DOI:10.13878/j.cnki.jnuist.2021.03.002

胡凯[1,2]　吴佳胜[1,2]　郑翡[1,2]　张彦雯[1,2]　陈雪超[1,2]　鹿奔[1,2]

视觉里程计研究综述

摘要

视觉里程计(Visual Odometry)作为视觉同步定位与地图构建技术(Visual Simultaneous Localization and Mapping)的一部分,主要通过相机传感器获取一系列拥有时间序列图像的信息,从而预估机器人的姿态信息,建立局部地图,也被称为前端,已经被广泛应用在了多个领域,并取得了丰硕的实际成果,它对于无人驾驶、全自主无人机、虚拟现实和增强现实等方面有着重要意义.本文在介绍经典视觉里程计技术框架模块中的各类算法的基础上,对近年来新颖的视觉里程计技术(VO)的研究和论文进行了总结,按照技术手段不同分为两大类——多传感器融合的视觉里程计(以惯性视觉融合为例)和基于深度学习的视觉里程计.前者通过各传感器之间的优势互补提高VO的精度,后者则是通过和深度学习网络结合改善VO的性能.最后通过比较视觉里程计现有算法,并结合VO面临的挑战展望了视觉里程计的未来发展趋势.

关键词

视觉里程计;多传感器融合;深度学习

中图分类号 TP24

文献标志码 A

收稿日期 2019-11-13

资助项目 国家自然科学基金(61773219,61701244);国家重点研发计划重点专项课题(2018YFC1405703);2020年江苏省大学生创新创业省级重点项目(2020103000492)

作者简介

胡凯,男,博士,副教授、高级实验师,研究方向为机器人.nuistpanda@163.com

吴佳胜(通信作者),男,硕士生,研究方向为机器人同步定位与地图构建.18795875237@163.com

1 南京信息工程大学 自动化学院,南京,210044

2 南京信息工程大学 江苏省大气环境与装备技术协同创新中心,南京,210044

0 引言

为了使得计算机能够和人一样通过感觉器官观察世界、理解世界和探索未知区域,视觉里程计(Visual Odometry,VO)技术应运而生.作为同步定位与地图构建(Simultaneous Localization and Mapping,SLAM)[1-3]的前端,它能够估计出机器人的位姿.一个优秀的视觉里程计技术能为SLAM的后端、全局地图构建提供优质的初始值,从而让机器人在复杂的未知环境中实现精准自主化来执行各种任务.传统的里程计,如轮式里程计因为轮子打滑空转而容易导致漂移,精确的激光传感器价格昂贵,惯性传感器虽然可以测量传感器瞬时精确的角速度和线速度,但是随着时间的推移,测量值有着明显的漂移,使得计算得到的位姿信息不可靠.而视觉里程计由于视觉传感器低廉的成本和长距离较为精准的定位在众多传统里程计中脱颖而出.

所谓视觉里程计就是从一系列图像流中恢复出相机的运动位姿,这一思想最早是由Moravec[4]提出的,他们不仅在论文中第一次提出了单独利用视觉输入的方法估计运动,而且提出了一种最早期的角点检测算法,并将其使用在行星探测车上,体现了现阶段视觉里程计的雏形,包括特征点检测及匹配、外点排除、位姿估计三大块,使得视觉里程计从提出问题阶段过渡到了构建算法阶段,Nister等[5]在CVPR上发表的论文中提出了一种利用单目或者立体视觉相机来获取图像的视觉里程计系统,宣告VO技术进入了优化算法阶段.随着ORB-SLAM[6]的问世,VO作为SLAM的前端成为了研究热潮,也代表着主流基于特征点法VO的一个高峰.Engle等[7]提出的LSD-SLAM则成功地把直接法的视觉里程计应用在了半稠密单目SLAM中.近年来涌现了各类的新颖视觉里程计系统,比如2019年Zheng等[8]提出了一种基于RGB-D传感器的自适应视觉里程计,可以根据是否有足够的纹理信息来自动地选择最合适的视觉里程计算法即间接法或者直接法来估计运动姿态.

本文重点对视觉里程计的已有研究工作进行综述,主要选取了近年来有代表性的或取得比较显著效果的方法进行详细的原理介绍和优缺点分析.根据是否需要提取特征点大致分为特征点法和直接法.也可以根据是否脱离经典的位姿估计模块方法分为经典视觉里程计和新颖视觉里程计.最后总结并提出未来的发展前景.

本文第 1 节介绍传统视觉里程计框架的算法.其中包括特征点法 VO 的关键技术和直接法视觉里程计中的相关算法.第 2、第 3 节综述最新的视觉里程计研究方法,包括第 2 节中惯性视觉传感器融合的易于工程实现轻量型的 VO,以及第 3 节中基于深度学习的视觉里程计可以通过高性能计算机实现精密建图等功能.第 4 节简要概括视觉里程计的各类标志性算法.第 5 节结合视觉里程计面临的挑战,展望了未来的发展方向.

1 传统视觉里程计

传统视觉里程计沿用了 Nister 等[5]的 VO 框架,即依据相邻帧之间特定像素几何关系估计出相机的位姿信息,包括位置(x,y,z)和滚转角(roll)、俯仰角(pitch)以及偏航角(yaw)三个方向信息.根据是否需要提取特征,分为特征点法和以灰度不变假设为前提的直接法.

1.1 特征点法

特征点法首先从图像中提取出关键特征点,然后通过匹配特征点的手段估计出相机运动.大致分为了两个部分,即特征点的提取匹配和相机运动的位姿估计.特征点法在视觉里程计中占据了主要地位,是因为其运行稳定,而且近年来研究者们设计了许多具有更好鲁棒性的图像特征,这些特征对于光照敏感性低,而且大多拥有旋转不变性和尺度不变性.线面特征的提出更是使得特征点法适应了纹理信息少的场景.特征点法示意如图 1 所示.

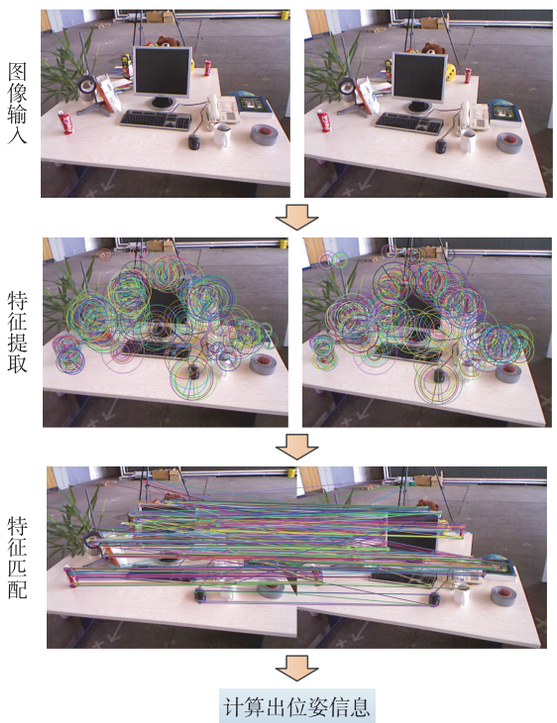

图 1 特征点法示意图

Fig. 1 Schematic diagram of feature-based method

1.1.1 特征提取及匹配

经典的特征算子有 SUSAN[9]、Harris[10]、FAST[11]、Shi-Tomasi[12]、SIFT[13]、SURF[14]、PCA-SIFT[15]、ORB[16],其中最为基础也是最为经典的是 Harris 和 SIFT(尺度不变特征变换)算法,现有的算法基本都是基于这两者,可以看作是 Harris 和 SIFT 的简化和改进.Harris 角点检测算法运用了微分运算和角点邻域的灰度二阶矩阵,而微分运算对图像密度和对亮度的变化不敏感性和二阶矩阵特征值不变性使得 Harris 角点检测算子拥有了光照变化不敏感、旋转不变性.后来出现了 SUSAN 算子,它的原理和步骤和 Harris 较为相似,但是 SUSAN 算子不仅拥有较好的边缘检测性能,在角点检测方面也有较好的效果,能够应用在需要轮廓提取的环境下.

