估计运动的方法，它的整体框架如图1-3所示。LSD-SLAM只选择了周围梯度足够大的点作为运动估计的优化目标，因此重建出的结果是半稠密的。此外，它还使用了一个半稠密的深度滤波器[6] 的提出了显著地降低了计算复杂度，允许实时操作的处理器，甚至能在一个智能手机上运行。LSD-SLAM算法在进行运动估计的过程中，每隔一段距离选取一帧关键帧。没有被选做关键帧的图像作为当前关键帧的参考帧，被用于改善关键帧的深度估计。

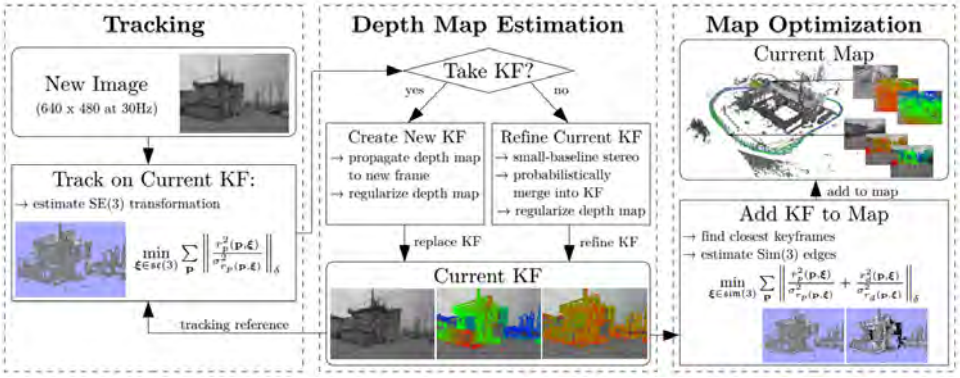


图 1-3 LSD-SLAM系统框架示意图[5]

首先，选取第一帧作为第一个关键帧。并将数据集中提供的第一帧对应的深度值作为初始深度信息。关键帧的结构为$K_i = (I_i, D_i, V_i)$，I_i表示原图像，$\Omega_i \to \mathbb{R}$，D_i表示逆深度图，$\Omega_{D_i} \to \mathbb{R}^+$，$V_i$表示深度倒数的方差，$\Omega_{D_i} \to \mathbb{R}^+$。深度和方差信息只对所有像素点的一个子集有定义，即周围强度的梯度足够大的点，因此是半稠密的。三维姿态用向量$\xi_{ji} \in \mathfrak{se}(3)$表示。相机姿态可通过最小化光学误差得到

$$E_p(\xi_{ji}) = \Sigma_{p\in\Omega_{D_i}} \| \frac{r_p^2(p,\xi_{ji})}{\sigma^2_{rp(p,\xi_{ji})}} \|_\delta \tag{1-1}$$

其中,

$$r_p(p,\xi_{ji}) := I_i(p) - I_j(\omega(p, D_i(p), \xi_{ji})) \tag{1-2}$$

$$\sigma^2_{rp(p,\xi_{ji})} := 2\sigma_I^2 + (\frac{\partial r_p(p,\xi_{ji})}{\partial D_i(p)})^2 V_i(p) \tag{1-3}$$

这里做了方差标准化，对于第i帧图像，ω函数将p点转换到第j帧相机坐标再变成图像坐标，r_p计算其差距，σ_I是高斯图像灰度噪声，r_p方差表示为r对逆深度的偏导平方乘以方差再加上σ_I。最后使用huber范数

$$\|r^2\|_\sigma = \begin{cases} \frac{r^2}{2\sigma} & |r| \le \delta \\ |r| - \frac{\delta}{2} & \text{otherwise} \end{cases} \tag{1-4}$$

没有被选为关键帧的图像作为当前关键帧的参考帧，将被用于改善关键帧的深度图。由于小基线的立体匹配的精度比较高，将其结果融合到现有的关键帧深度图中，从而使深度图得以改善。如果相机移动的距离现在的映射太远，就将最近的图片作为新的关键帧。定义关键帧之间的距离阈值为

$$dist(\xi_{ji} = \xi_{ji}^T W \xi_{ji}) \tag{1-5}$$

其中W是一个包含权值的对角阵。新的关键帧的深度图的计算方式是：首先从上一个关键帧向新关键帧投影得到初始化的深度图。然后用空间正则化迭代并去除异常值。

当一个新的图像被选为关键帧，它的深度图用前一个关键帧向它投影得到。之后，深度图利用平均逆深度值作为缩放因子进行缩放，并将这个缩放因子直接融入相机姿态。最后，新的关键帧将取代原来的关键帧用于后续的跟踪。

单目的SLAM算法相对于多目算法是内在规模矛盾的，即世界的决定规模是不可见的。因此，过长的跟踪轨迹将导致漂移，这是重建的一个主要误差来源。此外，所有距离都是基于缩放规模定义的，这将导致保证鲁棒性的核心（如huber范数）的定义不明确。

基于此，这篇文章的重点是提出了一个能观察到确定尺度的场景，所以姿态图直接在$\mathfrak{sim}(3)$上考虑图像之间的关联，也就是姿态的变换。考虑光学误差r_p和深度误差r_d，最小化最终误差为

$$E(\xi_{ji} = \Sigma_{p\in\Omega_{D_i}} \| \frac{r_p^2(p,\xi_{ji})}{\sigma^2_{rp(p,\xi_{ji})}} + \frac{r_d^2(p,\xi_{ji})}{\sigma^2_{rd(p,\xi_{ji})}} \|_\delta \tag{1-6}$$

其中，光学残差r_p^2和$\sigma_{r_p}^2$已在前边定义。深度残差和它的方差定义为

$$r_d(p,\xi_{ji}) = [p']_3 - D_j[p']_{1,2} \tag{1-7}$$

$$\sigma^2_{rd(p,\xi_{ji})} = V_j([p']_{1,2})(\frac{\partial r_d p,\xi_{ji}}{\partial D_j([p']_{1,2})})^2 + v_i(p)(\frac{\partial r_d(p,\xi_{ji})}{\partial D_j(p)})^2 \tag{1-8}$$

直接法需要灰度不变假设的支持。灰度不变是一个强假设，在现实当中很难成立。例如，当相机自动调整曝光时，图片的亮度会整体变亮或者整体变暗。在单面光照下的物体都会产生阴影，这样当相机从物体高光面移动到阴影面的时候，亮度也会产生明显的变化。因此，如果可以给直接法加入一些约束信息，提高直接法的鲁棒性，一定可以得到更准确的运动估计结果。

1.2.2 闭环检测研究现状

闭环检测问题可以定义为：输入一个查询图片和场景中的模板图片集，要在输入图片和模板图片集之间进行匹配，找到相似度较高的模板图片。然后可以通过设定阈值等信息，判断匹配到的模板图片和查询图片是否为同一地点。

现在还没有可以直接通过图像之间的信息，进行闭环检测的方法，因此现有的闭环检测方法都是基于图像特征的：首先从整张图片提取特征，然后对特征向量进行数据分析。可以将特征分为两类：全局特征和局部特征[7, 8]。

全局特征是从图片提取特征，它可以编码图像像素，形状特征和颜色等信息。例如，GIST特征[9]，提取不同方向和尺度的滤波器的响应。LDB特征[10]提取图片的强度和梯度信息，并将信息编码成二进制串。

局部特征是从图片的特殊的点出发，提取这个点邻域的信息。例如，ORB特征[4]是现在图片金字塔上提取FAST角点，提取到的角点就是图像中特殊的点，然后提取这些角点的BRIEF特征。在闭环检测中，经常使用词袋（Bag-of-Word，BoW）[11]模型，将局部特征处理成特征向量。

在闭环检测问题中。匹配的方法可以分为三种：最近邻搜索方法，稀疏优化方法和基于深度学习的方法。

FAB-MAP[12, 13]是一种最近邻搜索的方法。FAP-MAP用Bow的方法匹配当前的场景和之前的场景。通过提取图像中的关键点，并通过其周围的信息提取描述子。将描述子放到用训练数据训练的词袋当中，图像k就可以获得一个二值

的观测向量Z。

$$Z_k = \{z_1, \cdots, z_{|v|}\} \tag{1-9}$$

其中，v是词袋的大小。

在词袋聚类空间中的位置是物体e_i所代表的z_i在位置L_k的概率。

$$\{p(e_1 = 1|L_k), \cdots, p(e_{|v|} = 1|L_k)\} \tag{1-10}$$

每个概率是用特征检测的可靠性$p(z_i|e_i)$和环境中特征出现概率$p(e_i)$的先验知识得到的：

$$p(e_i = 1|L_k) = \frac{p(z_i|e_i = 1)p(e_i)}{\sum_{s_e \in \{0,1\}} p(z_i|e_i = s_e)p(e_i = s_e)} \tag{1-11}$$

$\mathcal{Z}^k$是到Z_k之前的所有观测向量的集合，在知道$\mathcal{Z}^k$的前提下，得到L_i的概率可以用贝叶斯公式求得：

$$p(L_i|\mathcal{Z}^k) = \frac{p(Z_k|L_i, \mathcal{Z}^{k-1})p(L_i|\mathcal{Z}^{k-1})}{p(Z_k|\mathcal{Z}^{k-1})} \tag{1-12}$$

其中，先验概率$p(L_i|\mathcal{Z}^{k-1})$是用简单的运动模型估计的，即相邻的观测相似的概率高。

FAB-MAP利用Chow Liu[14]树对特征的分布进行评估。Chow Liu算法是最大化信息熵的生成树。利用Chow Liu树对特征进行与BoW模型类似的聚类操作，将特征映射到聚类后的空间。在这个空间中，邻近的两个图片相似度最高。

ROMS(RObust Multimodal Sequence-based)[15]闭环检测方法是一种基于稀疏优化求解的方法。他的算法示意图如图1-4所示。对于给定收集到的模板图片的特征矩阵$\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_n] \in \mathbb{R}^{m \times n}$，查询图片$\mathbf{b} \in \mathbb{R}^m$，闭环检测问题可以定义成如下凸优化问题：

$$\min_{\mathbf{a}} \ ||\mathbf{D}\mathbf{a} - \mathbf{b}||_2 + \lambda||\mathbf{a}||_1 \tag{1-13}$$

其中$\lambda > 0$是折中参数，$\mathbf{a} \in \mathbb{R}^n$是$\mathbf{b}$和$\mathbf{D}$中所有模板图片的相似度向量。如果$a_i$值大，说明模板图片$\mathbf{d}_i$和查询图片$\mathbf{b}$越像。公式（1-13）的前半部分是一个凸优化损失函数，度量模板和查询图片之间的距离。第二部分是正则化项，防止结果过拟合。通过引入ℓ_1范数作为正则化项，可以使得到的解更稀疏，即每张查询图片只能得到几张相似度高的结果图片。这样可以使得闭环检测问题的解更加准确。

将问题推广为求解查询图片序列$\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_s] \in \mathbb{R}^{m \times s}$，此时，模板图片与查询图片之间的相似度矩阵为$\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_s] \in \mathbb{R}^{n \times s}$。

$$\min_{\mathbf{A}} \ ||(\mathbf{D}\mathbf{A} - \mathbf{B})^\top||_{2,1} + \lambda||\mathbf{A}||_{2,1} \tag{1-14}$$

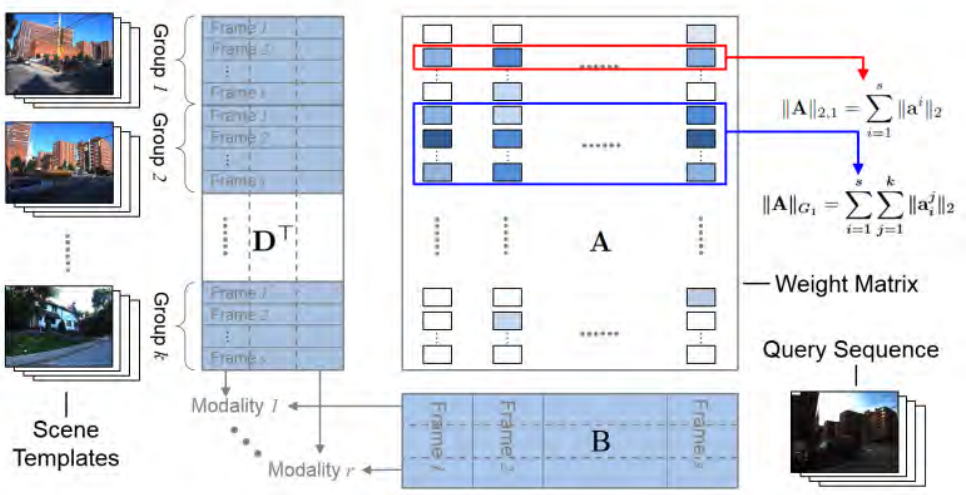


图 1-4 ROMS系统框架示意图[15]

考虑到在闭环检测问题中，相邻的帧比较相似，因此我们可以将模板图片分成组，每组中都是比较相似的图片。因此在求解的时候，希望得到的结果图片中，同组的图片尽量少。考虑加入组内的正则化项，得到组内稀疏的解。定义组范数为组内权重的ℓ_2范数的ℓ_1范数，即ℓ_2范数求和

$$\|\mathbf{A}\|_{G_1} = \sum_{i=1}^{s} \sum_{j=1}^{k} \|a_i^j\|_2 \tag{1-15}$$

其中，k是组的个数。将组范数加入到公式（1-14）中，得到问题最终需要优化的公式：

$$\min_{\mathbf{A}} \ \|(\mathbf{DA} - \mathbf{B})^{\top}\|_{2,1} + \lambda_1 \|\mathbf{A}\|_{2,1} + \lambda_2 \|\mathbf{A}\|_{G_1} \tag{1-16}$$

最终通过求解上述凸优化问题，求得闭环检测问题的解。

NetVLAD[16]是一种通过卷积神经网络提取图片的特征，通过直接度量特征间的距离求解闭环检测问题的方法，网络框架图，如图1-5所示。NetVLAD希望通过设计一个卷积神经网络得到一个从原始图片到特征空间的映射f_θ，θ是参数集合。通过这个映射，使得查询图片q和与其对应的模板图片I在特征空间中距离最短，即$d_\theta(q, I) = \|f_\theta(q) - f_\theta(I)\|$ 相比于q 与其他模板图片最小。类似的工作的CNN网络主要分为两部分：第一部分是提取局部特征，第二部分是将特征池化。NetVLAD遵循这个结构，作者将CNN的最后一个卷积层作为特征提取层，在之后加入作者提出的VLAD 池化层。

VLAD[17]是一种类似于BoW模型的特征聚类方法。它保存特征向量到它对应的聚类中心的残差和。假设有N个D维的局部特征向量$\{\mathbf{x}_i\}$和K个聚类中心$\{\mathbf{c}_k\}$，

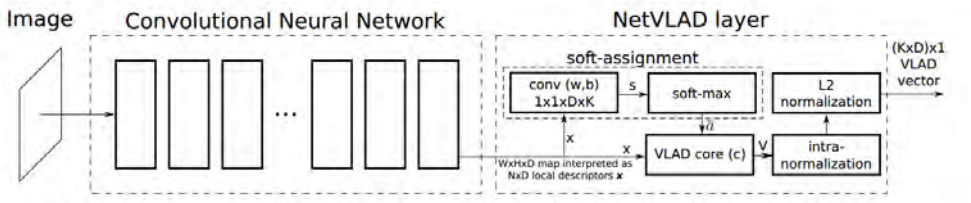


图 1-5 NetVLAD网络图[16]

VLAD保存的结果为：

$$\mathbf{V}(j,k)=\sum_{i=1}^{N}a_k(\mathbf{x}_i)(x_i(j)-c_k(j)) \tag{1-17}$$

其中，$\mathbf{V}\in\mathbb{R}^{K\times D}$，$x_i(j)$表示第$i$个特征$\mathbf{x}_i$的第$j$维，$c_k(j)$表示第$k$个聚类中心$\mathbf{c}_k$的第$j$维。当$\mathbf{x}_i$对应的聚类中心是第$k$个聚类中心时$a_k(\mathbf{x}_i)$为1，否则等于0。

但是，$a_k(\mathbf{x}_i)$不可微，无法反向传导。为了能使$a_k(\mathbf{x}_i)$可以进行反向传导，将$a_k(\mathbf{x}_i)$转化成$\bar{a}_k(\mathbf{x}_i)$求解：

$$\bar{a}_k(\mathbf{x}_i)=\frac{e^{-\alpha\|\mathbf{x}_i-\mathbf{c}_k\|^2}}{\sum_{k'}e^{-\alpha\|\mathbf{x}_i-\mathbf{c}_{k'}\|^2}} \tag{1-18}$$

将上式中的平方项展开可以发现$e^{-\alpha\|\mathbf{x}_i\|^2}$项可以约掉，于是公式（1-18）可以写成如下形式：

$$\bar{a}_k(\mathbf{x}_i)=\frac{e^{\mathbf{w}_k^{\top}\mathbf{x}_i+b_k}}{\sum_{k'}e^{\mathbf{w}_k^{\top}\mathbf{x}_i+b_{k'}}} \tag{1-19}$$

其中$\mathbf{w}_k=2\alpha\mathbf{c}_k$，$b_k=-\alpha\|\mathbf{c}_k\|^2$。公式（1-19）的形式可以看为一个卷积层加上一个softmax层。在这之后，作者又做了归一化处理，最终输出的是$K\times D\times 1$的向量。

在损失函数上，NetVLAD选择三元组损失函数。将训练集转换成三元组的形式$(q,\{p_i^q\},\{n_j^q\})$，其中q是查询图片，$\{p_i^q\}$是q可能匹配的正样本集合，$\{n_j^q\}$是与q不匹配的负样本集合。作者期望通过网络的训练，使得查询图片和正样本之间的距离尽量小，与负样本之间的距离尽量大。于是使用如下形式的损失函数：

$$L_\theta=\sum_h l(\min_i d_\theta^2(q,p_i^q)+m-d_\theta^2(q,n_j^2)) \tag{1-20}$$

