文章编号：2095-2163(2021)01-0089-05 中图分类号：TP391 文献标志码：A

一种视觉 SLAM 单目半稠密建图方法的实现

尚 任
（吉林化工学院 航空工程学院，吉林 吉林 132022）

摘 要： SLAM 即同时定位与地图构建，一直是机器人和计算机视觉的研究热点。尤其是视觉 SLAM 技术，21 世纪以来在理论和实践上均取得了明显的突破，已逐步迈向市场应用。建图作为 SLAM 的两大目标之一，可以满足更多的应用需求。本文在给定相机轨迹的情况下，提出一种视觉 SLAM 单目半稠密建图方法，利用极线搜索和块匹配技术，加入图像变换和逆深度高斯深度滤波器处理，以期避免单目稠密建图严重依赖纹理、计算量大的缺点，提高单目半稠密建图的准确性和鲁棒性。经测试显示，改进的单目半稠密建图方法在检测梯度变化明显像素点上更加准确，深度估计的平均误差和平方误差分别减少了 9% 和 47%，是一种可行有效的视觉 SLAM 单目半稠密建图解决方案。

关键词： 视觉 SLAM；建图；单目；极限搜索；块匹配；逆深度

A visual SLAM semi-dense mapping method on monocular camera

SHANG Ren
（School of Aeronautical Engineering, Jilin Institute of Chemical Technology, Jilin Jilin 132022, China）

【Abstract】 Simultaneous Localization and Mapping (SLAM) is the hot spot of Robots and Computer Vision. Visual SLAM had made a breakthrough since 21st century and applied to the market. Mapping is one of the targets of SLAM. Given location and poses of the monocular camera, the paper proposes a Visual SLAM semi-dense mapping method. Using searching on the epipolar line and block-matching, image transformation and inverse depth filter have been used in order to avoid depending on textures and large amounts of calculation. In this way, the accuracy and robustness of semi-dense mapping has been improved. From the testing result, the improved visual SLAM semi-dense mapping method on monocular camera has a better performance on gradient detecting of pixels. The average error and square error on depth estimation decrease about 9% and 47%. It is a feasible and effective solution to rebuild semi-dense mapping on monocular camera.

【Key words】 visual SLAM; mapping; monocular; extreme search; block match; inverse depth

0 引 言

同时定位与地图构建（Simultaneous Localization and Mapping，SLAM），于 1986 年提出[1]，是指搭载特定传感器的主体，在没有环境先验知识的情况下，在运动中估计自身轨迹，并同时建立环境的模型[2]。在视觉 SLAM 中，特定传感器以视觉传感器（即相机）为主体，研究者则需要根据连续拍摄的多幅图像，推断相机的运动以及周围的环境。

定位与建图是 SLAM 的两大目标。在视觉 SLAM 中，建图与定位同样重要，建图还具有很多其他的应用需求。对于单目相机的地图构建，国内外已涌现了不少的研究成果，提供了很多优秀的开源 SLAM 方案。PTAM[3] 的问世是视觉 SLAM 发展过程中的重要事件，首次提出并实现了跟踪与建图过程的并行化。ORB-SLAM[4] 是继 PTAM 之后业内著名的视觉 SLAM 系统，代表着主流特征点 SLAM 的一个高峰。ORB-SLAM 能够支持单目、双目、RGB-D 三种模式，有着良好的泛用性，但只能用于构建稀疏地图，在建图方面显得过于重量级。LSD-SLAM[5-6] 标志着单目直接法在 SLAM 中的成功应用，其核心贡献是将直接法应用到了半稠密单目地图重建中。还有很多单目 SLAM 开源系统（如 SVO[7]、DSO[8] 等等）。此外，罗鸿城[9]、谢瑒[10]、张剑华等人[11]、Vogiztzis 等人[12] 均对单目建图进行了探索。

本文旨在提出一种单目半稠密建图的方法，研究中在图像间仿射变换预处理后，利用块匹配的方法进行极线搜索[13]，并将像素点的逆深度进行高斯融合，采用高斯分布的逆深度滤波器的方式估计像素点的深度，具有良好的准确性和鲁棒性。

