DOI: 10.19650/j.cnki.cjsi.J2007047

融合改进 SuperPoint 网络的鲁棒单目视觉惯性 SLAM*

余洪山,郭 丰,郭林峰,王佳龙,付 强
(湖南大学 电气与信息工程学院 机器人视觉感知与控制技术国家工程实验室 长沙 410082)

摘 要:单目视觉惯性 SLAM 系统通过追踪人工设计的点特征来恢复位姿,如 Shi-Tomasi,FAST 等。然而光照或视角变化等挑战性场景中人工特征提取鲁棒性差,易导致位姿计算精度低甚至失败。启发于 SuperPoint 网络在特征提取的强鲁棒性,提出一种基于改进 SuperPoint 网络的单目 VINS 系统—CNN-VINS,旨在提升挑战性环境下 VINS 系统的鲁棒性。主要贡献包括:提出改进 SuperPoint 特征提取网络,通过动态调整检测阈值实现图像特征点均匀检测和描述,构建鲁棒精确的特征关联信息;将改进 SuperPoint 特征点提取网络与 VINS 系统的后端非线性优化、闭环检测模块融合,提出一个完整的单目视觉惯性 SLAM 系统;对网络的编码层和损失函数优化调整,并验证网络编码层对 VINS 系统定位精度的影响。在公共评测数据集 EuRoc 实验结果表明,相比国际公认的 VINS-Mono 系统,所提系统在光照剧烈变化的挑战性场景中定位精度提升 15%;对光照变化缓慢的简单场景,绝对轨迹误差均值保持在 0.067~0.069 m。

关键词: 单目视觉惯性系统;特征提取网络;同时定位与建图;位姿估计;特征编码
中图分类号: TP242 TH74 **文献标识码**: A **国家标准学科分类代码**: 510.80

Robust monocular visual-inertial SLAM based on the improved SuperPoint network

Yu Hongshan, Guo Feng, Guo Linfeng, Wang Jialong, Fu Qiang
(*National Engineering Laboratory of Robot Visual Perception and Control Technology, College of Electrical and Information Engineering, Hunan University, Changsha 410082, China*)

Abstract: Monocular visual-inertial SLAM (simultaneous localization and mapping) systems recover poses by tracking the hand-crafted point features, such as Shi-Tomas, FAST, and so on. However, the robustness of hand-crafted features is limited in some challenging scenes, such as severe illumination or perspective changes, which may lead to poor localization accuracy. Inspired by the excellent performance of SuperPoint network in feature extraction, a monocular VINS (i.e., CNN-VINS) is proposed, which is based on the self-supervised network and works robustly in challenging scenes. Our main contributions are summarized in three terms. An improved SuperPoint-based feature extraction network is proposed. The dynamical detection threshold adjustment algorithm is used to detect and describe feature points uniformly, which can establish accurate feature correspondence. The improved SuperPoint network is efficiently integrated into a complete monocular visual-inertial SLAM including nonlinear optimization and loop detection modules. In addition, to evaluate the performance of the feature extraction network encoder layer in terms of the localization accuracy of the VINS system, learn and optimize the intermediate shared encoder layer and loss function of the network. Experimental results on the public benchmark EuRoc dataset show that the localization accuracy of our method is increased 15% more than that of VINS-Mono in challenging scenes. In simple illumination change scenes, the mean absolute trajectory error is between 0.067~0.069 m.

Keywords: monocular visual-inertial systems; feature extraction network; simultaneous localization and mapping; pose estimation; feature encoding

收稿日期:2020-10-27 Received Date: 2020-10-27
*基金项目:国家自然科学基金(61973106, U1813205)、湖南省科技计划重点研发项目(2018GK2021)、航空科学基金(201705W1001)、郴州市科技计划项目资助

0 引　　言

同时定位与建图(simultaneous localization and mapping, SLAM)是移动机器人领域的热点研究内容,其中视觉惯性 SLAM 系统(visual inertial slam, VINS)融合视觉特征和惯性测量单元(inertial measurement unit, IMU)估计相机位姿,综合 IMU 和相机的互补特性,具备成本低、功耗少、可恢复尺度等优点,在机器人导航、自动驾驶和虚拟现实等领域中应用广泛[1]。

目前主流的视觉惯性 SLAM 包括传统的方法[1-6]和基于深度学习的方法[7-9]。传统方法主要依靠人工设计的点特征进行图像间匹配和跟踪,通过联合优化视觉特征点重投影误差和 IMU 测量误差来估计相机位姿。VINS-Mono[1]采用稀疏光流进行特征跟踪,利用非线性优化联合约束视觉几何和 IMU 测量误差,更新状态变量,其回环检测和位姿图复用具备较强的性能。但在光照视角变化剧烈的复杂场景中,由于系统依赖点特征提取,易导致特征跟踪失败。基于深度学习的方法大多通过对图像和 IMU 数据采用端到端的联合训练,直接估计相机位姿。VINet[7]是首个使用端到端网络训练的视觉惯性里程计系统,通过对校准误差的学习以增强自身位姿估计的鲁棒性;DeepVIO[8]利用二维光流预测和 IMU 预积分网络融合实现自监督端到端位姿估计。然而端到端的视觉惯性 SLAM 系统需要大量不同场景数据进行训练,泛化性较差,而且没有回环检测模块,无法校正系统长时间运动产生的累积误差。

传统视觉惯性 SLAM 系统通过跟踪人工设计的点特征构建图像间的特征关联。ORB(oriented FAST and BRIEF)[10]依靠计算量小、具备旋转和尺度不变性被 ORB-SLAM2[11]用于视觉前端的特征提取与匹配,然而在低纹理的场景中,依旧会产生大量的误匹配;VINS-Mono 采用 Shi-Tomas[12]检测特征点,通过光流算法追踪特征点,但在光照变化明显场景下的特征追踪性能较低;SIFT(scale-invariant feature transform)[13]特征的匹配精度很高,但计算成本较大,无法实时运行。虽然人工设计的点特征易于提取和描述,但由于场景、光照视角的不均衡使得人工设计的点特征稳定性较低,给算法鲁棒性带来较大影响。近年来,基于卷积神经网络(convolutional neural networks, CNN)的兴趣点检测网络在光照和视角变化的挑战性场景下表现优异。GCN(geometric correspondence network)[14-15]将卷积网络与递归网络结合训练几何对应网络,以检测关键点和生成描述符,同时该网络取代 ORB-SLAM2 系统的 ORB 特征提取器,位姿估计精度与原始系统相当,但对训练场景存在依赖性。SuperPoint[16]使用自监督网络同时预测关键点和描述符,单应性估计性能较高,在光照变化明显的复杂环境中特征提取很稳定。整体而言,在光照和视角变化明显的挑战性场景下,基于 CNN 的特征提取器优于人工设计的特征提取器[17]。鉴于 SuperPoint 网络具备很强的场景适应能力,本文将其作为基础框架改进并融合到 VINS 系统。