其中，$l(x)=\max(x,0)$，m是偏移参数。NetVLAD就是通过上述弱监督损失函数训练数据，得到端对端的求解闭环检测问题的网络。

闭环检测问题还有很多待解决的难点。首先，在现实世界中，可能会存在两个非常相似的地点，但他们并不是同一个地点，这通常被称之为感知混淆；

其次，真实的世界不是一成不变的。对于同一个地点，可能会由于光照的变化、季节的变化、天气的变化、视角的变化，使同一个地点采集到的信息产生很大的变化。现实当中不可能所有地方都是理想光照，可能会产生传感器与光源的角度不同，或者光照强度变化比较明显的情况，如一天中不同的时刻太阳光的强度和角度都会发生变化。由于光照的变化，就可能产生阴影、曝光过度或者曝光不足等问题，这些都可能导致传感器在同一地点采集的数据有很大的差别。现实当中的景物不是一成不变的，在大部分地区，由于季节的更迭，景物都会发生很大的变化。如冬天景物可能会被大雪覆盖，使得景物的形状发生变化。还有植被在不同季节的变化，会带来明显的颜色上和形状上的变化，使闭环检测系统在判断上产生偏差。还有天气的变化，例如雨天的积水，可能通过积水的反射，使传感器收集到其他场景的复杂信息，从而影响判断。此外，由于视角的变化，物理上同一地点的两张图片，在视觉上可能会有较大的变化，这也会给闭环检测带来不小的麻烦。除了这些自然现象造成的影响，诸如行人、车辆等一些人为的因素也会造成闭环检测的不准确。

1.3 本文研究内容

直接法虽然有速度快、避免有用信息浪费等优点，但是直接法基于灰度不变性这一强假设。在实际中，灰度不变的假设是不成立的。这样就会造成直接法产生一些误差。如果可以引入一种机制，这种机制可以判断直接法的估计是否准确。这样就可以提升直接法的鲁棒性，进而提高整个SLAM系统的估计精度。

闭环检测会受到感知混淆、光照等一些因素的影响，造成检测的结果不准确。高精度的闭环检测系统，对改善SLAM系统的重建结果有很大的帮助。本课题希望提出一种深度学习的方法，得到高精度的闭环检测的解。

本课题将从以上两个方面提高SLAM系统的准确度。具体的章节内容如下：

第2章，将提出一种基于可信控制点的置信度估计方法，该方法可以预测直接法估计出的深度信息是否准确，方法的框架图如图2-2所示。这样就可以删除深度估计不准确的点，只保留深度估计准确的点，提高直接法的鲁棒性。在方法中，本课题使用随机森林方法训练预测模型。本课题在TUM RGB-D数据集上进行了实验，验证了方法的有效性。同时，还与决策树、感知器等方法做了对比，以验证所选的随机森林方法的优越性。

第3章，将提出一个基于二阶统计信息的高精度闭环检测模型。阐述了二阶特征信息在闭环检测问题中的应用，网络图如图3-1所示。为了比较模型与现行的方法的精度，本课题对比了提出模型和LSD-SLAM中使用的FAB-MAP算法。为了验证模型与其他基于深度学习的闭环检测方法的优越性，实验对比了NetVLAD模型。

第4章，将第2、3章提出的方法集合到SLAM系统中。本课题基于LSD-SLAM，将第2章提出的基于可信控制点的置信度估计模型加入到LSD-SLAM的深度估计部分，框架图如图4-1所示，将第3章提出的模型替代原有的FAB-MAP算法，作为闭环检测的方法，框架图如图4-2所示。并设计实验对比了加入本课题提出的模型前后，SLAM算法的精度。最后，将本课题提出的两种模型都加入LSD-SLAM当中，框架图如图4-3所示，测试并对比了加入一种模型和加入多种模型对系统的影响。

第 2 章 基于可信控制点的置信度估计方法

2.1 引言

使用直接法估计深度和相机姿态，相较于特征法，在时间效率和信息利用的方法有较大的优势。但是由于灰度不变强假设的存在，直接法的估计结果并没有理想中的那么准确。基于此，本课题希望找到估计结果准确的点，并利用这些点进行进一步的估计和三维重建。本章主要提出了一种基于可信控制点的深度置信度预测算法。算法利用随机森林算法，训练了一个可以预测深度置信度的模型。通过设计可行性验证试验，本章证明了提出模型的可行性。通过准确性验证试验，验证了该模型可以较为准确的预测出深度置信度高的点。最后，本章还设计试验对比了几种不同的机器学习方法，证明选取的随机森林算法的合理性。

2.2 主要研究内容

2.2.1 立体匹配

立体匹配[18]是获得物体视差（深度）的常见方法。对视差或深度的估计是三维重建的关键步骤。

视差就是从有一定距离的两个点上观察同一个目标所产生的方向差异。从简单的情况讨论，假设两个相机的内部参数一致，如焦距、镜头等信息。引入坐标系，假设两个相机的X轴方向一致，坐标系以左相机为准，右相机相对左相机是简单的平移，用坐标$(T_X, 0, 0)$表示。

如图2-1所示，根据三角形的相似关系可以得出深度和视差的关系

$$\frac{d}{T_X} = \frac{f}{Z} \tag{2-1}$$

其中f为相机焦距，Z为物体深度。由此，可以得到深度信息和视差是成反比关系的。

立体匹配算法主要是通过在不同位置对物体进行拍摄一系列照片，根据图像的灰度等信息计算出匹配的代价，选择代价最小的视差，然后对得到的视差进行优化处理。

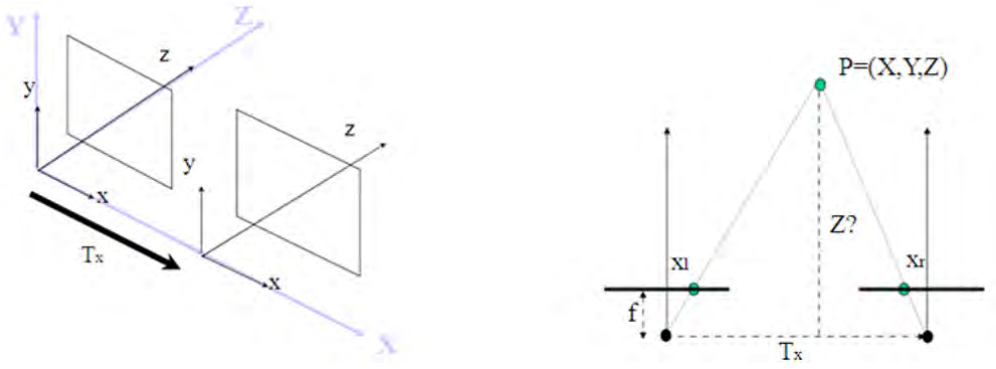


图 2-1 视差与深度示意图

一般使用局部匹配的方法计算匹配的代价。因为局部匹配方法值根据局部的信息计算匹配的代价，计算复杂度低，可以进行实时地匹配。本课题使用的就是基于局部立体匹配的方法。首先需要确定视差的取值范围及匹配的窗口大小。视差要确定合理的范围，避免不必要的计算。窗口不能选择过小，否则会导致窗口无法包含足够的灰度信息，窗口选择过大，则会导致视差的边缘不好，计算量也会增加。在指定的视差范围之内，枚举视差的大小。根据指定的视差，计算待匹配的两个窗口的匹配代价函数，得到匹配的代价。

确定视差的方法一般使用赢者通吃（Winner-Take-All，WTA）的方法，即选择匹配代价最小的视差值作为最终的视差。得到了视差之后，就可以根据视差与深度之间的关系计算出深度值，进而进行估计相机的运动，进行三维重建。

2.2.2 可信控制点在立体匹配中的应用

由于闭塞，图片缺乏纹理和图片中的重复结构等原因，使局部匹配算法造成混淆，导致立体匹配算法在这些情况下极易出错。

可信控制点，即视差估计可信的点。如果在匹配的过程中可以得知哪些点的估计是准确的，就可以增加这些点在匹配之中的权重，从而得到更鲁棒的匹配结果。本课题期望用一种学习的方法检测到立体匹配中的可信控制点。利用匹配点的置信度信息，提升视差估计的精度，进而使立体匹配达到更优的效果。基于可信控制点的高精度立体匹配[19]方法是一种利用可信控制点得到高精度的立体匹配的方法。除了匹配过程中计算得到的特征之外，该方法还使用了彩色图像的颜色、空间关系等信息，得到更为可靠的可信控制点。在立体匹配的过程中，该方法迭代交替更新可信控制点得到的置信度图和视差图，可以更

好的消除干扰信息，得到鲁棒的立体匹配结果。

本章实现的是一种利用匹配的特征和图片的特征进行监督学习的方法，利用方法训练出的模型可以有效地预测出可信控制点。图片像素之间的匹配特征，是通过在局部匹配过程中，计算局部的各种信息得到的。并没有使用边缘、角点、轮廓等几何特征信息，不会产生其他的时间消耗。

基于机器学习的方法在各种领域中都取得了较大的成功。本课题将利用机器学习方法训练可信控制点模型，通过可信控制点提高深度估计的准确度。

2.2.3 主要研究内容

现有的三维重建算法对深度（视差）的估计都不是非常准确，而深度值对于相机姿态预测有很大的影响，进而决定了三维重建的最终效果。LSD-SLAM是一种利用图像之间的匹配信息直接估计深度以及姿态的方法，虽然省去了提取特征的麻烦，可以直接利用整幅图像的信息，但是估计的准确度却很难保证。本章将基于LSD-SLAM的深度估计框架，利用图像特征，采用学习方法，获得深度估计的置信度图。

本章将通过研究和比较，从图像的多种特征中挑选一种或多种对逆深度图的准确性表征最多的[20]。利用随机森林算法预测出视差准确度高的点——可信控制点（Ground Control Points，GCPs）[21]。通过可信性验证试验，验证本章方法的可信性。然后通过定量的评估，验证方法的有效性。接着通过与其他机器学习的方法比较，阐述选择随机森林方法的理由。

2.3 置信度估计

2.3.1 随机森林算法

随机森林（Random Forest）[22]是一种比较新的机器学习模型，它是一种基于监督学习的机器学习模型。即在变量（列）的使用和数据（行）的使用上进行随机化，生成很多分类树，再汇总分类树的结果。随机森林在运算量没有显著提高的前提下提高了预测精度。随机森林对高维数据不敏感，可以很好地预测出多达上千个解释变量的作用，被誉为当前最好的算法之一。其他常用的机器学习算法还有决策树、SVM和感知器。其中决策树相当于只有一棵树的随机森林，在时间方面应该会优于随机森林，但是一棵树更容易产生过拟合的情况，在精度上应该不如随机森林算法。SVM是常用是机器学习算法之一，可以

解决线性问题和非线性问题。感知器是三层的神经网络，在训练过程中，通过反向传播自主学习参数。由于本章选取的训练特征数量很少且都是有实际意义的特征，因此并不适合深度网络。经过实验，多层神经网络在问题中并没有太大的作用。本文会在2.4.3 节比较这几种算法和随机森林在本问题中的性能。

随机森林算法的训练过程：首先，对训练集进行有放回的抽样N次，得到训练集的一个子集作为新训练集；接着，在新的训练集中无放回的随机抽出训练集的K个属性，训练一棵分类树（假设属性总量是X则要求$K << X$，一般取$K = \sqrt{X}$）。重复上述过程M次，得到M个分类器。即得到有M棵树的随机森林。

在使用随机森林算法进行分类时，使用M个分类器对数据分别进行分类，最后的分类结果由这M个分类器投票决定。对于分类问题，分类结果由投票中得票最高的类决定；对于回归问题，分类结果由投票结果取平均值得到。

随机森林算法的优点：第一，M个分类器之间无关，可并行训练数据；第二，对于某些出现概率小的噪点，其在抽样过程中抽到的概率较小，因此训练出的基分类器不受噪点影响。第三，可以处理高维度的数据，显而易见，每次抽取K（$K << X$）个属性进行训练，提高训练速度。同时，算法在训练的时候采用有放回的抽样策略，相对于单一的分类树不容易产生过拟合的问题。

2.3.2 特征选择

图像的特征提取是计算机视觉和图像处理中的一个概念。它指的是使用计算机提取图像信息，决定每个图像的点是否属于某个像素点集合。特征提取的结果会把图像上的点分为不同的子集。图像的特征是许多计算机图像分析算法的起点。因此，一个算法是否成功往往由它使用和定义的特征决定。所以特征提取最重要的一个特性是可重复性，即同一场景的不同图像所提取的特征应该是相同的。本章将利用从图像中提取到的特征判断像素深度的置信度。本章所使用的特征都可以在立体匹配的过程当中获得，并不会在特征提取上消耗过多的时间。

在描述特征之前，首先介绍一下记号。给定一对图像，代价函数记为$c(x_L, x_R, y)$，它计算的是左侧图像中(x_L, y)点和右侧图像中(x_R, y)点所有可能的匹配代价。视差定义为$d = x_L - x_R$，假设最小视差d_1是可以取到的。代价曲线是一个像素点能取到的所有匹配代价的集合。用c_1和c_2表示代价曲线上的最小代价和次小代价，c_2不一定是局部最小值。视差值d_1对应c_1。I为图像的RGB

值，I_L 为左侧图像的RGB 值。同理，I_R为右侧图像的RGB值。

在本课题中，使用了14个图像特征$f = f_1, \cdots, f_{14}$。其中，$f_1 \sim f_7$是从立体匹配的过程中获取的，具体如下：

边界距离（Distance from Border，DB）：像素到离它最近的边界的距离。距离边缘越近的点，越容易估计不准确。

代价（Cost）：匹配的代价，即NCC值。这里使用的是ZNCC方法。计算的是左侧图像上的坐标为(x, y)的点在视差为d的情况下的匹配代价。匹配的代价越低代表越有可能是正确的匹配。

$$ZNCC(x,y,d) = \frac{\Sigma\Sigma(\overline{I_L I_R}) - \overline{I_L I_R}}{\sqrt{\Sigma\Sigma(I_R - \overline{I_R})\Sigma\Sigma(I_L - \overline{I_L})}} \tag{2-2}$$

最大间隔（Maximum MargiN，MMN）：这个特征计算的是最大代价c_1与次小代价c_2之间的差值。选取的理由是较大的差值可能表明视差的计算更为准确。

$$C_{MMN} = c_2 - c_1 \tag{2-3}$$

胜者间隔（Winner MargiN，WMN）：用代价曲线的和标准化局部次小代价与最小代价的差。希望得到全局最小代价与选择的局部次小代价的差距大，并且总代价和也很大，以确保最小代价对应的视差的正确性。

$$C_{WMN} = \frac{c_2 - c_1}{\Sigma_d c(d)} \tag{2-4}$$

可得最大似然（Attainable Maximum Likelihood，AML）：这个特征基于代价曲线向一个可能的视差密度函数的转化。将所有匹配代价与最小代价代入高斯方程，其他代价值与最小代价相差越大，代表视差的估计可能越准确。

$$C_{AML} = \frac{1}{\Sigma_d e^{-\frac{(c(d)-c_1)^2}{2\sigma_{AML}^2}}} \tag{2-5}$$

最大似然（the Maximum Likelihood Measure，MLM）：用最小代价得到的指数值除以所有匹配代价带入高斯方程的值。分母越大，分子越小，代表越可能为正确的匹配点。

$$C_{MLM} = \frac{e^{-fracc_1 2\sigma_{MLM}^2}}{\Sigma_d e^{-\frac{c(d)}{2\sigma_{MLM}^2}}} \tag{2-6}$$

负熵（the Negative Entropy Measure，NEM）：是视差分布概率的负熵，使用先取负指数再进行标准化的方法获得概率。负熵是用来衡量置信度的。

$$p(d) = \frac{e^{-c_1}}{\Sigma_d e^{-c(d)}} \tag{2-7}$$

$$C_{NEM} = -\Sigma_d p(d)log(p(d)) \tag{2-8}$$

特征$f_8 \sim f_{14}$是通过深度图和RGB图的信息计算得到的。f_8和f_9分别是运动估计得到了几何误差和光学误差。f_{10}是当前参考帧和上一帧参考帧之间的估计值之差。f_{11}是深度的方差，f_{12} 和f_{13} 是像素的一阶导数和二阶导数。f_{14}是像素的深度。

2.3.3 基于可信控制点的置信度估计方法

本方法基于随机森林的可信控制点选择方法。采用监督学习的方法训练并预测置信度高的可信控制点。本方法是对直接法的改进，利用使用了直接法估计运动的LSD-SLAM算法获得数据。置信度的估计过程如图2-2所示。

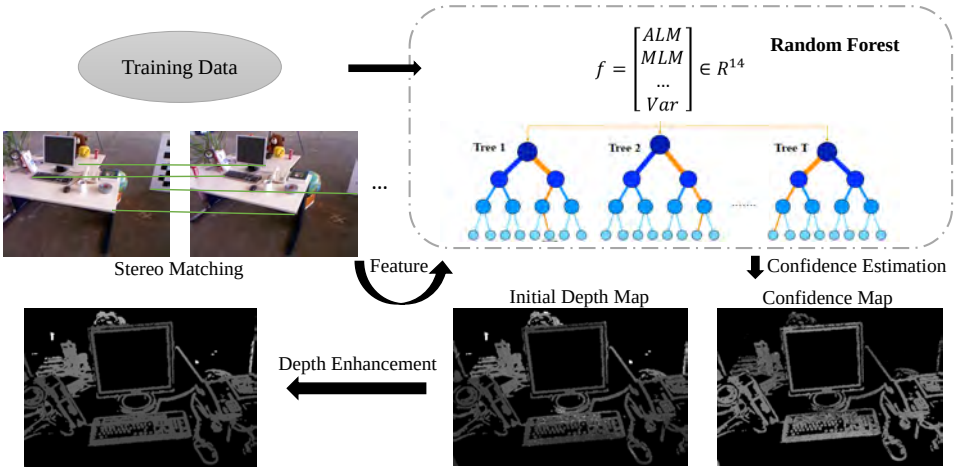


图 2-2 置信度估计过程示意图

首先，先从图像中提取到2.3.2 节中所述的特征信息，作为数据的属性。特征信息主要是像素点和邻近点关系的信息和匹配过程中计算的信息，即LSD-SLAM中关键帧与其对应的参考帧的匹配信息。其次，将提取到的深度信息与图像深度的真实值作比对，并将它们的差值归一化到[0, 1] 区间，将归一化后的值作为数据的标签。采用第2.3.1 节中提到的随机森林方法训练数据，得到一个预测可信控制点的模型。用训练好的模型对测试数据进行预测。将测试结果与之对应的数据标签作对比，并为它们之间的差值设定合适的阈值，预测结果在阈值范围内的即为可信控制点。

2.4 实验结果

2.4.1 可行性验证实验

实验使用的数据集为Middlebury Stereo Datasets。Middlebury[23]数据集包含27组图片，每组图片包含两张同一场景但存在一定视差的RGB彩色图片，和分别与它们相对应的视差图。

