

单目视觉里程计研究综述

祝朝政¹, 何 明¹, 杨 晟², 吴春晓¹, 刘 斌¹

ZHU Chaozheng¹, HE Ming¹, YANG Sheng², WU Chunxiao¹, LIU Bin¹

1. 中国人民解放军陆军工程大学 指挥控制工程学院, 南京 211117

2. 河海大学 计算机与信息学院, 南京 211100

1.College of Command Control Engineer, Army Engineering University, Nanjing 211117, China

2.College of Information and Computer, HoHai University, Nanjing 211110, China

ZHU Chaozheng, HE Ming, YANG Sheng, et al. Survey of monocular Visual Odometry. Computer Engineering and Applications, 2018, 54(7): 20-28.

Abstract: Monocular Visual Odometry can not only provide functions such as navigation and obstacle avoidance for mobile robots, but also can be applied to the current hot autopilot and other fields. This paper analyzes the basic principle of Visual Odometry, and studies the research status of monocular Visual Odometry technology at home and abroad. At the same time, the classical monocular Visual Odometries are analyzed and compared in detail, such as ORB-SLAM2, DSO and so on. Aiming at the issues of robustness and real-time which are generally concerned in the current visual odometry research, the development directions and trend of the next research work are discussed.

Key words: monocular Visual Odometry; feature method; direct method; visual SLAM

摘 要: 单目视觉里程计不仅能为移动机器人提供导航避障等功能, 在无人驾驶等领域中也有更广阔的应用价值。剖析了视觉里程计的基础原理, 研究了单目视觉里程计技术的国内外现状; 同时, 对ORB-SLAM2、DSO等典型单目视觉里程计进行深入分析和比较。针对当前视觉里程计研究中普遍关注的鲁棒性和实时性等问题, 探讨了下一步研究工作的发展方向和趋势。

关键词: 视觉里程计; 特征点法; 直接法; 视觉SLAM

文献标志码: A **中图分类号:** TP391.4 **doi:** 10.3778/j.issn.1002-8331.1712-0279

1 引言

移动机器人进入未知环境进行定位和导航是自主化的重要的一步, 因为未知环境的复杂性, 所以研究仅依靠机器人自身的传感器构建实时地图并进行定位具有重要意义^[1-2]。视觉传感器是机器人上常见的一类传感器, 具有精确度高, 成本低, 且数据信息丰富等特点, 因此利用视觉传感器来定位成为研究热点。视觉里程计(Visual Odometry, VO)这一概念^[3]由Nister提出, 指通过机器视觉技术, 分析相关图像序列来实时估计移动机器人的位姿(位置和姿态)过程, 能克服传统里程计的不足, 更加精准进行定位, 并且可以运行在全球定位系

统(Global Position System, GPS)无法覆盖或失效的环境中, 例如室内环境、星际探索^[3-4]等。

鉴于视觉里程计的特点和优势, VO在火星探测器上得到了成功应用^[4], 也在公共安全、虚拟现实(Virtual Reality, VR)^[5]、增强现实(Augmented Reality, AR)^[6]等领域凸显出其重要的应用价值。

1.1 视觉SLAM和VO的区别与联系

基于视觉的即时定位与地图构建(visual Simultaneous Location and Mapping, vSLAM)^[7]有两种主流方法: 滤波的方法, 使用基于概率分布进行视觉信息融合^[8]; 非滤波的方法, 选取关键帧进行全局优化^[9-10]。具

基金项目: 江苏省自然科学基金(No.BK20150721, No.BK20161469); 中国博士后基金(No.2015M582786, No.2016T91017); 江苏省重点研发计划(No.BE2015728, No.BE2016904); 江苏省科技基础设施建设计划(No.BM2014391); 国家重点研发计划(No.2016YFC0800606); 中国工程院重点咨询课题(No.2017-XZ-05)。

作者简介: 祝朝政(1995—), 男, 硕士研究生, CCF学生会员, 研究领域为机器视觉、嵌入式系统, E-mail: 370045744@qq.com; 何明(1978—), 男, 博士, 教授, 主要研究领域为大数据、人工智能。

收稿日期: 2017-12-19 **修回日期:** 2018-01-30 **文章编号:** 1002-8331(2018)07-0020-09

表1 经典的VO研究成果

方案名称	发布时间	传感器类型	处理方法	后端优化方法	特点
MonoSLAM ^[18]	2007年	单目	特征点	EKF滤波法	首个实时视觉SLAM,EKF+稀疏角点
PTAM ^[19]	2007年	单目	特征点	优化	关键帧+BA,首次以优化为后端
SVO ^[20]	2014年	单目	半直接法	优化	稀疏直接法,仅VO
DSO ^[21]	2016年	单目	直接法	优化	单目直接法,目前效果最好的直接法
LSD-SLAM ^[22]	2014年	单目	直接法	优化	直接法+半稠密地图
ORB-SLAM2 ^[16, 23]	2017年	单目为主	特征点法	优化	ORB特征点+三线结构
RTAB-MAP ^[24]	2013年	双目/RGB-D	特征点法	优化	支持较大场景的实用RGB-D
DTAM ^[25]	2011年	单目	直接法	优化	直接法,单目稠密地图,需要GPU
DVO ^[26]	2013年	RGB-D	直接法	优化	RGB-D直接法,稠密地图
RGB-D-SLAM-V2 ^[7]	2014年	RGB-D	特征点法	优化	完整的RGB-D稠密建图
Elastic Fusion ^[27-28]	2015年	RGB-D	直接法	优化	RGB-D在线重建,显示效果较好
OKVIS ^[9]	2015年	单目/多目+IMU	特征点法	优化	以优化为主的关键帧VIO
ROVIO ^[8, 29]	2015年	单目+IMU	直接法	EKF滤波法	以EKF为主的VIO
VINS-Mono ^[10]	2017年	单目+IMU	特征点法	优化	基于优化的紧耦合VIO框架
Kinect Fusion ^[30]	2011年	RGB-D	直接法	优化	首次实现RGB-D实时稠密重建

体关于这两种方法的评估详见^[11-12]。

vSLAM和VO两者的区别在于,后者仅关注局部轨迹的一致性,而前者关注的是全局机器人轨迹的一致性。理解什么时候产生回环和有效集成新的约束到当前地图是视觉SLAM主要研究问题。VO目标是增量式重建轨迹,可能只优化前 n 个路径的位姿,即基于窗口的捆绑调整。这个滑动窗口优化在SLAM中只能是建立一个局部地图。

vSLAM和VO两者的联系在于,后者可以视为前者中的一个模块,能增量式重建相机的运动轨迹,所以有些学者在研究中,将vSLAM视为VO展开研究。

如表1所示,自2007年并行跟踪与建图(Parallel Tracking and Mapping,PTAM)之后,由于发现了稀疏矩阵结构特殊性,后端研究都已经从EKF转换到优化的方式。同时,最近几年里,单目和双目相机都取得显著的进展^[13-17],大部分已经具备了大范围、室外环境的能力。

