移动机器人视觉里程计综述

丁文东 [1, 2] 徐 德 [1, 2, 3] 刘希龙 [1, 2] 张大朋 [1, 2] 陈 天 [1, 2]

摘 要 定位是移动机器人导航的重要组成部分. 在定位问题中, 视觉发挥了越来越重要的作用. 本文首先给出了视觉定位的数学描述, 然后按照数据关联方式的不同介绍了视觉里程计 (Visual odometry, VO) 所使用的较为代表性方法, 讨论了提高视觉里程计鲁棒性的方法. 此外, 本文讨论了语义分析在视觉定位中作用以及如何使用深度学习神经网络进行视觉定位的问题. 最后, 本文简述了视觉定位目前存在的问题和未来的发展方向.

关键词 视觉里程计, 视觉定位, 位姿估计, 导航, 移动机器人

引用格式 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

DOI 10.16383/j.aas.2018.c170107

Review on Visual Odometry for Mobile Robots

DING Wen-Dong[1, 2] XU De[1, 2, 3] LIU Xi-Long[1, 2] ZHANG Da-Peng[1, 2] CHEN Tian[1, 2]

Abstract Localization plays a key role in mobile robot′s navigation. Vision becomes more and more important for localization. Firstly, this paper gives the mathematical description of visual localization. Secondly, typical methods of visual odometry (VO) are introduced according to the data association modes. Thirdly, the methods to improve the robustness of visual odometry are discussed. Fourthly, the effect of semantic analysis on visual localization is described. How to use deep neural network in visual localization is also provided. Finally, existing problems and future development trends are presented.

Key words Visual odometry (VO), visual localization, pose estimation, navigation, mobile robot

Citation Ding Wen-Dong, Xu De, Liu Xi-Long, Zhang Da-Peng, Chen Tian. Review on visual odometry for mobile robots. *Acta Automatica Sinica*, 2018, **44**(3): 385−400

移动机器人想要完成自主导航[1], 首先要确定自身的位置和姿态, 即实现定位. 一方面, 一些移动机器人尤其是空中机器人[2] 的稳定运行需要位姿信息作为反馈, 以形成闭环控制系统. 另一方面, 随着移动机器人的快速发展, 移动机器人需要完成的任务多种多样, 例如物体抓取[3]、空间探索[4]、农业植保[5]、搜索救援[6] 等, 这些任务对移动机器人的定位提出了更高要求.

收稿日期 2017-02-27 录用日期 2017-09-07
Manuscript received February 27, 2017; accepted September 7, 2017
国家自然科学基金 (61503376, 61673383, 51405485, 51405486), 北京市自然科学基金 (4161002), 天津市支持科研院所来津发展项目 (16PTYJGX00050) 资助
Supported by National Natural Science Foundation of China (61503376, 61673383, 51405485, 51405486), Beijing Natural Science Foundation (4161002), and the Project of Development in Tianjin for Scientific Research Institutes Supported by Tianjin Government (16PTYJGX00050)
本文责任编委 侯增广
Recommended by Associate Editor HOU Zeng-Guang
1. 中国科学院自动化研究所精密感知与控制研究中心 北京 100190 2. 中国科学院大学 北京 101408 3. 天津中科智能技术研究院有限公司 天津 300300
1. Research Center of Precision Sensing and Control, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 2. University of Chinese Academy of Sciences, Beijing 101408 3. Tianjin Intelligent Technology Institute of CASIA Co., Ltd, Tianjin 300300

常用的定位方法有全球定位系统 (Global position system, GPS)、基于惯性导航系统 (Inertia navigation system, INS) 的定位、激光雷达定位、基于人工标志[7−8] 的定位方法、视觉里程计 (Visual odometry, VO) 定位[9] 等. GPS 定位装置接收多颗卫星的信号, 可解算出机器人的三维位置和速度. 定位精度在米量级, 误差不随时间积累, 但 GPS 信号被遮挡的地方无法使用. 基于 INS 的定位利用加速度计和陀螺仪经过积分计算出机器人的位置、速度、姿态等, 数据更新率高、短期精度和稳定性较好, 但定位误差会随时间积累. 激光雷达通过扫描获得机器人周围环境的三维点云数据, 根据这些数据实现机器人相对于环境的定位, 精度高, 实时性强, 但成本较高. 基于人工标志定位的方法利用二维码等作为路标实现机器人的定位, 二维码需要安装于环境中, 可以简单有效地完成定位, 但是一定程度上限制了这些定位方法的使用范围. 视觉里程计[9−10] 通过跟踪序列图像帧间的特征点估计相机的运动, 并对环境进行重建. 与轮式里程计类似, 视觉里程计通过累计帧间的运动估计当前时刻的位姿. VO 在系统运行中形成三维点云, 作为路标点保存在系统中. 在新的视角下, 基于这些点可利用 PnP (Perspective n points)[11] 方法进行定位. 视觉里程计具有广泛的

用途, 可应用于无人车[12]、无人机[13−15]、增强现实[16] 等.

本文针对 VO 展开讨论, 组织结构如下: 第 1 节简要介绍定位问题的数学描述. 第 2 节论述主流的视觉定位方法, 重点介绍三类视觉里程计的原理与特点. 第 3 节讨论在传感器建模和视觉里程计前端后端等方面的鲁棒性设计技巧. 第 4 节介绍结合视觉语义分析的位姿估计方法和深度学习网络在位姿估计中的应用. 第 5 节介绍位姿估计的性能评价方法, 常用的数据集和常用的工具库. 第 6 节给出视觉定位目前存在的问题和未来的发展方向.

1 定位问题数学描述

机器人 k 时刻的位姿为 $T_k = \begin{bmatrix} R_k & \boldsymbol{p}_k \\ \boldsymbol{0} & 1 \end{bmatrix}$, 其中, R_k 为机器人 k 时刻的姿态, $\boldsymbol{p}_k$ 为机器人 k 时刻的位置. 那么 $k+1$ 时刻的位姿为

$$T_{k+1} = T_k T_{k,k+1} \tag{1}$$

其中, $T_{k,k+1}$ 为 $k \sim k+1$ 时刻机器人的相对位姿, 初始状态下机器人的位姿为 T_0.

使用式 (1) 递推获得当前位姿. 因此, 该过程中不可避免地会出现误差, 且该误差具有累积现象. 为消除累积误差, 需要基于观测值进行滤波或 BA (Bundle adjustment) 优化.

为了保证系统的实时性, 视觉定位通常分为两部分: 1) 基于特征匹配的运动估计; 2) 对定位结果进行优化. 特征匹配针对位姿变化前后的图像获取对应特征点对, 利用 n ($n \geq 3$) 个匹配点对以及相机内参数得到相机的运动量. 当相机运动距离较大, 或能够跟踪到的点较少时, 则把这一帧图像作为关键帧保存下来. 优化部分利用特征点的重投影偏差最小化对关键帧对应的相机位姿及特征点在相机坐标系中的位置进行估计. 第 i 关键帧对应的投影矩阵为

$$P_i = K[R_i, \boldsymbol{p}_i] \tag{2}$$

其中, K 表示相机的内参数矩阵.

运动估计和优化均可采用

$$e = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} \| \boldsymbol{m}_{ij} - P_i \boldsymbol{M}_j \|^2 \tag{3}$$

其中, $\boldsymbol{M}_j$ 为路标点, $\boldsymbol{m}_{ij}$ 为 $\boldsymbol{M}_j$ 在第 i 帧中的图像坐标, e 表示误差, w_{ij} 表示路标点 $\boldsymbol{M}_j$ 在第 i 帧中权值. 如果点 j 在第 i 帧中可见, 则 $w_{ij} = 1$, 否则为 0.

运动估计部分利用式 (3) 获得相机的位姿 $[R, \boldsymbol{p}]$, 优化部分则对位姿 $[R, \boldsymbol{p}]$ 和路标点 M_j 同时进行优化.

在求解的过程中, 对该系统线性化, 然后可以用高斯−牛顿或 LM (Levenberg-Marquardt) 方法迭代求解. 由于点与点之间、位姿与位姿之间相对误差项是独立的, 相应矩阵具有稀疏性, 式 (3) 可以实时求解.

上述问题也可以建模为因子图 (Factor graph) 并使用图优化方法求解[17−18]. 图模型[19−20] 可直观地表示视觉定位问题, 图中的状态节点表示机器人的位姿或路标, 节点之间的边对应状态之间的几何约束. 图模型构建之后, 经过优化可得到与测量数据最匹配的状态参数, 进而形成路标点地图. 一个常用图网络优化工具为 g2o (General graph optimization)[21], 详见第 5.2 节.

2 VO 代表性方法

VO 系统中的数据关联表示了 3D 点在不同帧之间的关系. 在运动估计中, 使用当前帧图像和过往帧图像进行数据关联求解相机运动量, 通过递推每一步的运动量可以得到相机和机器人的位姿. 数据关联中的点所在空间有三种[10]:

1) 2D−2D: 当前帧的点和过往帧的点都是在图像空间中. 在单目相机的初始化过程中经常出现这种数据关联.

2) 3D−3D: 当前帧和过往帧的点都在 3D 空间中, 这种情形一般在深度相机 VO 系统的位姿估计或经过三角测量的点进行 BA 时出现.

3) 3D−2D: 过往帧的点在 3D 空间中, 当前帧的点在图像空间中, 这样问题转化为一个 PnP 问题.

在 VO 系统初始化时, 地图未建立, 系统无法确定当前状态, 采用 2D−2D 数据关联, 对基础矩阵或单应矩阵分解求解相机的相对位姿, 三角化求解路标点的三维坐标. 若地图中 3D 点可用, 优先使用 3D 点进行位姿估计. 此时, 将 3D 路标点投影到当前帧图像, 在局部范围内搜索完成图像点的匹配. 这种 3D−2D 的数据关联经常用于 VO 系统正常状态下的定位. 3D−3D 数据关联常用于估计和修正累积误差和漂移. 3D 路标点会出现在多帧图像中, 通过这些 3D 点之间的数据关联可以修正相机的运动轨迹以及 3D 点的三维位置.