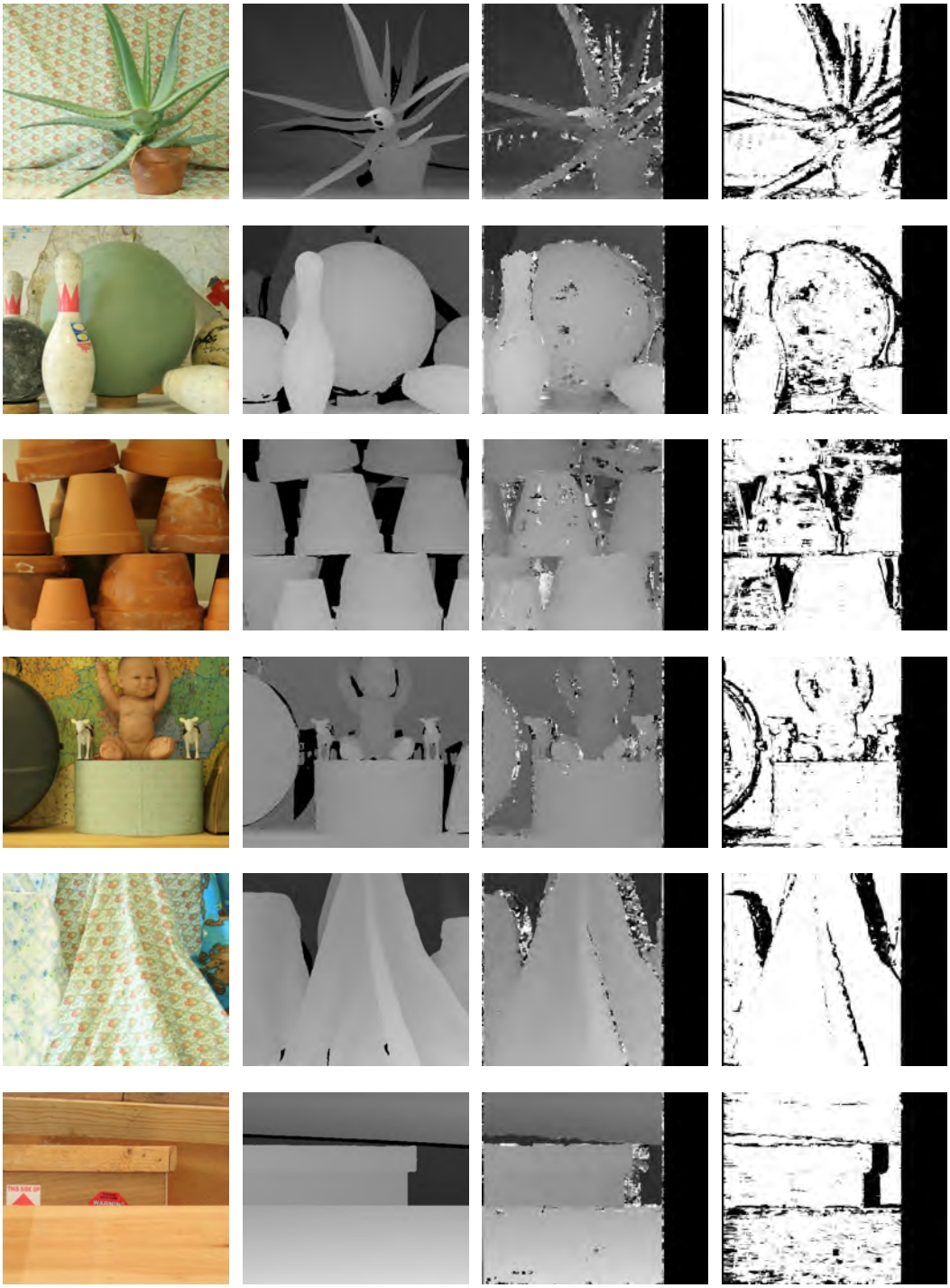
为了验证模型的可行性，本文首先做了一个二项分类实验。首先实现立体匹配算法，并计算出图像中的多种特征，利用随机森林算法训练并预测像素点匹配的准确性。首先，通过立体匹配算法计算出视差图。视差计算首先为视差设定一个合理的范围，通过枚举合法的视差值，计算在该视差值下的ZNCC值，选择得到ZNCC值最大的视差作为该点的视差。将计算得到的视差与数据集提供的视差图（真实图）进行对比，计算其差值作为视差的误差，并把视差误差小于等于1的点标记为0，误差大于1的点标记为1，并将其作为数据的标签。这里标签被标记为0的点为视差正确的点，被标记为1的点为视差计算错误的点。接着从图像中提取多种特征，并将其作为数据的属性。采用分类模式的随机森林算法进行训练和预测。

选用Middlebury数据集，随机选用9组图片作为训练集，另外选择的1组图片为测试集。将所有预测出的可靠点（即预测属于类别0的点）标记为白色，输出成预测结果图。对比真实图，计算出的视差和预测结果图，如图2-3所示。其中图2-3 a)列是其中一张输入的彩色图，图2-3 b)列是视差的真实图，图2-3 c)列是计算出的视差图，图2-3 d)列是视差计算准确与否的预测图。

对比真实图和计算出的视差图，可以观察到在物体的边缘分布很多视差不准确的点。对比每组的后三张图，可以发现视差计算错误的点，在预测结果图中也为负类点（黑色的点）；而视差计算正确的点，在预测结果图中也为正类点（白色的点）。因此，通过随机森林方法预测匹配点正确性的方法是可行的。

2.4.2 置信度评估

本实验使用2.3.2中提到的14种特征训练随机森林模型。使用的数据集为TUM RGB-D数据集[24]。TUM RGB-D数据集中的场景主要为桌子、电脑等室内场景。该数据集中包含了连续的场景图片，包括RGB彩色图片、Kinect测量出的深度图片以及相机参数信息。



a) b) c) d)

图 2-3 可行性分析结果图

首先，用LSD-SLAM方法计算出每个关键帧相对于其各个参考帧的特征值。将其中的深度值与Kinect的深度测量值（真实值）的差值归一化到在[0,1]区间内。其中归一化后的值越大，表示二者差距越小。接着，利用回归模式的随机森林算法进行训练和预测。

本实验的训练数据是从$fr2_desk$、$fr1_xyz$和$fr2_xyz$这三个序列中等比例随机选择了300,000个像素提取特征。由于选择的像素个数相较于每个序列整体的像素个数来说，数量非常少，并且LSD-SLAM算法是动态地选择关键帧，即每次运行算法，即使是同一个图像序列，选择的关键帧也会有所不同。因此，可以基本排除训练集和测试集重叠的情况。

为了分析每种特征在模型中的重要程度，首先进行了每种特种的重要度分析。每个特征的重要度如图2-4所示，图中横坐标的标号，对应特征向量f中特征的顺序。实验分别使用了$fr2_desk$, $fr1\_xyz$ 和$fr2_xyz$ 三个序列做训练集进行训练。从图中可以看出，每种特征的重要程度相差并不大。匹配的代价（f_2）、最大间隔（f_3）、可得最大似然（f_5）、最大似然（f_6）和深度（f_{14}）这几种特征相比于其他特征，在3个数据集上的重要程度都要更高。

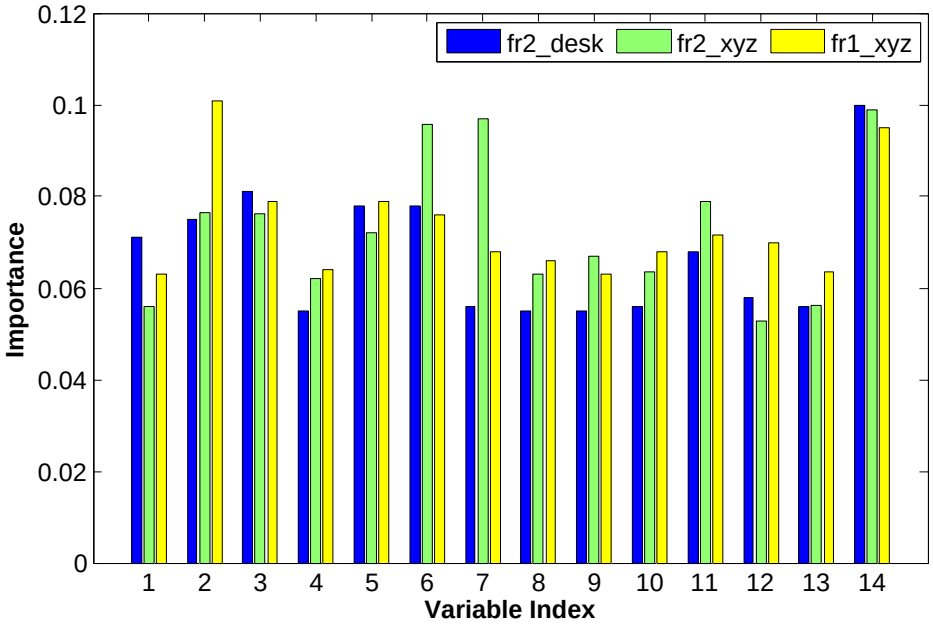


图 2-4 每种特征在模型中的重要程度

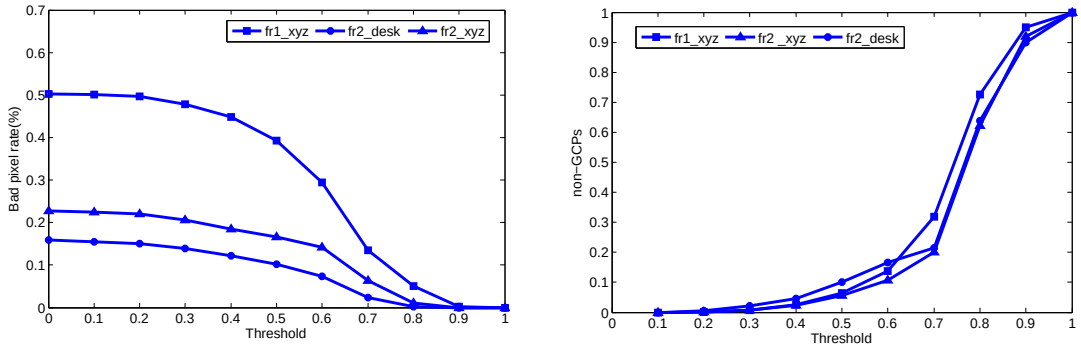
接着，本实验对随机森林中树的数量进行讨论分析。不同的数量的树的时间和准确度对比如表2-1所示。表中展示的是不同数量的树的随机森林模型，预

测每帧的平均时间和平均准确度的结果。可以看出，当树的数量超过30之后，准确度就没有很大的提升了，但是时间消耗却大幅的增加。折中时间效率和准确度，实验选取30棵树的随机森林进行后面的实验。

表 2-1 随机森林中树的棵树对比

树的数量	20	25	30	35	40
时间(ms)	132.0	180.5	241.4	291.8	376.9
结果(%)	81.45	81.53	81.56	81.57	81.58

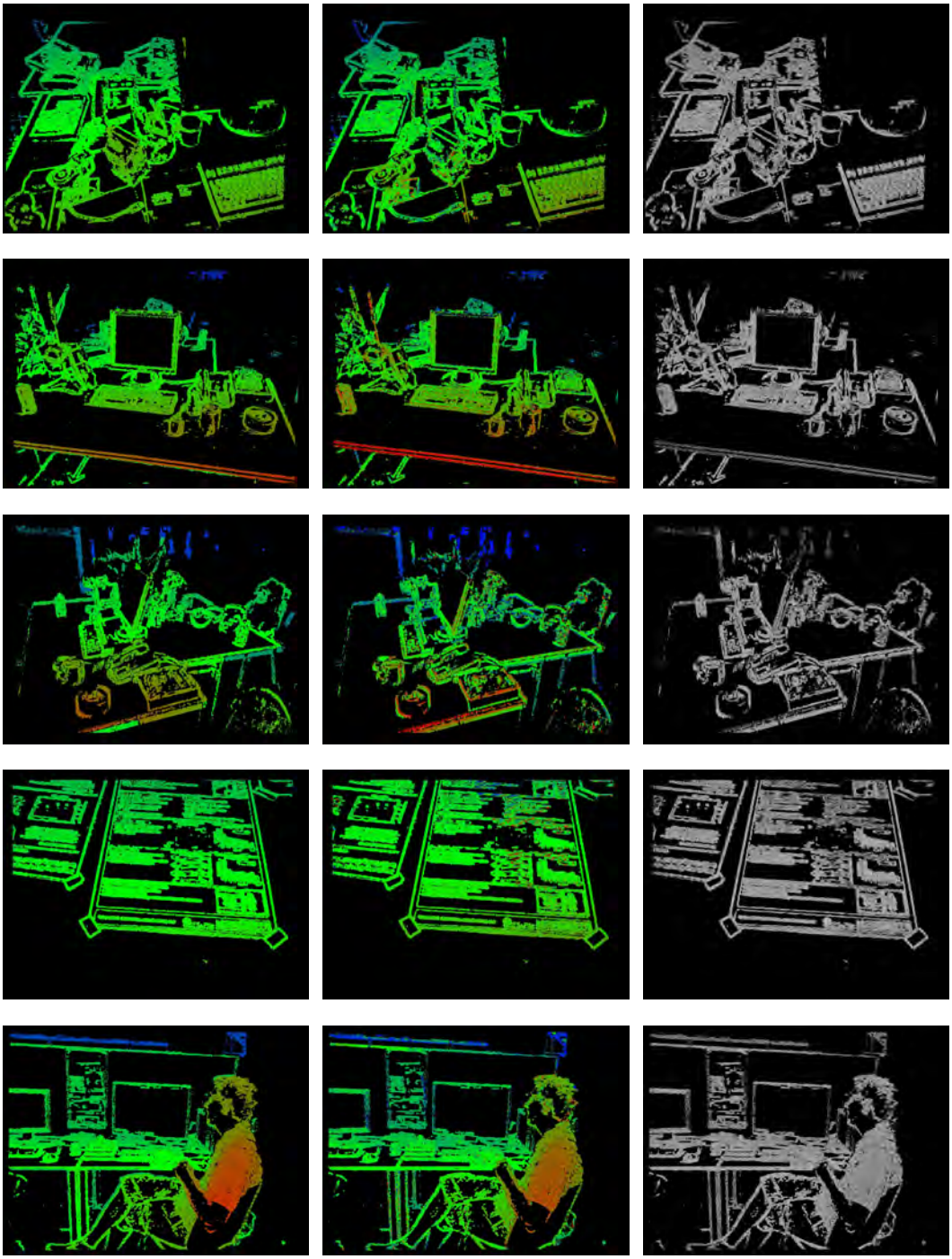
为了分析讨论选取可信控制点的阈值。如图2-5 a)所示的曲线表示在不同阈值下，模型错误判断的概率。阈值设置的越高，模型的错误率就会越低。如图2-5 b)所示是非可信控制点在不同阈值下的密度。阈值设置的越高，被选出来的可信控制点就会越少。本文希望得到高精度的可信控制点预测模型，同时，希望可信控制点的密度不要太小，否则过少的可信点可能会影响SLAM系统接下来对于深度的估计。因此，需要通过实验，选取合适的阈值。实验分别在*fr2_desk*、*fr1_xyz* 和*fr2_xyz* 三个序列上对阈值的选取做了讨论。可以看出，在三个序列上，可信控制点的密度和精度随阈值变化的趋势基本相同。如图所示根据结果，折中模型精度和可信控制点的密度，实验将阈值设置为0.7。



a) 不同阈值下的模型错误率　　b) 不同阈值下的可信控制点密度

图 2-5 可信控制点阈值分析

图2-6展示的是LSD-SLAM原始的深度估计图，其对应的置信度预测图和真实图。图2-6 a)列是用彩色化的真实图，图2-6 b)列是LSD-SLAM估计的深度图，图2-6 c)列是映射到0-255区间上的置信度图。通过观察实验结果可以看出，深度估计不准确的地方，得到的置信度就比较低，即置信度图灰度值小。对比观



a) b) c)

图 2-6 置信度估计

察彩色的半稠密的深度图和预测的置信度图进一步表明，本章提出的方法可以检测出不可信的深度估计。证明方法是可信的。

为了验证随机森林模型的准确度，实验选取了$fr1_xyz$、$fr2\_xyz$、$fr2\_desk$、$fr2\_deskwp$、$fr3\_sithalf$、$fr3_sitsyz$和$fr3_nostructure_texture_near_withloop$场景序列作为测试集。每个数据集的置信度评估结果如表2-2所示。为了保证得到的可信控制点的准确度和密度，根据前面的实验，将随机森林中树的棵树设置为30，将阈值设置为0.7。表中的可信深度表示像素点的真实值和LSD-SLAM的估计深度的误差归一化到[0, 1] 之后，归一化值大于0.7，不可信深度为归一化值小于等于0.7。第一列和第三列是正确估计的像素数量，第二列和第四列是估计不准确的像素数量。最后一行是各个数据集分别对于深度可信和不可信的像素点的平均预测准确度。从实验结果可以看出，本章提出的模型无论是对于深度可信的点还是对于深度不可信的点，预测的正确率都达到了较高的水平，分别为83.17%和89.15%。由此说明本章提出的可信控制点模型可以较为准确地预测出可信控制点。

表 2-2 可信控制点模型的准确度

	可信深度		不可信深度	
数据集	预测值> 0.7	预测值≤ 0.7	预测值≤ 0.7	预测值> 0.7
fr1_xyz	6,561,645	2,338,623	2,572,890	575,845
fr2_xyz	12,920,669	1,417,387	11,170,000	2,484,462
fr2_desk	20,591,387	4,686,548	2,655,680	160,108
fr2_deskwp	15,180,134	4,305,554	10,805,695	1,266,768
fr3_sithalf	1,137,471	81,181	2,101,169	157,178
fr3_sitxyz	6,406,993	1,357,164	2,072,123	119,684
fr3_nonear	207,983	50,454	169,175	89,369
Average	82.79%		85.76%	

2.4.3 随机森林方法的选择与讨论

立体匹配的置信度常用判断方法是决策树。本实验对比了随机森林模型和决策树算法。同时还对比了机器学习常用的SVM模型和感知器算法。对比实验采用的相同的训练集和相同的测试集。训练集与2.4.2中选用的相同，选择了$fr3_nostructure_texture_near_withloop$数据集。实验对比了不同方法的时间效率和结果精度，选择的阈值是0.7。

