基于单目视觉的同时定位与地图构建方法综述

刘浩敏[1)], 章国锋[1,2)*], 鲍虎军[1)]

[1)] (浙江大学 CAD&CG 国家重点实验室 杭州 310058)
[2)] (浙江大学工业信息物理融合系统协同创新中心 杭州 310058)
(zhangguofeng@cad.zju.edu.cn)

摘 要: 增强现实是一种在现实场景中无缝地融入虚拟物体或信息的技术, 能够比传统的文字、图像和视频等方式更高效、直观地呈现信息, 有着非常广泛的应用. 同时定位与地图构建作为增强现实的关键基础技术, 可以用来在未知环境中定位自身方位并同时构建环境三维地图, 从而保证叠加的虚拟物体与现实场景在几何上的一致性. 文中首先简述基于视觉的同时定位与地图构建的基本原理; 然后介绍几个代表性的基于单目视觉的同时定位与地图构建方法并做深入分析和比较; 最后讨论近年来研究热点和发展趋势, 并做总结和展望.

关键词: 增强现实; 同时定位与地图构建; 运动推断结构; 多视图几何; 摄像机跟踪
中图法分类号: TP391.41

A Survey of Monocular Simultaneous Localization and Mapping

Liu Haomin[1)], Zhang Guofeng[1,2)*], and Bao Hujun[1)]

[1)] (*State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou* 310058)
[2)] (*Collaborative Innovation Center for industrial Cyber-Physical System, Zhejiang University, Hangzhou* 310058)

Abstract: Augmented reality (AR) is a technique that allows to seamlessly composite virtual objects or information into real scene. Compared to traditional text, images and videos, AR is a more effective and intuitive way for information presentation and has wide applications. Simultaneous localization and mapping (SLAM) is a key fundamental technique for augmented reality, which provides the ability of self-localization in an unknown environment and mapping the 3D environment simultaneously. The localization and mapping enables fusion of virtual objects and real scenes in a geometrically consistent way. In this paper, we describe the basic principles of Visual SLAM, and introduce some state-of-the-art monocular SLAM methods with deep analysis and comparison. Finally, we discuss some research tendency in recent years and make conclusions.

Key words: augmented reality; simultaneous localization and mapping; structure-from-motion; multi-view geometry; camera tracking

对于很多应用来说, 传统的信息表达方式(如文字、图片、视频)和呈现方式(如二维浏览)显得低效、不够直观. 增强现实是一种在现实场景中无缝地融入虚拟物体或信息的技术, 它能比传统方式更为高效、直观地表达和呈现信息, 因而有着非常广阔的应用前景, 近年来已经在各种应用中崭露头角. 例如, 利用增强现实技术可以允许用户拿起智能手机或平板电脑即可观察所选中的家具在

收稿日期: 2016-04-30; 修回日期: 2016-05-11. 基金项目: 国家科技支撑计划(2012BAH35B02); 国家自然科学基金(61232011, 61272048); 中央高校基本科研业务费专项资金(2015XZZX005-05); 全国优秀博士学位论文作者专项资金资助项目(201245). 刘浩敏(1987—), 男, 博士研究生, 主要研究方向为运动推断结构、同时定位与地图构建; 章国锋(1981—), 男, 博士, 副教授, 博士生导师, 论文通讯作者, 主要研究方向为三维视觉、增强现实、计算机图形学等; 鲍虎军(1966—), 男, 博士, 教授, 博士生导师, CCF 常务理事, 主要研究方向为计算机图形学、三维视觉、虚拟现实、增强现实等.

自己房间里的摆放和搭配效果，比起只有图片或文字信息的展示方式，更为高效、直观，无需想象.

增强现实需要实时定位设备在环境中的方位. 定位方案虽然已经有很多种，但多数方案要么在实际应用中存在诸多局限，要么代价太高难以普及. 比如，GPS 无法在室内及遮挡严重的室外环境中使用，且定位精度较低；高精度的惯导系统成本太高且难以民用；基于无线信号的定位方案需要事先布置使用场景等. 基于视觉的同时定位与地图构建技术(visual simultaneous localization and mapping, V-SLAM)以其硬件成本低廉(一个普通摄像头即可)、小场景范围内精度较高、无需预先布置场景等优势，成为目前一个较常采用的定位方案. 尤其在增强现实应用中，由于虚拟物体的叠加目标通常为图像/视频，因此基于图像/视频等视觉信息的 V-SLAM 方案，对于确保虚实融合结果在几何上保持一致有着天然的优势.

同时定位与地图构建(simultaneous localization and mapping, SLAM)最早源于机器人领域[1-4]，其目标是在一个未知的环境中实时重建环境的三维结构并同时对机器人自身进行定位. 在计算机视觉领域，与之类似的技术是运动推断结构(structure-from-motion, SFM)[5]. 早期的 SFM 技术一般是离线处理的，后来随着技术的发展出现了实时的 SFM 技术，可以归入到 V-SLAM. SLAM 技术已经发展了几十年，研究人员已经做了大量的工作，而且也出现了一些关于 SLAM 的综述和教程[3-4,6-9]. 但是这些综述性文献大多偏向于介绍基于滤波的 SLAM 技术，或者只是对各类 SLAM 方法、原理和常用模块进行介绍，没有深入系统地对各类 SLAM 方法进行性能分析和比较；而且多数文献的发表时间也比较早(其中最经典的综述性文献是 Durrant-Whyte 等[3-4]于 2006 年撰写的关于 SLAM 的教程，距今已经 10 年了)，不能反映最新的 SLAM 技术发展潮流. 不同于这些文献，本文主要专注于对基于单目视觉的 SLAM 方法的分析和讨论，系统地介绍和分析目前 3 类主流单目 V-SLAM 方法的优缺点，并对它们的代表性系统进行性能分析和比较. 另外，本文也介绍和讨论了 V-SLAM 技术的最新研究热点和发展趋势，并进行总结和展望.

1 V-SLAM 的基本原理

V-SLAM 技术可以根据拍摄的视频信息推断出摄像头在未知环境中的方位，并同时构建环境地图，其基本原理为多视图几何原理[5]，图 1 所示为一个示意图. 其中只用单目摄像头的 V-SLAM 技术又称为单目 V-SLAM，也是本文要重点分析讨论的. V-SLAM 的目标为同时恢复出每帧图像对应的相机运动参数 $C_1\cdots C_m$，及场景三维结构 $\boldsymbol{X}_1\cdots\boldsymbol{X}_n$. 每个相机运动参数 C_i 包含了相机的位置和朝向信息，通常表达为一个 3×3 的旋转矩阵 $\boldsymbol{R}_i$ 和一个三维位置变量 $\boldsymbol{p}_i$. $\boldsymbol{R}_i$，$\boldsymbol{p}_i$ 将一个世界坐标系下的三维点 $\boldsymbol{X}_j$ 变换至 C_i 的局部坐标系

$$(\boldsymbol{X}_{ij},\boldsymbol{Y}_{ij},\boldsymbol{Z}_{ij})^{\mathrm{T}}=\boldsymbol{R}_i(\boldsymbol{X}_j-\boldsymbol{p}_i) \tag{1}$$

进而投影至图像中

$$\boldsymbol{h}_{ij}=(f_x\boldsymbol{X}_{ij}/\boldsymbol{Z}_{ij}+c_x,f_y\boldsymbol{Y}_{ij}/\boldsymbol{Z}_{ij}+c_y)^{\mathrm{T}} \tag{2}$$

其中，f_x，f_y 分别为沿图像 x，y 轴的图像焦距，(c_x,c_y) 为镜头光心在图像中的位置，通常假设这些参数已事先标定且保持不变. 由式(1)(2)可知，三维点在图像中的投影位置 $\boldsymbol{h}_{ij}$ 可表示为一个关于 C_i 和 $\boldsymbol{X}_j$ 的函数，记为

$$\boldsymbol{h}_{ij}=h(C_i,\boldsymbol{X}_j) \tag{3}$$

V-SLAM 算法需要将不同图像中对应于相同场景点的图像点匹配起来(如图 1 中，将对应于 $\boldsymbol{X}_1$ 的图像点 $\boldsymbol{x}_{11}$，$\boldsymbol{x}_{21}$，$\boldsymbol{x}_{31}$ 进行匹配). 通过求解优化如下目标函数

$$\underset{C_1\cdots C_m,\boldsymbol{X}_1\cdots\boldsymbol{X}_n}{\arg\min}\sum_{i=1}^{m}\sum_{j=1}^{n}\left\|h(C_i,\boldsymbol{X}_j)-\hat{\boldsymbol{x}}_{ij}\right\|_{\boldsymbol{\Sigma}_{ij}} \tag{4}$$

得到一组最优的 $C_1\cdots C_m$，$\boldsymbol{X}_1\cdots\boldsymbol{X}_n$，使得所有 $\boldsymbol{X}_j$ 在 C_i 图像中的投影位置 $\boldsymbol{h}_{ij}$ 与观测到的图像点位置 $\boldsymbol{x}_{ij}$ 尽可能靠近. 这里假设图像观测点符合高斯分布 $\boldsymbol{x}_{ij}\sim N(\hat{\boldsymbol{x}}_{ij},\boldsymbol{\Sigma}_{ij})$，$\|\boldsymbol{e}\|_{\boldsymbol{\Sigma}}=\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{e}$. 求解目标函数(4)的过程也称为集束调整(bundle adjustment, BA)[10]，

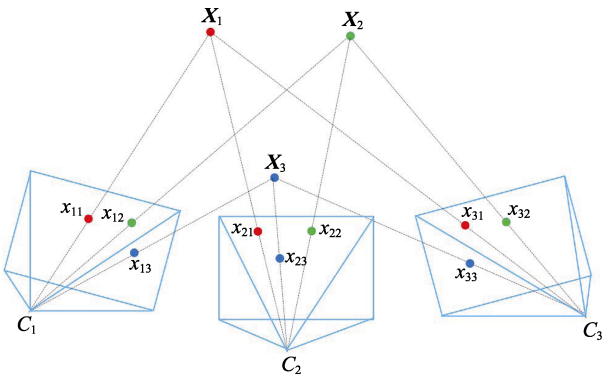


图 1 多视图几何原理

可利用线性方程的稀疏结构高效求解.

