doi:10.19306/j.cnki.2095-8110.2020.04.007

视觉/惯性组合导航技术发展综述

张礼廉,屈　豪,毛　军,胡小平

(国防科技大学智能科学学院,长沙 410073)

摘　要:视觉/惯性组合导航技术是自主导航领域的重要研究方向之一。首先介绍了视觉/惯性组合导航技术的发展概况,然后从纯视觉导航(里程计、同步定位与构图)以及组合导航(滤波、非线性优化)2 个层次介绍了传统的基于视觉几何与运动学模型的视觉/惯性组合导航方法,还介绍了近年来发展迅猛的基于机器学习的视觉/惯性组合导航方法。最后,简要介绍了视觉/惯性组合导航技术的典型应用及未来发展趋势。

关键词:视觉里程计;同步定位与构图;惯性导航;组合导航

中图分类号:V249　　**文献标志码**:A　　**开放科学(资源服务)标识码(OSID)**:

文章编号:2095-8110(2020)04-0050-14

A Survey of Intelligence Science and Technology Integrated Navigation Technology

ZHANG Li-lian, QU Hao, MAO Jun, HU Xiao-ping

(College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China)

Abstract: Visual-inertial integrated navigation technology is an important research direction in autonomous navigation. This paper first introduces the development of visual-inertial integrated navigation technology, then introduces the traditional visual-inertial integrated method based on visual geometry and kinematics model from the viewpoint of pure visual navigation (odometry, simultaneous localization and mapping) and integrated navigation (filtering, nonlinear optimization), then introduces the visual-inertial integrated navigation method based on machine learning which has developed rapidly in recent years. Finally, the typical applications and the future trend of visual-inertial integrated navigation technology are briefly introduced.

Key words: Visual odometry; Simultaneous localization and mapping; Inertial navigation; Integrated navigation

0　引言

随着无人机、无人车以及移动机器人的井喷式发展,导航技术成为了制约无人平台广泛应用的瓶颈技术之一。在应用需求的牵引下,视觉/惯性组合导航技术,特别是视觉与微惯性传感器的组合,逐渐发展成为当前自主导航及机器人领域的研究热点。本文介绍的视觉/惯性组合导航技术侧重于利用视觉和惯性信息估计载体的位置、速度、姿态等运动参数以及环境的几何结构参数,而不包含场景障碍物检测以及载体运动轨迹规划等。

视觉/惯性组合导航具有显著的优点:1)微惯

收稿日期:2020-05-18;**修订日期**:2020-06-19

基金项目:国家自然科学基金面上项目(61773394)

作者简介:张礼廉(1985-),男,博士,副教授,主要从事组合导航技术方面的研究。E-mail:lilianzhang@nudt.edu.cn

性器件和视觉传感器具有体积小、成本低的优点，随着制造技术的不断进步，器件越来越小，且成本越来越低。2)不同于卫星和无线电导航，视觉和惯性导航均不依赖外部设施支撑，可以实现自主导航。3)惯性器件和视觉器件具有很好的互补性，惯性导航误差随时间累积，但是在短时间内可以很好地跟踪载体快速运动，保证短时间的导航精度；而视觉导航在低动态运动中具有很高的估计精度，且引入了视觉闭环矫正可以极大地抑制组合导航误差，两者的组合可以更好地估计导航参数。

视觉和惯性组合导航技术近年来取得了长足的发展。孙永全和田红丽[1]从同步定位与构图(Simultaneous Localization and Mapping, SLAM)的角度对视觉/惯性组合导航技术的基本原理和标志性成果进行了详细分析。Huang[2]对基于滤波技术的视觉/惯性组合导航技术进行了全面的描述，特别是对滤波器的可观性和滤波状态的一致性问题进行了深入的探讨。Huang 和 Zhao 等[3]对基于激光和视觉传感器的 SLAM 技术进行了全面的介绍，该文引用的文献十分全面，但缺乏基本原理的阐述。当前随着基于机器学习的视觉/惯性组合导航算法性能不断提高，部分算法已达到甚至超过传统的基于模型的组合导航算法性能。因此，非常有必要按照基于模型的算法和基于机器学习的算法对视觉/惯性组合导航技术进行详细的分析。

1　视觉/惯性组合导航技术发展简述

传统的基于视觉几何与运动学模型的视觉和惯性导航技术研究成果非常丰富。本文主要从纯视觉导航以及组合导航 2 个层次梳理相关工作。

纯视觉导航技术主要有 2 个分支：一个分支是视觉里程计(Visual Odometry，VO)技术；而另一个分支是视觉同步定位与构图(Visual Simultaneous Localization and Mapping，VSLAM)技术。Scaramuzza 教授[4-5]对早期的 VO 技术进行了详细的介绍，并阐述了 VO 技术与 VSLAM 技术的区别与联系：VO 侧重于利用连续图像帧之间的位姿增量进行路径积分，至多包含滑动窗口内的局部地图；VSLAM 侧重于全局路径和地图的优化估计，支持重定位和闭环优化；通常 VO 可以作为 VSLAM 算法框架的前端。

目前，视觉里程计可以根据使用相机个数的不同分为单目、双目和多目视觉里程计。其中最具有代表性和影响力的主要有三种算法，分别是视觉里程计库(Library for Visual Odometry，LIBVISO)[6]、半直接单目视觉里程计(Semi-Direct Monocular Visual Odometry，SVO)[7]和直接稀疏里程计(Direct Sparse Odometry，DSO)[8]。这三种算法由于代码公开，易于使用，运动估计效果好，成为了研究者们广泛使用和对比的算法。

对于 VSLAM 算法，目前主流的方法可以分为两类：一类是基于滤波的方法；另一类是基于 Bundle Adjustment 的优化算法。这两类方法的开创性成果分别是 Davison 教授提出的 Mono SLAM 算法[9]和 Klein 博士提出的并行跟踪与构图(Parallel Tracking And Mapping，PTAM)算法[10]。在 2010 年国际机器人和自动化大会(IEEE International Conference on Robotics and Automation，ICRA)上，Strasdat 的文章[11]指出优化算法比滤波算法的性价比更高，从此以后基于非线性优化的 VSLAM 算法就渐渐多起来。其中代表性的工作是 ORB-SLAM[12]和 LSD-SLAM[13]，二者的主要区别是 ORB-SLAM 的前端采用稀疏特征，而 LSD-SLAM 的前端采用稠密特征。

当然，任何纯视觉导航算法都存在无法避免的固有缺点：依赖于场景的纹理特征、易受光照条件影响以及难以处理快速旋转运动等。因此，为了提高视觉导航系统的稳定性，引入惯性信息是很好的策略。

视觉/惯性组合导航技术与 VSLAM 算法类似，主要采用两种方案：一种是采用滤波技术融合惯性和视觉信息；另一种是采用非线性迭代优化技术融合惯性和视觉信息。

基于滤波技术的视觉/惯性组合导航算法，可以进一步分为松组合和紧组合两种框架。文献[14-15]使用了卡尔曼滤波器来融合双目相机和惯性器件输出。作为一种松组合方式，组合中没有充分使用惯性器件的输出来辅助图像特征点的匹配、跟踪与野值剔除。2007 年，Veth 提出了一种视觉辅助低精度惯性导航的方法[16]。该算法使用了多维随机特征跟踪方法，其最大的缺点是跟踪的特征点个数必须保持不变。同年，Mourikis 提出了基于多状态约束的卡尔曼滤波器(Multi-State Constraint Kalman Filter，MSCKF)算法[17]，其优点是在观测模型中不需要包含特征点的空间位置；但是 MSCKF 算法中存在滤波估计不一致问题：不可观的状态产生错误的可观性，如航向角是不可观的，但 MSCKF 通过扩展卡尔曼滤波(Extended

Kalman Filter,EKF)线性化后会使航向角产生错误的可观性。为了解决滤波估计不一致问题,李明阳等[18]提出了首次估计雅可比 EKF(the First Estimate Jacobian EKF,FEJ-EKF)算法;Huang 等[19]提出了基于可观性约束的无迹卡尔曼滤波(Unscented Kalman Filte, UKF)算法;Castellanos 等[20]提出了 Robocentric Mapping 滤波算法。这些算法均在一定程度上解决了滤波估计不一致问题。

2015 年,Bloesch 等提出了鲁棒视觉惯性里程计(Robust Visual Inertial Odometry,ROVIO)[21],该算法利用 EKF 将视觉信息和惯性测量单元(Inertial Measurement Unit,IMU)信息进行紧耦合,在保持精度的同时降低了计算量。Indelman 等基于 EKF,综合利用了 2 幅图像间的对极约束和 3 幅图像之间的三视图约束融合单目相机和惯性器件[22]。基于相同的观测模型,Hu 等给出了基于 UKF 的实现方法[23]。

