ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 3,May 2022,pp. 597－608
DOI:10. 16337/j. 1004-9037. 2022. 03. 010

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

# 基于几何-语义联合约束的动态环境视觉SLAM算法

沈晔湖[1]，陈嘉皓[1,2]，李　星[1]，蒋全胜[1]，谢　鸥[1]，牛雪梅[1]，朱其新[1]

(1. 苏州科技大学机械工程学院,苏州 215009;2. 北京工业大学人工智能与自动化学院,北京 100124)

**摘　要:** 传统视觉同步定位和地图构建(Simultaneous localization and mapping，SLAM)算法建立在静态环境假设的基础之上,当场景中出现动态物体时,会影响系统稳定性,造成位姿估计精度下降。现有方法大多基于概率统计和几何约束来减轻少量动态物体对视觉SLAM系统的影响,但是当场景中动态物体较多时,这些方法失效。针对这一问题,本文提出了一种将动态视觉SLAM算法与多目标跟踪算法相结合的方法。首先采用实例语义分割网络,结合几何约束,在有效地分离静态特征点和动态特征点的同时,进一步实现多目标跟踪,改善跟踪结果,并能够获得运动物体的轨迹和速度矢量信息,从而能够更好地为机器人自主导航提供决策信息。在KITTI数据集上的实验表明,该算法在动态场景中相较ORB-SLAM2算法精度提高了28%。

**关键词:** 几何约束;目标跟踪;机器视觉;视觉SLAM算法;实例语义分割

**中图分类号:** TP249　　**文献标志码:** A

## Dynamic Visual SLAM Based on Unified Geometric-Semantic Constraints

SHEN Yehu[1], CHEN Jiahao[1,2], LI Xing[1], JIANG Quansheng[1], XIE Ou[1], NIU Xuemei[1], ZHU Qixin[1]

(1. School of Mechanical Engineering, Suzhou University of Science and Technology, Suzhou 215009, China;2. College of Artificial Intelligence and Automation, Beijing University of Technology, Beijing 100124, China)

**Abstract:** Traditional visual simultaneous localization and mapping (SLAM) algorithms rely on the scene rigidity assumption. However, when dynamic objects exist in the scene, the stability of the SLAM system will be affected and the accuracy of pose estimation will be reduced. Currently, most of the existing methods apply probability strategies and geometric constraints to reduce the impact caused by a small number of dynamic objects. But when the number of dynamic objects in the scene is high, these methods will fail. In order to deal with this problem, a novel algorithm is proposed in this paper. It combines the dynamic visual SLAM algorithm with the multi-target tracking algorithm. Firstly, a semantic instance segmentation network together with geometric constraints is introduced to assist the visual SLAM module to effectively separate the static feature points from the dynamic ones, and at the same time, it can also achieve the better multi-target tracking performance. Furthermore, the trajectory and velocity information of the moving objects can also be estimated, which can provide decision information for autonomous robots navigation. The experimental results on KITTI dataset show that the localization accuracy of the proposed algorithm is improved by about 28% compared with ORB-SLAM2 algorithm in dynamic environments.

**Key words:** geometric constraints; target tracking; machine vision; visual SLAM algorithm; instance

**基金项目:** 国家自然科学基金(51975394, 61501451);江苏省高等学校大学生创新创业训练计划(201810332007Z)。
**收稿日期:** 2021-06-12;**修订日期:** 2021-11-10

semantic segmentation

## 引　言

同步定位和地图构建(Simultaneous localization and mapping,SLAM)的目标是使机器人能够在未知环境的移动过程中,完成自身定位和增量式地图构建[1]。传统的SLAM方法主要依赖于稳定性较好的距离传感器如激光雷达[2]。然而激光雷达获得的距离数据非常稀疏,这就造成SLAM构建得到的环境地图仅包含极少量的路标点。这个地图仅能被用来提高机器人的定位精度,无法用于路径规划等机器人导航的其他领域。此外激光雷达高昂的价格、较大的体积重量以及耗电量限制了其在某些领域的应用。因此,随着计算机、光电成像等技术的进步,有研究者提出了视觉SLAM[3]算法。相机在一定程度上克服了激光雷达在价格、体积、质量以及耗电量上的劣势,同时摄像机能够获取丰富的信息,但是相机也存在一些问题,例如对光线变化敏感,运算复杂度高等。为了克服单一传感器的缺点,研究者们提出了多传感器融合SLAM[4]算法,然而这类算法目前技术还不成熟,硬件成本较高,算法复杂。因此,目前视觉SLAM仍然是SLAM研究领域的一个重要方向。

现有的视觉SLAM算法大多基于环境静态假设,即场景是静态的,不存在相对运动的物体。但是在实际室外场景中大量存在行人、车辆等动态物体,从而限制了基于上述假设的SLAM系统在实际场景中运用。针对动态环境下视觉SLAM算法的定位精度和稳定性下降的问题,有学者提出了一些基于概率统计或者几何约束的算法,减少动态物体对视觉SLAM算法精度和稳定性的影响。例如当场景中存在少量动态物体时,可以使用RANSAC[5]等概率算法来剔除动态物体。但是当场景中出现大量动态物体时,上述算法将失效。因此,需要一种算法,能够在存在大量态度物体的场景中,有效地分离动态物体和静态物体。在现有的动态环境视觉SLAM算法中,对于动态物体的处理可以分为两类[6]:(1)剔除所有的动态特征点,只使用静态特征点,完成相机姿态估计和静态环境地图构建。(2)联合估计动态物体的运动,即同时估计相机的姿态、静态环境地图和场景中动态物体的运动。第一类方法未充分利用图像中所有有意义的信息,并且在动态物体较多的场景中难以使用。本文提出的算法属于第二类,该算法基于ORB-SLAM2[7],将动态视觉SLAM算法与多目标跟踪算法相结合。本文提出了几何-语义联合约束方法,实现了静态特征点和动态特征点的分离,并且估计出动态物体的运动速度。