针对上述研究现状及问题,为提升 VINS 系统在光照变化剧烈的复杂场景下定位精度,本文提出一种基于自监督特征提取网络的单目视觉惯性 SLAM 系统—CNN-VINS。与 VINS-Mono 等传统方法不同,本文采用深度神经网络提取和匹配特征点,通过动态调整阈值参数,使每帧图像提取均衡的特征点,防止由于场景光照变化过快,导致特征点提取不均衡问题。本文的主要贡献如下:

1)基于现有 VINS-Mono 框架提出一种鲁棒的单目视觉惯性 SLAM 系统,在光照变化强烈的复杂场景下可以保持良好的鲁棒性和定位精度;

2)提出改进 SuperPoint 特征提取网络模型检测特征点和匹配描述符,通过动态调整特征提取阈值,获取均衡的特征点数量,建立健壮可靠的特征点对应关系,保证 SLAM 系统后端优化环节的精度;

3)采用两种主流的轻量化架构对原始 SuperPoint 特征提取框架替换并预训练,对比验证编码层对网络特征提取性能和 SLAM 系统定位精度的影响。同时使用稀疏描述符替换稠密描述符损失函数,加速训练过程。

1 整体算法流程

本文提出的系统如图 1 所示。该系统基于著名的 VINS-Mono 框架搭建起来,工作于 3 个模块:视觉惯性前端、非线性优化后端和闭环检测。视觉惯性前端采用自监督特征提取网络完成特征点检测和匹配,实时跟踪特征点,建立特征关联信息。后端两个线程将前端获取的视觉特征信息和 IMU 测量数据按照时间戳对齐,然后进行滑动窗口优化和闭环全局优化,消除全局累积漂移误差,得到高精度估计位姿。过程如下:

1)视觉惯性前端。使用改进 SuperPoint 网络对图像提取特征点,动态调整阈值保证每帧图像上的特征点数量均衡,通过最近邻检索匹配描述符以实现相邻图像间的特征点跟踪,构建图像特征关联信息,利用随机抽样一致(random sample consensus, RANSAC)算法剔除异常点,保证每帧图像能正确跟踪一定数量的特征点,弱光照下特征提取与匹配结果如图 2 所示。实现图像特征和 IMU 数据对齐,对 IMU 数据预积分,计算 IMU 约束的误差项、协方差和雅克比矩阵,发布视觉特征观测和 IMU 测量误差到后端。

2)非线性优化后端。基于视觉特征观测和 IMU 测量信息估计系统运动状态,采用紧耦合的方式联合优化

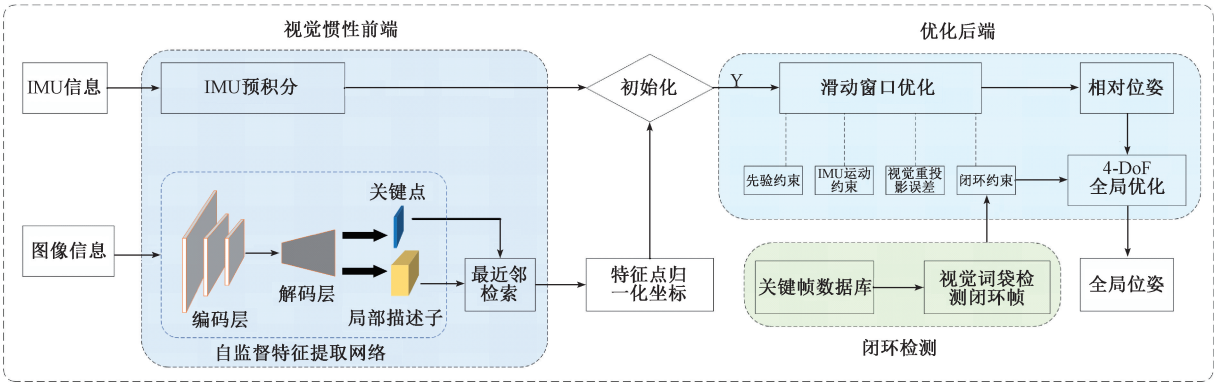
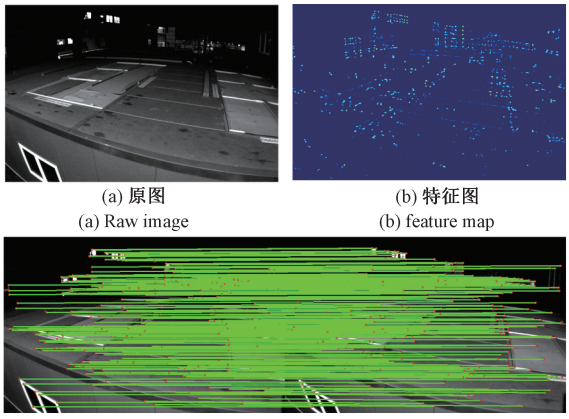


图 1 算法总体方案

Fig. 1 Overall scheme of algorithm



(a) 原图 (a) Raw image

(b) 特征图 (b) feature map

(c) 改进SuperPoint特征匹配 (c) Improved SuperPoint feature matching

图 2 弱光照场景特征匹配示例

Fig. 2 Low light scene feature matching example

得到初始相对位姿。由于优化变量维度较高,因此采用滑窗方式进行局部优化,通过固定优化图像帧数以减少计算量,利用迭代次数保证系统优化精度。局部滑动窗口内联合优化的误差包括:视觉重投影误差、IMU 预积分误差、边缘化先验约束和闭环约束。

3)闭环检测。该模块利用视觉词袋检索闭环帧,通过描述符匹配找到对应特征点,RANSAC 进行异常匹配对的剔除,然后将闭环图像送入滑窗内进行优化得到相对位姿,最后对相对位姿进行 4 自由度的全局轨迹优化得到全局位姿。

2 基于改进 SuperPoint 网络的特征点检测和描述

2.1 改进的 SuperPoint 网络模型

SuperPoint 是一种基于全卷积网络的自监督特征提取网络框架,由编码层、特征点检测层和描述符解码层 3 部分构成,可同时检测和描述特征点。该网络的编码层采用 VGG[18] 架构,结构简单但网络层数较多,导致训练时计算量大且需要大量的训练样本数据。此外,该网络提取的高维特征包含大量冗余信息,不利于检测层的解码过程。针对此问题,本文采用模型参数量少且特征编码信息丰富的 MobileNetV2[19]、GhostNet[20] 两种轻量化架构替换原始网络框架的类 VGG 编码层,从而改进 SuperPoint 网络的特征编码性能。MobileNetV2 与 GhostNet 两种架构采用类似的残差块结构,通过先扩张再压缩的方式,有利于增大特征图的感受野范围;同时加入一系列的线性操作使网络更关注特征数据中的关键信息,因此图像特征的鲁棒性得到了提高。