近年来,基于优化的算法得到了快速发展。Lupton 和 Sukkarieh 于 2012 年首次提出了利用无初值条件下的惯性积分增量方法来解决高动态条件下的惯性视觉组合导航问题[24]。文中采用了 Sliding Window Forced Independence Smoothing 技术优化求解状态变量。预积分理论的建立,使得基于优化的视觉/惯性组合导航算法得以实现。受此思想启发,Stefan 等采用 Partial Marginalization 技术,通过优化非线性目标函数来估计滑动窗口内关键帧的位姿参数[25]。其中,目标函数分为视觉约束和惯性约束 2 个部分:视觉约束由空间特征点的重投影误差表示,而惯性约束由 IMU 运动学中的误差传播特性表示。该方法不适用于长航时高精度导航,因为没有闭环检测功能,无法修正组合导航系统的累积误差。2017 年,Forster 等完善了计算关键帧之间惯性积分增量的理论,将该理论扩展到 Rotation Group,并分析了其误差传播规律[26]。该算法也未考虑闭环检测问题。同样基于预积分理论,沈劭劼课题组提出了视觉惯性导航(Visual-Inertial Navigation System, VINS)算法[27]。该算法具备自动初始化、在线外参标定、重定位、闭环检测等功能。ORB-SLAM 的设计者 Mur-Artal 等利用预积分理论,将惯性信息引入 ORB-SLAM 框架,设计了具有重定位和闭环检测等功能的视觉/惯性组合导航算法[28]。关于预积分理论,目前还缺乏积分增量合并以及相应的协方差矩阵合并方法。因此,文献[28]去掉了 ORB-SLAM 中的关键帧删除功能。表 1 汇总了基于视觉几何与运动学模型的视觉和惯性导航技术的主要研究成果。

表 1 基于模型的视觉/惯性组合导航技术

Tab. 1 Model based visual-inertial integrated navigation

分类方式		代表成果	主要特点
纯视觉导航	视觉里程计	LIB-VISO[6]	稀疏特征点 通过重投影误差约束优化相对位姿 利用滤波估计相机的绝对位姿
		SVO[7]	结合稀疏与稠密特征 通过 BA 优化相对位姿 有局部地图
		DSO[8]	稠密特征 通过光度误差约束优化相对位姿 后端滑动窗口优化
	视觉同步定位与构图	Mono-SLAM[9]	第 1 个 VSLAM 工作 稀疏特征 基于滤波技术优化位姿和特征空间位置参数
		PTAM[10]	第 1 个基于优化技术的 VSLAM 工作 稀疏特征 前端跟踪和后端优化
		ORB-SLAM[12]	基于稀疏特征点的 VSLAM 最佳代表 在 PTAM 基础上增加了地图管理和闭环优化等功能
		LSD-SLAM[13]	基于稠密光度误差的 VSLAM 最佳代表 具有地图管理和闭环优化功能
视觉惯性组合导航	基于滤波的信息融合	MSCKF[17]	滑动窗口技术 滤波状态只包含窗口内的相机位姿 EKF
		ROVIO[21]	滑动窗口技术 滤波状态包含窗口内的相机位姿和特征参数 IEKF
		SWVIO[23]	基于三视图的三焦张量几何约束 滤波状态只包含三视图的相机位姿 UKF
	基于迭代优化技术的信息融合	OKVIS[25]	基于关键帧优化的 VIO 采用预积分技术 采用 Partial Marginalization 技术
		VINS[27]	基于关键帧优化的 VIO 采用预积分技术 具有地图管理和闭环优化等功能
		VIM-SLAM[28]	在 ORB-SLAM 框架上引入惯性信息 采用预积分技术 具有地图管理和闭环优化等功能

基于模型的视觉/惯性组合导航技术需要信噪比较高的输入数据，算法的整体性能不仅受制于算法的基本原理，还取决于参数的合理性与精确度。相对而言，深度学习神经网络能够通过大数据训练的方式自适应地调节参数，对输入数据具有一定的容错性，因此已有研究人员开发了一系列基于深度学习的视觉/惯性组合导航技术，并已取得一定成果。

使用深度学习神经网络替换传统算法中的个别模块是较为直接的算法设计思路，如利用深度学习神经网络实现里程计前端中的特征点识别与匹配。Detone 等[29]提出了 SuperPoint 算法，该算法首先使用虚拟三维物体的角点作为初始训练集，并将特征点提取网络在此数据集上进行训练；对经过训练的网络在真实场景训练集中进行检测得到自标注点，并将标注有自标注点的真实场景图像进行仿射变化得到匹配的自标注点对，从而得到了最终的训练集；随后使用对称设计的特征点识别网络，将特征提取器读入的原始图像经过多层反卷积层转换为特征点响应图像，响应区域为相邻帧图像匹配特征点的位置。几何对应网络(Geometric Correspondence Network，GCN)[30]则是利用相对位姿标签值构建的几何误差作为匹配特征点空间位置估计值的约束；随后使用多视觉几何模型结合低层特征提取前端网络得到的匹配特征点，求解载体的运动信息。此类低层特征提取前端易于与传统实时定位与建图系统相结合，并且较为轻量，可植入嵌入式平台进行实时解算。

另一种思路是使用深度学习神经网络实现从原始数据到导航参数的整个转化过程。Kendall 团队基于图像识别网络 GoogleNet[31]开发了一种基于单张图像信息的绝对位姿估计网络 PoseNet[32]。首先，搭建绝对位姿回归数据集，配合高精度姿态捕捉设备，为单目相机拍摄的每一帧图像标注绝对位姿标签值；然后使用多层全连接层替换 GoogleNet 的多个 softmax 层，并构成位姿回归器，回归器的输出维度与使用欧拉角表示的位姿维数相同；通过长时间的训练，PoseNet 能较为准确地将训练数据集图像投影为对应位姿标签，然而没有额外的几何约束，网络收敛较为困难。Wang 等在位姿估计网络中引入相邻帧图像信息，构建基于深度学习的单目视觉里程计DeepVO[33]，为了能够同时处理相邻两帧图像的信息，将 FlowNet[34]网络的主体作为视觉特征提取器，并使用输入窗口大于 1 的长短时记忆(Long Short Term Memory，LSTM)网络联合时间轴上相邻多帧图像的高层信息，以此来优化里程计短时间内的估计精度；最后使用全连接层综合图像高层信息，并转化为相邻帧图像的相对位姿估计值。实验结果表明，DeepVO 相对于早期基于模型的视觉里程计 LIBVISO[6]性能具有一定提升，同时与同类型算法[35]相比，也有明显的性能提升。

与基于模型的视觉/惯性组合导航技术类似，为了提高导航算法的自主性与抗干扰能力，研究人员在基于深度学习的视觉导航技术中引入惯导数据，并为其设计单独的网络来提取有用的数据特征。牛津大学的 Clark 团队设计了一种端对端的视觉/惯性组合里程计网络 VINet[36]，使用双向光流提取网络 FlowNet-Corr[34]提取相邻帧图像的高层特征，使用单层全连接层对图像高层特征进行压缩，并使用多节点 LSTM 网络处理两帧图像间的惯性信息；随后将两种数据的高层特征在单维度上进行结合，构成视觉/惯性信息融合特征；最后使用全连接层将融合特征投影至 SE(3)空间中，得到相对位姿估计值。VINet 在道路与无人机数据中都显示出较为优秀的性能，同时为基于深度学习的组合导航技术提供了基础模板。

陈昶昊于 2019 年提出了基于注意力模型的视觉/惯性组合里程计网络 Attention-based VIO[37]，网络的基本框架与 VINet 类似，但视觉特征提取器使用更为轻量的 FlowNetsimple[34]卷积层，以此来提高网络运行效率。借鉴自然语言处理领域的注意力机制，使用 soft attention 和 hard attention 两种注意力网络剔除融合特征中的噪声高层特征，从而加快训练收敛，提高网络性能。表 2 汇总了基于机器学习的视觉/惯性组合导航技术的主要研究成果。

在国内，清华大学、上海交通大学、浙江大学、哈尔滨工程大学、国防科技大学、北京航空航天大学、北京理工大学、南京航空航天大学、西北工业大学、电子科技大学、中国科学院自动化研究所等高校和科研机构的多个研究团队近年来在惯性/视觉组合导航领域开展了系统性的研究工作，取得了诸多研究成果[38-44]。