例如, SVO (Semi-direct visual odometry)[22] 中除了初始化过程, 正常状态下系统处理当前的每一帧时三种数据关联先后被使用, 2D−2D 数据关联实现图像空间的特征点匹配, 通过 3D−2D 数据关联计算相机的位姿, 并经过 3D−3D 数据关联后利用 BA 进行优化. DTAM (Dense tracking and mapping)[23] 的目标函数中包含了多种数据关联的

误差, 包括图像空间的匹配误差和 3D 空间的位置误差. 当帧间运动较小, 成功匹配的 3D 点较多时, 估计位姿矩阵; 当帧间运动较大, 匹配 2D 点较多时, 估计基础矩阵. 按照 2D−2D 数据关联方式的不同, 视觉定位方法可以分为直接法、非直接法和混合法.

2.1 直接法

作为数据关联方式的一种, 直接法假设帧间光度值具有不变性, 即相机运动前后特征点的灰度值是相同的. 数据关联时, 根据灰度值对特征点进行匹配. 但这种假设与实际情况存在差异, 特征点容易出现误匹配. Engel 等[24−25] 使用了一种更精确的光度值模型, 该模型对相机成像过程建模了相机曝光参数、Gamma 矫正以及镜头衰减. 该模型使用辐照度不变性假设, 可以表示为 $I_i(\boldsymbol{m}) = G(t_i V(\boldsymbol{m}) B(\boldsymbol{m}))$, 其中像素点 $\boldsymbol{m}$ 的辐照度为 B, 镜头的衰减为 V, 曝光时间为 t_i, CCD (Charge coupled device) 的响应函数为 G. 对该模型进行逆向求解得到校正后的图像灰度值, 进行数据关联.

为了快速求解上述问题, Lucas 等[26] 引入 FAIA (Forward additional image alignment) 方法, 使用单一运动模型代替独立像素位移差. Baker 等[27] 提出统一的框架, 在 FAIA 基础上引入 FCIA (Forward composition image alignment), ICIA (Inverse compositional image alignment) 和 IAIA (Inverse additional image alignment)[27]. SVO 和 PTAM (Parallel tracking and mapping)[28] 利用 ICIA 实现块匹配, DPPTAM (Dense piecewise planar tracking and mapping)[29] 利用 ICIA 完成显著梯度点的半稠密重建.

LSD (Large scale direct) SLAM (Simultaneous localization and mapping)[30−31] 采用直接方法进行数据关联, 建立深度估计、跟踪和建图三个线程. 该方法对图像点建立随机深度图, 并在后续帧中对深度进行调整直至收敛. 该方法的初始化不需要两视几何约束, 不会陷入两视几何退化的困境, 但初始化过程需要多个关键帧之后深度图才会收敛, 此期间跟踪器产生的地图是不可靠的. LSD SLAM 通过权值高斯−牛顿迭代方法最小化光度值误差. 光度值误差是当前帧和参考关键帧之间所有对应点的灰度值差的平方和. LSD SLAM 建图对关键帧及非关键帧分开处理, 对于前者, 过往关键帧的深度图投影到当前关键帧, 并作为深度图的初始值; 对于后者, 则进行图像匹配并计算位姿, 对当前帧更新深度信息, 对深度信息进行平滑并移除外点.

DSO (Direct sparse odometry)[24] 系统基于直接法的拓展, 使用光度值误差最小化几何和光度学参数. DSO 对图像中有梯度、边缘或亮度平滑变化的点均匀采样以降低计算量. DSO 对光度学模型校正、曝光时间、透镜畸光晕和非线性响应都做了校准. 为了提高速度, 降低计算量, DSO 使用滑动窗口方法, 对固定帧数的位姿进行优化.

DPPTAM[29] 基于超像素对平面场景进行稠密重建. 该方法对图像中梯度明显的点进行半稠密重建, 然后对图像中其他点进行超像素分割, 通过最小化能量函数完成稠密重建, 该能量函数在第 3.3.2 节中介绍.

直接法使用了简单的成像模型, 适用于帧间运动较小的情形, 但在场景的照明发生变化时容易失败.

2.2 非直接法

另外一种帧间数据关联是非直接法, 又称为特征法, 该方法提取图像中的特征进行匹配, 最小化重投影误差得到位姿. 图像中的特征点以及对应描述子用于数据关联, 通过特征描述子的匹配, 完成初始化中 2D−2D 以及之后的 3D−2D 的数据关联. 常用的旋转、平移、尺度等不变性特征及描述子, 例如 ORB (Oriented FAST and rotated BRIEF)[32]、FAST (Features from accelerated segment test)[33]、BRISK (Binary robust invariant scalable keypoints)[34]、SURF (Speeded up robust features)[35], 可用于完成帧间点匹配.

PTAM[28] 是一个基于关键帧的 SLAM 系统, 是很多性能良好的 SLAM 系统的原型, PTAM 首先引入了跟踪和建图分线程处理的方法. 原始的版本经过修改之后增加了边缘特征、旋转估计和更好的重定位方法. PTAM 的地图点对应图像中的 FAST 角点, FAST 特征计算速度很快, 但没有形成特征描述子, 因此使用块相关完成匹配.

ORB 特征[32] 是一种快速的特征提取方法, 具有旋转不变性, 并可以利用金字塔构建出尺度不变性. 在整个定位过程以及建图的过程中, ORB SLAM[36] 使用了统一的 ORB 特征, 在跟踪的时候提取 ORB 特征, 完成点的匹配、跟踪、三角测量和闭环检测等关键过程.

DT (Deferred triangulation) SLAM[37] 在地图中的路标点不仅使用三维点, 而且使用二维图像特征点. 在位姿估计中, 目标函数中包括三维点的重建误差以及二维特征重投影误差. DT SLAM 维护了三个跟踪器, 每个跟踪器包含一种位姿估计方法: 位姿估计、本质矩阵估计和纯旋转估计. 当足够数量的 3D 点匹配存在时候, 可以使用位姿估计; 当 3D 点数量不足, 但是 2D 点数量较多的时候可以利用对极约束估计本质矩阵. 如果判定当前情况为纯旋转, 那么使用纯旋转估计.

当图像中没有足够的点特征时, 线特征是一个好的补充[38−39]. 通常使用的线段检测器有比较高的精度, 但是很耗时间. Gomez-Ojeda 等[40] 对每条线段计算 LBD (Line band descriptor) 描述子[41], 最小化点特征以及线段特征的重投影误差得到运动估计. Zhou 等[42−43] 使用消失点定义图像中的线结构, 使用 J-linkage[44] 将所得线段分类, 计算消失点的粗略值, 然后通过非线性最小二乘优化得到消失点在图像中的表示以及相机的方向.

Camposeco 等[45] 使用消失点来提高 VO 系统的精度, 首先使用线段检测器检测图像中的线段, 然后使用最小二乘法计算消失点, 将 EKF (Extended Kalman filter) 中的误差状态向量 (核心状态) 中增加消失点作为增广状态, 在更新 EKF 核心状态时同时更新增广状态方程. Gräter 等[46] 使用消失点提高单目 VO 系统的尺度计算的鲁棒性和精度. 但是由于计算实际的尺度值时使用了相机到地面的高度作为先验知识, 该方法仅限于平面运动机器人.

直接法和非直接法的优缺点对比详见表 1.

2.3 混合法

SVO[22] 是一种混合式的 VO, 该方法首先提取 FAST 特征, 使用特征点周围的图像块进行像素匹配, 并对帧间的相对位姿累积以初步估计当前位姿, 累积误差会导致系统产生漂移. SVO 通过匹配当前帧与地图中的点约束当前帧的位姿, 降低累积误差. SVO 初始化时使用单应矩阵分解求解相机的位姿, 假设初始化场景中的点分布在一个平面内, 因此适合平面场景的初始化.

3 鲁棒性改进措施

VO 系统在实际应用中的主要问题是鲁棒性不足, 限制条件过多. 本文从传感器的特性建模、系统的前端、后端等方面, 包括卷帘快门相机建模、系统初始化、运动模型假设、目标函数、深度图模型, 介绍增强鲁棒性的方法.

3.1 视觉传感器建模

很多现代的相机使用 CMOS (Complementary metal oxide semiconductor) 图像传感器, 成本较低, 但使用卷帘快门时, 图像中每一行像素曝光时间窗口不一样. 假设快门启动的时间为 t_0, 图像第 i 行的成像时刻为 t_i, 假设图像有 N_r 行, 传感器数据读出的时间为 t_s. 因此 $t_i = t_0 + t_s i/N_r$. 根据 Karpenko 等[47−48] 的分析可知, 在快门转动的时间段内, 平移运动的影响对于相机模型的影响较小, 可以忽略. 假设在快门开启时, 存在三维点 $\boldsymbol{M}$, 该点的成像时刻为 t_i, 对应图像空间中的点为 $\boldsymbol{m}_i$. 因此有

$$\lambda_i \boldsymbol{m}_i = K R_{0,i} \boldsymbol{M} \tag{4}$$

其中, $R_{0,i}$ 为 t_0 到 t_i 时刻的旋转矩阵, K 为内参数, λ_i 为常数. Kerl 等[49] 针对 RGBD (Red-green-blue depth) 图像使用 B 样条近似相机运动轨迹, 补偿卷帘快门的影响. 系统使用了深度值误差以及光度值误差优化计算相机的运动, 得到平滑连续的轨迹. Pertile 等[50] 使用 IMU (Inertial measurement unit) 来计算 $R_{0,i}$, 也就是从快门开启 t_0 到时刻 t_i 相机运动的旋转矩阵. 另外, Kim 等[51] 定义了行位姿, 相机的位姿依赖于图像行变量. 将滑动帧窗口方法扩展为近邻窗口, 该窗口包含固定个数的 B 样条控制点. 该系统使用 IMU 对相机在快门动作期间内估计相机的运动, 但是由于 CMOS 的快门时间戳和 IMU 的时间戳的同步比较困难, 且相机的时间戳不太准确, Guo 等[52] 对时间戳不精确的卷帘快门相机设计了一种 VIO (Visual inertial odometry) 系统, 其位姿使用线性插值方法近似相机的运动轨迹, 姿态使用旋转角度和旋转轴表示, 旋转轴不变, 对旋转角度线性插值, 使用 MSCKF (Multi-state constrained Kalman filter) 建模卷帘快门相机的测量模型.