为保证改进的 SuperPoint 网络满足实时运行的要求。本文采用多尺度、多单应性的方式对训练数据进行预处理,以提高特征点的可重复性和自适应能力;其次,通过设计描述符编码层结构和损失函数,减少描述符的冗余计算量。通过上述对原始 SuperPoint 网络的改进,以此验证和评估改进后的 SuperPoint 网络模型的特征提取性能和对 VINS 系统精度的影响。

1)编码层

本文设计的 3 种编码层详细结构如下:

(1)VGG 编码层。与原始 SuperPoint 的类 VGG 不同,本文在每个卷积层后加入 BatchNorm2d 层,保证训练过程中每层神经网络的输入保持相同分布。本文采用此架构作为对比基准,详细结构可参考文献[16]。

(2)MobileNetV2 编码层。MobileNetV2 架构在检测分类任务上取得不错的性能,相比于 VGG 架构,模型结构复杂,参数量少。体系结构如表 1 所示,该架构基于具有线性瓶颈的倒残差模块,本文设置每层的 bottleneck 模块仅重复 1 次,以减少模型参数量。其中 bottleneck 结构可见文献[19]。

表 1　MobileNetV2 编码层结构

Table 1　MobileNetV2 shared encode structure

卷积层	倍增系数	重复次数	步长
输入 1×480×640	–	–	–
Conv2d+BN+ReLU	0	1	2
bottleneck	1	1	1
bottleneck	6	1	2
bottleneck	6	1	1
bottleneck	6	1	2
bottleneck	6	1	1
bottleneck	6	1	1

(3)GhostNet 编码层。相比 MobileNetV2 架构,可使用较少的模型参数生成更丰富的特征图,该架构在检测和分类任务上效率和准确性较高。图 3 所示为基于 GhostNet 编码层的改进 SuperPoint 网络框架,详细的体系结构如表 2 所示。本文基于 GhostNet 架构的编码层,包含原始网络的前 7 层,除第一层卷积层外,每层都包含 Ghost bottleneck(G-bneck),其结构见文献[20]。为保证编码层聚合更多的图像细节信息,本文设置宽度乘数为 1.2,使整个网络结构通道更宽,信息更丰富。

2)特征点检测层

特征点检测层对共享特征图进行两层卷积操作,将特征图由 $60 \times 80 \times N$ 变为 60×80×65。softmax 操作使特征图取值介于 0~1 之间,将特征点检测变为二分类问题,特征图取值接近于 1 说明该位置处是真实的特征点。最后经过维度变换,将特征图变换为与输入图像相同的尺寸,直接对特征点检测层计算损失函数。

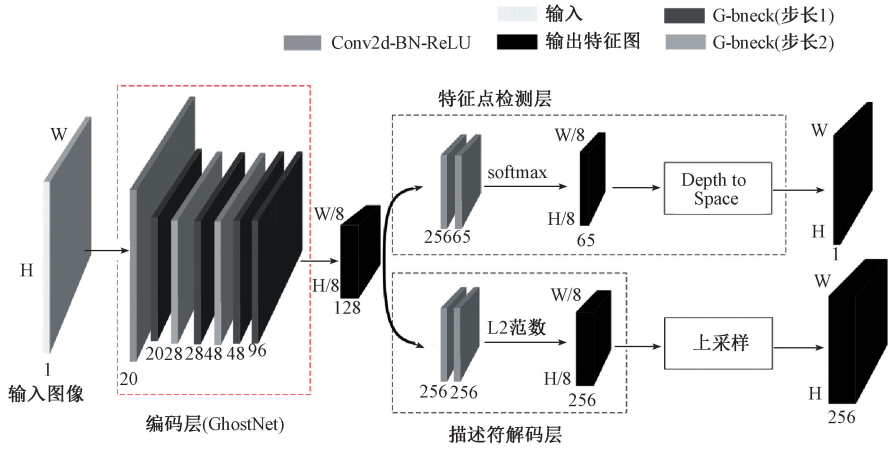


图 3　改进 SuperPoint 网络框架

Fig. 3　Network framework of the improved SuperPoint

表 2　GhostNet 编码层结构

Table 2　GhostNet shared encode structure

卷积层	扩张尺寸	扩张率	步长
输入 1×480×640	–	–	–
Conv2d+BN+ReLU	1	0	2
G-bneck	16	0	1
G-bneck	48	0	2
G-bneck	72	0	1
G-bneck	72	0.25	2
G-bneck	120	0.25	1
G-bneck	240	0	1

3)描述符解码层

原始 SuperPoint 的描述符解码层对描述符的特征图直接进行 3 次插值上采样,使用 L2 范数将特征图取值规范化为单位长度,最终得到 480×640×256 的稠密描述符向量。但描述符的维度过高,包含冗余的描述符信息,给计算带来很大的困扰,因此本文没有将 3 次插值直接放入网络模型中,仅在计算描述符损失函数前将其进行上采样,通过配合 2.2 节的稀疏描述符损失函数保证网络在反向传播过程中计算量大大减少。

2.2　构建损失函数

原始网络的稠密描述符损失函数计算量较大,训练过程中严重影响反向传播速度。为降低描述符损失函数的计算量,本文结合文献[21]提出的稀疏描述符损失函数,采用像素级的度量学习,以最近邻方式训练描述符。改进的 SuperPoint 框架同时检测和描述特征点,联合优化特征点检测层和描述符解码层的损失函数。网络反向传播损失函数包含 3 部分:原始特征点检测器损失 L_p、单应性变换后特征点检测器损失 L_{wp} 和稀疏描述符损失

L_d。最终损失由权重参数 λ 来平衡：

$$L_{\text{total}} = (L_p + L_{wp}) + \lambda L_d \tag{1}$$

1)特征点检测器损失

将原始图像 $\boldsymbol{I}$ 与经过单应变换 $\boldsymbol{H}$ 后的图像 $\boldsymbol{I}_w$ 输入到特征提取网络可得两组相互对应的特征点坐标($\boldsymbol{P}_m^I$,$\boldsymbol{P}_m^{I_w}$)与局部描述子向量($\boldsymbol{D}_m^I$,$\boldsymbol{D}_m^{I_w}$),其中 m 表示网络输出。

