基于拉普拉斯分布的双目视觉里程计

范涵奇 [1] 吴锦河 [1]

摘 要 针对相机在未知环境中定位及其周围环境地图重建的问题, 本文基于拉普拉斯分布提出了一种快速精确的双目视觉里程计算法. 在使用光流构建数据关联时结合使用三个策略: 平滑的运动约束、环形匹配以及视差一致性检测来剔除错误的关联以提高数据关联的精确性, 并在此基础上筛选稳定的特征点. 本文单独估计相机的旋转与平移. 假设相机旋转、三维空间点以及相机平移的误差都服从拉普拉斯分布, 在此 假设下优化得到最优的相机位姿估计与三维空间点位置. 在 KITTI 和 New Tsukuba 数据集上的实验结果表明, 本文算法能快速精确地估计相机位姿与三维空间点的位置.

关键词 视觉里程计, 运动估计, 光流, 拉普拉斯分布

引用格式 范涵奇, 吴锦河. 基于拉普拉斯分布的双目视觉里程计. 自动化学报, 2022, **48**(3): 865−876

DOI 10.16383/j.aas.c190860

Stereo Visual Odometry Based on Laplace Distribution

FAN Han-Qi[1] WU Jin-He[1]

Abstract In this paper, we present a stereo visual odometry algorithm to estimate the locations of the camera and the surrounding map of unknown environments. The proposed algorithm works fast and yields an accurate trajectory of the camera and environment map. We associate the features of frames by optical flow, and then we select stable features by applying three strategies, i.e. smooth motion constraints, circular matching, and disparity consistency, from the associations. Our algorithm estimates translations and orientations of the camera separately only from the selected stable features. We optimize camera poses and 3D points of the environmental map by assuming the uncertainties of these quantities obey the Laplace distributions which resist outliers and large errors. The experimental results on the KITTI and New Tsukuba datasets show that the proposed algorithm can quickly and accurately estimate the camera pose and 3D environment points.

Key words Visaul odometry (VO), motion estimation, optical flow, Laplace distribution

Citation Fan Han-Qi, Wu Jin-He. Stereo visual odometry based on Laplace distribution. *Acta Automatica Sinica*, 2022, **48**(3): 865−876

视觉里程计 (Visual odometry, VO) 是在未知环境中只通过相机获得的图像信息来实现机器人定位的技术. 近些年来, VO 广泛应用于机器人导航[1]、无人机[2−3] 和虚拟现实/增强现实[4] 等领域.

视觉里程计使用单目或多目相机作为视觉传感器. 单目相机为主要传感器的 VO 系统虽然成本较低, 但面临的主要问题是单目尺度的不确定性, 在计算过程中会发生尺度漂移[5−6]. 该问题通常使用多传感器来解决, 例如雷达和多目相机. 相比于单目相机, 双目相机可以直接测量三维空间点的位置, 避免了尺度的不确定性, 因此本文使用双目相机作为视觉传感器.

VO 系统分为前端和后端[7]. 前端通过特征点匹配构造数据关联来为后端优化提供初始位姿. 数据关联 (Data association) 是指在帧与帧之间的特征点、特征点与地图点以及地图点与地图点之间构建特征对应关系[8]. 数据关联错误是 VO 系统失败的主要原因之一. 基于图像特征点的 VO 系统数据关联方式主要分为两种. 一种是通过计算特征点的描述子来构建数据关联. 采用描述子匹配特征点的准确性与鲁棒性较高, 但是特征描述子的计算非常耗时. Mur-Artal 等[9] 测试尺度不变特征变换 (Scale-invariant feature teansform, SIFT)[10] 和加速健壮特征 (Speed-up robust feature, SURF)[11] 提取耗时约为 300 ms, 像这类比较耗时的特征提取算法会影响 VO 系统的实时性. 为了提高实时性, Mur-Artal 等在 ORB (Oriented fast and rotated brief)-SLAM

收稿日期 2019-12-18 录用日期 2020-07-12
Manuscript received December 18, 2019; accepted July 12, 2020
北京市教育委员会科研计划一般项目 (KM201710009007) 资助
Supported by Beijing Municipal Education Commission Scientific Research Program General Project (KM201710009007)
本文责任编委 吴毅红
Recommended by Associate Editor WU Yi-Hong
1. 北方工业大学信息学院 北京 100144
1. School of Information, North China University of Technology, Beijing 100144

(Simultaneous localization and mapping)[9] 中选择计算速度较快的 ORB 特征点作为图像特征, 帧与帧之间特征点通过特征描述子匹配. Cvišić 等[12] 在 SOFT (Stereo odometry based on feature tracking) 中提取 Corner 角点和 Blob 角点同时计算特征描述子, 并通过在连续帧中追踪同一特征点, 如果该特征能被追踪到则使用初始的描述子来提高数据关联的精确性. 由于相机帧率和图像分辨率越来越高, 导致特征提取的计算量越来越大, 即使使用 ORB 这类速度较快的特征描述子也可能会影响 VO 系统的实时性. 另一种方式只提取角点而不计算描述子, 角点之间的匹配关系通过稀疏的光流 (Optical flow) 跟踪来构建[13−14]. 稀疏的光流算法计算速度快, 但光流容易导致特征点误匹配从而使得数据关联错误, 因此本文使用光流来构建数据关联的同时采用其他技术尽可能剔除错误的数据关联.

VO 系统的后端优化前端估计的相机初始位姿, 本文在后端只优化相机位姿而不维护一个全局地图. 在优化过程中目标函数的设计影响着系统鲁棒性. 在经典的 Bundle adjustment (BA) 和扩展的卡尔曼滤波 (Extended Kalman filter, EKF) 算法中都假设误差服从高斯分布, 优化过程中对噪声敏感, 因而导致位姿估计的误差较大. 与高斯分布相比, 拉普拉斯分布对大噪声不敏感. 同时, 对于长尾数据 (Long tail data) 来说, 拉普拉斯分布比高斯分布更适合对大幅噪音的似然描述[15], 从而对异常点数据更加鲁棒.

本文在 VO 后端假设误差服从拉普拉斯分布, 进而构造优化问题的目标函数. Casafranca 等[16] 在此假设下构造因子图优化问题. Bustos 等[17] 通过构造旋转的平均只优化相机朝向, 在相机朝向已知的情况下得到全局最优的相机位置和三维空间点. 该方法优化相机朝向时不受相机位置与三维空间点的影响因此更简单并且能够处理纯旋转的相机运动, 但是该方法运算速度慢, 并且在求解相机位置时由于同时优化三维空间点和相机位置, 误差较大的三维点会影响相机位置的求解. 与 Casafranca 等[16] 和 Bustos 等[17] 的方法不同, 本文在后端分开优化求解相机朝向、三维空间点以及相机平移, 在此过程中假设相机位姿与三维空间点的误差都服从拉普拉斯分布. SOFT 等多个不同的算法证实分开估计相机的朝向与位置可以提高相机位姿估计的精确性.

近些年来, 基于直接法的双目视觉里程计越来越受研究者的欢迎, 例如, Stereo DSO (Stereo direct sparse odometry)[18]、SVO2 (Semidirect visual odometry 2)[19]和 FMD stereo SLAM (Fusing MVG (multiple view geometry) and direct SLAM)[20]. Stereo DSO 算法通过采用梯度较大的稀疏像素, 使用 Bundle adjustment优化来得到精度较高的相机位姿. 该算法速度较快、鲁棒性好, 并且可以生成稠密的三维点云, 但是对于场景的光照变化比较敏感. SVO2 扩展了半稠密直接法单目视觉里程计 SVO[13], SVO2 算法只对选择的关键帧提取特征点并且采用稀疏图像对齐算法匹配特征点, 因此 SVO2 速度快适用于计算平台受限场合的定位. SVO 采用深度滤波器估计深度, 由于初始深度估计具有较大误差在优化时可能无法收敛到真实深度, 进而影响相机位姿估计. FMD stereo SLAM 方法融合了多视角几何和直接法, 在前端采用直接法估计初始位姿而在后端采用多视角几何的方法来估计三维结构, 这种直接法和多视角几何法融合的方法同时兼顾了速度与精度. 与以上几种方法相比, 本文算法采用特征点法并且引入了误差服从拉普拉斯分布的假设来优化相机位姿.