1 SLAM 系统介绍

研究时通常使用便携式的传感器来完成 SLAM，在未知环境中进行实时建模[14]。视觉 SLAM 主要是指如何用相机解决定位和建图问题。按照工

作者简介：尚 任（1991-），女，硕士，讲师，主要研究方向：视觉 SLAM、电子与通信工程。

收稿日期：2020-10-26

作方式的不同,相机可以分为单目(Monocular)、双目(Stereo)和深度(RGB-D)三大类。此外,视觉SLAM中还可使用全景相机[15]、Event相机[16]等不同种类,但迄至目前还未成为主流,仅会将其应用在部分场景中。

在典型的视觉SLAM框架中,主要分为前端视觉里程计(Visual Odometry)、后端优化(Optimization)、回环检测(Loop Closure Detection)和建图(Mapping)四个模块。对此拟做阐释分述如下。

(1)视觉里程计。根据相邻图像的信息估计出粗略的相机运动以及局部地图,处理的结果将作为后端优化的初始值。视觉里程计的算法主要分为两大类:特征点法和直接法。基于特征点法的前端,直到现在也被公认为视觉里程计的主流方法。其中,特征点法稳定性强,对光照和动态物体不敏感,是目前比较成熟的解决方案。但特征点法中,关键点的提取与描述子的计算非常耗时,忽略了除特征点以外的所有信息,而且在特征缺失区域效果并不理想。直接法根据像素的亮度信息估计相机的运动,克服了特征点法的不足,但由于灰度不变假设,对光照、环境等外界条件敏感,稳定性相对较差。

(2)后端优化。负责优化整个问题,将视觉里程计估计的相机位姿以及回环检测的信息进行优化,返回优化后的结果,得到全局一致的轨迹和地图。后端优化部分主要有2种处理方法:批量(Batch)处理的非线性优化方法和渐进(Incremental)处理的滤波器方法。目前,视觉SLAM的主流是非线性优化方法。以非线性优化为主的后端考虑当前状态与之前所有状态的关系,在同等计算量的情况下,能够取得更好的精度和鲁棒性[17],但往往计算量过大,不适用于计算能力有限的平台。以扩展卡尔曼滤波为代表的滤波器方法,仅关注当前时刻和前一时刻的情况,虽然精度和鲁棒性上不如非线性优化,但是在计算资源受限,或待估计量较简单的场合,滤波器方法仍不失为一种有效的方式。

(3)回环检测。判断传感器是否经过先前位置,将检测到的回环信息提供给后端进行处理。目前,回环检测用得最多的方法是基于外观的方法,即仅根据2幅图像的相似性确定回环检测关系。较为经典的是词袋模型,该模型是一个非监督学习的过程。但是深度学习方法由于在学习与无监督聚类上的优异表现,其在性能方面则有望胜过目前的主流方法。

(4)建图模块。根据估计的轨迹,建立与需求对应的地图。建图是SLAM的两大目标之一。根据使用像素的数量,构建的地图分为3种:稀疏地图、稠密地图和半稠密地图。总地来说,稀疏地图只建模感兴趣的部分,稠密地图将建模所有看到过的部分,半稠密地图所建模的像素数量介于稀疏地图和稠密地图之间。地图的用途可以归纳为5点[14],分别是:定位、导航、避障、重建和交互。其中,定位需要用到稀疏地图,而导航、避障、重建则要用到稠密地图。

2 半稠密直接法

对于单目建图,特征点法和直接法均可以重建稀疏地图,而半稠密地图和稠密地图由于没有描述子,只能采用直接法重建。构建稀疏地图只能满足定位需求,应用有限,构建稠密地图虽然可以满足地图导航、避障和重建的应用需求,但对物体纹理有强烈的依赖性。故本文采用单目半稠密建图的折中方式。建图需要估计或者获取空间点的位置和深度。根据空间点深度的来源,可以把直接法分成3类:

(1)稀疏直接法,即空间点来自于稀疏关键点,不必计算描述子,速度最快,但只能计算稀疏的重构。

(2)半稠密直接法,即空间点来自部分像素,考虑到像素梯度为零的像素点对计算运动增量没有任何贡献,只使用带有梯度的像素点,能够重构一个半稠密结构。

(3)稠密直接法,即空间点来自于所有像素,需要计算所有像素,多数不能在现有CPU上实时计算,需要GPU加速。而且由于使用了梯度不明显的像素点,重构稠密地图的效果难以保证[14]。

本文采用半稠密直接法作为单目半稠密重建的重要环节,既保证了单目相机建图实时性的要求,又保证了建图结果的准确性。

3 图像匹配

本文沿用了直接法的思想,利用极限搜索和块匹配技术确定一幅图像中某像素出现在其他图像里的位置。当能够确定某个像素在各个图像中的位置,就可以像特征点法那样,利用三角测量法确定对应的深度。研究内容详见如下。

3.1 对极几何与极线搜索

单目SLAM仅已知二维像素坐标。对于稀疏地图的构建,可以利用特征匹配得到若干对配对好的

匹配点，从而通过这些二维像素点的对应关系，恢复出在两帧之间相机的运动。该问题可用对极几何解决。

而对于半稠密地图和稠密地图的构建，需要对每个或大部分像素点做匹配，这种情况下，没有描述子的存在，第一幅图像上的像素点只能在极线上搜索，以找到其在第二幅图像上比较相似的点。即沿着第二幅图像中的极线，逐个比较每个像素与第一幅图像像素点的相似程度。这种通过比较像素亮度来确定图像间匹配，从而估计相机运动的做法，与视觉里程计中的直接法思想相同。

3.2　块匹配

为避免直接法中单个像素没有区分度且灰度值不变假设过强的弊端，一种有效的方法是以图像块为计算单位，而不是单个像素点。研究时，可在第一幅图像待匹配像素点 p_1 周围取一个大小为 $n \times n$ 的小块，假设在不同图像间整个小块灰度值不变，比较相同大小像素块的亮度相对于比较单个像素的亮度要稳定可靠得多。对于单目相机的半稠密地图和稠密地图的构建，可在极线上取很多大小相等的像素块进行比较，使极线搜索的结果有更好的准确性和鲁棒性。

计算像素块之间差异的方法有若干种，其中将每个像素块均值去掉的处理方法更为可靠。常见的计算方法有 SAD、SSD、NCC 等[18]，去均值后，称为去均值的 SAD、去均值的 SSD、去均值的 NCC 等等。研究中，不妨将 p_1 周围的去均值像素块记为 $A_i, i = 1,\cdots,n$，把极线上的 n 个去均值像素块记为 $B_i, i = 1,\cdots,n$。去均值的 *NCC*（即去均值的归一化互相关）计算公式为：

$$S(A,B)_{NCC} = \frac{\sum_{i,j} A(i,j)B(i,j)}{\sqrt{\sum_{i,j} A(i,j)^2 \sum_{i,j} B(i,j)^2}}. \quad (1)$$

本文采用去均值的 *NCC* 来计算极线上像素块的相似性度量，在极限搜索中将得到一个沿着极线的 *NCC* 分布。由于图像的非凸性质，在搜索距离较长的情况下，*NCC* 分布通常会得到一个非凸函数，即这个分布存在很多峰值[12]。这种情况下，可使用概率分布来描述像素块的深度值，而不是用某个单一数值来描述深度。

3.3　图像变换

在块匹配中，研究假设像素块在相机运动时保持不变。这个假设在相机平移时能够保持成立，但在相机发生明显旋转时就有可能不成立。特别地，当相机绕着光心旋转较大角度，会出现相关性直接变成负数的情况，即使都是同一个像素块。在这种情况下，在块匹配前，做一次图像变换，把 2 帧图像间的运动考虑进来，是一种常见有效的预处理方式。