但是 Harris 算子和 SUSAN 算子都不具备尺度不变性,这是一个很大的缺陷.因为视觉传感器获得的图像除了旋转和光照变化之外,往往尺度都不具有一致性.

而 SIFT 算子充分考虑了在图像变换过程中出现的光照、尺度、旋转等变化.SIFT 算法在所有尺寸空间上通过高斯微分函数来识别出可能存在的尺度和旋转不变的特征点,使得 SIFT 具有尺度不变性.然后对 DoG 空间进行拟合处理,找到稳定的关键点的精确位置和尺寸.基于图像局部的梯度方向,分配给每个关键点位置方向信息,使得 SFIT 算子具有旋转不变性.此时关键点的位置、尺度和方向信息已确定,接下来需要描述符来描述关键点.其描述符根据图像局部梯度变换而来,这种表示允许比较大的局部形状的变形和光照变化.虽然 SIFT 有很多优点,但计算量极大,一般 SIFT 算法运用在不考虑计算成本的场景中.

SURF(Speeded Up Robust Features,加速稳健特征)在 SIFT 基础上进行改进,大大提升了运行速度.它采用了盒式滤波器来近似高斯滤波,对图像进行滤波之后,计算像素的黑塞(Hessian)矩阵行列式近似值,而盒式滤波器对图像的滤波转化成计算图像上不同区域间像素和的加减运算问题,只需要简单几次查找积分图就可以完成.SURF 节省了大量时间,兼顾了效果和精度.

Ke 等[15]通过对 SIFT 的描述子数据进行主成分分析,对数据进行了降维,最终也达到了加快算法的运行速度的目的.PCA-SIFT 构建了一个包含所有特征点和其描述子信息的特征矩阵,然后计算矩阵的协方差矩阵的特征向量,并选择前 n 个较大的特征向量构成投影矩阵,再把描述子向量与投影矩阵相乘即可降维.PCA-SIFT 对于旋转和光照有较好的不敏感性,但是由于 PCA-SIFT 不完全的仿射不变性,投影矩阵需要在特征比较明显的场景下才能起作用.

现阶段能够较为快速、稳定且准确地运用在视觉里程计上的是 ORB 算法,它充分考虑了 SLAM 系统需要的实时性、鲁棒性和准确性,为后端提供了较好的初始值.它采用了改进的 Fast 关键点检测,构建了图像金字塔,在每一个尺度层检测关键点,从而实现尺度不变性;特征的旋转不变性由灰度质心法实现.ORB 使用 BRIEF 描述子,它是一个二进制向量,是在提取关键点之后,在其邻域内选择 N 个点对比较像素大小,例如假设 $p_n(x_i,y_i)$,$q_n(x_j,y_j)$ 是某关键点邻域内的第 N 个点对. (x_i,y_i) 和 (x_j,y_j) 分别是 p_n,q_n 的坐标,若 $p_n(x_i,y_i)$ 的像素值小于 $q_n(x_j,y_j)$ 的像素值则取 0,否则为 1.经过 N 次比较后得到一个 N 维的描述子向量.

以上所述的都是传统的特征点法,它们由于环境因素导致的特征分布不均匀、纹理信息单一甚至是相机模糊等问题而提取不到足够的关键点,这种情况是普遍存在的,使得特征点法无法很好地运行,间接影响了后续位姿估计的精度.而线特征对于光照有着不敏感的特性,所以能够使得 VO 系统很好地适应弱纹理环境的场景.

Lu 等[17]提出了使用点线特征的视觉里程计,该算法吸收了直接法与特征点匹配法的优点,在纹理较少的环境中有不错的效果.通过增大算法的收敛域,该算法对于光照变化和快速运动的场景有更好的鲁棒性.在跟踪部分,同时处理点特征和线特征,点特征根据实际情况选用特征提取算法,比如 SURF;线特征跟踪部分,由于在针孔相机模型中,从世界坐标系中点或者线投影到相机成像平面的投影线始终保持直线,为了检测世界坐标系中的三维线,需要在对应的相机图像中检测它们的投影.所以使用 LSD [18]算法来提取线特征,效果如图 2 所示.

与点特征会有外点一样,线特征也会有离群点.RANSAC 算法在滤除离群点的同时检测三维线段的存在,对提取到的线特征的 MSLD 描述符[19]进行最近邻匹配,如图 3 所示.

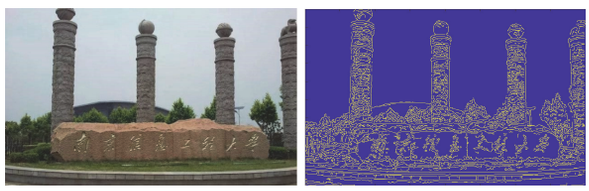

图 2　LSD 算法提取线特征效果

Fig. 2　Line feature extraction by LSD

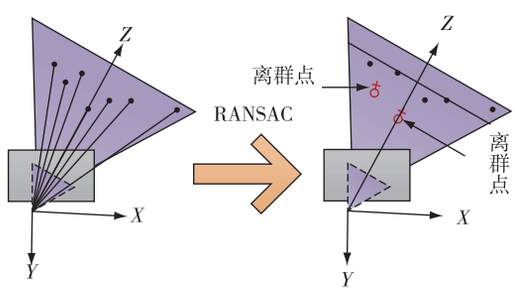

图 3　采样点选择示意图[19]

Fig. 3　Schematic diagram of sampling point selection[19]

虽然结合了线特征使得 VO 系统在复杂恶劣的环境中更加稳定、准确,但是同时也增加了计算量,会降低实时性.比如 Pumarola 等[20]提出了一种基于单目相机的融合点线特征的实时 PL-SLAM 系统,通过引入线特征提升了 ORB-SLAM 算法的精度,同时也增大了计算复杂度,尤其是在特征匹配阶段.PL-SLAM 在单目 ORB-SLAM 的基础进行改进,把 LSD 线段提取算法与 ORB 特征点提取算法融合,使得 ORB-SLAM 拥有了适应低纹理环境的能力,还提出了新的初始化策略,即在连续三帧图像中只能检测到线特征的情况下,估计出一个近似的初始化地图.

Gomez-Ojeda 等[21]则是通过点线特征的组合把 PL-SLAM 系统运用到立体视觉上,让线特征在视觉里程计系统上的使用更加泛化,再利用(Pseudo-Huber)损失函数来剔除误匹配的特征.最近文献[22]引入了与强角点即某些属性特别高的点相结合的边缘,提高具有很少或高频纹理的环境中的稳健性.Zhao 等[23]提出了一种由两个反投影平面的法线来表示线特征的参数化,从而使得线特征的重投影误差达到最小值,这种方法可以降低 PL-SLAM 系统对线特征的端点进行参数化造成计算冗余的负面影响.表 1 中列举了当前特征检测算法在旋转不变性、尺度不变性、光照不变性、可重复性、抗干扰性和计算效率几个方面的性能比较.

