ORB-SLAM：精确多功能单目 SLAM 系统

Contents

ORB-SLAM：精确多功能单目 SLAM 系统

ORB-SLAM: a Versatile and Accurate Monocular SLAM System

Taylor Guo, 2016 年 3 月 18 日-9:00

原文发表于 《IEEE Transactions on Robotics》 - 2015

摘要：

本文主要讲了 ORB-SLAM，一个基于特征识别的单目 slam 系统，可以实时运行，适用于各种场合，室内的或者室外的，大场景或小场景。系统具有很强的鲁棒性，可以很好地处理剧烈运动图像、可以有比较大的余地自由处理闭环控制、重定位、甚至全自动位置初始化。基于近年来的优秀算法，我们对系统做了精简，采用了所有 SLAM 相同功能：追踪，地图构建，重定位和闭环控制。选用了比较适合的策略，地图重构的方法采用云点和关键帧技术，具有很好的鲁棒性，生成了精简的、可追踪的地图，当场景的内容改变时，地图构建可持续工作。我们用最流行的图像数据集测试了 27 个图像序列。相比最新的单目 SLAM，ORB SLAM 性能优势明显。我们在网站上公布了源代码。

一 简介

由于比较强的匹配网络和初始化位置估计，BA 广泛应用于相机位置的准确估计和离散几何重构。在一段比较长的时间里，这种方法被认为不适合实时图像系统，比如 vSLAM。vSLAM 系统在构建环境的同时需要估计相机的轨迹。基于现有的硬件设备，现在可以获得比较好的计算结果，将 BA 应用于实时 SLAM 系统中：

- 在候选图像帧子集中（关键帧）匹配观测的场景特征（地图云点）。

- 由于关键帧数量的增长，需要做筛选避免冗余。
- 关键帧和云点的网络配置可以产生精确的结果，也就是，分布良好的关键帧集合和有明显视差、大量闭环匹配的观测云点。
- 关键帧和云点位置的初始估计，采用非线性优化的方法。
- 在构建局部地图的过程中，优化的关键是获得良好的稳定性。
- 本系统可以实时执行快速全局优化（比如位置地图）闭环回路。

B A 出现于 PTAM 中，第一次实时应用是视觉里程。尽管受制于小场景的应用，这个算法对关键帧的选择，特征匹配，云点三角化，每帧相机位置估计，追踪失败后的重定位非常有效。不幸的是几个关键因素限制了它的应用：缺少闭环控制和足够的阻塞处理，较差的视图不变特性和在形成地图过程中需要人工干预。

为了完成这些工作，我们采用的技术来源于，PTAM、place recognization、scale-aware loop closing 和大场景的视图关联信息。

单目 ORB SLAM 系统包含：

- 对所有的任务采用相同的特征，追踪、地图构建、重定位和闭环控制。这使得我们的系统更有效率、简单可靠。ORB 特征，在没有 GPU 的情况下可以应用于实时图像系统中，具有很好的旋转不变特性。
- 可应用于实时户外环境操作。由于其视图内容关联的特性，追踪和地图构建可在局部视图关联中处理，这可以独立于全局视图进行工作。
- 基于位置优化的实时闭环控制，我们称作 Essential Graph。它通过生成树构建，生成树由系统、闭环控制链接和视图内容关联强边缘进行维护。

- 实时相机重定位具有明显的旋转不变特性。这就使得跟踪丢失可以重做，地图也可以重复使用。
- 选择不同的模型可以创建不同的平面或者非平面的初始化地图；自动的、具有良好鲁棒性的初始化过程也是基于模型而选择。
- 大量地图云点和关键帧，需要经过严格的挑选，必须找到一个最合适的办法。好的挑选方法可以增强追踪的鲁棒性，同时去除冗余的关键帧 以增强程序的可操作性。

我们在公共数据集上对程序在室内和室外环境进行了评估 ,包括手持设备、汽车和机器人。相机的位置比现在最新的方法更精确，它通过像素扩展集进行优化、而不是特征的重映射。我们还讨论了提高基于特征的方法的准确性的原因。

闭环控制和重定位的方法是基于我们之前的工作论文 11。系统最初的版本是论文 12。本文中我们添加了初始化的方法 ,Essential graph 和其他方法。我们详细了描述了系统的各个板块，并且做了实验进行验证。就我们所知，这是目前最完整最可靠的单目 SLAM 系统。视频演示和源代码放在我们的项目网站上。

二 相关工作

A. 位置识别

论文 13 比较几种基于图像处理技术的位置识别的方法，其中有图像到图像的匹配，在大环境下比地图到地图或图像到地图方法尺度特性更好。图像方法中，词袋模型的效率更高，比如基于概率论的 FAB-MAP。由 BRIEF 特征描述子和 FAST 特征检测产生的二进制词袋可以用 DBoW2 获得。与 SURF 和 SIFT 相比，它的特征提取运算时间减小一个数量级。尽管系统运行效率高、鲁棒性好，采用 BRIEF 不具有旋转不变性和尺度不变性，系统只能运行在平面轨迹中，闭环检测也只能从相似的视角中获得。在我们之前的工作中，我们用 DBoW2 生成了基于 ORB 的词袋模型位置识别器。ORB 是具有旋转不变和尺度不变特性的二进制特征，它是一种高效的具有良好

针对视图不变的识别器。我们在 4 组不同的数据集上演示了位置识别功能，复用性好，鲁棒性强，从 10K 图像数据库中提取一个候选闭合回路的运算时间少于 39 毫秒。在我们的工作中，我们提出了一种改进版本的位置识别方法，采用内容相关的视图，检索数据库时返回几个前提而不是最好的匹配。

B. 地图初始化

单目 SLAM 通过图像序列生成初始化地图，单一图像并不能重建深度图。解决这个问题的一种方法是一开始跟踪一个已知的图像结构。在滤波方法中，用概率论方法从逆深度参数方法得到深度图中，初始化地图云点，与真实的位置信息融合。论文 10 中，采用类似的方法初始化像素的深度信息得到随机数值。

通过两个局部平面场景视图进行初始化的方法，从两个相关相机（视图）位姿进行 3D 重构，相机的位姿关系用单映射表示，或者计算一个基本矩阵，通过 5 点算法构建平面模型或者一般场景模型。两种方法都不会受到低视差的约束，平面上的所有的点也不需要靠近相机中心。另外，非平面场景可以通过线性 8 点算法来计算基本矩阵，相机的位姿也可以重构。

第四章详细讲述了一个全新的自动方法，它基于平面单映射或非平面的基本矩阵。模型选择的统计方法如论文 28 详细描述。基于相似变换理论，我们开发了初始化算法，选择退化二次曲线例子中基本矩阵，或单映射矩阵。在平面例子中，为了程序稳定性，如果选择情况模糊不清，我们尽量避免做初始化，否则方案可能崩溃。我们会延迟初始化过程，直到所选的方案产生明显的视差。

C. Monocular SLAM （单目 SLAM）

Mono-SLAM 可以通过滤波方案初始化。在这种方案，每一帧都通过滤波器估计地图特征位置和相机位姿，将其关联。处理连续的图像帧需要进行大量运算，线性误差会累积。由于地图

构建并不依赖于帧率，基于关键帧的方法，用筛选的关键帧估计地图，采用精确的 BA 优化。论文 31 演示了基于关键帧的地图方法比滤波器方法在相同的运算代价上更精确。

基于关键帧技术最具代表性的 SLAM 系统可能是 PTAM。它第一次将相机追踪和地图构建分开，并行计算，在小型场合，如增强现实领域非常成功。PTAM 中的地图云点通过图像区块与 FAST 角点匹配。云点适合追踪但不适合位置识别。实际上，PTAM 并不适合检测大的闭合回路，重定位基于低分辨率的关键帧小图像块，对视图不变性较差。

论文 6 展示了大场景的单目 SLAM 系统，前端用 GPU 光流，用 FAST 特征匹配和运动 BA；后端用滑动窗口 BA。闭环检测通过 7 自由度约束的相似变换位姿图优化，能够校正单目系统中的尺度偏移。我们采用这种 7 自由度的位姿图优化方法，应用到我们的 Essential Graph 方法中，如第三章 D 节里面详细描述。

论文 7 用，PTAM 的前端，通过内容相关的视图提取局部地图执行追踪。他们使用两个窗口优化后端，在内部窗口中采用 BA，在外部一个限制大小的窗口做位姿图。只有在外部窗口尺寸足够大可以包含整个闭环回路的情况下，闭环控制才能起作用。我们采用了基于内容相关的视图构建局部地图的方法，并且通过内容相关的视图构建位姿图，同时用它们设计前端和后端。另一个区别是，我们并没有用特别的特征提取方法做闭合回路检测（比如 SURF 方法），而是在相同的追踪和建图的特征上进行位置识别，获得具有鲁棒性的重定位和闭环检测。

论文 33 提出了 CD-SLAM 方法，一个非常复杂的系统，包括闭环控制，重定位，动态环境、大场景下运行。但它的地图初始化并没有讲。所以没法做精确性、鲁棒性和大场景下的测试对比。

论文 34 的视觉里程计方法使用了 ORB 特征做追踪，处理 BA 后端滑动窗口。相比之下，我们的方法更具一般性，他们没有全局重定位，闭环回路控制，地图也不能重用。他们使用了相机到地面的真实距离限制单目尺度漂移。

论文 25 与我们之前的工作论文 12 一样，也采用的相同的特征做追踪，地图构建和闭环检测。由于选择 BRIEF，系统受限于平面轨迹。从上一帧追踪云点，访问过的地图不能重用，与视觉里程计很像，因此系统不能扩展。我们在第三章 E 小节里面定性地做了比较。

论文 10 里的 LSD-SLAM，可以构建大场景的半稠密地图，特征提取并没有采用 BA 方法，而是直接方法（优化也是直接通过图像像素）。没有用 GPU 加速，构建了半稠密地图，可以运行在实时应用中，非常适合机器人应用，比其他基于特征的稀疏地图 SLAM 效果好。但它们仍然需要特征做闭环检测，相机定位的精度也明显比 PTAM 和我们的系统慢，我们将在第 8 章 B 小节演示实验结果。在 IX 章 B 小节进行讨论。

论文 22 提出了介于直接方式和基于特征的方法之间的半直接视觉里程 SVO 方法。不需要每帧都提取特征点，可以运行在较高的帧率下，在四轴飞行器上效果很好。然而，没有闭环检测，而且只使用了视角向下的摄像头。

最后，我们想讨论关键帧的选择。所有的的视觉 SLAM 用所有的云点和图像帧运行 BA 是不可行的。论文 31 在保留了尽可能多地非冗余关键帧和云点，性价比较高。PTAM 方法非常谨慎插入关键帧避免运算量增长过大。这种严格限制关键帧插入策略可能在未知地图追踪使失败。比较好的策略是在不同的场景下快速插入关键帧，然后再移除那些冗余的图像帧，避免额外的运算成本。

三　系统架构

A.　特征选择

我们系统设计的中心思想是对这些功能采用相同的特征，构建地图、追踪、位置识别、基于图像帧率的重定位和闭环回路检测。这使得我们的系统更有效率，没有必要极化特征识别的深度图，如论文 6,7 里讨论的。我们每张图像的特征提取远少于 33 毫秒，SIFT（~300ms），SURF(~300ms)，A-KAZE（~100ms）。为了获得一般性的位置识别方法，我们需要特征提取的旋转不变性，而 BRIEF 和 LDB 不具备这样的特性。

我们选择了我们选择了 ORB，它是旋转多尺度 FAST 角点检测具有 256 位特征描述子。他们计算和匹配的速度非常快，同时对视角具有旋转不变的特性。这样可以在更宽的基准线上匹配他们，增强了 BA 的精度。我们已经在论文 11 中演示了 ORB 位置识别的性能。本文的方案中也采用 ORB。