## 1　相关工作

传统静态环境视觉SLAM算法主要分为两类:一类是基于特征点的算法;另一类则是基于直接法的算法。特征点法从图像中提取具有代表性的特征点,通过帧间特征点的匹配,完成相机位姿估计和地图构建。Davison等[8]提出的MonoSLAM是第1个实时单目视觉SLAM系统。前端提取非常稀疏的特征点,然后利用卡尔曼滤波器作为后端,跟踪前端提取到的特征点。Klein等[9]提出了PTAM(Parallel tracking and mapping),它使用非线性优化作为后端,并且将建图和跟踪过程并行化。Mur-Artal等[7]提出了ORB-SLAM2,它提出了一种同时使用双目远近点和单目观测的方案,并证明了后端采用光束法平差(BA)[10]具有更高的精度。Campos等[11]在ORB-SLAM2算法的基础上改进,提出了ORB-SLAM3算法,这是一个基于特征点法的视觉惯导紧耦合SLAM算法,并且能够重用先前重建的地图信息。

直接法不需要提取特征点,直接采用图像中像素的灰度信息,通过最小化光度误差估计相机运动。LSD-SLAM(Large scale direct monocular SLAM)是Engel等[12]提出的SLAM算法,它将直接法应用到半稠密单目SLAM中,在不计算特征点的前提下构建半稠密地图。Forster等[13]提出了SVO(Semi-di-

rect visual odometry)算法,将特征点法和直接法相结合,提取关键点,但是不计算描述子,利用关键点周围4×4的小块进行匹配,估计自身运动。然而实际环境中不可避免地存在动态物体。为了减轻动态物体对相机定位和地图重建的影响,近年来,动态环境视觉SLAM算法逐渐出现。在动态环境视觉SLAM算法中,分离动态物体和静态物体是一个关键的问题。然而由于遮挡、运动模糊或丢失跟踪特征而导致的噪声、异常值或特征对应的缺失,使这个问题更加复杂。现有的针对该问题的方法可以分为4类:

(1)概率统计法

在静态场景中,连续图像之间的特征点的转换可以由同一个运动模型来描述。在动态场景中,连续图像之间的特征点可能来自多个运动模型,每个运动模型都是独立且不相同的。基于概率统计的动态分离算法对数据的一个子集进行采样,并采用RANSAC[5]等算法拟合出内点最多的模型,从而实现特征点的分离。

(2)子空间聚类法

子空间聚类是基于高维数据可以低维的子空间联合表示的原理。数据点的子空间可以用基向量来进行数据的低维表示。在子空间聚类框架下的三维运动分割问题是找到与每个运动物体相对应的子空间,并将数据拟合到子空间中,例如文献[14-15]都使用了子空间聚类法来实现动态物体的分离,并且得到了较为不错的效果。

(3)几何法

几何法将多视几何从静态场景扩展到包含独立移动对象的动态场景。在静态场景中,1个基本矩阵描述了相机相对于静态场景的运动,但在动态环境中,将有$n$个基本矩阵来描述$n$个物体的运动,其中包括1个描述相机与静态场景相对运动的基本矩阵,通过几何约束来求解这些不同物体的基本矩阵的方法,称为几何法。Vidal等[16]在文献提出多体上极性约束来求解这个问题。

(4)深度学习法。

通过深度学习算法分离动态物体的思路分为两种:一种是利用深度学习算法直接将动态物体分离出来,该方法依赖于预定义的运动刚体数量,从三维点云或者光流中生成运动物体的掩膜;另一种则是通过一些先验知识,将有可能运动的物体分离出来,例如语义分割,全景分割等。

目前基于上述方法的动态环境SLAM算法,通常将分离出来的动态特征点直接剔除[17-18]。Bescos等[17]提出了Dynaslam算法,这是一种通过深度学习剔除动态特征点,基于静态特征点构建地图的视觉SLAM算法,支持单目、双目和RGB-D三种模式。Yu等[18]提出了DS-SLAM算法,利用SegNet[19]网络作为一个单独的线程进行语义分割,同时另一个线程进行特征点提取,利用语义分割得到的掩膜剔除动态特征点,此外还有一个线程构建稠密的八叉树地图。

在最新研究中,研究者们通过实验发现联合估计动态物体的运动,对整个SLAM系统的精度有增益效果,因而出现了一些联合估计算法[20-21]。Alcantarilla等[20]提出了一种联合视觉和稠密场景流分离动态物体的算法,能够在复杂动态环境中保证系统的稳定性。另有一些学者提出了将静态背景和动态物体进行联合估计的算法。Zhang等[21]提出了VDO-SLAM算法,该算法构建了1个包含机器人姿态、静态和动态三维点以及运动物体的统一框架,并且提出了一种能够避免因为遮挡引起的语义分割失败的处理方法,该算法复杂度较高,无法实现实时运行。Ballester等[22]提出了DOT算法,这仅是一个前端算法,通过语义分割掩膜分离动态和静态特征点,并且根据估计的相机运动,通过最小化光度重投影误差来跟踪这些对象。

近些年随着深度学习领域的迅速发展,出现了一些端到端的视觉SLAM算法。Yang等[23]提出了D3VO,将主要任务分为3个部分,分别利用深度学习对深度、姿态和不确定性进行估计。D3VO将这3

种估计结合起来，组成了一个包含前端跟踪和后端非线性优化的视觉里程计结构。Czarnowski等[24]提出了DeepFactors，DeepFactors是第1个概率稠密的SLAM算法，在概率因子图公式中，将几何知识的先验知识与经典SLAM公式相结合。