表 2 基于机器学习的视觉/惯性组合导航技术

Tab. 2 Learning based visual-inertial integrated navigation

分类方式		代表成果	主要特点
纯视觉导航	绝对位姿估计网络	PoseNet[32]	采用 GoogleNet 作为视觉特征提取器，充分挖掘图像的高层特征 使用全连接层构成的位姿回归器，将图像高层特征抽象为位姿信息
	匹配特征提取网络	SuperPoint[29]	构建虚拟图像训练集，在训练过程中提高网络的泛化能力 使用对称网络设计，通过前向传播得到匹配特征点在图像中的位置
	视觉里程计网络	DeepVO[33]	采用 FlowNetSimple 作为特征提取器，同时读取两帧图像的高层特征 采用 LSTM 网络对视觉或者视觉惯性信息融合特征进行优化 使用 6 维度的全连接层将高层特征抽象为相对位姿估计值
视觉惯性组合导航	端对端视觉惯性组合里程计网络	VINet[36]	使用多节点的 LSTM 网络作为惯性信息特征提取器，提高网络解析能力 将视觉/惯性信息的高层特征在通道上进行结合，使用全连接层将特征投影为相对位姿
		Attention-based IO[37]	使用双向 LSTM 网络作为惯性信息特征提取器，提高网络的解析能力 引入两种注意力网络对数据噪声特征进行过滤

2 基于模型的视觉/惯性组合导航技术

基于模型的视觉/惯性组合导航技术的通用结构示意图如图 1 所示。

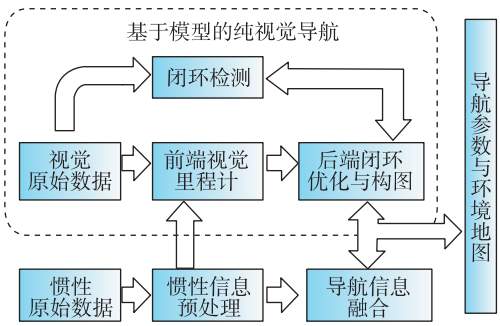

图 1 基于模型的视觉/惯性组合导航技术通用结构示意图

Fig. 1 Scheme of model based visual-inertial navigation technology

2.1 基于模型的纯视觉导航算法

基于模型的视觉导航算法是指以多视图几何等数学模型构建的 VO 和 VSLAM 算法。

(1)视觉里程计原理

载体在运动过程中，可以通过与其固联的摄像机获取图像流。由于载体运动，同一个静止的物体在不同帧图像中的成像位置将发生变化。根据摄像机的成像几何模型，可以利用同一物体在不同帧图像中成像位置的关系，恢复出相机在拍摄图像时的位置和姿态变化量。然后，将相邻帧图像的位置和姿态变化量进行积分，可以推算出摄像机运行的轨迹，如图 2 所示。

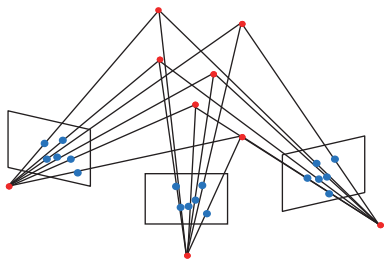

图 2 多视图几何示意图

Fig. 2 Scheme of multi-view geometry

摄像机的成像模型是从多视图中恢复出载体运动参数的基础。常用的相机模型包括透视模型(perspective model)、全景模型(omnidirectional model)和球形模型(spherical model)等。摄像机模型可以通过观察棋盘格或二维码等特征固定且尺度大小已知的物体进行离线标定。

视觉里程计根据特征利用的方式可以分为间接法和直接法两类。间接法通过最小化同一特征在不同图像中的位置投影误差来解算摄像机的运动参数；而直接法则基于光度(灰度)不变假设，通过最小化同一特征在不同图像中的光度误差来估计摄像机的运动参数。

间接法视觉里程计首先需要建立特征匹配关系，然后根据特征匹配对之间的坐标关系，解算出相机的运动参数。设载体在运动过程中，摄像机拍摄了 n 幅图像，表示为 $\boldsymbol{I}_{1:n}=\{\boldsymbol{I}_1,\cdots,\boldsymbol{I}_n\}$；同时，在导航环境中有 m 个特征，特征的空间坐标为 $\boldsymbol{p}_{1:m}=\{\boldsymbol{p}_1,\cdots,\boldsymbol{p}_m\}$；第 j 个特征在 k 时刻拍摄图像中的坐标为 $\boldsymbol{z}_{k,j}=\pi_k(\boldsymbol{p}_j)$，其中 π_k 表示相机在 k 时刻的投影模型，其与相机的成像模型和相机的位姿相关。

首先，通过特征匹配算法建立特征之间的对应关系 $\{\boldsymbol{z}_{k,j}\leftrightarrow\boldsymbol{z}_{k+1,j}\}$，间接法视觉里程的运动估计可以表示为最小化如下误差函数的过程

$$\underset{\boldsymbol{R}_{k,k+1},\boldsymbol{t}_{k+1}}{\operatorname{argmin}} \sum_j \left\| \boldsymbol{z}_{k,j} - \hat{\boldsymbol{z}}_{k,j} \right\| + \left\| \boldsymbol{z}_{k+1,j} - \hat{\boldsymbol{z}}_{k+1,j} \right\| \quad (1)$$

其中，$\boldsymbol{t}_{k,k+1}$ 为待估计的运动参数，表示摄像机在 $k+1$ 时刻相对于 k 时刻的平移量；$\boldsymbol{R}_{k,k+1}$ 也为待估计的运动参数，表示摄像机在 $k+1$ 时刻相对于 k 时刻的旋转矩阵；$\hat{\boldsymbol{z}}_{k,j}$ 和 $\hat{\boldsymbol{z}}_{k+1,j}$ 分别为将估计的特征位置 $\hat{\boldsymbol{p}}_j$ 投影到像平面形成的虚拟像点。式(1)也被称为重投影误差，重投影误差的几何表示如图 3 所示。广泛使用的间接法视觉里程计有 LIBVISO[6]。

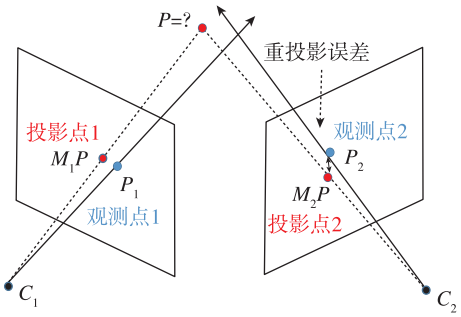

图 3　重投影误差示意图

Fig. 3　Scheme of reprojection error

与间接法不同，直接法视觉里程计则通过最小化光度误差估计摄像机的运动参数。通常，同一特征在短时间内拍摄的多幅图像中，其光度基本不变，并且摄像机在短时间内的位姿变化较小，同一特征在相邻帧图像中的成像位置变化不大。据此，直接法视觉里程计通过迭代优化算法在状态空间中进行搜索，使得同一特征在不同图像中的像点光度误差最小，从而解算得到摄像机运动参数，具体优化目标函数为

$$\underset{\boldsymbol{R}_{k,k+1},\boldsymbol{t}_{k+1}}{\arg\min} \sum_j \boldsymbol{I}_k(\boldsymbol{p}_j) - \boldsymbol{I}_{k+1}(\boldsymbol{p}_j) \quad (2)$$

其中，$\boldsymbol{I}_k(\boldsymbol{p}_j)$ 和 $\boldsymbol{I}_{k+1}(\boldsymbol{p}_j)$ 分别表示同一特征在相邻帧图像中的光度。若直接法里程计在运动估计过程中使用了整幅图像的像点光度，则为稠密视觉里程计算法；若仅使用部分像点光度，则为稀疏视觉里程计算法。由慕尼黑工业大学开发的 DSO 算法[8]就是一种稀疏直接法视觉里程计。

除直接法与间接法里程计外，Forster 等还提出了一种半直接法视觉里程计[7]。在 SVO 中使用了直接法进行运动解算，同时采用了间接法来估计特征的三维坐标，建立局部地图。

需要注意的是，在视觉里程计算法中，特征的三维坐标 $\boldsymbol{p}_j$ 通常是未知的，需要采用立体摄相机(如双目摄像机或 RGB-D 摄像机)测量得到，或采用三角测量算法从单目相机的多视图观测中估计出特征的三维坐标 $\hat{\boldsymbol{p}}_j$。因此，视觉里程计也具有局部建图功能。与 VSLAM 算法相比，VO 仅使用最新的若干帧图像进行运动估计和建图，以降低状态维度，提升算法效率。

(2)闭环优化与构图

视觉里程计是一种路径积分方法，因此具有累积误差。闭环优化是广泛使用的一种用于修正视觉里程计累积误差的方法。闭环修正依赖于构建的环境地图，其基本原理是：载体在移动过程中，将观测的视觉特征与地图中的视觉特征进行匹配，并通过匹配关系解算出载体在地图中所处的位置和姿态。由于建图误差和视觉里程计累积误差的影响，通过里程计估计的摄像机位姿与通过闭环检测估计的摄像机位姿之间具有差异，通过建立数学模型可以同时对里程计累积误差和建图误差进行修正。闭环优化与构图可以描述为一个最大后验概率(Maximum A Posteriori, MAP)问题，具体表达式为

$$\boldsymbol{X}^*, \boldsymbol{L}^* = \arg\max_{\boldsymbol{X},\boldsymbol{L}} p(\boldsymbol{X}, \boldsymbol{L} \mid \boldsymbol{Z}, \boldsymbol{U}) \quad (3)$$