由于 V-SLAM 需要进行图像特征的匹配，因此其稳定性严重依赖于场景特征的丰富程度. 例如，相机拍摄一面纯色的白墙，那么仅从图像无法恢复出相机的运动. 加入其他传感器信息能很大程度地解决这一问题. 目前最常用的是在 V-SLAM 中结合 IMU 数据(加速度、角速度). 这样的 SLAM 称为 VIN(visual-aided inertial navigation)或 VI-SLAM (visual-inertial SLAM). 将相邻 2 帧 (C_i, C_{i+1}) 间的所有 IMU 数据标记为集合 $Z_i = \{z_1 \cdots z_{n_i}\}$，VI-SLAM 方法[11-13]一般求解优化如下目标函数

$$\arg\min_{C_1 \cdots C_m, X_1 \cdots X_n} \sum_{i=1}^{m}\sum_{j=1}^{n} \left\| h(C_i, \boldsymbol{X}_j) - \hat{\boldsymbol{x}}_{ij} \right\|_{\boldsymbol{\Sigma}_{ij}} + \sum_{i=1}^{m-1} \left\| f(C_i, Z_i) - C_{i+1} \right\|_{\boldsymbol{\Gamma}_i} \tag{5}$$

与目标函数(4)相比，VI-SLAM 引入了一个运动方程，其中 $f(C_i, Z_i)$ 为 Z_i 作用于 C_i 后的运动参数，$\boldsymbol{\Gamma}_i$ 为运动方程的协方差矩阵. 常见的运动方程有连续时间系统(Continuous Time System)[14]、预积分(Preintegration)方程[15]等. 通常，VI-SLAM 需要求解每一时刻的运动速度 $\boldsymbol{v}_i$ 和 IMU 数据的偏移量 $\boldsymbol{b}_i$，即 $C_i = (\boldsymbol{R}_i, \boldsymbol{p}_i, \boldsymbol{v}_i, \boldsymbol{b}_i)$.

类似的原理同样可应用于其他传感器数据，如引入 GPS 数据 $\boldsymbol{p}_i^G$，只需在能量函数中再引入一项

$$\arg\min_{C_1 \cdots C_m, X_1 \cdots X_n} \sum_{i=1}^{m}\sum_{j=1}^{n} \left\| h(C_i, \boldsymbol{X}_j) - \hat{\boldsymbol{x}}_{ij} \right\|_{\boldsymbol{\Sigma}_{ij}} + \sum_{i=1}^{m-1} \left\| f(C_i, Z_i) - C_{i+1} \right\|_{\boldsymbol{\Gamma}_i} + \sum_{i=1}^{m-1} \left\| \boldsymbol{p}_i - \hat{\boldsymbol{p}}_i^G \right\|_{\boldsymbol{\Lambda}_i} \tag{6}$$

这里假设 GPS 观测值符合高斯分布

$$\boldsymbol{p}_i^G \sim N(\hat{\boldsymbol{p}}_i^G, \boldsymbol{\Lambda}_i).$$

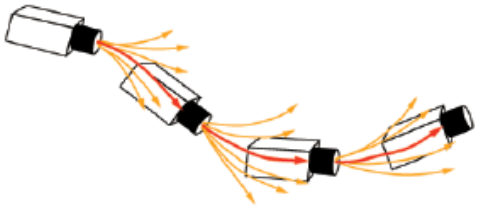
2　代表性单目 V-SLAM 系统

目前，国际上主流的 V-SLAM 方法大致可以分为 3 类：基于滤波器、基于关键帧 BA 和基于直接跟踪的 V-SLAM. 本节通过几个代表性的单目 V-SLAM 系统介绍这些方法，并分析其优劣.

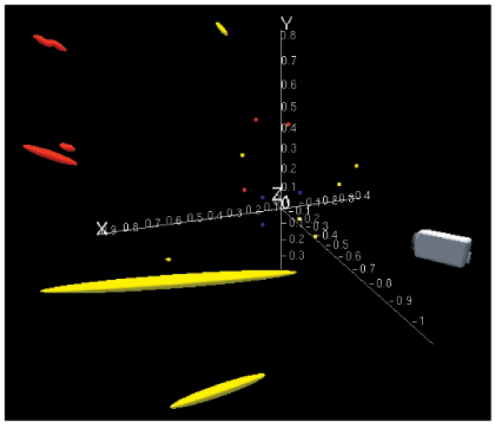
2.1　基于滤波器的 V-SLAM

基于滤波器的 V-SLAM 的基本思想如下：将每一时刻 t 的系统状态用一个高斯概率模型表达，$\boldsymbol{x}_t \sim N(\hat{\boldsymbol{x}}_t, \boldsymbol{P}_t)$，$\hat{\boldsymbol{x}}_t$ 为当前时刻系统状态估计值，$\boldsymbol{P}_t$ 为该估计值误差的协方差矩阵. 系统状态由一个滤波器不断更新. 不同的状态设计和滤波方式衍生出不同的 SLAM 系统. 本节介绍 2 款基于滤波器的 SLAM 系统 MonoSLAM[16]和 MSCKF[17].

MonoSLAM 是由 Davison 等发明的第一个成功基于单目摄像头的纯视觉 SLAM 系统. MonoSLAM 的状态 $\boldsymbol{x}_t$ 由 t 时刻的相机运动参数 C_t 和所有三维点位置 $\boldsymbol{X}_1 \cdots \boldsymbol{X}_n$ 构成，每一时刻的相机方位均带有一个概率偏差(如图 2a 所示)；同样，每个三维点位置也带有一个概率偏差，可以用一个三维椭球表示，椭球中心为估计值，椭球体积表明不确定程度(如图 2b 所示)；不同场景点之间，以及场景点和 C_t 之间均有概率关联. 在此概率模型下，场景点投影至图像的形状为一个投影概率椭圆(如图 2c



a. 相机运动模型



b. 场景点概率分布

c. 主动式特征匹配

图 2　MonoSLAM 的相机运动模型和三维点跟踪[16]

所示). MonoSLAM 为每帧图像中抽取 Shi-Tomasi 角点[18]，在投影椭圆中主动搜索(active search)[19]特征点匹配.

MonoSLAM 的滤波器选用的是扩展卡尔曼滤波器(extended Kalman filter, EKF). 在预测(Prediction)阶段，采用运动方程

$$C_t = f(C_{t-1}, \boldsymbol{a}_v \Delta t, \boldsymbol{a}_\omega \Delta t) \tag{7}$$

其中 $\boldsymbol{a}_v$ 和 $\boldsymbol{a}_\omega$ 分别为线性和旋转加速度，算法假设 $\boldsymbol{a}_v \sim N(0, \boldsymbol{\Gamma}_v)$，$\boldsymbol{a}_\omega \sim N(0, \boldsymbol{\Gamma}_\omega)$，$\Delta t$ 为相邻 2 帧时间差. 在更新(Update)阶段，采用投影方程

$$\hat{\boldsymbol{x}}_j = h(C_t, \boldsymbol{X}_j) + \boldsymbol{n}_j \tag{8}$$

其中，$\hat{\boldsymbol{x}}_j$ 为当前帧观测到 $\boldsymbol{X}_j$ 的图像点位置，$\boldsymbol{n}_j \sim N(0, \boldsymbol{\Sigma}_j)$. 与式(5)相比，MonoSLAM 每一时刻仅需估计当前时刻状态 x_t，之前所有时刻的相机运动参数 $\{C_1 \cdots C_{t-1}\}$ 全部不参与计算，由此简化计算量.

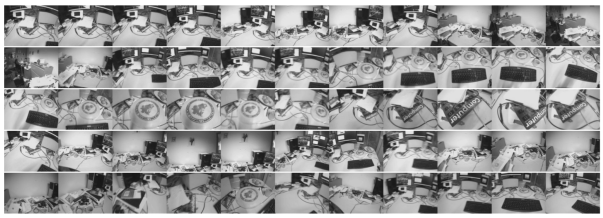
传统 EKF 方法最主要的局限性如下：1) 如果预测函数和更新函数为非线性(通常 V-SLAM 问题都是非线性的)，那么 EKF 并不能保证全局最优，与 Levenberg-Marquardt[20]等迭代的非线性优化技术相比，更容易造成误差累积. 如果上一帧处理完成时刻的估计值尚未精确，传递至下一帧的先验信息便带有误差；由于上一帧状态不再变化，所以先验信息中的误差便无法消除，误差不断向后传递造成误差累积. 2) 若将三维点引入状态变量，则每一时刻的计算复杂度为 $O(n^3)$，因此只能处理几百个点的小场景.

为了缓解EKF方法的计算复杂度问题, Mourikis 等[17]于 2007 年提出了 MSCKF. MSCKF 是一个 VI-SLAM 方法，在预测阶段，使用 IMU 数据进行传递系统状态[14]；在更新阶段, MSCKF 将邻近的 l 帧相机运动参数全部包含进一个状态变量集合 $C = \{C_{t-l+1} \cdots C_t\}$. C 中每个 C_i 的估计值均在不断优化，通常移出 C 前 C_i 已较为精确，由此缓解误差累积. 除此之外，MSCKF 对所有三维点进行消元(Marginalization)，将 C_i 与 $\boldsymbol{X}_j$ 间的二元约束转换为 $\{C_{t-l+1} \cdots C_t\}$ 间的多元约束，从而将 $O(n^3)$ 的计算复杂度简化为 $O(nl^3)$. 因为一般来说 $l << n$，而且为常数，所以该方法可以大大降低计算复杂度，将原来的跟三维点数成立方关系降到了线性关系.

