

基于深度特征的单目视觉惯导里程计

徐伟锋, 蔡述庭, 熊晓明

(广东工业大学 自动化学院, 广东 广州 510006)

摘要: 视觉里程计是SLAM (Simultaneous Localization and Mapping) 领域中的基石, 单目视觉里程计因其成本低廉和仅需较少的相机标定工作而占据着重要的地位, 但它存在着尺度不确定、尺度漂移、鲁棒性差等缺点。本文在ORB_SLAM3的基础上, 提出了一种基于深度特征的单目视觉惯导里程计, 简称DF-VIO (Visual Inertial Odometry Based on Deep Features), 它采用深度学习网络提取的深度特征替代传统的人工点特征, 并融合了人工线特征, 强化了系统在现实复杂场景下的鲁棒性; 另外, 系统提供了多种位姿跟踪方式, 除了基于恒速模型和跟踪参考关键帧的方式外, 还提供了一种基于深度学习网络的可重复性图的位姿跟踪方法, 进一步提高了系统位姿跟踪的精度。在公开数据集EuRoC上进行对比实验, 在纯视觉模式下, 平均轨迹误差下降了25.9%, 在视觉惯导模式下, 平均轨迹误差下降了8.6%, 证明了本文提出的系统在复杂的场景下能够具有更高的鲁棒性。

关键词: 视觉里程计; 深度学习; 惯导; 线特征

中图分类号: TP301

文献标志码: A

文章编号: 1007-7162(2023)01-0056-05

Visual Inertial Odometry Based on Deep Features

Xu Wei-feng, Cai Shu-ting, Xiong Xiao-ming

(School of Automation, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Visual odometry is the cornerstone in the field of SLAM. Monocular visual odometry occupies an important position because of its low cost and less camera calibration, but it has some shortcomings such as scale uncertainty, scale drift, poor robustness, and so on. To solve these problems, based on ORB-SLAM3, we process a monocular visual-inertial navigation odometer with depth features, referred to as DF-VIO (Visual Inertial Odometry Based on Deep Features), which uses depth features extracted by deep learning network to replace traditional artificial point features, and fuses artificial line features to enhance the robustness of the system in real complex scenes. Besides, the system provides a variety of pose tracking methods. In addition to the method based on the constant speed model and tracking reference keyframe, a pose tracking method based on the predicted repeatability map is also provided, which further improves the accuracy of system pose tracking. Comparative experiments are carried out on the open data set EuRoC, and the average trajectory error is reduced by 25.9% in pure visual mode and 8.6% in visual-inertial mode, which proves that the system proposed in this paper can be more robust in complex scenes.

Key words: visual inertial odometry; deep learning; inertial measurement unit; line features

单目视觉里程计是SLAM领域中的重要组成部分, 其作用是对机器人的位姿进行估计^[1]。然而视觉里程计存在不少缺点, 如在光照强度变化大或者特征较为单一的场景, 大多数的视觉里程计都无法提供让人满意的位姿估计结果^[2]。另外, 在机器人运动

过快、物体发生遮挡等情况下, 视觉里程计的精度会受到重大影响。有的学者采用传统图像方法增强单目视觉里程计的准确性^[3-4], 但仍然无法克服单目相机固有的缺陷。

由于单目相机的特殊性, 单目视觉里程计具有

收稿日期: 2021-02-18

基金项目: 广东省应用型科技研发专项(2017B090909004)

作者简介: 徐伟锋(1996-), 男, 硕士研究生, 主要研究方向为计算机视觉、深度学习等

通信作者: 蔡述庭(1979-), 男, 教授, 博士, 主要研究方向为FPGA加速、深度学习等, E-mail: shutingcai@126.com

尺度不确定、必须初始化和尺度漂移等特点^[5]。SLAM领域的学者们尝试使用多传感器融合技术去解决这些问题,由于惯导单元(Inertial Measurement Unit, IMU)不受各种环境因素的影响,因而成为主要研究热点。Mur-artal和Forster等^[6-7]对视觉惯导里程计进行了数学上的理论证明, Qin等^[8]研究的VINS系统是视觉惯导里程计的经典实现之一,证明了惯导与视觉融合能大幅度提高里程计的精度。