表 2-3 不同方法时间对比

方法	随机森林	决策树	SVM	感知器
时间(px/ms)	0.0383	0.0011	7.6696	0.0934

表 2-4 可信控制点模型的准确度

	可信深度			不可信深度			
方法	预测> 0.7	预测≤ 0.7	准确度	预测≤ 0.7	预测> 0.7	准确度	总准确度
随机森林	207,983	50,454	80.47%	169,175	89,369	65.43%	72.95%
决策树	253,827	4,610	98.21%	2,583	255,961	0.99%	49.59%
SVM	61,808	196,629	23.91%	192,000	66,544	74.26%	49.59%
感知器	51,758	206,400	20.04%	254,922	3,901	98.49%	59.32%

表2-3展示的是这几种模型在数据集上的时间对比。时间为估计每个像素点用的时间，利用总时间和像素点数相除求得，时间的单位是毫秒（ms）。从表中结果可以看出，在时间方面，随机森林方法虽然不如决策树算法，但是相较于其他方法，在计算效率上是有优势的。

表2-4展示的是在准确的方面的结果。其中第一列和第四列是估计准确的像素点个数，第二列和第五列是估计不准确的像素个数，第三列和第六列分别是深度可信点和深度不可信点的估计准确度，最后一列是整体的估计准确度。从结果中可以看出，随机森林算法的准确度，较其他方法来说有明显的优势。综合考虑估计精度和估计时间，本章选取的随机森林模型是这几种方法中最优的选择。

2.5 本章小结

本章提出了一种基于可信控制点的深度置信度预测模型。模型提出的目的是为了改进直接法的估计结果。直接法虽然有诸多优点，但是受到灰度不变强假设的制约，直接法的估计结果并没有理想中的那么鲁棒。这也是制约LSD-SLAM算法精度的一个重要原因。本章利用LSD-SLAM中关键帧和其参考帧进行立体匹配的过程中提取特征信息作为训练的特征，这些信息都是易提取的特征和图片的颜色信息，以深度估计值和深度估计值的差值作为标签，使用机器学习的方法学习了深度置信度的预测模型。

模型选用的机器学习方法是随机森林方法。随机森林方法可以在模型训练结束之后，给出每维特征的权值，便于之后的研究和分析。并且随机森林方法

的优越性已经在各个领域得到过验证。本章通过实验2.4.3，比较了随机森林算法、决策树算法、SVM 和感知器算法的时间效率和结果精度。验证了随机森林方法在本章研究的问题上，综合时间效率和准确度来说，是所比较方法中最好的。

在模型验证方面，本章首先利用传统的立体匹配算法，验证了模型的可行性。根据实验2.4.1中的结果，模型可以较为准确的预测出视差（深度）置信度高的点。接着，本章利用在LSD-SLAM的视觉里程计环节提取到的数据，对模型的准确度做了验证，并对模型的参数和阈值选取做了讨论。从结果上看，本章提出的模型可以较为准确的估计出深度的置信度，并通过讨论得到了折中准确度和可信控制点密度的参数和阈值。

第 3 章 基于卷积特征高阶统计的闭环检测

3.1 引言

上一章中，本课题提出了一种基于可信控制点的深度置信度预测模型，改善了SLAM中视觉里程计部分的估计精度。本章将针对闭环检测环节，提出一种基于高阶特征的闭环检测网络。本章尝试利用二阶特征进行闭环检测，二阶特征已经在其他领域取得了不错成绩的。在训练方面，本章采用的弱监督的训练方式，采用了三元组损失函数，并制定了三元组选择策略，使训练更加合理有效。通过设计实验，比较了本章提出模型与同为基于深度学习的NetVLAD。此外，本章还通过实验对于了提出模型和LSD-SLAM中作为闭环检测的FAB-MAP方法。

3.2 二阶特征及二阶特征在闭环检测中的应用

通过反向传播算法，神经网络可以将分布复杂的特征分成百上千个类别。现有的大部分网络都是通过改变网络的深度或宽度结构，来提高识别或分类的精度。只有少量的工作着手于特征的分布，提取高阶的特征信息。在数据不是很充足的情况下，神经网络加上的高阶特征方法，在结果的精确度上得到了不错的提升[25, 26]。

FisherNet[27]和NetVLAD是两种使用了二阶特征信息网络。费舍尔向量（Fisher Vector，FV）是一种对已有的特征采用高斯混合模型进行编码的方法，它可以得到二阶的特征信息。FisherNet是将FV实现成为一个可训练的池化层，添加在深度神经网络中，识别结果的精度得到了提升。NetVLAD 将卷积网络得到的特征，通过VLAD 池化层映射到VLAD 空间，得到了二阶的VLAD特征，显著地提高了地点识别的精度。这两个网络都是将方法近似简化成为卷积层、softmax层和池化层，这些都是在神经网络中已经存在的层，可以直接方便地将可训练的FV层或VLAD层加入已有的网络。

在所有的概率分布中，全局高斯分布是最接近方差，即二阶特征的。全局高斯分布并不能像上边的两个方法一样，直接分解，近似简化成神经网络中已有的层。但是，在G2DeNet中，已经实现并给出了高斯嵌入层正向和反向传播的推导。

虽然也有其他的全局的构造矩阵在神经网络中的正向、反向传播框架，并且有基于此的模型$DeepO_2P$[28]，$DeepO_2P$的核心也是在卷积神经网络中添加了一个可训练的二阶池化层。它基于奇异值分解和特征值分解的理论，并应用于图片分类。但是，$DeepO_2P$ 使用的是对数欧拉几何，对特征值矩阵要求严格，可能会对结果产生负面的影响。BCNN[29]是一种将两个CNN网络得到的特征做外积，池化，并正则化的网络。当两个CNN网络不同时，BCNN可以将来自不同网络的特征相关联；当两个CNN网络相同时，和$DeepO_2P$相类似，这一系列操作就会产生二阶非中心矩情况。并且上述两个网络都是做小规模分类，都没有应用在大规模的地点识别上。本课题使用的提取二阶信息的方法，能得到更近似真实协方差的估计，并且更加鲁棒，不会产生特殊情况造成的分歧。因而本课题提出的网络不会产生上述问题，并且可以应用在大规模的地点识别上。

3.3 基于二阶信息的深度网络

本课题使用MPN-COV[30]方法，在网络中加入二阶的特征信息。设最后一层卷积网络的输出为$\mathbf{X}$，首先计算出$\mathbf{X}$的样本协方差矩阵$\mathbf{P}$。然后将$\mathbf{P}$进行特征值分解，得到正交矩阵$\mathbf{U}$和对角矩阵$\mathbf{\Lambda}$。通过幂矩阵$\mathbf{Q} \triangleq \mathbf{P}^{\alpha}$，转化成求$\mathbf{P}$的特征值的幂。$\mathbf{Q}$作为二阶信息层的输出。本章提出的深度网络的网络示意图如图3-1所示。

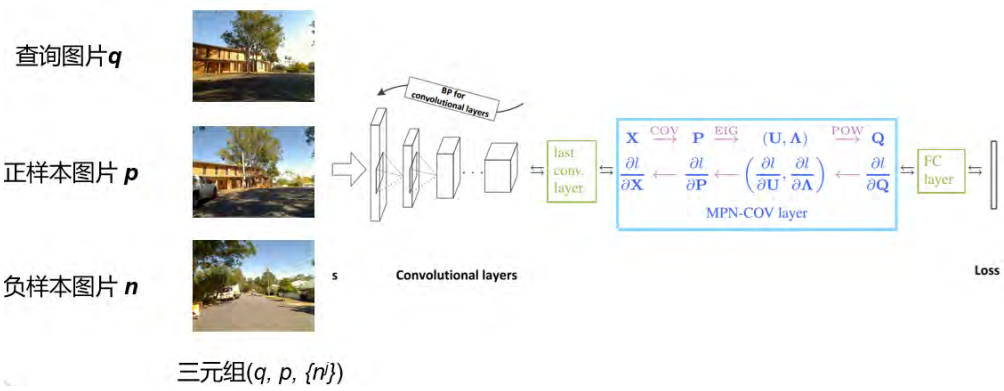


图 3-1 网络示意图

3.3.1 正向传播

设$\mathbf{X} \in \mathbb{R}^{d \times N}$是一个有$N$个$d$为特征的矩阵，它的样本协方差矩阵$\mathbf{P}$可以被计算为

$$\mathbf{X} \rightarrowtail \mathbf{P}, \quad \mathbf{P} = \mathbf{X}\bar{\mathbf{I}}\mathbf{X}^T \tag{3-1}$$

其中，$\bar{\mathbf{I}} = \frac{1}{N}(\mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T)$，$\mathbf{I}$是一个$N \times N$的单位矩阵，$\mathbf{1} = [1, \ldots, 1]^T$是一个$N$维的向量。$\mathbf{P}$是一个对称半正定矩阵，可以将其特征值分解为

$$\mathbf{P} \rightarrowtail (\mathbf{U}, \mathbf{\Lambda}), \quad \mathbf{P} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \tag{3-2}$$

其中，$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$是一个非增顺序排列的特征值对角矩阵。$\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_d]$是一个正交矩阵。其中$\mathbf{u}_i$是$\lambda_i$的特征向量。将矩阵的幂转化成特征值的幂，有

$$(\mathbf{U}, \mathbf{\Lambda}) \rightarrowtail \mathbf{Q}, \quad \mathbf{Q} \triangleq \mathbf{P}^\alpha = \mathbf{U}\mathbf{F}(\mathbf{\Lambda})\mathbf{U}^T \tag{3-3}$$

其中，α是一个正实数，$\mathbf{F}(\mathbf{\Lambda}) = \mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_d))$，$f(\lambda_i)$是特征值的幂函数，$f(\lambda_i) = \lambda_i^\alpha$。

3.3.2 反向传播

在反向传播过程中，首先通过$\frac{\partial l}{\partial \mathbf{Q}}$，求出$\frac{\partial l}{\partial \mathbf{U}}$和$\frac{\partial l}{\partial \mathbf{\Lambda}}$，根据求导的链式法则可以得到，

$$\mathrm{tr}((\frac{\partial l}{\partial \mathbf{U}})^T \mathrm{d}\mathbf{U} + (\frac{\partial l}{\partial \mathbf{\Lambda}})^T \mathrm{d}\mathbf{\Lambda}) = \mathrm{tr}((\frac{\partial l}{\partial \mathbf{Q}})^T \mathrm{d}\mathbf{Q}) \tag{3-4}$$

根据公式（3-3），可以得到$\mathrm{d}\mathbf{Q} = \mathrm{d}\mathbf{U}\mathbf{F}\mathbf{U}^T + \mathbf{U}\mathrm{d}\mathbf{F}\mathbf{U}^T + \mathbf{U}\mathbf{F}\mathrm{d}\mathbf{U}^T$，其中$\mathrm{d}\mathbf{F} = \mathrm{diag}(\alpha\lambda_1^{\alpha-1}, \ldots, \lambda_d^{\alpha-1})\mathrm{d}\mathbf{\Lambda}$。整理之后，得到

$$\begin{aligned} \frac{\partial l}{\partial \mathbf{U}} &= (\frac{\partial l}{\partial \mathbf{Q}} + (\frac{\partial l}{\partial \mathbf{Q}})^T)\mathbf{U}\mathbf{F} \\ \frac{\partial l}{\partial \mathbf{\Lambda}} &= \alpha(\mathrm{diag}(\alpha\lambda_1^{\alpha-1}, \ldots, \lambda_d^{\alpha-1})\mathbf{U}^T \frac{\partial l}{\partial \mathbf{Q}}\mathbf{U})_{\mathrm{diag}} \end{aligned} \tag{3-5}$$

其中$\mathbf{A}_{\mathrm{diag}}$是一种保留矩阵$\mathbf{A}$的对角值，并将其非对角的值都置为0的操作。接下来，用求得的$\frac{\partial l}{\partial \mathbf{U}}$和$\frac{\partial l}{\partial \mathbf{\Lambda}}$，根据公式（3-2）推导出求解$\frac{\partial l}{\partial \mathbf{P}}$的公式

$$\frac{\partial l}{\partial \mathbf{P}} = \mathbf{U}((\mathbf{K}^T \circ (\mathbf{U}^T \frac{\partial l}{\partial \mathbf{U}})) + (\frac{\partial l}{\partial \mathbf{\Lambda}}))\mathbf{U}^T \tag{3-6}$$

其中，$\circ$表示矩阵的克罗内克乘积。矩阵$\mathbf{K} = \{K_{ij}\}$，当$i \neq j$时，$K_{ij} = 1/(\lambda_i - \lambda_j)$，否则，$K_{ij} = 0$。最后用$\frac{\partial l}{\partial \mathbf{P}}$求出输入矩阵$\mathbf{X}$的梯度

$$\frac{\partial l}{\partial \mathbf{X}} = \bar{\mathbf{I}}\mathbf{X}(\frac{\partial l}{\partial \mathbf{P}} + (\frac{\partial l}{\partial \mathbf{P}})^T) \tag{3-7}$$

3.3.3 协方差正则化方法

样本协方差等于正态分布随机向量的极大似然估计解。虽然最大似然估

计（Maximum Likelihood Estimation，MLE）被广泛地运用在协方差估计中。但是它在样本维度大、数量少的情况下的表现很差。而这就是本课题研究所面临的问题。在大多数的卷积网络中，最后一层输出特征的维度基本都会大于特征个数。这就会导致协方差矩阵总是不满秩，使估计的鲁棒性降低。

有很多为了鲁棒地估计协方差的修正最大似然估计的方法。MPN-COV中也提出了一种鲁棒地估计协方差的方法，根据vN-MLE有，

当$\alpha = \frac{1}{2}$时，MPN-COV方法可以得到协方差矩阵的正则化最大似然估计的唯一解。

$$\mathbf{P}^{\frac{1}{2}} = \arg\min_{\mathbf{\Sigma}} \log|\mathbf{\Sigma}| + tr(\mathbf{\Sigma}^{-1}\mathbf{P}) + D_{vN}(\mathbf{I}, \mathbf{\Sigma}) \tag{3-8}$$

其中，$\mathbf{\Sigma}$是半正定矩阵。经典的MLE只包含了公式（3-8）的前两项，而vN-MLE中包含的第三项可以将协方差矩阵约束成为近似单位矩阵，可以得到比传统MLE更好的效果。

因为协方差矩阵的空间符合黎曼流形，需要选取适当的度量方法。主要有两种黎曼流形上的度量方法。仿射黎曼度量和Log-E度量。仿射黎曼度量有仿射不变形，但是计算效率低并且耦合。被更广泛使用的Log-E度量有相似不变形，计算效率高并且不耦合。MPN-COV使用的度量是Pow-E[31]度量与Log-E[32]度量有着紧密的联系。

对于任意两个协方差矩阵$\mathbf{P}$和$\tilde{\mathbf{P}}$，Pow-E度量$d_\alpha(\mathbf{P}, \tilde{\mathbf{P}}) = \frac{1}{\alpha}\|\mathbf{P}^\alpha - \tilde{\mathbf{P}}^\alpha\|_F$。当$\alpha$正向趋近于0时，Pow-E度量等于Log-E度量$\lim_{\alpha\to 0} d_\alpha(\mathbf{P}, \tilde{\mathbf{P}}) = \|(\mathbf{P}) - (\tilde{\mathbf{P}})\|_F$。

Log-E度量是被最广泛使用的黎曼度量，但是Log-E要求特征值严格为正，而Pow-E只要求特征值非负。虽然Log-E可以通过给特征值加一个极小的整数ϵ来保证特征值为正，但是ϵ的值很难进行选取，并且很可能会对结果产生负面的影响。即使Log-E是真实的距离，而Pow-E得到的只是近似的距离。但是考虑到Log-E会带来了弊端，Pow-E距离是相较而言的更好的选择。

3.4 弱监督训练

本文采用了三元组的损失函数。本课题需要解决的问题是判断两张图片是否来自同一个地点。如果采用传统的softmax函数作为损失函数，就需要将每个地点都作为一个类别进行训练。这样的话softmax函数的维数会非常大，不利于计算，而且将每个地点都作为一个类别的想法，在实际中也是不可能的。本课题采用的三元组损失函数，已经被应用在各个领域，比如用于人脸识别[33]，行人重识别[34]，并取得了不错的效果。

图 3-2 三元组损失函数整体思想

三元组损失函数的整体思想如图3-2所示。三元组(q,p,n)的构成为：作为参考的查询样本q，与查询样本同一类的正样本p，与查询样本不同类的负样本n。将三元组通过神经网络训练得到的特征分别记为$f_\theta(q),f_\theta(p)$和$f_\theta(n)$，将特征之间的距离函数记为d。三元组损失函数学习的目的就是尽可能拉近q和p特征之间的距离，尽量使q和n表达的特征之间的距离更远，即

$$d(f_\theta(q),f_\theta(p)) < d(f_\theta(q),f_\theta(n)) \tag{3-9}$$

在训练过程中，训练的目标是对于每个三元组(q_i,p_i,n_i)，都使得公式（3-9）成立，即查询样本和正样本之间的距离小于查询样本和负样本之间的距离。在此基础上，本课题希望$f_\theta(q)$与$f_\theta(p)$之间的距离尽量小，$f_\theta(q)$ 与$f_\theta(n)$之间的距离大。于是引入一个阈值m，其表示$d(f_\theta(q),f_\theta(p))$与$d(f_\theta(q),f_\theta(n))$ 之间的最小间隔。所以，目标函数定义为

$$L = max(d^2(f_\theta(q_i),f_\theta(p_i)) - d^2(f_\theta(q_i),f_\theta(n_i)) + m, 0) \tag{3-10}$$

从目标函数中可以看出，当$f_\theta(q_i)$与$f_\theta(p_i)$之间的距离和$f_\theta(q_i)$与$f_\theta(n_i)$之间的距离相差小于m时，即距离间隔不够大是，会产生损失；否则，损失函数为0。

在反向传播中，当损失函数不为零时，分别对$f_\theta(q_i),f_\theta(p_i),f_\theta(n_i)$求导，有

$$\begin{aligned}
\frac{\partial L}{\partial f_\theta(q_i)} &= 2\cdot(f_\theta(q_i)-f_\theta(p_i)) - 2\cdot(f_\theta(q_i)-f_\theta(n_i)) = 2\cdot(f_\theta(n_i)-f_\theta(p_i)) \\
\frac{\partial L}{\partial f_\theta(p_i)} &= 2\cdot(f_\theta(q_i)-f_\theta(p_i))\cdot(-1) = 2\cdot(f_\theta(p_i)-f_\theta(q_i)) \\
\frac{\partial L}{\partial f_\theta(n_i)} &= -2\cdot(f_\theta(q_i)-f_\theta(n_i))\cdot(-1) = 2\cdot(f_\theta(q_i)-f_\theta(n_i))
\end{aligned} \tag{3-11}$$

