E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2023)02-0522-13

论文引用格式:Xie T, Zhang X J, Ye Z C, Wang Z H, Wang Z, Zhang Y, Zhou X W and Ji X P. 2023. Visual localization system of integrated active and passive perception for indoor scenes. Journal of Image and Graphics,28(02):0522-0534(谢挺,张晓杰,叶智超,王子豪,王政,张涌,周晓巍,姬晓鹏. 2023. 面向室内场景的主被动融合视觉定位系统. 中国图象图形学报,28(02):0522-0534)[DOI:10.11834/jig.210603]

面向室内场景的主被动融合视觉定位系统

谢挺[1],张晓杰[2],叶智超[1],王子豪[1],王政[3],张涌[3],周晓巍[1],姬晓鹏[1*]

1. 浙江大学 CAD&CG 国家重点实验室, 杭州　310058; 2. 中国船舶工业系统工程研究院, 北京　100094;
3. 中讯邮电咨询设计院有限公司, 北京　100048

摘　要:**目的**　视觉定位旨在利用易于获取的 RGB 图像对运动物体进行目标定位及姿态估计。室内场景中普遍存在的物体遮挡、弱纹理区域等干扰极易造成目标关键点的错误估计,严重影响了视觉定位的精度。针对这一问题,本文提出一种主被动融合的室内定位系统,结合固定视角和移动视角的方案优势,实现室内场景中运动目标的精准定位。**方法**　提出一种基于平面先验的物体位姿估计方法,在关键点检测的单目定位框架基础上,使用平面约束进行 3 自由度姿态优化,提升固定视角下室内平面中运动目标的定位稳定性。基于无损卡尔曼滤波算法设计了一套数据融合定位系统,将从固定视角得到的被动式定位结果与从移动视角得到的主动式定位结果进行融合,提升了运动目标的位姿估计结果的可靠性。**结果**　本文提出的主被动融合室内视觉定位系统在 iGibson 仿真数据集上的平均定位精度为 2～3 cm,定位误差在 10 cm 内的准确率为 99%;在真实场景中平均定位精度为 3～4 cm,定位误差在 10 cm 内的准确率在 90% 以上,实现了 cm 级的定位精度。**结论**　提出的室内视觉定位系统融合了被动式和主动式定位方法的优势,能够以较低设备成本实现室内场景中高精度的目标定位结果,并在遮挡、目标丢失等复杂环境因素干扰下展示出鲁棒的定位性能。

关键词:视觉定位;数据融合;关键点检测;3 维重建;P*n*P 算法;卡尔曼滤波

Visual localization system of integrated active and passive perception for indoor scenes

Xie Ting[1], Zhang Xiaojie[2], Ye Zhichao[1], Wang Zihao[1], Wang Zheng[3], Zhang Yong[3], Zhou Xiaowei[1], Ji Xiaopeng[1*]

1. State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou　310058, China; 2. System Engineering Research Institute, Beijing　100094, China; 3. China Information Technology Designing & Consulting Institute, Beijing　100048, China

Abstract: **Objective** Visual localization is focused on the location and estimation of motion objects via easy-to-use RGB images. The feature-extracted information is challenged to meet the requirements of tasks in traditional computer vision methods in terms of feature extraction algorithms. The deep learning-based feature abstraction and demonstration ability can promote an emerging research issue for pose estimation in computer vision. In addition, the development and application of depth cameras and lasers-based sensors can provide more diverse manners to this issue as well. However, these sensors have some constraints of the entity and shape of the object and it need to be used in a structured environment. Multi-vision

收稿日期:2021-07-19;**修回日期**:2021-12-14;**预印本日期**:2021-12-21

*通信作者:姬晓鹏　xp. ji@cad. zju. edu. cn

基金项目:科技创新 2030 –“新一代人工智能”重大项目课题(2020AAA0108901)

Supported by: National Key R&D Program of China(2020AAA0108901)

ability is often challenged to the issues of installing and debugging problems. In contrast, sensors-visual applications are featured of low cost and less restrictions, and they are easy to be recognized and extended for multiple unstructured scenarios. Interferences are being existed in indoor scenes, such as object occlusion and weak texture areas, which can cause the incorrect estimation of the target points easily and affect the accuracy of visual localization severely. The different methods of camera-deployment can be divided into two categories based on visual object pose estimation method. 1) In order to get the target position data, one category of the two is based on monocular object positioning of pose estimation technology of using the deployment in cameras-fixed in the scene and detecting targets in the images of the relevant information. The pros of positioning result is stable and the cons of it is affected by light and fuzzy image easily, it cannot be dealt with object occlusion in the scene as well due to the limitation of observation angle; 2) The other category of two is oriented on scene reconstruction-based object pose estimation technology, which can use the camera fixed on the target itself to obtain the pose information of the target by detecting the feature points of the scene and matching the features with the 3D scene model constructed in advance. This scheme is derived of the status of texture features. 1) For rich textures and clear features scenes, the accurate positioning results-related can be obtained. 2) For non-texture features scenes and weak texture areas like walls scene, the positioning results are unstable, and other sensors such as inertial measurement unit (IMU) are needed to be positioning-aided. To achieve more precise positions of moving objects in indoor scenes, we propose an active and passive perception-based visual localization system, which combines the advantages of fixed and motion perspectives. **Method** First, a plane-prior object pose estimation method is proposed by our research team. Based on the monocular localization framework of keypoint detection, the plane-constraint is used to optimize the 3-DoF (degree of freedom) pose of the object and improve the localization stability under a fixed view. Second, we design a data fusion framework in terms of the unscented Kalman filtering algorithm. To improve the reliability of the pose estimation of the moving target, a fixed view-derived passive perception output and the active perception output are fused from a motion view. The active and passive-integrated indoor visual positioning system is composed of three aspects as mentioned below: 1) passive positioning module, 2) active positioning module, and 3) active and passive fusion module. The input of passive positioning module is oriented to RGB image-captured by indoor fixed camera, and the output is based on the target pose data-contained in the image. The input of the active positioning module is the RGB image shot on the perspective of the target to be located, and the output is based on the position and pose-relevant information of the target in the 3D scene. The active and passive fusion module is dealt with the integrating the positioning results of passive and active positioning, and the output is linked to more accurate positioning result of the target in the indoor scene. **Result** The average localization error of the indoor visual localization system proposed can reach 2～3 cm on the iGibson simulation dataset, and the accuracy of the 10 cm-within localization error can reach to 99%. In the real scenes, the average localization error can reach 3～4 cm, and the accuracy of the localization error within 10 cm is above 90%. Experimental results are shown our proposed system can obtain centimeter-level accurate positioning. The experimental results of real scenes illustrate that the active and passive fusion visual positioning system can reduce the external interference of passive positioning algorithm under fixed visual angle effectively due to the limitation of visual angle, object occlusion and other external disturbances, and it also can optimize the defects of single frame positioning algorithm with insufficient stability and large random error. **Conclusion** Our visual localization system has its potentials to the integrated advantages of passive-based and active-based methods, which can achieve high-precision positioning results in indoor scenes at a low cost. It also shows better robust performance under complex interference such as occlusion and target-missed. We develop a lossless Kalman filter based framework of active and passive fusion indoor visual positioning system for indoor mobile robot operation. Compared to the existing visual positioning algorithm, it can achieve high-precision target positioning results in indoor scenes with lower equipment cost. And, under the shade circumstances, the loss of target under complex environment factors shows robust positioning performance and the indoor scene visual centimeter-level accuracy-purified of positioning. The performance is tested and validated in simulation and the physical environment both. The experimental results show that the positioning system has its priority on high positioning accuracy and robustness for multiple scenarios further.