Dai 等[53] 对线性卷帘快门模型和均匀卷帘快门模型的相机计算了双视几何的本质矩阵. 线性卷帘快门模型中, 假设相机的运动为匀速直线运动, 均

表 1 直接法与非直接法优缺点对比

Table 1 The comparison between direct methods and indirect methods

	直接法	非直接法
目标函数	最小化光度值误差	最小化重投影误差
优点 1	使用了图像中的所有信息	适用于图像帧间的大幅运动
优点 2	使用帧间的增量计算减小了每帧的计算量	比较精确, 对于运动和结构的计算效率高
缺点 1	受限于帧与帧之间的运动比较小的情况	速度慢 (计算特征描述等)
缺点 2	通过对运动结构密集的优化比较耗时	需要使用 RANSAC 等鲁棒估计方法
容易失败	场景的照明发生变化	纹理较弱的地方

匀卷帘模型中, 相机的运动为一个匀角速度运动和一个匀速直线运动. 在全局快门相机中, 本质矩阵是一个 3×3 的奇异矩阵. 在使用线性卷帘模型的相机下, 本质矩阵为一个 5×5 的矩阵, 在使用均匀卷帘模型的相机下, 本质矩阵为一个 7×7 的矩阵. 因此, 在使用卷帘模型时, 5 点法无法求解本质矩阵. 线性卷帘模型和均匀卷帘模型分别需要 11 和 17 个点求解本质矩阵.

3.2　视觉里程计前端

3.2.1　初始化

单目系统初始化时完成运动估计常用的方法主要有两种: 1) 将当前场景视为一个平面场景[54], 估计单应矩阵并分解得到运动估计, 使用这种方法的有 SVO、PTAM 等. 2) 使用极线约束关系, 估计基础矩阵或者本质矩阵[55−56], 分解得到运动估计, 使用这种方法的有 DT SLAM 等. 初始化中遇到的普遍问题是双视几何中的退化问题. 当特征共面或相机发生纯旋转的时候, 解出的基础矩阵的自由度下降, 如果继续求解基础矩阵, 那么多出来的自由度主要由噪声决定. 为了避免退化现象造成的影响, 一些 VO 系统同时估计基础矩阵和单应矩阵, 例如 ORB SLAM 和 DPPTAM, 使用一个惩罚函数, 判断当前的情形, 选择重投影误差比较小的一方作为运动估计结果.

单目系统在初始化中还要完成像素点的深度估计, 单目系统无法直接从单张图像中恢复深度, 因此需要一个初始估计. 解决该问题的一种办法是跟踪一个已知的结构[57], 另外一种方法是初始化点为具有较大误差的逆深度[30−31], 在之后过程中优化直到收敛至真值.

VO 系统的初始化依赖于精确的相机标定和状态初始值. 对于系统的初始化, Shen 等[18−19] 在系统的运动中, 建立相邻两帧图像间的关系, 对从上一帧惯性坐标系至当前帧相机坐标系进行变换.

$$^{b}T_{k,k+1}\ ^{b}T_{c} = \ ^{b}T_{c}\ ^{c}T_{k,k+1} \tag{5}$$

根据相机和 IMU 多次运动分别获得的 IMU 测量的变换矩阵 $^{b}T_{k,k+1}$ 及相机测量的变换矩阵 $^{c}T_{k,k+1}$, 可以标定相机和 IMU 之间的变换矩阵 $^{b}T_{c}$.

3.2.2　运动模型

机器人的导航中, 实际的运动经常不符合恒速运动模型假设, 需设计应对失败的策略. ORB SLAM 的运动估计通过跟踪若干匹配的特征点来检测这种失败, 这种情况下可跟踪的点的数量较少. 因此 ORB SLAM 设置一定阈值, 如果能够跟踪的点的个数小于该阈值, 则会在一个更大的范围内进行特征的搜索匹配. DSO 系统中如果恒速模型失败, 会使用 27 种不同方向不同大小的旋转来尝试恢复. 这些尝试在较高的金字塔层上完成, 所以耗时很短. SVO 等方法假设当前时刻的位姿等于上一时刻的位姿, 通过最小化光度值误差估计帧间的位姿变化, 使用高斯–牛顿方法完成 ICIA 的迭代. ICIA 的使用也限制了帧间视差的最大值, 或需要较高的帧率 (典型的大于 70 fps). 表 2 给出几种常用运动模型在 VO 系统使用的情形.

表 2　常用运动模型先验假设

Table 2　The common used motion model assumption

运动模型假设	方法
恒速运动模型	ORB SLAM, PTAM, DPPTAM, DSO
帧位姿变化为 0	DPPTAM, SVO, LSD SLAM
帧间仿射变换	DT SLAM

3.3　视觉里程计后端

3.3.1　目标函数

上文讨论了直接法以及间接法中使用的目标函数, 目标函数的设计影响了 VO 系统鲁棒性. 在最大后验估计的定位问题中, 似然函数中如果假设噪声的分布为高斯分布, 那么目标函数中负对数似然函数等价于 ℓ_2 范数. 如果假设噪声的分布为拉普拉斯分布, 负对数似然函数对应 ℓ_1 范数. 在优化中, ℓ_2 范数对噪声敏感, 噪声的存在导致估计的结果与实际参数相差较大, 因而改用 M 估计器替换平方残差函数 $\rho(r_i)$. 表 3 给出几种常用鲁棒估计器的具体表达式.

Özyesil 等[58] 使用 ℓ_1 和 ℓ_2 两种范数结合的一种范数 IRLS (Iteratively reweighted least squares)[59], 通过迭代的方式解决带权重的 ℓ_p 范数 (参见表 3) 的优化问题. VO 系统常用的鲁棒目标函数如表 4 所示. 在恢复相机的运动中, 相机的位置估计容易被噪声干扰, 方向的估计在精度和鲁棒性方面则相对比较准确. Özyesil 等[58] 引入两步估计方法, 首先估计点对的相对方向, 然后从点对的相对方向中恢复每个点的 3D 位置. 位置估计的目标函数形式化为最小化方向的误差, 其中位置表示为方向和距离的乘积, 因为方向已知, 因此优化对象变为距离, 使用 IRLS 方法迭代优化目标值. Sünderhauf 等[60] 使用可切换约束的目标函数, 在优化中识别并丢弃外点. 另外该系统利用可切换的闭环检测约束以及可切换的先验约束, 避免对闭环检测的误报.

3.3.2　深度图

在基于直接法的 VO 系统 (DSO、LSD SLAM) 中, 常常需要估计点的深度, 原始的深度并不表现为

类高斯分布, 而是带有长拖尾. 在室外应用中, 存在很多无穷远点, 初始值难以设定, 因此使用高斯分布描述不准确. 逆深度 (原始深度的倒数) 的分布更加接近高斯分布, 具备更好的数值稳定性. 常用的深度图模型如表 5 所示.

表 3 常用的鲁棒估计器

Table 3 The common used robust estimators

类型	$\rho(x)$
ℓ_2	$\dfrac{x^2}{2}$
ℓ_1	$\lvert x\rvert$
$\ell_1-\ell_2$	$2\left(\sqrt{1+\dfrac{x^2}{2}}-1\right)$
ℓ_p	$\dfrac{\lvert x\rvert^{\nu}}{\nu}$
Huber	$\begin{cases}\dfrac{x^2}{2}, & 若\ \lvert x\rvert\le c\\ c\left(\lvert x\rvert-\dfrac{c}{2}\right), & 若\ \lvert x\rvert>c\end{cases}$
Cauchy	$\dfrac{C^2}{2}\ln\left(1+\left(\dfrac{x}{c}\right)^2\right)$
Tukey	$\begin{cases}\dfrac{c^2}{6}\left(1-\left[1-\left(\dfrac{x}{c}\right)^2\right]^3\right), & 若\ \lvert x\rvert\le c\\ \dfrac{c^2}{6}, & 若\ \lvert x\rvert>c\end{cases}$
t 分布	$\dfrac{\nu+1}{\nu+\left(\dfrac{r}{\sigma}\right)^2}$

表 4 VO 系统中的鲁棒目标函数设计

Table 4 The common used robust objection function in VO systems

VO 系统	目标函数框架
PTAM	Tukey biweight
DSO	Huber
DT SLAM	Cauchy distribution
DPPTAM	Reweighted Tukey
DTAM	Weighted Huber norm

像素点的深度估计方法有滤波器方法和非线性优化方法. 其中 SVO、DSO 将深度建模为一个类高斯模型, 然后使用滤波器估计. 另外一种方法对深度图构建一个能量函数, 例如 LSD SLAM、DTAM、DPPTAM 等, 然后使用非线性优化方法最小化能量函数. 该函数包括一个光度值误差项以及一个正则项, 用来平滑所得结果.

表 5 深度图模型

Table 5 The common used models of depth map

VO 系统	深度模型
SVO	高斯混合均匀模型
DSO	高斯模型
DT SLAM	极线分段约束[61]
DPPTAM 半稠密	一致性假设
DPPTAM 稠密	能量函数[1]
DTAM	能量函数[2]
LSD SLAM	能量函数[3]

[1] 光度值误差 + 图像空间平滑 + 平面块假设.
[2] 使用光度值误差和图像空间平滑 (正则).
[3] 光度值误差和关键帧间方法惩罚.

DPPTAM[29] 首先对图像中梯度明显的点估计深度, 由此得到半稠密的深度图. 梯度明显的点占图像所有点的比例较小, 因此要更新的点数较少, 可以实时完成位姿估计. 另外这些点还用于估计平面结构, 其深度图使用一致性假设, 包括三个方面.

1) 极线方向和梯度方向垂直的点的逆深度值是不可靠的.

2) 时间一致性. 相邻若干时刻同一个像素点的逆深度是相似的.

3) 空间一致性. 相邻像素的逆深度值是相似的.