如文献[16]所述,特征点检测器损失函数是对预测的特征点坐标和真实标签计算交叉熵损失。计算前需要先将伪真实特征点坐标($\boldsymbol{P}_G^I$,$\boldsymbol{P}_G^{I_w}$)转换为由(0,1)组成的图像特征点标签($\boldsymbol{l}_G^I$,$\boldsymbol{l}_G^{I_w}$),其中 G 表示真实标签。此时 L_p 与 L_{wp} 计算变为分类问题,特征点检测器损失函数如下：

$$\begin{cases} L_p = \dfrac{1}{H_c W_c}\text{CrossEntropy}(\boldsymbol{P}_m^I, \boldsymbol{l}_G^I) \\ L_{wp} = \dfrac{1}{H_c W_c}\text{CrossEntropy}(\boldsymbol{P}_m^{I_w}, \boldsymbol{l}_G^{I_w}) \end{cases} \tag{2}$$

式中：$W_c = W/8$；$H_c = H/8$；W 表示图像宽度；H 表示图像高度。

2)稀疏描述符损失

与原始 SuperPoint 网络使用稠密描述符不同,本文结合文献[21]提出的稀疏描述符替代稠密描述符损失函数,仅使用真实标签处的描述符,利用三重态损失联合确定描述符损失函数。相比于原始的稠密描述符,计算量大大减少,模型训练速度提升为原来的 2~3 倍。

根据真实特征点标签以及网络输出的描述符信息矩阵,选取真实特征点处的预测描述符矩阵：

$$\begin{cases} \boldsymbol{d}_m^I = sel(\boldsymbol{D}_m^I, \boldsymbol{l}_G^I) \\ \boldsymbol{d}_m^{I_w} = sel(\boldsymbol{D}_m^{I_w}, \boldsymbol{l}_G^{I_w}) \end{cases} \tag{3}$$

式中：$\boldsymbol{d}_m^I$ 表示原图真实标签处对应的描述符；$\boldsymbol{d}_m^{I_w}$ 为单应变换后的真实标签对应的描述符；选择函数 $sel(\boldsymbol{D}, \boldsymbol{l}_G)$ 表示从 $\boldsymbol{D}$ 中获取相应的描述符：

$$sel(\boldsymbol{D}, \boldsymbol{l}_G) = \begin{cases} 0, & \boldsymbol{l}_G(x,y) = 0 \\ \boldsymbol{D}_{x,y}, & \boldsymbol{l}_G(x,y) = 1 \end{cases} \tag{4}$$

根据已知两帧图像间的单应变换矩阵 $\boldsymbol{H}$,由式(3)可知对应的描述符向量 $\boldsymbol{d}_m^I$ 与 $\boldsymbol{d}_m^{I_w}$。若存在 N 对正匹配描述符向量,(x,y),(x',y') 表示描述符向量对应位置关系,因此正匹配损失函数：

$$L_{dp}(x_i, y_i) = \| \boldsymbol{d}_m^I(x_i, y_i) - \boldsymbol{d}_m^{I_w}(x_i', y_i') \|_2^2 \tag{5}$$

通过最近邻检索的方式,获取每对正匹配描述符向量周围区域对应的 M 个负样本描述符向量 $\boldsymbol{d}_{mn}^{I_w}$(不包括与 $\boldsymbol{d}_m^I$ 正对应的描述符 $\boldsymbol{d}_m^{I_w}$),则负匹配损失函数：

$$L_{dn}(x_i, y_i) = \frac{1}{M}\sum_{j=1}^{M}(\| \boldsymbol{d}_m^I(x_i, y_i) - \boldsymbol{d}_{mn}^{I_w}(x_j', y_j') \|_2^2) \tag{6}$$

式中：i 表示正匹配描述符样本个数；j 表示每个正匹配描述符中采集的负样本个数。

每个正样本对应 M 个负样本,共产生 $N \times M$ 个负样本对,设置合理的边界余量 m_d,这是衡量描述符之间相似度的重要指标,添加边界余量后,整幅图像的稀疏描述符损失如下：

$$L_d = \frac{1}{N}\sum_{i=1}^{N}\max(0, m_d + L_{dp}(x_i, y_i) - L_{dn}(x_i, y_i)) \tag{7}$$

2.3 预训练

网络模型的训练数据集为 MS-COCO 数据集[22],采用自监督方法进行特征提取网络训练,网络模型在 Pytorch 上搭建和训练。原始 SuperPoint 框架利用合成数据对特征点检测器进行初始化,通过单应性自适应在数据集上生成特征检测器的伪真值特征点标签,但本文选择直接使用预训练的初始化模型对 MS-COCO 数据集提取特征点真值,以保证特征点的真实标签准确性。优化器使用 Adam 优化器,学习率 $l_r = 0.000\ 1$。对训练数据预处理,转换为灰度图并放缩为 240×320,同时将尺度变换、旋转变换和透视变换相结合,构造随机单应性变换,并对训练数据进行随机单应性变换。其中描述符权重 $\lambda = 1.2$,边界余量 $m_d = 1.0$,该模型在真实训练数据上进行 120 K 次迭代。

3 融合改进 SuperPoint 网络和 VINS 的单目视觉惯性 SLAM 系统

3.1 视觉特征跟踪

原始 VINS-Mono 框架采用 Shi-Tomasi 检测特征点,然后基于 LK(lucas kanade)稀疏光流算法[23]进行特征跟踪,以保持单帧图像中跟踪和新检测的特征点之和在 100~300 之间。然而,光流跟踪存在很强的光度不变假设,且真实环境光照多变,在强光照变化下其特征关联的准确度较低。

本文采用对光照变化场景具备较强鲁棒性的深度卷积网络计算特征点位置和描述符,在强光照变化下也能获取一致的局部描述符,网络结构如 2.1 节所述。改进的 SuperPoint 特征提取网络输出结果被设计替换 VINS-Mono 中的 LK 光流跟踪和 Shi-Tomasi 角点。与光流跟踪不同,本文采用描述符最近邻检索完成特征点间的匹配(每帧图像描述符尺寸 $N_k \times 256$),实测两帧图像的描述符匹配时间通过 GPU 加速可控制在 10 ms 内,具体可见 4.3 节。视觉特征跟踪得到相邻图像关联的 n 个归一化特征点集,用于发布到后端对齐 IMU 测量数据和构建视觉重投影误差方程。本文的视觉特征跟踪算法流程如图 4 所示。