Key words: visual localization; data fusion; keypoint detection; 3D reconstruction; P*n*P algorithm; Kalman filter

0 引 言

高精度的室内目标定位技术在虚拟现实、增强现实和机器人导航等领域具有重要的应用价值。按照传感器信号来源,常见的室内定位技术可分为基于 Wi-Fi 等无线信号的定位技术、基于惯性导航的定位技术、基于地磁的定位技术和基于视觉信息的定位技术。其中,基于 Wi-Fi 等无线信号的定位技术对设备和场地有较高要求;基于惯性导航的定位技术只能通过惯性测量单元(inertial measurement unit,IMU)获取相对位置,无法获得绝对位置;而地磁定位技术时间开销较大且精度不高。在室内环境中,由于图像数据存在细节丰富、易于获取以及部署快捷等天然优势,基于视觉的室内定位方法得到广泛关注。室内视觉定位系统可以利用易于获取的 RGB 图像,对已知目标进行精确位姿估计。

按摄像头部署方式,基于视觉的室内定位方法可分为主动式定位(移动观测视角)和被动式定位(固定观测视角)两种。被动式定位方法利用部署在场景中的固定摄像头,通过检测图像中的目标关键点进行模板匹配,从而解算出目标的位姿数据。这种方案的优点是定位结果比较稳定,不易受到光照、模糊图像的影响,但由于观测视角的限制,无法处理场景中存在的物体遮挡情况。主动式定位方法则是利用固定在定位目标本身的摄像头,通过检测场景的特征点,并与事先构建的 3 维场景模型进行特征匹配来得到目标的位姿信息。这种方案的缺陷是过度依赖图像的纹理特征,对于纹理丰富、特征明显的场景可以得到比较准确的定位结果;而对于纹理特征缺失的场景,如墙面等弱纹理区域,定位结果非常不稳定。

在特定应用场景,例如室内移动机器人作业,既可以通过室内固定的监控摄像头进行目标的被动式定位,也可以通过定位目标自身(移动机器人)的移动摄像头进行主动式定位,这两类方法的定位结果有一定的共同性,而采用单一的主动式或被动式定位方法,都存在场景适应能力不足、定位精度受限等问题。

针对这类应用场景,本文提出一种主被动融合的室内场景定位系统。首先,基于单目标检测深度学习框架,提出一种基于平面先验的物体位姿估计方法,利用室内场景中普遍存在的平面约束,对运动目标进行 3 自由度(degree of freedom,DoF)位姿估计;其次,提出一个基于无损卡尔曼滤波(unscented Kalman filter,UKF)的主被动融合定位框架,对主动式和被动式定位模块得到的位姿结果进行融合,提升了室内场景下运动目标位姿估计结果的稳定性和精准性。为验证提出的主被动融合定位系统的性能,在仿真平台 iGibson 和真实室内场景进行实验。实验结果表明,本文提出的融合定位方法可以有效提升室内场景中移动目标的定位精度及准确率。

本文主要贡献如下:1)提出一个基于无损卡尔曼滤波的主被动融合室内视觉定位系统框架,可以有效解决弱纹理及遮挡条件下的室内移动目标的位姿估计问题;2)提出一种基于平面先验的物体位姿估计方法,可有效提升室内场景中运动目标的定位精度。

1 相关工作

视觉定位技术根据定位原理不同,可以分为被动式定位技术和主动式定位技术两种。

1.1 被动式视觉定位技术

被动式定位的目标是根据固定视角得到的图像进行目标的定位,并得到目标物体的 6-DoF 位姿数据。传统的单目标定位方法主要通过模板匹配技术获取目标的 3 维点位置信息,如基于图像的梯度响应图检测场景中的 3 维物体(Hinterstoisser 等,2013),或基于模型边缘轮廓的形状描述子估计目标位姿(Zhu 等,2014)。但传统方法对环境变化和物体遮挡关系比较敏感,无法处理结构复杂的图像。

随着深度学习的发展,卷积神经网络在复杂结构和环境变化的图像检测和识别领域展示出优秀性能,例如 PoseNet(Kendall 等,2015)使用卷积神经网络(convolutional neural network,CNN)直接回归目标物体的位姿数据,但受限于 RGB 图像的深度信息缺乏和庞大的解空间搜索规模,效果并不鲁棒。Schönberger 等人(2017)对手工特征和深度学习特征的 2D—3D 匹配性能进行评估,肯定了深度学习在图像匹配定位上的优秀效果。PoseCNN(Xiang 等,2018)则通过预测 2 维图像的深度图改善 3 维定位效果。也有一些方法通过离散化旋转空间将定位问题转化为分类问题(Sundermeyer 等,2018)进行处

理,从而提升姿态估计结果的精度。谢非等人(2020)基于端到端模型提出准确高效的机器人室内单目视觉定位算法。

基于目标关键点的定位方法可以得到更加稳定的定位结果,这类方法通过局部特征点检测提取目标关键点,然后基于 2D—3D 点的对应关系,利用 RANSAC(random sample consensus)算法和 P*n*P(perspective-*n*-point)算法求解得到物体位姿。一些方法基于随机森林预测 3D 坐标值(Michel 等,2017),并利用几何约束改进生成 2D—3D 点的对应关系。DenseFusion(Wang 等,2019)则基于 RGB-D 数据,使用网络对 RGB 图像特征和 3D 点云特征进行整合,得到 RGB-D 数据的像素密集特征表示,然后通过投票获取物体的 6-DoF 姿态。PVNet(pixel-wise voting network)(Peng 等,2019)通过对关键点进行投票得到遮挡情况下的局部特征向量,从而缓解图像中存在的目标遮挡和隔断问题。

单目视觉定位的一个缺陷是由于视角限制,位姿估计精度对拍摄距离和遮挡比较敏感。针对这一问题,刘昶等人(2012)基于共面二点一线特征进行视觉定位,MLOD(multi-view labelling object detector)(Deng 和 Czarnecki,2019)用多视图目标检测方法进行目标定位,取得了出色效果。

1.2 主动式视觉定位技术

主动式视觉定位方法指的是通过自身携带的移动摄像头拍摄环境图像来定位自身的方法,依据拍摄的当前观测图像,查询数据库中存在的图像位置信息,进行图像匹配来完成定位。经典的主动式定位方法包含场景建图和图像检索两个阶段。首先通过 3 维重建方法获取场景的 3 维点云模型,然后将输入的查询图像与点云模型中的 3 维点建立对应关系,使用 RANSAC 算法解算出查询图像的位置。

近年利用图像检索算法(Torii 等,2015)从场景图像数据库中检索近似图像,得到较准确的初始位姿成为主动式视觉定位的主流方法。InLoc(indoor visual localization)(Taira 等,2018)利用稠密的图像匹配和视角一致性来提升室内定位结果精度,但其算法依赖于全局匹配,计算代价较高,系统实时性较差。HFNet(hierarchical feature network)(Sarlin 等,2019)则提出将图像检索和局部特征提取融合到统一的网络框架中,以提升姿态估计算法的计算效率。