本文提出的算法相比传统的方法，在有较多动态物体的场景中具有更好的鲁棒性，提供了刚体速度的估计，并且由于算法复杂度低，因此运行速度较快，能够为导航提供更好的策略。

## 2 算法原理

### 2.1 算法流程

本文提出的算法首先完成特征点提取与匹配，利用实例语义分割网络对图像进行分割，并通过立体匹配得到场景的深度图；然后利用图像分割结果初步分离出静态特征点，并且计算基础矩阵。接着利用几何约束将动态特征点和静态特征点重新分离，估计相机位姿和构建地图，并且估计刚体运动速度，本文算法流程如图1所示。

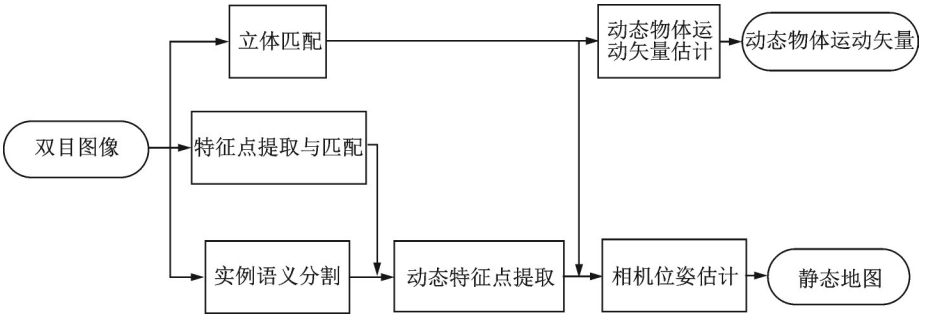



图1 算法流程图

Fig.1 Flow chart of the proposed algorithm

### 2.2 特征点提取与匹配

本文采用ORB特征点[25]。在提取特征之后，还需要根据描述子对前后帧图像进行特征匹配。本文计算一个特征点的描述子和其他描述子的汉明距离。汉明距离$d_{\text{hamming}}$定义为

$$d_{\text{hamming}}(i,j)=\sum_{k=1}^{N}\boldsymbol{X}_i^t(k)\oplus\boldsymbol{X}_j^{t+1}(k) \tag{1}$$

式中：$\boldsymbol{X}_i^t(k)$表示第$t$帧中第$i$个特征点的描述子中的第$k$位；$\boldsymbol{X}_j^{t+1}(k)$表示第$t+1$帧中第$j$个特征点的描述子中的第$k$位；$\oplus$表示异或操作；$N$为描述子矢量的维数。

当两个特征点同时满足式(2)和式(3)时，定义它们为匹配特征点对，有

$$j=\arg\min_{l} d_{\text{hamming}}(i,l)\qquad l=1,2,3,\cdots,n \tag{2}$$

$$i=\arg\min_{l} d_{\text{hamming}}(l,j)\qquad l=1,2,3,\cdots,n \tag{3}$$

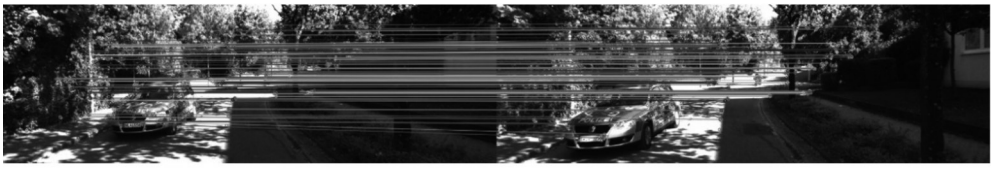
KITTI数据库[26]中特征匹配的例子如图2所示。

### 2.3 立体匹配

场景的深度信息在视觉SLAM系统中是一个非常重要的信息。本文采用SGBM(Semi-global block matching)立体匹配算法来获取深度信息[27]。

首先构建全局能量函数

(a) Consecutive images in Seq 02 of KITTI (The left image is the previous one, and the right image is the next one)



(b) Feature points matching result of Fig.2(a)

图2 特征匹配结果示例

Fig. 2 Examples of feature matching result

$$E(\boldsymbol{D})=\sum_{\boldsymbol{u}_p}(c(\boldsymbol{u}_p,\boldsymbol{D}(\boldsymbol{u}_p)))+\left(\sum_{\boldsymbol{u}_q\in\boldsymbol{N}_{\boldsymbol{u}_p}} l_1\delta(|\boldsymbol{D}(\boldsymbol{u}_p)-\boldsymbol{D}(\boldsymbol{u}_q)|=1)\right)+\left(\sum_{\boldsymbol{u}_q\in\boldsymbol{N}_{\boldsymbol{u}_p}} l_2\delta(|\boldsymbol{D}(\boldsymbol{u}_p)-\boldsymbol{D}(\boldsymbol{u}_q)|>1)\right) \tag{4}$$

式中:$\boldsymbol{D}$为视差图,$\boldsymbol{u}_p$,$\boldsymbol{u}_q$代表图像中的像素点$p$和$q$的坐标;$\boldsymbol{N}_{\boldsymbol{u}_p}$指像素点$\boldsymbol{u}_p$的相邻像素组成的集合;$c\left(\boldsymbol{u}_p,D\left(\boldsymbol{u}_p\right)\right)$为当前视差图$\boldsymbol{D}$中像素点$\boldsymbol{u}_p$匹配代价;$l_1$和$l_2$为惩罚系数;$\delta(\cdot)$为示性函数。

求解上述函数最优值是一个NP完全问题,因此将此问题近似分解成为一个线性问题进行求解,图3为图2(a)中左图对应的视差图。

图3 图2(a)中左图对应的视差图

Fig.3 Disparity map corresponding to the left image in Fig.2(a)