为了得到更好的训练效果，本课题选取k个负样本作为一个集合，而不是只选择一个负样本。这样做约束网络朝着更有利于地点区分的方向发展，为参数的训练增加了更多有利的约束。因此，公式中的目标函数将变为

$$L = \sum_j max(d^2(f_\theta(q_i),f_\theta(p_i)) - d^2(f_\theta(q_i),f_\theta(n_i^j)) + m, 0) \tag{3-12}$$

在反向传播中，也会添加相应的求和操作。

3.5 实验结果

3.5.1 实验介绍

本章实验分为两个部分：首先，对本章提出的模型精度进行验证实验。选用的数据集为Tokyo Time Machine（TokyoTM）、Tokyo24/7[35]数据集。主要对比的模型是同样用深度学习方法解决闭环检测问题的NetVLAD模型。由于本课题的目的在于得到高精度的SLAM算法，下一章会在LSD-SLAM的基础上，加上本章提出的模型进行对比和讨论。因此，在这里首先对比本章提出的闭环检测模型与LSD-SLAM中使用的FAB-MAP模型。实验的数据集是St Lucia[36] 数据集、NordLand[37]数据集和CMU_VL[38]数据集。

首先介绍第一部分实验用到的数据集。

a) b) c)

图 3-3 TokyoTM数据集图片示例。

a) 白天 b) 日落 c) 夜晚 d) 候选集对应图片

图 3-4 Tokyo24/7数据集图片示例。

TokyoTM数据集分为训练集和验证集，训练集的备选图片集有4.9万张图片，查询集有7277张图片；验证集的备选集有4.9万张图片，查询集有7186张图片。该数据集是利用Google Street View拍摄的街景图像生成的数据集。数据集中包含多个地点场景及地点的真实坐标，12 张图片为一组，包含了同一位置不同角度的照片，每个地点有多组图片。TokyoTM 数据集图片示例见图3-3，其中，图3-3 b)列与图3-3 a)列为相同时间地点，不同角度的图片，图3-3 b)列与图3-3 c) 列为相同地点角度，不同时间的图片。本次实验采用TokyoTM数据集作为训练数据集，并使用其验证集测试网络的效果。

Tokyo24/7数据集的备选图片集7.6万张图片，查询集有315张图片。其备选图片集也是来自Google Street View，查询集的图片是使用手机实际拍摄的图片，同样包含了所有图片的地点坐标。这是一个极富挑战性的数据集，它的询问数据集的图片包含了白天、日落和夜晚的图片，而备选数据集的只包含了白天的图像。Tokyo24/7 的示例图片见图3-4 所示。其中图3-4 a)、图3-4 b)、图3-4 c)分别为同一地点在白天、日落、夜晚的图片，图3-4 d) 为候选集中与之相近的图片。本次实验采用Tokyo24/7数据集作为测试数据集。

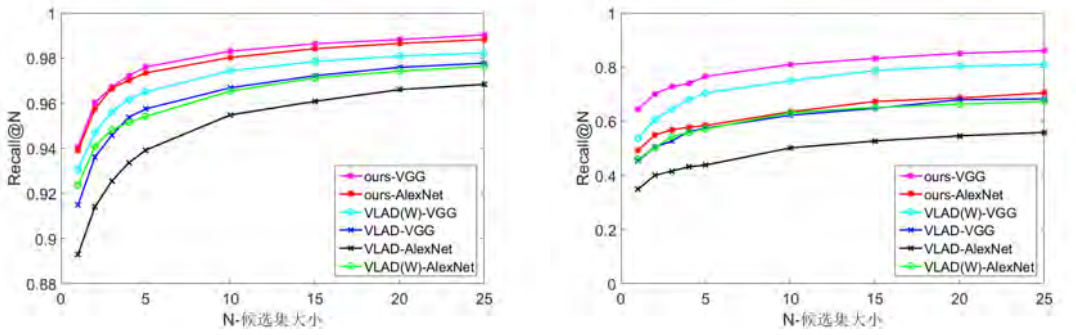
3.5.2 三元组的选择

本次实验采用TokyoTM数据集作为训练数据集。采用了动态选取三元组的方法。动态选取三元组的方法是：在与询问图片相同地点的图片中，选择特征距离和询问图片最小的一个样本作为正样本。为了加快训练的速度，随机选取100张与询问图片不同地点的样例，提取它们的特征，从中选选取10张与询问图片特征距离最近的10张图片作为负样本。

3.5.3 模型精度实验结果

实验用到的网络结构是VGG-16[39]+MPN-COV和AlexNet[40]+MPN-COV。损失函数使用的是上文中提到的三元组损失函数。直接使用网络输出的特征的欧氏距离作为对应样本地点的相似程度的判断标准。为了缩短训练的时间，得到更好的训练效果，本课题使用的是在ImageNet上预训练好的VGG-16和AlexNet。

为了验证本课题使用的网络，是否在结构上优于NetVLAD。首先不经过训练，直接测试精度。得到了如图3-5所示的实验结果。图3-6所示是经过了训练的实验结果。在实验中，本课题分别对比了NetVLAD和加了白化之后的NetVLAD，测试了网络在Tokyo24/7和TokyoTM的验证集



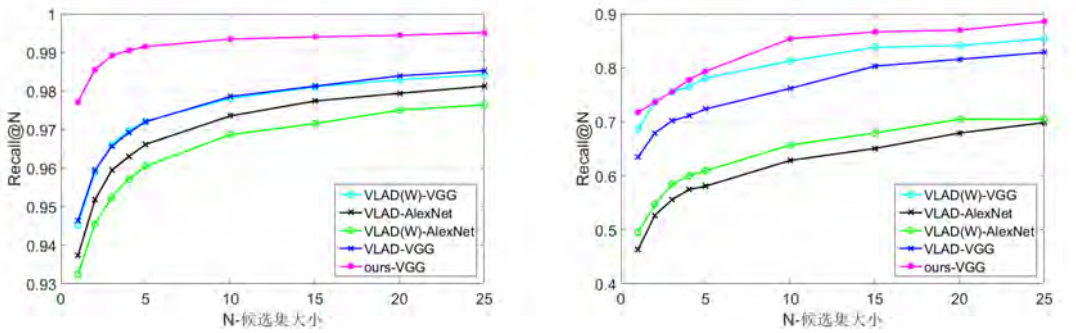
a) TokyoTM-val　　b) Tokyo24/7

图 3-5 未经训练的测试结果

上的结果。图中Recall@N的意义为，在备选集中选取前N个样本作为候选集，如果候选集中存在和测试样本相同地点的样本，则判断为识别成功。粉红色的线ours-VGG是VGG16+MPN-COV的结果，红色的线ours-AlexNet是AlexNet+MPN-COV的结果。蓝色的线是使用VGG16的VLAD(VLAD-VGG)的结果，浅蓝色的线是加了白化(VLAD(W)-VGG) 的结果，黑色的线是使用AlexNet 的VLAD(VLAD-AlexNet)的结果，绿色的线是加了白化(VLAD(W)-AlexNet) 之后的结果。

从实验结果可以看出，在未经训练的情况下，本课题的网络不论是在TokyoTM的验证集(图3-5 a))上，还是在Tokyo24/7(图3-5 b))数据集上，测试精度上高于NetVLAD。因此，本课题提出的网络结构在结果上优于NetVLAD。

在经过训练的网络上对上述数据集进行测试。对比的两个网络训练使用的训练集相同，训练的迭代次数都相同，均为30次。在测试结果中，图3-6 a)所示的是在TokyoTM验证集上的测试结果，本课题的网络的精度高于NetVLAD，相



a) TokyoTM-val　　b) Tokyo24/7

图 3-6 动态选择三元组训练的测试结果

比于图3-5 a)中的结果，经过训练之后的网络相比于经过同样训练的NetVLAD，结果有了更加明显的提高；在图3-6 b) 所示的Tokyo24/7测试集的测试结果中，本课题网络的结果也要高于NetVLAD。实验直接在用TokyoTM训练的网络模型上对Tokyo24/7测试集进行测试的。TokyoTM中都是白天光照下的图片，并没有Tokyo24/7中的日落或夜晚的场景。尽管如此，本章提出的模型还是得到了较高的精度，说明本章提出的模型对于光照等变化具有一定的鲁棒性。

总的来说，未经训练直接进行测试的实验，验证了所提出网络结构是有一定的优越性的。在经过了训练之后，仍然可以在测试中保持优势，说明本实验使用的训练方法有效。因此，本章提出的闭环检测模型，无论从网络结构还是模型的训练策略上来说，都是优越的。

3.5.4 SLAM常用算法对比实验

FAB-MAP是SLAM系统中常用的闭环检测方法，也是LSD-SLAM中使用的方法。为了验证本章提出的模型与SLAM实际应用的模型相比的优势。本实验分别在St Lucia数据集、NordLand数据集和CMU_VL数据集上，对本章提出的模型与FAB-MAP做了对比实验。实验数据集总览见表3-1。实验使用的模型是在TokyoTM数据集上训练的，并没有使用测试数据集进行参数微调。FAB-MAP使用的是开源的openFABMAP[41]代码。

表 3-1 实验数据集总览

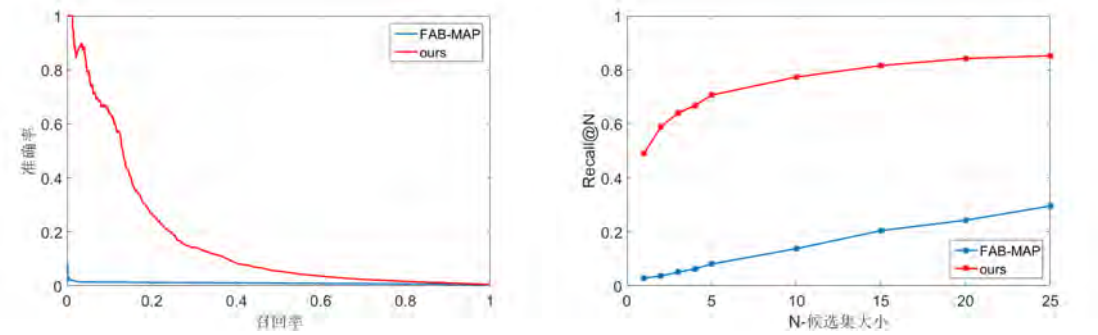
数据集	规模(帧)	分辨率(像素)	帧频(帧每秒)	特点
St Lucia	4 × 22,000	648×480	15	一天不同的时间段
NordLand	4 × 900,000	1920×1080	25	不同的季节
CMU_VL	5 × 13,000	1024×768	15	不同的月份

St Lucia是一个用车载单目相机拍摄的数据集。它是在两周期间，拍摄了St Lucia地区不同时间段下的视频。每个视频都用对应的GPS信息，可以用来作为数据的真实值使用。St Lucia 数据集包含了一天当中，不同时间段的图像。一天中的光照变化非常明显，在光照鲁棒性上对测试的模型极具挑战。同时，道路上的情况也非常复杂。会有行人、车辆等因素，对测试模型造成干扰。

本实验选取了2009年8月18日15:45和2009年9月10日10:00拍摄的两段视频进行测试。St Lucia数据集的视频的帧频为每秒15帧，实验每秒选取了一帧进行测试。如图3-7所示是下边一行是为2009年8月18日15:45拍摄的查询图片，上边

面一行是本章模型测出的2009年9月10日10:00拍摄的图像序列中最接近的图像。可以看出两个图片序列的光照情况有很大差别。

图 3-7 St Lucia数据集图片示例。



a) precision-recall b) recall@N

图 3-8 St Lucia数据集实验结果

如图3-8 a) 所示是本章提出的网络模型与FAB-MAP方法在St Lucia 数据集上的准确率-召回率曲线。可以看出，本章提出的网络，在结果上远高于FAB-MAP。如图3-8 b)所示是提出的网络模型和FAB-MAP，在相同大小的候选集下的召回率比较。Recall@N的意义和上文中相同。从实验结果可以看出，在候选集大小相同的情况下，本文提出的网络结构，在召回率上远高于FAB-MAP方法。

NordLand数据集包含四个10小时左右的火车车头拍摄的视频，视频为每秒15帧，同时记录了全程的GPS信息，可以作为真实值使用。四个视频分别在春夏秋冬四个季节进行拍摄，包含了明显的季节变化和光照变化，同时，由于数据集拍摄的都是同一个岛上的自然景色，数据集中的很多景象都比较相似，

也具有一定的感知混淆。因此，Nordland数据集也是一个富有挑战的数据集。

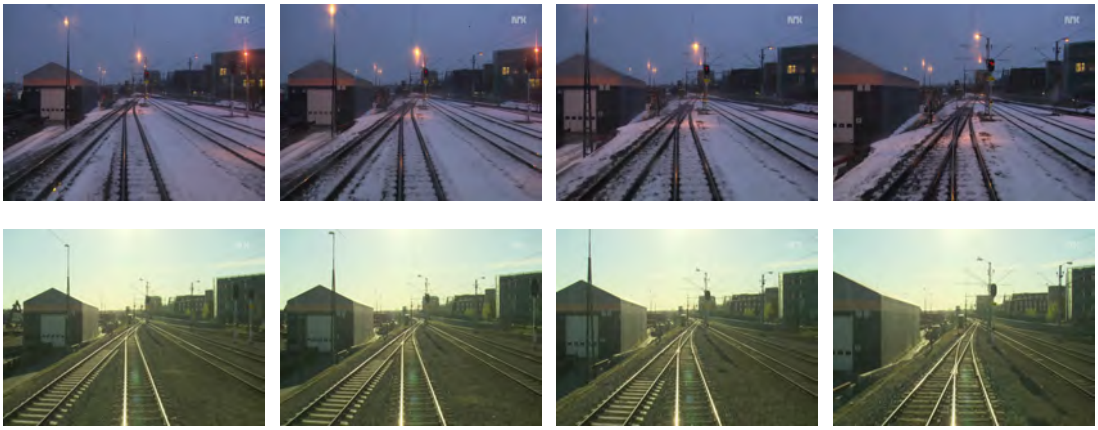
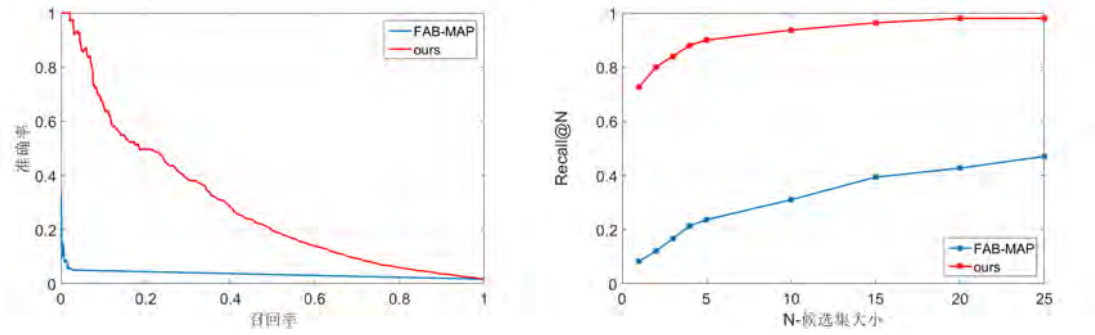


图 3-9 Nordland数据集图片示例。



a) precision-recall　　b) recall@N

图 3-10 Nordland数据集实验结果

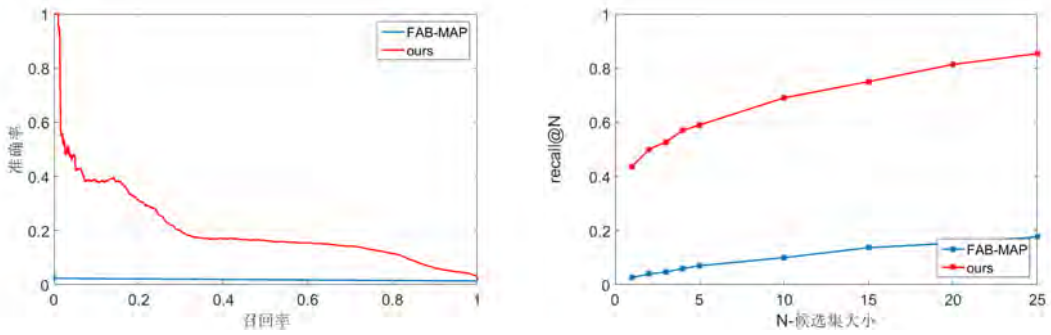
本实验选取了春天和冬天拍摄的图像作为测试，其中春天的图像是查询图像集，冬天的为备选图像集。实验选取了每秒一帧进行测试，每个视频选取了300帧。如图3-9所示，为Nordland数据集的示例图像，分别为春天拍摄的数据集和本章模型预测出的冬天拍摄的对应图像。

本章提出的网络模型与FAB-MAP方法在Nordland数据集上的准确率-召回率曲线对比，如图3-10 a)所示。本章提出的网络，优于FAB-MAP 方法的结果。如图3-10 b)所示是提出的网络模型和FAB-MAP，在相同大小的候选集下的召回率比较。从实验结果可以看出，在候选集大小相同的情况下，本文提出的网络结构，在召回率上优于FAB-MAP方法。

CMU-VL数据集是在不同的月份，使用车载摄像头在一个小镇拍摄的图像。数据集中包含了不同了月份、不同天气、不同植被的情况，给闭环检测

任务带来了光照，季节造成的植被，雨雪天气等挑战。数据集中同样也提供了GPS信息，可以作为验证的真实值使用。数据集中包含了左右两个视角的图像序列。本实验选取了左侧摄像头拍摄的图像序列进行实验。

图 3-11 CMU-VL数据集图片示例。



a) precision-recall　　b) recall@N

图 3-12 Nordland数据集实验结果

本实验选取了2010年9月15日和2010年12月21日拍摄的两段视频进行测试，其中12月21日的图像序列为备选集，9月15日拍摄的图像作为查询集。图3-11展示的是CMU-VL数据集中的查询图像，及其通过模型预测出的对应的备选集图像。

图3-12展示的是本章提出的网络模型和FAB-MAP方法在CMU-VL数据集上的结果。其中图3-12 a)是两种方法准确率- 召回率曲线的对比，本章提出的网络模型在结果上优于传统的FAB-MAP；图3-12 b)是在相同大小的候选集下的召回率比较，在CMU-VL数据集上，也得到了与前两个数据集类似的结果，本章提出了网络模型的结果优于SLAM广泛使用的FAB-MAP方法。