本文的组织结构如下:第 1 节简要介绍本文算法的框架. 第 2 节中详细阐述特征点的提取以及如何剔除错误的特征匹配并筛选稳定的特征点. 第 3 节为相机位姿的估计与优化. 第 4 节中通过实验验证了本文算法的有效性. 第 5 节为本文的结论.

1 系统概述

本节主要从整体上概述本文提出的算法. 算法主要由数据关联和相机位姿优化估计两部分组成. 数据关联是相机位姿估计的预处理过程. 在构造数据关联时, 特征点的选择与数据关联的准确性影响着相机位姿估计的精度. 在兼顾速度与精度的情况下如何选择稳定的特征点并剔除错误的关联是 VO 算法重要的一步. 位姿估计部分接收筛选的稳定特征点来优化求解相机位姿并重建出稀疏的环境地图.

1.1 特征选择与数据关联

为了提高 VO 算法的实时性, 需要尽可能快地提取每帧图像的特征点, 因此本文选择 FAST (Features from accelerated segment test)[21] 角点作为图像特征点. 数据关联主要通过稀疏光流算法来构建. 虽然稀疏的光流计算速度快, 但是往往会导致特征点的误匹配, 因此本文采用三个策略, 即平滑的运动约束、视差一致性检测以及环形匹配来尽可能剔除错误的关联, 进而提高算法的鲁棒性, 并在此基础上选择稳定的特征点.

1.2　相机位姿优化

本文算法的主要流程如图 1 所示. 其中, $\boldsymbol{P}$ 表示三维空间点的位置, $\boldsymbol{p}$ 表示三维空间点对应的在二维图像上的投影坐标, $\boldsymbol{R}$, $\boldsymbol{t}$ 分别表示相机的旋转和平移. 下标 l, r, k 为左右相机和图像帧的索引. 在相机位姿估计的过程中, 通常对于相机朝向的优化估计是比较困难的, 因此本文首先去优化估计相机的朝向以及三维空间点, 然后固定已经求得的相机朝向和三维空间点来优化求解相机平移. 在光流构建数据关联的过程中, 同一特征点可以连续地在多帧中被跟踪. 本文基于特征点能被连续跟踪的帧数选择不同的参考帧从而获得当前帧相机位姿的多个估计. 对于当前帧的多个相机位姿估计, 在相机位姿误差服从拉普拉斯分布的假设下构造优化问题的目标函数, 进而得到位姿的最优估计.

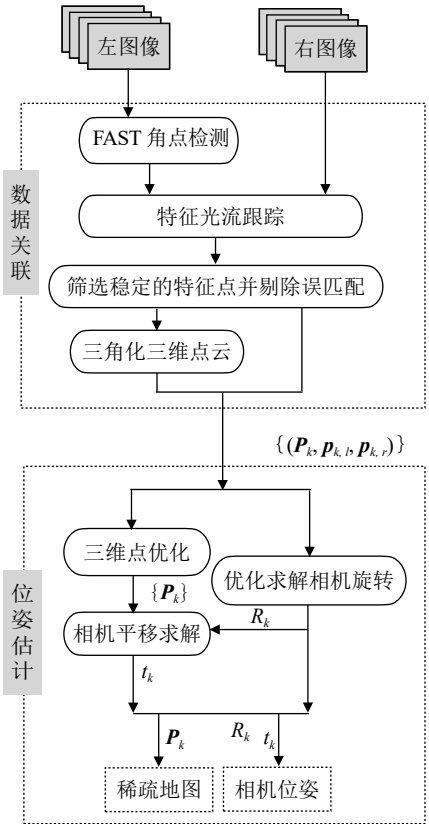


图 1　本文算法的流程图 (算法主要由数据关联与位姿估计两部分组成)

Fig. 1　Overview of the proposed algorithm framework (The algorithm mainly consists two parts: data association and pose estimation)

2　图像特征匹配

光流是三维运动在二维图像平面上的投影. 在使用光流构建数据关联时, 需要剔除错误的关联以提高数据关联的精确性, 并且需要在此基础上筛选稳定的特征点. 如果一个特征点能在连续多帧中跟踪到, 那么该特征点是稳定的.

一般而言, 相机连续帧之间的运动是平滑的, 即前后帧之间的运动量较小, 那么图像帧 I_k 的运动与前一帧 I_{k-1} 的运动是高度相似的, 这种现象称为平滑的运动约束 (Smooth motion constraint, SMC). Badino 等[22] 将该思想应用到三维空间中来剔除三维点的误匹配, 本文应用到二维的图像特征点上以提高光流跟踪的准确性.

由于当前帧的运动与前一帧是相似的, 因此可以使用前一帧的运动去预测特征点在当前帧上的位置. 即

$$\begin{cases} \boldsymbol{P}_k = \boldsymbol{T}_{k-2}^{k-1} \times \boldsymbol{P}_{k-1} \\ \boldsymbol{p}_k = \pi(\boldsymbol{P}_k) \end{cases} \tag{1}$$

式中, $\boldsymbol{P}_k$, $\boldsymbol{P}_{k-1}$ 分别代表同一组三维空间点在当前帧 I_k 和前一帧 I_{k-1} 视角下的位置表示, $\boldsymbol{T}_{k-2}^{k-1}$ 表示已经估计的前一帧的相对运动, 即帧 I_{k-1} 相对于帧 I_{k-2} 的相对位姿变换, $\pi(\cdot)$ 表示投影函数, 将三维空间点投影到二维的图像上, $\boldsymbol{p}_k$ 为三维空间点 $\boldsymbol{P}_k$ 在图像帧 I_k 上的投影.

由于 $\boldsymbol{p}_k$ 更接近于光流收敛的位置, 因此本文使用 $\boldsymbol{p}_k$ 作为光流跟踪过程中预测的前一帧图像特征点在当前帧上的初始位置, 在特征跟踪过程中可以消除一定特征点的误匹配, 从而提供更为可靠的特征点匹配信息.

在平滑的运动约束下单纯使用光流跟踪特征点依然存在误匹配, 为了剔除误匹配增加特征点匹配的正确性, 本文采用强约束环形匹配 (Circle matching)[12]. 图 2 中描述了环形匹配的整个过程. 如图 2 所示, 假设在图像帧 I_{k-1}^l 上检测到某一特征点 $\boldsymbol{x}$, 在连续帧之间的左右图像上使用光流按照 I_{k-1}^l, I_{k-1}^r, I_k^r, I_k^l, I_{k-1}^l 的顺序跟踪特征点 $\boldsymbol{x}$, 跟踪结束时在帧 I_{k-1}^l 上的特征点记为 $\boldsymbol{x}'$. 本文通过计算特征点 $\boldsymbol{x}$ 与 $\boldsymbol{x}'$ 之间的距离 $d(\boldsymbol{x}, \boldsymbol{x}')$ 来判断该环能否闭合. 如果距离 d 不大于一个像素, 则认为该环能闭合, 否则认为该特征点匹配出错, 进而剔除误匹配的特征点.