### 2.4 实例语义分割

本文采用Mask $R$-CNN网络[28],这是一个实例语义分割网络,能够有效地检测图像中的目标。为了针对多个动态物体的场景,本文提出了一种改进的Mask $R$-CNN网络输出结构设计。该改进的结构输出1张与输入图像具有相同大小的RGB图像。其中,$R$通道设计为类别通道,当检测到行人时,行人掩码范围内$R$通道置为64,检测到车辆时置为128,检测到非机动车时置为255,其余默认为0;$G$通道设计为实例通道,每个实例掩码值是该实例号与16的乘积。理论上最大识别类别为3,每种实例数量为16个,基本能够满足使用需求。当输入图像为图2(a)时,修正后的实例语义分割结果如图4所示。

### 2.5 动态特征点提取

在动态场景中,动态特征点影响相机位姿估计,造成视觉SLAM系统精度和稳定性下降,为了解决上述问题本文提出了一种迭代动态特征点提取算法。

图4 实例语义分割结果

Fig.4 Output of instance semantic segmentation

首先,利用2.4节的实例语义分割获得的动态对象实

例的掩膜,对所提取到的特征点进行一次筛选,即

$$\boldsymbol{U}_t^{\mathrm{ASP}}=\left\{\boldsymbol{u}_t^i \middle| \boldsymbol{u}_t^i\in\boldsymbol{U}_t,\ \boldsymbol{u}_t^i\notin\left(\boldsymbol{M}_t^1\cup\boldsymbol{M}_t^2\cup\cdots\cup\boldsymbol{M}_t^m\right)\right\} \tag{5}$$

式中:$\boldsymbol{U}_t^{\mathrm{ASP}}$为绝对静态特征点的像素坐标的集合;$\boldsymbol{u}_t^i$为第$t$帧中第$i$个绝对静态特征点的像素坐标;$\boldsymbol{U}_t$为第$t$帧的特征点的像素坐标的集合;$\boldsymbol{M}_t^m$为第$t$帧第$m$个实例的掩膜区域的集合。

当特征点符合对极约束时,认为该特征点为静态候选;当特征点不满足对极约束时,认为该特征点为动态候选,对极约束公式[29]为

$$\boldsymbol{u}_t^{i\mathrm{T}}\boldsymbol{K}^{-\mathrm{T}}\hat{\boldsymbol{t}}\boldsymbol{R}\boldsymbol{K}^{-1}\boldsymbol{u}_{t+1}^j=0 \tag{6}$$

即

$$\boldsymbol{u}_t^{i\mathrm{T}}\boldsymbol{F}\boldsymbol{u}_{t+1}^j=0 \tag{7}$$

式中:$\boldsymbol{u}_t^i$表示第$t$帧第$i$个特征点的像素坐标,$\boldsymbol{F}=\boldsymbol{K}^{-\mathrm{T}}\hat{\boldsymbol{t}}\boldsymbol{R}\boldsymbol{K}^{-1}$为基础矩阵;$\boldsymbol{u}_{t+1}^j$为第$t+1$帧与$\boldsymbol{u}_t^i$匹配的特征点的像素坐标。

利用$\boldsymbol{U}_t^{\mathrm{ASP}}$和$\boldsymbol{U}_{t+1}^{\mathrm{ASP}}$通过八点法[30]计算基础矩阵$\boldsymbol{F}$,然后可以对$\boldsymbol{U}_t$重新进行静态特征点和动态特征点的分离,当特征点同时满足以下3个条件时,该特征点被视为是动态特征点:(1) $\boldsymbol{u}_t^i\in\boldsymbol{M}_t^s$, $s=1,2,3,\cdots,n$;(2) $\boldsymbol{u}_t^{i\mathrm{T}}\boldsymbol{F}\boldsymbol{u}_{t+1}^j\geqslant N_{\mathrm{d}}$;(3) $\mathrm{card}\left(\boldsymbol{U}_t\cap\boldsymbol{M}_t^m\right)\geqslant N_{\mathrm{mat}}$。其中card()为集合的势,$N_{\mathrm{d}}$为动态阈值,$N_{\mathrm{mat}}$为匹配阈值。其余特征点都视为静态特征点,构成静态特征集合$\boldsymbol{S}_t$,动态特征点提取结果如图5所示,图中白色点为动态特征点,黑色点为静态特征点。

图5 动态特征提取结果

Fig. 5 Result of dynamic feature extraction

### 2.6 相机位姿估计

对于静态特征集合$\boldsymbol{S}_t$结合相机的成像原理可得

$$\lambda\,{}^{S}\boldsymbol{u}_t^i=\boldsymbol{K}\boldsymbol{T}\boldsymbol{p}_t^i \tag{8}$$

式中:$\lambda$表示尺度因子;$\boldsymbol{K}$为相机内参矩阵;$T\in\ SE(3)$为位姿变换矩阵;${}^{S}\boldsymbol{u}_t^i$为静态特征集合$\boldsymbol{S}_t$中的第$i$个元素;$\boldsymbol{p}_t^i$为与${}^{S}\boldsymbol{u}_t^i$对应特征点的三维空间坐标。

构建重投影误差并优化为

$$E(\boldsymbol{T})=\arg\min_{\boldsymbol{T}}\frac{1}{2}\sum_{i=1}^{n}\left\|{}^{S}\boldsymbol{u}_t^i-\frac{1}{s_t^i}\boldsymbol{K}\boldsymbol{T}\boldsymbol{p}_t^i\right\|_2^2 \tag{9}$$

PnP[31]可以在很少的匹配点中获得比较好的估计,因此本文选择PnP估计姿态,通过最小化式(9)的重投影误差来求解PnP,然后用Levenberg-Marquadt算法[32]进行优化。

### 2.7 动态物体运动矢量估计

本文2.4节虽然完成了实例分割,但是并没有建立相邻帧之间对应实例的联系,为此本文提出了一种简易的帧间物体匹配算法,具体步骤如下。

首先构建实例特征点集合,即

$$\boldsymbol{D}_t^m=\left\{\boldsymbol{u}_t^i \middle| \boldsymbol{u}_t^i\in\boldsymbol{U}_t^D\cap\boldsymbol{M}_t^m\right\} \tag{10}$$