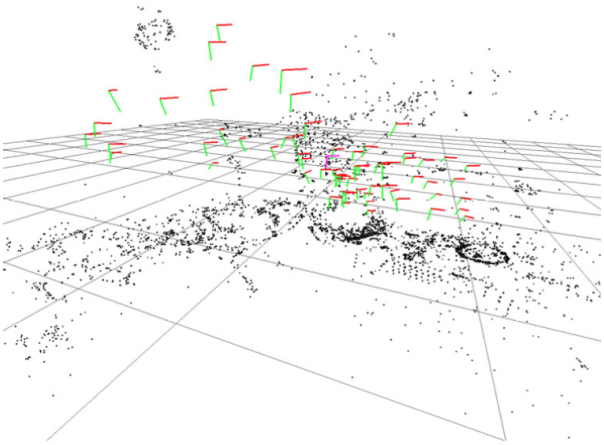
2.2 基于关键帧 BA 的 V-SLAM

PTAM 是实时 SFM 系统，也是首个基于关键帧 BA 的单目 V-SLAM 系统，由 Klein 等[21]于 2007 年提出并开源，又于 2009 年移植到 iPhone 3G 上[22].

PTAM的基本思想如下：将相机跟踪(Tracking)和地图构建(Mapping)作为 2 个独立的任务在 2 个线程并行执行. 地图构建线程仅维护原视频流中稀疏抽取的关键帧(如图 3a 所示)及关键帧中可见的三维点(如图 3b 所示)，这样就可以非常高效地求解目标函数(4)(即 BA)；有了 BA 恢复的精确三维结构，相机跟踪线程作为前台线程，仅需优化当前帧运动参数 C_t，足以达到实时的计算效率.



a. 关键帧图像



b. 关键帧和三维点

图 3 PTAM 选择的关键帧和重建结果[21]

具体来说，前台线程通过一个匀速运动模型预测当前帧方位，以搜索地图中的三维点在当前帧图像中对应的 FAST 角点[23]，并根据匹配关系优化当前帧方位

$$\arg\min_{C_t} \sum_{j=1}^{n} w_j \left\| h(C_t, \boldsymbol{X}_j) - \hat{\boldsymbol{x}}_j \right\|_{\boldsymbol{\Sigma}_j} \tag{9}$$

其中 w_j 是 Tukey 函数[24]对应的权重，用于缓解误匹配(Outliers)对结果的影响. 如果成功匹配点(Inliers)数不足(如因图像模糊、快速运动等)，则判断跟踪失败，开始重定位[25]——将当前帧与已有关键帧的缩略图进行比较，选择最相似的关键帧作为当前帧方位的预测，重复上述特征匹配和方位优化步骤. 如果跟踪成功，判断 C_t 是否满足关键帧条件，一旦符合，系统将当前帧作为新的关键帧传递

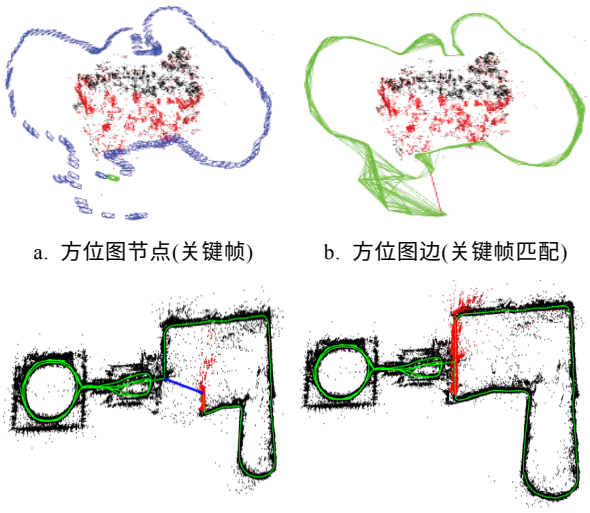
给后台构建地图. 后台线程沿极线(Epipolar Line)[5]匹配不同关键帧之间对应于相同场景点的图像特征点，通过三角化(Triangulation)[5]恢复这些场景点三维位置，并对所有关键帧和三维点运行 BA，恢复精确的三维地图.

PTAM 的开源对于 V-SLAM 的发展来说意义深远，目前市面上很多 V-SLAM 系统都是基于 PTAM 的算法框架改进而来. Mur-Artal 等[26]于 2015 年提出并开源的 ORB-SLAM 是目前性能最好的单目 V-SLAM 系统之一.

ORB-SLAM 基本延续了 PTAM 的算法框架，但对框架中的大部分组件都做了改进，归纳起来主要有 4 点: 1) ORB-SLAM 选用了 ORB 特征[27]，基于 ORB 描述量的特征匹配和重定位[28]，都比 PTAM 具有更好的视角不变性. 此外，新增三维点的特征匹配效率更高，因此能更及时地扩展场景. 扩展场景及时与否决定了后续帧是否能稳定跟踪. 2) ORB-SLAM 加入了循环回路的检测和闭合机制，以消除误差累积. 系统采用与重定位相同的方法[28]来检测回路(匹配回路两侧关键帧上的公共点)，通过方位图(Pose Graph)优化来闭合回路. 关键帧作为方位图的节点(如图 4a 所示)，每个关键帧被赋予一个相似变换 ξ_i 以矫正其方位；方位图的边表示关键帧之间存在特征匹配(如图 4b 所示)，2 团匹配的三维点云通过坐标对齐可求解一个相似变换 ξ_{ij}；系统选用 g2o[29]优化方位图，以闭合回路(如图 4c, 4d 所示)

$$\arg\min_{\xi_1,\cdots,\xi_m}\sum_{(\xi_{ij},\Sigma_{ij})}(\xi_{ij}\circ\xi_i^{-1}\circ\xi_j)^{\mathrm{T}}\Sigma_{ij}^{-1}(\xi_{ij}\circ\xi_i^{-1}\circ\xi_j) \quad (10)$$

其中，Σ_{ij} 为 ξ_{ij} 的协方差矩阵(文中设为单位阵)，操作符$\circ$按顺序连接 2 个相似变换. 与全局 BA 相比，方位图优化极大简化了全局优化的计算量，因此 ORB-SLAM 能处理大尺度场景. BA 在系统中只运行于局部场景(如图 4 中的红色三维点). 3) PTAM 需要用户指定 2 帧来初始化系统，2 帧间既要有足够的公共点，又要有足够的平移量. 平移运动为这些公共点提供视差(Parallax)，只有足够的视差才能三角化出精确的三维位置. ORB-SLAM 通过检测视差来自动选择初始化的 2 帧. 4) PTAM 扩展场景时也要求新加入的关键帧提供足够的视差，导致场景往往难以扩展. ORB-SLAM 采用一种更鲁棒的关键帧和三维点的选择机制——先用宽松的判断条件尽可能及时地加入新的关键帧和三维点，以保证后续帧的鲁棒跟踪；再用严格的判断条件删除冗余的关键帧和不稳定的三维点，以保证 BA 的效率和精度.



a. 方位图节点(关键帧)　　b. 方位图边(关键帧匹配)

c. 循环回路闭合前　　d. 循环回路闭合后

图 4　ORB-SLAM 的方位图优化和循环回路闭合[26]

浙江大学 CAD&CG 国家重点实验室计算机视觉组于 2013 年研发了 RDSLAM[30]，该系统在吸收 PTAM 的关键帧表达和并行跟踪/重建框架的基础上，采用 SIFT[31]特征点和在线的关键帧表达与更新方法，可以自适应地对动态场景进行建模，从而能够实时有效地检测出场景的颜色和结构等变化并正确处理. 此外，RDSLAM 对传统的 RANSAC[32]方法进行了改进，提出一种基于时序先验的自适应 RANSAC 方法，即使在正确匹配点比例很小的情况下也能快速可靠地将误匹配点去掉，从而实现复杂动态场景下的摄像机姿态的实时鲁棒求解. 该方法尤其适合处理场景在逐渐改变的情况，这是其他 V-SLAM 方法难以处理的.

2.3　基于直接跟踪的 V-SLAM

基于滤波器和基于关键帧 BA 的 V-SLAM 通常都需要在图像中提取并匹配特征点，因此对环境特征的丰富程度和图像质量(如模糊程度、图像噪声等)十分敏感. 相比之下，直接跟踪法(Direct Tracking)不依赖于特征点的提取和匹配，而是直接通过比较像素颜色来求解相机运动，因此通常在特征缺失、图像模糊等情况下有更好的鲁棒性. 本节介绍 2 款基于直接跟踪的代表性单目 V-SLAM 系统 DTAM[33]和 LSD-SLAM[34].

DTAM 是 Newcombe 等于 2011 年提出的单目 V-SLAM 系统，其最显著的特点是能实时恢复场景三维模型(如图 5a 所示). 基于三维模型，DTAM 既能允许 AR 应用中的虚拟物体与场景发生物理碰撞(如图 5b 所示)，又能保证在特征缺失、图像模糊等情况下稳定地直接跟踪. 具体来说，DTAM 预测一个

与当前帧相机方位 C_t 十分接近的虚拟相机 C_v，在 C_v 下绘制场景三维模型，由此求解 C_v 和 C_t 间的相对运动 $\boldsymbol{\xi}_{tv}$，

$$\arg\min_{\boldsymbol{\xi}_{tv}} \sum_{\boldsymbol{x}\in\Omega_v} \left\| r(\boldsymbol{x}, D_v(\boldsymbol{x}), \boldsymbol{\xi}_{tv}) \right\|_2^2 \tag{11}$$

其中，$r(\cdot)$ 为颜色残差，

$$r(\boldsymbol{x}, D_v(\boldsymbol{x}), \boldsymbol{\xi}_{tv}) = I_v(\boldsymbol{x}) - I_t(\omega(\boldsymbol{x}, D_v(\boldsymbol{x}), \boldsymbol{\xi}_{tv})) \tag{12}$$

I_v 和 D_v 分别为对三维模型在 C_v 下绘制得到的亮度图和深度图，Ω_v 为亮度和深度有效像素的集合，函数 $\omega(\boldsymbol{x}, D_v(\boldsymbol{x}), \boldsymbol{\xi}_{tv})$ 将虚拟帧 v 中的像素 x 投影至当前帧 t 中.