其中，$\boldsymbol{X}$ 表示摄像机在整个运动过程中的位置和姿态构成的状态向量；$\boldsymbol{L}$ 表示所有特征在参考系下的位置向量的集合；$\boldsymbol{Z}$ 表示特征在摄像机图像中的成像点位置的集合；$\boldsymbol{U}$ 表示里程计测量的运动参数。在大范围的导航应用中，式(3)中包含的状态量较多，因此需要对优化算法进行合理设计才能满足算法的实时性需求。目前，广泛使用的建图与闭环优化工具有 G2O[45]、GTSAM[46]和 Ceres[47]等。

2.2　基于模型的视觉/惯性组合导航算法

基于滤波技术和基于非线性迭代优化技术是视觉/惯性信息融合的两种典型方式。

(1)基于滤波技术的信息融合算法

基于滤波技术的信息融合算法主要考虑以下 3 个方面的问题：滤波器状态变量的选取、状态方程和观测方程的建立以及滤波算法的选取。

首先是滤波器状态变量的选取，常见的方式是将当前时刻的惯性导航参数、邻近 n 帧图像对应时刻的载体位姿参数以及这些图像所观测到的特征的空间位置参数加入到状态变量中。当前时刻惯性导航参数通常包含 IMU 的位置、姿态、速度和陀螺、加速度计的零偏等，其定义如下

$$\boldsymbol{x}_{\mathrm{IMU}}(t) = \left[(\boldsymbol{p}^{\mathrm{W}})^{\mathrm{T}} \quad (\bar{\boldsymbol{q}}_{WI})^{\mathrm{T}} \quad (\boldsymbol{v}^{\mathrm{W}})^{\mathrm{T}} \quad (\boldsymbol{b}_{\mathrm{g}})^{\mathrm{T}} \quad (\boldsymbol{b}_{\mathrm{a}})^{\mathrm{T}} \right]^{\mathrm{T}} \quad (4)$$

其中，$\boldsymbol{p}^{\mathrm{W}}$ 表示 IMU 在世界系下的位置；四元数 $\bar{\boldsymbol{q}}_{WI}$ 表示 IMU 坐标系 $\{I\}$ 在世界系中的姿态，与旋转矩阵 $\boldsymbol{R}_{WI}$ 对应；$\boldsymbol{v}^{\mathrm{W}}$ 表示 IMU 在世界系下的速度；$\boldsymbol{b}_{\mathrm{g}}$ 和 $\boldsymbol{b}_{\mathrm{a}}$ 分别为陀螺和加速度计的零偏。组合导航系统的状态变量通常具有如下形式

$$\boldsymbol{x}(t)=[\boldsymbol{x}_{\mathrm{IMU}}(t)^{\mathrm{T}}\quad \boldsymbol{x}_{\mathrm{cam}}^{1\ \mathrm{T}}\quad \cdots\quad \boldsymbol{x}_{\mathrm{cam}}^{n\ \mathrm{T}}\quad \boldsymbol{x}_{\mathrm{fea}}^{1\ \mathrm{T}}\quad \cdots\quad \boldsymbol{x}_{\mathrm{fea}}^{m\ \mathrm{T}}]^{\mathrm{T}} \tag{5}$$

其中，$\boldsymbol{x}_{\mathrm{cam}}^{n}$ 表示第 n 帧图像对应的相机位姿参数，$\boldsymbol{x}_{\mathrm{fea}}^{m}$ 表示第 m 个特征点对应的空间位置参数。不同算法中 n 的数值不尽相同。例如文献[15]中 n 为 1，文献[41]中 n 为 4，MSCKF 算法[17]中 n 为滑动窗口中的图像数目。特征点的个数 m 主要由检测的图像特征决定，但 MSCKF 算法中 m 的值为 0。

其次是状态方程和观测方程的建立。由于通常假设场景是固定的，即特征点的空间位置变化率为 0，因此系统的状态方程只与载体的运动参数有关。

典型的系统状态微分方程如式(6)所示

$$\dot{\boldsymbol{R}}_{WI}=\boldsymbol{R}_{WI}\hat{\boldsymbol{\omega}}_{WI}^{I},\ \dot{\boldsymbol{v}}^{\mathrm{W}}=\boldsymbol{a}^{\mathrm{W}},\ \dot{\boldsymbol{p}}^{\mathrm{W}}=\boldsymbol{v}^{\mathrm{W}},$$
$$\dot{\boldsymbol{b}}_{\mathrm{g}}=\boldsymbol{n}_{\mathrm{g}},\ \dot{\boldsymbol{b}}_{\mathrm{a}}=\boldsymbol{n}_{\mathrm{a}} \tag{6}$$

其中，$\hat{\boldsymbol{\omega}}_{WI}^{I}$ 为转动角速度，$\boldsymbol{a}^{\mathrm{W}}$ 为加速度，$\boldsymbol{n}_{\mathrm{g}}$ 和 $\boldsymbol{n}_{\mathrm{a}}$ 为陀螺和加速度计的测量噪声。系统的观测方程与具体采用的约束相关，是滤波算法设计的核心。如 2.1 节所述，通常采用的约束包括重投影误差约束[17,40]、两视图对极几何约束以及三视图三焦张量约束[22,41]等，而光度残差约束在基于滤波技术的视觉/惯性组合导航算法中比较罕见。

关于滤波器的选取，最常见的有 EKF[17-18,20-22] 和 UKF[19,23,40-41,43]，二者都是在卡尔曼滤波器(Kalman Filter，KF)的基础上发展起来的。EKF 通过偏导数得到雅可比矩阵，将状态方程和观测方程线性化，从而解决视觉/惯性融合中的非线性问题。为了克服 EKF 中高阶导数省略问题和雅克比矩阵计算难的问题，UKF 按一定间隔和概率在状态空间中选取采样点(sigma points)的方式，代入状态方程和观测方程，预测和更新状态值及其对应的协方差矩阵。

由于计算量的限制，一般不会将全局地图中的特征空间位置参数加入滤波器状态变量中，因此基于滤波技术的视觉/惯性组合导航算法通常无法构建全局地图，不支持闭环检测与优化。

(2)基于优化技术的信息融合算法

为了实现迭代优化算法框架下的视觉/惯性导航信息融合，必须解决惯性约束和视觉约束的统一表示问题。对于视觉信息，关键帧之间的位置和姿态约束可以通过它们共同观测的图像特征之间的匹配关系来确立。而对于惯性信息，2 个时刻间的位置和姿态约束可以通过 2 个时刻间的陀螺和加速计测量信息来建立。在视觉/惯性组合导航系统中，当前时刻的关键帧位姿参数是在前一时刻关键帧的位姿参数基础上，利用陀螺和加速度计测量值递推得到。由于关键帧的位姿参数属于迭代优化的状态变量，在优化过程中，每一次迭代都会改变，所以由前一时刻关键帧的位姿参数递推得到的当前帧的位姿参数，需要重新利用两帧之间的陀螺和加速度计测量值推算，处理效率非常低。为了避免该问题，需要设计一种不依赖于积分初值的惯性积分增量计算方法，使得在迭代优化过程中，前一时刻关键帧位姿参数变化之后，可以根据积分增量快速更新当前时刻的关键帧位姿参数。

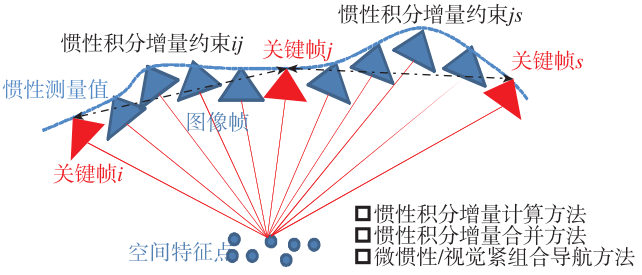

图 4 基于迭代优化技术的视觉/惯性组合导航示意图

Fig. 4 Scheme of visual-inertial integrated navigation based on iterative optimization

惯性预积分技术应运而生[24]，其核心思想是定义位置、姿态和速度积分增量，使得积分增量与积分初值无关。从系统的运动学模型式(6)出发，可以得到关键帧 $[t_i,t_j]$ 时刻间的位姿参数与惯性测量值之间的关系为

$$\boldsymbol{R}_j=\boldsymbol{R}_i\prod_{k=i}^{j-1}\mathrm{Exp}(\boldsymbol{\omega}_k\Delta t)$$
$$\boldsymbol{v}_j=\boldsymbol{v}_i+\boldsymbol{g}n\Delta t+\sum_{k=i}^{j-1}\boldsymbol{R}_k(\boldsymbol{a}_k-\boldsymbol{b}_{ak}-\boldsymbol{\eta}_k)\Delta t$$
$$\boldsymbol{p}_j=\boldsymbol{p}_i+\boldsymbol{v}_i n\Delta t+\frac{\boldsymbol{g}(n\Delta t)^2}{2}+\Delta t^2\sum_{k=i}^{j-1}\left\{\sum_{m=i}^{k-1}\boldsymbol{R}_m(\boldsymbol{a}_m-\boldsymbol{b}_{am}-\boldsymbol{\eta}_m)+\frac{\boldsymbol{R}_k(\boldsymbol{a}_k-\boldsymbol{b}_{ak}-\boldsymbol{\eta}_k)}{2}\right\} \tag{7}$$