随着深度学习技术的兴起,许多新的方法进入到人们的视野,如物体检测^[9]、语义分割^[10]、深度预测^[11]。这些新技术提供了许多提高里程计精度的新思路。例如:利用深度估计技术解决单目视觉里程计存在的尺度不确定问题^[12];采用物体检测技术构造新的多视图BA(Bundle Adjustment)约束,优化机器人位姿^[13];使用语义分割网络获取图像语义信息,构建语义地图^[14]。受上述技术的启发,本文设计了一个基于深度特征的视觉惯导里程计,采用深度学习网络提取深度特征点和图像高维度信息,提高了特征点匹配的准确性,同时融合了传统的人工线特征,进一步提高了位姿估计的精度。另外,本文还融合了IMU信息,解决单目相机所存在的初始化难、尺度不确定等问题。

1 系统介绍

1.1 系统框架

本文提出的里程计框架如图1所示,预处理部分主要包括特征提取及IMU数据的预积分处理。提取到的图像特征被位姿跟踪模块以及位姿优化模块所使用。

在位姿跟踪部分,本系统提供了3种位姿跟踪方式:恒速模型、跟踪参考关键帧和基于可重复性图的位姿跟踪。前两种方法是视觉里程计常用的方法,本系统增加了线特征的约束,提高了位姿跟踪的精度,第三种跟踪方法是深度特征网络所特有的,它能够大大提高位姿跟踪的成功率与精度。

最后,由于跟踪参考关键帧需要稳定的光照条件,恒速模型在现实中有时也难以得到满足,因而上述的位姿跟踪方法在复杂的日常环境中难以获得一个准确的位姿初始值,进而导致位姿跟踪失败,而IMU具有与环境因素无关的特性,所以本系统也支持融合IMU信息,能够根据IMU获取的速度与加速度信息求解较为精确的位姿初始值,作为位姿跟踪的初始值,加快迭代计算效率的同时提高位姿跟踪的成功率。

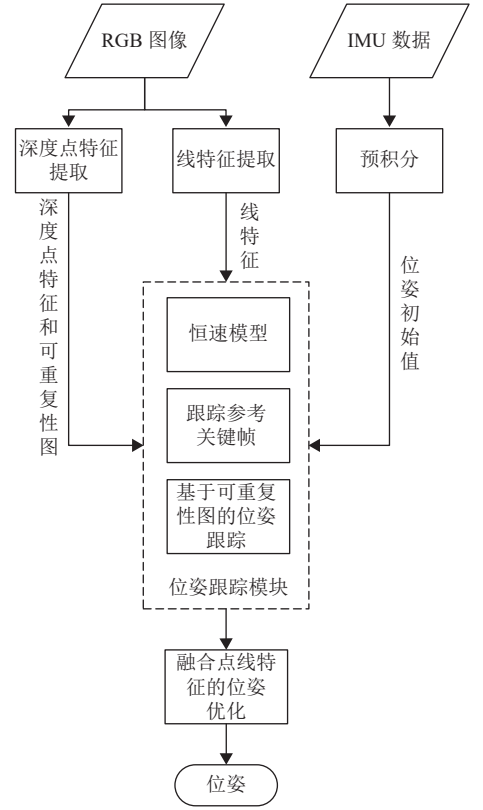


图1 系统框架图

Fig.1 System Framework

1.2 特征提取

本文选用DeTone D等设计的SuperPoint网络^[15]作为特征点提取器。它先用一个共享编码器对输入图像 $I \in \mathbb{R}^{H \times W}$ (H 和 W 分别是图像的高度和宽度)进行编码,然后分别用解码器进行解码,得到一个预测的可重复性单元 $\mathcal{H}' \in \mathbb{R}^{H_c \times W_c \times (C^2+1)}$ 和一个稠密描述子 $\mathcal{H}'' \in \mathbb{R}^{H_c \times W_c \times 256}$ 。其中 C 是网格的大小,在本文中设置为8。 $H_c = \frac{H}{C}$, $W_c = \frac{W}{C}$ 。 $\mathcal{H}'$ 可以标准化为

$$\mathcal{H}(h_c, w_c, y) = \frac{\exp(\mathcal{H}'(h_c, w_c, y))}{\sum_{k=1}^{\infty} \exp(\mathcal{H}'(h_c, w_c, k))} \quad (1)$$