从上边的实验结果来看，本章提出的网络模型在精度上高于FAB-MAP算法。如表3-2所示，是两个模型在上面的三个数据集上的测试时间对比，表中时间的单位是秒。从实验结果上看，本章提出的网络模型在使用GPU计算的情况下，在时间效率上优于openFABMAP算法。但是在只是用CPU计算的情况下，网络模型的速度不如openFABMAP。

表 3-2 模型运行时间对比

模型	St Lucia	NordLand	CMU_VL
ours(GPU)	187.32	96.33	115.93
ours(CPU)	1273.97	745.39	881.72
openFABMAP	417.65	180.93	711.21

综上，本章提出的闭环检测网络模型在精度上远高于SLAM常用的FAB-MAP算法。在只是用CPU计算的情况下，FAB-MAP算法的运行时间更短。但是在使用GPU计算的情况下，网络模型也能达到很好的效率。

3.6 本章小结

本章提出了一种高精度的闭环检测网络。闭环检测是SLAM中的重要一环，它可以对运动估计阶段累积的误差做及时有效地修正。因此，闭环检测模块的精度决定了SLAM系统是否能及时地发现并更正误差。

本章提出的闭环检测模型使用了深度学习的方法，并在其中加入了二阶的特征信息。不同于其他方法，本章加入的是符合高斯分布的协方差特征，使网络更适于图像的分布。在网络的训练方面，由于没有特定的标签，本章采用了三元组损失函数进行弱监督训练。通过训练是相同地点的图像特征不断得聚类。在实验方法，本章与同为用深度学习方法进行闭环检测的NetVLAD做对比，并取得了更好的结果。

本章还做了提出网络模型与LSD-SLAM中使用的FAB-MAP方法的精度对比。使用的是广泛应用于对比FAB-MAP方法的数据集。在多个数据集上，本章提出的网络模型在结果上都要更好。在时间效率上，使用CPU的深度学习网络不如传统的FAB-MAP高。如果可以加入GPU，模型也可以达到实时的效果。

第 4 章 基于学习方法的SLAM算法

4.1 引言

在前两章中，本课题分别针对视觉里程计部分和闭环检测部分进行了改进。本课题的方案都是基于LSD-SLAM中的不足提出的。本章将前两章提出的模型加入LSD-SLAM算法当中。尝试说明本课题提出的模型不仅在理论效果中，也在实际的SLAM系统中能使重建的效果得到可观的提升。首先，本章将分别对第二章提出的基于可信控制点的深度置信度预测模型和第三章提出的基于高阶特征的闭环检测网络进行实验，对比分别加入两个模型之后的重建效果及轨迹误差。然后将两个模型同时加入LSD-SLAM 当中，并验证模型的精确性和两个模型之间的可容性。

4.2 基于可信控制点的SLAM算法

由于模型误差以及图像噪声误差、量化误差等原因，现有的三维重建算法对深度（视差）的估计都不是非常准确，而深度值对于相机姿态预测有很大的影响，进而决定了三维重建的最终效果。现使用的深度估计的方法主要有两种，一种是基于特征的方法，一种是直接估计的方法。基于特征的方法通过提取FAST等关键点，再在关键点附近提取ORB、SURF等特征。通过特征之间的匹配，估计相机的运动，进而估计出深度。但是这种方法时间效率比较低，利用的图片信息也仅仅局限于特征点，还要求图片有足够强的纹理信息。另一种方法是直接法，它是直接利用像素灰度信息估计相机运动的方法。这种方法不需要计算关键点信息，效率较高，可以使用所有像素进行估计，避免的信息的遗失。但是直接法是基于灰度不变这一强假设的，其在实际当中不可能达到理想的效果。

LSD-SLAM使用的就是一种利用图像之间的立体匹配直接估计深度以及姿态的方法。LSD-SLAM算法没有使用全部的像素的进行深度估计，而是使用的是梯度变化明显的像素点估计相机的运动，建立半稠密的深度图。因为如果使用整张图片的所有像素点进行估计，其实会进行很多无用的计算，产生信息的冗余，同时也会影响算法的时间效率。这种方法虽然省去了提取特征的麻烦，并且使得SLAM系统可以实时地信息跟踪定位和重建，但是受噪声、光照变化

等一些不利因素影响，相机姿态估计的准确度可能会不如一些直接利用图像特征的算法。

如果可以在LSD-SLAM算法中利用可信控制点来改善相机姿态的估计和深度的计算，就能降低其在受噪声、光照等因素影响时产生的误差，理论上就能提高LSD-SLAM 算法对相机姿态估计的准确度，进而改善三维重建的效果。

根据第二章的分析及实验结果，首先，利用可信控制点改善立体匹配是有据可循的。其次，本课题通过2.4.1节中的可信控制点模型的可行性验证实验，证明了算法模型的可行性。第三，通过2.4.2节中对提出的可信控制点预测模型置信度的评估，验证了本课题提出的模型能得到较高的可信控制点预测准确度。同时通过实验讨论，选择了合适的模型参数，保证了模型的时间效率、可信控制点的密度和可信控制点的准确度。最后，根据2.4.3节中的对比实验，验证了本课题所选用的随机森林算法的优越性，即使用随机森林算法训练的模型能在保证时间效率的前提下，得到更加准确的置信度预测。

综上所述，利用基于随机森林的可信控制点选择方法对深度置信度的预测进行预测是可行且鲁棒的。因此，在LSD-SLAM算法中，加入可信控制点模型在理论上是可行的。本章在LSD-SLAM原有算法对深度估计的部分，加入可信控制点模型预测的置信度信息进行融合计算，使得置信度高的深度估计可以占得较大的权重。在对相机姿态进行估计的部分，利用预测结果增加可信度高的点的权重，使相机的姿态得到更为准确的估计。

本课题利用第二章提出的可信控制点模型改善LSD-SLAM的整体流程图如图4-1所示。首先利用LDS-SLAM算法提取出2.3.2节中提到的特征信息。由于特征中有算法估计的深度值，可以直接与真实深度信息进行计算，得到数据的标签。利用特征和数据标签，提前训练出随机森林模型。在深度估计分部分加入一个表示深度置信度的变量，这个变量由可信控制点模型预测得到。LSD-SLAM 算法在跟踪的过程中，根据距上一个关键帧的距离判断当前帧是否被定义为新的关键帧，即当前帧距离上一个关键帧足够远时，就将其定义为新的关键帧，其他帧作为当前关键帧的参考帧，不断地更新和改善对于关键帧的深度估计。在跟踪到一个当前关键帧的参考帧时，计算出所需的所有特征信息，这些信息可以在立体匹配中或者图像信息获得，并不需要消耗太多的计算时间。并用训练出的模型预测每个像素点的深度准确度。接着，将预测结果作为深度的置信度保存起来。后续参考帧在更新关键帧的深度时，将预测出的置信度作为一个权值，用它来对原来的深度信息和当前参考帧计算出的深度信息进行线性的融合，即使置信度高的深度估计在线性融合过程中占较大的权值。

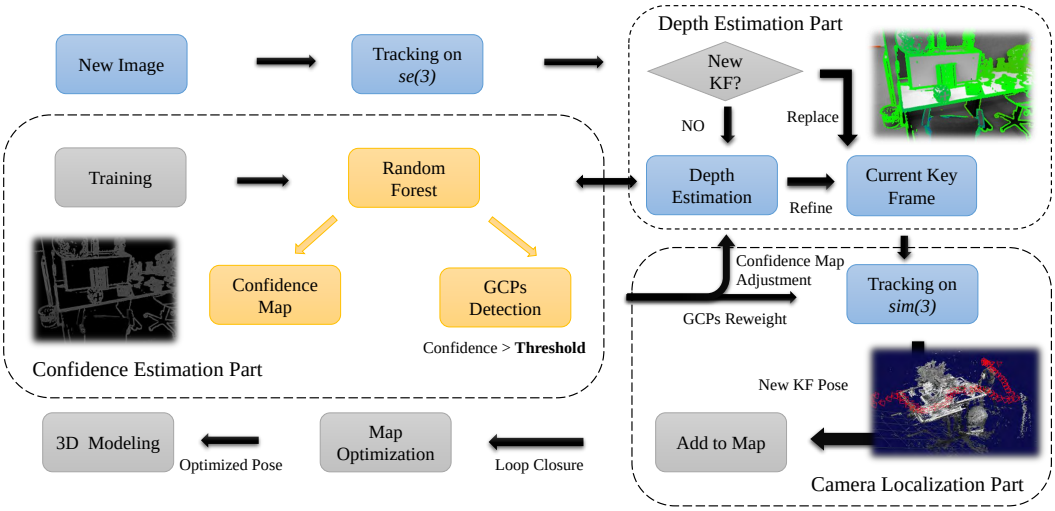


图 4-1 基于可信控制点（GCPs）的SLAM算法流程图

同时，用同样的方法融合原来的置信度和新计算出的置信度信息。在相机姿态估计的过程中，本算法增加了拥有较高置信度的像素点的权重，由此到达更好的姿态估计结果。最终使得重建的效果更加的鲁棒。

4.3 基于高阶特征的SLAM算法

前端里程计对相机运动的估计会不可避免的产生误差，即使通过上节的可信控制点模型，也只能找到估计相对准确的点，并不能得到理想的运动估计和深度估计结果。后端的非线性优化只能通过前面的状态和加入的噪声，得到一个最优化的估计结果，并不能真正的修正误差。这样持续下去就会因为误差的累积，导致估计的结果越来越不准确，最终导致重建结果的漂移。修正这种问题的一种方法就是闭环检测。

闭环检测是检测相机是否到过当前地点的模块，如果发现当前帧所在位置之前已经重建过，就将当前估计到的信息和之前的信息做融合，既能减少累计误差带来的漂移，又能进一步改善之前的估计结果。闭环检测虽然不能完全解决运动估计不准确的问题，但是可以在一定程度上提升系统的鲁棒性。

现有的闭环检测方法都没有达到很高的精度。LSD-SLAM中使用的闭环检测算法是FAB-MAP，通过3.5.4节的实验，可以看出，FAB-MAP的准确度比较低，在三个测试数据集上都得到了较差的结果，而且FAB-MAP使用的是传统的特征，对时间变化带来的景物变化、光照变化等干扰因素不鲁棒。并不能很理想的起到闭环检测的作用。

闭环检测的影响因素有很多。当两个不同的地点相似的时候，就会产生感知混淆。如果这时候闭环检测系统错误地将两个相似的不同地点判断为同一个地点，就会造成重建结果产生更大的偏差，导致闭环检测模块对系统产生负面的影响。在3.5.4节实验中使用的Nordland数据集拍摄的都是岛上火车道沿线的自然风光，基本上都是相似的火车道和树木等景色，从实验结果可看出，传统的算法并不能很好的检测出相同的地点。光照变化是重建过程当中经常会遇到的干扰因素，光照的变化对传统特征的影响很大，因此在光照变化明显的时候，基于传统特征进行闭环检测的算法，可能并不能检测出闭环，使得闭环检测系统并不能正确的行使自己的功能，可能会产生漂移的重建结果。对比实验中的St Lucia数据集就有很强的光照变化，在这种情况下，传统的方法几乎检测不出相同的地点。还有天气的变化等因素，会使得图片较之前而言，发生较大的变化，如果单纯的使用传统的特征信息，会很难发现闭环的存在。

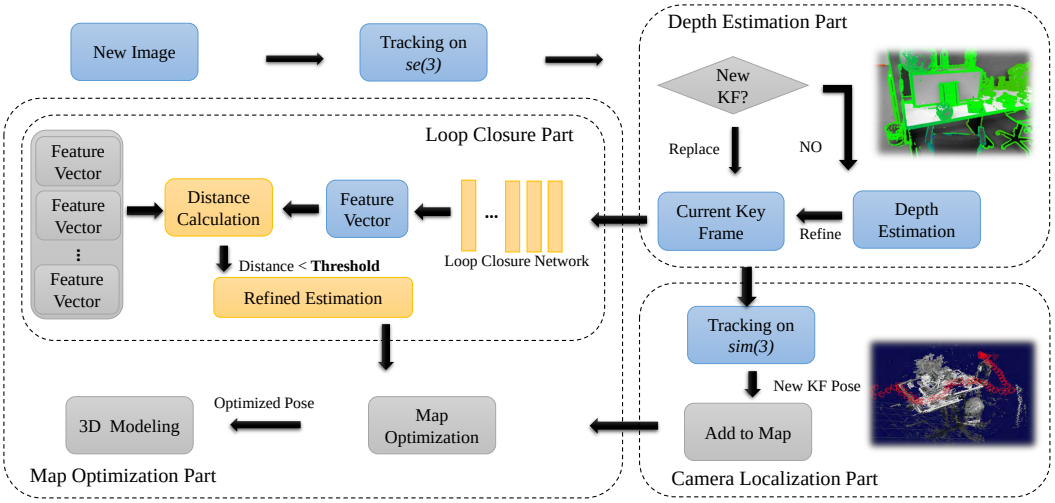


图 4-2 基于二阶特征闭环检测SLAM算法流程图

如果可以高准确性地进行闭环检测，就可以进一步地提升SLAM算法的准确性，同时减少错误的闭环检测给系统带来的副作用，进而提升重建结果。

第三章提出的闭环检测模型可以得到较高精度的闭环检测结果。模型没有使用传统的方法，使用了深度学习的方法，提取图片中的深层语义信息，使模型的检测结果受光照等干扰因素影响较小。模型还运用了在很多方向都使深度学习网络达到了更优结果的二阶的特征信息，并使用了符合高斯分布的二阶特征，更加接近图片的信息分布。在训练过程中，模型采用了弱监督学习，使同

一地点的输入样本可以更好的聚类。通过3.5.3节的实验结果可以看出，相较于同样适用深度学习进行闭环检测的NetVLAD模型，本课题提出的模型，无论是在结构上，还是在最终的测试结果上，都要更优。其中Tokyo24/7数据集包含了白天和夜晚的图片，模型依然可以达到较好的结果，说明本课题提出的模型对于明显的光照变化也具有一定的鲁棒性。通过3.5.4节的实验结果，可以对比地看出本课题模型的精度远高于LSD-SLAM使用的FAB-MAP方法。同时，模型并没有在测试数据集上进行微调，说明本课题提出的闭环检测模型具有一定的普适性。实验使用的Nordland数据集中的景物都很相似，但模型依然可以较为准确的找到相对应的图片，说明模型具有一定的抗感知混淆的能力。

将第三章提出的基于二阶特征的闭环检测模型嵌入LSD-SLAM系统当中，整体流程图如图4-2所示。用它替换原有的FAB-MAP算法。使用预先训练好的模型，并不需要在测试集上进行微调。在闭环检测环节通过关键帧的RGB图片提取模型特征，并计算与之前的特征欧氏距离。设定一定的阈值，当计算出的距离大于等于阈值的时候，说明当前的地点之前没有出现过，将特征存入备选集。当计算得到的距离小于阈值的时候，说明检测两个地点为同一个地点，将当前帧的信息融入之前的重建结果，达到高精度闭环检测的目的。

4.4 高精度的SLAM算法

近些年基于学习的方法已经应用在了各个领域，并且取得了很好的效果。

本章在LSD-SLAM算法原有的基础上，通过基于学习的方法对系统的视觉里程计和闭环检测模块做了改善。LSD-SLAM的视觉里程计模块使用的是直接法进行相机运动估计，进而估计图像中物体的深度，进行三维重建。为了改进因为使用直接法估计而带来的深度估计不准确的问题，本课题提出了一种基于学习的可信控制点预测模型。并通过实验证明了模型的可行性和有效性。

LSD-SLAM的闭环检测模块使用的是FAB-MAP算法，其准确度并不是很高。通过与本课题提出的基于二阶特征的闭环检测网络模型对比，本课题提出的模型在准确度上远高于FAB-MAP算法。通过与同样为使用深度学习方法进行闭环检测的NetVLAD模型对比，本课题提出的模型的结果也优于NetVLAD模型。

上两节分别从视觉里程计环节和闭环检测环节对LSD-SLAM算法做了改善，本节将上述两个模型都加入LSD-SLAM系统当中，得到了基于学习方法改进的高精度的SLAM算法。