主动式定位方法通常依赖于图像的强纹理特征,对纹理结构丰富的区域,特征点匹配成功率较高,得到的定位结果也比较准确,但在一些弱纹理或特征不显著的区域上的定位效果往往较差。针对这一问题,可以通过加入基于轨迹的滤波算法(Sattler 等,2017)或利用其他辅助信息来改善定位结果。有一些方法(DeTone 等,2018)通过增强特征点检测的性能来提升定位精度。场景无关相机定位方法 DSM(dense scene matching)(Tang 等,2021)和 KF-Net(Kalman filtering network)(Zhou 等,2020)主要使用稠密场景匹配,在图像和场景间构造 cost volume,通过 CNN 网络来估计稠密坐标。

1.3 数据融合定位技术

多传感器数据融合技术在导航中的应用比较广泛,通常结合卡尔曼滤波算法、图优化算法来实现多传感器输入下的高精度定位。卡尔曼滤波假定误差满足线性高斯分布,但实际系统并不满足线性近似假设。扩展卡尔曼滤波(extended Kalman filter,EKF)将非线性模型在状态估计值附近做泰勒级数展开,采用局部线性化方法获得状态估计。无损卡尔曼滤波基于无损变换(unscented transform,UT)对后验概率密度进行近似估计,可以有效解决非线性系统引起的滤波发散问题。

VINSMono(Qin 等,2018)基于单目视觉里程计和扩展卡尔曼滤波,提出了一套视觉信息和视觉惯性单元(inertial measurement unit,IMU)融合的松耦合框架,具有完善和鲁棒的初始化及闭环检测过程,以适应室内外环境。Li 等人(2017)基于 GPS(global positioning system)和图像融合,实现了对智能车辆的高精度定位。Xue 等人(2017)基于相机、激光雷达和地理信息系统(geographic information system,GIS)的融合定位方法,实现了无人车的障碍物感知和高精度自主定位。杨承凯等人(2009)探讨了多传感器融合中的卡尔曼滤波。

结合视觉和激光雷达、惯性测量单元等其他物理传感器的定位方法可以得到比较高的定位精度,但对于定位装置的要求也非常高,通常需要装配价格昂贵的传感器设备。因此,本文希望能够利用低成本的 RGB 摄像头,提出一套基于视觉的室内主被动融合定位系统以实现便捷、精准的室内目标定位。

本文提供了一个在特定应用场景(被动式定位技术与主动式定位技术的定位目标相同)可实现的系统性视觉定位解决思路,结合了被动式定位和主

动式定位两种方式的优点，当遇到视角盲区和强遮挡情况时，会更多地倾向于主动式定位的结果，而当遇到图像弱纹理情况时，更多地倾向于被动式定位的结果。仿真实验和实物实验结果都印证了本文的思路和猜想。

2 本文方法

提出的主被动融合室内视觉定位系统主要包含3个模块，即被动式定位模块、主动式定位模块和主被动融合模块，如图1所示。其中，被动式定位模块的输入为室内固定视角摄像头得到的RGB图像，输出为该图像包含目标的位姿数据；主动式定位模块的输入为待定位目标视角拍摄的RGB图像，输出为该目标在3维场景中的位姿信息；主被动融合模块负责融合被动式定位和主动式定位的定位结果，然后输出最终的位姿数据。

在场景定位中，本文定义全局坐标系为目标运动的水平面，这样直观且便于测量误差，而被动式定位和主动式定位结果都会转换到这一坐标系上。

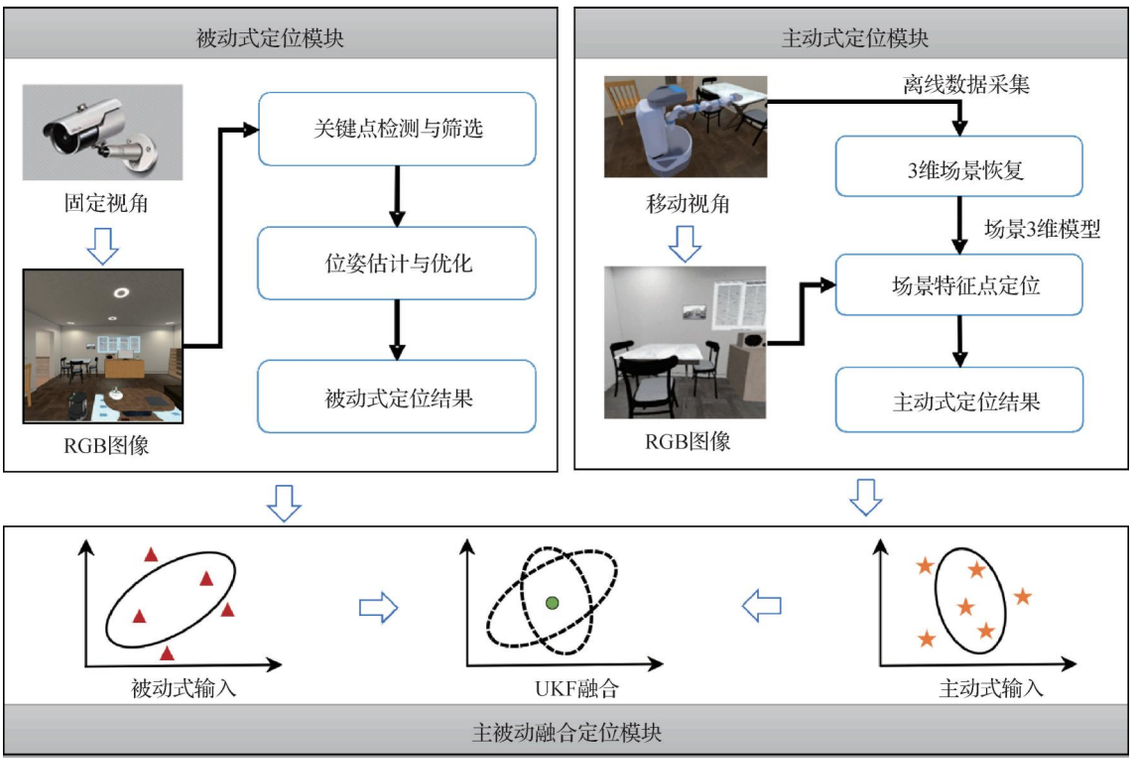


图1 主被动融合系统的整体框架

Fig. 1 Overview of the integrated active and passive perception system

2.1 被动式定位模块

被动式定位模块流程主要包括关键点检测与筛选、位姿估计与优化两个步骤，如图2所示。

2.1.1 关键点检测与筛选

本文使用卷积神经网络获取待定位目标的关键点信息。对固定视角输入的RGB图像，首先使用CenterNet(Zhou等，2019)提出的基于中心点的目标检测方法，构建网络模型，从图像中检测出目标的中心点和尺度信息，得到目标的2维边框，从原始图像中裁剪出来用于定位。然后，针对检测到的目标区域，基于预先定义的目标关键点信息，预测图像中对应于目标关键点分布的多幅热力图，每幅热力图都代表一个关键点位置的估计结果。在关键点检测过程中，采用网络输出的关键点热力图计算每个关键点的像素坐标及对应的置信度，并使用关键点置信度预先过滤掉一些错误的关键点，如图3所示。