表 1　特征检测算法性能比较

Table 1　Performance comparison of feature detection algorithms

特征检测算法	类别	旋转不变性	尺度不变性	光照不变性	抗干扰性	计算效率
HARRIS[10]	角点	是	否	是	弱	弱
SUSAN[9]	角点	是	否	是	弱	较强
Shi-Tomasi[12]	角点	是	否	是	弱	较强
FAST[11]	人工特征	否	否	否	弱	强
SIFT[13]	人工特征	是	是	是	强	弱
SURF[14]	人工特征	是	是	否	较强	较强
ORB[16]	人工特征	是	是	是	较强	强
LSD[18]	线特征	是	是	是	强	较强

1.1.2　位姿估计

位姿估计是视觉里程计系统中的核心,也是其重要目标.位姿估计也就是通过分析相机与空间点的几何关系,从而计算出把 $K-1$ 时刻的相机位姿变换到 K 时刻相机位姿的变换矩阵 $\boldsymbol{T}_{k,k-1}$. 根据时间序列把相邻时刻的运动串联起来,这样就构成了机器人或者相机的运动轨迹.无监督学习 VO、VISO2-Mono 和 VISO2-Stereo 在 KITTI 数据集上的运动轨迹如图 4 所示.

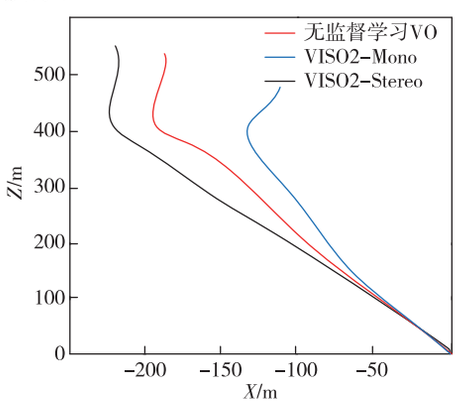

图 4　运动轨迹效果

Fig. 4　Motion track rendering

根据不同视觉传感器获得不同的图像信息而分成三种基本的运动估计计算方法.

1) ICP 方法.若相机能够通过某种方式获得深度信息如使用双目相机或者 RGB-D 相机,此时通常使用 ICP 算法(迭代最近点算法)[24]来解决.假设在相邻帧有一组匹配好的 3D 点 $P=\{p_1,p_2,\cdots,p_n\}$ 和 $P'=\{p'_1,p'_2,\cdots,p'_n\}$,位姿估计也就是想要找到一个旋转矩阵 $\boldsymbol{R}$ 和平移向量 $\boldsymbol{t}$ 使得 $\forall i,p_i=\boldsymbol{R}p'_i+\boldsymbol{t}$. 由此 3D-3D(3D 即三维图像的 3D 点)之间的位姿估计可以转换为求解最小化三维点之间误差的数学模型,即 $e_i=p_i-(\boldsymbol{R}p'_i+\boldsymbol{t})$.

现有的 ICP 求解方式分为两种,一种是线性代数的求解法比如奇异值分解法(SVD)[25],它可以分为三步:

①计算两组三维点的质心位置 q,q',定义两组点的质心为 $q=\frac{1}{n}\sum_{i=1}^{n}(p_i)$,$q'=\frac{1}{n}\sum_{i=1}^{n}(p'_i)$,然后计算各个点的去质心坐标 Q_i,Q'_i,即 $Q_i=p_i-q,Q'_i=p'_i-q'$.

②计算旋转矩阵 $\boldsymbol{R}^*$,估计质心坐标之间的最小欧式距离,即 $\boldsymbol{R}^*=\arg\min\frac{1}{2}\sum_{i=1}^{n}\|Q_i-\boldsymbol{R}Q'_i\|^2$.

③ 计算平移向量 $\boldsymbol{t}^*$,$\boldsymbol{t}^*=q-\boldsymbol{R}q'$.

另一种是非线性优化方法,通过迭代的方式找到最优值.它类似 Bundle Adjustment 方法,构建目标函数即式(1)之后,把相机位姿作为一个变量,不断迭代、更新、优化,得出一个最优的位姿:

$$\min_{\xi}=\frac{1}{2}\sum_{i=1}^{n}\|p_i-\exp(\boldsymbol{\xi}^{\wedge})p'_i\|_2^2\ , \tag{1}$$

其中在 $\boldsymbol{\xi}$ 右上角的倒三角符号 $\wedge$ 表示把 $\boldsymbol{\xi}$ 六维向量(前三维是平移向量,后三维是旋转向量)转换为一个四维矩阵.

2) 对极几何方法.若能获得的图像只有 2D 图像,如机器人使用单目相机而无法获得深度信息,此时使用对极几何方法解决.对极几何用在只知道匹配点的 2D 像素坐标的情况下,一般是机器人使用的相机传感器为单目相机.3D-3D 或者 3D-2D 问题都至少需要获得一组特征点是三维的,所以需要用至少两个单目相机或者能够获得深度的 RGB-D 相机,而解决 2D-2D 问题只需要一个单目相机,它以其低廉的价格在众多里程计方案中脱颖而出.为了探索 2D 点之间的几何关系,一般引入对极几何约束.如图 5 所示,p_{i-1},p_i 分别是图像 I_{i-1},I_i 中的由上述的特征匹配方法所得一个特征点,它们都是世界坐标

系中空间点 P 的投影.假设这是一次正确的匹配,其中 O_{i-1},O_i 是两个相机的光心,l_{i-1},l_i 分别是 I_{i-1},I_i 中的极线.

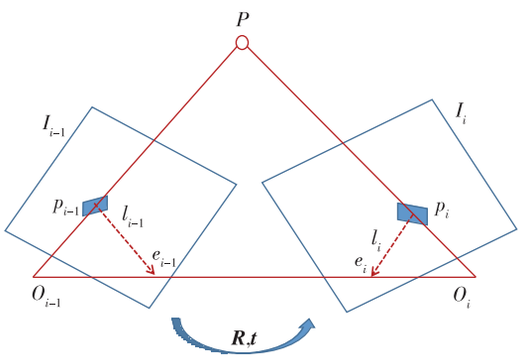

图 5　对极几何约束示意图

Fig. 5　Schematic diagram of polar geometric constraints

为了求解出它们之间的运动,即求解旋转矩阵 $\boldsymbol{R}$、平移向量 $\boldsymbol{t}$,引入对极几何约束,可以得出式(2),(3).其中 $\boldsymbol{t}^\wedge$ 表示向量 $\boldsymbol{t}$ 的反对称矩阵,上标 T 表示转置,$\boldsymbol{K}$ 为相机的内参矩阵.

$$\boldsymbol{p}_i^{\mathrm{T}}(\boldsymbol{K}^{-1})^{\mathrm{T}}\boldsymbol{t}^\wedge\boldsymbol{R}\boldsymbol{K}^{-1}\boldsymbol{p}_{i-1}=0, \tag{2}$$

$$\boldsymbol{E}=\boldsymbol{t}^\wedge\boldsymbol{R},\quad \boldsymbol{F}=(\boldsymbol{K}^{-1})^{\mathrm{T}}\boldsymbol{E}\boldsymbol{K}^{-1}, \tag{3}$$

从而可以从式(3)中的本质矩阵 $\boldsymbol{E}$、基础矩阵 $\boldsymbol{F}$ 解出 $\boldsymbol{R}$ 和 $\boldsymbol{t}$,常用八点法[26]或者复杂一点的五点法[27]来求解.如果相机画面中的特征点都落在同一平面上则需要单应矩阵来估计运动.