两个相机位姿在临近时刻 $k, k-1$ 形成一个刚性变换 $T_{k,k-1} \in R^{4 \times 4}$,记为:

$$T_{k,k-1} = \begin{bmatrix} R_{k,k-1} & t_{k,k-1} \\ 0 & 1 \end{bmatrix}$$

(1)

其中, $R_{k,k-1}$ 是旋转矩阵, $t_{k,k-1}$ 是平移矩阵。集合 $T_{1:n} = \{T_{1,0}, T_{2,1}, \dots, T_{n,n-1}\}$ 包含所有运动序列。最后,相机位姿集 $C_{0:n} = \{C_0, C_1, \dots, C_n\}$ 在 k 时刻初始坐标。当前位姿 C_n 能通过计算所有变换 $T_k(k=1, 2, \dots, n)$ 之间的联系得到,因此, $C_n = C_{n-1}T_n$, C_0 是 $k=0$ 时刻的相机位姿。

2 VO形式化描述

在 k 时刻,刚性机器人上的相机采集环境中运动图像。如果是单目VO,在 k 时刻采集到的图像集表示为 $I_{0:n} = \{I_0, I_1, \dots, I_n\}$ 。如果是双目VO,每个时刻都会有左右图像产生,表示为 $I_{l,0:n} = \{I_{l,0}, I_{l,1}, \dots, I_{l,n}\}$ 和 $I_{r,0:n} = \{I_{r,0}, I_{r,1}, \dots, I_{r,n}\}$,如图1所示。

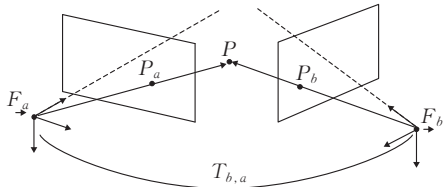


图1 VO问题的图例

假设相机坐标就是机器人的坐标。在立体视觉系统中,一般左相机为原点。

VO的主要工作就是计算从图像 I_k 到图像 I_{k-1} 相关变换 T_k ,然后集成所有的变换恢复出相机的全部轨迹 $C_{0:n}$ 。这意味着VO是一个位姿接着一个位姿,增量式重建轨迹。一个迭代优化基于前 m 位姿可以执行,之后可得到一个更准确的局部轨迹估计。

迭代优化通过基于前 m 帧最小化三维点在局部地图中的重投影误差(基于窗口的捆绑调整,因为它在 m 帧窗口上执行)。局部地图空间中3D点的深度值通过三角测量法进行估计,所以可构造一个最优化问题,调整 R, t 使得对于所有的特征点 z^j ,误差二范数累计最小,得到:

$$\min_{X^j, R, t} \sum_{j=1}^N \left\| \frac{1}{\lambda_1} CX^j - [Z_1^j, 1]^T \right\|^2 + \left\| \frac{1}{\lambda_2} C(RX^j + t) - [Z_2^j, 1]^T \right\|^2$$

(2)

这就是最小化重投影误差问题。实际操作中,在调整每个 X^j ,使得更符合每一次观测 Z^j ,也就是每个误差项都尽量小。由于此原因,也称为捆绑调整(Bundle Adjustment)。捆绑调整优化原理如图2所示。

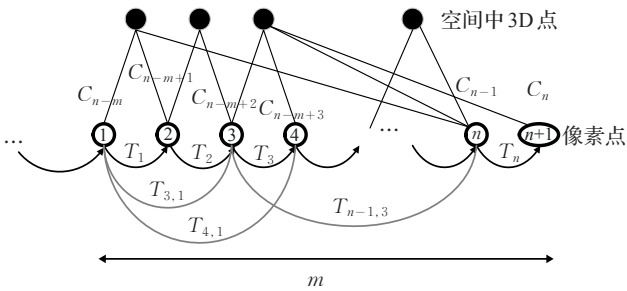


图2 捆绑调整优化原理(C 表示相机当前帧位姿, T 表示两个相机之间位姿的变换, m 表示相机总帧数)

3 VO 方法研究进展

在过去十多年里,大范围场景的VO^[22,31]取得了巨大的成功。从VO实现方法上来分,主要分为基于特征点法和基于直接法,也有人提出基于混合的半直接跟踪的方法,即两种方法混合使用。

3.1 基于特征点的方法

对于特征点法^[3,6,16,18,23,32-33],Nister是最早开展实时单目大场景VO相关工作^[3]。稀疏特征点法的VO是当前的主流方法^[32,34],它的基本思路是对于每帧新的图像 I_k (在立体相机中是一对图像),前两步是检测和匹配2D特征点,通过与之前帧进行对比匹配。二维特征点的重投影从不同图像帧提取共同的3D特征点,这就是图像对应关系(大部分VO实现的时候都假设相机已经校准)。第三步是计算时刻 $k-1$ 和 k 之间的相对运动 T_k 。根据对应关系是三维或二维,有三种不同的方法,包括2D-2D,对极约束、P3P、ICP^[35-38]等解决这个问题。相机位姿 C_k 是根据之前位姿变换 T_k 得到。除此之外,为了实现迭代优化(捆绑调整)获得更精确的局部轨迹估计,必须基于前 m 帧的深度估计构建局部地图。如图3是基于特征点法的VO系统的流程图。

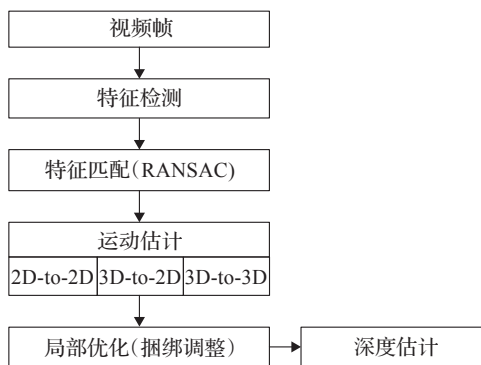


图3 基于特征点法的VO系统主要流程图

同时,注意由于存在噪声,错误的测量方法,以及对数据的错误假设等原因导致在特征匹配过程中会有一些匹配的局外点。即使在异常值的情况下,进行鲁棒估计是确保精确的运动估计的任务。由于局外点的比较分散性质,会使用随机采样一致性(RANSAC)来挑选最优匹配,而非最小二乘匹配算法。