在关键点检测网络模型训练中，通过预先定义的目标关键点3维位置，由其真实位姿值进行投影，得到关键点在图像上的像素坐标。

2.1.2 位姿估计与优化

对于室内场景中的运动目标，例如室内服务机器人，通常沿水平面进行移动，高度信息不会发生剧

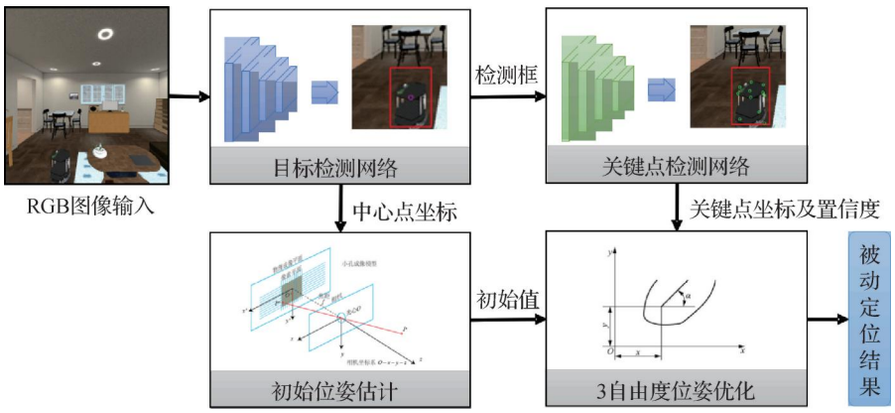
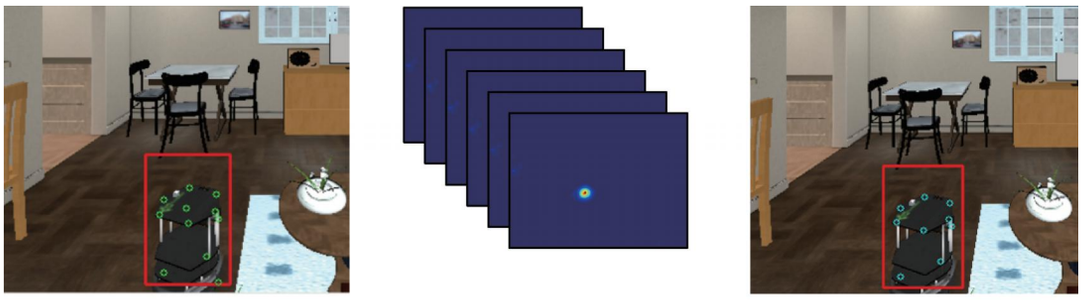


图 2　被动式方法定位流程图

Fig. 2　Pipeline of the passive localization



(a) 关键点检测　(b) 关键点热力图计算置信度　(c) 关键点过滤

图 3　关键点检测与筛选示意图

Fig. 3　Illustration of the keypoint detection and elimination

((a) keypoint detection; (b) keypoint thermodynamic diagram; (c) keypoint filter)

烈变化，即待定位目标处于高度已知的水平平面内。设物体系 z 轴与世界系 z 轴平行，则该物体在 3 维场景中的运动可以简化为 3-DoF 的位姿估计问题。利用这样的平面假设，可以缩小位姿估计的求解空间，提高其估计的稳定性和准确性。

假设摄像头固定，可以通过场景中设置的已知标记点得到固定摄像头在全局坐标系中的位置和姿态。在此基础上，可以将通过单目物体定位得到的物体相对于相机坐标系的位姿转换到全局坐标系中。图像中每个 2 维像素点对应的全局坐标系 3 维坐标值是固定的，可以根据目标的 2 维框中心位置来估计定位目标的位姿初值 (x_0, y_0)。

给定一组关键点的 2D—3D 对应关系，通过最小化目标函数可以得到物体在 3 维平面上的坐标 (x_0, y_0)，目标函数为

$$L = \sum_{i=0}^{n_v} w_i \left((\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2 \right) \tag{1}$$

式中，$(\hat{x}_i, \hat{y}_i)$ 为关键点检测与筛选后得到的第 i 个关键点的像素坐标，(x_i, y_i) 代表该关键点经重投影方程得到的重投影像素坐标，n_v 为定义的目标关键点数量，w_i 为关键点检测网络输出第 i 个关键点的置信度，用于平衡不同置信度下的关键点对优化过程的影响。

为避免式(1)的优化过程陷入局部最优，本文对 (x_0, y_0) 和目标对象的初始朝向角 θ_{init} 进行初始化，以 2 维候选框的中心坐标 (x_0, y_0) 代表目标的中心坐标，通过相机成像模型，并假设 $Z = 0$。因此，可以根据投影方程计算出目标中心在世界坐标系位置的初始值 $(X_{\text{init}}, Y_{\text{init}})$，具体为

$$\begin{bmatrix} x_0 \\ y_0 \\ 1 \end{bmatrix} = \boldsymbol{KT} \begin{bmatrix} X_{\text{init}} \\ Y_{\text{init}} \\ Z_{\text{init}} \\ 1 \end{bmatrix} \tag{2}$$

式中，$\boldsymbol{K} \in \mathbf{R}^{3\times4}$ 为相机内参矩阵，$\boldsymbol{T} \in \mathbf{R}^{4\times4}$ 为世界

坐标系相对于相机坐标系的位姿，$Z_{\text{init}}=0$。

接下来，计算目标对象的初始朝向角 θ_{init}。

朝向角初始估计值 θ_{init} 的取值范围为[0°, 360°)，将解空间以 k° 为步长进行采样，可计算出 $360/k$ 组重投影误差，从中选取使重投影误差(如式(1)所示)最小的 θ_{init} 作为初始朝向角。

位姿优化过程以位姿参数为待求解参数，以初始位姿为初始值，通过最小化重投影误差得到目标的最终位姿数据。

2.2 主动式定位模块

主动式方法定位流程包含3维场景恢复和场景特征点定位两个阶段，如图4所示。

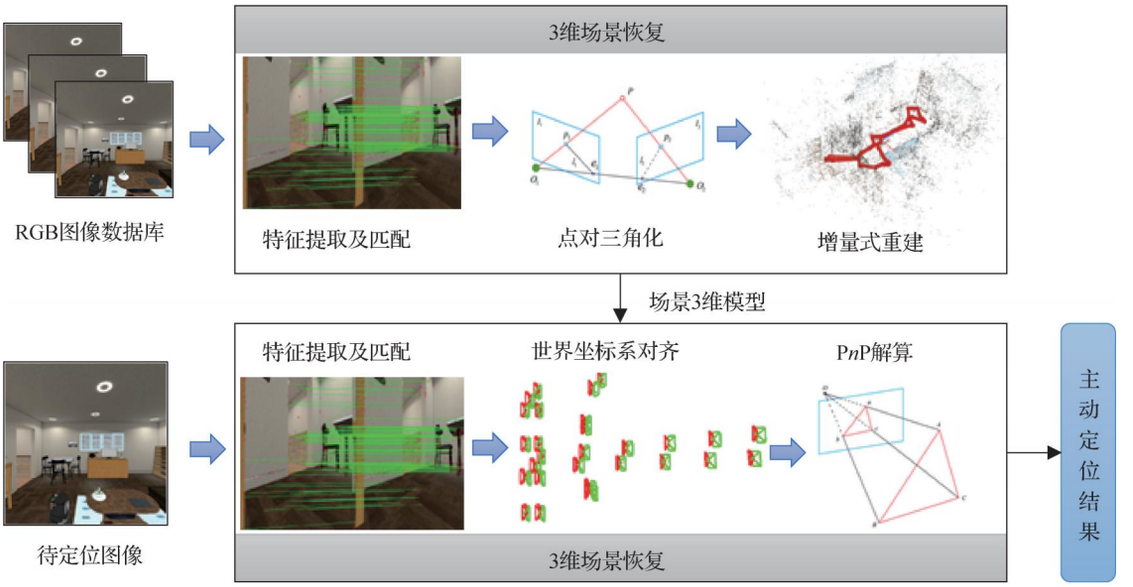


图4 主动式方法定位流程图

Fig. 4 Pipeline of the active localization

主动式定位以场景地图为先验，类似于SLAM(simultaneous localization and mapping)系统中的重定位部分。在定位过程中，用采集的图像数据(定位目标携带的移动摄像头拍摄)匹配场景地图，获得的是相对于场景地图的位姿，而场景地图是预先对齐到全局坐标系的，因此主动式定位能获取定位目标在全局坐标系中的位姿。