3) PnP 方法.PnP 方法用来解决相邻时刻仅有一个时刻的图像能获得深度信息的情况.PnP(Perspective-n-Point)是求解相邻两帧图像中特征点一帧是二维特征点而另一个是三维特征点的运动估计方法.PnP 求解的方式有很多种,其中 Moreno-Noguer 等[28]对此有很大的贡献.常用的解决方法有至少需要 6 对匹配点的直接线性变换(DLT),有只需要 3 对匹配点的 P3P,也有后续更为复杂的 EPnP 和 UPnP,还可以转化为非线性优化的方式,利用迭代法求解构建的最小二乘问题.其中 DLT 把旋转矩阵 $\boldsymbol{R}$ 和平移向量 $\boldsymbol{t}$ 定义成一个增广矩阵 $[\boldsymbol{R}|\boldsymbol{t}]$,根据空间点与其投影到相机成像平面对应的特征点之间的关系而求解位姿估计问题;P3P 则是利用给定的 3 个点之间形成的三角形相似性质来解决 3D-2D 位姿估计问题,把 2D 点转换成相机坐标系下的 3D 坐标,然后就变成了 3D-3D 的位姿估计问题,如图 6 所示.其中 O 为相机光心,A,B,C 分别为 3 个 3D 点,a,b,c 分别为 3 个 2D 点,L 为 3D 点的投影平面.

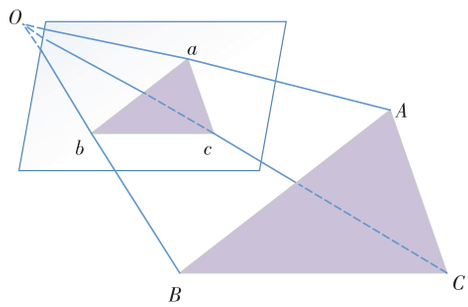

图 6　P3P 问题示意图

Fig. 6　Schematic diagram of P3P problem

1.2　直接法

虽然特征点法是主流方法,但是相比于直接法仍然有着很多缺点.比如特征点法需要十分耗时地提取特征,计算描述子的操作丢失了除了特征点以外的很多信息,而且更加适应低纹理信息场景.直接法不同于特征点法最小化重投影误差,而是通过最小化相邻帧之间的灰度误差估计相机运动,但是基于灰度不变假设:

$$I(x,y,z)=I(x+\Delta x,y+\Delta y,z+\Delta z). \tag{4}$$

例如假设空间有点 P 投影到相邻两帧图像上有 p_1,p_2 两点.它们的亮度分别为 $I_1(\boldsymbol{p}_{1,i})$ 和 $I_2(\boldsymbol{p}_{2,i})$,其中 i 表示当前图像中第 i 个点.则优化目标就是这两点的亮度误差 $\boldsymbol{e}_i$ 的二范数.此法可以应用在纹理信息较少、无法提取到足够的特征点的场景下,直接估计相机的运动.直接法示意如图 7 所示.

$$\begin{cases}\min\limits_{\xi}J(\boldsymbol{\xi})=\sum\limits_{i=1}^{N}\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{e}_i, & (5)\\ \boldsymbol{e}_i=I_1(\boldsymbol{p}_{1,i})-I_2(\boldsymbol{p}_{2,i}), & (6)\\ \boldsymbol{p}_{1,i}=\boldsymbol{T}\boldsymbol{P}_{2,i}, & (7)\\ \boldsymbol{T}=\exp(\boldsymbol{\xi}^\wedge), & (8)\end{cases}$$

其中 $\boldsymbol{T}$ 和 $\boldsymbol{\xi}$ 分别是 p_1,p_2 之间的转换矩阵及其李代数.式(8) $\boldsymbol{\xi}^\wedge$ 右上角的 $\wedge$ 表示把 $\boldsymbol{\xi}$ 转为一个四维矩阵,从而通过指数映射成为变换矩阵.

Ma 等[29]已经成功地把直接法用于 RGB-D 视觉传感器上.为了让计算量颇大的直接法能够实时地运行在单个 CPU 上,文献[30]提出了一种半稠密型深度滤波器公式,它能够大大降低计算复杂度,甚至还可以在智能手机上使用 AR 技术.LSD-SLAM[7]改进了传统直接法,将深度噪声加入到最小化光度误差的公式中得到较好的效果,它是单目直接法的标志性算法,是一种半稠密直接法.

直接法应用到完整的 V-SLAM 系统时,如果有恰当特征点的辅助,将会使得系统变得更加鲁棒和

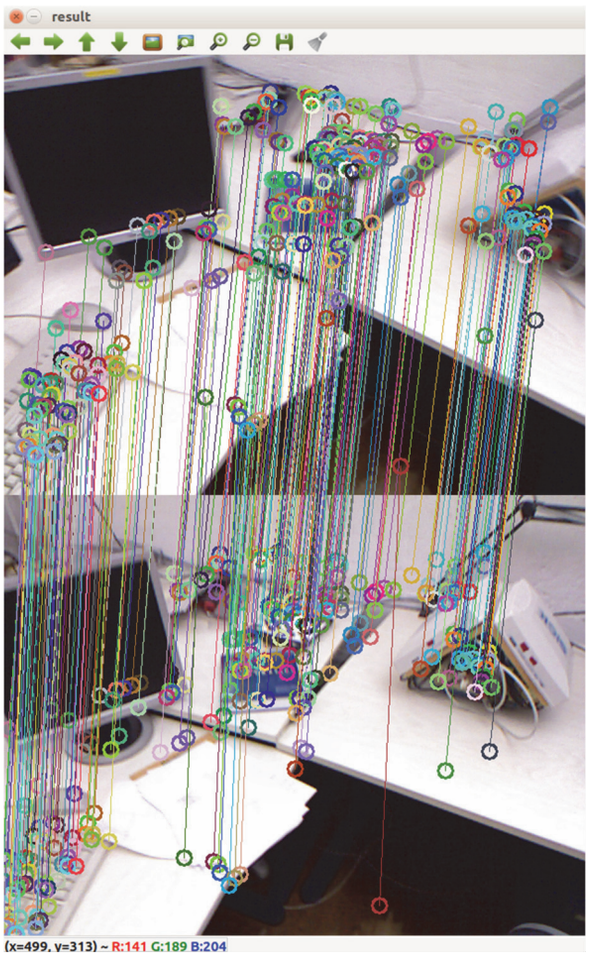

图 7　直接法示意图

Fig. 7　Schematic diagram of direct visual odometry

精准.如 Forster 等[31]提出一种半直接法 SVO(Semi direct monocular Visual Odometry),它结合了特征点和直接法,在追踪部分使用稀疏直接法对稀疏关键点获得粗略的位姿信息,并利用光流法来找到当前帧和地图点对应帧的像素块,优化后把关键帧地图点投影到当前帧.

2018 年,Zhang 等[32]提出的 DOVO 根据 ORB 特征获取的关键点数量和一个阈值 K 来评估使用 ORB 特征进行姿态估计的可靠性.如果关键点的数量小于阈值 K,则采用直接法保持摄像机的跟踪,并根据场景的光度优化光度误差来估计摄像机的姿态,否则通过优化重投影误差来计算姿态估计.实验结果表明,该方法保证了姿态的准确性和实时性.

Engel 等[33]提出一种纯使用直接法的视觉里程计 DSO(Direct Sparse Odometry).不同于传统的直接法,它将数据关联与位姿估计转换成一个统一的非线性优化问题.其第一创新点是通过光度标定改善由于相机参数改变引起的图像亮度变化问题,第二个创新点则是滑动窗口优化有效地控制了优化的计算量,又有良好的优化效果.实验结果表明,无论在跟踪精度还是鲁棒性方面,该方法在各种真实环境下都显著优于最先进的直接和间接方法.DSO 的出现将直接法的视觉里程计推上了一个新的高度.但是由于直接法相比于特征点法具有的非凸性,限制了 DSO 在处理视频时的效果.2019 年,Sun 等[34]提出的 FSMO(Fully Scaled Monocular direct sparse Odometry)基于 DSO 在原有的能量函数中增加了距离测量值,减少了直接法能量函数的非凸性带来的影响.这可以理解为式(5)的一种变形,即:

$$E_{\text{total}} = E_{\text{frame}} + \lambda \cdot E_{\text{dis}}, \tag{9}$$

其中 E_{frame} 表示光度误差,E_{dis} 是距离误差,λ 用来保持 E_{frame} 和 E_{dis} 在一个数量级上.