为保证每帧图像上均匀的提取特征点,本文采用非极大值抑制(non-maximum suppression, NMS)的方式设

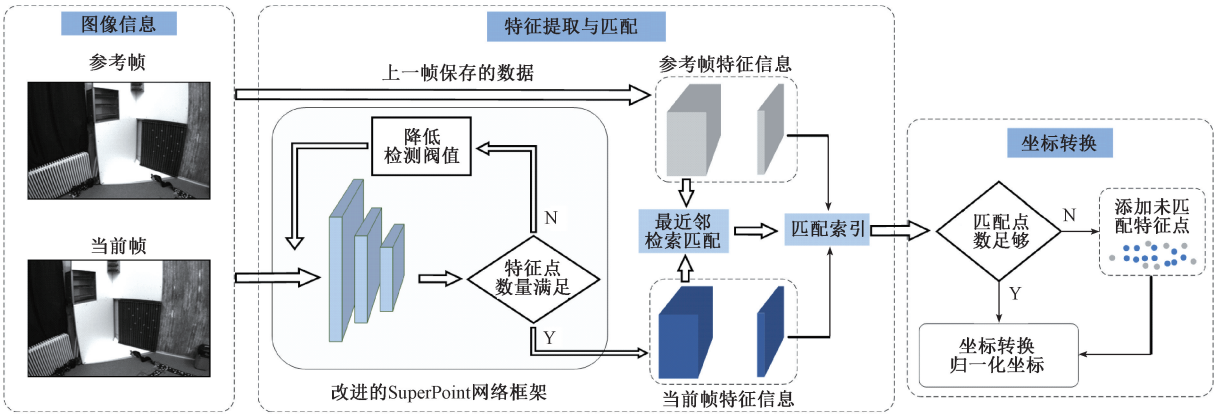


图 4　特征跟踪算法流程

Fig. 4　Feature tracking algorithm flow

定特征点间隔,经试验,在图像分辨率大小为 480×752 时,采用 $NMS=4$ 可以取得较好的特征提取效果。改进的 SuperPoint 特征提取网络是通过设置检测阈值 τ 来检测特征点,阈值设置后不再改变。然而,本文在光照或视角发生剧烈变化的真实数据集上测试时,固定阈值($\tau=0.015$)会导致特征点提取和匹配数量不足的问题,因此本文采用动态调整阈值 τ,保证每帧图像可以提取足够的特征点。由于跟踪的特征点数量设置在 100~300 之间,因此当系统检测到当前图像提取的特征点数量小于 100,便会通过降低检测阈值($\tau=0.008$)以获取更多的特征点,循环检测直到满足特征点数量足够条件,之后利用最近邻检索完成与上一帧图像的特征匹配,构建特征关联信息。

3.2　构建 VINS 系统

上一节中,通过改进的 SuperPoint 特征提取网络,提出的系统能获取准确视觉特征关联信息,特征点以像素坐标和归一化相机坐标进行存储。一旦将当前图像获取的所有特征点消息发布,后端便根据图像间的关联信息利用 SFM(structure from motion)求解滑动窗口内所有帧的相对位姿,利用视觉与 IMU 时间戳对齐获取相邻图像间的 IMU 测量数据,进行 IMU 预积分和测量误差的计算,完成 VINS 系统的初始化。

本文保留 VINS-Mono 的后端非线性优化、闭环检测模块。其中后端非线性优化是通过最小化视觉特征点的重投影误差、IMU 测量残差和滑动窗口边缘化先验约束的方式来联合优化相机位姿,优化的精度取决于迭代次数,考虑到时间约束,本文设置最大迭代次数为 8。闭环模块是基于视觉词袋算法进行回环检索,在关键帧数据库检测到闭环后会进行暴力匹配特征点求解相对位姿,将其作为闭环约束项添加到滑动窗口进行局部优化,当滑动窗口边缘化的图像也是闭环帧时,系统进行全局轨迹优化。

4　实验与分析

实验使用公开数据集 EuRoc 对整个单目视觉惯性 SLAM 系统进行定量评估,并在真实场景中与当前流行的 SLAM 系统进行定性对比。基于改进 SuperPoint 网络的视觉处理前端采用 Python 实现,而非线性优化后端和回环检测采用 C++实现,整体 SLAM 系统通过 ROS 节点进行消息传递。本实验的硬件平台是由 Intel Core i7-8700@3.2 GHz CPU、NVIDIA GTX 1660 显卡、16 GB 内存构成,运行 Ubuntu16.04 操作系统。

公开数据集 EuRoc[24] 由苏黎世联邦理工学院采集,是主流的测试单/双目 VIO 系统的数据集。数据集分简单、中等、困难 3 个等级,且同时发布图像消息和 IMU 消息,图像分辨率为 480×752,IMU 消息以 200 Hz 发布。

HPatches 数据集[25] 是一个评估局部描述符的基准,适用于特征匹配、检索和分类等任务,允许在不同应用场景中进行更真实且可靠的比较。HPatches 数据集本身包含视角变化(59 个序列)和光照变化(57 个序列)两部分,适用于网络模型进行特征提取和匹配的性能评估。

4.1　特征提取网络性能对比

为验证不同编码层输出的特征图对特征提取结果影响的大小,本文除使用类 VGG 架构训练特征提取网络模型外,另采用模型参数少、特征信息较丰富的 GhostNet 和 MobileNetV2 架构分别训练网络模型,以此对比编码层对模型特征提取性能的影响。其中描述符匹配计算采用最近邻(nearest neighbor, NN)检索,并使用 RANSAC 剔除匹配点对中的异常点,以保证匹配点对的正确对应关系。

为量化比较特征提取网络模型的特征提取和匹配性能,本文在 HPatches 数据集上进行性能评估。参考文献[16],采用特征点重复度(Repeatability)和描述符匹配平均准确率(mean average precision, mAP)作为特征点检

测和描述符匹配的性能指标。首先评估特征点检测性能,将 3 种网络模型和 Shi-Tomasi 进行特征点重复度比较,设置特征点检测数量 500,Shi-Tomasi 函数由 OpenCV 实现,模型应用非极大值抑制($NMS=4$),特征点检测阈值 0.015;然后评估描述符匹配能力,由于 Shi-Tomasi 不包含描述符,因此仅对 3 种网络模型进行实验,设置描述符最近邻匹配阈值 0.7。两种性能指标的实验结果如表 3 所示。表 3 中采用 SuperPoint_GhostNet、SuperPoint_MobileNet、SuperPoint_VGG 代表 3 种预训练的特征提取网络模型。

本文采用单应性估计准确度评估 3 种网络模型的整体性能,其中单应性估计准确度表示模型正确估计图像特征点数与真实单应变换后的目标图像特征点数之比。图 5 所示为模型单应性估计准确度随着正确度阈值的增大而变化的曲线,其中正确度阈值 ε 表示真实单应变换后的目标图像特征点与模型估计图像特征点之间的平均像素距离和。由于正确度阈值达到一定值后对单应性估计准确度影响变小,经测试本文设置最大正确度阈值 $\varepsilon=12$。

表 3 特征检测与匹配性能对比结果

Table 3 Comparison of feature detection and matching performance

方法	光照变化场景		视角变化场景	
	Rep.	mAP	Rep.	mAP
Shi-Tomasi	0.61	×	0.60	×
SuperPoint_VGG	0.68	0.83	0.58	0.73
SuperPoint_GhostNet(Ours)	0.68	0.81	0.56	0.65
SuperPoint_MobileNet(Ours)	0.67	0.83	0.56	0.69