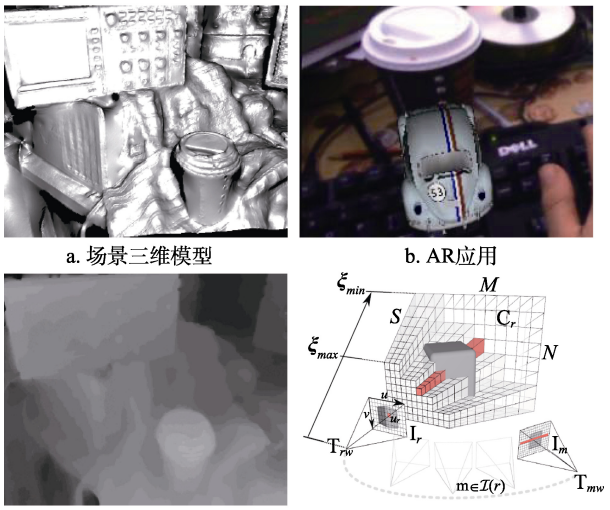


图 5 DTAM 的重建结果和逆深度表达[33]

为恢复场景三维模型，后台需要持续恢复参考帧 r 的深度图 D_r(如图 5c 所示). 这里选用的是逆深度(Inverse Depth)方式[35]表达深度. 如图 5d 所示，DTAM 将解空间离散为 $M\times N\times S$ 的三维网格，其中 $M\times N$ 为图像分辨率，S 为逆深度分辨率. DTAM 选择参考帧 r 周围的后续帧 $m\in N(r)$，对 r 中每个图像坐标为 x、逆深度为 $d=D_r(\boldsymbol{x})$ 的体素(Voxel)计算匹配代价

$$C(\boldsymbol{x}, d) = \frac{1}{|N(r)|} \sum_{m\in N(r)} \left\| r(\boldsymbol{x}, d, \boldsymbol{\xi}_{mr}) \right\|_1 \tag{13}$$

其中 $r(\cdot)$ 的求解见式(12). 仅通过 $\arg\min_d C(\boldsymbol{x}, d)$ 恢复的深度图在无纹理区域存在歧义，因此 DTAM 加入正则项(Regularization Term)

$$R(\boldsymbol{x}, d) = g(\boldsymbol{x}) \left\| \nabla D_r(\boldsymbol{x}) \right\|_\varepsilon \tag{14}$$

其中，$g(\boldsymbol{x})$ 为平滑权重，用于降低物体边界处的平滑程度

$$g(\boldsymbol{x}) = \mathrm{e}^{-\alpha \left\| \nabla I_r(\boldsymbol{x}) \right\|_2^\beta} \tag{15}$$

$\|\boldsymbol{x}\|_\varepsilon$ 为 Huber 范数，

$$\|\boldsymbol{x}\|_\varepsilon = \begin{cases} \|\boldsymbol{x}\|_2^2 / 2\varepsilon, & \text{if } \|\boldsymbol{x}\|_2 \leqslant \varepsilon \\ \|\boldsymbol{x}\|_1 - \varepsilon/2, & \text{otherwise} \end{cases} \tag{16}$$

用于保留深度不连续区域. DTAM 的能量函数为

$$\arg\min_{D_r} \int \left(R(\boldsymbol{x}, D_r(\boldsymbol{x})) + \lambda C(\boldsymbol{x}, D_r(\boldsymbol{x})) \right) \mathrm{d}\boldsymbol{x} \tag{17}$$

采用全变差(total variation, TV)技术[36]求解.

基于直接跟踪的 DTAM 对特征缺失、图像模糊有很好的鲁棒性，但由于 DTAM 为每个像素都恢复稠密的深度图，并且采用全局优化，因此计算量很大，即使采用 GPU 加速，模型的扩展效率仍然较低. Engel 等[37]于 2013 年提出了一套同样也是基于直接跟踪的视觉测量(visual odometry, VO)系统，该系统后扩展为 V-SLAM 系统 LSD-SLAM[34]，并开源了代码. 与 DTAM 相比，LSD-SLAM 仅恢复半稠密深度图(如图 6a 所示)，且每个像素深度独立计算，因此能达到很高的计算效率.

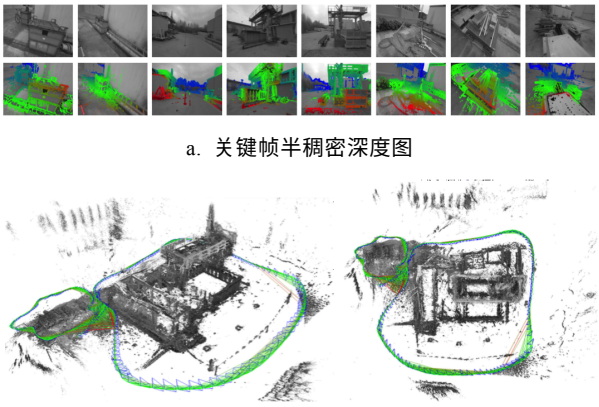


图 6 LSD-SLAM 的重建结果[34]

LSD-SLAM 采用关键帧表达场景，每个关键帧 k 包含图像 I_k、逆深度图 D_k 和逆深度的方差 V_k. 系统假设每个像素 x 的逆深度值服从高斯分布 $N(D_k(\boldsymbol{x}), V_k(\boldsymbol{x}))$. LSD-SLAM 的前台线程采用直接跟踪法恢复当前帧 t 与关键帧 k 之间相对运动 $\boldsymbol{\xi}_{tk}$，即求解优化式

$$\arg\min_{\boldsymbol{\xi}} \sum_{\boldsymbol{x}\in\Omega_k} \left\| \frac{r^2(\boldsymbol{x}, D_k(\boldsymbol{x}), \boldsymbol{\xi}_{tk})}{\sigma_r^2(\boldsymbol{x}, D_k(\boldsymbol{x}), \boldsymbol{\xi}_{tk})} \right\|_\delta \tag{18}$$

其中，Ω_k 是深度有效像素的集合；$r(\cdot)$ 的求解见式(12)；$\sigma^2(\boldsymbol{x}, \boldsymbol{\xi})$ 为 $r(\cdot)$ 的方差，用于减小深度误差对结果的影响，

$$\sigma_r^2(\boldsymbol{x}, D_k(\boldsymbol{x}), \boldsymbol{\xi}_{tk}) = 2\sigma_I^2 + \left(\frac{\partial r(\boldsymbol{x}, D_k(\boldsymbol{x}), \boldsymbol{\xi}_{tk})}{\partial D_k(\boldsymbol{x})}\right)^2 V_k(\boldsymbol{x}) \tag{19}$$

$\|\cdot\|_\delta$ 为阈值为 δ 的 Huber 范数，用于缓解 Outliers 对结果的影响. LSD-SLAM 的后台线程对关键帧中每个半稠密抽取的像素点 $\boldsymbol{x}$ (梯度显著区域)，在 I_t 中沿极线搜索 $I_k(\boldsymbol{x})$ 的对应点，得到新的逆深度观测值 $d_{\boldsymbol{x}}$ 及其方差 $\sigma_{\boldsymbol{x}}^2$，采用 EKF 更新 D_k 和 V_k

$$\begin{cases} D_k(\boldsymbol{x}) \leftarrow \dfrac{V_k(\boldsymbol{x}) d_{\boldsymbol{x}} + \sigma_{\boldsymbol{x}}^2 D(\boldsymbol{x})}{V_k(\boldsymbol{x}) + \sigma_{\boldsymbol{x}}^2} \\ V_k(\boldsymbol{x}) \leftarrow \dfrac{V_k(\boldsymbol{x})\sigma_{\boldsymbol{x}}^2}{V_k(\boldsymbol{x}) + \sigma_{\boldsymbol{x}}^2} \end{cases} \tag{20}$$

与 ORB-SLAM 类似，LSD-SLAM 也采用方位图优化，因此能闭合循环回路和处理大尺度场景. 系统为每个新加入的关键帧 k_i 选取距离最近的以及图像内容最相似[38]的已有关键帧集合 $\{k_j\}$，采用直接跟踪法求解所有 (k_i, k_j) 间的相似变换 $\boldsymbol{\xi}_{ij}$ 及其协方差 $\boldsymbol{\Sigma}_{ij}$ 构造方位图并求解目标函数(10).

2.4　分析和比较

上述各类单目 V-SLAM 系统的比较如表 1 所示. 需要指出的是，虽然MSCKF属于VI-SLAM而不是纯粹的 V-SLAM，但它采用基于滑动窗口和通过对所有三维点进行消元来降低计算复杂度的方法具有一定的代表性，可以直接用在 V-SLAM 上，因此这里还是将它与其他 V-SLAM 系统进行比较. 本节从如下方面分析和比较各类系统.

定位精度. 以 MonoSLAM 为代表的基于滤波器的 V-SLAM 由于容易在变量未精确时就进行消元，于是误差不断累积到下一帧，因此精度劣于以PTAM为代表的基于关键帧BA的V-SLAM[39]. MSCKF虽然也基于滤波器，但由于其推迟消元的机制，加上融合了 IMU 信息，通常能达到很高的精度. 比起同样基于关键帧BA的PTAM, RDSLAM和ORB-SLAM由于分别选用了匹配精度更高的 SIFT 特征和 ORB 特征，而且采用了比 PTAM 更为高效的 BA 实现，因此定位精度通常比 PTAM 更高；而且 ORB-SLAM 可以通过闭合回路消除误差累积，因此在大尺度场景下的定位精度最高. 基于直接跟踪法 DTAM 和 LSD-SLAM 对光照变化和动态干扰较为敏感，因此精度一般会劣于 ORB-SLAM. 根据文献[26]中在 TUM RGB-D 数据集[40]上的实验结果，LSD-SLAM 的定位误差比 ORB-SLAM 大 5~10 倍.

定位效率. MonoSLAM 的计算复杂度需要 $O(n^3)$，其中 n 为所有三维点个数，因此定位效率较低. MSCKF 的计算复杂度为 $O(nl^3)$，其中 n 为跟踪轨迹结束于当前帧的三维点个数，l 为系统维护的局部帧数. MSCKF 的定位效率和精度之间存在权衡，通常 l 越大精度越高，但效率就越低，反之亦然. PTAM 和 ORB-SLAM 的前台线程只需优化当前帧方位，定位效率最高. RDSLAM 虽然前台线程的定位方案与 PTAM 和 ORB-SLAM 类似，但定位效率受限于计算代价较高的SIFT特征.基于直接跟踪的 DTAM 和 LSD-SLAM 的定位效率主要取决于选取的图像分辨率，因此在效率和精度之间也存在权衡，通常分辨率越高精度越高，但效率越低，反之亦然.