B. 三个线程：追踪、局部地图构建和闭环控制

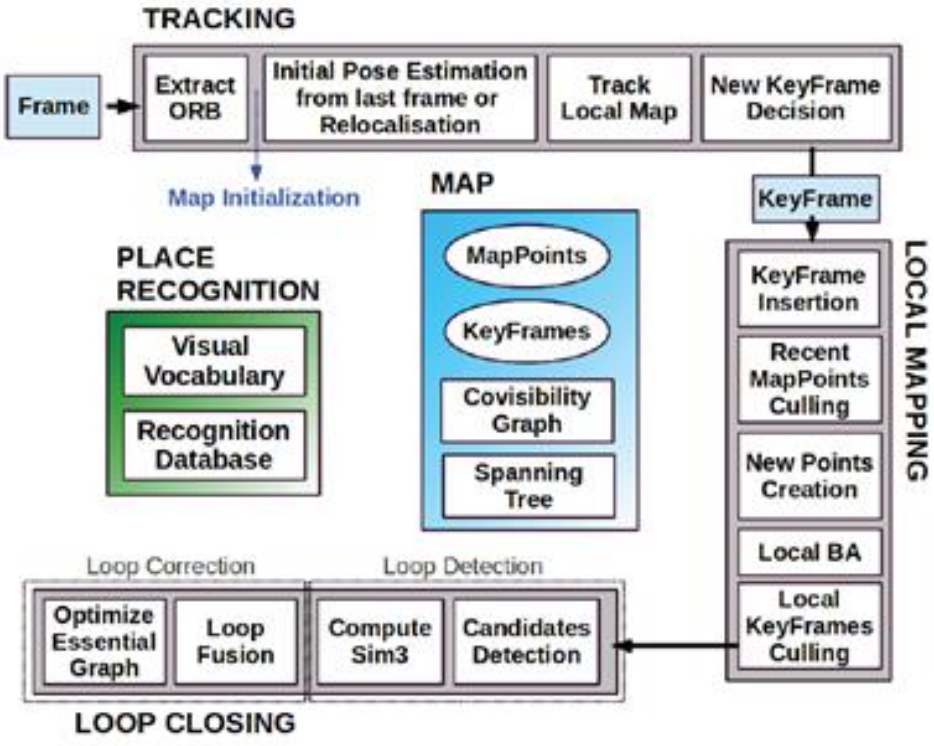
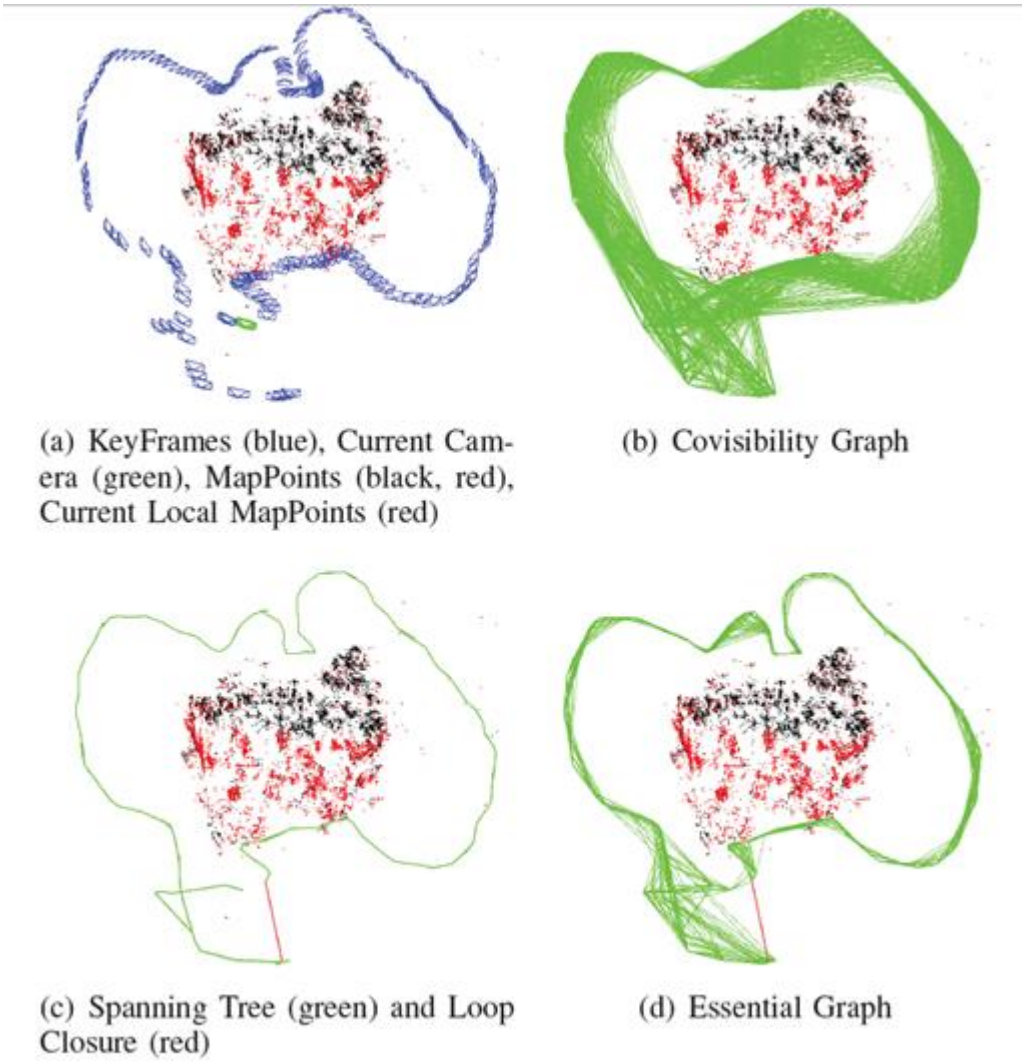


Fig. 1. ORB-SLAM system overview, showing all the steps performed by the tracking, local mapping and loop closing threads. The main components of the place recognition module and the map are also shown.

我们的系统如图一所示，整合了三个并行的线程：追踪、局部地图构建和闭环回路控制。追踪通过每帧图像定位相机位置，决定什么时候插入一个新的关键帧。我们先通过前一帧图像帧初始化特征匹配，采用运动 B A 优化位姿。如果追踪丢失，位置识别模块执行全局重定位。一旦获得最初的相机位姿估计和特征匹配，通过内容相似视图的关键帧提取一个局部可视化地图，如图 2a,b 所示。然后进行映射搜索局部地图云点的匹配，根据匹配优化相机位姿。最后，追踪线程觉得是否插入新的关键帧。所有的追踪步骤将在第 5 章详细解释。创建初始化地图将在第 4 章进行说明。



(a) KeyFrames (blue), Current Camera (green), MapPoints (black, red), Current Local MapPoints (red)

(b) Covisibility Graph

(c) Spanning Tree (green) and Loop Closure (red)

(d) Essential Graph

Fig. 2. Reconstruction and graphs in the sequence *fr3_long_office_household* from the TUM RGB-D Benchmark [38].

局部地图构建处理新的关键帧，在相机位姿的环境中执行局部 BA 优化重构。根据交互视图中已经连接的关键帧，搜索新关键帧中未匹配的 ORB 特征的对应关系，来三角化新的云点。有时尽管已经创建了新的云点，基于追踪线程过程中新收集的信息，为了获得高质量的云点，根据云点筛选策略可能会临时删除一些点。局部地图构建也负责删除冗余关键帧。我们将在第 6 章详细说明局部地图构建的步骤。

对每个新的关键帧都要进行闭环搜索，以确认是否形成闭环。如果闭环被侦测到，我们就计算相似变换来查看闭环的累积误差。这样闭环的两端就可以对齐，重复的云点就可以被融合。最后，相似约束的位姿图优化确保全局地图一致性。本文主要通过关键图像（Essential Graph）进行优化，它是一个交互视图中离散的子图像，第三章 D 小节详细描述。闭环检测和校验步骤将在第 7 章详细描述。

我们使用 g2o 库中的 Levenverg-Marquardt 算法执行全局优化。我们在附录中描述了每个优化的误差，计算成本和变量。

C. 地图云点、关键帧和选择标准

每个地图云点 p_i 保存：

- 它在世界坐标系中的 3D 位置 $X_{w,i}$
- 视图方向 n_i，也就是所有视图方向的平均单位向量（这个方向是以观察关键帧的光学中心到云点的方向）
- ORB 特征描述子 Di，已经观测到云点的关键帧与它关联的所有其他特征描述子相比，它汉明距离最小
- 根据 ORB 特征尺度不变性约束，可观测的云点的最大距离 d_{max} 和最小距离 d_{min}

每个关键帧 K_i 保存：

- 相机位姿 Tiw，从世界坐标转换成相机坐标
- 相机内参，包括主点和焦点距离
- 所有从图像帧提取的 ORB 特征，不管是否已经关联了地图云点，云点的坐标经过校准模型矫正过

地图云点和关键帧通过多重策略创建，在稍后挑选机制上将检测冗余的关键帧和错误匹配的云点或不可追踪的地图云点。系统在运行的过程中，地图扩展的弹性就比较好，在比较恶劣的环境下（比如视图旋转，快速移动）还是需要增强鲁棒性，在同一环境的重复访问情况下地图的大小是有限的，可以控制的。另外，与 PTAM 相比，由于包含的云点比较少，我们的地图包含的无效数据也更少一些。地图云点和关键帧的筛选过程将在第 6 章 B 节和 E 节分别解释。

D. 内容关联视图和重要视图

关键帧之间的视图内容相关信息在系统的几个功能上都非常有用，如论文 7 所述，它表示了一个间接的权重图像。每个节点都是一个关键帧，关键帧之间存在一个边缘，帧上面可以观察到相同的地图云点（至少有 15 个），对于边缘上相同的云点的数量我们赋予权重 θ。

为了矫正闭环回路，我们像论文 6 那样做位姿图优化，优化方法将闭环回路的误差分散到图像里。为了排除内容相关视图的边缘，可能非常密集，我们构建了关键图像（Essential Graph）保留所有的节点（关键帧），但是边缘更少，这可以保持一个比较强的网络以获得精确的结果。系统增量式地构建一个生成树，从第一个关键帧开始，它连接了边缘数量最少的内容相关视图的子图像。当新的关键帧插入时，它加入树中连接到老的关键帧上，新旧关键帧具有最多的相同的云点，但一个关键帧通过筛选策略删除时，系统会根据关键帧所在的位置更新链接。关键图像（Essential Graph）包含了一个生成树，具有高视图相关性（θ_{min}=100）的相关视图的边缘子集，闭环回路边缘产生一个相机的强网络。图 2 是一个相关视图的例子，生成树和关联的关键图像。

第 8 章 E 节的实验里，运行位姿图优化时，方案效果精确 BA 优化几乎没有增强系统效果。关键图像的效果和 θ_{min} 的效果如第 8 章 E 节所示。

E. 基于图像词袋模型的位置识别

采用 DBoW2，系统嵌入了图像词袋位置识别模块，执行闭环检测和位姿重定位。视觉单词离散分布于特征描述子空间，视觉单词组成视觉字典。视觉字典是离线创建的，用 ORB 特征描述子从大量图像中提取。如果图像都比较多，相同的视觉字典在不同的环境下也能获得很好的性能，如论文 11 那样。系统增量式地构建一个数据库，包括一个逆序指针，存储每个视觉字典里的视觉单词，关键帧可以通过视觉字典查看，所以检索数据库效率比较高。当关键帧通过筛选程序删除时，数据库也会更新。