仿射变换的矩阵形式如下：

$$\boldsymbol{T}_A = \begin{bmatrix} \boldsymbol{A} & t \\ \boldsymbol{O}^{\mathrm{T}} & 1 \end{bmatrix}, \quad (2)$$

现以第一幅图像为参考帧，推导参考帧与当前帧之间的仿射变换。根据相机模型，第一帧图像上的一个像素 $\boldsymbol{P}_R$ 与真实的三维点世界坐标 $\boldsymbol{P}_W$ 有以下关系：

$$d_R \boldsymbol{P}_R = K(\boldsymbol{R}_{RW}\boldsymbol{P}_W + \boldsymbol{t}_{RW}), \quad (3)$$

其中，d_R 表示像素 $\boldsymbol{P}_R$ 距离相机成像平面的深度。

类似地，对于当前帧，$\boldsymbol{P}_C$ 为真实三维点世界坐标 $\boldsymbol{P}_W$ 的投影为：

$$d_C \boldsymbol{P}_C = K(\boldsymbol{R}_{CW}\boldsymbol{P}_W + \boldsymbol{t}_{CW}), \quad (4)$$

2 帧图像之间的像素关系为：

$$d_C \boldsymbol{P}_C = d_R \boldsymbol{K}\boldsymbol{R}_{CW}\boldsymbol{R}_{RW}^{\mathrm{T}}\boldsymbol{K}^{-1}\boldsymbol{P}_R + \boldsymbol{K}\boldsymbol{t}_{CW} - \boldsymbol{K}\boldsymbol{R}_{CW}\boldsymbol{R}_{RW}^{\mathrm{T}}\boldsymbol{t}_{RW}. \quad (5)$$

当 $\boldsymbol{P}_R$ 与 d_R 已知时，可以计算出 $\boldsymbol{P}_C$ 的投影位置。再给 $\boldsymbol{P}_R$ 的 2 个分量各加一个增量 du、dv，就可以求得 $\boldsymbol{P}_C$ 的增量 du_c、dv_c。这样就可以算出局部范围内参考帧和当前帧一个线性的坐标变换，构成仿射变换。经过仿射变换后的块匹配，以期得到对旋转更好的鲁棒性。

4　深度估计

4.1　三角测量

研究已通过极线搜索和仿射变换后的块匹配技术得到了半稠密或稠密深度估计中某个像素在各个图中的位置，这样就可以利用三角测量（Triangulation）方法确定其深度。三角测量最早由高斯提出，是指通过不同位置对同一路标点进行观察，从观察到的位置推断路标点的距离。在视觉 SLAM 中，主要用三角化来估计像素点的距离（即深度）。由于噪声的存在，通常求得深度的最小二乘解。

4.2　高斯分布的逆深度滤波器

对像素点的深度估计，有滤波器方法或非线性优化两种求解思路。由于前端已经占用一定的计算量，建图方面通常采用计算量相对较少的滤波器方法。

对深度的分布假设有若干种方法，可假设深度值服从高斯分布，也可假设深度值服从均匀-高斯

混合分布[7,12]。但是深度的高斯分布并不准确,因为近处的点不会小于相机焦距,而且在一些室外场景中,会存在距离很远的点,这个分布并不和高斯分布一样是对称形状。在仿真中,研究发现假设深度的倒数(即逆深度)为高斯分布是比较有效的。逆深度在实际应用中也具有更好的数值稳定性,从而成为一种通用的技巧。本文采用的就是高斯分布的逆深度滤波器方法。这里,假设某个像素点的逆深度 $\frac{1}{d}$ 服从高斯分布,对应数学公式可写为:

$$P(\frac{1}{d}) = N(\mu, \sigma^2). \tag{6}$$

新观测的信息更新原来像素点的深度分布,融合后的逆深度分布仍然是高斯分布。采用逆深度的方差更新方式描述深度的不确定性。在实际工程中,当深度不确定性小于一定阈值时,即可认为深度数据已收敛。