2.2.1 3维场景恢复

3维重建的场景恢复过程主要基于运动恢复结构(structure from motion，SfM)技术。首先通过特征提取和特征匹配算法得到场景不同视角下图像之间的关联，初始选择两个基线距离较大且匹配较多的图像对进行重建，通过初始重建的3维点对其他图像帧进行定位。对新增的注册帧重新三角化出更多的3维点，然后不断迭代直到恢复出所有的图像位姿和点云表示的3维场景模型。

在完成场景的3维重建后，将场景地图对齐到全局坐标系。选取一些固定的标记点采集对应位置的图像，通过直接测量的方式获取标记点对应的全局坐标系3维坐标，然后将对应图像通过特征提取、与场景地图数据的特征匹配以及PnP算法得到在场景地图坐标系中的3维坐标，使用Umeyama算法(Umeyama，1991)计算得到两个坐标系之间的变换矩阵。

2.2.2 场景特征点定位

对输入的待定位图像，首先对其进行局部特征提取，并与已恢复场景地图中的局部特征进行匹配，得到待定位图像中2维局部特征位置与场景模型中3维结构之间的几何映射关系。然后在图像2D特征点和场景中3D特征点匹配基础上，通过求解PnP问题，得到该图像对应目标对象的6-DoF位姿信息。

2.3 主被动融合定位模块

无损卡尔曼滤波通过UT变换使非线性系统方程适用于线性假设下的标准卡尔曼体系，可以克服EKF估计精度低、稳定性差的缺陷。本文假设目标满足匀加速运动模型，定义其状态向量 $\boldsymbol{X}_k=(x_k,$

$y_k, v_k^x, v_k^y, a_k^x, a_k^y)^{\mathrm{T}}$，其中，$x_k, y_k$ 为当前时刻目标在 2 维平面上的坐标位置，v_k^x, v_k^y 为此时目标的速度值，a_k^x, a_k^y 为其加速度值。根据匀加速运动可以得到相邻时刻目标的运动状态向量，具体为

$$\boldsymbol{X}_{k+1} = \begin{pmatrix} x_k + v_k^x t + \dfrac{a_k^x t^2}{2} \\ y_k + v_k^y t + \dfrac{a_k^y t^2}{2} \\ v_k^x + a_k^x t \\ v_k^y + a_k^y t \\ a_k^x \\ a_k^y \end{pmatrix} \tag{3}$$

系统的状态方程和观测方程可以表示为

$$\boldsymbol{X}_{k+1} = f(\boldsymbol{X}_k, \boldsymbol{U}_k),\ U_k \sim \mathrm{N}(0, Q_k) \tag{4}$$

$$\boldsymbol{Z}_k = h(\boldsymbol{X}_k, \boldsymbol{W}_k),\ W_k \sim \mathrm{N}(0, R_k) \tag{5}$$

式中，$\boldsymbol{X}_k$ 为第 k 时刻目标的状态变量，包括位置、速度和加速度等，$\boldsymbol{U}_k$ 为系统噪声，Q_k 为系统噪声分布的协方差，$\boldsymbol{W}_k$ 为观测噪声，R_k 为观测噪声分布的协方差，观测向量 $\boldsymbol{Z}_k$ 为第 k 时刻主/被动式定位方法的预测定位结果 (x_k, y_k)。

使用无损卡尔曼滤波算法将被动式预测定位结果与主动式预测定位结果进行非线性融合，根据式(3)和式(4)，实际系统的状态方程可以表示为

$$\boldsymbol{X}_{k+1} = f(\boldsymbol{X}_k, \boldsymbol{U}_k) = f(\boldsymbol{X}_k) + \boldsymbol{U}_k \tag{6}$$

根据式(5)，观测方程可以表示为

$$\boldsymbol{Z}_k = h(\boldsymbol{X}_k, \boldsymbol{W}_k) = \boldsymbol{H}\boldsymbol{X}_k + \boldsymbol{W}_k \tag{7}$$

式中，初始噪声分布情况根据训练数据集中的样本采样估计、统计预测值和真实值的误差分布情况得到，包括被动式定位系统噪声 $\boldsymbol{Q}_p$、观测噪声 $\boldsymbol{R}_p$ 和主动式定位系统噪声 $\boldsymbol{Q}_a$、观测噪声 $\boldsymbol{R}_a$，即

$$\boldsymbol{Q}_k = \begin{cases} \boldsymbol{Q}_a & \text{主动式定位} \\ \boldsymbol{Q}_p & \text{被动式定位} \end{cases}$$

$$\boldsymbol{R}_k = \begin{cases} \boldsymbol{R}_a & \text{主动式定位} \\ \boldsymbol{R}_p & \text{被动式定位} \end{cases}$$

$$\boldsymbol{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

主被动定位的结果按时间整理成观测值队列，按顺序加入到融合定位的滤波系统中，分别按照各自的噪声分布更新协方差矩阵和状态向量。

具体融合定位实现过程如下：

1)求初始状态均值和方差矩阵

$$\begin{cases} \bar{\boldsymbol{X}}_0 = E(\boldsymbol{X}_0) \\ \boldsymbol{P}_{X,0} = E[(\boldsymbol{X}_0 - \bar{\boldsymbol{X}}_0)(\boldsymbol{X}_0 - \bar{\boldsymbol{X}}_0)^{\mathrm{T}}] \end{cases} \tag{8}$$

式中，$\bar{\boldsymbol{X}}_0$ 为状态量初始均值；$\boldsymbol{P}_{X,0}$ 为状态量初始方差矩阵。

2)通过 UT 变换，获取状态量采样点和对应权值，各采样点可以表示为

$$\boldsymbol{\chi}_{i,k} = \begin{cases} \bar{\boldsymbol{X}}_k & i = 0 \\ \bar{\boldsymbol{X}}_k + (\sqrt{(n+\lambda)\boldsymbol{P}_{X,k}}) & i = 1, \cdots, n \\ \bar{\boldsymbol{X}}_k - (\sqrt{(n+\lambda)\boldsymbol{P}_{X,k}}) & i = n+1, \cdots, 2n \end{cases} \tag{9}$$

各采样点对应的权值可以表示为

$$w_i = \begin{cases} \dfrac{\lambda}{n+\lambda} & i = 0 \\ \dfrac{1}{2(n+\lambda)} & i \neq 0 \end{cases} \tag{10}$$