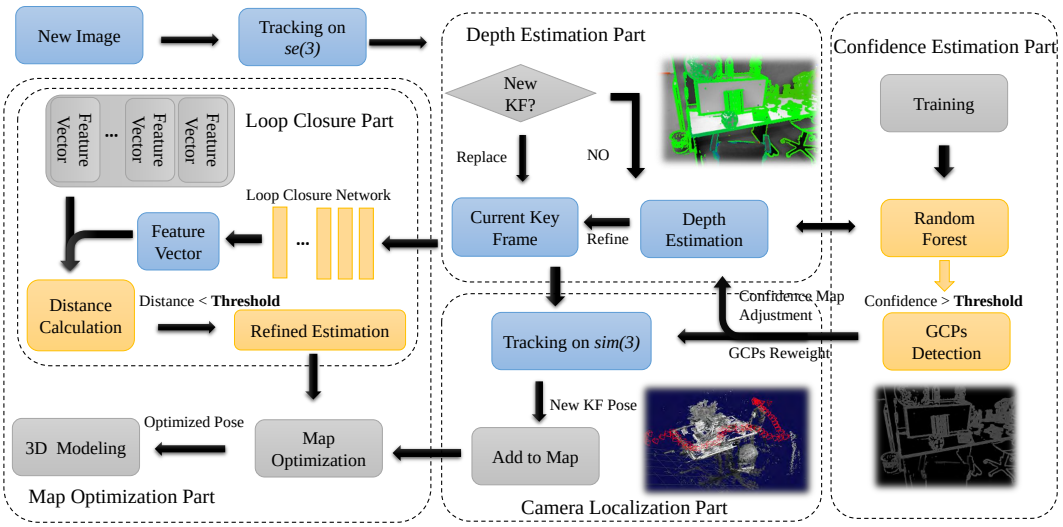


图 4-3 SLAM算法流程图

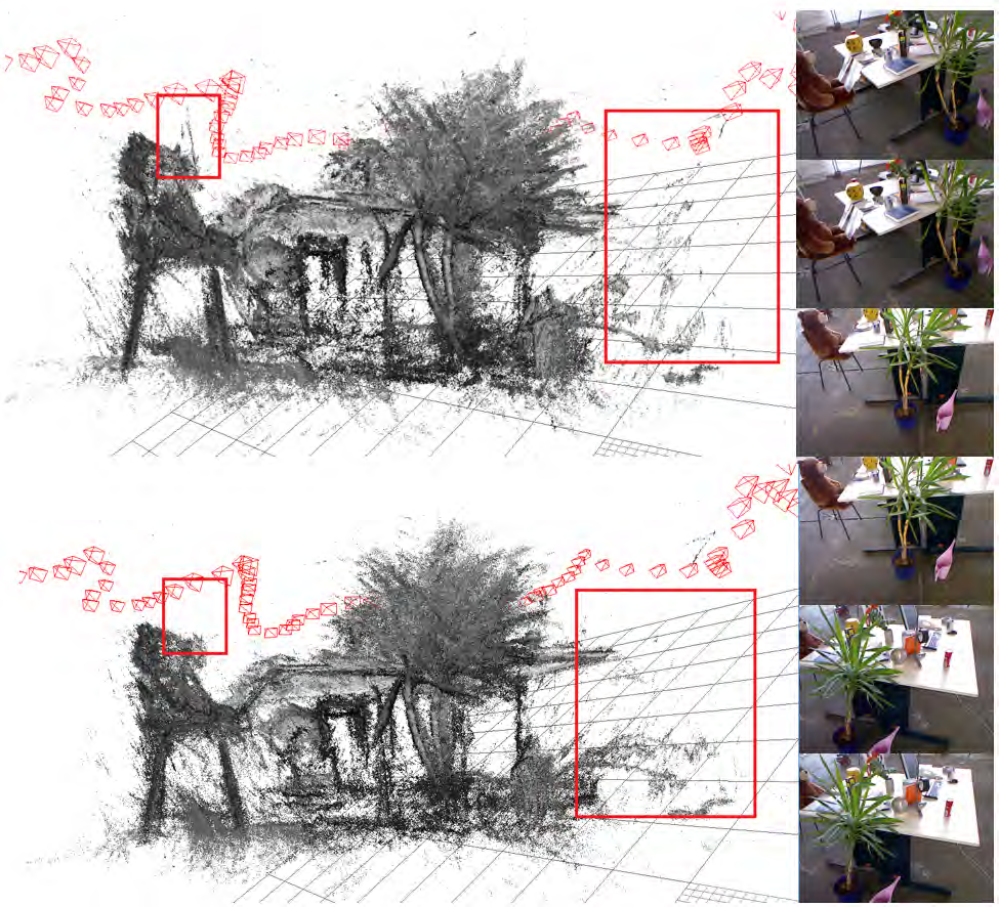


图 4-4 加入置信度估计的重建结果对比

算法的整体流程示意图如图4-3所示。在当前帧为参考帧时，计算其和当前关键帧的匹配特征，然后利用可信控制点模型预测深度估计的置信度。利用置信度信息改进深度和相机姿态的估计，使得到的深度估计更为准确。在跟踪到关键帧的时候，利用本课题提出的基于二阶特征的闭环检测模型进行闭环检测。模型可以达到较高精度的检测准确率，能及时地发现闭环的出现，使模型可以及时地融合信息，得到更为准确的重建结果。

4.5 实验结果

本节实验分为三个部分：分析讨论加入深度置信度预测模型的LSD-SLAM算法，并与原始的LSD-SLAM算法作对比。接着，对加入了基于二阶特征信息的闭环检测网络的LSD-SLAM算法进行对比分析。最后，将两个模型都加入LSD-SLAM算法中，对比原始LSD-SLAM、只加入置信度信息的LSD-SLAM、只更换闭环检测方法的LSD-SLAM和加入两种模型的LSD-SLAM的重建效果，并进行分析。

首先对加入可信控制点模型的SLAM算法做实验和分析。

使用LSD-SLAM方法计算出各个参考帧与其对应的关键帧之间的特征值，测试本课题的方法的准确性。使用的数据集为TUM RGB-D数据集。TUM RGB-D数据集中的场景主要为桌子、电脑等室内场景。该数据集中包含了连续的场景图片，包括RGB彩色图片、Kinect测量出的深度图片以及相机参数信息。首先，用LSD-SLAM方法计算出每个关键帧相对于其各个参考帧的特征值。特征中包含逆深度值，将其取倒数可得到深度值信息。将深度值与Kinect的深度测量值（真实值）的差值归一化到在[0,1]区间内。其中归一化后的值越大，表示二者差距越小。我们将阈值设定为0.7，即大于等于0.7表示计算的深度值准确，小于0.7 判定为不准确。接着，利用回归模式的随机森林算法进行训练和预测。

模型的训练集和第2.4.2节实验中选取的一样，都是随机地从*fr2_desk*，*fr1_xyz*和*fr2_xyz*中等比例的选取300,000个像素点的特征进行训练。在进行跟踪重建之前，就将模型训练好，并不需要在重建的过程中训练模型，使加入置信度估计的SLAM算法也可以达到实时地跟踪定位和重建。

首选，本实验对加入可信控制点模型的SLAM算法做直观的重建对比分析。图4-4是原始的LSD-SLAM算法和加入置信度估计的SLAM算法的重见效果的对比。重建的数据集是*fr2_desk*。图的左侧一列分别是没有加入置信度信息（上）和加入了置信度信息的（下）的重建结果。右侧一列是数据集的图片示例。从

圈出的部分可以看出，加入置信度估计的SLAM 算法的噪声更少，重建结果更接近实际物体，重建效果要明显好于原始的LSD-SLAM算法。

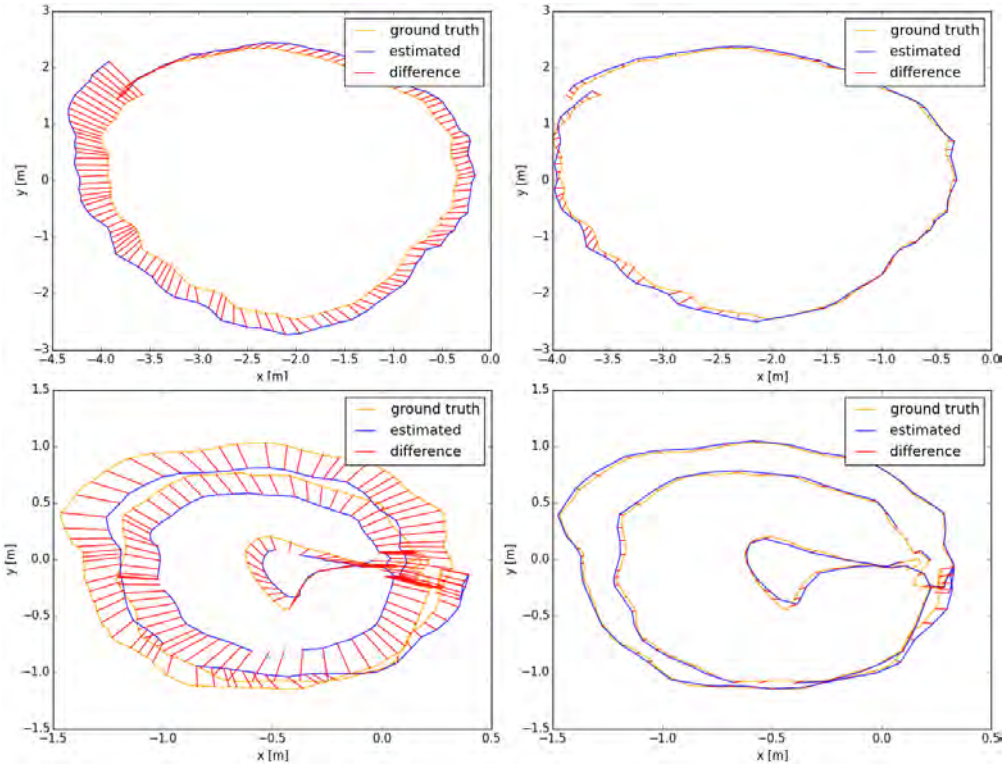


图 4-5 跟踪估计误差度

如图4-5所示是算法的跟踪轨迹误差图。本实验分别在*fr2_dishes*和*fr3_nostructure_texture_near*两个数据集上进行测试。其中黄色的线是真实的轨迹，蓝色的线是算法估计出的轨迹，红色的线代表前两者的差距。左侧一列是LSD-SLAM 的轨迹估计结果和误差，右侧一列是加入了深度置信度预测模型的跟踪轨迹估计结果和误差。从图中可以看出，加入了置信度信息之后，轨迹估计的误差在两个数据集上都有了明显的下降。

表4-1是各种模型在TMU-RGBD数据集上的轨迹跟踪均方根误差（RMSE）结果。其中的“-”表示该方法无法完成对应数据集的重建，结果不存在。加粗的是每个数据集在LSD-SLAM框架的方法中的最好结果。实验分别对比了本文致力于改进的LSD-SLAM、DVO-SLAM、RGBD-SLAM 和本文提出的GCP-SLAM。在实验中，分别对比了只将深度置信度信息融入深度估计、只将置信度融入相机姿态估计和将置信度即用于深度估计又用于相机姿态估计的结果。从实验结果可以看出，本文提出的加入基于可信控制点的深度置信度预测模型的SLAM 算法，在结果上要优于LSD-SLAM。

表 4-1 轨迹跟踪RMSE(cm)误差

数据集	GCP-SLAM			LSD-SLAM	DVO-SLAM	RGBD-SLAM
	depth	sim(3)	depth&sim(3)			
fr1_xyz	1.6	4.0	**1.5**	6.0	1.16	1.34
fr1_desk	33.4	31.0	**30.1**	39.2	2.10	2.58
fr1_floor	39.4	31.9	**27.1**	34.2	5.50	9.00
fr2_xyz	1.11	**0.84**	0.86	1.23	1.18	2.61
fr2_deskwp	6.16	4.65	**4.43**	31.73	-	6.85
fr3_sitxyz	6.03	6.53	**6.01**	6.94	-	-
fr3_longoff	35.2	30.4	**28.5**	36.9	3.50	-

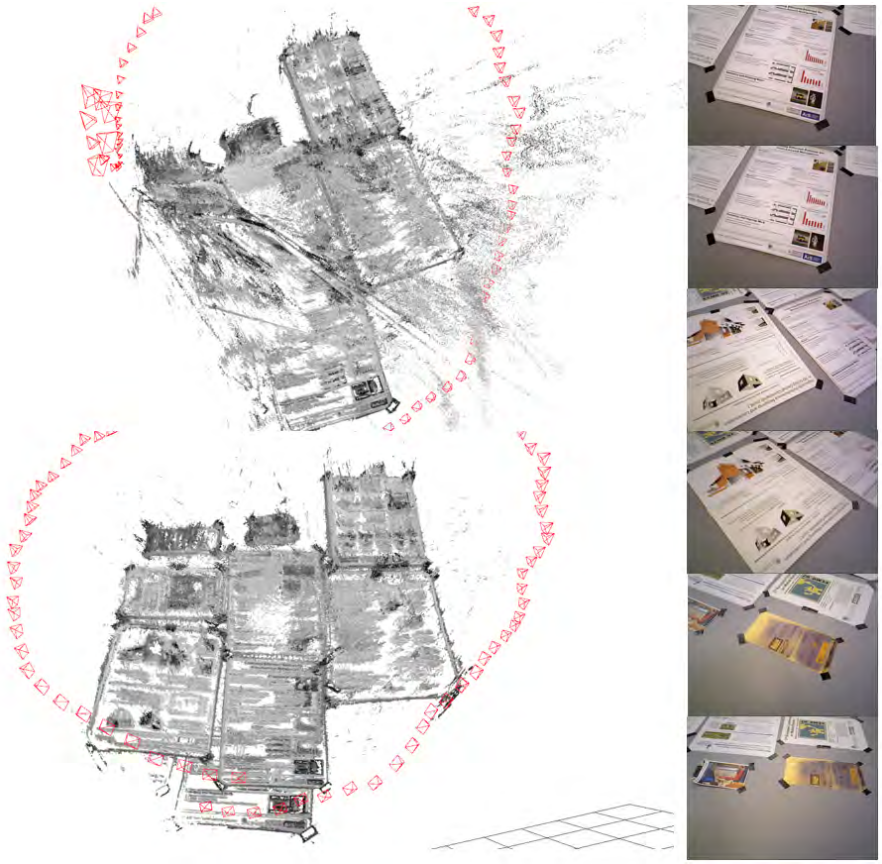


图 4-6 不同闭环检测方法的重建结果对比

接下来分析加入基于二阶特征的闭环检测模型的SLAM算法。

模型使用的是VGG16+MPN-COV的结构，并利用TokyoTM模型进行微调，使模型更适用于闭环检测问题。每当定义一个关键帧时，利用网络模型提取关键帧的特征，并将特征与之前相机经过地点的特征作对比，当两个特征之间的距离低于阈值的时候，就判断发生了闭环。将之前的信息与当前的深度信息做以融合，从而使得重建结果更加准确。

实验对比了分别使用本文提出模型和FAB-MAP算法进行闭环检测的方法。重建结果对比如图4-6所示。实验使用的数据集是*fr3_nostructure_texture_near_withloop*。图的左边一列是重建结果对比，上边的是使用FAB-MAP作为闭环检测方法的重建结果，下面是使用本文提出的闭环检测网络模型作为闭环检测方法的重建结果。图的右侧一列是数据集的图片示例，数据集的拍摄内容是贴在地上的数张海报，海报本身非常相似，容易产生感知混淆。图4-7是两种方法的轨迹误差，标记方法与图4-5相同。结果显示，本文提出的方法显著地减小了轨迹跟踪误差。对比重建结果也可以看出，FAB-MAP 方法受到感知混淆的影响，检测出了错误的闭环，导致重建结果出现了较大程度的漂移。而本文提出的基于高阶特征的闭环检测模型可以较好的避免感知混淆，更为准确的重建出了三维模型。

将可信控制点模型和基于二阶特征的闭环检测模型都加入LSD-SLAM中，模型的设置均不作改变。得到了如图4-8所示的重建结果。其中图4-8 a)是原始的LSD-SLAM的重建结果，从结果中可以看出，由于感知混淆的产生，重建结果产生了很大的误差。图4-8 b)是加入深度置信度估计模型的LSD-SLAM算法，加入置信度信息之后，重建的效果得到了明显的提升，但是红色方框的部分仍然产生了很大的重建误差。图4-8 c)是将FAB-MAP算法替换成本文提出的闭环检测模型后得到的结果。图4-8 d)是加入了本文提出的两种模型之后的重建结果，对比图4-8 c)可以看出，在红色框的部分，加入两种模型的重建结果要更准确，重建的噪声更小。因此，加入基于学习方法的SLAM系统的重建准确度要更高，重建的噪声也更少。

如图4-9所示是四种方法的轨迹误差图。其中图4-9 a)是使用原始的LSD-SLAM方法得到的轨迹误差，图4-9 b)是加入置信度估计的LSD-SLAM算法的轨迹误差，图4-9 c)是将闭环检测模块替换为本文提出的闭环检测模型的LSD-SLAM算法的轨迹误差，图4-9 d)是加入了本文提出的两种模型的LSD-SLAM算法的轨迹误差。从图中可以看出，两种方法都可以提高算法的重建精度，并且模型之间不冲突，效果可以叠加，得到更加精准的重建结果。

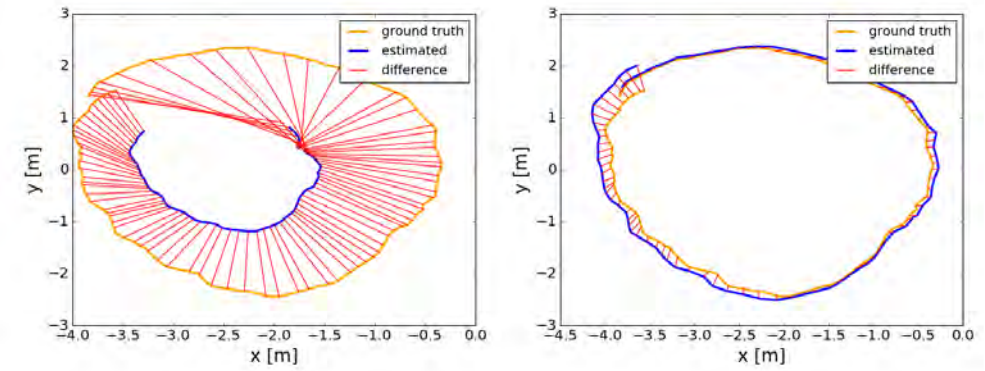
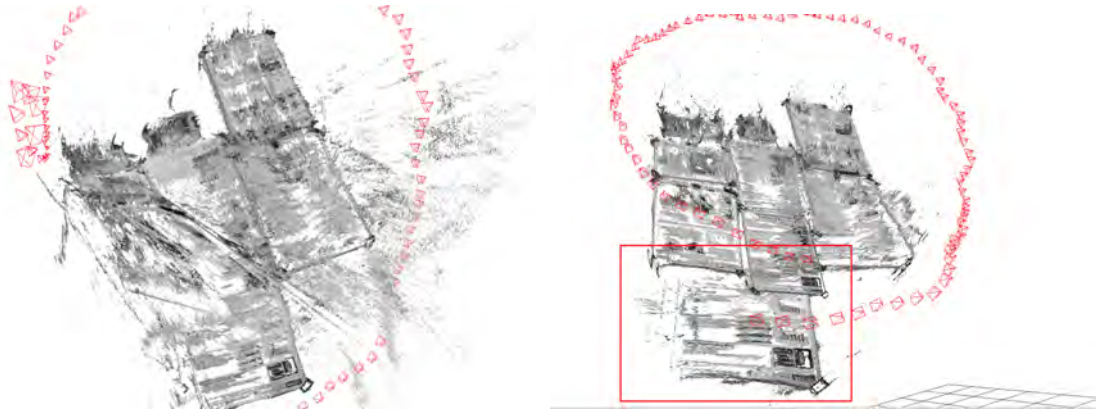
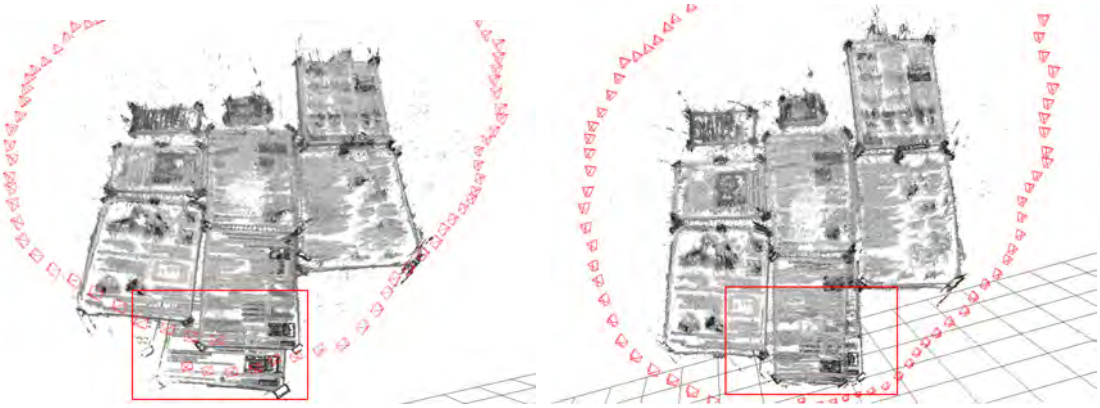