式中:$\boldsymbol{D}_t^m$表示在第$t$帧属于第$m$实例的动态特征点的像素坐标集合;$\boldsymbol{U}_t^D$表示2.5节中提取动态特征点的像素坐标集合。

结合$\boldsymbol{D}_t^m$和2.2节得到的帧间特征点匹配结果,计算得到$t+1$帧中与$\boldsymbol{D}_t^m$匹配的特征点的像素坐标集合$\hat{\boldsymbol{D}}_{t+1}^m$,当满足式(11)时,$\alpha$和$\beta$是同一个实例。

$$\mathrm{card}\left(\boldsymbol{D}_{t+1}^{\alpha}\cap\vec{\boldsymbol{D}}_{t+1}^{\beta}\right)\geqslant N_{\mathrm{mat}} \tag{11}$$

由于每个实例中的特征点数量无法满足后续实例跟踪的需求,所以需要在图像实例区域内提升特征点的数量。本文使用基于Lucas-Kanade光流[33]的反向增量光流金字塔实现上述工作。优化目标如下

$$E(\boldsymbol{p})=\sum_{\omega\times\omega}\left[\boldsymbol{I}\left(\omega(\boldsymbol{x};\boldsymbol{p})\right)-T(\boldsymbol{x})\right]^2 \tag{12}$$

上述优化目标通过迭代法进行优化,即在图像$\boldsymbol{I}$中对原始参数$\boldsymbol{p}$上进行增量$\Delta\boldsymbol{p}$映射运算,迭代优化如下目标函数为

$$E(\Delta\boldsymbol{p})=\sum_{\omega\times\omega}\left[\boldsymbol{I}\left(\omega(\boldsymbol{x};\boldsymbol{p}+\Delta\boldsymbol{p})\right)-T(\boldsymbol{x})\right]^2 \tag{13}$$

通过对每个实例进行光流跟踪算法,得到每个实例像素间的对应关系,类似式(9)构造重投影误差函数,通过Levenberg-Marquadt算法优化[32],可得动态物体与相机之间的相对位姿$\boldsymbol{T}^m\in SE(3)$。$\boldsymbol{T}^m$包含了一个旋转矩阵$\boldsymbol{R}$和一个平移向量$\boldsymbol{t}$,取$\boldsymbol{T}^m$中最后1列前3行为平移向量$\boldsymbol{t}$,取$\boldsymbol{T}^m$三阶顺序主子式为旋转矩阵$\boldsymbol{R}=\begin{bmatrix}R_{11}&R_{12}&R_{13}\\R_{21}&R_{22}&R_{23}\\R_{31}&R_{32}&R_{33}\end{bmatrix}$,根据欧拉角与旋转矩阵转换公式

$$\begin{cases}\theta_x=\arctan 2\left(R_{32},R_{33}\right)\\\theta_y=\arctan 2\left(-R_{31},\sqrt{{R_{32}}^2+{R_{33}}^2}\right)\\\theta_z=\arctan 2\left(R_{21},R_{11}\right)\end{cases} \tag{14}$$

式中:$\theta_x$、$\theta_y$、$\theta_z$分别表示滚转角、俯仰角和偏航角,由于本文实验数据为室外城市道路场景,因此只保留偏航角。从而动态物体的运动矢量为$\left[\theta_z,\boldsymbol{t}\right]$,动态物体的运动矢量估计如图6所示。

图6 动态物体运动矢量估计

Fig.6 Motion estimation of dynamic objects

## 3 实验结果

### 3.1 度量标准

本文实验采用绝对轨迹误差(Absolute trajectory error,ATE)和相对位姿误差(Relatire pose error,RPE)[34]作为度量标准,ATE的度量公式为

$$\mathrm{RMSE}(\boldsymbol{E}_{\mathrm{AT}})=\left(\frac{1}{M_A}\sum_{t=1}^{M_A}\left\|\boldsymbol{E}_{\mathrm{AT}}^{t}\right\|^2\right)^{\frac{1}{2}} \tag{15}$$

式中:$\boldsymbol{E}_{\mathrm{AT}}^{t}$为第$t$帧的绝对轨迹误差;$M_A$为输入双目图像的总数目。

RPE的度量公式为

$$\mathrm{RMSE}(\boldsymbol{E}_{\mathrm{RP}})=\left(\frac{1}{M_R}\sum_{i=1}^{M_R}\left\|\boldsymbol{E}_{\mathrm{RP}}^{t}\right\|^2\right)^{\frac{1}{2}} \tag{16}$$

式中:$\boldsymbol{E}_{\mathrm{RP}}^{t}$为第$t$帧和第$t+1$帧对位姿误差;$M_R$为计算所得相位姿误差的总数。

### 3.2 数据集和实验平台

本文在KITTI odometry数据集和KTTI tracking数据集[26]上对提出的算法进行测试。KITTI

odometry数据集每一组都有彩色双目图像,以及对应的灰度图和雷达数据,其中大部分采集于高速公路上,一般用于里程计和SLAM算法的测试。KTTI tracking数据集主要采集于城市道路中,图像内包含较多的动态物体,一般应用于图像语义分割测试,数据集内除了包含彩色双目图像外,还有图像语义信息以及GPS/IMU传感器信息。本文采用的实验平台配置如表1所示。

**表1 实验平台配置**

**Table 1 Experimental platform**

| 类别 | 具体配置 |
|---|---|
| CPU | i9-9900K |
| GPU | RTX 2080Ti |
| 内存 | 32 GB |
| 操作系统 | Ubuntu 18.04 |
| 开发语言 | C++ |