场景尺度. MonoSLAM 由于其 $O(n^3)$ 的计算复杂度，因此只适用于几百个点的小场景. MSCKF 只维护局部地图，对场景尺度没有限制. PTAM 和 RDSLAM

表 1　各类单目 **V-SLAM** 系统比较

	基于滤波器		基于关键帧 BA			基于直接跟踪	
	MonoSLAM	MSCKF	PTAM	ORB-SLAM	RDSLAM	DTAM	LSD-SLAM
定位精度	✓	✓✓✓	✓✓	✓✓✓	✓✓	✓✓	✓
定位效率	✓	✓✓	✓✓✓	✓✓✓	✓✓	✓✓	✓✓
场景尺度	✓	✓✓✓✓	✓✓	✓✓✓✓	✓✓✓	✓	✓✓✓✓
特征缺失鲁棒性	✓	✓✓✓	✓	✓	✓	✓✓	✓✓
重定位能力	×	×	✓✓	✓✓✓	✓✓✓	✓✓	✓✓✓
快速运动鲁棒性	✓✓	✓✓✓✓	✓✓✓	✓✓✓✓	✓✓✓✓	✓✓✓	✓
扩展效率	✓✓✓	✓✓✓✓	✓✓	✓✓✓	✓✓✓	✓	✓
近似纯旋转扩展鲁棒性	✓✓✓	✓✓✓✓	✓	✓✓	✓	✓	✓
场景变化鲁棒性	✓	✓✓	✓	✓	✓✓✓	✓	✓
回路闭合能力	✓	×	×	✓✓✓	✓✓	×	✓✓✓

的场景尺度主要限制于全局 BA 和特征点匹配效率: PTAM 通常只能实时处理几千个点的中等尺度场景; RDSLAM 由于 BA 的效率更高, 而且其基于 KDTree 的匹配策略[41]也比 PTAM 将三维点投影到当前帧搜索的策略更高效, 因此能处理的尺度更大一些, 可以实时处理几万个点的场景规模. ORB-SLAM 和 LSD-SLAM 用高效的方位图优化替代全局 BA, 适用于较大尺度的场景. DTAM 由于需要维护和渲染整个场景的三维模型, 也只适用于小场景.

特征缺失鲁棒性. 特征缺失(比如场景缺乏丰富的纹理或者由于图像模糊导致)对几乎所有 V-SLAM 都有较大影响. 其中, 只有基于直接跟踪的 DTAM 和 LSD-SLAM 能通过利用稠密或半稠密的图像信息缓解特征依赖, 但也不能完全消除. MSCKF 属于 VI-SLAM, 即使出现短时间特征缺失的情况, 也可利用 IMU 数据来跟踪方位, 因此对临时特征缺失的鲁棒性最好.

快速运动和扩展效率. 能否处理相机快速运动, 一方面依赖于所采用的匹配方法在大运动情况下的鲁棒性, 另一方面也依赖于场景地图的扩展效率. MonoSLAM 采用 EKF 的方式预测特征点的位置并进行主动搜索, 这对于太大的运动会失败. PTAM 采用主要依靠运动预测的方式将关键帧上三维位置已知的特征点投影到当前帧来, 并通过基于金字塔模型的匹配来增加鲁棒性, 对快速运动的鲁棒性会比 MonoSLAM 高. 但扩展场景需要对三维位置未知的特征点进行暴力搜索特征匹配, 该步骤十分耗时, 只能放到后台进行, 因此对于快速场景扩展的效率会比 MonoSLAM 差. 与 PTAM 类似, DTAM 也先采用基于金字塔模型的整张图像对齐方法[42]估计一个大概的相机姿态, 然后将整个模型绘制到这个视点下与当前帧图像进行进一步的对齐优化. MSCKF, RDSLAM 和 ORB-SLAM 均采用对视角变化具有不变性的特征描述量(MSCKF 和 RDSLAM 选用 SIFT 特征, ORB-SLAM 选用 ORB 特征), 允许高效的全局特征匹配(RDSLAM 采用 KDTree[41], ORB-SLAM 采用词袋模型[43]), 因此对于大运动情况具有较好的鲁棒性(RDSLAM 实际上每帧都进行全局重定位). LSD-SLAM 假设相机平缓运动, 采用前一帧的相机姿态作初始化进行直接的图像对齐, 因此对快速运动最敏感. 类似于 PTAM, RDSLAM 和 ORB-SLAM 均在后台线程扩展场景, 但由于它们高效的特征匹配, RDSLAM 和 ORB-SLAM 扩展场景的效率优于 PTAM. MSCKF 因为利用了 IMU 信息, 即使匹配点还没有三角化也能扩展场景, 因此扩展场景的效率最高. LSD-SLAM 和 DTAM 由于需要恢复半稠密或稠密的深度图, 一方面计算量较大, 另一方面对于新扩展部分的深度往往需要处理至少若干帧才能收敛, 因此扩展效率最差.

重定位能力. 实际跟踪的时候难免遇到失败的情况, 这时需要用重定位技术从失败状态中恢复. MonoSLAM, MSCKF 不支持重定位. PTAM 是通过维护关键帧的缩略图, 跟丢的时候将当前帧与已有关键帧的缩略图进行比较, 找到最匹配的关键帧恢复初始的方位后, 再把地图中的三维点投影到当前帧来进行特征匹配和方位优化来实现重定位. 该方法的缺点是, 如果当前帧跟已有的关键帧的视角不够接近, 就不容易重定位成功. RDSLAM, ORB-SLAM 和 LSD-SLAM 都是采用对视角变化具有不变性的特征描述量并结合高效的检索方法(RDSLAM 采用 KDTree[41], ORB-SLAM 采用词袋模型[43], LSD-SLAM 采用 FAB-MAP 方法[38])来实现重定位, 即使当前帧与关键帧有较大的视角差异也能重定位, 鲁棒性会比 PTAM 的基于缩小模糊图匹配方式更好.

近似纯旋转扩展鲁棒性. 在实际使用中, 相机可能会以近似纯旋转的方式转向拍摄旁边的场景内容. MonoSLAM 因为每帧同时优化三维点和相机方位, 所以对近似纯旋转有很好的鲁棒性. MSCKF 也类似, 不同的是, MSCKF 仅当跟踪轨迹结束时才三角化出三维点, 并且立即消元, 而且用了 IMU 信息, 因此对近似纯旋转最鲁棒. PTAM 和 RDSLAM 在近似纯旋转扩展场景时容易因为视差不够无法三角化新的三维点, 导致无法加入新的关键帧, 从而造成后续帧跟踪丢失, 鲁棒性较差. ORB-SLAM, DTAM 和 LSD-SLAM 对纯旋转扩展的鲁棒性很大程度取决于后台场景地图的扩展及优化的效率, 能否及时得到高质量的三维地图是后续帧能否稳定跟踪的关键. 因为 ORB-SLAM 不需要恢复稠密深度, 计算效率高, 所以在近似纯旋转扩展情况下的鲁棒性要比 DTAM 和 LSD-SLAM 更好.

场景变化的鲁棒性. Mono-SLAM, MSCKF, PTAM, ORB-SLAM, DTAM, LSD-SLAM 等系统都假设场景是静止不变的, 如果场景变化很大就会跟踪失败. 不过这些系统都采用了 RANSAC[32]或鲁棒函数(如 Huber, Tukey 等)等策略提高鲁棒性,

如果场景中动态物体上的匹配点数相对于静态匹配点数的比例不是很大，一般能当作 Outliers 剔除掉；但是如果场景在不断改变，而且大部分区域都发生了变化，那么这些系统就会失败. MSCKF 因为使用了 IMU 信息，理论上对动态变化的鲁棒性应该会好些. RDSLAM 由于会在线地检测场景的变化，识别出改变的三维点并更新关键帧，而且其提出的基于时序先验的自适应 RANSAC 方法也能很好地处理误匹配点，因此能很好地处理场景逐渐改变和动态物体干扰比较多的情况，对场景变化的鲁棒性最高. 当然，如果整个场景在很短的时间内完全发生改变，那么 RDSLAM 也无法处理.

回路闭合能力. MonoSLAM 没有显式的回路检测步骤，但如果系统状态一致(即协方差真实反映误差)，回路发生时仍有可能在当前帧的投影概率椭圆内搜索到回路上点的匹配点. RDSLAM 虽然没有显式的回路检测，但是由于采用 SIFT 特征与地图里的三维点进行全局匹配，回路发生时理论上可以检测出特征点的匹配关系并通过 BA 消除误差累积来闭合；如果误差累积过大，匹配点有可能会被当作 Outliers 剔除掉，从而造成无法有效的闭合. MSCKF, PTAM 和 DTAM 均无回路闭合机制. ORB-SLAM 和 LSD-SLAM 显式检测回路构建并优化方位图，回路闭合能力最强.

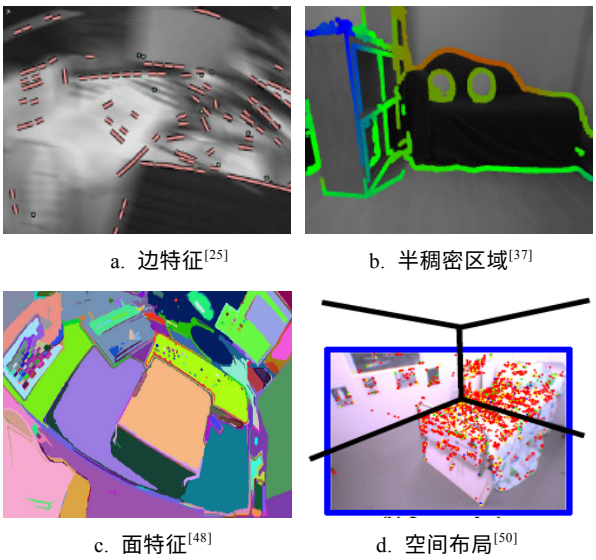
3　近年研究热点与发展趋势

随着技术的发展和越来越多开源系统的出现，V-SLAM 技术在逐渐趋于成熟. 然而，时至今日仍有很多实际问题尚待进一步研究解决.