VO研究的主要问题是如何根据图像来估计相机运动。通常情况下由于灰度值极易受到光照、形变等影响,不同图像间变化可能非常大,因此仅凭灰度值是不够的,所以需要对图像提取特征点。根据维基百科关于特征(计算机视觉)的定义,在计算机视觉和图像处理中,特征是一组与计算任务有关的信息,计算任务取决于具体的应用。特征也可能是一般邻域操作或者特征检测应用到图像的结果。特征在图像中可能拥有特殊结构,例如角点、边缘,或者区块物体^[39]。不过,一般更容易找出两幅图像中出现同一个角点,同一边缘则稍微困难些,同一区块则是最为困难的。所以,一种直观的特征提取方式就是辨认寻找不同图像角点,确定它们的对应关系。在这种情况下,角点就是所谓的特征。

然而在实际情况中,单纯的角点依然不能满足需求。因此研究人员设计了许多更加稳定的局部图像特征,如SIFT^[40]、SUFT^[41]等。虽然SIFT和SUFT充分考虑了图像变换过程中的各种问题,但是也带来了较大的计算量,一般来说很难实时在CPU上计算。不过近几年来,诸如ORB^[42]、BRISK^[43]等一些易于计算的特征提取/描述算法的流行,逐渐替代了之前追踪效果不好的Harris角点或计算复杂的SIFT/SUFT,成为VO的首选。

ORB由于融合了FAST和BRIEF各自的优点,使得其在尺度、旋转、亮度等方面具有良好的特性。同时,该组合也非常高效,使得ORB特征是目前实时性最好的方案^[16]。一般来说特征都是由关键点和描述子组成。其中,FAST角点提取:ORB为了在描述子中增加旋转不变性,在原版的FAST^[44]基础上,增加了特征点的主方向。新BRIEF描述子:对前一步提取的关键点周围像素区域进行描述,由于在角点提取的时候增加了主方向,所以相对于原始的BRIEF^[45]描述子,ORB的描述子具有较好的旋转不变性。

本文主要针对特征点提取三种主要方法进行比较,分别是SIFT、SURF以及ORB,这三种方法在OpenCV里面都已实现。如表2所示。

表2 不同特征之间性能比较

	特征类型		
	ORB	SURF	SIFT
计算速度	√√√√√	√√√	√
旋转鲁棒性	√	√√√	√
模糊鲁棒性	√	√√√	√
尺度变换鲁棒性	×	√√√	√

基于特征点的实时VO早期较为成功的就是Klein等人提出的单目VO框架——PTAM^[19]。虽然它的性能不是十分完善,但它提供了一个完整通用的框架,将整个里程计的实现分为前端、后端,分别包括跟踪和建图过程的并行化。目前多数的VO框架都是基于它实现的,包括目前最稳定的第二代基于ORB的即时定位与

地图构建(Simultaneous Location and Mapping based on ORB,ORB-SLAM2)^[16]。同时它也是第一个使用非线性优化的系统,在此之前的传统VO都是基于滤波器^[18]实现。不过,它也存在场景小,缺乏全局重定位功能,导致实用性较差。

由于光流法也具备跟踪特征点的特性,并且相对其他特征点匹配的方法可以节省部分计算量,所以也有人提出基于光流的特征点法^[46-47],虽然可以大幅提高VO的速度,但是要求相机运动较缓或者帧率较高。

现有研究中实用性最好的基于特征的VO方法是ORB-SLAM2^[16],它提出了一个更为完整的VO框架,如图4所示。包括跟踪、建图和回环检测三个线程。其中,跟踪线程主要负责对新一帧图像提取ORB^[42]特征点,并粗略估计相机位姿。建图线程主要是基于Bundle Adjustment对局部空间中的特征点与相机位姿的优化,求解误差更小的位姿与特征点的空间位姿。而回环检测线程负责实现基于关键帧的回环检测,可以有效消除累计误差,同时还可以进行全局重定位。同时它还兼容单目、双目和RGB-D相机等模式,这使它具有良好的泛用性。

对于初始化方面,作者提出了一种自动的初始化地图策略,同时计算单应矩阵(假设一个平面场景)^[31]和本质矩阵(假设非平面的场景)^[32],根据启发式的准则判断属于对应情况来初始化位姿。这也是文献[16]最大的贡献。ORB-SLAM与PTAM计算优势除了选取的ORB特征更加高效之外,还取了上一帧能观测的地图点进行匹配,而不是直接使用所有地图点来匹配新的帧。

3.2 基于直接跟踪的方法

特征点法一直是长期以来比较经典的方法,不过其鲁棒性主要建立于特征点的描述上。一方面越是增强鲁棒性,增加特征点描述的复杂性,越会导致算法复杂度的大幅提高;另一方面,特征点没法应用在特征点较弱的场景,例如墙面、天空等。所以基于像素灰度不变性假设估计相机运动的直接法在近年发展迅猛^[20-21]。直接法从光流^[48]发展而来,能够在不提特征(或不计算特征描述)的情况下,通过最小化光度误差(特征点法中最

小化特征点的重投影误差),来估计相机运动和像素的空间位置。可以有效地解决特征点法所面临的问题。总的来说,根据空间点 P 多少情况,直接法又分为三类:稀疏直接法、半稠密的直接法和稠密直接法。

早期直接的VO方法很少基于跟踪和建图框架,多数都是人工选择关键点^[49-51]。直到RGB-D相机的出现,研究人员发现直接法对RGB-D相机^[26],进而对单目相机^[21-22]都是非常有效的手段。近期出现的一些直接法都是直接使用图像像素点的灰度信息和几何信息来构造误差函数,通过图优化求解最小化代价函数,从而得到最优相机位姿,且处理大规模地图问题用位姿图表示^[21,52]。为了构建半稠密的三维环境地图,Engel等人^[22]提出了大尺度的直接单目即时定位与地图构建(Large-Scale Direct monocular Simultaneous Location and Mapping, LSD-SLAM)算法,相比之前的直接的VO方法,该方法在估计高精度相机位姿的同时能够创建大规模的三维环境地图。由于单目VO存在尺度不确定性和尺度漂移问题,地图直接由关键帧直接的Sim(3)变换组成,能够准确地检测尺度漂移,并且整个系统可在CPU上实时运行。与ORB-SLAM2类似,LSD-SLAM也采用位姿图优化,因此能形成闭环回路和处理大尺度场景。系统为每个新加入的关键帧在已有关键帧集合(地图)中选取距离最近的关键帧位置。LSD-SLAM主要流程图如图5所示。