### 3.3 实验结果与分析

#### 3.3.1 定性分析结果

将ORB-SLAM2和本文提出算法在KITTI odometry数据集和KITTI tracking数据集进行测试,两者在KITTI tracking数据集上的轨迹对比结果如图7所示。在图7中,灰色虚线表示相机自运动真实轨迹,蓝色实线表示ORB-SLAM2估计得到的相机自运动轨迹,红色实线表示本文所提算法估计得到轨迹。对比本文所提算法和ORB-SLAM2发现,本文所提算法在KITTI tracking数据集上均比ORB-SLAM2接近真实轨迹。图7中本文所提算法对比ORB-SLAM2在00、03、04序列平均定位误差分别降低了0.03、0.15和0.14 m,这证明了本文将动态物体进行单独处理对定位的有益效果。

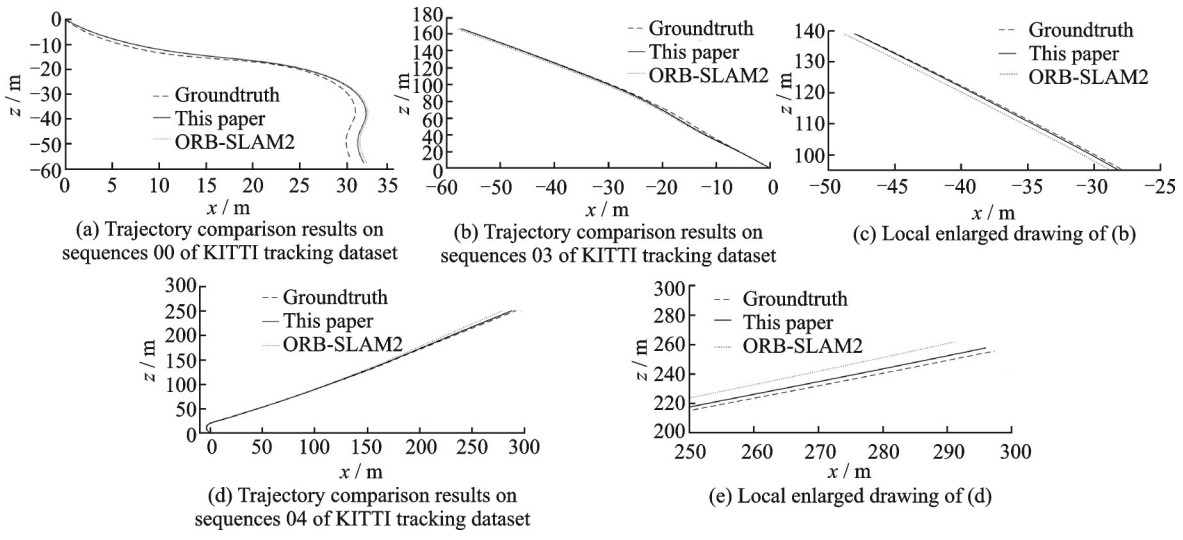



图7 本文算法和ORB-SLAM2在KITTI tracking数据集00、03、04序列轨迹对比

Fig.7 Trajectory comparison results of our proposed algorithm and ORB-SLAM2 on sequences 00, 03 and 04 of KITTI tracking dataset

#### 3.3.2 定量分析结果

本文提出算法、ORB-SLAM2算法和Dynaslam算法在KITTI odometry数据集01-10序列上的结果如表2所示。从表2中可以看出在KITTI odometry数据集01序列中,本文算法对比ORB-SLAM2和Dynaslam算法有较大的提升,然而在02和03序列中提升并不明显。这是由于01序列的场景为高速公路,该场景中包含较多的运动物体,因此ORB-SLAM2算法在该序列下绝对轨迹误差较大,Dynaslam算法使用语义分割网络对动态物体进行了剔除,相对ORB-SLAM2算法绝对轨迹误差较小,而本文算法在该序列下表现更佳。02序列和03序列的场景为城市道路,运动物体较少,因此本文算法和上述两种算法差距不大。

表3为本文算法和ORB-SLAM2算法在KITTI tracking数据集的对比结果,通过对比发现本文算法在

表2 本文算法、ORB-SLAM2和Dynaslam在KITTI odometry数据集上对比结果

Table 2 Comparison results of proposed algorithm, ORB-SLAM2 and Dynaslam on the KITTI odometry dataset m

| 序列编号 | ATE | | |
|---|---|---|---|
| | ORB-SLAM2[7] | Dynaslam[17] | 本文算法 |
| 00 | **1.3** | 1.4 | **1.3** |
| 01 | 10.1 | 9.4 | **6.1** |
| 02 | 5.6 | 6.7 | **4.4** |
| 03 | **0.6** | **0.6** | **0.6** |
| 04 | **0.2** | **0.2** | **0.2** |
| 05 | 0.8 | 0.8 | **0.7** |
| 06 | 0.8 | 0.8 | **0.7** |
| 07 | 0.5 | 0.5 | **0.4** |
| 08 | 3.6 | 3.5 | **2.8** |
| 09 | 3.2 | **1.6** | **1.7** |
| 10 | **1.0** | 1.2 | **1.0** |
| 均值 | 2.52 | 2.42 | **1.81** |

表3 本文算法和ORB-SLAM2在KITTI tracking数据集上对比结果

Table 3 Comparison results of our proposed algorithm and ORB-SLAM2 on the KITTI tracking dataset m

| 序列编号 | ATE | |
|---|---|---|
| | ORB-SLAM[7] | 本文算法 |
| 00 | 1.32 | **1.29** |
| 01 | 1.95 | **1.89** |
| 02 | 0.95 | **0.92** |
| 03 | 0.74 | **0.59** |
| 04 | 1.44 | **1.30** |
| 05 | **1.23** | **1.23** |
| 06 | **0.19** | **0.19** |
| 均值 | 1.117 | **1.059** |