对于其他点的深度估计通过最小化一个由光度值误差、深度距离和梯度正则项组成的能量函数完成. 光度值误差同直接法中光度值不变性假设. 另外两项为正则项, 深度距离计算了被估计深度距离分段平面的距离. 梯度正则计算了深度图的梯度, 用于平滑深度图. DTAM[23] 中的能量函数除光度值误差、梯度正则外, 还使用了一个对偶项, 避免了线性化目标函数并迭代优化导致的重建结果损失深度图细节, 这样还可以使用原始对偶方法快速完成优化. 原始对偶方法不同于原始方法以及对偶优化方法, 基本思想是从对偶问题的一个可行解开始, 同时计算原问题和对偶问题, 求出原问题满足松弛条件的可行解, 这个可行解就是最优解.

4 语义分析与深度学习

上文介绍了改进视觉里程计鲁棒性的措施, 视觉语义分析以及深度学习的应用同样对提高系统的鲁棒性具有帮助. 本节围绕语义分析和深度学习方面的相关问题展开介绍.

4.1 语义分析

语义分析根据结构型数据的相似特性对像素 (区域) 进行标记, 对场景中的区域分类. 粗粒度的

语义分析应该包括物体检测、区域分割等. 语义分析和位姿估计之间相互影响, 可以体现在两个方面: 1) 语义分析能够提高位姿及建图的精度[62]; 2) VO 的测量结果降低语义分析的难度.

在基于稀疏特征的 VO 系统中, 场景重建为稀疏点云; 在稠密的 VO 系统中, 场景重建为连续的表面; 而在含有语义分析的系统中会建立一个语义地图, 该地图中组成元素为物体, 而不是度量地图中的稠密或稀疏的点. SLAM++ 系统[62] 中, 语义地图表示为一个图网络, 其中节点有两种: 1) 相机在世界坐标系的位姿; 2) 物体在世界坐标系的位姿. 物体在相机坐标系的位姿作为网络中的一个约束, 连接相机节点和物体节点. 另外网络中还加入了平面结构等约束提高定位的精度.

MO-SLAM (Multi object SLAM)[63] 对于场景中重复出现的物体进行检测, 该方法不需要离线训练以及预制物体数据库. 系统将重建的路标点分类, 标记该点所属的物体类别. 一个物体表示为一个路标点集合, 相同的物体的不同实例的路标点之间存在如下关系

$$\boldsymbol{P}^m_{\boldsymbol{I}_j} = E^m_{j1}\boldsymbol{P}^m_{\boldsymbol{I}_1} \tag{6}$$

其中, $\boldsymbol{P}^m_{\boldsymbol{I}_j}$ 表示物体 O_m 的实例 $\boldsymbol{I}_j$ 在系统中的路标点, $\boldsymbol{P}^m_{\boldsymbol{I}_1}$ 表示物体 O_m 的实例 $\boldsymbol{I}_1$ 路标点. 系统对于生成的关键帧建立 ORB 描述子的单词树, 在新的关键帧和候选关键帧之间进行汉明距离匹配. 如果匹配点的数量不够, 那么识别线程停止处理当前帧, 等待下一个关键帧. 使用 RANSAC (Random sample consensus) 框架初始化一个位姿变换, 使用式 (6) 最小化重投影误差. 另外目标函数中增加同类物体不同实例的空间变换约束以提高精度. Choudhary 等[64] 对 SLAM 系统增加了在线物体发现和物体建模方法, 利用检测到的物体作为路标点帮助机器人定位, 有利于系统回环检测. Dame 等[65] 利用 3D 形状先验完成稠密重建, 在 PTAM 系统基础上使用一个滑动窗口进行物体检测, 添加物体的位姿约束至目标函数, 以提高系统定位精度.

高层特征具备更好的区分性, 同时帮助机器人更好完成数据关联. DARNN[66] 引入数据联合 (Data association, DA) 下的 RNN (Recurrent neural network), 同时对 RGBD 图像进行语义标注和场景重建. 将 RGB 图像和深度图像分别输入全卷积网络, 在反卷积层加入数据联合 RNN 层, 将不同帧图像的特征进行融合, 同时能够融合 RGBD 图像和深度图像. 该文章使用 KinectFusion[67] 完成相机的跟踪, 估计当前相机的 6DOF 位姿, 将 3D 场景表示为 3D 体素, 保存于 TSDF (Truncated signed distance function). McCormac 等[68] 使用 ElasticFunsion 完成 SLAM 的稠密重建及位姿估计任务, 使用 FCN (Fully convolutional network) 完成语义分割, 不同的种类使用面元 (Surfel) 表示, 使用贝叶斯更新器跟踪分割该面元的概率分布, 使用 SLAM 生成的点匹配更新面元的概率分布. 针对建图规模大、稠密重建速度慢和室外环境建图困难等问题, Vineet 等[69] 使用基于 CRF (Conditional random field) 的体积平均场方法进行图像分割, 同时基于 KinectFusion 方法完成稠密重建.

4.2 深度学习方法

人类可以不监督的完成认知任务, 通过在代理任务 (例如本体运动估计) 的监督学习可以解决其他的任务 (例如深度理解), 避免了显式的监督学习. 一些任务学习的泛化能力强, 可以作为其他任务的基础. 另外深度网络的应用中, Zamir 等[70] 提出了一种多任务学习的方法, 经过特征匹配任务训练的网络不需要重新调整参数就完成相机位姿的估计, 此过程体现了深度网络的抽象能力. 该网络表现为一种通用的能够泛化至新的任务的深度网络感知系统.

基于深度学习的方法要解决的一个基本问题是如何得到训练使用的大规模数据集合, KITTI (Karlsruhe institute of technology and Toyota technological institute) 和 TUM (Technische Universität München) 数据集中除了图像序列, 还给出了图像的深度和相机采集图像时的位姿, 详见第 5.3 节. 如果不存在 VICON 或高精度 IMU 等数据作为真值, 只有单纯图像序列的数据集, 可以使用 SFM (Structure from motion) 方法计算每一帧图像的对应相机运动参数.

现有的深度学习还无法完成一个完整的视觉定位系统, 但有望能够解决传统的 VO 方法难以解决的问题, 例如重定位[71]、长极线匹配[72−73]、数据融合[74] 等. 在一个完整的 VO 系统中, 深度网络一般作为一个辅助系统, 利用高层次的语义分析, 目标识别的功能形成基于语义级的定位约束提高系统的精度和鲁棒性. 表 6 为一些深度学习网络定位系统的特点, 包括要解决的问题, 输出结果等.

在视差大 (基线宽), 而运动模型预测不好的状态下, 由于搜索区域较大, VO 系统中容易发生点匹配失效. 另外一些情况, 例如局部外观变化或自遮挡, 点匹配也容易失效. Choy 等[72] 针对该问题结合 CNN (Convolutional neural network) 和 RNN 网络, 利用物体的形状信息对单帧图像完成三维重建. 由于 LSTM (Long short term memory) 网络可以学习长期历史信息, 在训练中网络针对同一物体不同视角的图像的信息进行处理, 输出物体的一个 3D 栅格. 如果已知物体的外表和形状, 使用这些先验信息, 在大视差下仍然可以完成特征匹配以及

表 6 深度网络定位系统特点

Table 6 The comparison of the learning based localization methods

定位系统	目标函数	输入数据	输出结果	网络类型	面向的问题
LSM[1][79]	SFA[2]	2 帧图像	位姿	CNN	运动估计
PoseNet	位姿误差 (7)	RGB 单帧图像	位姿	GoogLeNet[3]	重定位
3D-R2N2[5][72]	体素交叉熵[4]	单/多帧图像	图像重建	CNN + LSTM	三维重建
LST[6][74]	位姿误差	IMU	位姿	LSTM	数据融合
MatchNet	相似度交叉熵[7]	2 帧图像	匹配度	CNN + FC	图像块匹配
GVNN	光度值误差	当前/参考图像	位姿	CNN + SE3[8]	视觉里程计
HomographNet	图像点误差	2 帧图像	单应矩阵	CNN	估计单应矩阵
SFM-Net[80]	相机运动误差	RGBD 图像	相机运动、三维点云	全卷积	相机运动估计和三维重建
SE3-Net[81]	物体运动误差	点云数据	物体运动	卷积 + 反卷积	刚体运动

[1] Learning to see by moving.
[2] SFA 使用图像的密集像素匹配目标 (损失) 函数, 详见文献 [82].
[3] GoogLeNet 是一种 22 层的 CNN 网络, 常用于分类识别等.
[4] 体素是一个三维向量, 对应像素 (二维), 具有三维坐标, 表示点对应的空间位置的颜色值, 详见文献 [83].
[5] 3D-R2N2(3D Recurrent reconstruction neural network).
[6] Learning to fuse.
[7] 文章在全连接层中使用 Softmax 层, 因此输出为 0/1 值, 全连接层输入为拼接的特征点对, 目标函数为 sofmax 输出值的交叉熵误差.
[8] 除了使用 SE3 层, 还有包括投影层, 反投影层等.

三维重建. 使用深度网络进行深度图估计可以省略中间步骤, 例如形状外表的学习和特征匹配, 直接进行三维重建[72, 75−76], 但需要使用预知的 3D 模型数据.