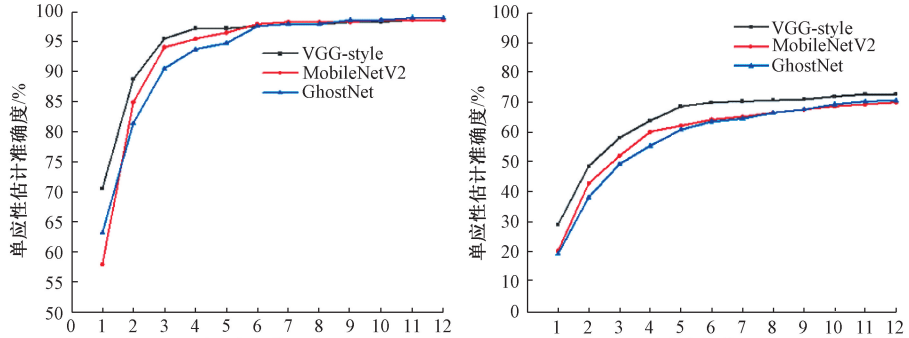


图 5 单应性估计准确度对比

Fig. 5 Comparison of the accuracy of homography estimation

3 种特征提取网络在 HPatches 数据集上直观的特征匹配结果如图 6 所示,前两行是光照变化场景,后两行是视角变化场景,设置特征点检测阈值 $\tau=0.015$,描述符匹配阈值 0.7,采用 RANSAC 筛选匹配点。图中深色线表示正确匹配对,浅色线表示误匹配对,左上角数字表示特征点正确匹配率。

从表 3 和图 5 可以分析得出:

1)在光照变化场景下,3 种模型特征点重复度和描述符匹配精度性能相当,特征点重复度指标高于传统人工设计的检测器 Shi-Tomasi;在视角变化场景下,3 种模型特征点重复度和传统特征提取算法 Shi-Tomasi 性能相当,其中 SuperPoint_GhostNet 的描述符匹配精度较低。

2)综合来看,3 种模型在光照变化场景下特征提取性能更突出。其中 SuperPoint_VGG 在视角变化场景下的单应性估计准确度更高,高于其他两个模型 3%~6%;除此之外,在光照变化场景下,当正确度阈值 $\varepsilon=6$ 时,三者的单应性估计能力相当。

上述实验结果得到两个有用的结论:1)特征提取网络的特征点检测能力高于传统人工设计的检测器,如 Shi-Tomasi;2)不同的编码层对特征提取性能会产生一定影响,本文测试的性能差距在 3%~6%之间。同时引出问题:在实际 VINS 系统中,特征提取网络能否在真实复杂环境下保持出色的性能?接下来,将 3 种模型融合到 VINS 系统中进行真实环境测试。

4.2 VINS 系统精度评估

在公开数据集 EuRoc 上,实验对比本文系统、VINS-Mono 系统的定位精度。由于 3 种不同编码层模型的特征提取性能不同,参见上述 4.1 节,为定量评估,需要将 3 种模型分别进行测试,其中原始 VINS-Mono 代码使用作者提供的默认参数[1]。实验采用 EuRoc 公开数据集,共测试 11 个序列,利用 evo 工具评估里程计全局轨迹误差。本文选用绝对位姿误差(absolute pose error, APE)作为评价指标,直接比较估计值与真实值间的绝对轨迹

(a) 原始图像对
(a) Images pairs

(b) GhostNet编码层
(b) GhostNet encoder

(c) MobileNetV2编码层
(c) MobileNetV2 encoder

(d) 类VGG编码层
(d) VGG-style encoder

图 6　HPatches 数据集特征匹配效果

Fig. 6　HPatches dataset feature matching effect

误差。表 4 所示为 4 种方法在 EuRoc 中等和困难序列上的轨迹平移均方根误差(root mean square error, RMSE),黑色粗体数字表示性能最佳,×表示发生漂移。

表 4　EuRoc 中等/困难序列的均方根误差

Table 4　The RMSE results on the medium/difficult EuRoc dataset

m

序列	VINS-Mono	CNN-VINS(Ours)		
		VGG-style	GhostNet	MobileNetV2
MH_03	0. 077	0. 079	**0. 069**	0. 076
MH_04	0. 151	**0. 094**	0. 117	0. 149
MH_05	0. 126	0. 179	0. 131	**0. 121**
V1_02	**0. 053**	×	0. 055	0. 064
V1_03	0. 085	0. 084	**0. 083**	0. 090
V2_02	0. 103	0. 091	**0. 076**	0. 089
V2_03	0. 197	0. 135	0. 144	**0. 123**
均值	0. 113	0. 110	**0. 096**(↓15%)	0. 102

如表 4 所示,在 EuRoc 复杂场景序列数据集上基于 SuperPoint_GhostNet 的 VINS 系统在 3 个序列上表现出最佳性能;基于 SuperPoint_MobileNet 的 VINS 系统在两个序列上表现出最佳性能;而在 4. 1 节网络模型性能评估上表现突出的 SuperPoint_VGG 则表现不佳,甚至在序列 V1_02 上出现初始化失败导致轨迹漂移的情况,侧面验证 4. 1 节最后提出的问题。实验结果表明,本文的 CNN-VINS 系统在 EuRoc 复杂场景序列上精度较高,与原始 VINS-Mono 绝对位姿误差的差距在厘米级,例如 MH_04,V2_02,V2_03 等序列。在 EuRoc 数据集的简单场景序列上,本文提出的 CNN-VINS 系统与原始 VINS-Mono 算法的位姿估计性能相当,绝对轨迹均方根误差的平均值最低为 0. 067 m,实验结果如表 5 所示。

表 5　EuRoc 简单序列的均方根误差

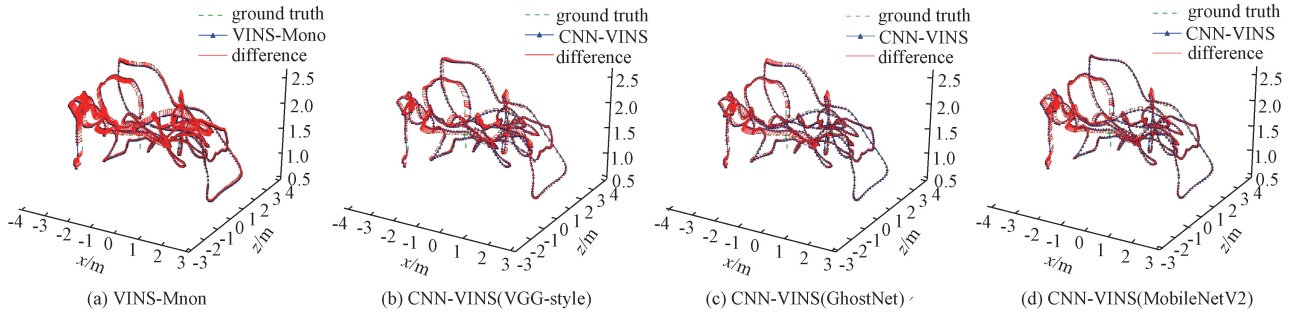
Table 5　The RMSE results on the easy EuRoc dataset

m

序列	VINS-Mono	CNN-VINS(Ours)		
		VGG-style	GhostNet	MobileNetV2
MH_01	0. 084	0. 073	0. 083	**0. 061**
MH_02	0. 072	0. 079	**0. 069**	0. 080
V1_01	0. 059	**0. 055**	0. 060	0. 057
V2_01	**0. 057**	0. 062	0. 063	0. 073
均值	0. 068	**0. 067**	0. 069	0. 068