KITTI tracking数据集00-06序列中也优于ORB-SLAM2,这一结果与3.3.1节中定性分析结果一致。

本文算法还与目前典型的动态环境下的SLAM算法:VDO-SLAM[21]在KITTI tracking数据集上进行了对比,结果如表4所示。$RPE_t$为平移部分的相对位置误差,单位为米/帧(m/f);$RPE_r$为旋转部分的相对位置误差,单位为度/帧((°)/f)。从表4可以看出,本文算法的平移位姿误差优于VDO-SLAM算法,旋转位姿误差略大于VDO-SLAM算法,总体来说两者定位精度接近。

本文还比较了本文算法在KITTI tracking数据集上计算速度,如表5所示。和VDO-SLAM算法类似,由于实例语义分割和立体匹配均可以单独的线程形式在GPU上实时运行,因此只需测试特征点提取匹配以及相机/动态物体运动估计的时间。由表5可知,本文算法速度远快于VDO-SLAM算法,帧率为VDO-SLAM的近3倍,基本达到了实时运行,在性能和速度上找到了较好的平衡点。

3.3.3 消融实验

本文所提算法基于几何-语义联合约束,主要包括几何约束模块和语义分割模块。对本文所提算法进行了消融实验,实验结果如表6所示。通过表6单独使用几何约束模块或者语义分割模块,定位误差均远大于本文提出算法,这表明本文提出算法结合使用几何约束和语义分割的重要性。

**表4 本文算法和VDO-SLAM在KITTI tracking数据集上对比结果**

**Table 4 Egomotion comparison with VDO-SLAM on the KITTI tracking dataset**

| 序列编号 | VDO-SLAM[21] | | 本文算法 | |
|---|---|---|---|---|
| | $RPE_t$/(m•f$^{-1}$) | $RPE_r$/((°)•f$^{-1}$) | $RPE_t$/(m•f$^{-1}$) | $RPE_r$/((°)•f$^{-1}$) |
| 00 | 0.05 | **0.05** | **0.04** | 0.06 |
| 01 | 0.12 | **0.04** | **0.07** | **0.04** |
| 02 | **0.04** | **0.02** | 0.04 | 0.03 |
| 03 | 0.09 | **0.04** | **0.06** | **0.04** |
| 04 | 0.11 | **0.05** | **0.07** | 0.06 |
| 05 | 0.10 | **0.02** | **0.06** | 0.03 |
| 06 | **0.02** | 0.05 | **0.02** | **0.04** |
| 均值 | 0.075 | **0.038** | **0.053** | 0.042 |

**表5 本文算法和VDO-SLAM计算速度对比**

**Table 5 Runtime comparison of the proposed algorithm and VDO-SLAM**

| 类别 | VDO-SLAM[21] | 本文算法 |
|---|---|---|
| 平均帧率 | 5.4 | 15.7 |

**表6 消融实验结果**

**Table 6 Results of ablation study** m

| 序列编号 | ATM | |
|---|---|---|
| | 不使用几何约束模块 | 不使用语义分割模块 |
| 00 | 1.4 | 1.4 |
| 01 | 9.3 | 9.8 |
| 02 | 6.8 | 5.7 |
| 03 | 0.7 | 0.7 |
| 04 | 0.2 | 0.3 |
| 05 | 0.8 | 0.8 |
| 06 | 0.8 | 0.8 |
| 07 | 0.5 | 0.5 |
| 08 | 3.5 | 3.1 |
| 09 | 1.7 | 2.8 |
| 10 | 1.0 | 1.0 |
| 均值 | 2.43 | 2.45 |

## 4 结束语

本文所提的算法采用实例语义分割网络,结合几何约束,在有效分离静态特征点和动态特征点的同时,能够实现多目标跟踪,获得动态物体的运动矢量信息。与ORB-SLAM2算法相比,在存在较多动态物体的场景中,本文所提算法具有更高的稳定性和准确度,在KITTI odometry数据集上准确度提高了28%。与VDO-SLAM算法相比,由于本文所提算法复杂度低,因此具有更高的计算速度,约为其3倍,并且能获得接近的定位精度。由于本文算法基于实例语义分割网络和几何约束,因此算法精度容易受到实例语义分割网络精度的影响,并且实例语义分割网络的执行速度较慢也是限制本文算法实际应用的问题之一。上述算法中存在的不足,将在以后的工作中进一步完善。在以后的工作中将使用精

度更高、速度更快的语义分割网络,并且采用选取关键帧的策略,仅在关键帧中进行语义分割,从而进一步提高运行速率。

## 参考文献


[1] LEONARD J, DURRANT-WHYTE F. Simultaneous map building and localization for an autonomous mobile robot[C]// Proceedings IROS'91: IEEE/RSJ International Workshop on Intelligent Robots and Systems'91. Osaka, Japan: IEEE, 1991: 1442-1447.

[2] THRUN S, BURGARD W, FOX D.概率机器人[M].曹红玉,谭志,史晓霞,译.北京:机械工业出版社,2017: 115.
THRUN S, BURGARD W, FOX D. Probabilistic robotics[M]. CAO Hongyu, TAN Zhi, SHI Xiaoxia, translate. Beijing: China Machine Press, 2017 : 115.

[3] 高翔,张涛,刘毅,等.视觉SLAM十四讲:从理论到实践[M].2版.北京:电子工业出版社,2019: 15.
GAO Xiang, ZHANG Tao, LIU Yi, et al. 14 lectures on visual SLAM: From theory to practice[M]. 2nd ed. Beijing: Electronic Industry Press, 2019: 15.

[4] REHDER J, NIKOLIC J, SCHNEIDER T, et al. Extending kalibr: Calibrating the extrinsics of multiple IMUs and of individual axes [C]//Proceedings of IEEE International Conference on Robotics and Automation (ICRA). Stockholm, Sweden : IEEE, 2016: 4304-4311.

[5] FISCHLER M A, BOLLES R C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography[J]. Communications of the ACM, 1981, 24(6): 381-395.