3.1　缓解特征依赖

V-SLAM 最大的局限在于过于依赖场景特征. 基于直接跟踪的方法(第 2.3 节)通过直接对比像素颜色，避免了对特征缺失/图像模糊非常敏感的特征提取和匹配过程，从而很大程度上缓解了特征依赖. 然而，稠密或半稠密的直接跟踪会引入很大的计算量，若要运行在计算性能较低的移动设备上，就需要将图像降采样至很小的分辨率[44]，那么必然会降低跟踪精度. Forster 等[45]提出半直接视觉测量(semi-direct VO, SVO)，只对稀疏的特征点进行直接跟踪，能达到很高的效率.

V-SLAM 对场景特征的依赖，本质上是由于使用了过于底层的局部特征(点特征)，如果能利用边缘、平面等更为高层的图像信息，也能有效地缓解特征依赖. Klein 等[25]早在 2008 年就提出使用边特征来对抗图像模糊，如图 7a 所示. LSD-SLAM 实际上也是隐式地利用了图像边缘信息. 如图 7b 所示，大部分的半稠密区域均位于物体边缘. 近年来还有许多研究工作也将边缘信息用于 SFM 与 VO 中[46-47]. Concha 等[48]提出使用面特征将图像中颜色一致的区域近似为平面(如图 7c 所示)，称之为超像素(Superpixel). Concha 等[48-49]分别将超像素集成进 PTAM 和 LSD-SLAM 中，提高原系统鲁棒性. 此外，更为高层的空间布局对 V-SLAM 来说也是非常有价值的图像信息. 通常室内房间可以近似一个三维盒子(如图 7d 所示)，通过恢复房间盒子的三维参数辅助相机跟踪[50]或恢复高质量三维地图[51].



a. 边特征[25]　b. 半稠密区域[37]　c. 面特征[48]　d. 空间布局[50]

图 7　高层图像信息

3.2　稠密三维重建

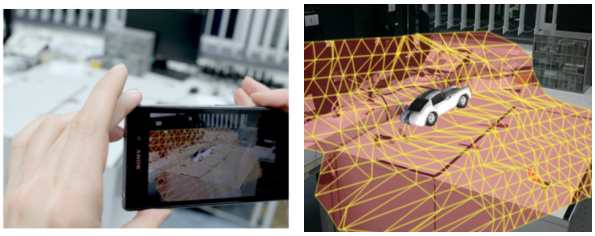
早期的 V-SLAM 大多只能实时重建出非常稀疏的三维点云. 若要获得更为稠密的三维信息，往往需要离线处理[52]，或是将实时数据传到性能强大的服务器进行计算[53]. 微软公司于 2010 年推出 RGB-D 摄像头 Kinect，能实时捕获彩色图像及深度图. 深度图不仅能辅助相机跟踪，还能提供稠密的场景三维信息. Newcombe 等[54]率先提出基于 RGB-D 的稠密 SLAM 系统 KinectFusion. KinectFusion 使用 TSDF(truncated signed distance function)[55]表达三维场景，算法在重建的三维物体上叠加一个均匀的三维网格，每个网格体素存储其中心到最近的三维物体表面间的距离. 基于 TSDF 绘制某一视角下的深度图，可使用 Marching Cube 算法[56]抽取 TSDF 为 0 的三维物体表面并在该视角下进行深度绘制，或直接进行光线投射(Ray Cast)至 TSDF 为

0 的深度. 为跟踪相机运动, KinectFusion 绘制全局模型在上一帧视角下的深度图, 并使用 ICP 算法[57]与当前帧深度图对齐, 求解 2 帧间相对运动. 恢复的相机方位将当前帧深度图注册至全局坐标下, 与全局模型融合, 更新 TSDF.

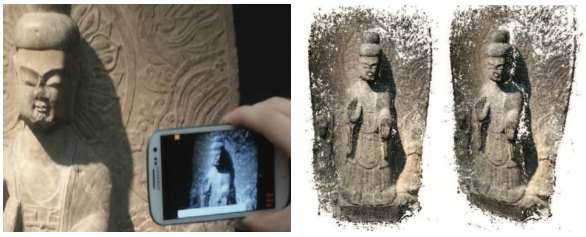
KinectFusion 最明显的局限性如下: 1) 预设的三维网格限制了场景尺度; 2) 基于 ICP 的相机跟踪严重依赖于场景几何特征的丰富程度; 3) 循环回路无法闭合. 为解决上述问题, Whelan 等开展了一系列研究工作: 文献[58]提出让三维网格随相机运动, 移出三维网格的区域不参与计算, 从而突破了场景尺度的限制; 文献[59]融合了 ICP, FOVIS[60]和 RGB-D[61] 3 种相机跟踪方法, 缓解跟踪对几何特征的依赖, 减小误差累积; 文献[62]采用 DBoW[43]检测循环回路从而构建方位图, 并采用文献[63]提出的方法形变方位图以闭合循环回路, 重建的三维模型也相应形变; 文献[64]提出使用面元(Surfel)[65]表达场景, 面元包含了三维点的位置、法向等信息, 形变直接作用于面元上而非通过方位图间接作用于三维模型, 从而获得更高的模型精度.

近年来出现了许多 V-SLAM 方法, 只用单目摄像头就能实时重建稠密的三维信息. 由于目前深度传感器仍未大范围普及, 尤其对于移动 AR 的应用来说, 深度摄像头获取深度的精度和范围受限于移动设备的成本和功耗, 因此基于单目摄像头的 V-SLAM 比基于 RGB-D 的 SLAM 更为实用. 第 2.3 节介绍的 DTAM 和 LSD-SLAM 都属于这类 V- SLAM 系统. Schöps 等[44]将 LSD-SLAM 移植到手机上, 并拟合大致的场景三维网格, 用于 AR 应用中的虚拟物体与真实场景间的物理碰撞 (如图 8a 所示). Pizzoli 等[66]提出一个实时的三维扫描系统 REMODE, 能够显示实时扫描的稠密三维点云引导用户拍摄. Tanskanen 等[67]和 Kolev 等[68]将手机作为三维扫描设备(如图 8b 所示), 利用手机上的 IMU 获得扫描物体的真实尺度. Pradeep 等[69]提出的 MonoFusion 系统能实时扫描出物体的三维模型. Ondruska 等[70]提出 MobileFusion, 用手机就能实现上述功能(如图 8c 所示). Schöps 等[71]利用谷歌推出的平板电脑 Project Tango 上的鱼眼摄像头, 实现大场景三维模型的实时重建 (如图 8d 所示).

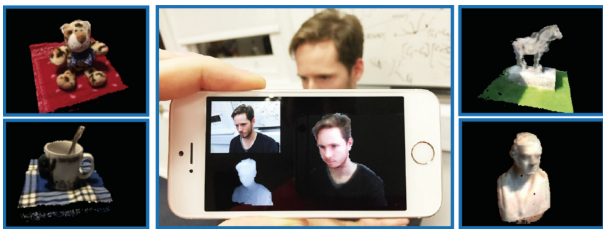
基于单目摄像头的稠密三维重建的难点在于需要实时恢复稠密的深度图, 这一过程通常都需要引入很大的计算量, 关键是如何权衡重建精度和计算效率. 通常, 算法选择 2 帧具有视差的图像



a. 基于稠密三维重建的 AR 应用[44]



b. 扫描稠密三维点云[67]



c. MobileFusion[70]

d. 大场景稠密三维重建[71]

图 8 基于单目摄像头的稠密三维重建

(I, I'), 通过立体匹配恢复粗略的深度图. 一个标准的方法是为 I 中的每个像素 x 在 I' 中对应的极线上搜索最相似的像素 x', 通常在选择局部图像块(Patch)采用 NCC(normalized cross correlation)作为相似性度量. 为加速匹配过程, 文献[67-68]对图像建造高斯金字塔, 先在低分辨率图像立体匹配, 用于限制高分辨率图像的搜索范围. MonoFusion[69]采用 PatchMatch 技术[72]为每个像素随机生成一个深度值, 再利用相邻像素深度值相近的特点将可靠的深度值向周围像素传递. 上述局部匹配在缺少纹理或重复纹理的区域存在歧义. DTAM 和 REMODE 引入正则项(见式(14))对深度图进行全局优化, 缓解匹配歧义. REMODE 还采用了文献[73]中提出的深度滤波方法, 将每个像素的深度及其属于 Inlier 的概率显式表达为一个混合概率模型, 不断融合每帧的深度测量更新此概率模型的参数.

Kolev 等[68]采用面元表达[65], 通过对深度图对应像素的深度值、法向和可见性进行一致性判断, 剔除 Outliers 或更新面元参数. MonoFusion, MobileFusion 和 Schöps 等[71]提出的基于 Project Tango 的三维重建方法都采用 TSDF[55]来融合深度图、剔除 Outliers.

上述方法虽然都能实时重建出稠密的三维信息, 但大多依赖于 GPU 并行计算. 然而在很多 AR 应用中, 往往 GPU 需要用来绘制虚拟物体. 因此如何进一步提高效率, 只用 CPU 就能恢复稠密或半稠密的三维信息[34,74], 仍值得进一步研究.

3.3　多传感器融合

基于单一传感器的定位方案不可避免地都有各自的固有局限: 仅基于图像的 V-SLAM 依赖场景纹理特征; 仅基于 IMU 的定位通常有严重的误差累积; 仅基于深度的 SLAM 依赖于场景几何特征, 且设备获取深度的精度和范围受限于设备的成本和功耗. 只有将不同传感器数据融合起来, 才能互相取长补短, 达到最高的精度和鲁棒性. 如今大多数移动设备都配有单目摄像头和 IMU, 有的甚至配有双目、鱼眼或深度摄像头, 如何融合这些多传感器数据成为近年来的一个研究热点.

如第 1 节所述, 融合多传感器数据本质上就是一个非线性优化问题, 每种类型的传感器信息作为能量函数中的一项. 难点在于运动参数需要实时地恢复出来, 而变量个数会随着运行时间不断增加. 比如在 t 时刻, 待优化的相机变量为 $C=\{C_1, C_2,\cdots,C_t\}$, 那么在 30 帧/s 的应用中, 每秒会增加 30 组待优化的运动参数, 很快便会超出计算设备的极限. 因此, 求解完整的优化问题并不现实, 需要简化.