式中，n 为状态向量维数，根据实际应用场景确定；$\boldsymbol{\chi}_{i,k}$ 为采样得到的第 i 个点状态量；$\bar{\boldsymbol{X}}_k$ 为第 k 时刻状态量均值；λ 为缩放尺度参数；$\boldsymbol{P}_{X,k}$ 为第 k 时刻状态量方差矩阵。

3)使用上述采样点集进行预测，计算预测估计值 $\bar{\boldsymbol{X}}_{k+1}$，以及计算 $k+1$ 时刻的预测协方差 $\bar{\boldsymbol{P}}_{X,k+1}$ 和预测测量值 $\bar{\boldsymbol{Z}}_{k+1}$，可表示为

$$\begin{cases} \bar{\boldsymbol{X}}_{k+1} = \sum_{i=0}^{2n} w_i \boldsymbol{\chi}_{i,k} \\ \bar{\boldsymbol{P}}_{X,k+1} = \sum_{i=0}^{2n} w_i (\boldsymbol{\chi}_{i,k} - \bar{\boldsymbol{X}}_{k+1})(\boldsymbol{\chi}_{i,k} - \bar{\boldsymbol{X}}_{k+1})^{\mathrm{T}} + \boldsymbol{Q}_k \\ \bar{\boldsymbol{Z}}_{k+1} = \sum_{i=0}^{2n} w_i \boldsymbol{\gamma}_{i,k} = \sum_{i=0}^{2n} w_i \boldsymbol{H} \boldsymbol{\chi}_{i,k} \\ \boldsymbol{P}_{Z,k+1} = \sum_{i=0}^{2n} w_i (\boldsymbol{\gamma}_{i,k} - \bar{\boldsymbol{Z}}_{k+1})(\boldsymbol{\gamma}_{i,k} - \bar{\boldsymbol{Z}}_{k+1})^{\mathrm{T}} + \boldsymbol{R}_k \\ \boldsymbol{P}_{XZ,k+1} = \sum_{i=0}^{2n} w_i (\boldsymbol{\chi}_{i,k} - \bar{\boldsymbol{X}}_{k+1})(\boldsymbol{\gamma}_{i,k} - \bar{\boldsymbol{Z}}_{k+1})^{\mathrm{T}} \end{cases} \tag{11}$$

式中，$\boldsymbol{\gamma}_{i,k}$ 为采样点集通过观测函数得到的观测点集。

4)计算卡尔曼增益 $\boldsymbol{K}_{k+1}$，更新协方差 $\boldsymbol{P}_{X,k+1}$ 和状态向量 $\boldsymbol{X}_{k+1}$，得到状态转移方程

$$\begin{cases} \boldsymbol{K}_{k+1} = \boldsymbol{P}_{XZ,k+1} \boldsymbol{P}_{Z,k+1}^{-1} \\ \boldsymbol{P}_{X,k+1} = \bar{\boldsymbol{P}}_{X,k+1} - \boldsymbol{K}_{k+1} \bar{\boldsymbol{P}}_{Z,k+1} \boldsymbol{K}_{k+1}^{\mathrm{T}} \\ \boldsymbol{X}_{k+1} = \bar{\boldsymbol{X}}_{k+1} + \boldsymbol{K}_{k+1} (\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1}) \end{cases} \tag{12}$$

将式中状态向量 $\boldsymbol{X}_{k+1}$ 包含的坐标值作为 $k+1$

时刻主被动融合定位的位置坐标。至此,完成了基于无损卡尔曼滤波的主被动融合定位方法。

数据融合的过程中,观测值来自于被动式定位和主动式定位的定位结果,共用同一个状态向量进行更新,但是被动式定位和主动式定位有各自的协方差矩阵。状态向量更新时,先判断观测数据来源是被动式定位还是主动式定位,选择协方差矩阵,然后对主被动定位方法得到的初步定位结果采用阈值筛选的方法去除偏差较大的观测值,得到有效观测。

对于被动式定位结果,基于重投影误差和关键点置信度进行评价和筛选,误差函数为

$$S_p(x,y) = w_1 \frac{1}{n}\sum \left| \boldsymbol{P}_{2\mathrm{d}}^{\mathrm{proj}} - \boldsymbol{P}_{2\mathrm{d}}^{\mathrm{pred}} \right| + w_2 \frac{1}{n}\sum x_{\mathrm{conf}} \tag{13}$$

式中, $\boldsymbol{P}_{2\mathrm{d}}^{\mathrm{pred}}$ 和 $\boldsymbol{P}_{2\mathrm{d}}^{\mathrm{proj}}$ 为检测和位姿投影的关键点 2 维投影坐标值,重投影误差体现了定位检测的精度结果。x_{conf} 为检测网络输出的关键点平均置信度, w_1, w_2 为手工调节的权值。

对于主动式定位结果,基于检测到的特征点数量和匹配点数量比例进行评价和筛选,误差函数为

$$S_a(x_2,y_2) = w_1 \frac{n_{\mathrm{inliers}}}{n_{\mathrm{sift}}} + w_2 n_{\mathrm{sift}} \tag{14}$$

式中, n_{sift} 为图像中检测到的 SIFT(scale-invariant feature transform)特征点数量, n_{inliers} 为匹配到场景 3 维点的图像特征点数量。

3 实验结果分析

3.1 实验平台与数据

为验证提出的主被动融合方法的有效性,本文分别在仿真平台和真实场景进行实验。仿真数据和真实环境获取的数据示例如图 5 所示。

(a) 仿真数据主动式/被动式视角　　(b) 真实场景主动式/被动式视角

图 5　仿真数据和真实数据示例

Fig. 5　Simulation data and real data samples

((a) simulation data from *active*/passive perspective; (b) real data from active/passive perspective)

本文的仿真实验平台基于 iGibson(Xia 等, 2020)机器人仿真环境构建。iGibson 是美国斯坦福大学开发的大型交互式机器人模拟平台,包含 15 个依据真实房屋构建的可交互、高拟真度的室内场景。本文选择其中 8 个室内场景,记录机器人运动时的主动视角图像、被动视角(全局摄像头)图像和真实位姿,共生成 16 组轨迹数据,其中,8 组样本作为训练数据,用于被动式关键点检测网络的训练和主动式定位中室内场景的离线 3 维重建,其余 8 组用于模型测试,场景面积约为 40 ～50 m^2。

本文在室内实际场景中利用 Turtlebot3 和 Realsense 采集图像数据进行实物实验,共生成 10 组数据,其中 5 组样本作为训练数据,用于被动式定位的模型训练以及主动式定位的场景 3 维重建,其余 5 组用于模型测试。其中,物体位姿定位的真实值(ground truth)根据在地面标记的记号点利用直尺直接测量得到。

3.2 评价标准

在算法评估方面,本文使用平均定位精度和定位准确率作为量化指标。

平均定位精度(average accuracy)指测试集中每个样本的定位结果到真实值(ground truth)的欧氏距离均值。具体为

$$error = \frac{1}{n}\sum_{k=1}^{n} \sqrt{(\hat{x}-x_{\mathrm{gt}})^2 + (\hat{y}-y_{\mathrm{gt}})^2} \tag{15}$$