DSO^[21](Direct Sparse Odometry,直接稀疏里程计)也是由LSD-SLAM的作者Engel提出的,该方法不仅从鲁棒性、精准度还是计算速度都远远超过之前的ORB-SLAM和LSD-SLAM等方法的效果。因为采用新的深度估计机制滑动窗口优化代替原来的卡尔曼滤波方法,所以在精度上有了十足的提高。另外,与LSD-SLAM相比,DTAM^[25]引入直接法计算基于单目相机的实时稠密地图。相机的位姿使用深度图直接匹配整个图像得到。然而,从单目视觉中计算稠密深度需要大量的计算力,通常是使用GPU并行运算,例如开源的REMODE^[53]。因此,也有不少研究人员在这方面做了一些平衡诞生了

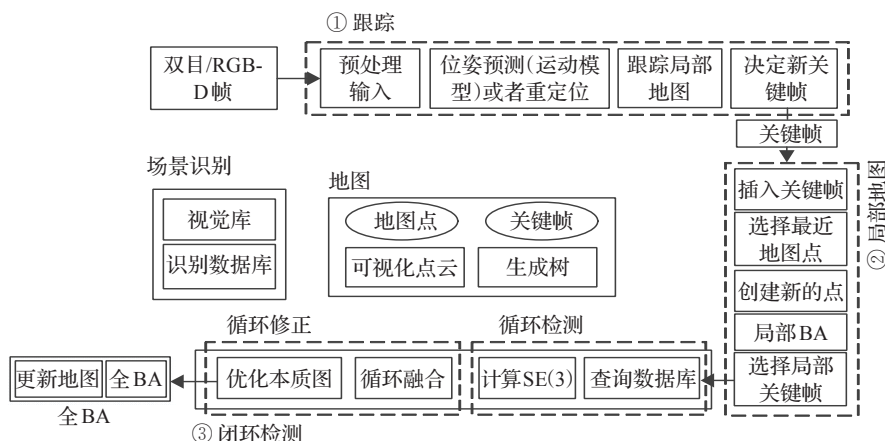


图4 ORB-SLAM2框架结构图

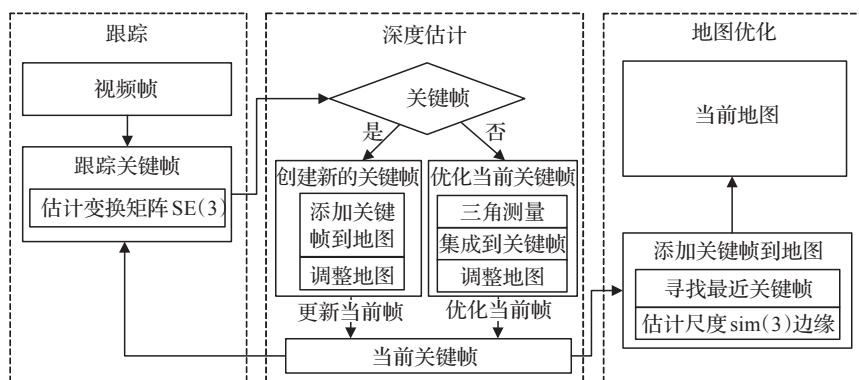


图5 LSD-SLAM的模块流程图

能达到更快计算速度的方法,例如文献[54]和文献[20]。

3.3 基于混合的半直接跟踪的方法

虽然基于直接跟踪的方法已经非常流行,但是低速以及没法保证最优性和一致性也是限制直接法的问题所在。因此有人在基于特征的方法和基于直接跟踪的方法两者各自优点的基础上,提出了一种混合的半直接方法即半直接视觉里程计(Semi-direct Visual Odometry, SVO)^[20],虽然SVO依旧是依赖于特征一致性,但是它的思路主要是还是通过直接法来获取位姿,因此避免了特征匹配和外围点处理,极大地缩短了计算时间,算法速度非常快。在嵌入式无人机平台(ARM Cortex A9 1.6 GHz CPU)上可以做到55 f/s,而在普通笔记本上(Intel i7 2.8 GHz CPU)上可以高达300 f/s。

深度估计是构建局部点云地图的核心,SVO也是采用概率模型建图。不过跟LSD-SLAM等方法不同的是,SVO的深度滤波是采用高斯分布和均匀分布的混合模型^[55],而LSD-SLAM是基于高斯分布模型。首先使用直接法求解位姿进行匹配,其次使用了经典的Lucas-Kanade光流法^[48]匹配得到子像素精度,然后结合点云地图计算最小化重投影误差进行优化。如图6所示。

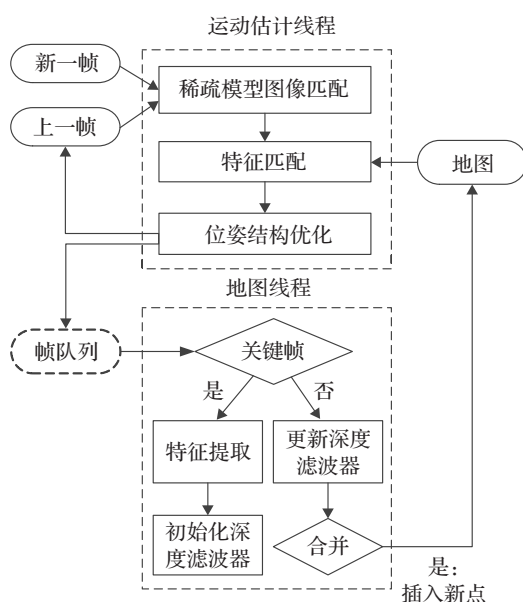


图6 SVO的模块流程图

整个过程相较于传统的特征点法,只有选择关键帧时需要依靠特征,并且去掉了匹配描述子的计算,以及使用RANSAC去除局外点的步骤,所以更加高效。同时它相比于直接法不是对整幅图像进行直接匹配从而获得相机位姿,而是通过在整幅图像中提取的图像块来进行位姿的获取,这样能够增强算法的鲁棒性。SVO最大贡献就是在巧妙设计了三次优化的方法(优化灰度误差,优化特征点预测位置,优化重投影误差)来满足精度问题的同时,也保持较为优秀的计算速度。另外,其代码结构较为简洁,非常适合深入研究学习。后来Forster又证明了该方法可以拓展到多目相机系统^[56],跟踪边缘,包括运动的先验知识,同时也支持多种相机,例如鱼眼相机和透视相机。不过在半直接法方面的研究,除了Forster最新的研究^[56]之外,目前还未出现其他有大影响力的论文。