式中: h_c, w_c, y 分别为 $\mathcal{H}'$ 中3个通道中的任意一位。

其他网络的详细信息可见文献^[15]。

另外,本文选用文献^[16]设计的LSD线段提取算法提取线特征,以提高系统在弱纹理场景下的稳定性。为了减少短线段的误匹配对位姿跟踪精度造成影响,本文只筛选像素坐标系下长度大于30的线段。

1.3 位姿跟踪

在位姿跟踪部分,除了经典SLAM框架中的跟踪关键帧和根据恒速模型估算位姿以外,本文还采用了文献^[17]中提出的基于预测的可重复性图的位姿跟踪算法,与传统的直接法直接最小化图像的光

度误差不同,该方法通过最小化基于预测的可重复性图的重投影误差项来位姿跟踪。由于该方法构造的误差函数为凸函数,比传统直接法的误差函数更容易收敛到最优解,因此能够取得更高的精度。

像素级的可重复性图 $\mathcal{R}$ 和分块级的可重复性图 $\mathcal{R}_d$ 可分别定义为

$$\begin{aligned}\mathcal{R} &= -\lg(s(\mathcal{R}(\cdot, \cdot, -1))) \\ \mathcal{R}_d &= \mathcal{R}(\cdot, \cdot, C^2 + 1)\end{aligned}\quad (2)$$

式中: $\mathcal{R} \in \mathbb{R}^{H \times W}$, $s: \mathbb{R}^{H_C \times W_C \times C^2} \rightarrow \mathbb{R}^{H \times W}$, s 为从可重复性图像素块到2D预测值的映射函数。

本文定义了新的基于可重复性的残差替代传统直接法的光度残差:

$$e_i^{\text{repeat}} = \mathcal{R}_{(\cdot)}(\pi_{(\cdot)}((\xi_k \oplus \mathbf{T}_k) p_i)) \quad (3)$$

式中: $\mathcal{R}_{(\cdot)}$ 为两种不同分辨率的可重复性图之一, $\pi_{(\cdot)}$ 为对应分辨率下的相机投影方程, ξ_k 为位姿增量, $\mathbf{T}_k$ 为当前机器人的位姿, p_i 为3D路标点,其对应的像素点为 $u_{i,k}^\pi$ 。可重复性残差项 e_i^{repeat} 的雅克布方程可以表示为

$$J_i = J_{\text{repeat}} - J_{\text{proj}} - J_{\text{pose}} = \frac{\partial \mathcal{R}_{(\cdot)}(u_{i,k}^\pi)}{\partial u_{i,k}^\pi} \cdot \frac{\partial u_{i,k}^\pi}{\partial p_i^c} \cdot \frac{\partial ((\xi_k \oplus \mathbf{T}_i) p_i)}{\partial \xi_k} \quad (4)$$

式中: J_{proj} 为相机投影方程对路标点的偏导数, J_{pose} 为投影点对位姿增量的偏导数, J_{repeat} 为可重复性图对像素点的偏导数。

因此,误差函数可表示为

$$\widehat{\xi}_k = \arg \min_{\xi_k} \sum_{i \in \mathcal{Q}_k} \|\mathcal{R}(\pi(\xi_k \oplus \mathbf{T}_k) p_i)\|_\gamma \quad (5)$$

式中: $\mathcal{Q}_k$ 被初始化为上一帧跟踪到的路标点集合,若集合过小,则将其其他共视程度较高的帧中的路标点加入该集合,使集合大小满足迭代的最低要求。