2　惯性视觉融合

不管特征点法还是直接法要准确地估计出图像之间的变换都需要消耗很大的计算量,所以实际应用中,为了易于工程的实现,一个机器人往往携带多种传感器.由于惯性传感器(IMU)能够在短时间内精确测量传感器的角速度和加速度,但是如果长时间应用累积误差严重.IMU 与相机传感器结合,称为视觉惯性里程计 VIO (Visual-Inertial Odometry),可以分为基于滤波和基于优化的两大类 VIO,也可以根据两个传感器数据应用的方法不同分为松耦合和紧耦合.松耦合是指 IMU 和相机分别进行位姿估计,紧耦合是指相机数据和 IMU 数据融合,共同构建运动方程和观测方程进行位姿估计.

现阶段基于非线性优化的方案有 VINS-Mono、OKVIS 等,还有基于滤波的紧耦合算法,它需要把相机图像特征向量加入到系统的状态向量中,使得状态向量的维度非常高,从而也会消耗更大的计算资源,MSCKF(Multi-State Constraint Kalman Filter)[35]和 ROVIO(RObust Visual Inertial Odometery)[36]是其中具有代表性的算法.传统的基于 EKF(扩展卡尔曼滤波)的视觉里程计与 IMU 数据融合时,EKF 的更新是基于单帧观测的,每个时刻的状态向量保存的是当前帧的位姿信息、速度、变换矩阵和 IMU 的误差等,使用 IMU 做预测步骤,视觉信息作为更新步骤.而 MSCKF 以类似滑动窗口(sliding window)的形式,使一个特征点在几个位姿都被观察到,从而建立约束,再进行滤波更新,它避免了仅依赖两帧相对的

运动估计带来的计算负担和信息损失,大大提高了收敛性和鲁棒性.图 8 为 MSCKF 的滑动窗口原理图.

ROVIO 是一种基于单目相机的 EKF 滤波 VIO,它直接优点是计算量小,但是需要根据设备型号调整到适合的参数,参数也是影响精度的重要因素.ROVIO 应用在 SLAM 系统中时没有闭环,也没有建图的线程,所以误差会有漂移.

针对基于滤波的松耦合,为了降低计算量,通过把图像信息当作一个黑盒,将 VO 的位姿估计结果与 IMU 数据进行融合,来减小状态向量维度.是一个很好的思路.

Weiss[37]在他的博士论文中详细介绍了视觉和 IMU 基于 EKF 的融合过程以及多传感器下构建的融合框架.其中,Ethzasl_SSF 和 Ethzasl_MSF 都是基于滤波的松耦合中优秀的开源算法.Ethzasl_SSF 主要是处理视觉与单个惯性传感器的融合问题,而 Ethzasl_MSF 提出了与多传感器融合框架,会使用深度相机、激光、IMU 等一系列传感器的数据来最终输出一个稳定的姿态.滤波器的状态向量是 24 维,如式(10),相比于紧耦合的方法精简很多.

$$x=\{p_w^i,v_w^i,q_w^i,b_w,b_a,\lambda,q_v^w,q_c^i,p_c^i\}\ , \tag{10}$$

其中除了不同坐标系变换的位置和四元数之外,还加入了陀螺仪 b_w 和加速度计 b_a 的偏差和,以及单目视觉尺度缩放的视觉比例因子 λ .

OKVIS (Open Key-frame-based Visual-Inertial SLAM)[38]和香港科技大学沈邵劼课题组的 VINS[39]是基于优化方法的 VIO 现阶段效果最好的算法.OKVIS 是基于关键帧优化的 VIO,它将视觉和 IMU 的误差项和状态量放在一起进行优化.在 VO 和 SLAM 中,通过最小化相机帧中观察到的地标的重投影误差来进行非线性优化以找到相机位姿和地标位置.图 9 上部为纯视觉 VO 示意图,下部为加上 IMU 后 VO 的示意图.IMU 对相邻两帧的位姿之间添加约束,而且对每一帧添加了状态量(陀螺仪和加速度计的偏差及速度).对于这样的新结构,文献[38]建立了一个包含重投影误差和 IMU 误差项的统一损失函数进行联合优化:

$$J(x)=\underbrace{\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j\in J(i,k)}\boldsymbol{e}_r^{i,j,kT}\boldsymbol{W}_r^{i,j,k}\boldsymbol{e}_r^{i,j,k}}_{\text{visual(视觉)}}+\underbrace{\sum_{k=1}^{K-1}\boldsymbol{e}_s^{kT}\boldsymbol{W}_s^k\boldsymbol{e}_s^k}_{\text{inertial(惯性传感器)}}\ , \tag{11}$$

其中,k 表示关键帧序列号,i 表示相机序列号,j 表示特征点的序列号,$\boldsymbol{W}_r^{i,j,k}$ 表示特征的信息矩阵,$\boldsymbol{W}_s^k$ 表示 IMU 误差的信息矩阵,而 $\boldsymbol{e}_r^{i,j,k}$ 为视觉重投影误差,$\boldsymbol{e}_s^k$ 为 IMU 误差项.

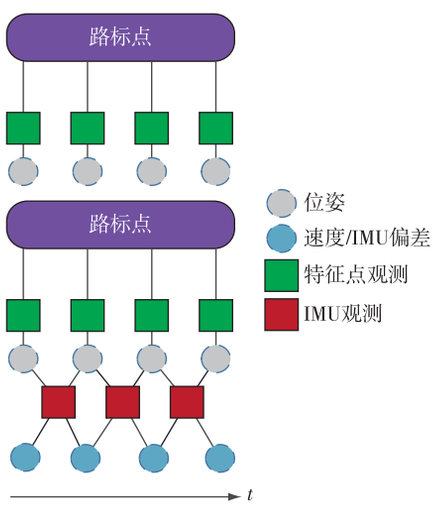

图 9 OKVIS 视觉与 IMU 的融合结构示意图

Fig. 9 Structure of OKVIS vision fused with IMU

IMU 误差项的实现和文献[40]一致,OKVIS 优化也是预积分的思路.OKVIS 将前后帧 IMU 测量值做积分,因为积分会用到 IMU 的偏差,而偏差是状态量,每次迭代时是变化的.所以每次迭代时会根据

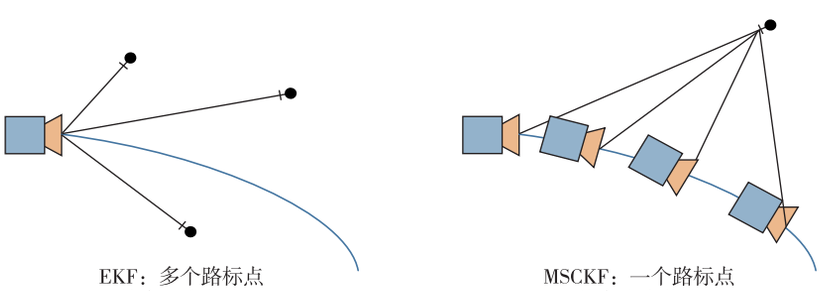

图 8 EKF 点约束状态与 MSCKF 点约束状态对比

Fig. 8 Comparison in point constraining between EKF and MSCKF

状态量相对于偏差的雅可比重新计算预积分值,当偏差变化太大时,不能再用雅可比近似计算预积分值,这时会根据 IMU 测量值重新进行积分.