3.4 优缺点分析

特征点法一直是长期以来比较经典的方法,不过其鲁棒性主要建立于特征点的描述上,一方面越是增强鲁棒性,增加特征点描述的复杂性,也会导致算法复杂度的大幅提高;另一方面,特征点没法应用在特征点较弱的场景,例如墙面上。直接法是近几年比较新颖的方法,它可以适应于特征不够明显的场景下,例如走廊或者光滑的墙面上^[57],具有较强鲁棒性。由于跳过了特征描述和匹配的步骤,直接法,尤其是稀疏直接法,往往能够在极快的速度下运行。它也适用于那些需要构建半稠密或者稠密的地图的需求场景,这是特征点法是无法实现的。但是直接法也存在非凸性、单个像素没有区分度和灰度不变性假设性太牵强等问题,因此其研究和应用仍然没有特征点法成熟,目前只适合于运动较小,图像整体亮度变化不大的情形。

虽然基于特征点的VO^[16,23]更为主流,但是从慕尼黑大学TUM组公布的实验结果来看,直接法的VO^[20-22]在近几年也取得很大的突破,其中基于稀疏的直接法^[21]已经比稀疏的特征点法^[16]具有更快更好的效果。直接方法使用了图像上的所有信息,甚至是像素梯度很小的区域,因此即使在场景纹理很差,失焦,运动模糊的情况下

的性能也优于基于特征的方法。根据文献[21]对基于直接跟踪的方法和基于特征的方法进行的噪声实验对比,基于直接跟踪的方法对几何噪声较为敏感,例如卷帘快门相机等;而基于特征的方法对光学噪声更为敏感,例如模糊等。因此,在普通手机设备上(一般为卷帘快门相机),基于特征的方法效果可能更好;而在基于全局快门相机的机器人中,基于直接跟踪的方法可能越来越流行。

基于混合的半直接跟踪的方法^[20],由Forster最早提出,具有速度快,适合于地图不确定性的模型,同时不受运动模型假设的影响的优点;然而由于跟踪的特征比较少,有些情况下可能会丢失。作者不仅发布了惊艳的实验测试视频,并开源了其代码框架。虽然其开源的代码效果鲁棒性不是很好,不过由于代码规范性较好,依旧很适合初学者进行阅读。为了更好地对比了解当前各个方法的进展情况,本文分别选取了当前基于特征的方法、基于直接跟踪的方法、基于混合的半直接跟踪的方法中最具代表性的方法进行实验测试,结果分别如图7所示。

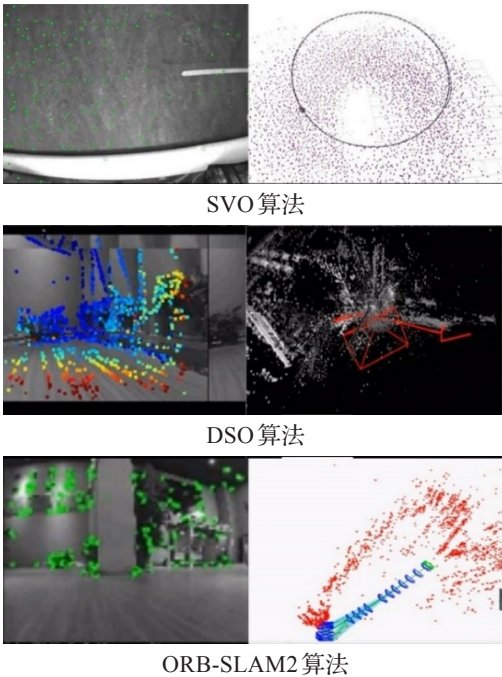


图7 三类VO方法实现效果对比

4 VO主要发展趋势及研究热点

目前下表中所列的国内外学术科研机构,对VO展开了不同侧重的研究,如表3所示。

虽然VO问题研究本质上是增量地计算相机位姿问题,为上层应用提供自身的一个位姿估计^[3],但是如何进一步提高精度、效率、鲁棒性等问题一直是研究人员不懈的追求。围绕着上述三个问题,目前已经形成探索新型传感器、多传感器数据融合、应用机器学习、探究新的缓解特征依赖和降低计算复杂度等几个方面入手的研究热点。

表3 国内外前沿机构的研究方向

研究机构	侧重方向
清华大学	基于平面的特征匹配、基于深度学习的语义地图
苏黎世大学	直接法、基于新型视觉传感器的VO
慕尼黑大学	直接法
香港科技大学	视觉惯导融合
谷歌	RGB-D相机与惯导融合
苹果	视觉惯导融合
苏黎世联邦理工学院	视觉惯导融合
浙江大学	双目立体视觉
美国国家航空航天局	双目立体视觉

4.1 探索新型传感器

随着2010年微软推出RGB-D相机Kinect的兴起,它具有能够实时获取深度图的特性,能够简化大量的计算,也逐渐成为一种稠密三维重建系统的实现方式^[7,24,26-28,30],但是一方面由于其有效距离较短,另一方面容易受到外界光源的干扰无法在室外场景中使用,限制了它没法真正解决VO问题。文献[58]提出基于事件相机的VO算法,并且基于扩展卡尔曼滤波器与无结构的测量模型,集成了IMU作为数据融合的补充,以精确得到6自由度相机的位姿。未来随着新型传感器的出现,势必会引发一阵新的热点。

4.2 多传感器数据融合

对于很多移动机器人来说,IMU和视觉都是必备的传感器,它们可以数据融合互补,满足移动机器人系统的鲁棒性和定位精度的需求。单目摄像头和惯导融合^[8-10,31,59]也是一个近几年比较流行的一个趋势,苹果公司在WWDC 2017大会上推出的ARKit,主要就是基于EKF对单目相机和惯导数据融合的思路实现,为开发者做室内定位提供良好的基础平台支撑。后来又有人提出了用优化关键帧^[60]方式对多目相机和惯导数据进行融合的思路^[9]。数据融合分为紧耦合和松耦合。一方面,有时候为了限制计算复杂度,许多工作遵循松耦合的原则。文献[31]集成IMU作为独立姿态和相关偏航测量加入到视觉的非线性优化问题。相反,文献[61]使用视觉位姿估计维护一个间接IMU的EKF。类似的松耦合算法还有文献[62]和[63],相机的位姿估计使用非线性优化集到了因子图,包括惯导和GPS数据。另一方面,由于松耦合方法本质上是忽略了不同传感器内部之间的相关性,所以紧耦合方法是将相机和IMU数据合并,将所有状态都联合估计成一个共同问题,因此需要考虑它们之间的相关性。文献[9]将两类方法进行比较,实验表明这些传感器内部的相关性对于高精度的视觉惯导系统(VINS)是非常关键的,所以高精度视觉惯导系统都是采用紧耦合来实现。