为了提高计算的速度,先将可重复性图上的分块级局部特征与3D路标点进行关联,然后使用列文伯格-马夸尔特法迭代优化,使得误差函数收敛,得到精度粗糙的位姿增量。

此时,在像素级的可重复性图上应用半径搜索算法,寻找与3D路标点的描述子最接近的像素点 $\tilde{u}_{i,k}$,随后,只要最小化重投影误差即可求出机器人当前位姿 $\mathbf{T}_k$,其对应的旋转矩阵为 $\mathbf{R}_k$,平移向量为 $\mathbf{t}_k$ 。对于每个3D路标点 p_i ,重投影误差为

$$e_{i,k}^{\text{repro}} = \pi(\mathbf{R}_k p_i + \mathbf{t}_k) - \tilde{u}_{i,k} \quad (6)$$

可以推导出点特征重投影误差函数为

$$E^{\text{point}} = \sum_{i \in \mathcal{Q}_k} \|(e_{i,k}^{\text{repro}})^T \boldsymbol{\Omega}_{i,k}^{-1} e_{i,k}^{\text{repro}}\|_\gamma \quad (7)$$

式中: $\boldsymbol{\Omega}_{i,k}$ 为像素点 $\tilde{u}_{i,k}$ 的协方差矩阵,它决定了特征点在优化中的权重,在ORB-SLAM框架中,它取决于特征点在图像金字塔的哪一层中被检测到,而在基于深度学习的特征提取网络中,没有图像金字塔的概念,因此本文用特征点的局部显著性作为该点的协方差矩阵,其定义为

$$\boldsymbol{\Omega} = E[(u - u_{\text{peak}})(u - u_{\text{peak}})^T] = \sum_{i \in \mathcal{V}_p} \frac{\mathcal{R}(u_i)}{\sum_{k \in \mathcal{V}_p} \mathcal{R}(u_k)} (u_i - u_{\text{peak}})(u_i - u_{\text{peak}})^T \quad (8)$$

式中: u_i 为可重复性图中的第 i 个像素点, u_{peak} 为可重复性图中的峰值像素点,而 $\mathcal{V}_p$ 为该点附近具有正响应值的像素点集合。

为了进一步提升视觉里程计的精度,本文还加入了线特征的信息来进行位姿估计。

假定三维空间下线特征线段的两个端点分别为 $P, Q \in \mathbb{R}^3$,相机坐标系下线特征线段的两个端点分别为 $p_d, q_d \in \mathbb{R}^2$,它们对应的齐次坐标分别为 $p_d^h, q_d^h \in \mathbb{R}^3$,则归一化线性系数为

$$I = \frac{p_d^h \times q_d^h}{|p_d^h \times q_d^h|} \quad (9)$$

本文将线特征的重投影误差定义为线段两个端点的重投影误差之和,即

$$E^{\text{line}}(p_d, q_d, I) = E_{\text{pl}}^2(p_d, I) + E_{\text{pl}}^2(q_d, I) \quad (10)$$

$$E_{\text{pl}}^2(p_d, I) = I^T p_d, E_{\text{pl}}^2(q_d, I) = I^T q_d \quad (11)$$

本文将整个系统的误差函数设置为点特征的重投影误差加上线特征的重投影误差,即

$$E = E^{\text{point}} + E^{\text{line}} \quad (12)$$

通过迭代优化重投影误差函数,即可得到更加精确的机器人位姿。

1.4 IMU融合

这部分采用与ORB_SLAM3^[18]相同的融合策略。系统的状态向量定义为 $\mathbf{S}_i = \{\mathbf{T}_i, v_i, b_i^s, b_i^a\}$,其中 $\mathbf{T}_i$ 是第 i 帧机器人的位姿, $\mathbf{T}_i = [\mathbf{R}_i, p_i]$, v_i 是机器人对应时刻的速度, b_i^s 是陀螺仪的零偏, b_i^a 是加速度计的零偏。对其进行预积分,可得到测量向量信息矩阵 $\boldsymbol{\Omega}_{i,i+1}$ 的预积分为

$$r_{X_{i+1}} = [r_{\Delta R_{i+1}}, r_{\Delta v_{i+1}}, r_{\Delta p_{i+1}}] \quad (13)$$