VINS 也是类似的思路.VINS-Mono[39] 是 VINS 开源的单目视觉惯性 SLAM 方案,是基于滑动窗口优化实现的 VIO,使用 IMU 预积分构建紧耦合框架,是具备自动初始化、在线外参标定、重定位、闭环检测,以及全局位姿图优化功能的一套完整的 SLAM 系统.该算法的前端(VO)是 Harris 角点加 LK 光流跟踪,闭环检测添加了 BoW 词袋算法.VINS-Mono 主要设计用于状态估计和自主无人机的反馈控制,但它也能够为 AR 应用提供精确定位,与 ROS 完全集成.此外,团队还开源了 IOS 版本 VINS-Mobile[41],致力于部分 AR 的 APP 的研究.

3 基于深度学习的视觉里程计

除了与别的传感器进行融合这一思路之外,由于视觉里程计获得的都是图像信息,而深度学习在对图像识别、检测、分割方面的发展尤为迅速,从而为两者结合提供了良好的基础.深度学习与视觉里程计的多方面都有所结合,相比传统视觉里程计的人工特征,深度学习有着较为优秀的自动特征表征能力,且和直接法一样不需要计算描述子,也不需要传统视觉里程计复杂的工程,比如相机参数标定,各种传统的人工特征或者角点在面临光照变化、动态场景或者是图像纹理较为单一的情况时都有一定的敏感性,对于这些问题,基于深度学习的特征提取技术使得视觉里程计的性能有了一定的改善.用来解决位姿估计问题的深度学习技术大致分为监督学习和无监督学习两种.

监督学习网络中,最开始 Kendall 等[42] 提出 PoseNet,他们使用 CNN 粗略地预测出相机运动的速度和方向,使用 SFM 自动生成训练样本的标注,在没有大量数据标签情况下,通过迁移学习实现了输出较为精准的位姿信息.Costante 等[43] 用稠密光流代替 RGB 图像作为 CNN 的输入.该系统设计了三种不同的 CNN 架构用于 VO 的特征学习,实现了算法在图像模糊和曝光不足等条件下的鲁棒性.然而,同 PoseNet 一样,实验结果表明训练数据对于算法影响很大.当图像序列帧间运动较大时,算法误差很大,这主要是由于训练数据缺少高速训练样本.

目前效果较好的 DeepVO[44],网络结构如图 10 所示.它能够从序列原始图像直接映射出其对应的位姿,是基于 RCNN(递归卷积神经网络)的,它分为两个部分:首先通过卷积神经网络(CNN)学习图像的特征,然后通过深度递归神经网络学习(RNN)隐式地学习图像间的动力学关系及内在联系,运用各种大小的卷积核来更好地提取感兴趣特征使网络能够更好地提取各种特征.通过这种 CNN 网络学习得到的特征描述不仅将原始的高维 RGB 图像压缩成一个紧凑的描述,而且提高了后续的连续图像序列的训练效果.将 CNN 的最后一个卷积层(Conv6)提取到的特征传递给下一部分的 RNN,第二部分的 RNN 为了能够发现和利用图像之间的相关性,使用了一种 LSTM(Long Short-Term Memory)在较长的运动轨迹上实现这个目的.它能够随着时间的推移仍然保持隐藏状态的记忆,并且在隐藏状态之间存在反馈回路,使得当前的隐藏状态与之前状态存在的函数关系.因此它能够找出输入图像和姿态之间的联系.RNN 在 k 时刻的状态更新公式如下:

$$\boldsymbol{h}_k = H(\boldsymbol{W}_{xh}\boldsymbol{X}_k + \boldsymbol{W}_{hh}\boldsymbol{h}_{k-1} + \boldsymbol{b}_h)\ , \tag{12}$$

$$\boldsymbol{y}_k = \boldsymbol{W}_{hy}\boldsymbol{h}_k + \boldsymbol{b}_y\ , \tag{13}$$

其中 $\boldsymbol{h}_k$ 和 $\boldsymbol{y}_k$ 为 k 时刻的隐藏状态和输出,$\boldsymbol{W}$ 项表示相应的权矩阵,$\boldsymbol{b}$ 项表示偏置向量,H 为非线性激活函数(如 sigmoid).RNN 根据 CNN 生成的视觉特征,随着相机的移动和图像的获取在每一步输出一个 6 维姿态估计,包括位置信息和姿态信息.

还有一种不需要数据标签的无监督学习方法,一定程度上解决了监督学习由于缺少训练样本带来的问题.文献[45]通过无监学习的方式进行单一图像的深度估计,该方法采用双目数据集,通过多重目标损失训练网络产生视差图.通过模型的训练,该网络对单个图像深度估计达到高精度,超过了最先进的监督学习方法.Zhou 等[46] 提出了一种用于深度和相机运动估计的无监督深度学习框架,其深度预测和姿态估计结果较好.

2019 年,Liu 等[47] 创建了一种单目视觉里程计的无监督训练框架,每次位姿估计时无需根据实际值计算尺度因子,而是通过把单目图像和深度信息输入到训练网络来获得绝对尺度.他们的框架不需要相机真实的姿态来训练网络,实际姿态只用来评估系统的性能.该网络基于 RCNN 的框架,图像通过卷积层提取特征后,把特征输入到 LSTM 网络中分别输出旋转矩阵和平移向量.为了寻到最优参数,需要最小化损失函数,其描述了 RCNN 生成的姿态与预期结果之间的距离.利用 RCNN 得出的变换矩阵

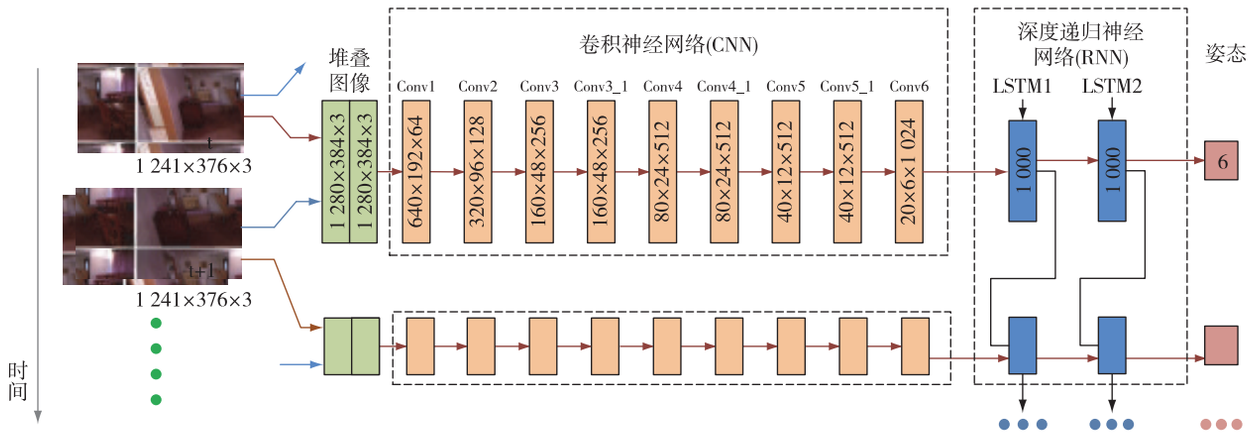

图 10　DeepVO 结构示意图[44]

Fig. 10　DeepVO structure[44]

和对应的点云值计算二维空间损失函数和三维空间损失函数.对于二维损失函数,首先通过变换矩阵和深度值把当前帧 I_t 的点投影到下一帧 I_{t+1} ,于是得到一个新帧 I'_t ,二维损失函数如式(14)所示.三维损失函数是把当前帧的点云 C_t 变换到相邻帧后得到 C'_t ,计算变换前后点云图的差距,如式(15)所示.最终通过加权系数 λ_{2D} 和 λ_{3D} 融合两个损失函数,如式(16)所示.

$$L_{2D}=\sum_{t=1}^{N-1}\overline{|I_t-I'_t|}-\overline{|I_{t+1}-I'_{t+1}|},\tag{14}$$

$$L_{3D}=\sum_{t=1}^{N-1}\overline{|C_t-C'_t|}-\overline{|C_{t+1}-C'_{t+1}|},\tag{15}$$

$$L=\lambda_{2D}L_{2D}+\lambda_{3D}L_{3D}.\tag{16}$$

现阶段基于深度学习的 VO 并不能取代传统的基于几何方法的 VO,而是一种可行的补充.因为深度学习提取得到的都是表象特征,但几何特征对 VO 是至关重要的.