表 4、5 结果表明,特征提取网络用于 VINS 系统的视觉处理前端可提高复杂环境下位姿估计的精度,特别在光照变化明显的复杂场景中,其系统鲁棒性和定位精度得到提升,其中基于 SuperPoint_GhostNet 的 CNN-VINS 相较于 VINS-Mono 定位精度提升 15%。而对于光照变化不明显的简单场景,本文的 CNN-VINS 系统可达到主流 VINS-Mono 系统的性能,定位轨迹均方根误差的均值保持在 0. 067~0. 069 m 之间。

图 7 所示为 VINS-Mono 和本文 3 种方法在 EuRoc 数据集的 V2_02 序列上运动轨迹估计的定性比较。在 V2_02 序列下,无人机运动剧烈且光线明暗变化不一致,提取特征困难。相比于 VINS-Mono 包含较多的累积误差,本文融合改进 SuperPoint 网络的 CNN-VINS 系统绝对轨迹均方根误差介于 0.076~0.091 m,定位精度提升超过 12%。



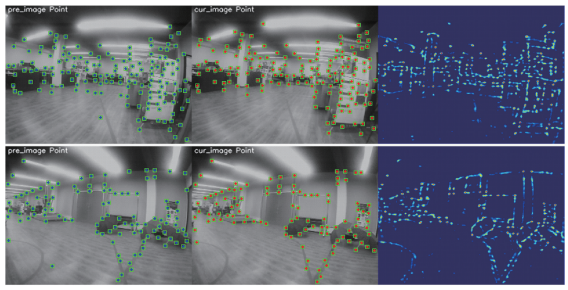
(a) VINS-Mnon (b) CNN-VINS(VGG-style) (c) CNN-VINS(GhostNet) (d) CNN-VINS(MobileNetV2)

图 7 定性比较轨迹误差

Fig. 7 Qualitative comparison of trajectory errors

4.3 真实场景数据评估

利用真实室内环形场景测试 CNN-VINS 系统鲁棒性,采用小觅双目相机作为视觉惯性传感器,型号 D1010-IR-120/Color,同时获取图像和 IMU 数据。实验过程通过手持相机沿规划的环形路径运动,速度相对平稳。通过 ROS 记录整个真实环形室内场景的图像和 IMU 数据,彩色图像大小为 640×480。图 8 所示为基于 SuperPoint_GhostNet 的 VINS 系统在测试过程中的特征提取图。从图 8 中看出,前后帧的特征点重复度较高,特征图稳定,不会由于光照不均衡而发生明显变化。



(a) 参考帧 (b) 当前帧 (c) 特征图

(a) Reference frame (b) Current frame (c) Feature map

图 8 运动过程特征提取图

Fig. 8 Feature extraction of motion processes

为验证系统的鲁棒性和适用性,使用 VINS-Mono 和基于 SuperPoint_GhostNet 的 CNN-VINS 进行真实场景的位姿估计。室内环形场景的轨迹总长约为 51 m,手持相机运动一周回到起点。由于难以获取相机的真实运动位姿,实验中将两种算法均去除闭环模块,通过起点与终点的闭合误差定量判断位姿估计误差,最终运动轨迹如图 9 所示,其中闭合误差表示起点与终点距离。图 9 的放大区域提供了 VINS-Mono 与本文算法的起始点处运动轨迹,两者的闭合误差分别为 33.9 cm(VINS-Mono),15.2 cm(CNN-VINS),本文算法具有较好的闭合误差精度。

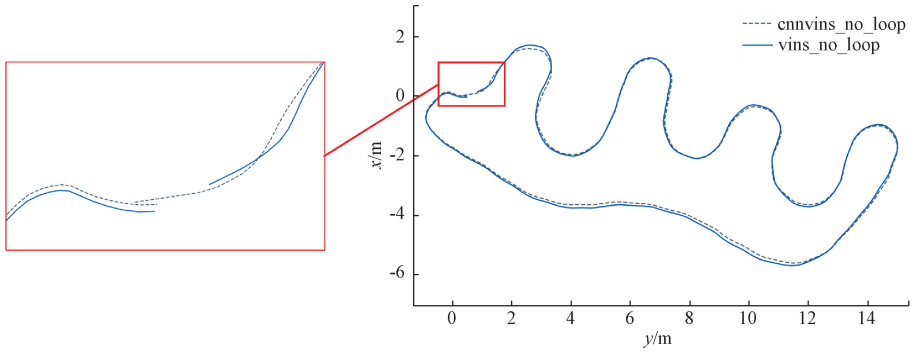


图 9 室内环境下的运动轨迹

Fig. 9 The trajectories in the indoor environment

在上述真实环形环境下,场景图像数量为 1 156 帧,每帧提取/跟踪关键点数量设置为 150 个,使用四种特征提取与跟踪方法 LK + Shi-Tomasi、SuperPoint_VGG、SuperPoint_GhostNet 和 SuperPoint_MobileNet 进行运行时

间评估。其中 LK+Shi-Tomasi 运行在 CPU 上,其余 3 种方法在 GPU 上运行,每阶段消耗总时间和每秒帧数(frames per second, FPS)如表 6 所示。表 6 结果显示,本文算法的平均运行速率小于 LK+Shi-Tomasi,然而 VINS 系统后端非线性优化过程的处理速率在 10 Hz 左右,因此本文 CNN-VINS 系统在真实环境下可实时运行。

表 6　平均运行时间

Table 6　Average running time

功能模块	LK+Shi-Tomasi	SuperPoint		
		VGG-style	GhostNet (Ours)	MobileNetV2 (Ours)
特征提取/角点检测/s	13.0	34.5	34.6	25.5
特征匹配/光流跟踪/s	4.5	5.6	3.2	3.1
总时间/s	17.5	40.1	37.8	28.6
平均速率/Hz	60	28	30	40