关键帧在视图上可能会存在重叠，检索数据库时，可能不止一个高分值的关键帧。DBoW2 认为是图像重叠的问题，就将时间上接近的所有图像的分值相加。但这并没有包括同一地点不同时间的关键帧。我们将这些与内容相关视图的关键帧进行分类。另外，我们的数据库返回的是分值高于最好分值 75%的所有关键帧。

词袋模型表示的特征匹配的另外一个优势在论文 5 里详细介绍。如果我们想计算两个 ORB 特征的对应关系，我们可以通过暴力匹配视觉字典树上某一层（6 层里面选第 2 层）的相同节点（关键帧）里的特征，这可以加快搜索。在闭环回路检测和重定位中，我们通过这个方法搜索匹配用作三角化新的云点。我们还通过方向一致性测试改进对应关系，具体如论文 11，这可以去掉无效数据，保证所有对应关系的内在方向性。

四 全自动地图初始化

地图初始化的目的是计算两帧之间的相关位姿来三角化一组初始地图云点。这个方法与场景图像不相关（平面的或一般的）而且不需要人工干预去选择一个好的两个视图对应关系，比如

具有明显视差。我们建议并行计算两个几何模型，平面视图的单映射和非平面视图的基本矩阵。我们用启发式的方法选择模型，并从相关位姿中进行重构。当个视图之间的关系比较确定时，我们才发挥作用，检测低视差的情况或已知两部分平面模糊的情况（如论文 27 所示），避免生成一个有缺陷的地图。这个算法的步骤是：

1.查找最初的对应关系：

从当前帧 Fc 提取 ORB 特征（只在最好的尺度上），与参考帧 Fr 搜索匹配 Xc <-->Xr。如果找不到足够的匹配，就重置参考帧。

2.并行计算两个模型：

在两个线程上并行计算单映射 Hcr 和基本矩阵 Fcr：

$X_c=H_{cr}X_r$ $X_c^TF_{cr}X_r=0$ （1）

在《多视图几何》里详细解释了分别使用归一化直接线性变换 DLT 和 8 点算法计算原理，通过 RANSAC 计算。为了使两个模型的计算流程尽量一样，将两个模型的迭代循环次数预先设置成一样，每次迭代的云点也一样，8 个基本矩阵，4 个单映射。每次迭代我们都给每一个模型 M（H 表示单映射，F 表示基本矩阵）计算一个分值 S_M：

$$S_M = \sum_i \left(\rho_M\left(d_{cr}^2(\mathbf{x}_c^i, \mathbf{x}_r^i, M)\right) + \rho_M(d_{rc}^2\left(\mathbf{x}_c^i, \mathbf{x}_r^i, M\right))\right)$$

$$\rho_M(d^2) = \begin{cases} \Gamma - d^2 & \text{if} \quad d^2 < T_M \\ 0 & \text{if} \quad d^2 \geq T_M \end{cases} \tag{2}$$

其中，d_{cr}^2 和 d_{rc}^2 是帧和帧之间对称的传递误差。T_M 是无效数据的排除阈值，它的依据是 X^2 的 95%（T_H=5.99, T_F=3.84，假设在测量误差上有 1 个像素的标准偏差）。τ 等于 T_H，两个模型在有效数据上对于同一误差 d 的分值相同，同样使得运算流程保持一致。

单映射和基本矩阵的分值最高。如果没有足够的有效数据，模型没有找到，算法流程重启，从第一步开始重新寻找。

3．模型的选择：

如果场景是平面的，靠近平面有一个低视差，可以通过单映射来描述。同样地，我们也可以找到一个基本矩阵，但问题不能很好地约束表示，从基本矩阵重构运动场景可能会导致错误的结果。我们应该选择单映射作为重构的方法，它可以从二维图像正确初始化或者检测到的低视差的情况而不进行初始化工作。另外一方面，非平面场景有足够的视差可以通过基本矩阵来表示，但单映射可以用来表示二维平面或者低视差的匹配子集。在这种情况下我们应该选择基本矩阵。我们用如下公式进行计算：

$$R_H = \frac{S_H}{S_H + S_F} \qquad (3)$$

如果 R_H>0.45，这表示二维平面和低视差的情况。其他的情况，我们选择基本矩阵。

4. 运动和运动结构重构

一旦选择好模型，我们就可以获得运动状态。在单映射的例子中，我们通过论文 23 的方法，提取 8 种运动假设。这个方法通过测试选择有效的方案。如果在低视差的情况下，云点跑到相机的前面或后面，测试就会出现错误从而选择一个错的方案。我们建议直接三角化 8 种方案，检查两个相机前面具有较少的重投影误差情况下，在视图低视差情况下是否大部分云点都可以看到。如果没有一个明确的方案胜出，我们就不执行初始化，重新从第一步开始。这种方法就提供了一个清晰的方案，在低视差和两个交叉的视图情况下，初始化程序更具鲁棒性，这是我们方案鲁棒性的关键所在。在基本矩阵的情况下，我们用校准矩阵和本证矩阵进行转换：

$$\mathbf{E}_{rc} = \mathbf{K}^T \mathbf{F}_{rc} \mathbf{K} \qquad (4)$$

我们通过论文 2 中的单值分解方法提取四种运动假设。针对单映射，我们三角化四种方案进行重构。

5. 捆集调整

最后我们执行一个全局捆集调整，如附录所示，用来优化初始化重构。

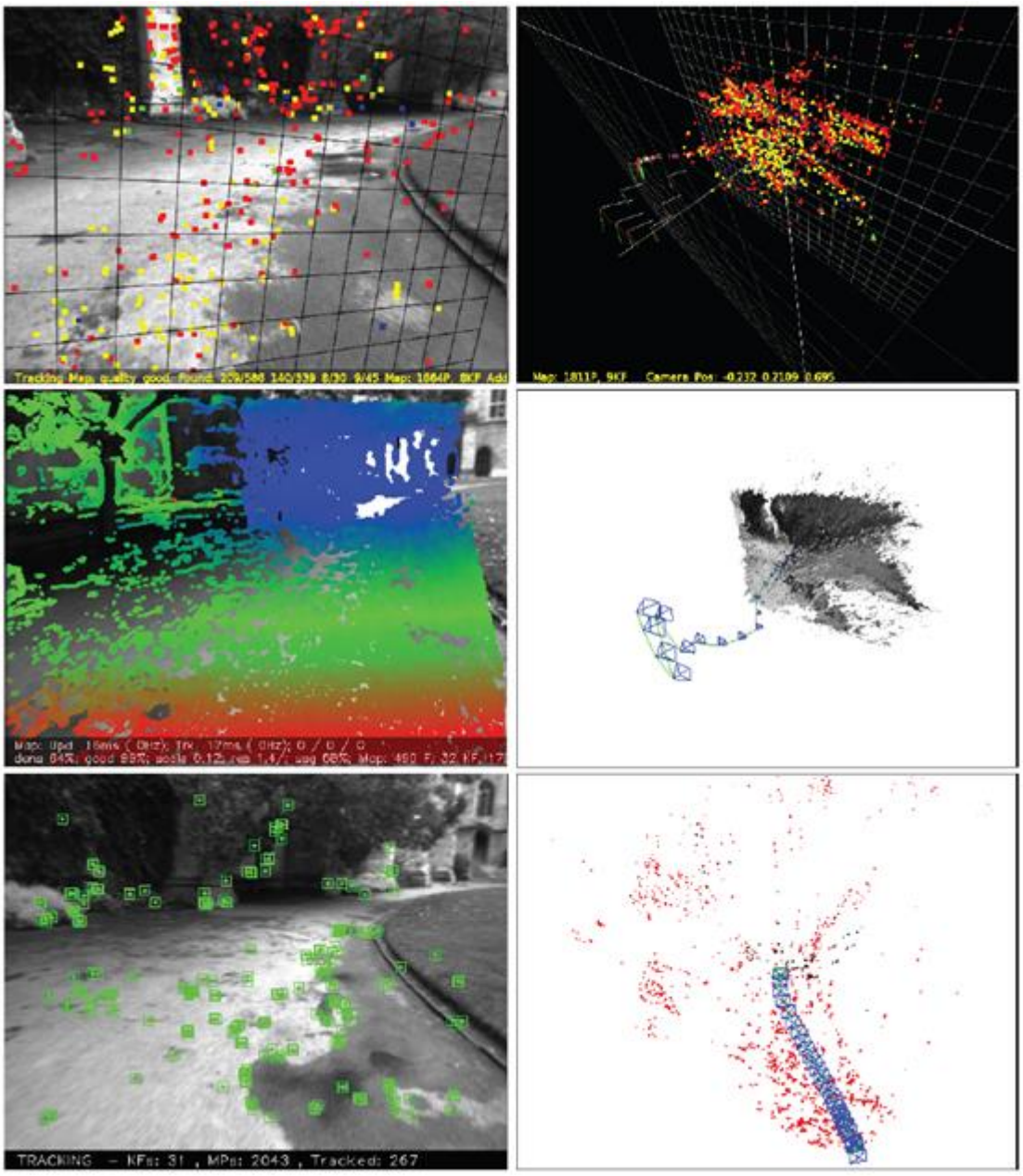


Fig. 3. Top: PTAM, middle LSD-SLAM, bottom: ORB-SLAM, some time after initialization in the NewCollege sequence [39]. PTAM and LSD-SLAM initialize a corrupted planar solution while our method has automatically initialized from the fundamental matrix when it has detected enough parallax. Depending on which keyframes are manually selected, PTAM is also able to initialize well.

图 3 所示的论文 39，室外环境下初始化工作具有很大挑战。PTAM 和 LSD-SLAM 初始化二维平面所有云点，我们的方法是有足够视差才进行初始化，从基本矩阵正确地初始化。

五 追踪

这章我们详细介绍追踪线程通过相机图像每一帧执行的具体步骤。相机的位置优化，如前几步提到的，由局部捆集调整构成，如附录中详细描述。

A. ORB 特征提取

我们在 8 层图像金字塔，提取 FAST 角点，尺度因子为 1.2。如果图像分辨率从 512*384 到 752*480，我们发现提取 1000 个角点比较合适，如果分辨率提高，如 KITTI 数据集，论文 40，我们提取 2000 个角点。为了确保单映射分布，我们将每层分成网格，每格提取至少 5 个角点。检测每格角点的时候，如果角点数量不够，就提高阈值。角点的数量根据单元格变化，即使单元格检测不出角点（没有纹理或对比度低的情况）。再根据保留 FAST 的角点计算方向和 ORB 特征描述子。ORB 特征描述子用于所有的特征匹配，而不是像 P T A M 那样根据图像区块关联性进行搜索。

B. 前一图像帧作初始位姿估计

如果上一帧的追踪成功，我们就用同样的速率运动模型计算相机位置，搜索上一帧观测到的地图云点。如果没有找到足够的匹配（比如，运动模型失效），我们就加大搜索范围搜索上一帧地图云点附近的点。通过寻找到的对应关系优化相机位姿。

C. 全局重定位优化初始化位姿估计

如果跟踪丢失，我们把当前帧转换成图像词袋，检索图像数据库，为全局重定位查找关键帧。我们计算 ORB 特征和每个关键帧的地图云点的对应关系，如第三章 E 节描述。接着，我们对每个关键帧执行 RANSAC 迭代计算，用 PnP 算法找出相机位置。如果通过足够的有效数据找到相机位姿，我们优化它的位姿，搜索候选关键帧的地图云点的更多的匹配。最后，相机位置被进一步优化，如果有足够的有效数据，跟踪程序将持续执行。