其中，$\boldsymbol{g}$ 是重力矢量，$\boldsymbol{\eta}$ 是加速计测量噪声，n 是积分时段内惯性传感器的采样个数。从式(7)可以看出，t_j 时刻关键帧的位姿参数与 t_i 时刻关键帧的位姿参数以及 $[t_i,t_j]$ 时刻间的惯性测量值有

关。为了消除 t_i 时刻关键帧的位姿参数的影响，定义 t_i 和 t_j 时刻关键帧之间的状态变量增量计算公式如下

$$\begin{aligned}\Delta\boldsymbol{R}_{ij} &= \boldsymbol{R}_i^{\mathrm{T}}\boldsymbol{R}_j = \prod_{k=i}^{j-1}\mathrm{Exp}(\boldsymbol{\omega}_k\Delta t)\\ \Delta\boldsymbol{v}_{ij} &= \boldsymbol{R}_i^{\mathrm{T}}(\boldsymbol{v}_j - \boldsymbol{v}_i - \boldsymbol{g}\Delta t_{ij})\\ &= \sum_{k=i}^{j-1}\boldsymbol{R}_{ik}(\boldsymbol{a}_k - \boldsymbol{b}_{ak} - \boldsymbol{\eta}_k)\Delta t\\ \Delta\boldsymbol{p}_{ij} &= \boldsymbol{R}_i^{\mathrm{T}}(\boldsymbol{p}_j - \boldsymbol{p}_i - \boldsymbol{v}_i\Delta t_{ij} - \frac{\boldsymbol{g}\Delta t_{ij}^2}{2})\\ &= t^2\sum_{k=i}^{j-1}\left\{\sum_{m=i}^{k-1}\boldsymbol{R}_{im}(\boldsymbol{a}_m - \boldsymbol{b}_{am} - \boldsymbol{\eta}_m) + \frac{\boldsymbol{R}_{ik}(\boldsymbol{a}_k - \boldsymbol{b}_{ak} - \boldsymbol{\eta}_k)}{2}\right\}\end{aligned} \tag{8}$$

式中，$\Delta t_{ij} = t_j - t_i$。从式(8)可以看出，惯性积分增量 $\Delta\boldsymbol{R}_{ij}$、$\Delta\boldsymbol{v}_{ij}$、$\Delta\boldsymbol{p}_{ij}$ 仅与$[t_i, t_j]$时刻间的陀螺测量值 $\boldsymbol{\omega}$ 和加速度计测量值 $\boldsymbol{a}$ 有关，与积分的初值 $\boldsymbol{R}_i$、$\boldsymbol{v}_i$ 和 $\boldsymbol{p}_i$ 无关。

通过预积分对惯性信息进行预处理之后，就可以建立统一视觉约束和惯性约束的优化目标函数。以 VINS 为例，其目标函数具有如下形式[27]

$$\min_{\boldsymbol{X}}\{\|\boldsymbol{r}_p - \boldsymbol{H}_p\boldsymbol{X}\|^2 + \sum_{k\in B}\|\boldsymbol{r}_B(\hat{\boldsymbol{z}}_{b_{k+1}}^{b_k}, \boldsymbol{X})\|^2_{\boldsymbol{P}_{b_{k+1}}^{b_k}} + \sum_{(i,j)\in C}\|\boldsymbol{r}_C(\hat{\boldsymbol{z}}_l^{c_j}, \boldsymbol{X})\|^2_{\boldsymbol{P}_l^{c_j}}\} \tag{9}$$

其中，3 个残差项依次是边缘化的先验信息、IMU 测量残差以及视觉的观测残差，$\boldsymbol{X}$ 是待优化的状态向量，包含关键帧的相机位姿、特征的空间位置、惯性器件的零偏等。

当然，一个完整的视觉/惯性组合导航系统还包含系统初始化、闭环修正与优化等。此处不再赘述，感兴趣的读者可以查阅文献[25-28]。

3　基于机器学习的视觉/惯性组合导航技术

深度学习神经网络是机器学习概念的重要分支，具有参数学习与非线性模型拟合的能力，利用深度学习解决组合导航问题，实质上是使用神经网络对原始数据与导航参数之间的关系进行建模，并通过长时间训练来优化模型的参数。为了增强深度学习网络的可解释性，需对网络不同功能模块使用不同种类的网络进行建模。图 5 所示为基于深度学习的视觉/惯性组合导航技术的通用结构示意图。

3.1　前端网络

(1)视觉特征提取器

与基于模型的组合导航技术类似，基于深度学习的导航技术也存在前端，即处理原始数据的模块。针对图像这种高维度的信息，需从中捕获高层特征来解析相机运动信息。

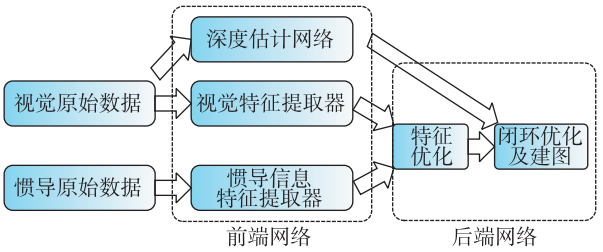

图 5　基于机器学习的视觉/惯性组合导航技术通用结构示意图

Fig. 5　Scheme of learning based visual-inertial navigation technology

根据多视觉几何原理，相邻帧图像隐含移动载体的运动信息即导航参数，并且原始图像与导航参数之间存在非线性关系。因此可以使用多层神经网络提取原始图像数据的高层特征，随后使用神经网络回归器直接将高层特征映射为导航参数。如式(10)所示，构建深度学习神经网络来拟合图像与导航参数之间的非线性模型，其中 f_{enc} 代表原始图像数据的特征提取器，它通过多层卷积神经网络提取图像 I_c 数据中的高层特征 a，并通过回归器将其投影至标签空间变为导航参数估计值 $\hat{y}$，随后结合导航参数标签值 y 构建训练误差 L，以此来优化特征提取器与回归器网络的参数。

$$\begin{aligned}a &= f_{\mathrm{enc}}(I_c)\\ \hat{y} &= f_{\mathrm{reg}}(a)\\ L &= \|y - \hat{y}\|\end{aligned} \tag{10}$$

文献[33, 37]使用单输入的光流估计网络(FlowNetSimple)[34]的卷积层部分搭建视觉特征提取器，并将网络的输入层通道数设置为 6，接收时间轴上相邻两帧的 RGB 图像。为了能对相邻图像的高层信息进行更充分的解析，文献[36]使用双输入的光流估计网络(FlowNetCorr)[34]的卷积层部分搭建视觉特征提取器，为前后两帧图像分别构建卷积神经网络，解析 2 张图像中的高层特征，并使用 correlate 操作融合两帧图像的高层特征。FlowNetCorr 的层数较多，训练成本较大，因此在基于深度学习的视觉里程计中一般选用 FlowNetSimple 的卷积层部分搭建视觉特征提取器。上述两种视觉特征提取器依据成熟的卷积神经网络进行设计，同时 Dosovitskiy 等[34]已公开这两种卷积神经网络的预训练参数，有

利于开发基于视觉信息的深度学习导航技术。然而 FlowNetCorr 与 FlowNetSimple 都属于层数较多的卷积神经网络，参数量较大，其中 FlowNetCorr 参数占磁盘空间 149M，FlowNetSimple 占 148M，因此这两种卷积神经网络不适用于包含深度信息的全导航参数估计算法。针对此问题，文献[49-50]设计了仅由 6 层卷积核构成的视觉特征提取器，并且使用均值池化操作将视觉高层特征直接压缩为 6 维度的相邻图像帧位姿；但较少的层数也导致提取器的解析能力较弱，在深度以及位姿估计任务中的性能也有一定局限性，训练收敛速度较慢。