双目相机系统中, 视差是一个非常重要的量. 在使用以上两种方法剔除特征点的误匹配后, 为了进一步提高特征点匹配的精确性, 通过检测双目相机的视差来再次剔除误匹配. 本文在光流跟踪过程中只计算稀疏特征点之间的视差. 对于已经校正好

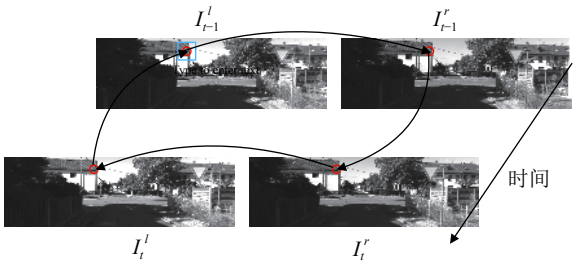


图 2 环形匹配 (使用光流依照箭头所示顺序跟踪特征点)

Fig. 2 Circular matching (Use optical flow to track feature points following the order as arrows direct)

的左右相机的图像帧来说, 左右匹配的特征点一定位于水平的极线 (Epipolar line) 上. 从图像特征点的角度来看, 假设 (x^l, y^l) 与 (x^r, y^r) 是左右相机上一对匹配的特征点, 则视差的 y 分量应等于 0. 即: $y^l = y^r$, 但实际上在图像校正与特征点跟踪过程中由于噪声的存在, 等式不是严格成立的. 因此, 本文采用阈值 th_d 来筛除掉大于阈值的匹配点. 此外, 本文以左边相机为参考系, 则视差在 x 方向上的分量应该为正. 本文使用视差一致性 (Disparity consistency) 来描述这种现象. 即

$$\begin{cases} y^l - y^r \leq th_d \\ x^l - x^r > 0 \end{cases} \tag{2}$$

其中, (x^l, y^l), (x^r, y^r) 为左右图像上匹配的特征点. 如图 3 所示, 使用特征选择策略后能明显地剔除误匹配, 进而提高数据关联的精确性.

为了精确地估计相机位姿, 本文尽可能选择稳定的特征点. 特征点的稳定性通过该特征能被观测到的图像帧数来度量. 一个特征点能在连续帧中被观测到的次数越多, 那么该特征点越稳定. 如图 4 所示, 使用光流构建数据关联时同一特征点可以被连续跟踪多帧. SOFT 使用 age 来描述特征点的稳定性, 本文同样采用 age 来表示一个特征点的稳定性. 以往的几项研究[12, 23] 证实选择稳定的特征点可以提高相机位姿估计的精确性.

3 相机位姿估计

本节主要描述了相机位姿估计与稀疏点云的重建. 本文将位姿估计分为两部分. 首先, 根据帧与帧之间二维的匹配点直接估计相机的相对旋转; 然后固定已知的相机旋转来估计平移.

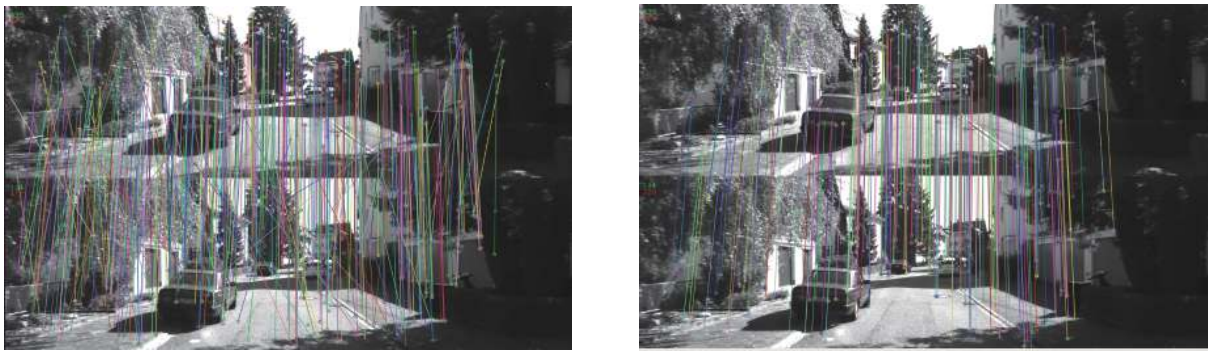
3.1 拉普拉斯分布假设下目标函数的构造

一般而言, SLAM 可以表示为最大后验估计 (Maximum a posteriori estimation, MAP) 问题. 假定要估计的未知变量为 $\boldsymbol{X}$, 其中, $\boldsymbol{X}$ 包括相机位姿和环境中路标点的位置. 由观测方程 $z_j = h_j(\boldsymbol{X}_j) + \epsilon_j$ 可得到一个观测集合 $\boldsymbol{Z} = \{z_j : j = 1, 2, \cdots, m\}$, 其中, $\boldsymbol{X}_j \subseteq \boldsymbol{X}$ 是估计变量的子集,$h_j(\cdot)$ 为观测模型,ϵ_j 为观测的噪音.

在 MAP 估计中, 通过最大化后验概率来估计变量 $\boldsymbol{X}$, 即

$$\boldsymbol{X}^* \approx \arg\max_{\boldsymbol{X}} p(\boldsymbol{X}|\boldsymbol{Z}) = \arg\max_{\boldsymbol{X}} p(\boldsymbol{Z}|\boldsymbol{X}) p(\boldsymbol{X}) \tag{3}$$

其中, $p(\boldsymbol{Z}|\boldsymbol{X})$ 为似然, $p(\boldsymbol{X})$ 为先验. 在没有先验信息的情况下, $p(\boldsymbol{X})$ 是一个常量, MAP 退化为最大似然估计 (Maximum likelihood estimation, MLE). 假定观测的噪音互不相关, 即观测 $\boldsymbol{Z}$ 相互独立, 则式 (3) 分解为



(a) 单纯使用光流匹配特征点的结果
(a) Feature matching only used optical flow

(b) 采用特征点选择策略剔除误匹配后的结果
(b) Correct matchings after applying three stable feature constraints

图 3 特征点匹配的对比

Fig. 3 Comparison of feature matching

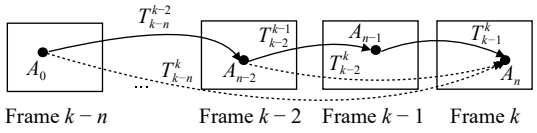


图 4　光流特征跟踪与 *age* 说明 (同一特征点可以在连续帧中被跟踪, *age* 值越大该特征点越稳定, T 表示两帧之间的位姿变换)

Fig. 4　Optical flow feature tracking and *age* description (The same feature point can be tracked in consecutive frames. The larger the *age*, the more stable the feature point, T represents the pose transformation between two frames)

$$\boldsymbol{X}^* = \arg\max_{\boldsymbol{X}} p(\boldsymbol{X}) \prod_{j=1}^{m} p(z_j|\boldsymbol{X}) = \arg\max_{\boldsymbol{X}} p(\boldsymbol{X}) \prod_{j=1}^{m} p(z_j|\boldsymbol{X}_j) \tag{4}$$

假设观测的噪音 ϵ_j 服从具有信息矩阵 $\boldsymbol{\Omega}_j$ 的零均值拉普拉斯分布, 则式 (4) 中的似然为

$$p(z_j|\boldsymbol{X}_j) \propto \exp(-\|\boldsymbol{\Omega}_j(z_j - h_j(\boldsymbol{X}_j))\|_1) \tag{5}$$

由于最大后验估计等价于最小化后验的负对数, 则有

$$\boldsymbol{X}^* = \arg\min_{\boldsymbol{X}} -\lg\left(p(\boldsymbol{X}) \prod_{j=1}^{m} p(z_j|\boldsymbol{X}_j)\right) = \arg\min_{\boldsymbol{X}} \|\boldsymbol{\Omega}_j(z_j - h_j(\boldsymbol{X}_j))\|_1 \tag{6}$$

由式 (6) 可知, 在噪音服从拉普拉斯分布的假设下构造了以 L_1 范数来度量误差的非线性优化问题. 因此在相机位姿与三维点云优化时, 本文在拉普拉斯分布的假设下采用 L_1 范数来度量误差.

传统上在高斯分布假设下采用 L_2 范数作为目标函数来优化估计相机位姿. 尽管 L_2 范数应用广泛, 但是拉普拉斯分布假设下的 L_1 范数提供了另一种选择. Moreno 等[24] 详细对比了 L_1 范数和 L_2 范数在不同噪音水平下的表现, 证实了 L_1 范数对较大的异常值更加鲁棒.