式中:

$$r_{\Delta R_{i+1}} = \lg(\Delta \mathbf{R}_{i,i+1}^T \mathbf{R}_i^T \mathbf{R}_{i+1}) \quad (14)$$

$$r_{\Delta v_{i+1}} = \mathbf{R}_i^T (v_{i+1} - v_i - g \Delta t_{i,i+1}) - \Delta v_{i,i+1} \quad (15)$$

$$r_{\Delta p_{i+1}} = \mathbf{R}_i^T \left(p_j - p_i - v_i \Delta t - \frac{1}{2} g \Delta t^2 \right) - \Delta p_{i+1} \quad (16)$$

式中: g 为加速度常数。

则3D点 x_j 在第 i 帧图像的重投影误差可表示为

$$r_{ij} = u_{ij} - \mathbf{II} \left(\mathbf{T}_{CB} \mathbf{T}_i^{-1} \oplus x_j \right) \quad (17)$$

式中: u_{ij} 为3D点 x_j 在第 i 帧的投影点, $\mathbf{II}$ 为相机投影矩阵, $\mathbf{T}_{CB}$ 为机器人坐标到相机坐标的刚体变换矩阵。因此对于给定的 $k+1$ 个关键帧及对应的状态量 $\bar{\mathbf{S}}_k \triangleq \{\mathbf{S}_0, \dots, \mathbf{S}_k\}$ 和给定的 l 个3D点集及其对应的状态量 $\mathcal{X} \triangleq \{x_0, \dots, x_{l-1}\}$,IMU误差函数可以表示为

$$\min_{\bar{\mathbf{S}}_k, \mathcal{X}} \left(\sum_{i=1}^k \|r_{i-1}\|_{\mathbf{\Omega}_{i-1}^2}^2 + \sum_{j=0}^{l-1} \sum_{i \in \mathcal{K}_j} \rho_{\text{Hub}}(\|r_{ij}\|_{\mathbf{\Omega}_{ij}}) \right) \quad (18)$$

式中: $\mathcal{K}_j$ 为第 j 个3D点的观测关键帧集合, ρ_{Hub} 为用于减少误匹配点对影响的鲁棒Huber核函数。

通过优化式(18),可以得到一个比较精准的位姿初始值。

1.5 位姿优化

假设第 j 个地图点表示为 $X_j \in \mathbb{R}^3$,对于第 i 个关键帧,其相机投影方程为

$$x_{i,j} = \pi(X_j, \mathbf{T}_i, \mathbf{K}) \quad (19)$$

式中: $\mathbf{T}_i$ 为机器人当前位姿, $\mathbf{K}$ 为相机参数,又知观测值为 $x_{i,j}$,则误差可定义为

$$\omega_{ij} = x_{ij} - x_{ijj} \quad (20)$$

类似的,对于线特征,假设在地图上线段的两个端点分别为 P_j 和 Q_j ,对应的投影点的齐次坐标可分别表示为

$$p_{i,j}^h = \pi(P_j, \mathbf{T}_i, \mathbf{K}) \quad (21)$$

$$q_{i,j}^h = \pi(Q_j, \mathbf{T}_i, \mathbf{K}) \quad (22)$$

因此,对于图像上对应的观测点 $p_{i,j}$ 和 $q_{i,j}$,误差可定义为

$$e'_{i,j} = (\tilde{\mathbf{I}}_{i,j})^T (\mathbf{K}^{-1} p_{i,j}^h) \quad (23)$$

$$e''_{i,j} = (\tilde{\mathbf{I}}_{i,j})^T (\mathbf{K}^{-1} q_{i,j}^h) \quad (24)$$