(2)惯性信息特征提取器

基于紧耦合优化的视觉/惯性组合里程计前端一般采用预积分的方式处理惯性信息[26-27]，从而提高非线性优化线程的效率。然而预积分数值中存在由零偏和噪声产生的误差，而且这些误差会随着时间而积累。陈昶昊等[51]提出了一种基于双向 LSTM 网络的惯性里程计，借鉴零速检测原理，物体运动的绝对速度与加速度振动模态有关，通过一段时间内的 IMU 线加速度与角速度测量值估计物体的速度，从而抑制位姿解算误差随时间积累。上述过程如式(11)所示，其中 f_{enc} 代表双向 LSTM 网络，$I_{\text{imu}}^{i\cdots i+n}$ 代表第 i 与第 $i+n$ 帧之间的 n 帧 IMU 测量值，包括线加速度与角速度。引出 f_{enc} 最后时刻的输出作为惯导数据高层特征 a_{imu}^{i+n}，并通过回归器 f_{reg} 得到第 i 与第 $i+n$ 帧之间的相对位姿 T_i^{i+n}。

$$\begin{aligned} a_{\text{imu}}^{i+n} &= f_{\text{enc}}(I_{\text{imu}}^{i\cdots i+n}) \\ T_i^{i+n} &= f_{\text{reg}}(a_{\text{imu}}^{i+n}) \end{aligned} \tag{11}$$

文献[36-37]使用惯性信息与图像的融合特征进行姿态解算，实验结果表明，添加惯性信息的里程计网络收敛较快并且测试精度较高。为了进一步提高里程计的解算精度，文献[37]设计了两种注意力网络，注意力网络输出与原始数据高层特征同尺寸的权重掩膜，并通过改变特征元素的相对大小，从而调整网络的训练方向，规避噪声特征对网络性能的影响。文献[37]的实验结果表明，基于深度学习的惯导信息特征提取器在多种惯导信息噪声的环境下也具有较为稳定的性能。然而由于深度学习神经网络参数对于训练数据具有一定的依赖性，对于不同场景数据的泛化能力较差，这限制了基于深度学习的特征提取器的应用范围。文献[52]使用迁移学习的方法，找到不同场景中惯导数据的共有特征并结合其物理模型，在没有标签数据的情况下，利用低精度的手持设备数据也能得到精度较高的位姿解算结果。

原始数据的高层特征，需使用位姿回归器将高层特征投影至标签空间中。常见的位姿回归器由多层全连接层组成，全连接层的输出通道数与位姿估计值的形式有关。现阶段基于深度学习的视觉里程计都采用欧拉角来表示姿态，因此一般将位姿回归器中最后一层的全连接层设置为 6[33,37]，也可以将位姿回归的过程解耦，分别设置 3 维度的位置回归器与姿态回归器。

(3)深度估计网络

除了提取原始数据的特征以外，前端还需给出当前视角内特征点的深度信息。基于模型的视觉/惯性组合导航技术使用多视觉几何模型联合相邻帧图像的匹配特征点，求得相对位姿以及无尺度的特征点深度值。然而，在纹理缺失以及光线较暗的部分，特征点识别算法失效导致无法得到较为准确的深度值。深度学习神经网络通过前向传播直接得到原始图像像素点对应的深度值，同时设计具有几何约束的误差项来校正神经网络参数，从而提高深度估计的精度。文献[50,55-56]构建了类 U-NET 的深度估计网络，使用多层卷积神经网络构建特征提取器，其中文献[54-55]使用主流的 PackNet 和 ResNet 网络作为特征提取器，在训练前使用预训练参数进行初始化，便于训练的收敛；随后使用深度解码器将特征提取器解析的图像高层特征变为与原图尺寸一致的深度估计值，深度解码器由多层反卷积层构成，同时将特征提取器输出图像的不同层次的特征输入到深度解码器对应的反卷积层中，强化深度估计图像的多尺度细节。为了提高深度估计网络的性能，现有两种思路：1)改进网络的结构，例如将网络的高低特征联结[50,55-56]，增强输出的深度图像质量；2)在设计误差函数时添加约束条件，例如文献[55]引入时空最小误差，剔除在连续两帧图像中因相机旋转而移出视场范围的像素点，避免了在计算重投影误差时出现局部异常极大值的现象；文献[56]则在总误差中引入极线误差，使得网络能够充分利用相邻帧的点线特征，从而增强网络性能。

3.2 特征优化

特征优化环节对应基于模型的视觉/惯性组合导航技术中的非线性优化模块，该模块利用前端提取的低层特征以及里程计估计的位姿参数构建几

何误差函数，使用特定的非线性优化算法降低误差函数值，以此得到优化的导航参数。同样地，特征优化环节也设计了特定的网络来优化前端网络得到的数据特征或者导航参数估计值。

借鉴传统 SLAM 窗口优化的思想，文献[33，36-37]在视觉特征提取器的最后一层卷积层中添加 LSTM 网络，以综合前后多帧原始数据的高层特征，优化当前时刻的高层特征。上述过程如式(12)所示，其中 f_{lstm} 的每一时刻都引出隐藏变量，使得经优化的特征与未优化特征的尺寸保持一致。

$$\tilde{a}_{1\cdots m}=f_{\text{lstm}}(a_{1\cdots n}),\quad m=n \tag{12}$$

同时 LSTM 网络采用多层次级联设计，并添加多个节点以增加网络的解析能力。然而，此类算法属于端对端优化算法，不具有可解释性。为了能在优化原始数据高层特征的过程中考虑到几何模型的因素，文献[49-50，53-55]在总误差中设计了重投影误差，耦合了深度估计网络与位姿估计网络参数的优化过程。然而，以上工作都仅将重投影模型体现在总误差函数中，没有构建显示的网络结构对重投影模型进行求解，网络设计依旧欠缺一定的可解释性，因此很难确定网络是否拟合出了图像像素值、深度与相对位姿之间存在的重投影模型；同时从以上文献的算法性能验证实验可以看出，以上算法相对于端对端的里程计或者深度估计网络的性能并没有显著的提升，这从另一个侧面说明了以上算法在构建网络时并没有充分利用重投影模型原理。鉴于此，Tang 等[57]构建了可微重投影约束层(BA-Layer)，对重投影模型的每个参数进行显示建模，从而对输出的导航参数进行优化。分别设计了基础深度生成网络以及多尺度特征提取网络，将时间上相邻的一组图像代入基础深度生成网络得到每一帧图像的深度图像族，并使用与深度图像族对应的可微系数，将深度图像族加权组合为深度估计值图像；同时使用多尺度特征提取网络得到图像帧的高层特征，随后构建特征级的重投影误差，并代入 BA-Layer 层中进行优化。BA-Layer 层根据前一时刻的状态优化量计算雅克比矩阵、正规方程、阻尼系数以及海森矩阵，进而得到状态量的变化量，从而得到当前时刻的状态优化量。为了确保 BA-Layer 层的可微性，固定了特征级重投影误差的优化步数，同时使用多层全连接层将特征级重投影误差转化为阻尼系数。从实验结果来看，相比于使用光度重投影误差与几何重投影误差的位姿估计方法，该文设计的相对位姿估计网络的旋转角与平移矢量测试精度更高。首先，这说明 BA-Layer 能对重投影误差进行有效建模。其次，文献提到使用几何重投影误差的位姿估计方法在室内环境中可能无法进行有效的特征匹配，光度重投影误差则会增加优化函数的非凸性，导致优化算法对初值设置较为敏感。相比较而言，BA-Layer 使用经卷积神经网络解析的高层特征进行导航参数的求解，相比于特征点、光流等底层特征，高层特征具有较高的稳定性，因此算法的鲁棒性较好。此外，卷积神经网络具有较强的非线性拟合能力，可以在训练过程中对状态初值进行隐式估计，不需要人为指定。

Chen 等[58]则提出了一种基于深度学习的卡尔曼滤波算法 DynaNet。该算法首先假设视觉/惯性组合里程计是一个马尔科夫过程，即当前时刻的状态量与前一时刻的状态量有关，并且能用线性模型来描述状态传递过程。DynaNet 算法使用 LSTM 网络估计状态传递矩阵以及协方差传递误差，并使用卷积神经网络得到视觉/惯性原始数据的高层特征以及测量误差；随后构建卡尔曼滤波方程，经过迭代得到当前时刻的状态量估计值；最后结合状态量的标签值构建训练误差，经过多轮训练得到精度更高的状态量估计值。相比于 Tang 等的工作，DynaNet 使用深度学习神经网络重构线性卡尔曼滤波方程，但鉴于深度学习具有强大的非线性拟合能力，DynaNet 的状态传递矩阵估计网络也能对位姿求解过程进行建模。从实验结果来看，DynaNet 的位姿解算精度高于基于模型的 ORB-SLAM[12]以及基于深度学习的 VO-Feat[50]，这证明了经过精心设计的深度学习网络具有超越基于模型的导航算法的能力；同时也说明了使用深度学习重构传统卡尔曼滤波模型能有效提升深度学习框架求解位姿问题的能力。

3.3　闭环优化与建图

基于深度学习的里程计虽设计了非线性优化模块，然而无法消除长时间的积累误差，因此需借鉴传统视觉/惯性组合里程计算法的闭环优化思想，在整体算法中设置闭环优化环节。Li 等[59]将优化窗口内的第一帧与最后一帧视为一对短闭环节点，并联合其对应的图像构建重投影误差。在得到全局的绝对位姿之后，使用基于 DBoW2[60]开发的场景识别算法检测全局图像帧中的闭环节点，并使用相对位姿估计网络估计闭环节点的位姿代入位