为了提高 L_2 范数对较大误差的鲁棒性, 通常结合鲁棒的核函数 (例如 Huber) 来减小误差较大的数据对优化结果的影响. Casafranca 等[25] 在后端采用 L_1 范数度量误差实现了因子图 SLAM, 并在实验中证实了 L_1 范数在消除大噪音数据影响方面比 Huber 核更加鲁棒. 同时相比于 Huber 核函数, L_1 范数在优化时并不需要调整任何内核参数. Bahreinian 等[26] 研究了在位姿图优化中通过最小化 L_1 范数能有效剔除离群点的影响. 综上, 尽管与鲁棒 Huber 核函数类似, 但是 L_1 范数在消除大噪音数据的影响方面仍然更加有效, 即 L_1 范数对大噪音数据优化的鲁棒性更好.

3.2　旋转的估计

不同于一般的双目 VO 算法只使用左侧图像信息来估计相机位姿, 为了尽可能多地估计当前帧的旋转, 本文使用左右相机的图像信息来分别估计当前帧的旋转.

如图 4 所示, 本文根据当前帧上特征点的 *age* 值选择不同的参考帧来估计当前帧的相机朝向. 对于首次提取的某一特征点的 *age* 值记为 0, 如果该特征点能跟踪到下一帧则 *age* 值加 1, 以此类推, *age* 值越大说明该特征点能被连续跟踪的帧数越多. 如果当前帧中某一特征点的 *age* 值为 n, 那么说明该特征点可以在前 n 帧中被跟踪到, 因此本文选择前 n 个图像帧分别作为参考帧来估计当前帧相机的朝向. 对于当前帧 I_k 相机朝向的 n 个不同估计, 使用集合 $\{{}_1\boldsymbol{R}_k, {}_2\boldsymbol{R}_k, \cdots, {}_n\boldsymbol{R}_k\}$ 来表示, 其中 $\boldsymbol{R}_k$ 代表当前帧 I_k 相机的绝对朝向. 为了得到相机朝向的最优估计, 假设噪声服从拉普拉斯分布, 则有

$$\boldsymbol{S}^* = \arg\min_{\boldsymbol{S}\in SO(3)} \sum_{i=1}^{n} \|\boldsymbol{R}_i - \hat{\boldsymbol{S}}\|_1 \tag{7}$$

其中, $\boldsymbol{R}_i \in \{{}_1\boldsymbol{R}_k, {}_2\boldsymbol{R}_k, \cdots, {}_{n-1}\boldsymbol{R}_k, {}_n\boldsymbol{R}_k\}$, $\hat{\boldsymbol{S}}$ 为相机旋转. 通过采用 $SO(3)$ 上的迭代加权最小二乘 (Iterative reweighted least squares, IRLS) 算法获得相机朝向的最优估计 $\boldsymbol{S}^*$. 算法 1 描述了 $SO(3)$ 上的 IRLS 算法.

算法 1. $SO(3)$ 上的 IRLS 算法

输入. 集合 $\{{}_1R_k, {}_2R_k, \cdots, {}_{n-1}R_k, {}_nR_k\}$

输出. 当前帧 I_k 的旋转 R_k

1) Set $R \leftarrow {}_1R_k$, Choose $\varepsilon > 0$
2) **loop**
3)　Compute $\Delta_R \leftarrow \frac{1}{n}\sum_{i=1}^{n} \lg({}_iR_k R^{\mathrm{T}})$
4)　**if** $\|\Delta_R\| < \varepsilon$ **then**
5)　　Set $S_0 \leftarrow R$
6)　　break
7)　**else**
8)　　$R \leftarrow R \times \exp(\Delta_R)$
9)　**end if**
10) **end loop**
11) Choose $\epsilon > 0$, $j = 0, 1, \cdots$
12) **loop**

13) Compute $e_i \leftarrow \lg_{S_j}({}_iR_k)$
14) Set $\delta \leftarrow \dfrac{\sum_{i=0}^{n} e_i/\|e_i\|}{\sum_{i=0}^{n} 1/\|e_i\|}$
15) **if** $\delta < \epsilon$ **then**
16) Set $R_k \leftarrow S_j$
17) return R_k
18) **else**
19) $S_{j+1} \leftarrow \exp(\delta) \times S_j$
20) **end if**
21) **end loop**

为了估计当前帧 I_k 的绝对旋转 $\boldsymbol{R}_k$, 本文首先估计选择的参考帧 I_{k-i} 与当前帧 I_k 之间的相对旋转 $\boldsymbol{R}_{k-i}^{k}$. 由于参考帧 I_{k-i} 的绝对相机旋转 $\boldsymbol{R}_{k-i}$ 已经求得, 因此可以得到当前帧 I_k 绝对的相机旋转, 即: ${}_i\boldsymbol{R}_k = \boldsymbol{R}_{k-i} \times \boldsymbol{R}_{k-i}^{k}$, $i \in \{1, \cdots, n\}$.

对于参考帧与当前帧之间的相对旋转, 传统上应用对极约束根据帧与帧之间二维特征点的匹配关系, 通过分解本质矩阵 (Essential matrix)[27] 来恢复, 但是对于纯旋转的相机运动来说, 无法从本质矩阵中恢复相机的运动. 针对这个问题, Fathian 等[28] 使用四元数来表示相机的相对旋转, 提出了 QuEst 算法来从二维的匹配点中估计相机的运动. 该算法需要 5 对点来求解相机旋转, 但通常匹配特征点的数量多于 5 对, 因此本文采用 QuEst 结合随机采样一致性算法 (Random sample consensus, RANSAC)[29] 来精确地估计两个相机视角之间相对的旋转变换.

3.3 稀疏点云重建

重建的三维空间点的精确性影响着相机位置的估计. 由第 3.2 节, 在相机运动的过程中基于特征点的 *age* 值, 可以在前 n 帧中多次观测到同一三维空间点. 如图 5 所示, 首先根据双目相机的视差可以三角化得到同一三维空间点的 n 个位置表示. 然后由于前 n 帧的相机位姿已知, 因此可以使用左右相机前后帧之间特征点的匹配关系再次三角化得到 $2n-2$ 个同一三维空间的位置, 即三角化得到同一三维空间点的 $3n-2$ 个位置. 本文使用集合 $\{{}_1\boldsymbol{P}, {}_2\boldsymbol{P}, \cdots, {}_{3n-2}\boldsymbol{P}\}$ 表示同一空间点的多次观测. 为了得到最优三维空间点的位置, 本文将三维空间点视为欧氏空间 $\mathbf{R}^3$ 中的三维向量, 在同一三维空间点的多次观测中得到该三维空间点的最优位置, 即

$$\boldsymbol{P}^* = \arg\min_{\hat{\boldsymbol{P}} \in \mathbf{R}^3} \sum_{i=1}^{3n-2} \|\boldsymbol{P}_i - \hat{\boldsymbol{P}}\|_1 \tag{8}$$

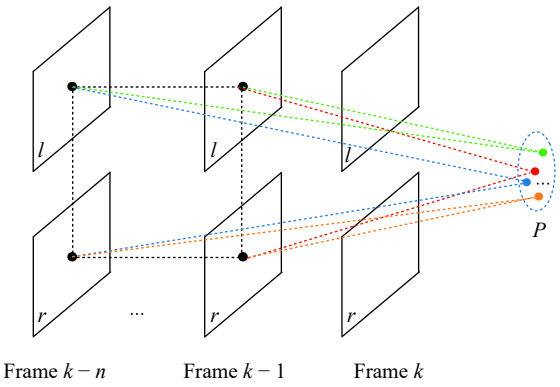


图 5 同一三维空间点的多次三角化 (基于特征点的 *age* 值, 在前 n 帧中根据双目相机的视差以及左右相机连续帧间特征点的匹配关系多次三角化同一三维空间点)

Fig. 5 Multiple triangulations of the same 3D space point (Based on the *age* of the features, the same 3D space point is triangulated multiple times in the first n frames according to the disparity of the stereo camera and the matching relationship between the feature points of the left and right camera consecutive frames)

其中, $\boldsymbol{P}_i \in \{{}_1\boldsymbol{P}, {}_2\boldsymbol{P}, \cdots, {}_{3n-2}\boldsymbol{P}\}$. 通过采用 $\mathbf{R}^3$ 上的 IRLS 算法优化三维空间点, 得到三维空间点的最优位置 $\boldsymbol{P}^*$.