4　标志性视觉里程计系统归纳

目前的基于传统视觉里程计框架的算法中,ORB-SLAM2[48]应对各种场景已经有了较好的鲁棒性和实时性,但是面对光照变化大、图像纹理信息较少的情况或者在动态场景中,传统 VO 框架无法很好地发挥作用.使用点线特征结合的 PL-SLAM 使得 ORB-SLAM2 不管是否缺少纹理特征信息都能精确地估计位姿.而当面对动态环境时本文综述了两个思路.一是不考虑计算成本只求精度的时候,可以使用 DeepVO、SFM-Net 等基于深度学习的高性能的视觉里程计,而有些问题,误差由于 VO 只用相机获取信息的途径而无法消除避免,多传感器融合可以很好的减小此类问题对定位带来的影响,例如 MSCKF.表 2 中分析了现有的 VO 算法的贡献和特点,包括视觉惯性传感器融合和基于深度学习的视觉里程计算法.表 3 中收集各类 VO 算法的开源代码.

5　总结与展望

本文对视觉里程计的三个模块即像素跟踪模块、外点排除模块和位姿估计模块进行了综述,介绍了近几年的视觉里程计算法.其中着重介绍了像素跟踪模块,包括传统的基于点特征法和直接法的视觉里程计(VO),还对比较新颖的线特征和线特征运用在 VO 系统中的优势进行了介绍.在外点排除和运动估计模块简略地介绍了相关理论知识.最后结合最新的算法详细介绍了当前较为火热的两个 VO 发展趋势,即以视觉惯性传感器融合为例的多传感器融合的 SLAM 前端算法,以及基于深度学习的视觉里程计.

挑战和机遇是一对“双胞胎”,VO 技术也是如此,面临挑战时往往会带来机遇.未来视觉里程计可能的发展趋势如下:

1)结合地图语义信息.由于环境中普遍存在动态场景造成的实际样本和检测样本之间误差降低了目前大部分的算法模型的位姿估计和轨迹的精度,通过结合语义地图的方式将从几何和语义两个方面来感知场景,使得应用对象对环境内容有抽象理解,获取更多的信息,从而来减小动态场景带来的误差,还可以为 SLAM 中的回环检测带来更多信息从而提高精度,但是计算成本会增加很多.适合通过高性能的计算设备用于实现精密地图构建、场景理解等功能的场合.

表 2　优秀 VO 算法的贡献和特点

Table 2　Contribution to and advantages of excellent VO algorithms

算法	VO 类型	贡献和创新
LSD-SLAM[30]	直接法	1)提出了一种基于相似变换空间李代数 sim(3)的直接跟踪方法,实现了对尺度漂移的显式检测; 2)使用一种概率方法在图像跟踪过程中减少了噪声对深度图信息的影响
DSO[33]	直接法	1)引入了光度标定的概念,建立了精细的相机成像模型; 2)结合了一个完全直接的概率模型(最小化光度误差)和所有模型参数的一致联合优化
ORB-SLAM2[48]	特征点法	1)首个基于单目、双目和 RGB-D 相机的开源方案; 2)使用具有视点不变性和光照不变性 ORB 特征; 3)根据平面和非平面选择相应模型,实现自动初始化
PL-SLAM[19]	点线特征法	1)允许同时处理点特征和线特征; 2)提出了一个新的基于线特征的地图初始化方法
MSCKF[35]	VIO	1)建立了一个能表示从多个相机姿态观察到的静态特征时产生的几何约束的测量模型; 2)计算复杂度仅取决于线性特征的数量
ROVIO[36]	VIO	1)提出了一种利用直接强度误差作为视觉测量的视觉惯性滤波框架; 2)在高度动态的手持设备和无人机上的应用取得到了较好的性能
DeepVO[44]	基于深度学习	1)通过卷积神经网络自动学习有效特征及其描述; 2)利用 RNN 对运动模型和数据关联模型进行隐式建模; 3)首个通过深度神经网络实现单目 VO 端到端的方法
Unsupervised DL VO[47]	基于深度学习	1)创建无监督的训练框架,省去复杂的标签工作; 2)无需相机姿态的真实值即能恢复绝对尺度

表 3　各类开源 VO 系统实现代码地址表

Table 3　Code addresses of various VO systems

算法	代码地址
ORB-SLAM2[48]	https://github.com/raulmur/ORB_SLAM2
LSD-SLAM[30]	https://github.com/tum-vision/lsd_slam
DSO[33]	https://github.com/JakobEngel/dso
PL-SLAM[19]	https://github.com/rubengooj/pl-slam
MSCKF[35]	https://github.com/KumarRobotics/msckf_vio
DeePVO[44]	https://github.com/ChiWeiHsiao/DeepVO-pytorch

2)多机器人协同的视觉里程计系统.单个机器人可能无法快速熟悉环境特征及其相对于环境特征的位置,也可能在执行任务的过程中损坏.为了稳定的精准导航,开发分布式系统来实现视觉里程计将是一个发展方向.使用多个机器人可以有很多优点,例如可以减少探索一个环境所需的时间、不同的信息来源将提供更多的信息、分布式系统对故障更健壮等.但是多个机器人 VO 的缺点就是必须将每个机器人生成的地图合并成一张全局地图,同时还需要自我定位与其他机器人协作.由于单个地图以及机器人之间的相对姿态的不确定性,使得地图合并变得更加困难.

参考文献

References

[1] Patra S,Aggarwal H,Arora H,et al.Computing egomotion with local loop closures for egocentric videos[C]//IEEE Winter Conference on Applications of Computer Vision (WACV),2017:454-463

[2] Wang S, Zhang Y, Zhu F. Monocular visual SLAM algorithm for autonomous vessel sailing in harbor area [C]//25th Saint Petersburg International Conference on Integrated Navigation Systems (ICINS),2018:1-7

[3] He H, Jia Y H, Sun L. Simultaneous location and map construction based on RBPF-SLAM algorithm[C]//Chinese Control and Decision Conference (CCDC),2018:4907-4910

[4] Moravec H P.Obstacle avoidance and navigation in the real world by a seeing robot rover[D].Palo Alto:Stanford University,1980

[5] 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

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

[7] Engel J,Schöps T,Cremers D.LSD-SLAM:large-scale direct monocular SLAM[C]//European Conference on Computer Vision,2014:834-849

[8] Zheng E H,Wang T T,Liu Z,et al.An adaptive visual odometry based on RGB-D sensors[C]//4th International Conference on Electromechanical Control Technology and Transportation (ICECTT),2019:154-158

[9] Smith S,Brady J.SUSAN:a new approach to low level image processing[J].International Journal of Computer Vision,1997,23(1):45-78