姿节点图中，最后使用 G2O 工具包[45]对位姿节点图进行优化得到经过校正的绝对位姿。上述过程如式(13)所示，其中 r 代表相对位姿估计值误差汇总，e_{ij} 代表相对位姿重定位值 $\hat{T}_{ij}$ 与推算值 $T_i^{-1}T_j$ 之间的误差。

$$r(T_1,T_2\cdots T_m)=\sum_{i,j\in\varepsilon}e_{ij}^{\mathrm{T}}e_{ij} \tag{13}$$

在得到经过优化的绝对位姿之后，需结合关键帧的深度信息构建全局的三维立体模型，然而基于重投影误差估计的关键帧深度值不具有全局一致的尺度，因此还需设计更多的几何约束使得网络在长时间的训练过程中逐渐恢复关键帧的尺度。Guizilini 等[54]提出了在训练误差函数中添加训练数据集中的速度标签，使得相对位姿估计网络输出的相对平移量具有与标签值一致的尺度。Bian 等[61]则使用深度估计网络同时估计参考帧与目标帧的深度，随后使用匹配特征点对应的空间点坐标构建投影误差。

4 视觉/惯性组合导航技术的典型应用及发展趋势

视觉/惯性组合导航技术是机器人、计算机视觉、导航等领域的研究热点，在国民经济和国防建设中取得了广泛的应用，但也面临着诸多挑战。

4.1 视觉/惯性组合导航技术的典型应用

国民经济领域，在无人机、无人车、机器人、现实增强、高精度地图等应用的推动下，视觉/惯性组合导航技术取得了快速发展。例如 Google 的 Tango 项目和无人车项目、微软的 Hololens 项目、苹果的 ARKit 项目、百度无人车项目、大疆无人机项目、高德高精度地图项目等大型应用项目都成立了视觉/惯性组合导航技术相关的研究小组，极大地促进了视觉/惯性组合导航技术在国民经济中的应用。以 Google 的 Tango 项目为例，其导航定位核心算法是基于滤波框架的 MSCKF 算法；微软的 Hololens 项目则是以 KinectFusion 为基础的 SLAM 算法。

国防建设领域，由于视觉/惯性组合导航技术不依赖外部人造实施，在卫星拒止环境中有着重要的应用价值。例如美国陆军研发的一种新型联合精确空投系统采用惯性/视觉组合导航技术解决高精度定位问题。嫦娥三号巡视器也采用视觉与惯性组合实现定姿定位。李丰阳等[62]总结了视觉/惯性组合导航技术在地面、空中、水下和深空等多种场景中的应用。

4.2 视觉/惯性组合导航技术的未来发展趋势

视觉/惯性组合导航技术取得了广泛的应用，但在复杂条件下的可靠性还有待加强，其未来的发展主要体现在以下 4 个方向：

1)提升信息源的质量。首先是提升惯性器件(特别是基于微机电系统(Micro-Electro-Mechanical System，MEMS)工艺的微惯性器件)的零偏稳定性和环境适应性等性能指标；其次是提升视觉传感器的光照动态适应性、快速运动适应性等性能指标；此外，还可以引入更多的传感器，如磁传感器、超声波传感器、激光雷达等，提升复杂条件下组合导航系统的综合性能。

2)提升信息融合算法的水平。视觉和惯性信息各有特点，不同条件下信息的质量也不尽相同，需要设计智能的信息融合机制。目前的算法大多是基于静态场景假设，但在实际应用中，场景都有一定的动态性，动态环境下的视觉/惯性组合导航是提升复杂条件下导航可靠性的重要研究方向。此外，目前基于滤波的信息融合算法仍然存在滤波状态发散或者状态收敛到错误值的情况，需要对系统的可观性进行分析，提升状态估计的一致性。对于优化框架的信息融合算法，目前的预积分理论还有待完善，特别是在 SLAM 的地图管理中删除关键帧时，与关键帧相关的积分增量及对应的协方差需要合并，目前还缺乏协方差合并方法；而且基于 BA 的优化算法计算量较大，对于大尺度的闭环优化，计算耗时太久，存在错失闭环优化的情况，急需提升 BA 算法的效率。

3)发展新的导航理论。大自然中许多动物具有惊人的导航本领，例如：北极燕鸥每年往返于相距数万 km 远的南北两极地区；信鸽能够在距离饲养巢穴数百 km 远的地方顺利返回巢穴。模仿和借鉴动物导航本领的仿生导航技术逐渐成为了导航领域研究的热点。胡小平等[63]对仿生导航技术进行了全面的总结。此外，随着多平台集群应用的普及，利用组网编队中平台间导航信息交互来提升位置、速度、姿态等参数估计精度的协同导航技术方兴未艾。谢启龙等[64]从无人机、机器人、无人水下潜航器、导弹 4 个应用层面梳理了协同导航技术的国内外发展现状。

4)扩充导航系统的功能。目前的视觉/惯性组合导航侧重于导航参数的估计，对于引导和控制等

关注较少。随着机器学习技术在视觉/惯性组合导航领域的应用，可以将机器学习在环境理解、避障检测、引导控制等方面的成果融入到导航系统中。

参考文献

[1] 孙永全，田红丽．视觉惯性 SLAM 综述[J]．计算机应用研究，2018，36(12)：1-6.
Sun Yongquan，Tian Hongli. Overview of visual inertial SLAM[J]. Application Research of Computers，2018，36(12)：1-6(in Chinese).

[2] Huang G. Visual-inertial navigation：a concise review[C]// Proceedings of International Conference on Robotics and Automation (ICRA)，2019：9572-9582.

[3] Huang B，Zhao J，Liu J，et al. A survey of simultaneous localization and mapping with an envision in 6G wireless networks[J]. arXiv：1909.05214v3，2019.

[4] Scaramuzza D，Fraundorfer F. Visual odometry：part I—the first 30 years and fundamentals[J]. IEEE Robotics and Automation Magazine，2011，18(4)：80-92.

[5] Fraundorfer F，Scaramuzza D. Visual odometry：part II—matching，robustness，and applications[J]. IEEE Robotics and Automation Magazine，2012，19(2)：78-90.

[6] Kitt B，Geiger A，Lategahn H，et al. Visual odo-metry based on stereo image sequences with RANSAC-based outlier rejection scheme[C]// Proceedings of IEEE Intelligent Vehicles Symposium，2010：486-492.

[7] Forster C，Pizzoli M，Scaramuzza D，et al. SVO：fast semi-direct monocular visual odometry[C]// Proceedings of International Conference on Robotics and Automation (ICRA)，2014：15-22.

[8] Engel J，Koltun V，Cremers D，et al. Direct Sparse Odometry[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2018，40(3)：611-625.

[9] 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.

[10] Klein G，Murray D W. Parallel tracking and mapping for small AR workspaces[C]// Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR)，2007：1-10.

[11] Strasdat H，Montiel J M，Davison A J，et al. Real-time monocular SLAM：why filter? [C]// Proceedings of International Conference on Robotics and Automations (ICRA)，2010：2657-2664.

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

[13] Engel J，Schops T，Cremers D，et al. LSD-SLAM：large-scale direct monocular SLAM[C]// Proceedings of European Conference on Computer Vision (ECCV)，2014：834-849.

[14] Carrillo L R，Lopez A E，Lozano R，et al. Combiningstereo vision and inertial navigation system for a quad-rotor UAV[J]. Journal of Intelligent and Robotic Systems，2012，65(1)：373-387.

[15] Kelly J，Saripalli S，Sukhatme G S，et al. Combined-visual and inertial navigation for an unmanned aerial vehicle[C]// Proceedings of Field and Service Robotics，2008：255-264.

[16] Veth M，Raquet J F. Fusing low-cost image and inertial sensors for passive navigation[J]. Annual of Navigation，2007，54(1)：11-20.

[17] Mourikis A I，Roumeliotis S I. A multi-state constraint Kalman filter for vision-aided inertial navigation[C]// Proceedings of International Conference on Robotics and Automations (ICRA)，2007：3565-3572.

[18] 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.

[19] Huang G，Mourikis A I，RoumeliotisS I，et al. A quadratic-complexity observability-constrained unscented Kalman filter for SLAM[J]. IEEE Transactions on Robotics，2013，29(5)：1226-1243.

[20] Castellanos J A，Martinez-Cantin R，Tardós J D，et al. Robocentric map joining：improving the consistency of EKF-SLAM[J]. Robotics and Autonomous Systems，2007，55(1)：21-29.

[21] Bloesch M，Omari S，Hutter M，et al. Robust visual inertial odometry using a direct EKF-based approach[C]// Proceedings of Intelligent Robots and Systems，2015：298-304.

[22] Indelman V，Gurfil P，Rivlin E，et al. Real-time vision-aided localization and navigation based on three-view geometry[J]. IEEE Transactions on Aerospace and Electronic Systems，2012，48(3)：2239-2259.