有学者尝试多传感器的融合,首先是杨绍武提出的多相机传感器的融合^[64],还有双目立体视觉与惯导、速

度等数据融合^[65],其次是Akshay提出的基于点云特征的GPS-Lidar融合算法,在3D城市建模过程中能有效地降低的位置测量误差^[66]。

4.3 应用机器学习

神经网络等机器学习方法近年来在众多领域中引起了广泛的学术轰动,VO领域也不例外,在匹配跟踪部分,文献[67]提出了一种数据驱动模型(即3DMatch),通过自监督的特征学习从现有的RGB-D重建结果中获得局部空间块的描述子,进而建立局部3D数据之间的对应关系。对于优化匹配误差,传统的RANSAC可能被一种新的Highway Network架构替代,它基于多级加权残差的跳层连接(Multilevel Weighted Residual Shortcuts)的方式,计算每个可能视差值的匹配误差,并利用复合损失函数进行训练,支持图像块的多级比较。在精细化步骤中可用于更好地检测异常点。文献[68]针对这种新架构应用立体匹配基准数据集进行实验,结果也表明匹配置信度远远优于现有算法。

单目VO缺乏尺度信息一直是研究人员最为关注的问题,近期有德国研究人员Keisuke等人针对低纹理区域等单目VO恢复尺度容易失败的情况,提出一种将CNN预测的深度信息与单目直接计算的深度信息进行融合的方法,实验表明,它解决了单目VO的一个尺度信息丢失问题^[69]。

2016年Muller提出了基于光流(直接法)和深度学习的VO^[70],光流的帧作为CNN的输入,计算旋转和平移,顺序增量式的旋转和平移构建相机运动轨迹地图。实验证明该方法比现有的VO系统具有更高的实时性。

4.4 探究新的缓解特征依赖

VO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),因此目前出现了不少研究提出了利用边、平面^[71]等更为高层的图像信息来缓解特征依赖。理论上由于边可以携带方向、长短、灰度值等信息,所以具备更为鲁棒的特性,基于边的特征在室内场景(规则物品较多)应具有更好的鲁棒性。文献[72]提出了一种结合点与边缘优点的单目VO算法。该算法不仅在TUM提供的单目公开数据集^[21]中表现优异,而且在低纹理的环境中,可以大幅降低运动估计误差。文献[6]主要应用了图模型和图匹配机制对平面物体进行跟踪,并且设计一种解决最优解寻找问题的新策略,该策略能预测物体姿态和关键点匹配。

4.5 降低计算复杂度

目前基于RGB-D相机的实时恢复稠密场景已经较为完善^[17,24,26-28,30],近年来,由于AR应用研究的爆发,较早的开始研究为AR提供基础技术支撑的是谷歌的Schöps等人,他们提出基于TSDF来融合深度图在Project Tango上实现的三维重建方法^[74],其主要计算复杂度在于半稠密或稠密重建所需计算的深度估计点太多。所

以目前多数的优化手段大多集中在优化深度估计步骤,例如DTAM^[25]引入正则项对深度图进行全局优化,降低错误匹配概率。REMODE^[53]则采用了深度滤波模型,不断优化每帧的深度测量更新概率模型的参数。上述方法虽然都能实时重建出稠密的三维点云地图,但大多还都依赖于GPU的并行加速运算。因此,如何提高计算效率,降低计算复杂度,只用CPU即可恢复出基于单目的半稠密或稠密三维点云地图,仍是未来一个热门课题。

5 结论

本文从VO与视觉SLAM的对比分析入手,对VO问题进行形式化。随后重点探究实现VO的各类方法的研究进展,实验对比分析各自优劣。最后结合国内外一流的科研机构在研方向,总结今后发展研究热点。目前多数研究人员只关注白天等视野较好的理想场景,但是场景变化(白天黑夜、四季变化等)问题在实际室内外场景中很常见,如何让VO系统在这样的环境下依旧实现高鲁棒性应是科研人员一个重要研究内容。此外,为了达到实时效果,VO的计算复杂度也不能太高。

未来也可以在以下领域展开新的应用研究:在消防领域,消防人员可以对大型室内火灾救援现场进行定位,并实时绘制出运动轨迹帮助救灾人员标记已经搜救完成的地方,将搜救工作效率最大化;在反恐领域中,针对突发事件中对警犬的行为特征进行检测分析,以便实现在复杂人群中通过警犬对突发事件实现前方预警。将VO应用到该场景中,对警犬的位姿进行准确的定位,进而以一种低成本的方式提供预警功能。

参考文献:

- [1] Durrantwhyte H, Bailey T. Simultaneous localization and mapping: Part I[J]. IEEE Robotics & Automation Magazine, 2006, 13(3): 108-117.
- [2] Durrantwhyte H, Bailey T. Simultaneous localization and mapping: Part II[J]. IEEE Robotics & Automation Magazine, 2006, 13(3): 108-117.
- [3] Nister D, Naroditsky O, Bergen J. Visual odometry[C]// Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004: 652-659.
- [4] Matthies L, Maimone M, Johnson A, et al. Computer vision on mars[J]. International Journal of Computer Vision, 2007, 75(1): 67-92.
- [5] Malleson C, Gilbert A, Trumble M, et al. Real-time full-body motion capture from video and IMUs[C]// Proceedings of International Conference on 3D Vision, 2017.
- [6] Wang T, Ling H. Gracker: A graph-based planar object tracker[J]. IEEE Transactions on Pattern Analysis & Machine