Doumanoglou 等[77−78] 利用隐类型霍夫森林 (Latent class Hough forest, LCHF) 同时进行物体识别和位姿估计, LCHF 在训练中使用正样本和回归保持类分布在叶节点上. 在测试中类分布作为隐变量被迭代更新. Doumanoglou 等[77] 通过稀疏自编码器提取对应的特征向量, 然后对特征向量构成 HF. 在 Hough 空间中统计各节点投票数, 得到最终的物体类别的位姿. 使用深度网络可以从单帧图像中估计物体的位姿, 该网络在识别物体的同时估计物体的位姿. Wohlhart 等[84] 使用 3D 描述子表示物体的特征和物体的位姿, 使用欧拉距离计算描述子之间的相似度. 使用深度网络完成位姿估计的一种方法是利用其他任务训练的网络及参数, 迁移至定位估计, 例如 (PoseNet[71], FuseNet[85]). 使用端到端的训练方式中, 图像对应的相机位姿数据作为回归结果, 损失函数为

$$L_i = \|\boldsymbol{p}_i - \hat{\boldsymbol{p}}_i\|_2 + \beta \cdot \left\| \boldsymbol{q}_i - \frac{\hat{\boldsymbol{q}}_i}{\|\hat{\boldsymbol{q}}_i\|_2} \right\|_2 \tag{7}$$

其中, $\boldsymbol{p}_i$ 和 $\hat{\boldsymbol{p}}_i$ 为位置的真值和预测值, $\boldsymbol{q}_i$ 和 $\hat{\boldsymbol{q}}_i$ 为姿态四元数的真值和预测值. 针对单帧图像, Kendall 等[71] 训练一个端到端网络, 迁移学习针对分类任务训练的网络 (GoogLeNet), 修改末端结构为回归层, 利用 SFM 标注的数据集重新训练. DeTone 等[86] 训练 HomograghNet 用于估计帧间单应矩阵, 通过产生随机透视变换, 对数据集中的图像做变换, 原始图像和变换后的图像一同输入网络进行训练.

Liu 等[87] 从深度值的连续性出发, 将深度值预测转化为条件随机场问题, 使用深度结构化学习模式, 构造连续条件随机场的一元和二元势函数. 根据相邻区域的像素的深度估计一致性信息, 点的深度差作为一元势函数, 计算区域间颜色差异, 颜色直方图差异和纹理差异, 这些差异构成二元势函数.

Handa 等[85, 88−89] 提出了空间变换层, SO(3) 层对应旋转变换, 参数可以表示为一个三维向量, SE(3) 层在 SO(3) 层的基础上增加了一个平移, 参数为一个 6 维向量. Sim(3) 层在 SE(3) 的顶层有一个尺度因子, 投影层将 3D 点投影到图像平面, 参数为焦距和光心位置.

双塔结构的网络 (例如 MatchNet[90], LSM[83]) 的输入为当前帧图像和参考帧图像, 双塔 CNN 网络使用了相同的参数, 为保证在训练结束后仍然保持相同的参数, 在训练时同步更新两个子网络参数. Xiang 等[66] 双塔结构输入的两个通道分别是 RGB 图像和深度图像, 在卷积层后使用数据联合融合两个通道的卷积信息和 RNN 处理帧间的信息实现深度重建.

另外一种常用结构为编解码器结构, 例如 FuseNet[80]、3D-R2N2[77], 使用卷积层作为编码器, 反卷积层作为解码器, LSTM 置于编码器和解码器

中, 并融合来自深度图像和 RGB 图像信息. Choy 等[72] 利用 LSTM 网络存储信息的特点, 卷积层作为编码器, 经过 LSTM 网络, 数据进入反卷积层. 编码器将图像转换至低维的特征空间, 然后更新网络状态, 通过反卷积层解码隐含层得到重建的三维点.

5　定位方法性能评价

本节介绍视觉定位方法的验证方法. 首先介绍一些性能的评价方法, 然后介绍相关的数据集和工具库.

5.1　性能评价

如果验证数据集中提供了相机位姿的真值, 那么可以直接比较测量值和真值, 称为绝对轨迹误差. 这时进行性能评价是比较直接的, 但是实际上运动相机在连续采集图像过程中难以获得相机位姿的真值, 参见表 6. 为完成算法的验证, Engel 等[25] 使用一个闭环的运动, 相机运动的开始和结束在同一个位置, 被测试算法只需要比较开始和最终状态下的位姿就可以计算出整个算法的漂移的大小. Engel 等[25] 给出了一种统一计算尺度误差、位置、姿态的误差的方法. 该方法首先通过最小化测量结果和实际值之间的位姿, 计算出初始时刻位姿 T_s 和结束时刻位姿 T_e. 然后计算两者之间的漂移 $T_{e,s} = (T_e)^{-1}T_s$. 为了避免分别计算尺度、位置和旋转的漂移, 文章定义了对齐误差.

$$e_a = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|T_s\boldsymbol{p}_i - T_e\boldsymbol{p}_i\|_2^2} \tag{8}$$

这种测量方式可以应用于具有不同的观测方式的定位系统, 被评估的系统可以是双目系统也可以是 VIO 系统, 对于尺度、位置、旋转的误差影响是均衡的.

另外一种难于验证的情形是相对位姿的验证, Burgard 等[91−92] 提出了一种基于图模型的相对位姿计算方法, 但该方法是基于二维空间中三自由度的运动, 我们将之拓展至三维空间六自由度的运动. 两个位姿之间的相对误差为

$$\varepsilon(\delta) = \frac{1}{n}\sum_{i,j}(\delta_{i,j} \ominus \delta_{i,j}^*)^2 \tag{9}$$

其中, $\ominus$ 表示标准运动组合算子 $\oplus$ 的逆算子. 我们假设对于一个 SE(3) 量的扰动量 ΔT, 对应的李代数表示为 $\delta\boldsymbol{\xi} = [\delta\boldsymbol{\rho}, \delta\boldsymbol{\phi}]$, 一个原始的位姿 $T_1 = [R_1, \boldsymbol{P}_1]$, 扰动之后的位姿为

$$T_2 = T_1 \oplus \Delta T = \begin{bmatrix} R_1\exp(\delta\boldsymbol{\rho}^\wedge) & \boldsymbol{P}_1 + \delta\boldsymbol{\phi} \end{bmatrix} \tag{10}$$

其中, $\exp(*^\wedge)$ 表示 $\mathfrak{so}(3)$ 李代数计算出反对称矩阵, 然后进行指数变换. $\Delta T = T_2 \ominus T_1$, 因此

$$\delta_{i,j} \ominus \delta_{i,j}^* = \|\delta\boldsymbol{\rho}\|_2 + \|\delta\boldsymbol{\phi}\|_2 = \|\ln\left((R_1^{-1}R_2)^\vee\right)\|_2 + \|\boldsymbol{P}_2 - \boldsymbol{P}_1\|_2 \tag{11}$$

5.2　开源库及相关工具

视觉方面, ORB、BRISK 等特征描述子、LK 光流法[26] 等在 OpenCV[93] 均有实现. 另外一个重要的问题是相机和 IMU 的标定问题, 相机的标定中对于针孔相机 OpenCV Calib 和 MATLAB 相机标定工具箱使用了标准的模型. Kalibr[94] 是一个工具箱, 它能够标定多目相机系统、相机 IMU 相对位姿和卷帘快门相机. 常用的 SFM 工具有 Bundler[95]、OpenMVG[96] 和 MATLAB 多视几何工具箱[97] 等. Bunlder 增量式地处理一组图像, 提取其中的特征点进行匹配, 完成三维重建并输出一个稀疏的场景结构. OpenMVG 则偏重于多视几何问题的求解.

优化方面, Sophus 库为三维空间的刚体变换及李群李代数一个 C++ 的实现. Eigen 为线性代数和 (稀疏) 矩阵的实现, 对 LAPACK 实现了 C++ 的封装. g2o[21] 是一个针对非线性最小二乘优化问题的 C++ 代码实现. VO 问题可以用图表示, g2o 把非线性最小二乘问题表示为一个图或超图, 图的边可以连接多个节点, 一个超图是图的拓展问题, 其他的优化实现还包括 ceres[98]、GTSAM[99]、iSAM[100]、SLAM++[101] (这里的 SLAM++ 不同于文献 SLAM++[62], 前者是一个非线性优化方法, 后者对应一种语义 SLAM 系统). 常用的优化开源库及其使用场合, 如表 7 所示.

表 7　视觉定位系统工具库

Table 7　The common used tools in visual localization

分类	算法库
优化	Eigen, g2o[21], ceres[98], GTSAM[99], iSAM[100], SLAM++[101]
空间变换	Eigen, ROS TF, OpenCV Transform, Sophus
标定	OpenCV Calib, Kalibr, MATLAB Calibration Toolbox
特征	OpenCV Feature, VLFeat[102]
可视化	PCL Visialization, Pangolin, rviz
SFM	Bundler[95], opencvMVG[96], 多视几何 Matlab 工具箱[97]

5.3 验证数据集

大规模数据的存在使得深度网络在各种视觉任务中达到较好的效果, 同样在机器人的定位技术发展的同时产生多种可用的数据集. 这些数据使得研究者在没有机器人硬件平台的情况下仍然可以开发出可以实际应用的方法. 我们从数据集的发布时间, 数据的类型, 相机的类型, 真值的来源等方面介绍几个 VO 系统中常用的验证数据集, 如表 8 所示.

这些数据集具有不同的特点, COLD 数据集采集了来自不同光照条件下 (白天、晚上、多云) 的图像. 该数据包含了室内的一些常见物体的图像, 一些语义地图方法使用它作为验证数据集, 验证语义建图方法的效果. ICL NUM 数据规模适于训练深度网络, 完成图像的匹配, 图像的光流计算等.

6 未来发展方向

综上所述, 移动机器人的视觉方法仍然存在多个方面的问题, 鲁棒性方面的问题主要集中在如何完成图像的配准以及系统初始化、卷帘快门等问题, 效率方面主要集中在如何实时的完成稠密、半稠密重建、图像点的选择、如何进行边缘化等问题.

随着深度学习在物体检测、语义分割、物体跟踪等方向的发展, 环境中语义和环境理解更多地与视觉定位相结合提高视觉定位的鲁棒性, 并建立更精简的地图. 另外, 嵌入式 VO 系统以及组合定位也将成为视觉定位系统的发展方向.