5 建图

本次研究主要着重于改进单目半稠密建图的方法,故选用给定相机轨迹的数据进行讨论,根据一段视频系列估计某幅图像的深度。研究时选用了 REMODE[13,19] 数据集和 EuRoC 数据集。

5.1 数据集介绍

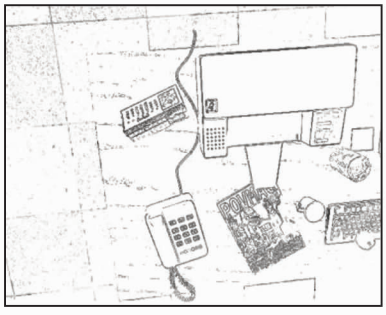
REMODE 数据集提供了一架无人机采集的单目俯视图像,共 200 张,同时提供了每张图像的真实位姿。EuRoC 数据集是双目和 IMU 数据集,包含 2 个场景:苏黎世联邦理工学院的一个厂房和一个普通房间。本文仅用双目中的一个相机的厂房场景数据和对应的图像真实位姿数据,共 200 张。

5.2 建图测试结果分析

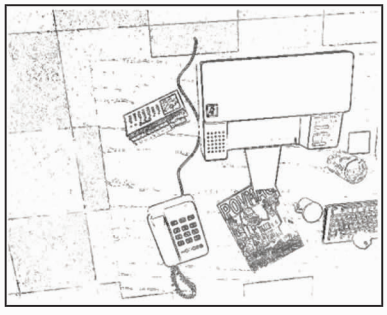
建图测试结果验证分为对照组和改进组。其中,对照组采用未优化的单目半稠密建图方法(极限搜索+块匹配+高斯分布的深度滤波器),改进组采用改进的单目半稠密建图方法(极限搜索+图像变换+块匹配+高斯分布的逆深度滤波器)。

对照组方法和改进组方法在 REMODE 和 EuRoC 数据集中表现如图 1 所示,数据集测试结果分析见表 1。

(a) REMODE数据集原始图像
(a) One original image of REMODE data set

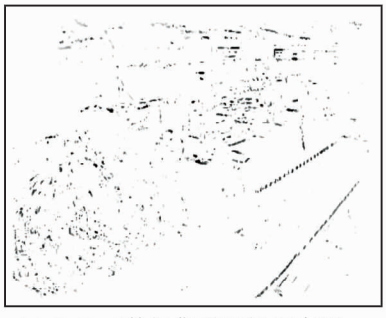


(b) REMODE数据集对照组深度图
(b) Depth image of REMODE data set



(c) REMODE数据集对照组深度图
(c) Improved depth image of REMODE data set

(d) EuRoC数据集原始图像
(d) One original image of EuRoC data set



(e) EuRoC数据集对照组深度图
(e) Depth image of EuRoC data set

(f) EuRoC数据集对照组深度图
(f) Improved depth image of EuRoC data set

图 1 数据集原始图像和深度图

Fig. 1 Original image and depth image of data sets

表 1　数据集测试结果分析

Tab. 1　Testing result of data sets

数据集	对照组			改进组		
	用时/s	平均误差	平方误差	用时/s	平均误差	平方误差
REMODE	0.29	2.91	8.63	0.14	2.90	8.62
EuRoC	22.12	3.14	16.07	7.19	2.86	8.52

可以看出,改进的方法有更好的准确性和鲁棒性。具体来说,在检测梯度变化明显的像素点上,对于相对简单的 REMODE 数据集室内普通房间场景,对照组和改进组两种方法效果大致相同,深度的平均误差和平方误差相差无几,但在 EuRoC 数据集复杂工厂室内场景中,改进的方法显然要更加准确,深度图效果更为突出,平均误差减少了 9%,平方误差减少了 47%。在用时上,由于改进的半稠密建图方法对图像中每个像素做了仿射变换,使改进组方法用时多于对照组方法。

6　结束语

在视觉 SLAM 中,相比于 RGB-D 建图,单目建图的计算量大,结果也并不可靠。但是单目视觉 SLAM 的应用能够适用于复杂室外环境,而且性价比高,是颇具学术价值的研究方向。在今后的研究工作中,将致力于提出完整的单目 SLAM 系统,以期得到更加准确有效的定位及建图解决方案。

参考文献

[1] BARFOOT T. State estimation for robotics - A matrix lie group approach[M]. Cambridge: Cambridge University Press, 2016.