3.4 平移的估计

本文使用优化后的三维点云与二维点的匹配关系来求相机的平移. 三维空间点投影到二维图像平面的整个过程为

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \approx \pi(\boldsymbol{P}; \boldsymbol{R}, \boldsymbol{t}) = \begin{bmatrix} f & 0 & \alpha_x \\ 0 & f & \alpha_y \\ 0 & 0 & 1 \end{bmatrix} [\boldsymbol{R}|\boldsymbol{t}] \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{9}$$

其中, $[u, v, 1]^{\mathrm{T}}$ 为投影点的齐次坐标表示, f 为相机焦距, 投影函数用 π 来表示, (α_x, α_y) 为图像的主点, $[\boldsymbol{R}|\boldsymbol{t}]$ 为相机两个视角之间的运动. $[x, y, z, 1]^{\mathrm{T}}$ 为三维点云的齐次坐标表示.

本文固定已经估计的旋转和优化后的三维空间点云来优化估计相机的平移, 即

$$\boldsymbol{t}^* = \arg\min_{\hat{\boldsymbol{t}}} \sum_{i=1}^{n} \sum_{c \in \{l, r\}} \|{}^c\boldsymbol{p}_i - {}^c\pi(\boldsymbol{P}, \boldsymbol{R}, \hat{\boldsymbol{t}})\|_1 \tag{10}$$

其中, l, r 表示左右相机. 整个过程中 IRLS 算法只迭代优化相机的平移, 通常仅需迭代少数几次即可收敛. 与 $SO(3)$ 上的 IRLS 算法相比, 对于 $\mathbf{R}^3$ 上的 IRLS 算法, 只在于计算误差以及迭代更新的方式

不同, $\mathbf{R}^3$ 上直接通过两个向量之差来定义误差.

4　实验结果与分析

本文在 KITTI 数据集[30] 以及 New Tsukuba 数据集[31] 上测试提出的算法. 实验环境为装有 Ubuntu 18.04, 配置为 Intel® Core™ i7-4770 CPU, 8 GB RAM 的计算机. KITTI 数据集提供了大尺度动态场景下的 22 个立体图像序列. 其中前 11 序列 (00 ~ 10) 提供了真实轨迹. New Tsukuba 数据集为人工合成的小场景的静态室内环境.

本文主要在精度与速度两个方面来评估算法. 在精度方面, 主要通过评估估计轨迹与真实轨迹之间的均方根误差 (Root mean square error, RMSE)、均值 (Mean) 与标准差 (Standard deviation, STD) 来反映本文算法的精确性, 其中均方根误差代表绝对轨迹误差, 均方根误差越小则位姿估计越准确. 本文使用光流跟踪特征点选择的窗口块大小为 9×9, 视差一致性检测的阈值 $th_d = 1$, 验证是否构成回环匹配的阈值 $d \leq 1$.

由于 VISO2-S[32] 与 ORB-SLAM2[33] 的作者开放了源代码, 因此本文选择这两种算法来与本文算法 (算法 1) 进行对比. 表 1 中对比了本文算法与 ORB-SLAM2在 KITTI 数据集 00 ~ 10 序列共 11 个序列上的测试结果. 从表 1 中可知, 除了 01 序列外, 本文算法在其余的 10 组序列上位姿估计的结果较精确. 01 序列采集的是高速公路的场景, 由于场景纹理较少并且场景纹理相似, 导致提取的特征点分布不均匀, 且大部分集中在中间的一条线上, 只使用图像特征点在此场景中估计相机位姿是困难的, 其他算法在此序列上也存在同样的问题, 导致位姿估计的误差较大. 相比于 ORB-SLAM2, 本文算法的 RMSE 在多数序列上更小. 由于在 05 和 06 序列上有闭环, ORB-SLAM2 通过闭环检测与校正来进一步提高位姿估计的准确性, 而本文算法没有回环校正. 虽然在 05 和 06 序列上 ORB-SLAM2 的 RMSE 较小, 但是本文算法的 RMSE 非常接近 ORB-SLAM2.

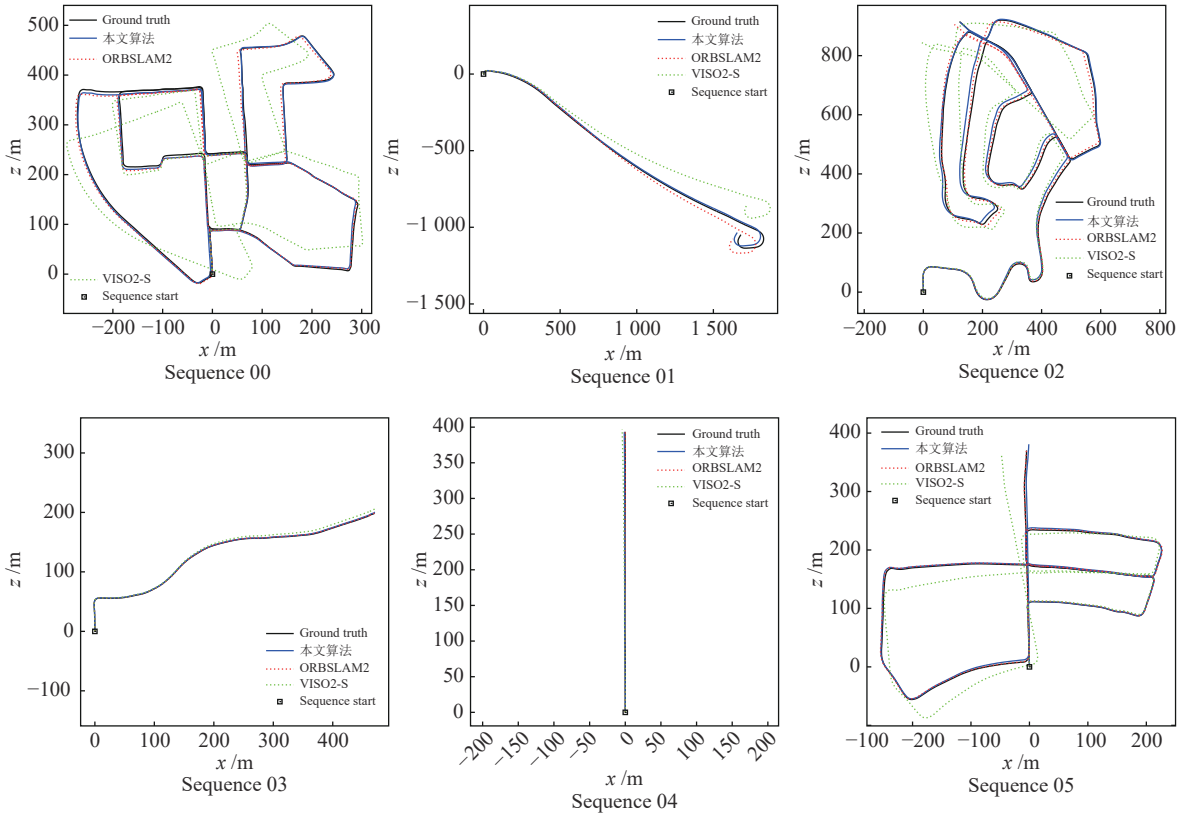
表 1 中同时对比了本文算法与 VISO2-S 在 KITTI 数据集 00 ~ 10 序列上估计的相机运动轨迹与真实轨迹的 RMSE、Mean 和 STD. VISO2-S 算法通过最小化重投影误差并采用高斯牛顿法迭代优化求解相机的旋转与平移. 通常对于相机的旋转优化是比较困难的, 相机旋转影响着相机平移的估计, 因此采用迭代法同时优化求解得到相机旋转与平移的误差较大. 本文算法通过分开独立的优化相机的旋转与平移提高来位姿估计的精确性, 从表中数据可知本文算法优于 VISO2-S 算法.