6.1 嵌入式系统

随着移动处理的发展, 嵌入式系统的性能变得更加接近 PC, 但是计算能力仍然较弱. 而移动机器人和无人机等常常使用嵌入式系统作为视觉处理系统. 使用 SIMD (Single instruction multiple data) 指令可对 3D 重建和后端的优化进行加速. 除了 SIMD, 另外一种加速方法是使用 GPU. 早期的 VO 方法只能进行实时稀疏的三维点云重建, GPU 的使用使得单目视觉能够实时完成稠密重建. 嵌入式系统的 GPU 和 CPU 共享 RAM 存储器, 不需要像 PC 机那样消耗很长的时间完成数据在 CPU 和 GPU 之间的交换. Jetson TK1, TX1/2[112] 使得开发者可以在嵌入式系统中使用 GPU, 便于在无人机和移动机器人对功耗和载重等要求严格的系统完成视觉定位算法. Pizzoli 等[113] 对深度图建立深度滤波器, 使用正则化方法, 利用 GPU 实时完成稠密三维点云重建. DTAM[23] 使用 GPU 针对特征缺失和图像模糊等情况下实现稳定的跟踪.

6.2 组合定位

由于单一定位方法难以满足机器人对定位精度的要求, 所以组合定位方式[114] 应运而生. 一种组合定位方式是以 INS 为主, 引入另一种辅助定位方式以修正惯性测量数据的累积误差[115], 例如 GPS、视觉定位等. 另一种组合定位方式以视觉定位为主, 配合 GPS、INS 等, 改善定位精度和鲁棒性. 第一种方

表 8 VO 系统常用验证数据集

Table 8 The common used dataset in VO system

名称	发布时间	数据类型	相机类型	真值来源	传感器	文献
KITTI VO	2012	png	双目	GPS	激光	[103]
TUM-Monocular	2012	jpg	单目	无	否	[25]
TUM-RGBD	2012	png + d	RGBD	无	否	[104]
ICL NUM	2014	png	双目	[4]	无	[105]
EuRoC MAV	2016	ROS[1] + ASL[2]	双目	VICON[3]	IMU	[106]
Scene Flow	2016	png	双目	[4]	无	[107]
COLD[5]	2009	JPEG	全向[6]	无	激光[7]	[108]
NYU depth	2011/2012	png + d	RGBD	无	无	[109], [110]
PACAL 3D +	2014	JPEG	单目	[4]	无	[111]

[1] ROS 中使用的一种 bag 记录文件, 使用 ROS 可以广播文件中的数据为消息.

[2] ASL 为该数据集自定义格式.

[3] 除了使用 VICON 另外还有 Laser tracker 以及 3D structure scan, 具体为 Vicon motion capture system (6D pose), Leica MS50 laser tracker (3D position), Leica MS50 3D structure scan.

[4] 合成数据集, 存在真值.

[5] 这里有一个 COLD 数据的拓展数据集, 详见 http://www.pronobis.pro/data/cold-stockholm

[6] 系统配备了普通的相机以及全向相机 (Omnidirectional camera).

[7] 除了激光雷达, 还有一个轮式里程计 (码盘).

式实时性好, 较常见于无人机系统. 第二种方式信息量丰富, 抗干扰能力强, 在移动机器人系统中较常采用.

视觉信息和 IMU 数据融合在数据交互的方式上主要可以分为两种方式, 松耦合[115−117] 和紧耦合[18−19]. 松耦合的方法采用独立的惯性定位模块和定位导航模块, 两个模块更新频率不一致, 模块之间存在一定的信息交换. 在松耦合方式中以惯性数据为核心, 视觉测量数据修正惯性测量数据的累积误差. 松耦合方法中视觉定位方法作为一个黑盒模块, 由于不考虑 IMU 信息的辅助, 因此在视觉定位困难的地方不够鲁棒, 另外该方法无法纠正视觉测量引入的漂移.

紧耦合方式使用 IMU 完成视觉 VO 中的运动估计, IMU 在图像帧间的积分的误差比较小, IMU 的数据可用于预测帧间运动, 加速完成点匹配, 完成 VO 位姿估计. 相对于松耦合, 紧耦合的另外一个优点是 IMU 的尺度度量信息可以用于辅助视觉中的尺度的估计.

6.3 语义分析与深度学习

语义分析和深度学习网络在视觉定位中的作用越来越重要. 在未来发展中, 语义分析与视觉定位的结合可能表现有以下几种形式: 通过语义分割完成图像的区域分割, 物体检测结果和图像区域的分割结果建立新的约束实现相机更加精确的定位. 另外可以通过对重建的三维点云分割建立更加紧凑的语义地图, 降低对空间资源的需求.

通过深度卷积网络的特征提取有望取代手工设计的特征提取和匹配, 通过离线或在线的训练, 定位系统利用的特征更加贴近应用场景, 提高在相应的应用场景下的鲁棒性和定位精度. 通过 RNN 网络在未来有望取代视觉里程计的帧间数据关联, 通过 LSTM 等网络的记忆特性, 使得深度网络更加方便地处理图像帧序列并保存其中的历史信息. 通过深度网络的端到端的训练实现场景识别, 有望实现大规模的建图, 消除定位过程的累积误差.

7 结束语

本文首先简述了定位问题, 对定位问题进行建模, 按照数据关联方式分类介绍了几种常用的 VO 系统. 然后围绕鲁棒性展开介绍几个方面的 VO 系统的特点, 这些方面在不同程度上影响了系统的鲁棒性. 接着介绍了语义分析在视觉定位中作用以及如何使用深度网络进行视觉定位. 本文最后介绍了性能评价的方法, 相关的开源库、开源工具, 以及验证数据集.

在过去的多年里, 视觉定位系统取得了许多进步, 无论是早期的基于特征方法, 还是采用光度值匹配的直接法都得到了较快发展. 稀疏矩阵及相关的优化工具使得 VO 系统可以使用图优化方法代替滤波器方法, 显著提升精度的同时保持实时性. 视觉系统的研究已经取得很多进展, 但是系统的鲁棒性和资源消耗等方面还存在需要提高的地方. 例如, 应对成像模型尤其是卷帘快门相机的建模方法、控制优化规模同时不损失过多的精度、尺度漂移等, 虽然有一些解决方法能够在一定程度上提高系统的性能, 但仍存在提升的空间.

深度学习在场景识别中的进展, 为我们提供了许多使用深度学习网络完成定位的思路. 语义分析与视觉定位的结合、深度学习应用于视觉定位、嵌入式视觉定位系统和组合定位等都是未来定位和视觉定位系统的重要发展方向, 这些方向有望在进一步提升系统鲁棒性的同时降低所需的计算资源.

References

1 Burri M, Oleynikova H, Achtelik M W, Siegwart R. Real-time visual-inertial mapping, re-localization and planning onboard MAVs in unknown environments. In: Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hamburg, Germany: IEEE, 2015. 1872−1878

2 Dunkley O, Engel J, Sturm J, Cremers D. Visual-inertial navigation for a camera-equipped 25g Nano-quadrotor. In: Proceedings of IROS2014 Aerial Open Source Robotics Workshop. Chicago, USA: IEEE, 2014. 1−2

3 Pinto L, Gupta A. Supersizing self-supervision: learning to grasp from 50 K tries and 700 robot hours. In: Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA). Stockholm, Sweden: IEEE, 2016. 3406−3413

4 Ai-Chang M, Bresina J, Charest L, Chase A, Hsu J C J, Jonsson A, Kanefsky B, Morris P, Rajan K, Yglesias J, Chafin B G, Dias W C, Maldague P F. MAPGEN: mixed-initiative planning and scheduling for the mars exploration rover mission. *IEEE Intelligent Systems*, 2004, **19**(1): 8−12

5 Slaughter D C, Giles D K, Downey D. Autonomous robotic weed control systems: a review. *Computers and Electronics in Agriculture*, 2008, **61**(1): 63−78

6 Kamegawa T, Yarnasaki T, Igarashi H, Matsuno F. Development of the snake-like rescue robot "kohga". In: Proceedings of the 2004 IEEE International Conference on Robotics and Automation. New Orleans, LA, USA: IEEE, 2004. 5081−5086

7 Olson E. AprilTag: a robust and flexible visual fiducial system. In: Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA). Shanghai, China: IEEE, 2011. 3400−3407

8 Kikkeri H, Parent G, Jalobeanu M, Birchfield S. An inexpensive method for evaluating the localization performance of a mobile robot navigation system. In: Proceedings of the

2014 IEEE International Conference on Robotics and Automation (ICRA). Hong Kong, China: IEEE, 2014. 4100−4107

9 Scaramuzza D, Faundorfer F. Visual odometry: Part I: the first 30 years and fundamentals. *IEEE Robotics and Automation Magazine*, 2011, **18**(4): 80−92

10 Fraundorfer F, Scaramuzza D. Visual odometry: Part II: matching, robustness, optimization, and applications. *IEEE Robotics and Automation Magazine*, 2012, **19**(2): 78−90

11 Hesch J A, Roumeliotis S I. A direct least-squares (DLS) method for PnP In: Proceedings of the 2011 International Conference on Computer Vision (ICCV). Barcelona, Spain: IEEE, 2011. 383−390

12 Craighead J, Murphy R, Burke J, Goldiez B. A survey of commercial and open source unmanned vehicle simulators. In: Proceedings of the 2007 IEEE International Conference on Robotics and Automation. Roma, Italy: IEEE, 2007. 852−857

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

14 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**(1−2): 21−39

15 Lee G H, Achtelik M, Fraundorfer F, Pollefeys M, Siegwart R. A benchmarking tool for MAV visual pose estimation. In: Proceedings of the 11th International Conference on Control Automation Robotics and Vision (ICARCV). Singapore, Singapore: IEEE, 2010. 1541−1546

16 Klein G, Murray D. Parallel tracking and mapping for small AR workspaces. In: Proceedings of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR). Nara, Japan: IEEE, 2007. 225−234

17 Leutenegger S, Lynen S, Bosse M, Siegwart R, Furgale P. Keyframe-based visual-inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 2015, **34**(3): 314−334

18 Yang Z F, Shen S J. Monocular visual-inertial state estimation with online initialization and camera-IMU extrinsic calibration. *IEEE Transactions on Automation Science and Engineering*, 2017, **14**(1): 39−51

19 Shen S J, Michael N, Kumar V. Tightly-coupled monocular visual-inertial fusion for autonomous flight of rotorcraft MAVs. In: Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA). Seattle, WA, USA: IEEE, 2015. 5303−5310

20 Concha A, Loianno G, Kumar V, Civera J. Visual-inertial direct SLAM. In: Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA). Stockholm, Sweden: IEEE, 2016. 1331−1338

21 Kümmerle R, Grisetti G, Strasdat H, Konolige K, Burgard W. G2o: a general framework for graph optimization. In: Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA). Shanghai, China: IEEE, 2011. 3607−3613

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

23 Newcombe R A, Lovegrove S J, Davison A J. DTAM: dense tracking and mapping in real-time. In: Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV). Barcelona, Spain: IEEE, 2011. 2320−2327

24 Engel J, Koltun V, Cremers D. Direct sparse odometry. arXiv: 1607.02565, 2016.

25 Engel J, Usenko V, Cremers D. A photometrically calibrated benchmark for monocular visual odometry. arXiv: 1607. 02555, 2016.