- Intelligence, 2017, 99.
- [7] Endres F, Hess J, Sturm J, et al. 3-D mapping with an RGB-D camera[J]. IEEE Transactions on Robotics, 2017, 30(1): 177-187.
- [8] Bloesch M, Omari S, Hutter M, et al. Robust visual inertial odometry using a direct EKF-based approach[C]// Proceedings of International Conference on Intelligent Robots and Systems, 2015: 298-304.
- [9] Leutenegger S, Lynen S, Bosse M, et al. Keyframe-based visual-inertial odometry using nonlinear optimization[J]. International Journal of Robotics Research, 2015, 34(3): 314-334.
- [10] Qin T, Li P, Shen S. VINS-Mono: A robust and versatile monocular visual-inertial state estimator[J]. arXiv: 1708.03852v1, 2017.
- [11] Strasdat H, Montiel J M M, Davison A J. Visual SLAM: Why filter?[J]. Image & Vision Computing, 2012, 30(2): 65-77.
- [12] 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, 2010: 2657-2664.
- [13] Handa A, Chli M, Strasdat H, et al. Scalable active matching[C]// Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2010: 1546-1553.
- [14] Civera J, Grasa O G, Davison A J, et al. 1-Point RANSAC for extended Kalman filtering: Application to real-time structure from motion and visual odometry[J]. Journal of Field Robotics, 2010, 27(5): 609-631.
- [15] Mei C, Sibley G, Cummins M, et al. RSLAM: A system for large-scale mapping in constant-time using stereo[J]. International Journal of Computer Vision, 2011, 94(2): 198-214.
- [16] Mur-Artal R, Tardós J D. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras[J]. IEEE Transactions on Robotics, 2016, 33(5): 1255-1262.
- [17] 高翔. 视觉SLAM十四讲[M]. 北京: 电子工业出版社, 2017.
- [18] Davison A J, Reid I D, Molton N D, et al. MonoSLAM: Real-time single camera SLAM[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(6): 1052.
- [19] Klein G, Murray D. Parallel tracking and mapping for small AR workspaces[C]// Proc of IEEE & ACM Int Sympo on Mixed & Augmented Reality, 2007: 1-10.
- [20] Forster C, Pizzoli M, Scaramuzza D. SVO: Fast semi-direct monocular visual odometry[C]// Proceedings of IEEE International Conference on Robotics and Automation, 2014: 15-22.
- [21] Engel J, Koltun V, Cremers D. Direct sparse odometry[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 40(3): 611-625.
- [22] Engel J, Schöps T, Cremers D. LSD-SLAM: Large-scale direct monocular SLAM[C]// Proceedings of European Conference on Computer Vision, 2014: 834-849.
- [23] Mur-Artal R, Montiel J M M, Tardós J D. ORB-SLAM: A versatile and accurate monocular SLAM system[J]. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.
- [24] Labbé M, Michaud F. Online global loop closure detection for large-scale multi-session graph-based SLAM[C]// Proceedings of International Conference on Intelligent Robots and Systems, 2014: 2661-2666.
- [25] 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, 2011: 2320-2327.
- [26] Kerl C, Sturm J, Cremers D. Dense visual SLAM for RGB-D cameras[C]// Proceedings of International Conference on Intelligent Robots and Systems, 2014: 2100-2106.
- [27] Whelan T, Salas-Moreno R F, Glocker B, et al. ElasticFusion: Real-time dense SLAM and light source estimation[J]. International Journal of Robotics Research, 2016, 35(14): 1697-1716.
- [28] Whelan T, Leutenegger S, Moreno R S, et al. ElasticFusion: Dense SLAM without a pose graph[J]. International Journal of Robotics Research, 2016, 35(14): 1-9.
- [29] Bloesch M, Burri M, Omari S, et al. Iterated extended Kalman filter based visual-inertial odometry using direct photometric feedback[J]. International Journal of Robotics Research, 2017, 36(10): 1053-1072.
- [30] Izadi S, Kim D, Hilliges O, et al. KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera[C]// Proceedings of ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, 2011: 559-568.
- [31] Konolige K, Agrawal M, Solà J. Large-scale visual odometry for rough terrain[C]// Proceedings of International Symposium on Robotics Research, November 26-29, 2011: 201-212.
- [32] Quijada S D, Zalama E, García-Bermejo J G, et al. Fast 6D odometry based on visual features and depth[M]// Intelligent Autonomous Systems 12. Berlin Heidelberg: Springer, 2013: 5-16.
- [33] Tang C, Wang O, Tan P. GlobalSLAM: Initialization-robust Monocular Visual SLAM[J]. arXiv: 1708.04814v1, 2017.
- [34] Scaramuzza D, Fraundorfer F. Visual Odometry[Tutorial][J].

- IEEE Robotics & Automation Magazine, 2011, 18(4): 80-92.
- [35] Hartley R I. In defense of the eight-point algorithm[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1997, 19(6): 580-593.
- [36] Besl P J, McKay N D. A method for registration of 3-D shapes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992, 14(2): 239-256.
- [37] Penate-Sanchez A, Andrade-Cetto J, Moreno-Noguer F. Exhaustive linearization for robust camera pose and focal length estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(10): 2387-2400.
- [38] Lepetit V, Moreno-Noguer F, Fua P. EPnP: An accurate $O(n)$ solution to the PnP problem[J]. International Journal of Computer Vision, 2009, 81(2): 155-166.
- [39] Wikipedia. Feature (computer vision) [EB/OL]. (2016-07-09) [2017-11-01]. [https://en.wikipedia.org/wiki/Feature_\(computer_vision\)](https://en.wikipedia.org/wiki/Feature_(computer_vision)).
- [40] Lowe D G. Distinctive Image features from scale-invariant key points[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.
- [41] Bay H, Tuytelaars T, Gool L V. SURF: Speeded up robust features[C]//Proceedings of European Conference on Computer Vision, 2006: 404-417.
- [42] 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, 2012: 2564-2571.
- [43] Leutenegger S, Chli M, Siegwart R Y. BRISK: Binary robust invariant scalable keypoints[C]//Proceedings of International Conference on Computer Vision, 2011: 2548-2555.
- [44] Rosten E, Drummond T. Machine learning for high-speed corner detection[C]//Proceedings of European Conference on Computer Vision, 2006: 430-443.
- [45] Calonder M, Lepetit V, Strecha C, et al. BRIEF: Binary robust independent elementary feature[C]//Proceedings of European Conference on Computer Vision, 2010: 778-792.
- [46] Kitt B, Geiger A, Lategahn H. Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme[C]//Proceedings of Intelligent Vehicles Symposium, 2010: 486-492.
- [47] Geiger A, Ziegler J, Stiller C. StereoScan: Dense 3D reconstruction in real-time[C]//Proceedings of IEEE Intelligent Vehicles Symposium, 2011: 963-968.
- [48] Baker S, Matthews I. Lucas-Kanade 20 years on: A unifying framework[J]. International Journal of Computer Vision, 2004, 56(3): 221-255.
- [49] Favaro P, Jin H, Soatto S. A semi-direct approach to structure from motion[C]//Proceedings of International Conference on Image Analysis and Processing, 2001: 250-255.
- [50] Benhimane S, Malis E. Integration of Euclidean constraints in template based visual tracking of piecewise-planar scenes[C]//Proceedings of International Conference on Intelligent Robots and Systems, 2007: 1218-1223.
- [51] Silveira G, Malis E, Rives P. An efficient direct approach to visual SLAM[J]. IEEE Transactions on Robotics, 2008, 24(5): 969-979.
- [52] Gokhool T, Meilland M, Rives P, et al. A dense map building approach from spherical RGBD images[C]//Proceedings of International Conference on Computer Vision Theory and Applications, 2014: 656-663.
- [53] Pizzoli M, Forster C, Scaramuzza D. REMODE: Probabilistic, monocular dense reconstruction in real time[C]//Proceedings of IEEE International Conference on Robotics and Automation, 2014: 2609-2616.
- [54] Engel J, Cremers D. Semi-dense visual odometry for a monocular camera[C]//Proceedings of IEEE International Conference on Computer Vision, 2014: 1449-1456.
- [55] Vogiatzis G, Hernández C. Video-based, real-time multi-view stereo[J]. Image & Vision Computing, 2011, 29(7): 434-441.
- [56] Forster C, Zhang Z, Gassner M, et al. SVO: Semidirect visual odometry for monocular and multicamera systems[J]. IEEE Transactions on Robotics, 2017, 33(2): 249-265.
- [57] Lovegrove S, Davison A J, Ibañez-Guzmán J. Accurate visual odometry from a rear parking camera[C]//Proceedings of Intelligent Vehicles Symposium, 2011: 788-793.
- [58] Zhu A Z, Atanasov N, Daniilidis K. Event-based visual inertial odometry[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017: 5816-5824.
- [59] Lin Y, Gao F, Qin T, et al. Autonomous aerial navigation using monocular visual-inertial fusion[J]. Journal of Field Robotics, 2018, 35(4): 23-51.
- [60] Gui J, Gu D, Wang S, et al. A review of visual inertial odometry from filtering and optimization perspectives[J]. Advanced Robotics, 2015, 29(20): 1289-1301.
- [61] Weiss S, Achtelik M W, Lynen S, et al. Real-time onboard visual-inertial state estimation and self-calibration of MAVs in unknown environments[C]//Proceedings of IEEE International Conference on Robotics and Automation, 2012: 957-964.
- [62] Dellaert F, Ranganathan A, Kaess M. Fast 3D pose estimation with out-of-sequence measurements[C]//Proceedings of IEEE International Conference on Intelligent Robots and Systems, 2007: 2486-2493.