表 2 中对比了不同算法在大场景的 KITTI 数据集上的运行时间, 部分数据来源于文献 [34] 与 KITTI Benchmark[30]. 本文只统计特征处理与位姿优化估计所耗费的时间. 从表中数据可知, 由于 ORB-SLAM2 采用局部 BA 与全局 BA 来优化相机位姿, 因此 ORB-SLAM2 位姿优化估计部分耗时较多. 由于 VISO2-S 对左右每幅图像提取 Coner 角点和 Blob 角点并计算描述子, 通过描述子来构建图像帧之间的数据关联, 最后通过最小化重投影误差求解相机位姿, 因此 VISO2-S 主要耗时在特征处理部分. FB-KLT 方法提取 Shi-Tomasi 角点并

表 1　本文算法、ORB-SLAM2 以及 VISO2-S 估计的轨迹与真实轨迹之间 RMSE、Mean、STD 的对比

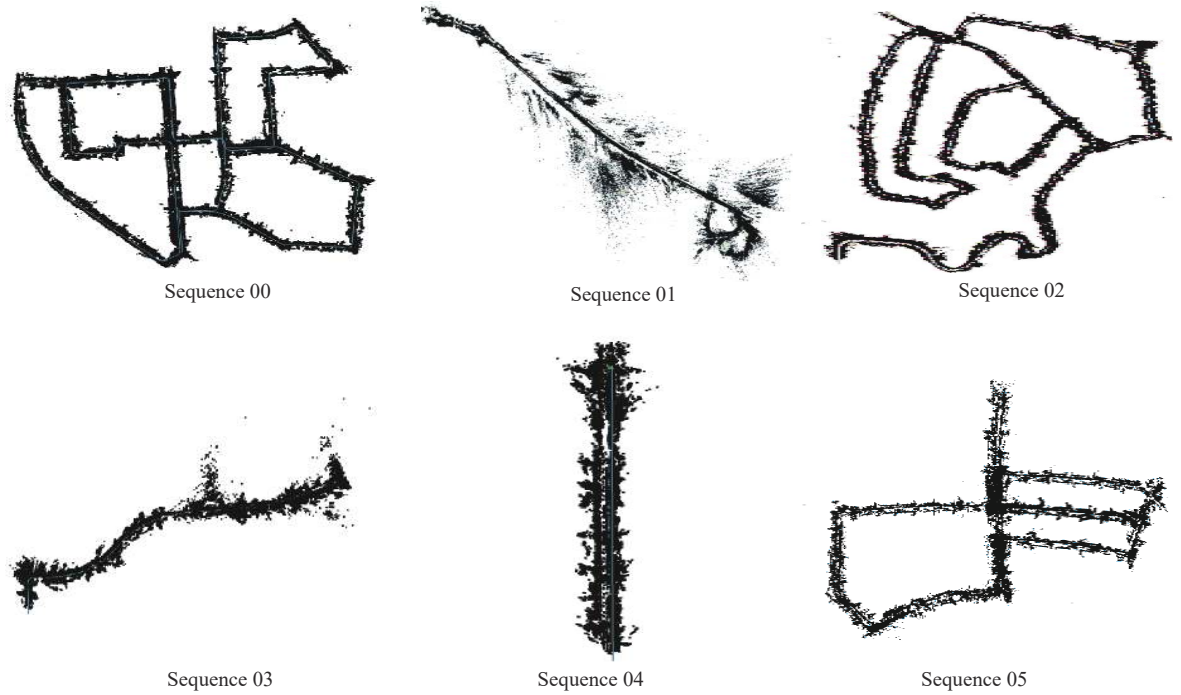
Table 1　Comparison of RMSE, Mean, STD between the trajectory estimated by ours, ORB-SLAM2, and VISO2-S and the real trajectory

序列	RMSE (m)			Mean (m)			STD (m)		
	本文算法	ORB-SLAM2	VISO2-S	本文算法	ORB-SLAM2	VISO2-S	本文算法	ORB-SLAM2	VISO2-S
00	**5.248**	7.410	32.119	**4.696**	6.733	27.761	**2.343**	3.095	16.153
01	**33.938**	38.426	132.138	**28.257**	29.988	105.667	**18.797**	24.027	79.341
02	**11.365**	13.081	34.759	**10.332**	11.300	31.594	**4.733**	6.589	14.491
03	**1.031**	1.662	1.841	**0.909**	1.486	1.672	**0.486**	0.745	0.771
04	**0.495**	0.529	0.975	**0.426**	0.487	0.861	**0.207**	0.253	0.457
05	4.207	**1.569**	12.437	3.061	**1.421**	10.561	2.885	**0.664**	6.567
06	2.839	**2.059**	7.758	2.538	**1.759**	6.941	1.272	**1.072**	4.245
07	3.655	**1.903**	12.277	3.079	**1.813**	9.399	1.971	**1.393**	7.898
08	**13.001**	13.112	20.645	**12.555**	12.853	18.786	**2.594**	3.376	8.562
09	**4.668**	6.081	19.491	**3.561**	5.212	15.326	**3.018**	3.312	12.041
10	**2.817**	4.811	11.789	**2.628**	4.958	8.074	**1.013**	2.594	8.589



(a) 在数据集序列上估计的相机运动轨迹与真实轨迹的对比

(a) Comparison between estimated camera motion trajectory and real trajectory on the sequences