一个常见的简化方法为设置一个滑动窗口, 窗口中包含邻近的 l 帧运动参数 $\{C_{t-l+1}\cdots C_t\}$, 每次只优化这 l 组运动参数和相应的三维结构. 移出窗口的变量采用消元法, 得到关于剩余窗口内变量的先验约束, 加入后续的优化中. 基于滤波器的 SLAM 都属于这种类型(对于标准的滤波器来说 $l=1$). 近年来, 有很多研究工作试图提高基于滤波器的精度. MSCKF 通过加大滑动窗口大小 l 缓解由尚未精确的状态被消元而产生的误差累积. Li 等[75]提出了改进版本 MSCKF 2.0. 通过对 MSCKF 进行可见性分析(Observability Analysis)后发现, 如果线性化点(Linearization Point)不断变化, 会在原系统的不可见的艏摇(Yaw, 即绕重力方向的旋转)方向上错误地产生信息, 最终导致误差累积. 于是, MSCKF 2.0 采用 First-Estimate Jacobian[76]固定线性化点, 消除艏摇方向上的错误信息, 提高估计的精度. 文献[77-78]采用了不同的方法进行可见性分析(也都得到同样的结论)及消除艏摇方向上的错误信息.

另一类方法采用迭代优化[11-13,79]. EKF 可以看成只进行一次迭代, 由于 SLAM 问题通常非线性, 因此迭代优化通常比 EKF 精度更高, 但显然计算量更大. 为简化计算, 文献[11-12]采用预积分技术[15]来避免每次迭代 IMU 数据的重复积分. 文献[79]提出一种增量优化器 iSAM2, 每次只优化当前帧影响到的局部变量，可以极大地减少计算量. 文献[13]提出在滑动窗口同时保留邻近的关键帧和非关键帧, 移出滑动窗口的信息根据对当前帧的影响选择消元或直接丢弃.

3.4　其他实际问题

虽然目前 V-SLAM 算法在理想条件下已经能达到很高的精度和计算效率, 但在实际应用中往往无法满足理想条件. 比如, 多数单目 V-SLAM 系统, 尤其是基于关键帧 BA 的单目 V-SLAM 系统(第 2.2 节), 要求用户以平移的方式运动才能扩展场景. 然而, 普通用户可能难以理解如何平移, 尤其对于增强现实的应用来说, 用户很容易以接近纯旋转的方式观看虚实融合效果, 从而导致跟踪失败; 又比如, 用户在观看的时候可能不会非常平缓地运动, 实际使用过程很容易出现快速的移动和突然转动的情况, 而且这种情况下往往伴随着运动模糊, 这对目前的 V-SLAM 系统来说挑战性很大. 如何支持近似纯旋转[80-82]和快速扩展场景, 提高对运动模糊的容忍能力, 对于提高系统鲁棒性、改进增强现实应用体验尤为重要. 另外, 移动设备上的摄像头往往存在滚动快门(Rolling Shutter)现象, 图像中每一行在不同时刻曝光, 即对应不同的相机运动. 如何处理滚动快门现象, 也成为近年来的一个研究热点[83-86]. 此外, 对于画面中出现的动态物体, 如果动态物体所占画面区域较大且特征丰富, 那么传统的 RANSAC 方法很容易失败, 需要更鲁棒的误匹配点剔除机制[30,87]或者利用 IMU 信息来解决歧义.

4　结　语

近年来, 移动终端、头戴设备的快速发展为

增强现实技术提供了很好的硬件展示平台. SLAM 作为增强现实的关键基础技术, 近年来也得到了快速发展. 随着各种硬件传感器的发展和普及, 目前 SLAM 技术正朝着多传感器融合的方向发展, 试图通过利用各种传感器的优势互补性来达到尽可能高的精度和鲁棒性.

对于移动增强现实应用来说, 由于通常采用的传感器、CPU 等硬件设备的性能严重受限于价格、功耗等因素, 这就要求 SLAM 算法具有很高的鲁棒性和计算效率. 如何更进一步提高 SLAM 算法的鲁棒性和计算效率, 通过软件算法节约硬件成本, 将会是一个很有价值的研究方向. 此外, 在实际应用中很多理想条件往往无法满足, 一个鲁棒的 SLAM 系统需要能够处理各种各样的复杂情况, 才能很好地满足实际应用要求.

参考文献(References):

[1] Smith R C, Cheeseman P. On the representation and estimation of spatial uncertainty[J]. International Journal of Robotics Research, 1986, 5(4):56-68

[2] Smith R, Self M, Cheeseman P. Estimating uncertain spatial relationships in robotics[M] //Autonomous Robot Vehicles. New York: Springer, 1990: 167-193

[3] Durrant-Whyte H, Bailey T. Simultaneous localization and mapping: Part I[J]. IEEE Robotics & Automation Magazine, 2006, 13(2): 99-110

[4] Bailey T, Durrant-Whyte H. Simultaneous localization and mapping(SLAM): Part II[J]. IEEE Robotics & Automation Magazine, 2006, 13(3): 108-117

[5] Hartley R, Zisserman A. Multiple view geometry in computer vision[M]. Cambridge: Cambridge University Press, 2004

[6] Aulinas J, Petillot Y R, Salvi J, *et al*. The SLAM problem: a survey[J]. CCIA, 2008, 184(1): 363-371

[7] Ros G, Sappa A, Ponsa D, *et al*. Visual SLAM for driverless cars: a brief survey[C] //Proceedings of IEEE Workshop on Navigation, Perception, Accurate Positioning and Mapping for Intelligent Vehicles. Los Alamitos: IEEE Computer Society Press, 2012: Article No.3

[8] He Junxue，Li Zhanming. Survey of vision-based approach to simultaneous localization and mapping[J]. Application Research of Computer, 2010, 27(8): 2839-2843(in Chinese)
(何俊学, 李战明. 基于视觉的同时定位与地图构建方法综述[J]. 计算机应用研究, 2010, 27(8): 2839-2843)

[9] Liang Mingjie, Min Huaqing, Luo Ronghua. Graph-based SLAM: a survey[J]. Robot, 2013, 35(4): 500-512(in Chinese)
(梁明杰, 闵华清, 罗荣华. 基于图优化的同时定位与地图创建综述[J]. 机器人, 2013, 35(4): 500-512)

[10] Triggs B, Mclauchlan P F, Hartley R I, *et al*. Bundle adjustment — a modern synthesis[C] //Proceedings of International Workshop on Vision Algorithms: Theory and Practice. Heidelberg: Springer, 1999: 298-372

[11] Indelman V, Williams S, Kaess M, *et al*. Information fusion in navigation systems via factor graph based incremental smoothing[J]. Robotics and Autonomous Systems, 2013, 61(8): 721-738

[12] Forster C, Carlone L, Dellaert F, *et al*. IMU preintegration on manifold for efficient visual-inertial maximum－a-posteriori estimation [C] //Proceedings of Robotics: Science and Systems. Rome: Robotics: Science and Systems, 2015: Article No.6

[13] Leutenegger S, Lynen S, Bosse M, *et al*. Keyframe-based visual-inertial odometry using nonlinear optimization[J]. The International Journal of Robotics Research, 2015, 34(3): 314-334

[14] Chatfield A B. Fundamentals of high accuracy inertial navigation[M]. Reston: American Institute of Astronautics and Aeronautics, 1997

[15] Lupton T, Sukkarieh S. Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions[J]. IEEE Transactions on Robotics, 2012, 28(1): 61-76

[16] Davison A J, Reid I D, Molton N D, *et al*. MonoSLAM: real-time single camera SLAM[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6):1052-1067

[17] Mourikis A, Roumeliotis S. A multi-state constraint Kalman filter for vision-aided inertial navigation[C] //Proceedings of IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2007: 3565-3572

[18] Shi J, Tomasi C. Good features to track[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 1994: 593-600

[19] Davison A J. Active search for real-time vision[C] //Proceedings of IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2005, 1: 66-73

[20] Moré J. The Levenberg-Marquardt algorithm: implementation and theory[J]. Numerical Analysis, 1978, 630(1): 105-116

[21] Klein G, Murray D. Parallel tracking and mapping for small AR workspaces[C] //Proceedings of IEEE and ACM International Symposium on Mixed and Augmented Reality. Los Alamitos: IEEE Computer Society Press, 2007: 225-234

[22] Klein G, Murray D. Parallel tracking and mapping on a camera phone[C] //Proceedings of IEEE and ACM International Symposium on Mixed and Augmented Reality. Los Alamitos: IEEE Computer Society Press, 2009: 83-86

[23] Rosten E, Drummond T. Machine learning for high-speed corner detection[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2006, 1: 430-443

[24] Huber P J. Robust statistics[M]. Hoboken: Wiley, 2009

[25] Klein G, Murray D. Improving the agility of keyframe-based SLAM[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2008, 2: 802-815

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

[27] Rublee E, Rabaud V, Konolige K, *et al*. ORB: an efficient alternative to SIFT or SURF[C] //Proceedings of IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2011: 2564-2571

[28] Mur-Artal R, Tardós J D. Fast relocalisation and loop closing in keyframe-based SLAM[C] //Proceedings of IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2014: 846-853

[29] Kümmerle R, Grisetti G, Strasdat H, *et al*. g2o: a general framework for graph optimization[C] //Proceedings of IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2011: 3607-3613

[30] Tan W, Liu H, Dong Z, *et al*. Robust monocular SLAM in dynamic environments[C] //Proceedings of IEEE International Symposium on Mixed and Augmented Reality. Los Alamitos: IEEE Computer Society Press, 2013: 209-218

[31] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110

[32] Fischler M A, Bolles R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography[J]. Communications of the ACM, 1981, 24(6): 381-395

[33] Newcombe R A, Lovegrove S J, Davison A J. DTAM: dense tracking and mapping in real-time[C] //Proceedings of IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2011: 2320-2327