[2] 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.

[3] KLEIN G, MURRAY D. Parallel tracking and mapping for small AR workspaces [C]// ISMAR 2007 6th IEEE and ACM International Symposium on mixed and Augmented Reality. Nara, Japan:IEEE, 2007: 225-234.

[4] MUR-ARTAL R, MONTIEL J, TARDOS J D. Orb-slam: A versatile and accurate monocular slam system[J]. arXiv preprint arXiv: 1502.00956, 2015.

[5] ENGEL J, SCHOEPS T, CREMERS D. Lsd-slam: Large-scale direct monocular SLAM[M]//FLEET D, PAJDLA T, SCHIELE B, et al. Computer Vision - ECCV 2014. Lecture Notes in Computer Science. Cham:Springer, 2014:834-849.

[6] ENGEL J, STURM J, CREMERS D. Semi - dense visual odometry for a monocular camera[C]//Proceedings of the IEEE International Conference on Computer Vision. Washington, DC, USA:IEEE, 2013:1449-1456.

[7] FORSTER C, PIZZOLI M, SCARAMUZZA D. Svo: Fast semi-direct monocular visual odometry[C]//2014 IEEE International Conference on Robotics and Automation (ICRA). Hong Kong: IEEE, 2014:15-22.

[8] ENGEL J, KOLTUN V, CREMERS D. Direct sparse odometry [J]. arXiv preprint arXiv: 1607. 02565, 2016.

[9] 罗鸿城. 基于卷积神经网络的实时单目稠密建图方法研究[D]. 武汉:华中科技大学,2019.

[10] 谢瑒. 无人机自主导航的单目视觉 SLAM 技术研究[D]. 南京:南京航空航天大学,2019.

[11] 张剑华,王燕燕,王曾媛,等. 单目同时定位与建图中的地图恢复融合技术[J]. 中国图象图形学报,2018,23(3):372-383.

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

[13] PIZZOIL M, FORSTER C, SCARAMUZZA D. Remode: Probabilistic, monocular dense reconstruction in real time[C]// 2014 IEEE International Conference on Robotics and Automation (ICRA). Hong Kong:IEEE, 2014:2609-2616.

[14] 高翔, 张涛, 刘毅,等. 视觉 SLAM 十四讲, 从理论到实践[M]. 2 版. 北京:电子工业出版社, 2019.

[15] PRETTO A, MENEGATTI E, PAGELLO E. Omnidirectional dense large - scale mapping and navigation based on meaningful triangulation [C]//2011 IEEE International Conference on Robotics and Automation(ICRA 2011). Shanghai, China:IEEE, 2011: 3289-3296.

[16] RUECKAUER B, DELBRUCK T. Evaluation of event - based algorithms for optical flow with ground - truth from inertial measurement sensor [J]. Frotiers in neuroscience, 2016, 10 (137):1-17.

[17] STRASDAT H, MONTIEL J M, DAVISON A J. Visual slam: Why filter? [J]. Image and Vision Computing, 2012,30(2):65-77.

[18] HIRSCHMULLER H, SCHARSTEIN D. Evaluation of cost functions for stereo matching[C]//2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis: IEEE, 2007:1-8.

[19] HANDA A, NEWCOMBE R A, ANGELI A, et al. Real-time camera tracking: When is high frame - rate best? [M]// FITZGIBBON A, LAZEBNIK S, PERONA P, et al. Computer Vision-ECCV 2012. Lecture Notes in Computer Science. Berlin/ Heidelberg:Springer, 2012:222-235.