式中: $\tilde{\mathbf{I}}_{i,j}$ 为将观测点 $p_{i,j}$ 和 $q_{i,j}$ 代入式(9)得到的观测值。

最后,可以将误差整合为

$$C = \sum_{i,j} \rho(e_{i,j}^T \mathbf{\Omega}_{i,j}^{-1} e_{i,j} + e'_{i,j}{}^T \mathbf{\Omega}_{i,j}'^{-1} e'_{i,j} + e''_{i,j}{}^T \mathbf{\Omega}_{i,j}''^{-1} e''_{i,j}) \quad (25)$$

式中: ρ 为Huber核函数, $\mathbf{\Omega}_{i,j}$, $\mathbf{\Omega}_{i,j}'$, $\mathbf{\Omega}_{i,j}''$ 分别为点特征和线段的两个端点的协方差矩阵。

通过迭代优化最小化误差函数 C ,即可优化机器人估计的位姿。

2 实验评估

为了更好地评估本文提出的系统在具有挑战性的场景下的鲁棒性,本文选取了EuRoC数据集^[19]来测试本文提出的系统,由于本文的系统是基于ORB_SLAM3提出的,因此将其作为对照目标,对比两个系统在EuRoC数据集上的性能。评价指标方面,本文选用了里程计的两个最重要的指标——精度与效率,对系统进行评估。

2.1 位姿跟踪精度

作为视觉里程计,位姿估计的准确度始终是放在首位的,除了需要提供准确的位姿估计以外,还需要具备较强的鲁棒性,以适应现实的场景,因此本文将两个方案在每个数据集上测试5次,取其平均值作为该数据集的轨迹误差,结果如表1所示。

表1 位姿误差
Table 1 Trajectory error

模式	ORB_SLAM3		DF-VIO	
	视觉	视觉惯导	视觉	视觉惯导
MH01-easy	0.022 9	0.040 7	0.023 0	0.041 2
MH02-medium	0.046 4	0.047 0	0.033 8	0.042 6
MH03-medium	0.028 0	0.058 5	0.025 7	0.048 0
MH04-difficult	0.084 9	0.080 5	0.082 9	0.073 3
MH05-difficult	0.074 6	0.079 4	0.059 1	0.062 3
V101-easy	0.032 8	0.025 6	0.033 1	0.025 1
V102-medium	0.015 9	0.060 9	0.014 9	0.058 6
V103-difficult	0.093 6	0.052 8	0.038 8	0.050 3
V201-easy	0.027 1	0.049 6	0.014 3	0.047 6
V202-medium	0.017 3	0.055 0	0.019 8	0.054 1
V203-difficult	0.070 1	0.064 3	0.034 8	0.058 3
平均	0.046 7	0.055 8	0.034 6	0.051 0

从表1可以看出,在纯视觉模式下,本文提出的系统在大多数数据集上估计的位姿更加准确,得到的轨迹误差更小,这一点在标记为“difficult”的数据集上体现的更加突出,本文提出的系统在以上数据集的平均误差只有ORB_SLAM3的74.1%。这与预期相符合,由于深度点特征包含图像的高维度信息,误匹配概率较低,并且添加了线特征约束,位姿求解与优化的结果更准确,因而误差更小。

在视觉惯导的模式下,本文提出的系统在数据集上的性能仍然有所提升,但提升的幅度不是很大,这是因为在此模式下,机器人的位姿主要由IMU提

供的速度与加速度来估计,位姿的精度本来就比较
高,而深度点特征与线特征只是用于优化位姿,因此
本文提出的方法对位姿的影响变小了。

2.2 位姿跟踪效率

实时性对于视觉里程计来讲也是一个重要的指
标,为了探讨提出的方法对系统实时性的影响,本文
测试了位姿跟踪时间、特征提取时间和局部光束平
差法(Bundle Adjustment, BA) 时间,结果如表2所示。

表 2 运行时间测试

Table 2 Run time testms

模块	ORB_SLAM3	DF-VIO
位姿跟踪	13.808 46	13.184 43
局部BA	167.424 54	250.339 62
点特征提取	35.956 94	68.448 39
线特征提取	—	12.948 08

从表2可以看出,本文提出的系统减少了位姿跟
踪的时间,这是因为本文提出的方案增加了基于预
测的可重复性图的位姿跟踪方法,在增加了位姿跟
踪的成功率的同时,减少了因跟踪失败而切换至其
他跟踪方法所消耗的时间。