[34] Engel J, Schöps T, Cremers D. LSD-SLAM: large-scale direct monocular SLAM[C] //Proceedings of Computer Vision – ECCV 2014. Heidelberg: Springer, 2014: 834-849

[35] Civera J, Davison A J, Montiel J M M. Inverse depth parametrization for monocular SLAM[J]. IEEE Transactions on Robotics, 2008, 24(5): 932-945

[36] Chambolle A, Pock T. A first-order primal-dual algorithm for convex problems with applications to imaging[J]. Journal of Mathematical Imaging and Vision, 2011, 40(1): 120-145

[37] Engel J, Sturm J, Cremers D. Semi-dense visual odometry for a monocular camera[C] //Proceedings of IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2013: 1449-1456

[38] Glover A, Maddern W, Warren M, *et al*. OpenFABMAP: an open source toolbox for appearance-based loop closure detection [C] //Proceedings of IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2012: 4730-4735

[39] Strasdat H, Montiel J M M, Davison A J. Real-time monocular slam: Why filter?[C] //Proceedings of IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2010: 2657-2664

[40] Sturm J, Engelhard N, Endres F, *et al*. A Benchmark for the evaluation of RGB-D SLAM systems[C] //Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. Los Alamitos: IEEE Computer Society Press, 2012: 573-580

[41] Arya S, Mount D M, Netanyahu N S, *et al*. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions[J]. Journal of the ACM, 1998, 45(6): 891-923

[42] Lovegrove S, Davison A J. Real-time spherical mosaicing using whole image alignment[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2010, 3: 73-86

[43] Galvez-Lopez D, Tardos J D. Bags of binary words for fast place recognition in image sequences[J]. IEEE Transactions on Robotics, 2012, 28(5): 1188-1197

[44] Schöps T, Engel J, Cremers D. Semi-dense visual odometry for AR on a smartphone[C] //Proceedings of IEEE International Symposium on Mixed and Augmented Reality. Los Alamitos: IEEE Computer Society Press, 2014: 145-150

[45] Forster C, Pizzoli M, Scaramuzza D. SVO: fast semi-direct monocular visual odometry[C] //Proceedings of IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2014: 15-22

[46] Nurutdinova I, Fitzgibbon A. Towards pointless structure from motion: 3D reconstruction and camera parameters from general 3D curves[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 2363-2371

[47] Tarrio J J, Pedre S. Realtime edge-based visual odometry for a monocular camera[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 702-710

[48] Concha A, Civera J. Using superpixels in monocular SLAM[C] //Proceedings of IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2014: 365-372

[49] Concha A, Civera J. DPPTAM: dense piecewise planar tracking and mapping from a monocular sequence[C] //Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. Los Alamitos: IEEE Computer Society Press, 2015: 5686-5693

[50] Salas M, Hussain W, Concha A, *et al*. Layout aware visual tracking and mapping[C] //Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. Los Alamitos: IEEE Computer Society Press, 2015: 149-156

[51] Concha A, Hussain W, Montano L, *et al*. Incorporating scene priors to dense monocular mapping[J]. Autonomous Robots, 2015, 39(3): 279-292

[52] Furukawa Y, Ponce J. Accurate, dense, and robust multiview stereopsis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(8): 1362-1376

[53] Wendel A, Maurer M, Graber G, *et al*. Dense reconstruction on-the-fly[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2012: 1450-1457

[54] Newcombe R A, Izadi S, Hilliges O, *et al*. KinectFusion: real-time dense surface mapping and tracking[C] //Proceedings of IEEE International Symposium on Mixed and Augmented Reality. Los Alamitos: IEEE Computer Society Press, 2011: 127-136

[55] Curless B, Levoy M. A volumetric method for building complex models from range images[C] //Proceedings of the 23rd Annual Conference on Computer Graphics and Inter-Active Techniques. New York: ACM Press, 1996: 303-312

[56] Lorensen W E, Cline H E. Marching cubes: a high resolution 3D surface construction algorithm[C] //Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM Press, 1987: 163-169

[57] Yang C, Medioni G. Object modelling by registration of multiple range images[J]. Image and Vision Computing, 1992, 10(3): 145-155

[58] Whelan T, Kaess M, Fallon M, *et al*. Kintinuous: spatially extended KinectFusion[R]. Cambridge: Massachusetts Institute of Technology, MIT-CSAIL-TR-2012-020, 2012

[59] Whelan T, Johannsson H, Kaess M, *et al*. Robust real-time visual odometry for dense RGB-D mapping[C] //Proceedings of

IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2013: 5724-5731

[60] Huang A S, Bachrach A, Henry P, *et al*. Visual odometry and mapping for autonomous flight using an RGB-D camera[C] //Proceedings of International Symposium on Robotics Research. Heidelberg: Springer, 2011: Article No.2

[61] Steinbrücker F, Sturm J, Cremers D. Real-time visual odometry from dense RGB-D images[C] //Proceedings of IEEE International Conference on Computer Vision Workshop. Los Alamitos: IEEE Computer Society Press, 2011: 719-722

[62] Whelan T, Kaess M, Leonard J J, *et al*. Deformation-based loop closure for large scale dense RGB-D SLAM[C] //Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. Los Alamitos: IEEE Computer Society Press, 2013: 548-555

[63] Sumner R W, Schmid J, Pauly M. Embedded deformation for shape manipulation[J]. ACM Transactions on Graphics, 2007, 26(3): Article No.80

[64] Whelan T, Leutenegger S, Salas-Moreno R F, *et al*. ElasticFusion: dense SLAM without a pose graph[C] //Proceedings of Robotics: Science and Systems. Rome: Robotics: Science and Systems, 2015: Article No.1

[65] Pfister H, Zwicker M, Van Baar J, *et al*. Surfels : surface elements as rendering primitives[C] //Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM Press, 2000: 335-342

[66] Pizzoli M, Forster C, Scaramuzza D. REMODE: probabilistic, monocular dense reconstruction in real time[C] //Proceedings of IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2014: 2609-2616

[67] Tanskanen P, Kolev K, Meier L, *et al*. Live metric 3D reconstruction on mobile phones[C] //Proceedings of IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2013: 65-72

[68] Kolev K, Tanskanen P, Speciale P, *et al*. Turning mobile phones into 3D scanners[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 3946-3953

[69] Pradeep V, Rhemann C, Izadi S, *et al*. MonoFusion: real-time 3D reconstruction of small scenes with a single Web camera[C] //Proceedings of IEEE International Symposium on Mixed and Augmented Reality. Los Alamitos: IEEE Computer Society Press, 2013: 83-88

[70] Ondruska P, Kohli P, Izadi S. MobileFusion: real-time volumetric surface reconstruction and dense tracking on mobile phones[J]. IEEE Transactions on Visualization and Computer Graphics, 2015, 21(11): 1251-1258

[71] Schöps T, Sattler T, Hane C, *et al*. 3D modeling on the go: interactive 3D reconstruction of large-scale scenes on mobile devices[C] //Proceedings of International Conference on 3D Vision. Los Alamitos: IEEE Computer Society Press, 2015: 291-299

[72] Bleyer M, Rhemann C, Rother C. PatchMatch stereo-stereo matching with slanted support windows[C] //Proceedings of the British Machine Vision Conference. Guildford: BMVA Press, 2011: 1-11

[73] Vogiatzis G, Hernández C. Video-based, real-time multi-view stereo[J]. Image and Vision Computing, 2011, 29(7): 434-441

[74] Mur-Artal R, Tardos J D. Probabilistic semi-dense mapping from highly accurate feature-based monocular SLAM[C]// Proceedings of Robotics: Science and Systems. Rome: Robotics: Science and Systems, 2015: Article No.41

[75] Li M, Mourikis A I. High-precision, consistent EKF-based visual-inertial odometry[J]. The International Journal of Robotics Research, 2013, 32(6): 690-711

[76] Huang G P, Mourikis A, Roumeliotis S. Analysis and improvement of the consistency of extended Kalman filter based SLAM[C] //Proceedings of IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2008: 473-479

[77] Hesch J A, Kottas D G, Bowman S L, *et al*. Camera-IMU-based localization: observability analysis and consistency improvement[J]. International Journal of Robotics Research, 2014, 33(1): 182-201

[78] Huang G, Kaess M, Leonard J J. Towards consistent visual-inertial navigation[C] //Proceedings of IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2014: 4926-4933

[79] Kaess M, Johannsson H, Roberts R, *et al*. iSAM2: incremental smoothing and mapping using the Bayes tree[J]. International Journal of Robotics Research, 2012, 31(2): 216-235

[80] Gauglitz S, Sweeney C, Ventura J, *et al*. Live tracking and mapping from both general and rotation-only camera motion[C] //Proceedings of IEEE International Symposium on Mixed and Augmented Reality. Los Alamitos: IEEE Computer Society Press, 2012: 13-22

[81] Pirchheim C, Schmalstieg D, Reitmayr G. Handling pure camera rotation in keyframe-based SLAM[C] //Proceedings of IEEE International Symposium on Mixed and Augmented Reality. Los Alamitos: IEEE Computer Society Press, 2013: 229-238

[82] Herrera C, Kim K, Kannala J, *et al*. DT-SLAM: deferred triangulation for robust SLAM[C] //Proceedings of International Conference on 3D Vision (3DV). Los Alamitos: IEEE Computer Society Press, 2014: 609-616

[83] Hedborg J, Forssén P E, Felsberg M, *et al*. Rolling shutter bundle adjustment[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2012: 1434-1441

[84] Lovegrove S, Patron-Perez A, Sibley G. Spline fusion: a continuous-time representation for visual-inertial fusion with application to rolling shutter cameras[C] //Proceedings of the British Machine Vision Conference. Guildford: BMVA Press, 2013: Article No.93

[85] Li M, Kim B H, Mourikis A I. Real-time motion tracking on a cellphone using inertial sensing and a rolling-shutter camera[C] //Proceedings of IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2013: 4712-4719

[86] Kerl C, Stuckler J, Cremers D. Dense continuous-time tracking and mapping with rolling shutter RGB-D cameras[C] //Proceedings of IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 2264-2272

[87] Kerl C, Sturm J, Cremers D. Robust odometry estimation for RGB-D cameras[C] //Proceedings of IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2013: 3748-3754