D. 局部地图追踪

一旦我们获得了相机位姿和一组初始特征匹配的估计 ,我们可以将地图和图像帧对应起来，搜索更多地图云点的对应关系。为了减少大地图的复杂性，我们只映射局部地图。局部地图包含一组关键帧 K1，它们和当前关键帧有相同的地图云点，与相邻的关键帧组 K2 图像内容相关。局部地图有一个参考关键帧 Kref∈K1，它与当前帧具有最多相同的地图云点。针对 K1, K2 可见的每个地图云点，我们通过如下步骤，在当前帧中进行搜索：

1. 计算当前帧中的地图云点映射集 x。如果位于图像边缘外面的点，就丢掉。
2. 计算当前视图射线 v 和地图云点平均视图方向 n 的夹角。如果 n< $\cos(60^{o})$，就丢掉。
3. 计算地图云点到相机中心的距离 d。如果它不在地图云点的尺度不变区间，即 d∉【d_{min},d_{max}】，就丢掉。
4. 计算每帧图像的尺度因子，比值为 d/d_{min}。
5. 对比地图云点的特征描述子 D 和当前帧中还未匹配的 ORB 特征，在尺度因子，和靠近 x 的云点作最优匹配。

相机位姿最后通过当前帧中获得所有的地图云点进行优化。

E. 新关键帧的判断标准

最后一步是决定当前帧是否可以作为关键帧。局部地图构建的过程中有一个机制去筛选冗余的关键帧，我们尽可能快地插入关键帧，这可以使跟踪线程对相机的运动更具鲁棒性，尤其是旋转。我们根据以下要求插入新的关键帧：

1. 每次全局重定位过程需要超过 20 个图像帧。
2. 局部地图构建处于空闲状态，或者上一个关键帧插入时，已经有超过 20 个关键帧。
3. 当前帧跟踪至少 50 个地图云点。
4. 当前帧跟踪少于参考关键帧云点的 90%。

与 PTAM 中用关键帧的距离作为判断标准不同，我们加入一个最小的视图变换，如条件 4 要求。条件 1 确保一个好的重定位，条件 3 保证好的跟踪。当局部地图构建处于忙状态（条件 2 的后半部分）的同时插入关键帧的时候，就会发信号去暂停局部捆集调整，这样就可以尽可能快地去处理新的关键帧。

六 局部地图构建

这章我们将描述根据每个新的关键帧 K_i 构建局部地图的步骤。

A. 关键帧插入

我们先更新内容相关的视图，添加节点 K_i，更新关键帧间具有相同地图云点产生的边缘。我们还要更新生成树上 K_i 和其他关键帧的链接。然后，计算表示关键帧的词袋，用于数据关联来三角化新的云点。

B. 地图云点筛选

保留在地图里的地图云点，在最初创建的 3 个关键帧上，需要经过严格的检测，保证它们可以被跟踪，不会由于错误的数据被错误地被三角化。一个云点必须满足如下条件：

1. 跟踪线程必须要找到超过 25%的关键帧。

2. 如果超过一个关键帧完成地图云点创建过程，它必须至少是能够被其他 3 个关键帧可被观测到。

一旦一个地图云点通过测试，它只能在被少于 3 个关键帧观测到的情况下移除。在局部捆集调整删除无效观测数据的情况下，关键帧才能被筛除掉。这个策略使得我们的地图包含很少的无效数据。

C. 新地图云点标准

在内容相关的视图 Kc 之间三角化 ORB 特征向量，可以创建新的地图云点。对 Ki 中每个未匹配的 ORB 特征，和其他关键帧中没有匹配的云点，我们查找一个匹配。这个匹配过程在第三章 E 节详细解释，丢掉那些不满足极对约束的匹配。ORB 特征对三角化后，将要获得新的云点，这时要检查两个相机视图的景深，视差，重映射误差，和尺度一致性。起初，一个地图云点通过 2 个关键帧进行观测，但它却是和其他关键帧匹配，所以它可以映射到其他相连的关键帧，按照第 5 章 D 节的方法搜索对应关系。

D. 局部捆集调整

局部捆集调整优化当前处理的关键帧 Ki，在交互视图集 Kc 中所有连接到它的关键帧，和所有被这些关键帧观测到的地图云点。所有其他能够观测到这些云点的关键帧但没有连接到当前处理的关键帧的这些关键帧会被保留在优化线程中，但会被修复。被标记为无效数据的观测将在优化中间阶段和最后阶段被丢弃。附录有详细的优化细节。

E. 局部关键帧筛选

为了使重构保持简洁，局部地图构建尽量检测冗余的关键帧，删除它们。这会大有帮助，随着关键帧数量的增加，捆集调整的复杂度增加，但关键帧的数量也不会无限制地增加，因为它要在同一环境下在整个运行执行操作，除非场景内容发生变化。我们会删除 Kc 中所有 90%的地图云点可以至少被其他 3 个关键帧在同一或更好的尺度上观测到的关键帧。

七 闭环控制

闭环控制线程获取 Ki，上一个局部地图构建的关键帧，用于检测和闭合回环。具体步骤如下所示：

A. 候选回环检测

我们先计算 K_i 的词袋向量和它在视图内容相关的所有邻近图像（$\theta_{min}=30$）中计算相似度，保留最低分值 S_{min}。然后，我们检索图像识别数据库，丢掉那些分值低于 S_{min} 的关键帧。这和 DBoW2 中均值化分值的操作类似，可以获得好的鲁棒性，它计算的是前一帧图像，而我们是使用的内容相关信息。另外，所有连接到 Ki 的关键帧都会从结果中删除。为了获得候选回环，我们必须联系检测 3 个一致的候选回环（内容相关图像中的关键帧）。如果对 Ki 来说环境样子都差不多，就可能有几个候选回环。

B. 计算相似变换

单目 SLAM 系统有 7 个自由度，3 个平移，3 个旋转，1 个尺度因子，如论文 6。因此，闭合回环，我们需要计算从当前关键帧 Ki 到回环关键帧 Kl 的相似变换，以获得回环的累积误差。计算相似变换也可以作为回环的几何验证。

我们先计算 ORB 特征关联的当前关键帧的地图云点和回环候选关键帧的对应关系，具体步骤如第 3 章 E 节所示。此时，对每个候选回环，我们有了一个 3D 到 3D 的对应关系。我们对每个候选回环执行 RANSAC 迭代，通过 Horn 方法（如论文 42）找到相似变换。如果我们用足够的有效数据找到相似变换 Sil，我们就可以优化它，并搜索更多的对应关系。如果 Sil 有足够的有效数据，我们再优化它，直到 Kl 回环被接受。

C. 回环融合

回环矫正的第一步是融合重复的地图云点，插入与回环闭合相关的相似视图的新边缘。先通过相似变换 Sil 矫正当前关键帧位姿 Tiw，这种矫正方法应用于所有 Ki 相邻的关键帧，执行相似变换，这样回环两端就可以对齐。

回环关键帧所有的地图云点和它的近邻映射到 Ki，通过映射在较小的区域内搜索它的近邻和匹配，如第 5 章 D 节所述。所有匹配的地图云点和计算 Sil 过程中的有效数据进行融合。融合过程中所有的关键帧将会更新它们的边缘，这些视图内容相关的图像创建的边缘用于回环控制。

D. 本征图像优化

为了有效地闭合回环，我们通过本征矩阵优化位姿图，如第三章 D 节所示，这样可以将回环闭合的误差分散到图像中去。优化程序通过相似变换校正尺度偏移，如论文 6。误差和成本计算如附录所示。优化过后，每一个地图云点都根据关键帧的校正进行变换。

八 实验

我们对实验进行了评估，用 Newcollege 室外大场景机器图像序列，评估整个系统性能；用 TUM RGB-D 的 16 个手持式室内图像序列，评估定位精度，重定位和程序运行能力；用 KITTI 的 10 个汽车户外图像数据集，评估实时大场景操作，定位精度和位姿图优化效率。

我们的电脑配置为 Intel Core i7-4700MQ （4 核@2.40GHz）和 8GB RAM，用于实时处理图像。ORB-SLAM 有 3 个主线程，它和其他 ROS 线程并行处理。

A. Newcollege

Newcollege 是一个 2.2 公里的校园和相邻的公园的机器人图像序列。它是双目相机，帧率 20fps，分辨率 512x38。它包含几个回环和快速的旋转，这对单目视觉非常具有挑战性。据我们所知，没有单目系统可以处理整个图像序列。例如论文 7，可以形成回环，也可以应用于大场景环境，但对单目结果只能显示序列的一小部分。

图 4 显示的回环闭合线程通过有效数据支持相似变换。图 5 对比了回环闭合前后的重构状况。红色是局部地图，回环闭合延伸回环的两端后的状况。图 6 是实时帧率状态下处理图像序列的整个地图。后边的大回环并没有完全闭合，它从另外一个方向穿过，位置识别程序没能发现闭合回环。

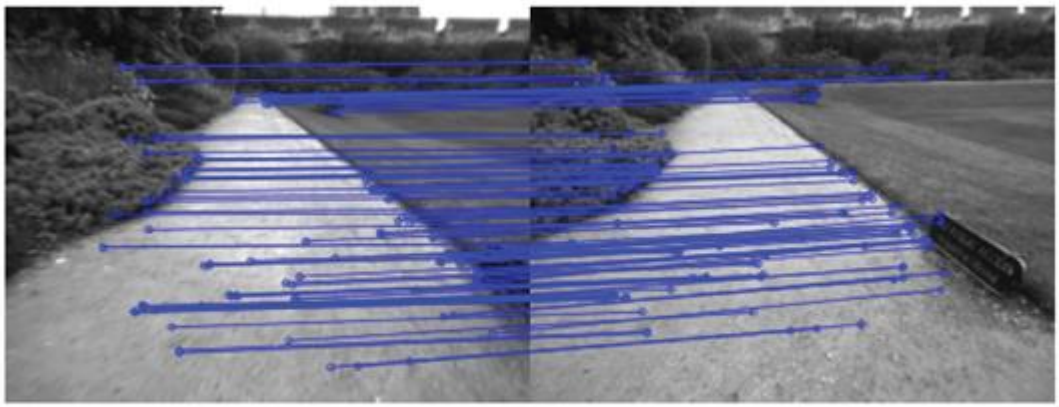


Fig. 4. Example of loop detected in the NewCollege sequence. We draw the inlier correspondences supporting the similarity transformation found.

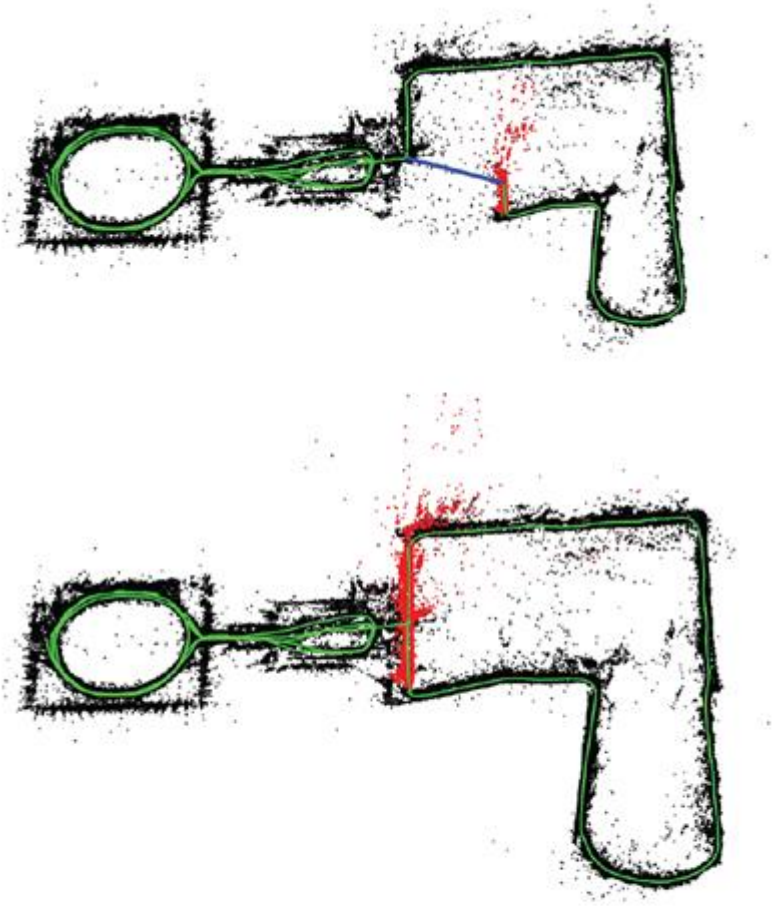


Fig. 5. Map before and after a loop closure in the NewCollege sequence. The loop closure match is drawn in blue, the trajectory in green, and the local map for the tracking at that moment in red. The local map is extended along both sides of the loop after it is closed.