图 4-7 不同闭环检测方法的轨迹误差对比



a) LSD-SLAM　　b) 加入置信度模型的LSD-SLAM

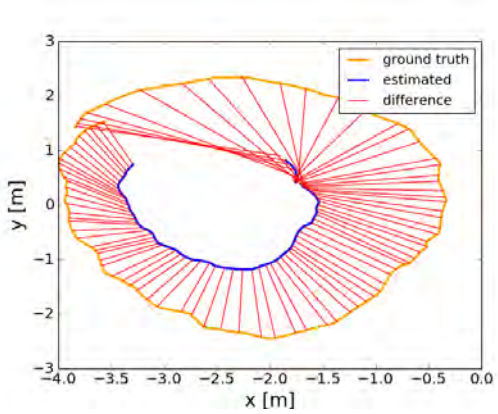


c) 加入闭环检测模型的LSD-SLAM　　d) 加入两种模型的LSD-SLAM

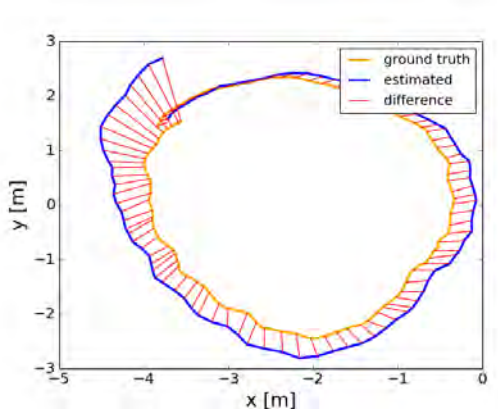
图 4-8 重建结果对比

4.6 本章小结

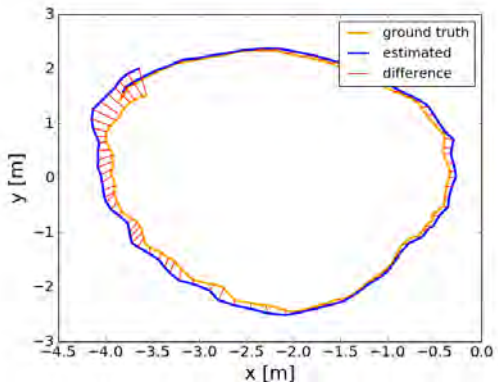
本章综合前两章提出的模型，并利用其改进了LSD-SLAM算法。使模型的效果不仅仅存在在理论中，也可以正确有效得利用到SLAM系统当中。本章将第二章提出的基于可信控制点的深度置信度预测模型应用在深度估计和相机姿态估计当中，得到了更加准确的重建结果。将第三章提出的闭环检测网络模型应用到了LSD-SLAM当中，使重建结果的漂移现象进一步的得到了减小。本章利用以上提到的模型对LSD-SLAM的视觉里程计环节和闭环检测环节做了改进，在保留LSD-SLAM优势的情况下，使重建效果得到了提升。



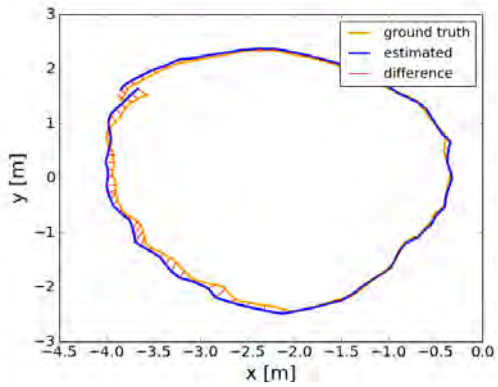
a) LSD-SLAM



b) 加入置信度模型的LSD-SLAM



c) 加入闭环检测模型的LSD-SLAM



d) 加入两种模型的LSD-SLAM

图 4-9 重建结果对比

结　论

本课题使用基于学习的方法对SLAM中的视觉里程计和闭环检测两个环节进行了提升和改进，并将改进的算法模型应用到了LSD-SLAM当中，提升了LSD-SLAM的重建精度和鲁棒性。

本课题的主要研究内容和贡献主要分为以三个部分：

(1) 本课题提出了一种基于可信控制点的深度置信度预测模型。模型利用立体匹配过程中提取到的特征，利用随机森林算法训练一个深度置信度的预测模型。置信度高的点就是深度估计更为准确的点。本课题讨论比较了使用不同的机器学习算法训练可信控制点模型的效果，通过实验证明本课题选用的随机森林算法在时间效率和准确度上是比较方法中最优的。

(2) 本课题提出了一个基于二阶特征的闭环检测网络。本课题将在计算机视觉领域取得了不错成绩的二阶特征，加入到闭环检测的问题中去，并取得了不错的效果。在训练方面，本课题采用了三元组的损失函数，对网络进行弱监督训练。图像的特征在训练过程中不断地进行聚类操作，使相同地点的特征不断聚集，不同地点的特征尽量远。通过实验对比验证了与其他基于深度学习方法的模型相比，本课题提出的模型达到了更高的精度。

(3) 本课题将以上两点提出的模型加入到LSD-SLAM算法当中，在SLAM跟踪重建的过程中，使用提出的可信控制点模型预测参考值估计出深度的置信度。利用这个置信度信息，对估计出的信息进行加权。使置信度高的深度的估计在深度的融合中占较大的权重，达到改进深度估计的效果。在SLAM的闭环检测环节，将提出的基于二阶特征的闭环检测模型加入到系统当中，是系统在各种特殊情况下都能得到较为准确的闭环检测解。

在今后的研究中，对于深度置信度估计模型，可以尝试更多的机器学习方法训练模型，或采用深度学习的方法直接预测图像中物体的深度。对于闭环检测网络，可以尝试加入更高阶的特征。已经有了使用二阶以上特征进行其他方面研究的成果，并得到了不错的效果。可以尝试将二阶以上的特征加入闭环检测网络，争取得到更鲁棒的结果。

参考文献

[1] Triggs B, McLauchlan P F, Hartley R I, Fitzgibbon A W. Bundle adjustment—a modern synthesis[C] // International workshop on vision algorithms. 1999 : 298 – 372.

[2] Kümmerle R, Grisetti G, Strasdat H, Konolige K, Burgard W. g 2 o: A general framework for graph optimization[C] // Robotics and Automation (ICRA), 2011 IEEE International Conference on. 2011 : 3607 – 3613.

[3] Mur-Artal R, Montiel J M M, Tardos J D. ORB-SLAM: a versatile and accurate monocular SLAM system[J]. IEEE Transactions on Robotics, 2015, 31(5) : 1147 – 1163.

[4] Rublee E, Rabaud V, Konolige K, Bradski G. ORB: An efficient alternative to SIFT or SURF[C] // Computer Vision (ICCV), 2011 IEEE international conference on. 2011 : 2564 – 2571.

[5] Engel J, Schöps T, Cremers D. LSD-SLAM: Large-scale direct monocular SLAM[C] // European Conference on Computer Vision. 2014 : 834 – 849.

[6] Engel J, Sturm J, Cremers D. Semi-dense visual odometry for a monocular camera[C] // Proceedings of the IEEE international conference on computer vision. 2013 : 1449 – 1456.

[7] Stumm E S, Mei C, Lacroix S. Building location models for visual place recognition[J]. The International Journal of Robotics Research, 2016, 35(4) : 334 – 356.

[8] Zhang H, Reardon C, Parker L E. Real-time multiple human perception with color-depth cameras on a mobile robot[J]. IEEE Transactions on Cybernetics, 2013, 43(5) : 1429 – 1441.

[9] Latif Y, Huang G, Leonard J J, Neira J. An Online Sparsity-Cognizant Loop-Closure Algorithm for Visual Navigation.[C] // Robotics: Science and Systems. 2014.

[10] Arroyo R, Alcantarilla P F, Bergasa L M, Romera E. Towards life-long visual localization using an efficient matching of binary sequences from images[C] // Robotics and Automation (ICRA), 2015 IEEE International Conference on. 2015 : 6328 – 6335.

[11] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos[C] // null. 2003 : 1470.

[12] Cummins M, Newman P. FAB-MAP: Probabilistic localization and mapping in the space of appearance[J]. The International Journal of Robotics Research, 2008, 27(6) : 647 – 665.

[13] Cummins M, Newman P. Highly scalable appearance-only SLAM-FAB-MAP 2.0.[C] // Robotics: Science and Systems : Vol 5. 2009 : 17.

[14] Chow C, Liu C. Approximating discrete probability distributions with dependence trees[J]. IEEE transactions on Information Theory, 1968, 14(3) : 462 – 467.

[15] Zhang H, Han F, Wang H. Robust Multimodal Sequence-Based Loop Closure Detection via Structured Sparsity.[C] // Robotics: Science and Systems. 2016.

[16] Arandjelovic R, Gronat P, Torii A, Pajdla T, Sivic J. NetVLAD: CNN architecture for weakly supervised place recognition[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016 : 5297 – 5307.

[17] Jégou H, Douze M, Schmid C, Pérez P. Aggregating local descriptors into a compact image representation[C] // Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. 2010 : 3304 – 3311.

[18] Scharstein D, Szeliski R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms[J]. International journal of computer vision, 2002, 47(1-3) : 7 – 42.

[19] Shi C, Wang G, Yin X, Pei X, He B, Lin X. High-accuracy stereo matching based on adaptive ground control points.[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2015, 24(4) : 1412 – 23.

[20] Hu X, Mordohai P. A quantitative evaluation of confidence measures for stereo vision[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 34(11) : 2121 – 2133.

[21] Spyropoulos A, Komodakis N, Mordohai P. Learning to detect ground control points for improving the accuracy of stereo matching[C] // Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. 2014 : 1621 – 1628.

[22] Breiman L. Random forests[J]. Machine learning, 2001, 45(1) : 5 – 32.

[23] Scharstein D, Pal C. Learning conditional random fields for stereo[C] // Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. 2007 : 1 – 8.

[24] Sturm J, Engelhard N, Endres F, Burgard W, Cremers D. A benchmark for the evaluation of RGB-D SLAM systems[C] // Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on. 2012 : 573 – 580.

[25] Cimpoi M, Maji S, Vedaldi A. Deep filter banks for texture recognition and segmentation[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2015 : 3828 – 3836.

[26] Wang Q, Li P, Zuo W, Zhang L. RAID-G: Robust Estimation of Approximate Infinite Dimensional Gaussian with Application to Material Recognition[C] // Computer Vision and Pattern Recognition. 2016 : 4433 – 4441.

[27] Tang P, Wang X, Shi B, Bai X, Liu W, Tu Z. Deep FisherNet for Object Classification[J], 2016.

[28] Ionescu C, Vantzos O, Sminchisescu C. Matrix Backpropagation for Deep Networks with Structured Layers[C] // IEEE International Conference on Computer Vision. 2015 : 2965 – 2973.

[29] Lin T-Y, RoyChowdhury A, Maji S. Bilinear cnn models for fine-grained visual recognition[C] // Proceedings of the IEEE International Conference on Computer Vision. 2015 : 1449 – 1457.

[30] Li P, Xie J, Wang Q, Zuo W. Is Second-order Information Helpful for Large-scale Visual Recognition?[J], 2017.

[31] Dryden I L, Koloydenko A, Zhou D. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging[J]. Annals of Applied Statistics, 2009, 3(3) : 1102 – 1123.

[32] Arsigny V, Fillard P, Pennec X, Ayache N. Geometric means in a novel vector space structure on symmetric positive-definite matrices[J]. SIAM journal on matrix analysis and applications, 2007, 29(1) : 328 – 347.

[33] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015 : 815 – 823.

[34] Hermans A, Beyer L, Leibe B. In Defense of the Triplet Loss for Person Re-Identification[J], 2017.

[35] Torii A, Arandjelović R, Sivic J, Okutomi M, Pajdla T. 24/7 place recognition by view synthesis[C] // CVPR. 2015.

[36] Glover A J, Maddern W P, Milford M J, Wyeth G F. FAB-MAP+ RatSLAM: Appearance-based SLAM for multiple times of day[C] // Robotics and Automation (ICRA), 2010 IEEE International Conference on. 2010 : 3507 – 3512.

[37] Sünderhauf N, Neubert P, Protzel P. Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons[C] // Proc. of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA). 2013 : 2013.

[38] Badino H, Huber D, Kanade T. Real-time topometric localization[C] // Robotics and Automation (ICRA), 2012 IEEE International Conference on. 2012 : 1635 – 1642.

[39] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.

[40] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C] // International Conference on Neural Information Processing Systems. 2012 : 1097 – 1105.

[41] Glover A, Maddern W, Warren M, Reid S, Milford M, Wyeth G. OpenFABMAP: An open source toolbox for appearance-based loop closure detection[C] // Robotics and automation (ICRA), 2012 IEEE international conference on. 2012 : 4730 – 4735.

[42] Natarajan B K. Sparse approximate solutions to linear systems[J]. SIAM journal on computing, 1995, 24(2) : 227 – 234.

[43] Çetin M, Karl W C. Feature-enhanced synthetic aperture radar image formation based on nonquadratic regularization[J]. IEEE Transactions on Image Processing, 2001, 10(4) : 623 – 631.

[44] Chartrand R. Exact reconstruction of sparse signals via nonconvex minimization[J]. IEEE Signal Processing Letters, 2007, 14(10) : 707 – 710.

[45] Zuo W, Meng D, Zhang L, Feng X, Zhang D. A generalized iterated shrinkage algorithm for non-convex sparse coding[C] // Proceedings of the IEEE international conference on computer vision. 2013 : 217 – 224.

[46] Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X. Residual Attention Network for Image Classification[J]. arXiv preprint arXiv:1704.06904, 2017.

[47] Wang J, Song Y, Leung T, Rosenberg C, Wang J, Philbin J, Chen B, Wu Y. Learning fine-grained image similarity with deep ranking[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014 : 1386 – 1393.

[48] Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification[J]. Journal of Machine Learning Research, 2009, 10(Feb) : 207 – 244.

[49] Schultz M, Joachims T. Learning a distance metric from relative comparisons[C] // Advances in neural information processing systems. 2004 : 41 – 48.

[50] Wang Q, Li P, Zhang L. G2DeNet: Global Gaussian Distribution Embedding Network and Its Application to Visual Recognition[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2017 : 6507 – 6516.

[51] Torii A, Sivic J, Pajdla T, Okutomi M. Visual place recognition with repetitive structures[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2013 : 883 – 890.

攻读硕士学位期间发表的论文及其他成果

（一）发表的学术论文

[1] Feng A., Zhang W., Yan Z., Zuo W. GCP-SLAM: LSD-SLAM with Learning-Based Confidence Estimation[C]. PSIVT 2017: Image and Video Technology 262-275.（EI 收录号：20180904845043）

哈尔滨工业大学学位论文原创性声明及使用授权说明

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《基于学习方法的高精度SLAM算法研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：冯爱迪　　日期：2018 年 06月 24日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：冯爱迪　　日期：2018 年 06月 24日

导师签名：张大朋　　日期：2018 年 06月 24日

致　谢

两年的研究生生活转瞬即逝，转眼间已经在哈尔滨工业大学已经度过了六年的时光，度过了我人生的四分之一。我的母校见证了我人生中最为重要的一个阶段，见证了我的成长。感谢母校有着严谨又自由的学术氛围，让我没有虚度这六年的时光。感谢同样陪我度过了六年的计算机科学与技术学院，感谢各科老师的悉心指导和关怀，使我六年以来都没有后悔过选择计算机专业，并想在可以预见到的未来，继续从事计算机相关的工作。祝母校越来越好，为国家培养出更多优秀的人才。也祝计算机科学与技术学院能有更多的成就，更好的科研环境，培养出更过专业过硬的人才。

感谢我的导师张大鹏老师，我的主要指导老师左旺孟老师。左老师治学严谨，几乎将自己的全部时间都投身科研事业，是我学习的榜样。左老师总是能及时发现我研究过程中存在的问题，并总能提出新颖可行的解决方案。感谢左老师两年以来对我的宽容，即使我有做的不够好的地方，也没有很严厉的批评过我。回首过去两年，觉得自己并没有付出最大程度上的努力，其实还应该做的更好。感谢张宏志老师两年以来对我的照顾，张老师总是可以把实验室的大事小情安排的合理且周到，让我们能在课题组更好的科研和生活，同时也让我们学到了严谨细致的做事方式和态度，相信一定会受益终身。感谢闫子飞老师，闫老师总是能将实验的琐事处理得很好，让我们能用最少的时间解决其他事情，将更多的时间投身科研。闫老师总是细心地、不耐其烦地为我们修改论文，发现论文中的每一个问题，认真帮我们斟酌语句，让我们的科研工作更加顺利。

感谢感知计算中心的王宽全老师、邬向前老师、马琳老师和袁永峰老师在开题和中期答辩上给我提出的宝贵意见，让我能及时发现自己的误区，更加顺利地进行研究工作。感谢实验室的师兄师姐师弟师妹对我的帮助。感谢和我一起进行课题研究的张玮奇师姐，从师姐那里得到了很多的帮助，也从师姐身上学到了很多研究的技巧和方法。感谢和我同级的小伙伴的陪伴，能有你们陪伴着走过这两年是我的幸运。

感谢我的父母的养育之恩，作为子女我还有很多不足的地方。这六年来也没有多少时间陪伴他们。感谢两年以来和我一起成长的同学，我的室友许家欢、叶丽华、杨婧，感谢你们的包容和陪伴；我的朋友王文茹、孙震，感谢你们的关心和帮助。感谢这两年一起学习的所有同学。

感谢两年来带给了我很多力量的歌手们，在我感觉到疲惫或者想要放弃的时候，他们的作品总能给我力量，让我能继续坚持下去。

最后，感觉两年时间成长了很多的我自己。两年的时间虽然不长，但明显感到自己的思想成熟了很多，生活态度也变得更加的积极。

从课题开始到结束，所有对我提供过帮助的老师和同学，在这里真诚地向你们表示感谢。