式中, n 为样本数量, $\hat{x}$, $\hat{y}$ 为样本的定位结果, x_{gt}, y_{gt} 为真实的目标坐标位置。

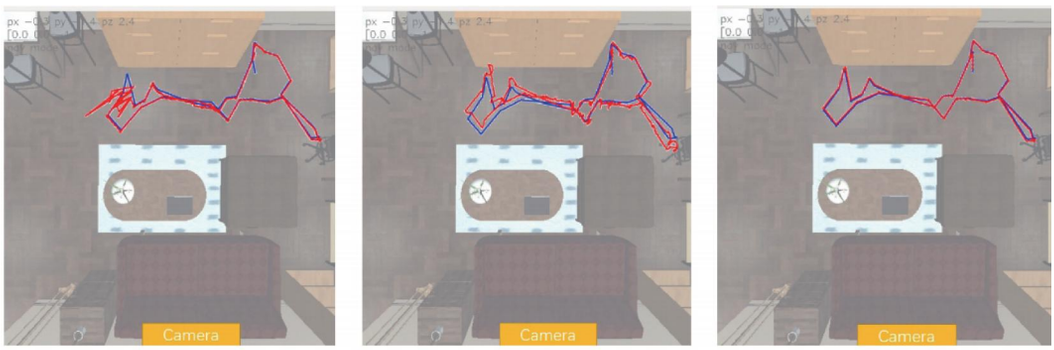
定位准确率指能达到指定定位精度的正确样本占所有样本的比例。实验设置了小于 15 cm、小于 10 cm 和小于 5 cm 三段范围的准确率指标来定量评估不同算法的定位精度。

3.3 网络参数设置

被动式定位方法中目标检测的网络框架采用 Resnet-18，关键点检测网络框架采用 Dla-34。模型训练输入的图像分辨率为 512 × 512 像素，批大小（batch size）设置为 8，学习率设置为 0.000 1。

3.4 仿真实验结果

在仿真实验平台上的可视化结果如图 6 所示，其中蓝色轨迹为 iGibson 记录真实值，红色轨迹为不同定位方法输出预测值，黄色区域为场景中固定视角的摄像头。



(a) 被动式方法定位轨迹　(b) 主动式方法定位轨迹　(c) 融合定位轨迹

图 6　仿真环境被动式/主动式/融合定位方法可视化结果

Fig. 6　Visualization of passive/active/fusion localization methods in the simulation environment

((a) passive localization trajectory; (b) active localization trajectory; (c) fusion localization trajectory)

从定位轨迹结果可以看到，被动式定位方法的特点是在固定摄像头靠近中心的视野区域定位精度较高，而对于偏离视野中心距离的区域以及存在较严重遮挡的区域，其定位精度会严重下降。主动式定位方法对于纹理贫乏、特征不明晰的图像定位效果较差，其轨迹吻合度相对于被动式定位方法精度要差，但不容易产生较大估计误差，具备比较好的稳定性。经过 UKF 融合后的结果能准确恢复机器人的运动轨迹，对于主/被动存在定位误差部分进行了有效修正，且其轨迹的平滑程度和鲁棒性都得到了明显提升。

表 1 给出了不同算法在仿真数据集上的定位精度指标。从定位精度评估结果可以看到，基于 UKF 的主被动融合定位方法在平均定位精度和准确率上有明显提升。与 SSD-6D 的方法相比，被动式定位方法的平均定位精度和准确率都要高出一些，主要优势来自平面场景的 3 自由度优化。与基于 EKF 的融合定位方法相比，平均定位精度基本一致，但 10 cm/5 cm 的准确率有明显提升，特别是对遮挡比较严重的区域有更好的定位效果。

仿真实验结果表明，本文提出的主被动融合定位方法可以在仿真场景中实现 cm 级的精准定位，平均定位精度为 2.01 cm，误差在 10 cm 内的准确率可以达到 99.0% 以上。

表 1　仿真数据集定位结果

Table 1　Localization results on simulation dataset

方法	平均定位精度/cm	准确率/%		
		<15 cm	<10 cm	<5 cm
SSD-6D	3.5	90.60	85.30	81.70
被动式定位	2.4	98.50	97.90	87.90
主动式定位	7.33	89.20	68.40	35.90
主被动融合定位(EKF)	2.04	99.00	98.50	90.10
主被动融合定位(UKF)	**2.01**	**99.10**	**99.10**	**92.50**

注：加粗字体表示各列最优结果。

3.5 真实场景实验结果

为进一步验证本文方法的可用性,在真实室内场景的实物数据集上对不同算法的定位精度进行测试,结果如表 2 所示。从定位结果可以看出,基于 UKF 的主被动融合定位方法在平均定位精度和准确率上均取得了最优结果。图 8 给出了实际场景的定位准确率曲线,可以看到,在定位准确率上,基于 UKF 的方法相比主被动定位方法和 EKF 融合方法都有一定提升。

真实场景的实验结果表明,主被动融合视觉定位系统能有效降低固定视角下被动式定位算法由于视角局限性、物体遮挡等外界干扰,同时可以有效克服单帧的定位算法稳定性不足、随机误差大的缺陷。本文提出的主被动融合定位方法同样能够在真实室内环境中实现 cm 级的目标精准定位,误差在 10 cm 内和 15 cm 内的准确率分别为 93.1% 和 97.1% 。

表 2 实物数据集定位结果

Table 2 Localization results on real dataset

方法	平均定位精度/cm	准确率/%		
		< 15 cm	< 10 cm	< 5 cm
SSD-6D	4.02	92.10	85.30	75.60
被动式定位	3.67	95.60	90.60	80.90
主动式定位	7.26	82.60	56.80	32.10
主被动融合定位(EKF)	3.52	95.10	91.30	81.30
主被动融合定位(UKF)	**3.47**	**97.10**	**93.10**	**82.50**

注:加粗字体表示各列最优结果。

4 结 论

本文针对室内移动机器人作业等特定应用场景,提出了一个基于无损卡尔曼滤波的主被动融合室内视觉定位系统框架,与现有的视觉定位算法相比,能够以较低设备成本实现室内场景中高精度的目标定位结果,并在遮挡、目标丢失等复杂环境因素干扰下展示出鲁棒的定位性能,实现室内场景的纯视觉 cm 级精准定位。在仿真和实物环境下都进行了测试和验证,实验结果表明,定位系统具备很高的定位准确率和鲁棒性,能够用于多种场景。

但是,本文的研究工作还存在不少可以改进的地方。一方面,单目视觉定位方法存在对物体先验模型的依赖性,在无先验条件下进行定位是一个难题,而基于机器学习的定位方法需要大量的标注数据,如何对原始数据进行快速智能标注也是一个需要解决的问题;另一方面,从数据融合算法的角度看,本文方法主要是在数据级融合和特征级融合上展开,在更高一层的目标级融合研究上还存在一些欠缺。

在未来工作中,将继续深入研究如何与视觉里程计等 SLAM 方法进行更深层次的数据融合,实现不同场景的模型迁移,将数据融合的系统定位算法应用于更加广阔的场景。

参考文献(References)

Deng J and Czarnecki K. 2019. MLOD: a multi-view 3D object detection based on robust feature fusion method//Proceedings of 2019 IEEE Intelligent Transportation Systems Conference (ITSC). Auckland, New Zealand: IEEE: 279-284 [DOI: 10.1109/ITSC.2019.8917126]

DeTone D, Malisiewicz T and Rabinovich A. 2018. SuperPoint: self-supervised interest point detection and description// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA: IEEE: 337-349 [DOI: 10.1109/CVPRW.2018.00060]

Hinterstoisser S, Lepetit V, Ilic S, Holzer S, Bradski G, Konolige K and Navab N. 2013. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes//Proceedings of the 11th Asian Conference on Computer Vision. Daejeon, Korea (South): Springer: 548-562 [DOI: 10.1007/978-3-642-37331-2_42]

Kendall A, Grimes M and Cipolla R. 2015. PoseNet: a convolutional network for real-time 6-DOF camera relocalization//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV).