5　结　　论

本文提出一种鲁棒的单目 VINS 系统,该系统基于改进 SuperPoint 网络和 VINS-Mono 框架发展而来,通过将特征提取网络融入视觉处理前端,有效提升了 VINS 系统在光照变化剧烈的挑战性场景下定位精度。此外,通过设置 SuperPoint 网络的不同编码层,对比验证得出基于全卷积编码层的网络特征提取性能更高,特别是单应性矩阵的估计能力。在公开 EuRoc 数据集上实验结果表明,在光照变化分布不均衡的复杂场景下,相比于 VINS-Mono 系统,本文系统定位精度提升 15%;光照无明显变化的场景下,与 VINS-Mono 定位精度相当,绝对轨迹均方根误差的平均值保持在 0.068 m 左右。在真实室内环形场景下,本文系统具有更好的闭合误差精度 15.2 cm。

未来工作将考虑将深度预测和特征提取融合为一个网络模型作为纯视觉单目 SLAM 系统的视觉处理前端,以解决纯视觉 SLAM 系统尺度无法观测的问题。同时采用精度更高的描述符匹配方法获取特征点对应关系,增强特征点跟踪的鲁棒性和精度。

参考文献

[1] QIN T, LI P L, SHEN SH J. VINS-mono: A robust and versatile monocular visual-inertial state estimator[J]. IEEE Transactions on Robotics, 2018, 34(4): 1004-1020.

[2] 齐乃新, 张胜修, 杨小冈, 等. 基于相机状态方程多模增广的改进 MSCKF 算法[J]. 仪器仪表学报, 2019, 40(5): 89-98.
QI N X, ZHANG SH X, YANG X G, et al. An improved MSCKF algorithm based on multi-mode augmentation method for the camera state equation[J]. Chinese Journal of Scientific Instrument, 2019, 40(5): 89-98.

[3] 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.

[4] MOURIKIS A I, ROUMELIOTIS S I. A multi-state constraint Kalman filter for vision-aided inertial navigation[C]. IEEE International Conference on Robotics and Automation, 2007: 3565-3572.

[5] BLOESCH M, OMARI S, HUTTER M, et al. Robust visual inertial odometry using a direct EKF-based approach[C]. IEEE International Conference on Intelligent Robots and Systems, 2015: 298-304.

[6] 潘林豪, 田福庆, 应文健, 等. 单目相机-IMU 外参自动标定与在线估计的视觉-惯导 SLAM[J]. 仪器仪表学报, 2019, 40(6): 56-67.
PAN L H, TIAN F Q, YING W J, et al. VI-SLAM algorithm with camera-IMU extrinsic automatic calibration and online estimation[J]. Chinese Journal of Scientific Instrument, 2019, 40(6): 56-67.

[7] CLARK R, WANG S, WEN H K, et al. VINet: Visual-inertial odometry as a sequence-to-sequence learning problem[C]. AAAI, 2017: 3995-4001.

[8] HAN L, LIN Y, DU G, et al. DeepVIO: Self-supervised deep learning of monocular visual inertial odometry using 3D geometric constraints[C]. IEEE International Conference on Intelligent Robots and Systems, 2019: 6906-6913.

[9] SHAMWELL E J, LEUNG S, NOTHWANG W D. Vision-aided absolute trajectory estimation using an unsupervised deep network with online error correction[C]. IEEE International Conference on Intelligent Robots and Systems, 2018: 2524-2531.

[10] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: An efficient alternative to SIFT or SURF[C]. IEEE International Conference on Computer Vision, 2011: 2564-2571.

[11] MUR-ARTAL R, TARDOS J D. ORB-SLAM2: An open-source SLAM system for monocular, stereo and RGB-D cameras[J]. IEEE Transactions on Robotics, 2017, 33(5): 1255-1262.

[12] SHI J, TOMASI C. Good features to track[C]. IEEE Conference on Computer Vision and Pattern Recognition, 1994: 593-600.

[13] LOWE D G. Object recognition from local scale-invariant features[C]. IEEE International Conference on Computer

Vision, 1999: 1150-1157.

[14] TANG J, FOLKESSON J, JENSFELT P. Geometric correspondence network for camera motion estimation[J]. IEEE Robotics and Automation Letters, 2018, 3(2): 1010-1017.

[15] TANG J, ERICSON L, FOLKESSON J, et al. GCNv2: Efficient correspondence prediction for real-time SLAM[J]. IEEE Robotics and Automation Letters, 2019, 4(4): 3505-3512.

[16] DETONE D, MALISIEWICZ T, RABINOVICH A. SuperPoint: Self-supervised interest point detection and description[C]. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018: 337-349.

[17] DAI Z, HUANG X, CHEN W, et al. A comparison of CNN-based and hand-crafted keypoint descriptors[C]. IEEE International Conference on Robotics and Automation, 2019: 2399-2404.

[18] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [C]. International Conference on Learning Representations, 2015: 345-358.

[19] SANDLER M, HOWARD A, ZHU M, et al. MobileNetV2: Inverted residuals and linear bottlenecks [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4510-4520.

[20] HAN K, WANG Y, TIAN Q, et al. GhostNet: More features from cheap operations[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2020: 1577-1586.

[21] JAU Y Y, ZHU R, SU H, et al. Deep keypoint-based camera pose estimation with geometric constraints[C]. ArXiv Preprint, 2020, ArXiv:2007.15122.

[22] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context [C]. IEEE International Conference on Computer Vision, 2014: 740-755.

[23] LUCAS B D, KANADE T. An iterative image registration technique with an application to stereo vision[C]. In Proceedings of the 7th International Joint Conference on Artificial intelligence, 1981: 674-679.

[24] BURRI M, NIKOLIC J, GOHL P, et al. The EuRoC micro aerial vehicle datasets[J]. International Journal of Robotics Research, 2016, 35(10): 1157-1163.

[25] BALNTAS V, LENC K, VEDALDI A, et al. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3852-3861.

作者简介

余洪山，分别在 2001 年、2004 年和 2007 年于湖南大学获得学士、硕士、博士学位，现为湖南大学电气与信息工程学院教授，机器人视觉感知与控制技术国家工程实验室副主任，主要研究方向为自主机器人和机器学习。

E-mail: yuhongshancn@163.com

Yu Hongshan received his B. Sc. degree, M. Sc. degree and Ph. D. degree all from Hunan University in 2001, 2004, and 2007, respectively. He is currently a professor College of Electrical and Information Engineering at Hunan University, and an associate dean at National Engineering Laboratory for Robot Visual Perception and Control. His main research interests include autonomous mobile robot and machine vision.

郭丰（通信作者），2019 年于河南大学获得学士学位，现为湖南大学电气与信息工程学院硕士研究生，主要研究方向为视觉 SLAM 和移动机器人位姿估计。

E-mail: 1637850405@qq.com

Guo Feng (Corresponding author) received his B. Sc. degree from Henan University in 2019. He is currently a master student College of Electrical and Information Engineering at Hunan University. His main research interests include visual SLAM and pose estimation.