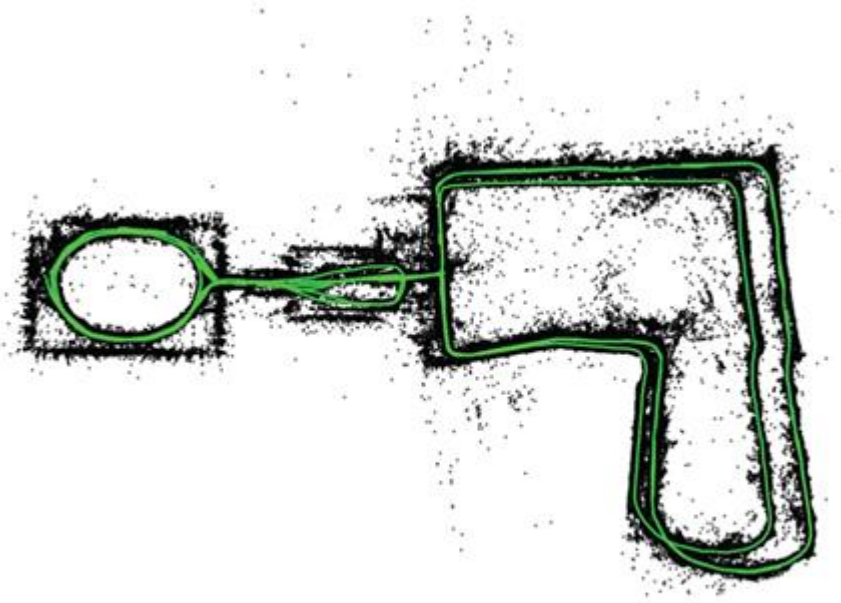


Fig. 6. ORB-SLAM reconstruction of the full sequence of NewCollege. The bigger loop on the right is traversed in opposite directions and not visual loop closures were found, therefore they do not perfectly align.

我们统计了每个线程所用时间。表 1 显示了追踪和局部地图构建所用的时间。跟踪的帧率大概在 25-30Hz，在做局部地图跟踪时的要求最高。如有必要，这个时间可以减少，需要限制局部地图关键帧的数量。局部地图构建线程中要求最高的是局部捆集调整。局部捆集调整的时间根据机器人的状态变动，向未知环境运动或在一个已经建好地图的区域运动是不同的，因为在未知环境中如果跟踪线程插入一个新的关键帧，捆集调整会被中断，如第 5 章 E 节所示。如果不需要插入新的关键帧，局部捆集调整会执行大量已经设置的迭代程序。

TABLE I
TRACKING AND MAPPING TIMES IN NEWCOLLEGE

Thread	Operation	Median (ms)	Mean (ms)	Std (ms)
TRACKING	ORB extraction	11.10	11.42	1.61
	Initial Pose Est.	3.38	3.45	0.99
	Track Local Map	14.84	16.01	9.98
	Total	30.57	31.60	10.39
LOCAL MAPPING	KeyFrame Insertion	10.29	11.88	5.03
	Map Point Culling	0.10	3.18	6.70
	Map Point Creation	66.79	72.96	31.48
	Local BA	296.08	360.41	171.11
	KeyFrame Culling	8.07	15.79	18.98
	Total	383.59	464.27	217.89

TABLE II
LOOP CLOSING TIMES IN NEWCOLLEGE

			Loop Detection (ms)		Loop Correction (s)		
Loop	KeyFrames	Essential Graph Edges	Candidates Detection	Similarity Transformation	Fusion	Essential Graph Optimization	Total (s)
1	287	1347	4.71	20.77	0.20	0.26	0.51
2	1082	5950	4.14	17.98	0.39	1.06	1.52
3	1279	7128	9.82	31.29	0.95	1.26	2.27
4	2648	12547	12.37	30.36	0.97	2.30	3.33
5	3150	16033	14.71	41.28	1.73	2.80	4.60
6	4496	21797	13.52	48.68	0.97	3.62	4.69

表 2 显示了 6 个闭合回路的结果。可以看到回环检测是如何亚线性地增加关键帧的数量。主要是由于检索数据库的效率高，它值比较了具有相同图像单词的图像子集，由此可见用于位置识别的词袋模型的作用。我们的本征图像包含的边缘是关键帧数量的 5 倍，它是一个稀疏图。

B. TUM RGB-D 定位精度对比

论文 38 里，评测的 TUM RGB-D 数据集，用于评估相机定位精度，它通过外部运动捕捉系统提供了具有精确基准的几个图像序列。我们去掉那些不适合纯单目 SLAM 系统的图像序列，这些序列包含强烈的旋转，没有纹理或没有运动。

为了方便比较，我们还运行了直接的半稠密 LSD-SLAM（论文 10）和 PTAM（论文 4）作为对比。我们还比较了 RGBD-SLAM（论文 43）轨迹作为对比。为了在相同的基准下比较 ORB-SLAM，LSD-SLAM 和 PTAM，我们用相似变换对齐关键帧轨迹，在尺度未知的情况下，检测轨迹的绝对误差（论文 38）。对 RGBD-SLAM 我们通过相机坐标变换对齐轨迹，同样的方法检测尺度是否覆盖良好。LSD-SLAM 从随机深度值进行初始化，执行聚类，因此与基准对比的时候，我们会丢掉前 10 个关键帧。对于 PTAM，我们从一个好的初始化中，手动选择两个关键帧。表 3 是对 16 个图像序列运行 5 次的中等结果。

ORB-SLAM 可以运行所有的图像序列，除了 *fr3 nostructure texture far (fr3 nstr tex far)* 以外。这是一个平面的场景，相机的轨迹可能有两种解释，比如论文 20 中的描述。我们的初始化方法检测到模糊的视图，为了安全而不进行初始化。PTAM 初始化有时会选择对的方案，有些可能会选择错的方案，但导致的错误可能不能接受。我们没有注意到 LSD-SLAM 的 2 种不同的重构方案，但在这个图像序列出现的错误非常多。其他的图像序列，PTAM 和 LSD-SLAM 的鲁棒性比我们的方法差，容易跟踪丢失。

关于精度问题，ORB-SLAM 和 PTAM 非常相似，ORB-SLAM 在图像序列 *fr3 nostructure texture near withloop (fr3 nstr tex near)*中检测大的闭环时，可以获得更高精度。非常意外的一个结果是 PTAM 和 ORB-SLAM 都非常明显地表现出精度高于 LSD-SLAM 和 RGBD-SLAM。一个可能的原因是它们没有用传感器的测量优化位姿图从而减少了对地图的优化，但我们采用捆集调整，同时通过传感器测量优化相机和地图，通过运动结构的经典算法来解决传感器测量，如论文 2 所示。我们在第 9 章 B 节进一步讨论了这个结果。另一个有趣的结果是在图像序列 *fr2 desk with person* 和 *fr3 walking xyz* 中，LSD-SLAM 对动态物体的鲁棒性相比 ORB-SLAM 差一些。

我们注意到 RGBD-SLAM 在图像序列 fr2 上尺度上有偏差，可以用 7 个自由度对齐轨迹明显减少误差。最后我们注意到论文 10 提到在 f2_xyz 上 PTAM 的精度比 LSD-SLAM 低，RMSE 是 24.28cm。然而，论文没有给出足够的细节说明如何获得这些结果的，我们没有办法复现它。

TABLE III
KEYFRAME LOCALIZATION ERROR COMPARISON IN THE TUM RGB-D BENCHMARK [38]

	Absolute KeyFrame Trajectory RMSE (cm)			
	ORB-SLAM	PTAM	LSD-SLAM	RGBD-SLAM
fr1_xyz	**0.90**	1.15	9.00	1.34 (1.34)
fr2_xyz	0.30	**0.20**	2.15	2.61 (1.42)
fr1_floor	**2.99**	X	38.07	3.51 (3.51)
fr1_desk	**1.69**	X	10.65	2.58 (2.52)
fr2_360_kidnap	3.81	**2.63**	X	393.3 (100.5)
fr2_desk	**0.88**	X	4.57	9.50 (3.94)
fr3_long_office	**3.45**	X	38.53	-
fr3_nstr_tex_far	ambiguity detected	4.92 / 34.74	18.31	-
fr3_nstr_tex_near	**1.39**	2.74	7.54	-
fr3_str_tex_far	**0.77**	0.93	7.95	-
fr3_str_tex_near	1.58	**1.04**	X	-
fr2_desk_person	**0.63**	X	31.73	6.97 (2.00)
fr3_sit_xyz	**0.79**	0.83	7.73	-
fr3_sit_halfsph	**1.34**	X	5.87	-
fr3_walk_xyz	**1.24**	X	12.44	-
fr3_walk_halfsph	**1.74**	X	X	-

C. TUM RGB-D 重定位性能比较

我们在 TUM RGB-D 数据集上进行了重定位性能的对比实验。在第一个实验中，我们通过 fr2_xyz 的前 30 秒构建了一个地图，执行了全局重定位，评估了位姿精度。对 PTAM 进行了相

同的实验。图 7 是创建初始地图的关键帧 ,重定位的图像帧位姿和这些帧的基准。可以看出 PTAM 由于重定位方法较少的不变性而只重定位关键帧附件的图像帧。表 4 显示了相对基准的调用和误差。ORB-SLAM 比 PTAM 可以更精准地多定位 2 倍的图像帧。在第 2 个实验中，我们采用 fr3_sitting_xyz 初始化地图，从 fr3_walking_xyz 重定位所有的图像帧。这是一个颇具挑战性的实验，由于图像中有人移动造成了阻塞。这里，PTAM 并没有重定位，而 ORB-SLAM 重定位 78% 的图像帧，如表 4 所示。图 8 显示了 ORB-SLAM 重定位的一些挑战。