(下转55页)

- by active agents as a model for daphnia swarming[J]. Physical Review E: Statistical, Nonlinear, and Soft Matter Physics, 2006, 73(6).
- [7] Ru Lining, Li Zhuchun, Xue Xiaoping. Cucker-Smale flocking with randomly failed interactions[J]. Journal of the Franklin Institute, 2015, 352(3): 1099-1118.
- [8] Cucker F, Dong Jiugang. A general collision-avoiding flocking framework[J]. IEEE Transactions on Automatic Control, 2011, 56(5): 1124-1128.
- [9] Dong Jiugang. Flocking under hierarchical leadership with a free-will leader[J]. International Journal Robust Non-linear Control, 2013, 23: 1891-1898.
- [10] Erban R, Haskovec J, Sun Yongzheng. On Cucker-Smale model with noise and delay[J]. SIAM J Appl Math, 2015, 76(4): 1535-1557.
- [11] Liu Yicheng, Wu Jianhong. Flocking and asymptotic velocity of the Cucker-Smale model with processing delay[J]. Journal of Mathematical Analysis and Applications, 2014, 415(1): 53-61.
- [12] Breder C M. Equations descriptive of fish schools and other animal aggregations[J]. Ecology, 1954, 35: 361-370.
- [13] Grunbaum D, Okubo A. Modeling social animal aggregations[J]. Biomathematics, 1994, 100: 296-325.
- [14] Okubo A. Dynamical aspects of animal grouping: Swarms, schools, flocks, and herds[J]. Adv Biophys, 1986, 22: 1.
- [15] Cho J, Ha S Y, Huang F, et al. Emergence of bi-cluster flocking for the Cucker-Smale model[J]. Math Mod Meth Appl Sci, 2016, 26(6): 1191-1218.
- [16] Cho J, Ha S Y, Huang F, et al. Emergence of bi-cluster flocking for agent-based models with unit speed constraint[J]. Analysis and Applications, 2016, 14: 39-73.
- [17] Li Z, Ha S Y. On the Cucker-Smale flocking with alternating leaders[J]. Quart Appl Math, 2015, 73: 693-709.
- [18] Li Z, Ha S Y, Xue X. Emergent phenomena in an ensemble of Cucker-Smale particles under joint rooted leadership[J]. Math Mod Meth Appl Sc, 2014, 24: 1389-1419.
- [19] Reynolds C W. Flocks, herds, and schools: A distributed behavioral model[J]. Computer Graphics, 1987, 21: 25-34.
- [20] Vicsek T, Czirok A, Ben-Jacob E, et al. Novel type of phase transition in a system of self-driven particles[J]. Physical Review Letters, 1995, 75: 1226-1229.
- [21] Cucker F, Smale S. Emergent behavior in flocks[J]. IEEE Transactions on Automatic Control, 2007, 52(5): 852-862.
- [22] Cucker F, Smale S. On the mathematics of emergence[J]. Japanese Journal of Mathematics, 2007, 2(1): 197-227.
- [23] Shen J. Cucker-Smale flocking under hierarchical leadership[J]. SIAM J Appl Math, 2007, 68: 694-719.
- [24] Cucker F, Dong J G. On the critical exponent for flocks under hierarchical leadership[J]. Math Mod Meth Appl Sci, 2009, 19: 1391-1404.

(上接28页)

- [63] Indelman V, Williams S, Kaess M, et al. Factor graph based incremental smoothing in inertial navigation systems[C]//Proceedings of International Conference on Information Fusion, 2012: 2154-2161.
- [64] Yang S, Scherer S A, Yi X, et al. Multi-camera visual SLAM for autonomous navigation of micro aerial vehicles[J]. Robotics & Autonomous Systems, 2017, 93: 116-134.
- [65] Usenko V, Engel J, Stücker J, et al. Direct visual-inertial odometry with stereo cameras[C]//Proceedings of IEEE International Conference on Robotics and Automation, 2016: 1885-1892.
- [66] Shetty A P. GPS-LiDAR sensor fusion aided by 3D city models for UAVs[Z]. 2017.
- [67] Zeng A, Song S, Niebner M, et al. 3DMatch: Learning local geometric descriptors from RGB-D reconstructions[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017: 199-208.
- [68] Shaked A, Wolf L. Improved stereo matching with constant highway networks and reflective confidence learning[C]//Proceedings of Conference on Computer Vision and Pattern Recognition, 2016.
- [69] Tateno K, Tombari F, Laina I, et al. CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction[C]//Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2017: 6565-6574.
- [70] Muller P, Savakis A. Flowdometry: An optical flow and deep learning based approach to visual odometry[C]//Proceedings of Conference on Applications of Computer Vision, 2017: 624-631.
- [71] Gao X, Zhang T. Robust RGB-D simultaneous localization and mapping using planar point features[J]. Robotics & Autonomous Systems, 2015, 72: 1-14.
- [72] Yang S, Scherer S. Direct monocular odometry using points and lines[C]//Proceedings of Conference on IEEE International Conference on Robotics and Automation, 2017: 3871-3877.
- [73] Schöps T, Sattler T, Häne C, et al. 3D Modeling on the Go: Interactive 3D reconstruction of large-scale scenes on mobile devices[C]//Proceedings of Conference on International Conference on 3D Vision, 2015: 291-299.