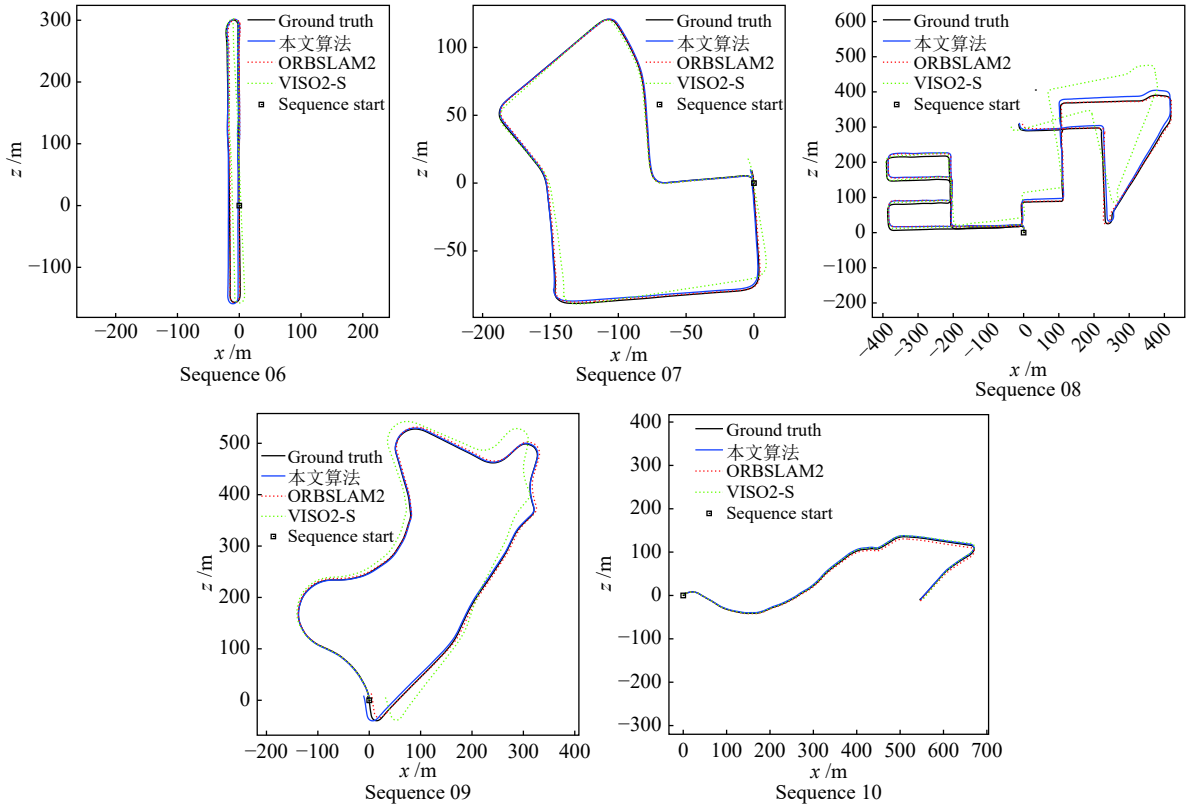


(b) 本文算法重建的稀疏环境地图

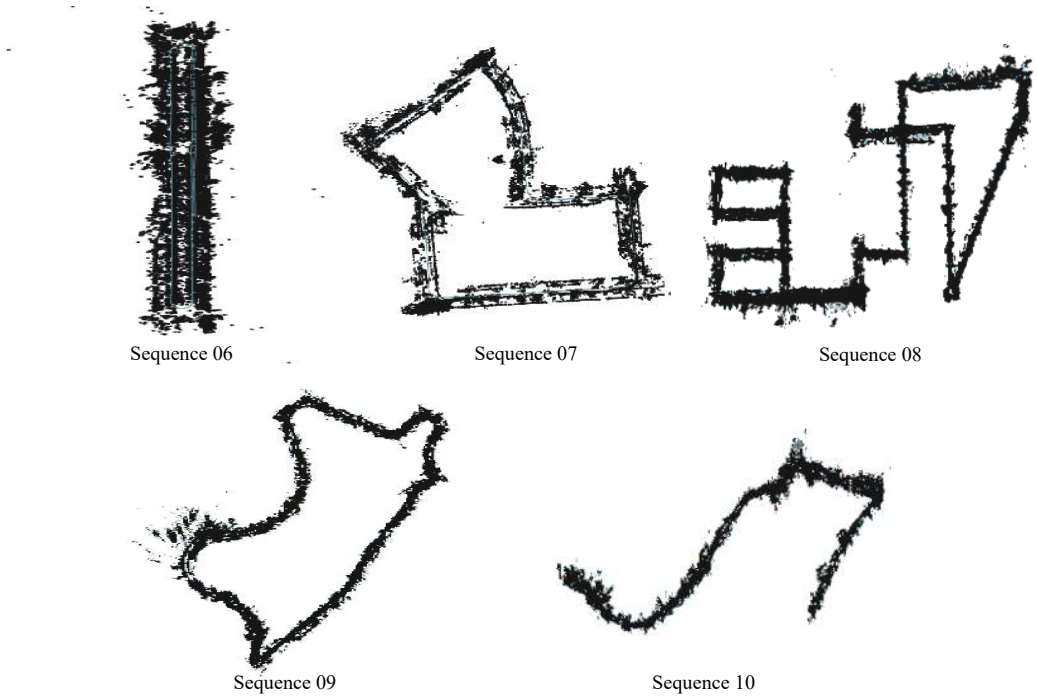
(b) Reconstructed sparse environment map by ours algorithm

图 6 在 KITTI 数据集序列 00 ~ 05 上的实验

Fig. 6 Experiments on the KITTI sequence 00 ~ 05



(a) 在数据集序列上估计的相机运动轨迹与真实轨迹的对比
(a) Comparison between estimated camera motion trajectory and real trajectory on the sequences



(b) 本文算法重建的稀疏环境地图
(b) Reconstructed sparse environment map by ours algorithm

图 7　在 KITTI 数据集序列 06 ~ 10 上的实验

Fig. 7　Experiments on the KITTI sequence 06 ~ 10

表 2 计算时间的对比 (部分数据摘自 KITTI Benchmark[30]) (ms)

Table 2 Computation time comparison (Partial data from KITTI Benchmark[30]) (ms)

方法	特征处理	运动估计	总耗时	计算平台
orbslam2[33]	11.4	109.2	120.6	3.5 GHz (1核)
VISO2-S[32]	34.5	3.3	37.8	3.5 GHz (1核)
FB-KLT[34]	40.8	1.1	41.9	3.5 GHz (1核)
本文	10.73	18.88	29.61	3.5 GHz (1核)

通过高斯牛顿法求解相机位姿, 从表中实验数据看出该方法主要耗时在特征处理部分. 本文算法在特征处理部分速度最快, 耗时主要在位姿优化估计部分. 总体而言, 从表 2 中可以看出本文算法耗时最少.

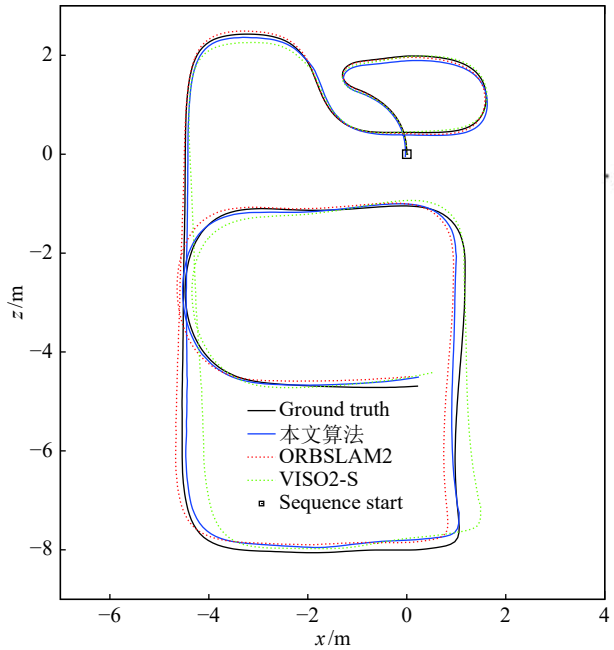
在室内场景的 New Tsukuba 数据集上, 本文算法估计的 RMSE 为 0.389, ORB-SLAM2 估计的 RMSE 为 1.015, VISO2-S 估计的 RMSE 为 0.916, 可以看出本文算法表现较好.

在图 6 (a) 和图 7 (a) 中, 给出了本文算法、ORB-SLAM2 和 VISO2-S 估计的相机运动轨迹与真实相机轨迹在 KITTI 数据集上的对比. 而图 6 (b) 和图 7 (b) 为本文算法在 KITTI 数据集上重建的周围环境的稀疏地图.

在图 8(a) 中, 给出了 New Tsukuba 数据集上本文算法、ORB-SLAM2 以及 VISO2-S 估计的轨迹与真实轨迹的对比. 从轨迹对比图中可以看出, 本文算法估计的轨迹更接近于真实轨迹. 图 8(b) 为本文算法在 New Tsukuba 数据集上重建的稀疏环境地图.

5 结束语

本文基于拉普拉斯分布提出了一种快速精确的双目视觉里程计算法. 为了提高相机位姿与三维空间点估计的精确性, 在相机旋转、三维空间点以及相机平移这些量的误差服从拉普拉斯分布的假设下来构造优化问题的目标函数. 通过在 $SO(3)$ 以及 $\mathbf{R}^3$ 上的优化来得到位姿和三维空间点的最优估计. KITTI 以及 New Tsukuba 数据集上的实验结果表明本文提出的算法能够快速准确地估计相机位姿与三维空间点的位置. 本文算法尚有一定的局限性, 在构建数据关联时只使用了光流跟踪, 由于光流自身的原因, 对于亮度变化明显的环境不够鲁棒. 未来的工作中, 将进一步研究提高光流跟踪的精确性并尝试将视觉信息与惯性测量结合起来, 增加系统的鲁棒性.