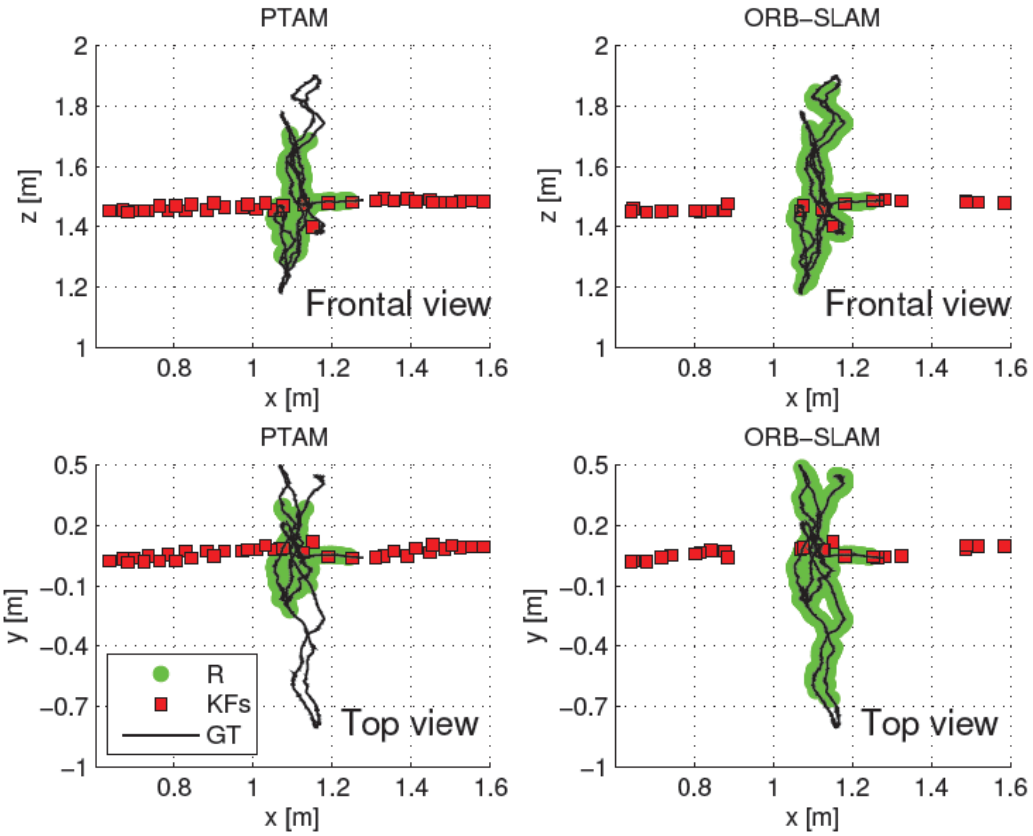


Fig. 7. Relocalization experiment in *fr2_xyz*. Map is initially created during the first 30 seconds of the sequence (KFs). The goal is to relocalize subsequent frames. Successful relocalizations (R) of our system and PTAM are shown. The ground truth (GT) is only shown for the frames to relocalize.

TABLE IV
RESULTS FOR THE RELOCALIZATION EXPERIMENTS

System	Initial Map		Relocalization		
	KFs	RMSE (cm)	Recall (%)	RMSE (cm)	Max. Error (cm)
fr2_xyz. 2769 frames to relocalize					
PTAM	37	0.19	34.9	0.26	1.52
ORB-SLAM	24	0.19	**78.4**	0.38	1.67
fr3_walking_xyz. 859 frames to relocalize					
PTAM	34	0.83	0.0	-	-
ORB-SLAM	31	0.82	**77.9**	1.32	4.95

Fig. 8. Example of challenging relocalizations (severe scale change, dynamic objects) that our system successfully found in the relocalization experiments.

D. TUM RGB-D 系统运行实验对比

之前的重定位实验表明我们的系统可以从不同的视角定位地图，在中等动态环境中的鲁棒性也较好。这个特性和关键帧筛选程序可以在不同的视角和局部动态环境中一直运行到程序结束。

在全静态场景情况下，即使相机从不同视角观测场景，ORB-SLAM 也可以使关键帧数量保持在一个有限的水平内。我们演示了一个图像序列，相机 93 秒以内都在拍摄同一张桌子，但视角一直变换，形成一个轨迹。我们对比了我们地图的关键帧数量和 PTAM 生成的关键帧，如图 9 所示。可以看到 PTAM 一直插入关键帧，ORB-SLAM 会删除冗余的关键帧。

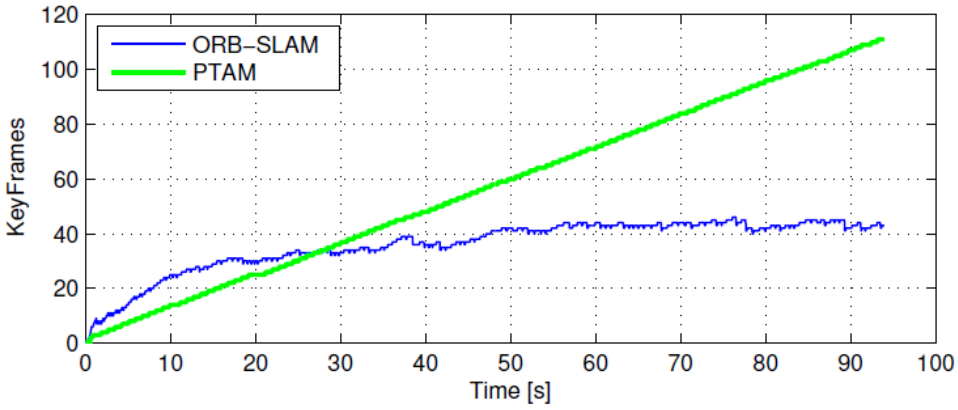


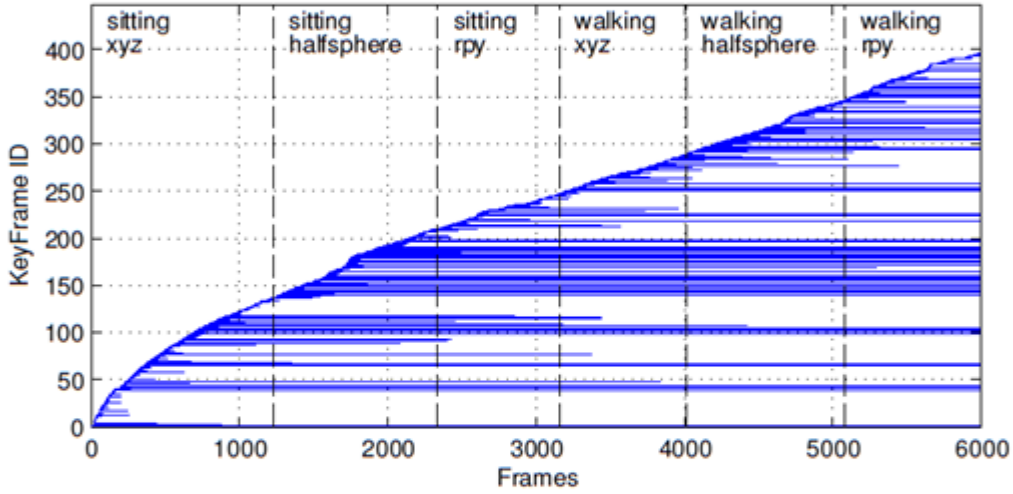
Fig. 9. Lifelong experiment in a static environment where the camera is always looking at the same place from different viewpoints. PTAM is always inserting keyframes, while ORB-SLAM is able to prune redundant keyframes and maintains a bounded-size map.

当然，在整个程序运行过程中，静态环境下的正常操作是任何 SLAM 系统的一个基本要求，更引人注意的是动态环境下的状况。我们在几个 *fr3* 的图像序列中分析了 ORB-SLAM 系统的状况，图像序列有：*sitting xyz, sitting halfsphere, sitting rpy, walking xyz, walking halfspehere* 和 *walking rpy*。所有的视频中，相机都对着桌子，但运动轨迹不同，有人在移动，椅子也被移动了。图 10（a）是地图中所有关键帧的总数量，图 10（b）显示从图像帧中创建或删除关键帧，从关键帧到地图构建需要多久时间。可以看到前 2 个图像序列中新看到（增加）场景时地图的大小一直在增加。图 10（b）是前 2 个视频中创建的关键帧。在视频 sitting_rpy 和 walking_xyz 中，地图没有增加，地图是通过已有场景创建。相反，在最后两个视频中，有更

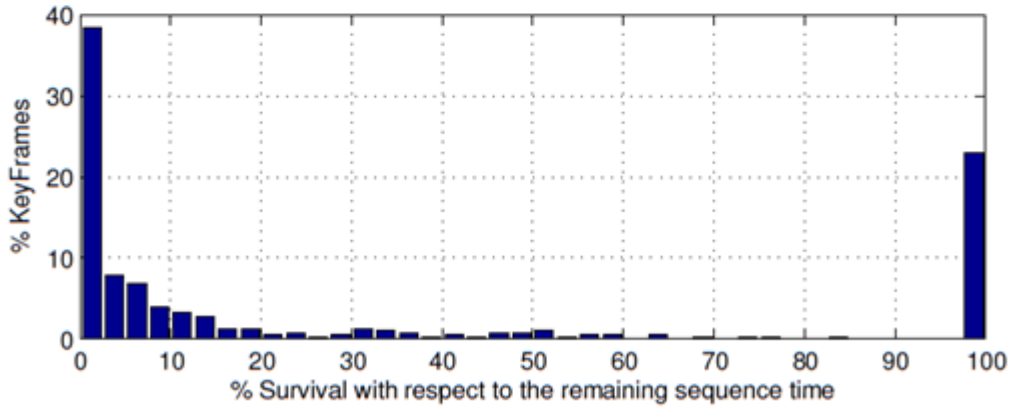
多的关键帧插入但没有在场景中表示出来，可能由于场景的动态变化。图 10（C）关键帧的柱状图，它们是从视频中挑选出来的。大部分的关键帧被筛选程序删除了，只有一小部分留下来了。ORB-SLAM 有大量关键帧的生成策略，在未知环境下非常有用；后面系统会生成一个小的子集来代表这些关键帧。

在整个实验中，我们系统的地图根据场景上内容来增加，而不是根据时间，它可以存储场景的动态变化，对场景的理解非常有用。

(a) Evolution of the number of keyframes in the map



(b) Keyframe creation and destruction. Each horizontal line corresponds to a keyframe, from its creation frame until its destruction

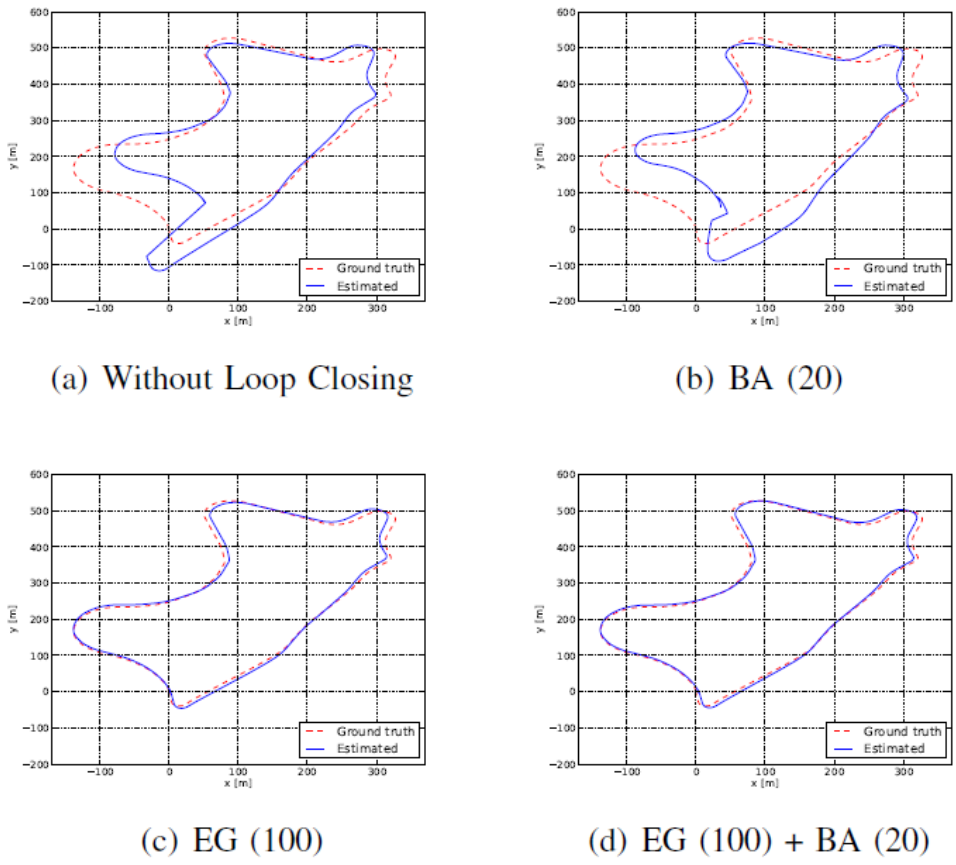


(c) Histogram of the survival time of all spawned keyframes with respect to the remaining time of the experiment

Fig. 10. Lifelong experiment in a dynamic environment from the TUM RGB-D Benchmark.