26 Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence. Vancouver, BC, Canada: ACM, 1981. 674−679

27 Baker S, Matthews I. Lucas-Kanade 20 years on: a unifying framework. *International Journal of Computer Vision*, 2004, **56**(3): 221−255

28 Klein G, Murray D. Parallel tracking and mapping for small AR workspaces. In: Proceedings of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR). Nara, Japan: IEEE, 2007. 225−234

29 Concha A, Civera J. DPPTAM: dense piecewise planar tracking and mapping from a monocular sequence. In: Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hamburg, Germany: IEEE, 2015. 5686−5693

30 Engel J, Sturm J, Cremers D. Semi-dense visual odometry for a monocular camera. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, NSW, Australia: IEEE, 2013. 1449−1456

31 Engel J, Schöps T, Cremers D. LSD-SLAM: large-scale direct monocular SLAM. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 834−849

32 Rublee E, Rabaud V, Konolige K, Bradski G. ORB: an efficient alternative to SIFT or SURF. In: Proceedings of the 2011 IEEE International Conference on Computer Vision. Barcelona, Spain: IEEE, 2011. 2564−2571

33 Rosten E, Porter R, Drummond T. Faster and better: a machine learning approach to corner detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(1): 105−119

34 Leutenegger S, Chli M, Siegwart R Y. Brisk: binary robust invariant scalable keypoints. In: Proceedings of the 2011 International Conference on Computer Vision. Barcelona, Spain: IEEE, 2011. 2548−2555

35 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

36 Mur-Artal R, Montiel J M M, Tardós J D. Orb-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 2015, **31**(5): 1147−1163

37 Herrera C D, Kim K, Kannala J, Pulli K, Heikkilä J. DT-SLAM: deferred triangulation for robust SLAM. In: Proceedings of the 2nd International Conference on 3D Vision (3DV). Tokyo, Japan: IEEE, 2014. 609−616

38 Yang S C, Scherer S. Direct monocular odometry using points and lines. arXiv: 1703.06380, 2017.

39 Lu Y, Song D Z. Robust RGB-D odometry using point and line features. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 3934−3942

40 Gomez-Ojeda R, Gonzalez-Jimenez J. Robust stereo visual odometry through a probabilistic combination of points and line segments. In: Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA). Stockholm, Sweden: IEEE, 2016. 2521−2526

41 Zhang L L, Koch R. An efficient and robust line segment matching approach based on LBD descriptor and pairwise geometric consistency. *Journal of Visual Communication and Image Representation*, 2013, **24**(7): 794−805

42 Zhou H Z, Zou D P, Pei L, Ying R D, Liu P L, Yu W X. StructSLAM: visual slam with building structure lines. *IEEE Transactions on Vehicular Technology*, 2015, **64**(4): 1364−1375

43 Zhang G X, Suh I H. Building a partial 3D line-based map using a monocular SLAM. In: Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA). Shanghai, China: IEEE, 2011. 1497−1502

44 Toldo R, Fusiello A. Robust multiple structures estimation with J-linkage. In: Proceedings of the 10th European Conference on Computer Vision. Marseille, France: Springer, 2008. 537−547

45 Camposeco F, Pollefeys M. Using vanishing points to improve visual-inertial odometry. In: Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA). Seattle, WA, USA: IEEE, 2015. 5219−5225

46 Gräter J, Schwarze T, Lauer M. Robust scale estimation for monocular visual odometry using structure from motion and vanishing points. In: Proceedings of the 2015 IEEE Intelligent Vehicles Symposium (IV). Seoul, South Korea: IEEE 2015. 475−480

47 Karpenko A, Jacobs D, Baek J, Levoy M. Digital Video Stabilization and Rolling Shutter Correction Using Gyroscopes, Stanford University Computer Science Technical Report, CTSR 2011-03, Stanford University, USA, 2011.

48 Forssén P E, Ringaby E. Rectifying rolling shutter video from hand-held devices. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, CA, USA: IEEE, 2010. 507−514

49 Kerl C, Stüeckler J, Cremers D. Dense continuous-time tracking and mapping with rolling shutter RGB-D cameras. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 2264−2272

50 Pertile M, Chiodini S, Giubilato R, Debei S. Effect of rolling shutter on visual odometry systems suitable for planetary exploration. In: Proceedings of the 2016 IEEE Metrology for Aerospace (MetroAeroSpace). Florence, Italy: IEEE, 2016. 598−603

51 Kim J H, Cadena C, Reid I. Direct semi-dense SLAM for rolling shutter cameras. In: Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA). Stockholm, Sweden: IEEE, 2016. 1308−1315

52 Guo C X, Kottas D G, DuToit R C, Ahmed A, Li R P, Roumeliotis S I. Efficient visual-inertial navigation using a rolling-shutter camera with inaccurate timestamps. In: Proceedings of the 2014 Robotics: Science and Systems. Berkeley, USA: University of California, 2014. 1−9

53 Dai Y C, Li H D, Kneip L. Rolling shutter camera relative pose: generalized epipolar geometry. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 4132−4140

54 Faugeras O D, Lustman F. Motion and structure from motion in a piecewise planar environment. *International Journal of Pattern Recognition and Artificial Intelligence*, 1988, **2**(3): 485−508

55 Tan W, Liu H M, Dong Z L, Zhang G F, Bao H J. Robust monocular SLAM in dynamic environments. In: Proceedings of the 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). Adelaide, SA, Australia: IEEE, 2013. 209−218

56 Lim H, Lim J, Kim H J. Real-time 6-DOF monocular visual SLAM in a large-scale environment. In: Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA). Hong Kong, China: IEEE, 2014. 1532−1539

57 Davison A J, Reid I D, Molton N D, Stasse O. MonoSLAM: real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, **9**(6): 1052−1067

58 Özyesil O, Singer A. Robust camera location estimation by convex programming. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 2674−2683

59 Daubechies I, DeVore R, Fornasier M, Güntürk C S. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 2010, **63**(1): 1−38

60 Sünderhauf N, Protzel P. Switchable constraints for robust pose graph SLAM. In: Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vilamoura, Portugal: IEEE, 2012. 1879−1884

61 Chum O, Werner T, Matas J. Epipolar geometry estimation via RANSAC benefits from the oriented epipolar constraint. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR). Cambridge, UK: IEEE, 2004. 112−115

62 Salas-Moreno R F, Newcombe R A, Strasdat H, Kelly P H J, Davison A J. SLAM++: simultaneous localisation and mapping at the level of objects. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Portland, OR, USA: IEEE, 2013. 1352−1359

63 Dharmasiri T, Lui V, Drummond T. Mo-SLAM: multi object SLAM with run-time object discovery through duplicates. In: Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Daejeon, South Korea: IEEE, 2016. 1214−1221

64 Choudhary S, Trevor A J B, Christensen H I, Dellaert F. SLAM with object discovery, modeling and mapping. In: Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Chicago, IL, USA: IEEE, 2014. 1018−1025

65 Dame A, Prisacariu V A, Ren C Y, Reid I. Dense reconstruction using 3D object shape priors. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Portland, OR, USA: IEEE, 2013. 1288−1295

66 Xiang Y, Fox D. DA-RNN: semantic mapping with data associated recurrent neural networks. arXiv: 1703.03098, 2017.

67 Newcombe R A, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison A J, Kohi P, Shotton J, Hodges S, Fitzgibbon A. KinectFusion: real-time dense surface mapping and tracking. In: Proceedings of the 10th IEEE International Symposium on Mixed and Augmented Reality (ISMAR). Basel, Switzerland: IEEE, 2011. 127−136

68 McCormac J, Handa A, Davison A, Leutenegger S. SemanticFusion: dense 3D semantic mapping with convolutional neural networks. arXiv: 1609.05130, 2016.

69 Vineet V, Miksik O, Lidegaard M, Nießner M, Golodetz S, Prisacariu V A, Kähler O, Murray D W, Izadi S, Pérez P, Torr P H S. Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In: Proceedings of the 2015 IEEE International Conference on in Robotics and Automation (ICRA). Seattle, WA, USA: IEEE, 2015. 75−82

70 Zamir A R, Wekel T, Agrawal P, Wei C, Malik J, Savarese S. Generic 3D representation via pose estimation and matching. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, Netherlands: Springer, 2016. 535−553

71 Kendall A, Grimes M, Cipolla R. PoseNet: a convolutional network for real-time 6-DOF camera relocalization. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 2938−2946

72 Choy C B, Xu D F, Gwak J, Chen K, Savarese S. 3D-R2N2: a unified approach for single and multi-view 3D object reconstruction. arXiv: 1604.00449, 2016.

73 Altwaijry H, Trulls E, Hays J, Fua P, Belongie S. Learning to match aerial images with deep attentive architectures. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 3539−3547

74 Rambach J R, Tewari A, Pagani A, Stricker D. Learning to fuse: a deep learning approach to visual-inertial camera pose estimation. In: Proceedings of the 2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). Merida, Mexico: IEEE, 2016. 71−76

75 Kar A, Tulsiani S, Carreira J, Malik J. Category-specific object reconstruction from a single image. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 1966−1974

76 Vicente S, Carreira J, Agapito L, Batista J. Reconstructing PASCAL VOC. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA: IEEE, 2014. 41−48

77 Doumanoglou A, Kouskouridas R, Malassiotis S, Kim T K. Recovering 6D object pose and predicting next-best-view in the crowd. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 3583−3592

78 Tejani A, Tang D, Kouskouridas R, Kim T K. Latent-class hough forests for 3D object detection and pose estimation. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 462−477

79 Agrawal P, Carreira J, Malik J. Learning to see by moving. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 37−45

80 Vijayanarasimhan S, Ricco S, Schmid C, Sukthankar R, Fragkiadaki K. SfM-Net: learning of structure and motion from video. arXiv: 1704.07804, 2017.