(a) 估计的相机运动轨迹与真实轨迹的对比
(a) Comparison between estimated camera motion trajectory and real trajectory

(b) 本文算法重建的稀疏环境地图
(b) Reconstructed sparse environment map by ours algorithm

图 8 在 New Tsukuba 数据集上的实验

Fig. 8 Experiments on New Tsukuba sequence

References

1 Lin Y, Gao F, Qin T, Gao W L, Liu T B, Wu W, Yang Z F, Shen S J. Autonomous aerial navigation using monocular visual-inertial fusion. *Journal of Field Robotics*, 2018, **35**(1): 23−51

2 Faessler M, Mueggler E, Schwabe K, Scaramuzza D. A monocular pose estimation system based on infrared LEDs. In: Proceedings of the 2014 International Conference on Robotics and Automation. Hong Kong, China: IEEE, 2014. 907−913

3 Meier L, Tanskanen P, Heng L, Lee G H, Fraundorfer F, Pollefeys M. PIXHAWK: A micro aerial vehicle design for autonomous flight using onboard computer vision. *Autonomous Robots*, 2012, **33**(1): 21−39

4 Klein G, Murray D W. Parallel tracking and mapping for small AR workspaces. In: Proceedings of the 6th International Symposium on Mixed and Augmented Reality. Nara, Japan: IEEE/ACM, 2007. 1−10

5 Strasdat H, Montiel J, Davison A J. Scale drift-aware large scale monocular SLAM. *Robotics: Science and Systems*, 2010, **2**(3): 27−34

6 Engel J, Schops T, Cremers D. LSD-SLAM: Large-scale direct monocular SLAM. In: Proceedings of the 13th European Conference on Computer Vision. Zürich, Switzerland: Springer, 2014. 834−849

7 Cadena C, Carlone L, Carrillo H, Latif Y, Scaramuzza D, Neira J, Reid I, Leonard J J. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 2016, **32**(6): 1309−1332

8 Ding Wen-Dong, Xu De, Liu Xi-Long, Zhang Da-Peng, Chen Tian. Review on visual odometry for mobilerobots. *Acta Automatica Sinica*, 2018, **44**(3): 385−400
(丁文东, 徐德, 刘希龙, 张大朋, 陈天. 移动机器人视觉里程计综述. 自动化学报, 2018, **44**(3): 385−400)

9 Mur-Artal R, Montiel J M, Tardos J D. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 2015, **31**(5): 1147−1163

10 Lowe D G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, **60**(2): 91−110

11 Bay H, Tuytelaars T, Van Gool L. SURF: Speeded up robust features. In: Proceedings of the 9th European Conference on Computer Vision. Graz, Austria: Springer, 2006. 404−417

12 Cvisic I, Petrovic I. Stereo odometry based on careful feature selection and tracking. In: Proceedings of the 7th European Conference on Mobile Robots. Lincoln, Lincolnshire, United Kingdom: IEEE, 2015. 1−6

13 Forster C, Pizzoli M, Scaramuzza D. SVO: Fast semi-direct monocular visual odometry. In: Proceedings of the 2014 International Conference on Robotics and Automation. Hong Kong, China: IEEE, 2014. 15−22

14 More V, Kumar H, Kaingade S, Gaidhani P, Gupta N. Visual odometry using optic flow for unmanned aerial vehicles. In: Proceedings of the 1st International Conference on Cognitive Computing and Information Processing. Noida, India: IEEE, 2015. 1−6

15 Jing L P, Wang P, Yang L. Sparse probabilistic matrix factorization by laplace distribution for collaborative filtering. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina: AAAI Press 2015. 25−31

16 Casafranca J J, Paz L M, Pinies P. A back-end L1 norm based solution for factor graph SLAM. In: Proceedings of the 2013 International Conference on Intelligent Robots and Systems. Tokyo, Japan: IEEE/RJS, 2013. 17−23

17 Bustos A P, Chin T, Eriksson A, Reid I. Visual SLAM: Why bundle adjust? In: Proceedings of the 2019 International Conference on Robotics and Automation. Montreal, Canada: IEEE, 2019. 2385−2391

18 Wang R, Schworer M, Cremers D. Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras. In: Proceedings of the 2017 International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 3903−3911

19 Forster C, Zhang Z C, Gassner M, Werlberger M, Scaramuzza D. SVO: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics*, 2017, **33**(2): 249−265

20 Tang F L, Li H P, Wu Y H. FMD stereo SLAM: Fusing MVG and direct formulation towards accurate and fast stereo SLAM. In: Proceedings of the 2019 International Conference on Robotics and Automation. Montreal, Canada: IEEE, 2019. 133−139

21 Rosten E, Drummond T. Machine learning for high-speed corner detection. In: Proceedings of the 9th European Conference on Computer Vision. Graz, Austria: Springer, 2006. 430−443

22 Badino Hernán. A robust approach for ego-motion estimation using a mobile stereo platform. In: Proceedings of the 1st International Workshop on Complex Motion. Günzburg, Germany: Springer, Berlin, Heidelberg, 2004. 198−208

23 Fan H Q, Zhang S. Stereo odometry based on careful frame selection. In: Proceedings of the 10th International Symposium on Computational Intelligence and Design. Hangzhou, China: IEEE, 2017. 177−180

24 Moreno L, Blanco D, Muñoz M L, Garrido S. L1-L2 norm comparison in global localization of mobile robots. *Robotics and Autonomous Systems*, 2011, **59**(9): 597−610

25 Casafranca J J, Paz L M, Pinies P. L1 factor graph SLAM: Going beyond the L2 norm. In: Proceedings of the 2013 Robust and Multimodal Inference in Factor Graphs Workshop, IEEE International Conference on Robots and Automation. Karlsruhe, Germany: IEEE, 2013. 17−23

26 Bahreinian M, Tron R. A computational theory of robust localization verifiability in the presence of pure outlier measurements. In: Proceedings of the 58th Conference on Decision and Control. Nice, France: IEEE, 2019. 7824−7831

27 Nister D. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, **26**(6): 756−777

28 Fathian K, Ramirez-Paredes J P, Doucette E A, Curtis J W,

Gans N R. QuEst: A quaternion-based approach for camera motion estimation from minimal feature points. *International Conference on Robotics and Automation*, 2018, **3**(2): 857−864

29 Fischler M A, Bolles R C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981, **24**(6): 381−395

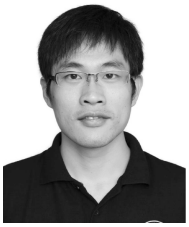
30 Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 2013, **32**(11): 1231−1237

31 Martull S, Peris M, Fukui K. Realistic CG stereo image dataset with ground truth disparity maps. *Scientific Programming*, 2012, **111**(431): 117−118

32 Geiger A, Ziegler J, Stiller C. StereoScan: Dense 3D reconstruction in real-time. In: Proceedings of the 2011 IEEE Intelligent Vehicles Symposium. Baden-Baden, Germany: IEEE, 2011. 963−968

33 Murartal R, Tardos J D. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 2017, **33**(5): 1255−1262

34 Wu M, Lam S, Srikanthan T. A framework for fast and robust visual odometry. *IEEE Transactions on Intelligent Transportation Systems*, 2017, **18**(12): 3433−3448



范涵奇 北方工业大学信息学院副教授. 2011 年于浙江大学 CAD & CG 国家重点实验室获得博士学位. 主要研究方向为计算机视觉与视觉 SLAM. 本文通信作者.
E-mail: fhq@ncut.edu.cn
(**FAN Han-Qi** Associate professor at the School of Information, North China University of Technology. He received his Ph.D. degree in computer science from the CAD & CG State Key Laboratory of Zhejiang University in 2011. His research interest covers computer vision and visual SLAM. Corresponding author of this paper.)

吴锦河 北方工业大学信息学院硕士研究生. 主要研究方向为视觉 SLAM.
E-mail: jhe_wu@163.com
(**WU Jin-He** Master student at the School of Information, North China University of Technology. His main research interest is visual SLAM.)