[6] SAPUTRA M, MARKHAM A, TRIGONI N. Visual SLAM and structure from motion in dynamic environments[J]. ACM Computing Surveys (CSUR), 2018, 51(2): 1-36.

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

[8] DAVISON A, REID I, MOLTON N, et al. Monoslam: Real-time single camera slam[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(6): 1052-1067.

[9] KLEIN G, MURRAY D. Parallel tracking and mapping for small AR workspaces [C]//Proceedings of IEEE & ACM International Symposium on Mixed & Augmented Reality. Cambridge, UK: ACM, 2008: 225-234.

[10] TRIGGS B, MCLAUCHLAN P, HARTLEY R, et al. Bundle adjustment—A modern synthesis [C]//Proceedings of International Workshop on Vision Algorithms. Berlin: Springer, 1999: 298-372.

[11] CAMPOS C, ELVIRA R, RODRÍGUEZ J, et al. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM[J]. IEEE Transactions on Robotics, 2021, 37: 1-17.

[12] ENGEL J, SCHOPS T, CREMERS D. LSD-Slam: Largescale direct monocular Slam[C]//Proceedings of European Conference On Computer Vision. Zurich: Springer, 2014: 834-849.

[13] FORSTER C, ZHANG Z, GASSNER M, et al. SVO: Semidirect visual odometry for monocular and multicamera systems [J]. IEEE Transactions on Robotics, 2016, 33(2): 249-265.

[14] ELHAMIFAR E, VIDAL R. Sparse subspace clustering: Algorithm, theory, and applications[J]. IEEE Transactions on Software Engineering, 2013, 35(11): 2765-2781.

[15] XU X, CHEONG L F, LI Z.3D Rigid motion segmentation with mixed and unknown number of models[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 1-16.

[16] VIDAL R, MA Y, SOATTO S, et al.Two-view multibody structure from motion[J]. International Journal of Computer Vision, 2006, 68(1): 7-25.

[17] BESCOS B, FACIL J, CIVERA J, et al. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 1-1.

[18] YU C, LIU Z, LIU X, et al. DS-SLAM: A semantic visual SLAM towards dynamic environments [C]//Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid, Spain: IEEE, 2018: 1168-1174.

[19] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(12): 2481-2495.

[20] ALCANTARILLA P, YEBES J, ALMAZAN J, et al. On combining visual SLAM and dense scene flow to increase the robustness of localization and mapping in dynamic environments [C]//Proceedings of IEEE International Conference on

Robotics & Automation. St. Paul, Minnesota: IEEE, 2012: 1290-1297.
[21] ZHANG J, HENEIN M, MAHONY R, et al. VDO-SLAM: A visual dynamic object-aware SLAM system[EB/OL].(2020-01-16)[2021-06-20].https://arXiv.org/abs/2005.11052.
[22] BALLESTER I, FONTAN A, CIVERA J, et al. DOT: Dynamic object tracking for visual SLAM[EB/OL]. (2020-02-14)[2011-06-20]. https://arXiv.org/abs/2010.0005X1.
[23] YANG N, VON STUMBERG L, WANG R, et al. D3VO: Deep depth, deep pose and deep uncertainty for monocular visual odometry [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 1278-1289.
[24] CZARNOWSKI J, LAIDLOW T, CLARK R, et al. DeepFactors: Real-time probabilistic dense monocular SLAM [J]. IEEE Robotics and Automation Letters, 2020, 5(2): 721-728.
[25] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: An efficient alternative to SIFT or SURF [C]//Proceedings of International Conference on Computer Vision. Barcelona, Spain: IEEE, 2011: 2564-2571.
[26] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: The KITTI dataset [J]. International Journal of Robotics Research, 2013, 32(11): 1231-1237.
[27] HEIKO H. Stereo processing by semiglobal matching and mutual information [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2008, 30(2): 328-341.
[28] HE K, GKIOXARI G, DOLLAR P, et al. Mask R-CNN [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2020, 42(2): 386-397.
[29] HARTLEY R, ZISSERMAN A. Multi view geometry in computer vision [M]. 2nd ed. London: Cambridge University Press, 2003: 190.
[30] HARTLEY R. In defense of the eight-point algorithm[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1997, 19(6): 580-593.
[31] LEPETIT V, MORENO-NOGUER F, FUA P. EPnP: An accurate O($n$) solution to the PnP problem[J]. International Journal of Computer Vision, 2009, 81(2): 155-166.
[32] 王宜举,修乃华.非线性最优化理论与方法[M].3版.北京:科学出版社,2020: 110-118.
WANG Yiju, XIU Nahua. Nonlinear optimization theory and method [M]. 3rd ed. Beijing: Science Press, 2020: 110-118.
[33] BAKER S, MATTHEWS I. Lucas-Kanade 20 years on: A unifying framework[J]. International Journal of Computer Vision, 2004, 56: 221-255.
[34] STURM J, ENGELHARD N, ENDRES F, et al. A benchmark for the evaluation of RGB-D SLAM systems [C]// Proceedings of Intelligent Robots and Systems (IROS), IEEE/RSJ International Conference on. Macau, China: IEEE, 2012: 573-580.

**作者简介:**

**沈晔湖**(1982-),**通信作者**,男,副教授,研究方向:图像处理、计算机视觉,E-mail: yehushen@mail.usts.edu.cn。

**陈嘉皓**(1997-),男,硕士研究生,研究方向:机器人导航技术。

**李星**(1997-),男,硕士研究生,研究方向:视觉SLAM技术。

**蒋全胜**(1978-),男,副教授,研究方向:信号与信息处理。

**谢鸥**(1983-),男,副教授,研究方向:水下机器人技术。

**牛雪梅**(1980-),女,副教授,研究方向:并联机器人智能控制。

**朱其新**(1971-),男,教授,研究方向:先进控制理论及应用。

(编辑:刘彦东)