另外,深度特征的提取时间比传统点特征提取
的时间要长,再加上提取线特征的时间,约为
ORB_SLAM3的提取时间的2倍。位姿优化部分由于
添加了线特征约束,因而局部光束平差法的时间也
有所增加,大概增加了50%。

3 结论

本文从目前视觉里程计的问题出发,设计了基
于深度特征的视觉惯导里程计,并融合了线特征,提
高了系统在机器人剧烈运动场景或者弱纹理场景下
的鲁棒性,在EuRoC数据集上进行测试,位姿估计效
果优异。在纯视觉模式下,本文提出的视觉里程计
的平均轨迹误差对比方案下降了25.9%,在惯导模
式下,由于本文所提出的特征只用于优化位姿,平均
轨迹误差下降了8.6%,这证明了本文提出的方法能
够提高里程计的鲁棒性及准确性。

另外,本文也对里程计进行了时间测试,从测试
结果可知由于本文添加了线特征约束,而且提取深
度点特征和人工线特征也相对比较耗时,因此本文
提出的系统的特征提取时间变成了原来的两倍,而
优化时间也增加了一半,尽管依然能满足实时性的
要求,但这在实时性敏感的场景可能会带来一些问
题,未来可以考虑使用硬件改进来解决这一问题。

参考文献:

[1] CADENA C, CARLONE L, CARRILLO H, *et al.* Past, present, and future of simultaneous localization and mapping: toward the robust-perception age [J]. *IEEE Transactions on Robotics*, 2016, 32(6): 1309-1332.

[2] PARK S, SCHÖPS T, POLLEFEYS M. Illumination change robustness in direct visual slam[C]//2017 IEEE international conference on Robotics and Automation (ICRA) . Singapore: IEEE, 2017: 4523-4530.

[3] 池鹏可, 苏成悦. 移动机器人中单目视觉里程计的研究[J]. *广东工业大学学报*, 2017, 34(5): 40-44.

CHI P K, SU C Y. A Research on monocular visual odometry for mobile robots [J]. *Journal of Guangdong University of Technology*, 2017, 34(5): 40-44.

[4] 汝少楠, 何元烈, 叶星余. 基于稀疏直接法闭环检测定位的视觉里程计[J]. *广东工业大学学报*, 2021, 38(3): 48-54.

RU S N, HE Y L, YE X Y. Visual odometry based on sparse direct method loop-closure detection [J]. *Journal of Guangdong University of Technology*, 2021, 38(3): 48-54.

[5] STRASDAT H, MONTIEL J, DAVISON A J. Scale drift-aware large scale monocular SLAM[C]// *Robotics: Science and Systems VI*. Zaragoza: MIT Press, 2010.

[6] MUR-ARTAL R, TARDÓS J D. Visual-inertial monocular SLAM with map reuse [J]. *IEEE Robotics and Automation Letters*, 2017, 2(2): 796-803.

[7] FORSTER C, CARLONE L, DELLAERT F, *et al.* On-manifold preintegration for real-time visual-inertial odometry [J]. *IEEE Transactions on Robotics*, 2016, 33(1): 1-21.

[8] QIN T, SHEN S. Online temporal calibration for monocular visual-inertial systems[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) . Madrid: IEEE, 2018: 3662-3669.

[9] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection[EB/OL]. arXiv: 2004.10934 (2020-04-23). <https://arxiv.org/abs/2004.10934>.

[10] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston: IEEE, 2015: 3431-3440.

[11] EIGEN D, FERGUS R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C]//*Proceedings of the IEEE International Conference on Computer Vision*. Santiago: IEEE, 2015: 2650-2658.

[12] TORRES-CAMARA J M, ESCALONA F, GOMEZ-DONOSO F, *et al.* Map slammer: densifying scattered KSLAM 3D maps with estimated depth[C]//*Iberian Robotics Conference*. Porto: Springer, Cham, 2019: 563-574.