E. KITTI 大场景大回环行性能对比

里程计通过 KITTI 数据集的 11 个视频对比，它是在住宅区驾驶汽车，基准精度非常高，有一个 GPS 和一个 Velodyne Laser Scanner。对单目系统非常有调整性，它里面有快速旋转，区域内有大量树叶，这使数据关联变得更困难，车速比较快，视频的帧率只有 10fps。除了视频 01 外，ORB-SLAM 可以处理其他所有的视频，01 是高速路上的视频，可追踪的物体非常少。视频 00,02,05,06,07,09，有闭环回路，系统可以检测到，并使它闭合。视频 09 的闭环只能在视频的最后几个图像帧里检测到，并不是每次都能成功检测到。



(a) Without Loop Closing (b) BA (20)

(c) EG (100) (d) EG (100) + BA (20)

Fig. 13. Comparison of different loop closing strategies in KITTI 09.

对于轨迹和基准的定性比较如图 11 和 12 所示。在 TUM RGB-D 对比中，我们可以通过相似变换对齐轨迹的关键帧和基准。图 11 是定性比较的结果，图 12 是论文 25 中的最新单目 SLAM

在视频 00,05,06,07 和 08 上执行的结果。除了 08 有一些偏移以外，ORB-SLAM 在这些视频上的轨迹都很精准。

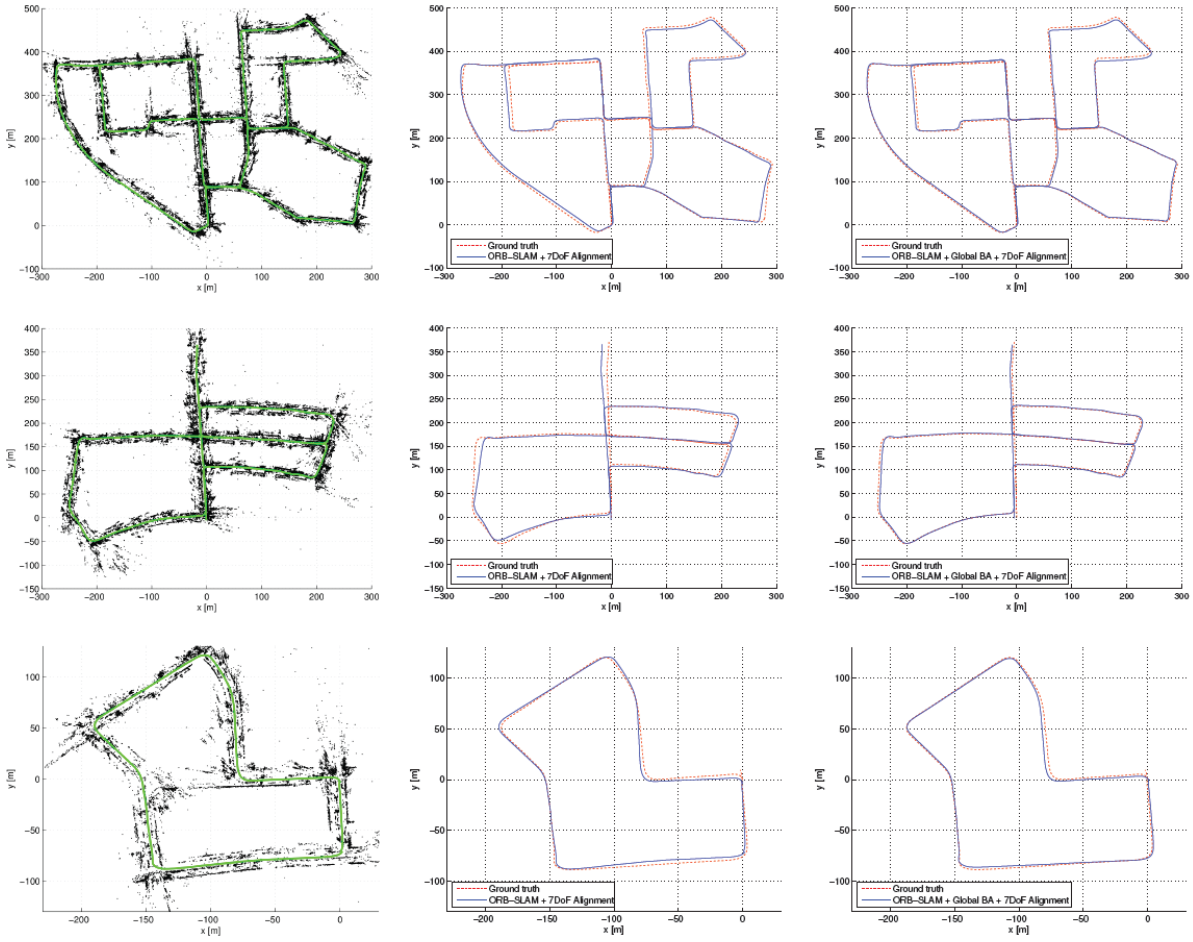
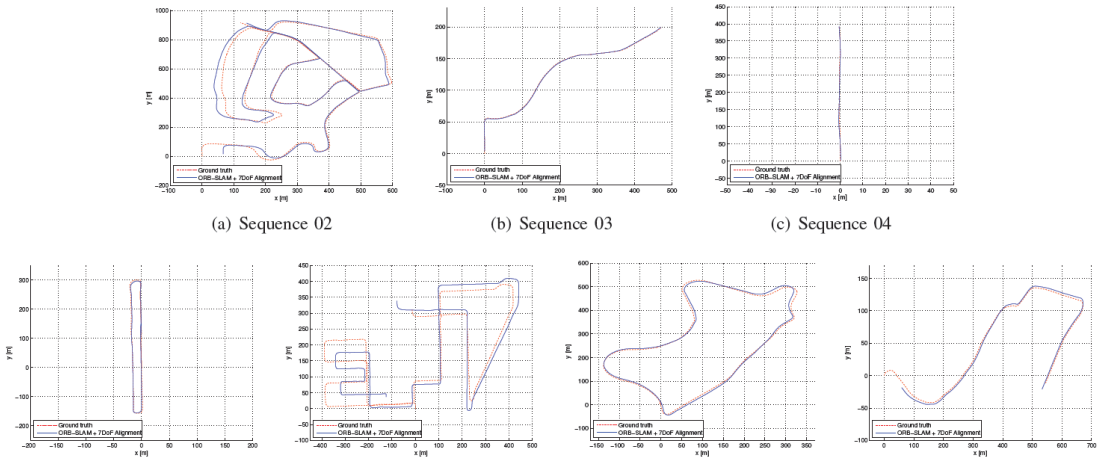


Fig. 11. Sequences 00, 05 and 07 from the odometry benchmark of the KITTI dataset. Left: points and keyframe trajectory. Center: trajectory and ground truth. Right: trajectory after 20 iterations of full BA. The output of our system is quite accurate, while it can be slightly improved with some iterations of BA.



(a) Sequence 02 (b) Sequence 03 (c) Sequence 04

(d) Sequence 06 (e) Sequence 08 (f) Sequence 09 (g) Sequence 10

Fig. 12. ORB-SLAM keyframe trajectories in sequences 02, 03, 04 ,06, 08, 09 and 10 from the odometry benchmark of the KITTI dataset. Sequence 08 does not contains loops and drift (especially scale) is not corrected.

表 5 显示了每个视频的关键帧轨迹的中等 RMSE 误差。我们基于地图尺寸提供了轨迹的误差。结果表明我们的轨迹误差吃地图尺寸的 1%左右。大致范围低的是视频 03 的 0.3%高的是视频 08 的 5%。视频 08 中没有闭环，漂移也没办法纠正，因为闭环控制需要获得更精确的重构。

TABLE V
RESULTS OF OUR SYSTEM IN THE KITTI DATASET.

		ORB-SLAM		+ Global BA (20 its.)	
Sequence	Dimension (m×m)	KFs	RMSE (m)	RMSE (m)	Time BA (s)
KITTI 00	564 × 496	1391	6.68	5.33	24.83
KITTI 01	1157 × 1827	X	X	X	X
KITTI 02	599 × 946	1801	21.75	21.28	30.07
KITTI 03	471 × 199	250	1.59	1.51	4.88
KITTI 04	0.5 × 394	108	1.79	1.62	1.58
KITTI 05	479 × 426	820	8.23	4.85	15.20
KITTI 06	23 × 457	373	14.68	12.34	7.78
KITTI 07	191 × 209	351	3.36	2.26	6.28
KITTI 08	808 × 391	1473	46.58	46.68	25.60
KITTI 09	465 × 568	653	7.62	6.62	11.33
KITTI 10	671 × 177	411	8.68	8.80	7.64

我们还确认了全局捆集调整的 20 层迭代如何增强地图重构，在每个视频的最后，细节如附录所示。我们还注意到全局捆集调整可以稍微增加闭环轨迹的精度，但这对开环轨迹有负面影响，这意味着我们的系统已经非常精确了。在有些应用中，如果需要非常精确的结果我们的算法会提供一组匹配，需要定义一个比较强的相机网络，一个初始估计，这样全局捆集调整迭代次数就会变少。

最后讲一下闭环控制和用于本征图像的边缘的 θ_{min} 的效率。视频 09 是个闭环，时间长。我们评估的不同的闭环策略。表 6 是关键帧轨迹 RMSE 和不同情况下没有闭环控制优化所用的时间，如果直接采用全局捆集调整（20 层或 100 层迭代），如果只用位姿图优化（10 层迭代不同数量的边缘），如果先用位姿图优化再执行全局捆集调整。结果表明，在闭环控制之前，方案就比较糟糕，以至于捆集调整会有融合问题。100 层迭代后误差就非常大。本征图像优化融合速度更快，结果也更精确。θ_{min} 对精度影响并不大，减少边缘的数量会明显降低精度。位姿图优化后再执行一个捆集调整可以增加精度，但时间也增加了。

TABLE VI
COMPARISON OF LOOP CLOSING STRATEGIES IN KITTI 09

Method	Time (s)	Pose Graph Edges	RMSE (m)
-	-	-	48.77
BA (20)	14.64	-	49.90
BA (100)	72.16	-	18.82
EG (200)	0.38	890	8.84
EG (100)	0.48	1979	8.36
EG (50)	0.59	3583	8.95
EG (15)	0.94	6663	8.88
EG (100) + BA (20)	13.40	1979	7.22

First row shows results without loop closing. Number between brackets for BA (Bundle Adjustment) means number of Levenberg-Marquardt (LM) iterations, while for EG (Essential Graph) is the $\theta_{\min}$ to build the Essential Graph. All EG optimizations perform 10 LM iterations.