81 Byravan A, Fox D. SE3-Nets: learning rigid body motion using deep neural networks. arXiv: 1606.02378, 2016.

82 Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). San Diego, CA, USA: IEEE, 2005. 539−546

83 Lengyel E S. Voxel-based Terrain for Real-time Virtual Simulations [Ph. D. dissertation], University of California, USA, 2010. 67−82

84 Wohlhart P, Lepetit V. Learning descriptors for object recognition and 3D pose estimation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 3109−3118

85 Hazirbas C, Ma L N, Domokos C, Cremers D. FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture. In: Proceedings of the 13th Asian Conference on Computer Vision. Taipei, China: Springer, 2016. 213−228

86 DeTone D, Malisiewicz T, Rabinovich A. Deep image homography estimation. arXiv: 1606.03798, 2016.

87 Liu F Y, Shen C H, Lin G S. Deep convolutional neural fields for depth estimation from a single image. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 5162−5170

88 Handa A, Bloesch M, Pătrăucean V, Stent S, McCormac J, Davison A. Gvnn: neural network library for geometric computer vision. *Computer Vision-ECCV 2016 Workshops*. Cham: Springer, 2016.

89 Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial transformer networks. In: Proceedings of the 2015 Advances in Neural Information Processing Systems. Montreal, Canada: Curran Associates, Inc., 2015. 2017−2025

90 Han X F, Leung T, Jia Y Q, Sukthankar R, Berg A C. MatchNet: unifying feature and metric learning for patch-based matching. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 3279−3286

91 Burgard W, Stachniss C, Grisetti G, Steder B, Kÿmmerle R, Dornhege C, Ruhnke M, Kleiner A, Tardös J D. A comparison of SLAM algorithms based on a graph of relations. In: Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems. St. Louis, MO, USA: IEEE, 2009. 2089−2095

92 Kümmerle R, Steder B, Dornhege C, Ruhnke M, Grisetti G, Stachniss C, Kleiner A. On measuring the accuracy of SLAM algorithms. *Autonomous Robots*, 2009, **27**(4): 387−407

93 Kaehler A, Bradski G. Open source computer vision library [Online], available: https://github.com/itseez/opencv, February 2, 2018

94 Furgale P, Rehder J, Siegwart R. Unified temporal and spatial calibration for multi-sensor systems. In: Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Tokyo, Japan: IEEE, 2013. 1280−1286

95 Snavely N, Seitz S M, Szeliski R. Photo tourism: exploring photo collections in 3D. *ACM Transactions on Graphics*, 2006, **25**(3): 835−846

96 Moulon P, Monasse P, Marlet R. OpenMVG [Online], available: https://github.com/openMVG/openMVG, December 9, 2017

97 Capel D, Fitzgibbon A, Kovesi P, Werner T, Wexler Y, Zisserman A. MATLAB functions for multiple view geometry [Online], available: http://www.robots.ox.ac.uk/~vgg/hzbook/code, October 14, 2017

98 Agarwal S, Mierle K. Ceres solver [Online], available: http://ceres-solver.org, January 9, 2018

99 Dellaert F. Factor Graphs and GTSAM: a Hands-on Introduction, Technical Report, GT-RIM-CP&R-2012-002, February 10, 2018

100 Kaess M, Ranganathan A, Dellaert F. iSAM: incremental smoothing and mapping. *IEEE Transactions on Robotics*, 2008, **24**(6): 1365−1378

101 Polok L, Ila V, Solony M, Smrz P, Zemcik P. Incremental block cholesky factorization for nonlinear least squares in robotics. In: Proceedings of the 2013 Robotics: Science and Systems. Berlin, Germany: MIT Press, 2013. 1−7

102 Vedaldi A, Fulkerson B. VLFeat: an open and portable library of computer vision algorithms [Online], available: http://www.vlfeat.org/, November 5, 2017

103 Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the KITTI vision benchmark suite. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI, USA: IEEE, 2012. 3354−3361

104 Sturm J, Engelhard N, Endres F, Burgard W, Cremers D. A benchmark for the evaluation of RGB-D slam systems. In: Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robot and Systems (IROS). Vilamoura, Portugal: IEEE, 2012. 573−580

105 Handa A, Whelan T, McDonald J, Davison A J. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In: Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA). Hong Kong, China: IEEE, 2014. 1524−1531

106 Burri M, Nikolic J, Gohl P, Schneider T, Rehder J, Omari S, Achtelik M W, Siegwart R. The EuRoC micro aerial vehicle datasets. *The International Journal of Robotics Research*, 2016, **35**(10): 1157−1163

107 Mayer N, Ilg E, Häusser P, Fischer P, Cremers D, Dosovitskiy A, Brox T. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the 2016 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 4040−4048

108 Pronobis A, Caputo B. COLD: the COsy localization database. *The International Journal of Robotics Research*, 2009, **28**(5): 588−594

109 Silberman N, Hoiem D, Kohli P, Fergus R. Indoor segmentation and support inference from RGBD images. In: Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: ACM, 2012. 746−760

110 Silberman N, Fergus R. Indoor scene segmentation using a structured light sensor. In: Proceedings of the 2011 IEEE International Conference on Computer Vision Workshop. Barcelona, Spain: IEEE, 2011. 601−608

111 Xiang Y, Mottaghi R, Savarese S. Beyond PASCAL: a benchmark for 3D object detection in the wild. In: Proceedings of the 2014 IEEE Winter Conference on Applications of Computer Vision (WACV). Steamboat Springs, CO, USA: IEEE, 2014. 75−82

112 Nikolskiy V P, Stegailov V V, Vecher V S. Efficiency of the tegra K1 and X1 systems-on-chip for classical molecular dynamics. In: Proceedings of the 2016 International Conference on High Performance Computing and Simulation (HPCS). Innsbruck, Austria: IEEE, 2016. 682−689

113 Pizzoli M, Forster C, Scaramuzza D. REMODE: probabilistic, monocular dense reconstruction in real time. In: Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA). Hong Kong, China: IEEE, 2014. 2609−2616

114 Faessler M, Fontana F, Forster C, Mueggler E, Pizzoli M, Scaramuzza D. Autonomous, vision-based flight and live dense 3D mapping with a quadrotor micro aerial vehicle. *Journal of Field Robotics*, 2016, **33**(4): 431−450

115 Weiss S, Achtelik M W, Chli M, Siegwart R. Versatile distributed pose estimation and sensor self-calibration for an autonomous MAV. In: Proceedings of the 2012 IEEE International Conference on Robotics and Automation (ICRA). Saint Paul, MN, USA: IEEE, 2012. 31−38

116 Weiss S, Siegwart R. Real-time metric state estimation for modular vision-inertial systems. In: Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA). Shanghai, China: IEEE, 2011. 4531−4537

117 Lynen S, Achtelik M W, Weiss S, Chli M, Siegwart R. A robust and modular multi-sensor fusion approach applied to MAV navigation. In: Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Tokyo, Japan: IEEE, 2013. 3923−3929

丁文东 中国科学院自动化研究所博士研究生. 2013 年获得武汉理工大学信息工程学院电子科学与技术学士学位. 主要研究方向为视觉测量及定位技术.
E-mail: dingwendong2013@ia.ac.cn
(**DING Wen-Dong** Ph. D. candidate at the Institute of Automation, Chinese Academy of Sciences. He received his bachelor degree in electronic science and technology from Wuhan University of Technology in 2013. His research interest covers visual localization and measurement.)

徐 德 中国科学院自动化研究所研究员. 1985 年、1990 年获得山东科技大学学士、硕士学位. 2001 年获得浙江大学博士学位. 主要研究方向为机器人视觉测量, 视觉伺服, 显微视觉技术. 本文通信作者. E-mail: de.xu@ia.ac.cn
(**XU De** Professor at the Institute of Automation, Chinese Academy of Sciences. He received his bachelor degree and master degree from Shandong University of Technology in 1985 and 1990, and received his Ph. D. degree from Zhejiang University in 2001. His research interest covers robot vision measurement, visual servoing, and microvisual technology. Corresponding author of this paper.)

刘希龙 中国科学院自动化研究所副研究员. 2009 年获得北京交通大学学士学位. 2014 年获得中国科学院自动化研究所博士学位. 主要研究方向为图像处理, 模式识别, 视觉测量.
E-mail: xilong.liu@ia.ac.cn
(**LIU Xi-Long** Associate professor at the Institute of Automation, Chinese Academy of Sciences. He received his bachelor degree from Beijing Jiaotong University in 2009, and his Ph. D. degree from the Institute of Automation, Chinese Academy of Sciences in 2014. His research interest covers image processing, pattern recognition, visual measurement, and visual scene cognition.)

张大朋 中国科学院自动化研究所副研究员. 2003 年、2006 年获得河北科技大学学士、硕士学位. 2011 年获得北京航空航天大学博士学位. 主要研究方向为机器人视觉测量, 医疗机器人.
E-mail: dapeng.zhang@ia.ac.cn
(**ZHANG Da-Peng** Associate professor at the Institute of Automation, Chinese Academy of Sciences. He received his bachelor degree and master degree from Hebei University of Science and Technology, in 2003 and 2006, and received his Ph. D. degree from Beijing University of Aeronautics and Astronautics, in 2011. His research interest covers robot vision measurement and medical robot.)

陈 天 中国科学院自动化研究所硕士研究生. 2016 年获得北京邮电大学学士学位. 主要研究方向为视觉定位及三维重建技术.
E-mail: chentian2016@ia.ac.cn
(**CHEN Tian** Master student at the Institute of Automation, Chinese Academy of Sciences. She received her bachelor degree from Beijing University of Posts and Telecommunications in 2016. Her research interest covers visual localization and reconstruction.)