[23] Hu J S，Chen M Y. A sliding-window visual-IMU odemeter based on tri-focal tensor geometry[C]// Proceedings of International Conference on Robotics and Automations (ICRA)，2014.

[24] 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.

[25] Leutenegger S, Furgale P, Rabaud V, et al. Keyframe-based visual-inertial SLAM using nonlinear optimization[J]. International Journal of Robotics Research, 2015, 34(3): 314-334.

[26] Forster C, Carlone L, Dellaert F, et al. On-manifold preintegration for real-time visual-inertial odometry [J]. IEEE Transactions on Robotics, 2017, 33(1): 1-21.

[27] Qin T, Li P, Shen S, et al. VINS-Mono: a robust and versatile monocular visual-inertial state estimator [J]. IEEE Transactions on Robotics, 2018, 34(4): 1004-1020.

[28] Murartal R, Tardos J D. Visual-inertial monocular SLAM with map reuse[J]. IEEE Robotics and Automation Letters, 2017, 2(2): 796-803.

[29] Detone D, Malisiewicz T, Rabinovich A, et al. Superpoint: self-supervised interest point detection and description[C]// Proceedings of Computer Vision and Pattern Recognition (CVPR), 2018.

[30] Tang J, Folkesson J, Jensfelt P, et al. Geometric correspondence network for camera motion estimation [C]// Proceedings of International Conference on Robotics and Automations (ICRA), 2018, 3(2): 1010-1017.

[31] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]// Proceedings of Computer Vision and Pattern Recognition (CVPR), 2015: 1-9.

[32] Kendall A, Grimes M, Cipolla R, et al. PoseNet: a convolutional network for real-time 6-DOF camera relocalization[C]// Proceedings of International Conference on Computer Vision (ICCV), 2015: 2938-2946.

[33] Wang S, Clark R, Wen H, et al. DeepVO: towards end-to-end visual odometry with deep recurrent convolutional neural networks[C]// Proceedings of International Conference on Robotics and Automations (ICRA), 2017: 2043-2050.

[34] Dosovitskiy A, Fischery P, Ilg E, et al. FlowNet: learning optical flow with convolutional networks [C]// Proceedings of International Conference on Computer Vision (ICCV), 2015: 2758-2766.

[35] Costante G, Mancini M, Valigi P, et al. Exploring representation learning with CNNs for frame-to-frame ego-motion estimation[C]// Proceedings of International Conference on Robotics and Automations (ICRA), 2016: 18-25.

[36] Clark R, Wang S, Wen H, et al. VINet: visual inertial odometry as a sequence to sequence learning problem [C]// Proceedings of International Conference on Artificial Intelligence, 2017: 3995-4001.

[37] Chen C, Rosa S, Miao Y, et al. Selective sensor fusion for neural visual inertial odometry[C]// Proceedings of Computer Vision and Pattern Recognition (CVPR), 2019: 10542-10551.

[38] Guo K, Xu F, Yu T, et al. Real-time geometry, albedo and motion reconstruction using a single RGBD camera [J]. ACM Transactions on Computer Graphics, 2017, 36(3): 1-32

[39] Zou D, Wu Y, Pei L, et al. StructVIO: visual-inertial odometry with structural regularity of man-made environments[J]. IEEE Transactions on Robotics, 2019, 35(4): 999-1013.

[40] Xian Z, Hu X, Lian J. Fusing stereo camera and low-cost inertial measurement unit for autonomous navigation in a tightly-coupled approach[J]. Jounal of Navigation, 2015, 68(3): 434-52.

[41] Kong X, Wu W, Zhang L, et al. Tightly-coupled stereo visual-inertial navigation using point and line features[J]. Sensors, 2015, 15(6): 12816-12833.

[42] Dong Z, Zhang G, Jia J, et al. Efficient keyframe-based real-time camera tracking[J]. Computer Vision and Image Understanding, 2014, 118: 97-110.

[43] 崔乃刚，王小刚，郭继峰．基于 Sigma-point 卡尔曼滤波的 INS/Vision 相对导航方法研究[J]. 宇航学报，2009，30(6)：2220-2225.
Cui Naigang, Wang Xiaogang, Guo Jifeng. Research on relative navigation method based on INS/Vision using Sigma-point Kalman filter[J]. Journal of Astronautics, 2009, 30(6): 2220-2225(in Chinese).

[44] 杜光勋，全权，蔡开元．视觉与惯性传感器融合的隐式卡尔曼滤波位置估计算法[J]. 控制理论与应用，2012,29(7)：833-840.
Du Guangxun, Quan Quan, Cai Kaiyuan. Implicit Kalman filter for position estimation with visual and inertial sensor fusion[J]. Control Theory & Applications, 2012, 29(7): 833-840(in Chinese).

[45] Kummerle R, Grisetti G, Strasdat H, et al. G2o: a general framework for graph optimization[C]// Proceedings of International Conference on Robotics and Automations (ICRA), 2011: 3607-3613.

[46] Kaess M, Johannsson H, Roberts R, et al. iSAM2: incremental smoothing and mapping using the Bayes tree[J]. The International Journal of Robotics Re-

search, 2012, 31(2): 216-235.

[47] Agarwal S, Mierle K. Ceres solver [DB/OL]. http://www.ceres-solver.org/tutorial.html.

[48] Kelly J, Sukhatme G. Visual-inertial sensor fusion: localization, mapping and sensor-to-sensor self-calibration [J]. International Journal of Robotics Research, 2011,30(1): 56-79.

[49] Zhou T, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video[C]// Proceedings of Computer Vision and Pattern Recognition (CVPR), 2017: 6612-6619.

[50] Zhan H, Garg R, Weerasekera C S, et al. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction [C]// Proceedings of Computer Vision and Pattern Recognition (CVPR), 2018: 340-349.

[51] Chen C, Lu X, Markham A, et al. IONet: learning to cure the curse of drift in inertial odometry[C]// Proceedings of International Conference on Artificial Intelligence (ICAI), 2018: 6468-6476.

[52] Chen C, Miao Y, Lu X, et al. MotionTransformer: transferring physical motion between domains for neural inertial tracking[C]// Proceedings of International Conference on Artificial Intelligence (ICAI), 2019.

[53] Godard C, Aodha O M, Brostow G J, et al. Unsupervised monocular depth estimation with left-right consistency[C]// Proceedings of Computer Vision and Pattern Recognition (CVPR), 2017: 6602-6611.

[54] Guizilini V, Ambrus R, Pillai S, et al. 3D packing for self-supervised monocular depth estimation[J]. arXiv preprint arXiv:1905.02693, 2019.

[55] Godard C, Aodha O M, Firman M, et al. Digging into self-supervised monocular depth estimation[C]// Proceedings of International Conference on Computer Vision (ICCV), 2019: 3828-3838.

[56] Shen T, Luo Z, Zhou L, et al. Beyond photometric loss for self-supervised ego-motion estimation[C]// Proceedings of International Conference on Robotics and Automations (ICRA), 2019: 6359-6365.

[57] Tang C, Tan P. BA-Net: dense bundle adjustment network[J]. arXiv preprint arXiv:1806.04807, 2018.

[58] Chen C, Lu C X, Wang B, et al. DynaNet: neural Kalman dynamical model for motion estimation and prediction[J]. arXiv preprint arXiv:1908.03918, 2019.

[59] Li Y, Ushiku Y, Harada T. Posegraph optimization for unsupervised monocular visual odometry[C]// Proceedings of International Conference on Robotics and Automations (ICRA), 2019: 5439-5445.

[60] Gálvez-López 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.

[61] Bian J, Li Z, Wang N, et al. Unsupervised scale-consistent depth and ego-motion learning from monocular video[C]// Proceedings of Advances in Neural Information Processing Systems, 2019: 35-45.

[62] 李丰阳, 贾学东, 董明. 惯性/视觉组合导航在不同应用场景的发展[J]. 导航定位学报, 2016, 4(4): 30-35.
Li Fengyang, Jia Xuedong, Dong Ming. Development of vision/inertial integrated navigation in different application scenarios[J]. Journal of Navigation and Positioning, 2016, 4(4): 30-35(in Chinese).

[63] 胡小平, 毛军, 范晨, 等. 仿生导航技术综述[J]. 导航定位与授时, 2020, 7(4): 1-10.
Hu Xiaoping, Mao Jun, Fan Chen, et al. Bionic navigation technology: a survey[J]. Navigation Positioning and Timing, 2020, 7(4): 1-10(in Chinese).

[64] 谢启龙, 宋龙, 鲁浩, 等. 协同导航技术研究综述[J]. 航空兵器, 2019, 26(4): 23-30.
Xie Qilong, Song Long, Lu Hao, et al. Review of collaborative navigation technology[J]. Aero Weaponry, 2019, 26(4): 23-30(in Chinese).