九 结论和讨论

A. 结论

通过实验，我们的系统可以应用于室内和室外场景，用在汽车、机器人和手持设备上。室内小场景的精度在 1 厘米，室外大场景的应用是几个厘米。

PTAM 被认为是目前最精准的单目 SLAM。根据离线运动结构问题，PTAM 后端捆集调整不能同时加载。PTAM 的主要贡献是给机器人 SLAM 的研究带来了新的技术，具有良好的实时性。我们的主要贡献是扩张了 PTAM 对于应用环境的多样性和对系统的交互性。因此，我们挑选了新的方法和算法，整合了前人优秀的地方，如论文 5 的闭环检测，论文 6,7 的闭环控制程序和内容关联视图，论文 37 的 g2o 优化框架，论文 9 的 ORB 特征。目前所知，我们的精度最高，是最可靠最完整的单目 SLAM 系统。我们的生成和挑选关键帧策略可以创建较少的关键帧。弹性的地图探测在条件较差的轨迹上非常有用，比如旋转和快速移动。在相同场景的情况下，地图在只有新内容的情况下才会增长，保存了不同的视图外观。

最后，ORB 特征可识别剧烈的视角变换的位置。也可以快速提取特征和匹配特征，使得实时追踪和地图构建更精确。

B. 基于离散/特征方法与稠密/直接方法对比

论文 44 的 DTAM 和论文 10 的 LSD-SLAM 可以进行环境的稠密或半稠密重构，相机位姿可以通过图像像素的亮度直接优化。直接方法不需要特征提取，可以避免人工匹配。他们对图像模糊，低纹理环境和论文 45 的高频纹理的鲁棒性更好。稠密的重构，对比稀疏的地图云点，比如 ORB-SLAM 或 PTAM，对相机定位更有效。

然而，直接方法有他们自己的局限。首先，这些方法采用一个表面反馈模型。光电描记法的一致性要求限制了匹配的基准，这要比其他特征窄得多。这对重构的精度影响很大，需要更宽的观测减少深度的不确定。直接方法，如果准确的建模，可能会受到快门，自动增益和自动曝光的影响（如 TUM RGB-D 的对比测试）。最后，由于直接方法计算要求较高，像 DTAM 中地图增量式地扩张，或像 LSD-SLAM 丢掉传感器测量，根据位姿图优化地图。

相反 ,基于特征的方法可以在更宽的基准上匹配特征 ,主要得益于其较好地视图不变特性。捆集调整和相机位姿优化，地图云点通过传感器测量进行融合。在运动结构估计中，论文 46 已经指出了基于特征的方法相比直接方法的优势。在我们的实验中，直接提供了证据，第 8 章 B 节，表明精度更高。未来单目 SLAM 应该会整合两种最好的方法。

C. 后续工作

系统精度可以进一步增强。这些云点，如果没有足够的视差是不放入地图中的，但对相机的旋转非常有用。

另外一种方法是将稀疏地图更新到稠密地图 ,重构更有作用。由于我们关键帧的选择机制，关键帧有高精度位姿和更多的内容相关视图信息。所以，ORB-SLAM 稀疏地图是一个非常优秀的初始估计框架，比稠密地图更好。论文 47 有详细描述。

附录： 非线性优化

捆集调整

地图云点 3D 位置 $X_{w,j} \in R^3$，关键帧位姿 $T_{iw} \in SE(3)$

W 表示世界坐标，通过匹配的关键点 $X_{i,j} \in R^2$ 减少重投影误差。

地图云点 j 在关键帧 i 中的误差是：

$$\mathbf{e}_{i,j} = \mathbf{x}_{i,j} - \pi_i(\mathbf{T}_{iw}, \mathbf{X}_{w,j}) \qquad (5)$$

其中 πi 是影射函数：

$$\pi_i(\mathbf{T}_{iw}, \mathbf{X}_{w,j}) = \begin{bmatrix} f_{i,u}\frac{x_{i,j}}{z_{i,j}} + c_{i,u} \\ f_{i,v}\frac{y_{i,j}}{z_{i,j}} + c_{i,v} \end{bmatrix} \qquad (6)$$

$$\begin{bmatrix} x_{i,j} & y_{i,j} & z_{i,j} \end{bmatrix}^T = \mathbf{R}_{iw}\mathbf{X}_{w,j} + \mathbf{t}_{iw}$$

其中，$R_{iw} \in SO(3)$，$t_{iw} \in R^3$，分别表示 T_{iw} 的旋转和平移部分

（$f_{i,u}$，$f_{i,v}$），（$c_{i,u}$，$c_{i,v}$）分别是相机 i 的焦点距离和主点。

代价函数：

$$C = \sum_{i,j} \rho_h(\mathbf{e}_{i,j}^T \mathbf{\Omega}_{i,j}^{-1} \mathbf{e}_{i,j}) \qquad (7)$$

ρ_h 是 Huber 鲁棒代价函数，$\Omega_{i,j}=\delta_{i,j}^2 I_{2x2}$ 是协方差矩阵，与检测关键点的尺度有关。在全局捆集调整中（在初始化地图中），我们优化了所有云点和关键帧。

参考文献

[1] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment a modern synthesis," in Vision algorithms: theory and practice, 2000, pp. 298–372.

[2] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, 2nd ed. Cambridge University Press, 2004.

[3] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd, "Real time localization and 3d reconstruction," in Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 1, 2006, pp. 363–370.

[4] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR), Nara, Japan, November 2007, pp. 225–234.

[5] D. G´alvez-L´opez and J. D. Tard´os, "Bags of binary words for fast place recognition in image sequences," IEEE Transactions on Robotics, vol. 28, no. 5, pp. 1188–1197, 2012.

[6] H. Strasdat, J. M. M. Montiel, and A. J. Davison, "Scale drift-aware large scale monocular SLAM." in Robotics: Science and Systems (RSS), Zaragoza, Spain, June 2010.

[7] H. Strasdat, A. J. Davison, J. M. M. Montiel, and K. Konolige, "Double window optimisation for constant time visual SLAM," in IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, November 2011, pp. 2352–2359.

[8] C. Mei, G. Sibley, and P. Newman, "Closing loops without places," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Taipei, Taiwan, October 2010, pp. 3738–3744.

[9] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, November 2011, pp. 2564– 2571.

[10] J. Engel, T. Sch¨ops, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in European Conference on Computer Vision (ECCV), Zurich, Switzerland, September 2014, pp. 834–849.

[11] R. Mur-Artal and J. D. Tard´os, "Fast relocalisation and loop closing in keyframe-based SLAM," in IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, June 2014, pp. 846–853.

[12] ——, "ORB-SLAM: Tracking and mapping recognizable features," in MVIGRO Workshop at Robotics Science and Systems (RSS), Berkeley, USA, July 2014.

[13] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, and J. D. Tard´os, "A comparison of loop closing techniques in monocular SLAM,"Robotics and Autonomous Systems, vol. 57, no. 12, pp. 1188–1197, 2009.

[14] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, New York City, USA, June 2006, pp. 2161–2168.

[15] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," The International Journal of Robotics Research, vol. 30, no. 9, pp. 1100–1123, 2011.

[16] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary Robust Independent Elementary Features," in European Conference on Computer Vision (ECCV), Hersonissos, Greece, September 2010, pp. 778–792.

[17] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in European Conference on Computer Vision (ECCV), Graz, Austria, May 2006, pp. 430–443.

[18] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in European Conference on Computer Vision (ECCV), Graz, Austria, May 2006, pp. 404–417.

[19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints,"International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[20] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 1052–1067, 2007.

[21] J. Civera, A. J. Davison, and J. M. M. Montiel, "Inverse depth parametrization for monocular SLAM," IEEE Transactions on Robotics, vol. 24, no. 5, pp. 932–945, 2008.

[22] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in Proc. IEEE Intl. Conf. on Robotics and Automation, Hong Kong, China, June 2014, pp. 15–22.

[23] O. D. Faugeras and F. Lustman, "Motion and structure from motion in a piecewise planar environment," International Journal of Pattern Recognition and Artificial Intelligence, vol. 2, no. 03, pp. 485–508, 1988.

[24] W. Tan, H. Liu, Z. Dong, G. Zhang, and H. Bao, "Robust monocular SLAM in dynamic environments," in IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Adelaide, Australia, October 2013, pp. 209–218.

[25] H. Lim, J. Lim, and H. J. Kim, "Real-time 6-DOF monocular visual SLAM in a large-scale environment," in IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, June 2014, pp. 1532–1539.

[26] D. Nist´er, "An efficient solution to the five-point relative pose problem," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 6, pp. 756–770, 2004.

[27] H. Longuet-Higgins, "The reconstruction of a plane surface from two perspective projections," Proceedings of the Royal Society of London. Series B. Biological Sciences, vol. 227, no. 1249, pp. 399–410, 1986.

[28] P. H. Torr, A. W. Fitzgibbon, and A. Zisserman, "The problem of degeneracy in structure and motion recovery from uncalibrated image sequences," International Journal of Computer Vision, vol. 32, no. 1, pp. 27–44, 1999.

[29] A. Chiuso, P. Favaro, H. Jin, and S. Soatto, "Structure from motion causally integrated over time," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 4, pp. 523–535, 2002.

[30] E. Eade and T. Drummond, "Scalable monocular SLAM," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, New York City, USA, June 2006, pp. 469–476.

[31] H. Strasdat, J. M. M. Montiel, and A. J. Davison, "Visual SLAM: Why filter?" Image and Vision Computing, vol. 30, no. 2, pp. 65–77, 2012.

[32] G. Klein and D. Murray, "Improving the agility of keyframe-based slam," in European Conference on Computer Vision (ECCV), Marseille, France, October 2008, pp. 802–815.

[33] K. Pirker, M. Ruther, and H. Bischof, "CD SLAM-continuous localization and mapping in a dynamic world," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), San Francisco, USA, September 2011, pp. 3990–3997.

[34] S. Song, M. Chandraker, and C. C. Guest, "Parallel, real-time monocular visual odometry," in IEEE International Conference on Robotics and Automation (ICRA), 2013, pp. 4698–4705.

[35] P. F. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," in British Machine Vision Conference (BMVC), Bristol, UK, 2013.

[36] X. Yang and K.-T. Cheng, "LDB: An ultra-fast feature for scalable augmented reality on mobile devices," in IEEE International Symposium on Mixed and Augmented Reality (ISMAR), 2012, pp. 49–57.

[37] R. Kuemmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g2o: A general framework for graph optimization," in IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China, May 2011, pp. 3607–3613.

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

[39] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman, "The new college vision and laser data set," The International Journal of Robotics Research, vol. 28, no. 5, pp. 595–599, 2009. [40] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," The International Journal of Robotics Research, vol. 32, no. 11, pp. 1231–1237, 2013.

[41] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate O(n) solution to the PnP problem," International Journal of Computer Vision, vol. 81, no. 2, pp. 155–166, 2009.

[42] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions," Journal of the Optical Society of America A, vol. 4, no. 4, pp. 629–642, 1987.

[43] F. Endres, J. Hess, J. Sturm, D. Cremers, and W. Burgard, "3-d mapping with an rgb-d camera," IEEE Transactions on Robotics, vol. 30, no. 1, pp. 177–187, 2014.

[44] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, November 2011, pp. 2320–2327.

[45] S. Lovegrove, A. J. Davison, and J. Ibanez-Guzm´an, "Accurate visual odometry from a rear parking camera," in IEEE Intelligent Vehicles Symposium (IV), 2011, pp. 788–793.

[46] P. H. Torr and A. Zisserman, "Feature based methods for structure and motion estimation," in Vision Algorithms: Theory and Practice. Springer, 2000, pp. 278–294.

[47] R. Mur-Artal and J. D. Tardos, "Probabilistic semi-dense mapping from highly accurate feature-based monocular SLAM," in Robotics: Science and Systems (RSS), Rome, Italy, July 2015.

[48] H. Strasdat, "Local Accuracy and Global Consistency for Efficient Visual SLAM," Ph.D. dissertation, Imperial College, London, October 2012.