Santiago, Chile: IEEE: 2938-2946 [DOI: 10.1109/ICCV.2015.336]

Li Y C, Hu Z Z, Hu Y Z and Wu H W. 2017. Accurate localization based on GPS and image fusion for intelligent vehicles. Journal of Transportation Systems Engineering and Information Technology, 17(3): 112-119 [DOI: 10.16097/j.cnki.1009-6744.2017.03.017]

Liu C, Zhu F and Xia R B. 2012. Monocular pose determination from coplanar two points and one line features. Application Research of Computers, 29(8): 3145-3147 (刘昶, 朱枫, 夏仁波. 2012. 基于共面二点一线特征的单目视觉定位. 计算机应用研究, 29(8): 3145-3147) [DOI: 10.3969/j.issn.1001-3695.2012.08.090]

Michel F, Kirillov A, Brachmann E, Krull A, Gumhold S, Savchynskyy B and Rother C. 2017. Global hypothesis generation for 6D object pose estimation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 115-124 [DOI: 10.1109/CVPR.2017.20]

Peng S D, Liu Y, Huang Q X, Zhou X W and Bao H J. 2019. PVNet: pixel-wise voting network for 6DoF pose estimation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 4556-4565 [DOI: 10.1109/CVPR.2019.00469]

Qin T, Li P L and Shen S J. 2018. VINS-Mono: a robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics, 34(4): 1004-1020 [DOI: 10.1109/TRO.2018.2853729]

Sarlin P E, Cadena C, Siegwart R and Dymczyk M. 2019. From coarse to fine: robust hierarchical localization at large scale//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 12708-12717 [DOI: 10.1109/CVPR.2019.01300]

Sattler T, Leibe B and Kobbelt L. 2017. Efficient and effective prioritized matching for large-scale image-based localization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(9): 1744-1756 [DOI: 10.1109/TPAMI.2016.2611662]

Schönberger J L, Hardmeier H, Sattler T and Pollefeys M. 2017. Comparative evaluation of hand-crafted and learned local features//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 6959-6968 [DOI: 10.1109/CVPR.2017.736]

Sundermeyer M, Marton Z C, Durner M, Brucker M and Triebel R. 2018. Implicit 3D orientation learning for 6D object detection from rgb images//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 712-729 [DOI: 10.1007/978-3-030-01231-1_43]

Taira H, Okutomi M, Sattler T, Cimpoi M, Pollefeys M, Sivic J, Pajdla T and Torii A. 2018. InLoc: indoor visual localization with dense matching and view synthesis//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7199-7209 [DOI: 10.1109/CVPR.2018.00752]

Tang S T, Tang C Z, Huang R, Zhu S Y and Tan P. 2021. Learning camera localization via dense scene matching//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1831-1841 [DOI: 10.1109/CVPR46437.2021.00187]

Torii A, Arandjelović R, Sivic J, Okutomi M and Pajdla T. 2015. 24/7 place recognition by view synthesis//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 1808-1817 [DOI: 10.1109/CVPR.2015.7298790]

Umeyama S. 1991. Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(4): 376-380 [DOI: 10.1109/34.88573]

Wang C, Xu D F, Zhu Y K, Martín-Martín R, Lu C W, Li F F and Savarese S. 2019. DenseFusion: 6D object pose estimation by iterative dense fusion//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3338-3347 [DOI: 10.1109/CVPR.2019.00346]

Xia F, Shen W B, Li C S, Kasimbeg P, Tchapmi M E, Toshev A, Martín-Martín R and Savarese S. 2020. Interactive Gibson benchmark: a benchmark for interactive navigation in cluttered environments. IEEE Robotics and Automation Letters, 5(2): 713-720 [DOI: 10.1109/LRA.2020.2965078]

Xiang Y, Schmidt T, Narayanan V and Fox D. 2018. PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes//Proceedings of the 14th Robotics: Science and Systems. Pittsburgh, USA: [s.n.]: #19

Xie F, Wu J, Huang L, Zhao J, Liu X X and Qian W X. 2020 A robot indoor monocular vision positioning algorithm based on end-to-end model. Journal of Chinese Inertial Technology, 28(4): 493-498, 560 (谢非, 吴俊, 黄磊, 赵静, 刘锡祥, 钱伟行. 2020. 基于端到端模型的机器人室内单目视觉定位算法. 中国惯性技术学报, 28(4): 493-498, 560) [DOI: 10.13695/j.cnki.12-1222/o3.2020.04.012]

Xue J R, Wang D, Du S Y, Cui D X, Huang Y and Zheng N N. 2017. A vision-centered multi-sensor fusing approach to self-localization and obstacle perception for robotic cars. Frontiers of Information Technology and Electronic Engineering, 18(1): 122-138 [DOI: 10.1631/FITEE.1601873]

Yang C K, Zeng J and Huang H. 2009. Discussion of Kalman filter in multi-sensor fusion. Modern Electronics Technique, 32(14): 159-161, 164 (杨承凯, 曾军, 黄华. 2009. 多传感器融合中的卡尔曼滤波探讨. 现代电子技术, 32(14): 159-161, 164) [DOI: 10.3969/j.issn.1004-373X.2009.14.050]

Zhou L, Luo Z X, Shen T W, Zhang J H, Zhen M M, Yao Y, Fang T and Quan L. 2020. KFNet: learning temporal camera relocalization

using kalman filtering//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4918-4927 [DOI: 10.1109/CVPR42600.2020.00497]

Zhou X Y, Wang D Q and Krähenbühl P. 2019. Objects as points [EB/OL]. [2021-07-15]. http://arxiv.org/pdf/1904.07850.pdf

Zhu M L, Derpanis K G, Yang Y F, Brahmbhatt S, Zhang M, Phillips C, Lecce M and Daniilidis K. 2014. Single image 3D object detection and pose estimation for grasping//Proceedings of 2014 IEEE International Conference on Robotics and Automation (ICRA). Hong Kong, China: IEEE: 3936-3943 [DOI: 10.1109/ICRA.2014.6907430]

作者简介

谢挺,男,硕士研究生,主要研究方向为计算机视觉和目标定位。E-mail: 21921254@zju.edu.cn

姬晓鹏,通信作者,男,博士研究生,主要研究方向为计算机视觉和3维重建。E-mail: xp.ji@cad.zju.edu.cn

张晓杰,男,硕士研究生,主要研究方向为舰载航空保障指挥控制技术。E-mail: 18911990998@163.com

叶智超,男,博士研究生,主要研究方向为计算机视觉和机器人SLAM。E-mail: yezhichao@zju.edu.cn

王子豪,男,硕士研究生,主要研究方向为计算机视觉和物体位姿估计。E-mail: zihaowang@zju.edu.cn

王政,男,研究员,主要研究方向为数字化智能化转型。E-mail: wangz378@chinaunicom.com

张涌,男,高级工程师,主要研究方向为ICT运营。E-mail: yong-zhang@chinaunicom.com

周晓巍,男,研究员,博士生导师,主要研究方向为计算机视觉及其在增强现实和机器人领域的应用。E-mail: xwzhou@zju.edu.cn