[10] Harris C, Stephens M. A combined corner and edge detector[J].Proc Alvey Vision Conf,1988,1988(3):147-151

[11] Rosten E,Drummond T.Machine learning for high-speed corner detection[M]//Computer Vision-ECCV2006.Berlin, Heidelberg: Springer Berlin Heidelberg, 2006:

430-443
[12] Shi J B,Tomasi C.Good features to track[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,2002:593-600
[13] Lowe D G.Distinctive image features from scale-invariant keypoints[J].International Journal of Computer Vision,2004,60(2):91-110
[14] Bay H, Tuytelaars T, van Gool L. SURF: speeded up robust features [M] // Computer Vision-ECCV 2006. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 404-417
[15] Ke Y,Sukthankar R.PCA-SIFT:a more distinctive representation for local image descriptors [C] // IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. DOI: 10. 1109/CVPR. 2004.1315206
[16] Rublee E,Rabaud V,Konolige K,et al.ORB:an efficient alternative to SIFT or SURF [C] // International Conference on Computer Vision,2011:2564-2571
[17] Lu Y,Song D Z.Robust RGB-D odometry using point and line features [C] // IEEE International Conference on Computer Vision (ICCV),2015:3934-3942
[18] von Gioi R G,Jakubowicz J,Morel J M,et al.LSD:a fast line segment detector with a false detection control[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2010,32(4):722-732
[19] Wang Z H,Wu F C,Hu Z Y.MSLD:a robust descriptor for line matching[J].Pattern Recognition,2009,42(5): 941-953
[20] Pumarola A,Vakhitov A,Agudo A,et al.PL-SLAM:Real-time monocular visual SLAM with points and lines[C]// IEEE International Conference on Robotics and Automation (ICRA),2017:4503-4508
[21] Gomez-Ojeda R,Moreno F A,Zuniga-Noel D,et al.PL-SLAM:a stereo SLAM system through the combination of points and line segments[J].IEEE Transactions on Robotics,2019,35(3):734-746
[22] Forster C,Zhang Z C,Gassner M,et al.SVO:semidirect visual odometry for monocular and multicamera systems [J]. IEEE Transactions on Robotics, 2017, 33 (2): 249-265
[23] Zhao L,Huang S D,Yan L,et al.A new feature parametrization for monocular SLAM using line features[J].Robotica,2015,33(3):513-536
[24] Milella A,Siegwart R.Stereo-based ego-motion estimation using pixel tracking and iterative closest point[C]// Fourth IEEE International Conference on Computer Vision Systems,2006:21-21
[25] Gong P L,Zhang Q F,Zhang A Q.Stereo vision based motion estimation for underwater vehicles [C] // 2009 Second International Conference on Intelligent Computation Technology and Automation,2009:745-749
[26] Lowe D G.Distinctive image features from scale-invariant keypoints[J].International Journal of Computer Vision,2004,60(2):91-110
[27] Parra I, Sotelo M A, Llorca D F, et al. Robust visual odometry for vehicle localization in urban environments [J].Robotica,2010,28(3):441-452
[28] Moreno-Noguer F, Lepetit V, Fua P. Accurate non-iterative O(n) solution to the PnP problem[C]//IEEE 11th International Conference on Computer Vision,2007: 1-8
[29] Ma L N,Kerl C,Stuckler J,et al.CPA-SLAM:Consistent-Plane-model Alignment for direct RGB-D SLAM[C]// IEEE International Conference on Robotics and Automation (ICRA),2016:1285-1291
[30] Engel J,Sturm J,Cremers D.Semi-dense visual odometry for a monocular camera[C]//IEEE International Conference on Computer Vision,2013:1449-1456
[31] Forster C,Pizzoli M,Scaramuzza D.SVO:fast semi-direct monocular visual odometry [C] // IEEE International Conference on Robotics and Automation (ICRA),2014: 15-22
[32] Zhang Z T, Wan W G. DOVO: mixed visual odometry based on direct method and orb feature [C] // International Conference on Audio, Language and Image Processing (ICALIP),2018:344-348
[33] Engel J, Koltun V, Cremers D. Direct sparse odometry [J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,40(3):611-625
[34] Sun J M,Wang Y Q,Shen Y Y.Fully scaled monocular direct sparse odometry with a distance constraint[C]// 5th International Conference on Control,Automation and Robotics (ICCAR),2019:271-275
[35] 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
[36] Bloesch M, Omari S, Hutter M, et al. Robust visual inertial odometry using a direct EKF-based approach[C] // IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),2015:298-304
[37] Weiss S M.Vision based navigation for micro helicopters [D].Zürich:ETH Zürich,2012
[38] Leutenegger S,Lynen S,Bosse M,et al.Keyframe-based visual-inertial odometry using nonlinear optimization[J]. The International Journal of Robotics Research,2015,34 (3):314-334
[39] Qin T, Li P L, Shen S J. VINS-mono: a robust and versatile monocular visual-inertial state estimator [J]. IEEE Transactions on Robotics,2018,34(4):1004-1020
[40] Forster C,Carlone L,Dellaert F,et al.IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation [C] // Robotics: Science and Systems, 2015.DOI:10.15607/RSS.2015.XI.006
[41] Li P L,Qin T,Hu B T,et al.Monocular visual-inertial state estimation for mobile augmented reality[C]//IEEE International Symposium on Mixed and Augmented Reality (ISMAR),2017:11-21
[42] Kendall A,Grimes M,Cipolla R.PoseNet:a convolutional network for real-time 6-DOF camera relocalization[C]// IEEE International Conference on Computer Vision (ICCV),2015:2938-2946
[43] Costante G,Ciarfuglia T A.LS-VO:learning dense optical subspace for robust visual odometry estimation[J].IEEE Robotics and Automation Letters,2018,3(3):1735-1742

[44] Wang S,Clark R,Wen H K,et al.DeepVO:towards end-to-end visual odometry with deep recurrent convolutional neural networks[C]//IEEE International Conference on Robotics and Automation (ICRA),2017:2043-2050
[45] Godard C,Aodha O M,Brostow G J.Unsupervised monocular depth estimation with left-right consistency[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2017:270-279
[46] Zhou T H,Brown M,Snavely N,et al.Unsupervised learning of depth and ego-motion from video[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2017:6612-6619
[47] Liu Q,Li R H,Hu H S,et al.Using unsupervised deep learning technique for monocular visual odometry[J].IEEE Access,2019,7:18076-18088
[48] 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

A survey of visual odometry

HU Kai[1,2] WU Jiasheng[1,2] ZHENG Fei [1,2] ZHANG Yanwen[1,2] CHEN Xuechao[1,2] LU Ben[1,2]
1 School of Automation,Nanjing University of Information Science & Technology,Nanjing 210044
2 Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET),Nanjing University of Information Science & Technology,Nanjing 210044

Abstract Visual Odometry (VO),which is an important part of visual simultaneous localization and mapping technology,is mainly used for robot pose estimation and local map building through time series images captured by camera sensors.Known as the front end,VO has been widely used in many fields and achieved fruitful practical results,and it is of great significance for unmanned driving,autonomous drones,virtual reality,and augmented reality,etc.In this paper,we summarize the recent research results on the novel visual odometry technology based on introduction of various algorithms in the framework module of classical VO.According to their technical means,the novel methods are divided into two categories,including VO integrated with multiple sensors (take VIO as an example),and VO based on deep learning.The former improves the accuracy of VO by complementing the advantages of various sensors,while the latter is combined with deep learning network.Finally,the existing algorithms of visual odometry are compared,and the future development trend is forecasted based on the challenges faced by VO.

Key words Visual Odometry (VO);multi-sensor fusion;deep learning