基于改进 SLAM 框架的动态场景三维语义地图构建方法研究

张鹏飞[1] 李宏伟[2*] 赵亚帅[1] 张彭昱[1] 冯彬彬[2]

[1](郑州大学信息工程学院 河南 郑州 450001)

[2](郑州大学地球科技与技术学院 河南 郑州 450001)

摘 要 三维语义信息是智能机器理解世界的重要因素,是人工智能的重要一环。提出一种基于 ORB-SLAM2 改进的 SLAM 框架,可以更好地适应于动态复杂环境下低纹理和感知混叠等问题的处理。结合用于语义分割的卷积神经网络提供的语义信息,通过贝叶斯方法进行语义关联,实现在 Octomap 中的优化定位与更新,构建一致的三维语义地图。基于公开数据集的测试结果表明,该方法在复杂环境下,整体建图精度和速度相较于传统视觉 SLAM 算法有一定提升,而且降低光照变换产生的影响,具有较高的应用价值。

关键词 视觉 SLAM 三维重建 语义分割 语义地图

中图分类号 TP242.6 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.11.033

CONSTRUCTION METHOD OF 3D SEMANTIC MAP OF DYNAMIC SCENE BASED ON IMPROVED SLAM FRAMEWORK

Zhang Pengfei[1] Li Hongwei[2*] Zhao Yashuai[1] Zhang Pengyu[1] Feng Binbin[2]

[1](*School of Information Engineering, Zhengzhou University, Zhengzhou 450001, Henan, China*)

[2](*School of Earth Science and Technology, Zhengzhou University, Zhengzhou 450001, Henan, China*)

Abstract Three-dimensional semantic information is an important factor for intelligent machines to understand the world and an important part of artificial intelligence. This paper proposes an improved SLAM framework based on ORB-SLAM2, which can better adapt to the processing of low texture and perceptual aliasing problems in dynamic and complex environments. Combined with the semantic information provided by the convolutional neural network for semantic segmentation, the Bayesian method was used for semantic association to achieve optimized positioning and update in Octomap, and a consistent three-dimensional semantic map was built. The test results based on the public data set show that in a complex environment, the overall mapping accuracy and speed of this method are improved compared with traditional visual SLAM algorithms, and the impact of light transformation is reduced, which has high application value.

Keywords Visual SLAM Three-dimensional reconstruction Semantic segmentation Semantic map

0 引 言

在实际应用中,智能机器广泛要求 3D 场景。这是因为智能机器在对所处环境完全未知的情况下,对所处状态及自身位置没有任何的标注信息,无法有效地完成任务。移动设备普遍配置有激光、深度相机和 GPS,可以获取原始的数据,却无法有效整理成结构化、层次化的数据信息直接应用。而智能机器在各种应用中,又需要依赖计算机几何图形及其组合。所以如何将视觉数据通过计算与学习构建三维场景,并进行语义分析,最终可以转化为机器理解的信息,是目前计算机视觉在人工智能领域研究的重点。

本文主要研究解决的两个基本问题,即智能机器

收稿日期:2021-06-07。中国工程科技发展战略河南研究院战略咨询研究项目(2020HENZT07)。**张鹏飞**,硕士生,主研领域:视觉 SLAM,三维重建,深度学习。**李宏伟**,教授。**赵亚帅**,硕士生。**张彭昱**,硕士生。**冯彬彬**,硕士生。

如何正确感知和理解环境中的“对象在哪里”与“对象是什么”的问题。SLAM 就是要解决在一个未知环境中如何对环境进行实时重建,并同时对移动设备进行定位,最后同步创建环境地图。视觉 SLAM 算法可通过利用传感器中获取的连续图像帧序列进行位姿估计,以达到三维重建的效果。但是视觉 SLAM 算法只能获取稀疏或者稠密的空间点信息,不能提供更高级的环境语义信息,不能构建强大、准确和详细的三维地图。为此要在 SLAM 的基础上增强对环境的认知,就需要对三维环境中的语义信息进行标注。

基于卷积神经网络(CNN)的语义分割算法,为我们解决语义信息标注提供了可能。语义分割网络可以构建一个场景理解小型系统[1],在室内外不同的场景中,可以通过利用不同类别的物体之间呈现的特定空间关系,来提高语义分割性能。然后将改进的 SLAM 系统弹性融合的几何信息与使用卷积神经网络的语义分割算法相结合[2],实现三维语义建模。SLAM 系统提供了从 2D 帧到全局一致的 3D 地图的对应关系,利用 CNN 网络的 SemanticFusion,从多角度进行语义预测,可以将标签融合到一个稠密的语义注释地图中。

本文结合一种半直接法的视觉 SLAM 算法和深度语义分割方法,解决三维建模方法、进行语义分割信息标注,构建结合深度学习卷积神经网络算法进行图像分割、语义标注的新的视觉 SLAM 系统,最终建立包含语义信息的三维场景地图。同时利用本文算法能够让智能机器可以识别不同深度的物体而更好地完成导航任务。

1　相关工作

三维语义重建的基础是 SLAM 算法,视觉 SLAM 通过图像对齐的方法,可以包含稀疏、半稠密和稠密类型。视觉 SLAM 系统的架构主要由视觉里程计、后端优化、回环检测和建图组成[3]。前端任务是对位姿进行状态估计,这需要抓取与跟踪传感器的图像数据进行支撑。后端任务是优化,将前端的关键帧数据进行深度图估计,实现地图优化。后端可以进行反馈于前端,实现交互,并进行回环检测,最后实现建图。

起初,许多单目 SLAM 方法都是通过滤波来实现的,如 MONOSLAM 等[4],而非线性优化因其计算量大而不受欢迎。目前 SLAM 算法大致有两个方向,一种是基于特征的方法,只能重建稀疏点云,如 ORB-SLAM[5]。其思想是将如何从图像中进行几何信息估计,划分为提取特征观察值、利用位姿信息进行计算两个紧密的步骤。整个问题虽然得到简化,但受到只能使用符合特征类型信息的限制。为了获得密集的重建,估计的相机姿态可以用于多视角立体重建密集的地图。另外一种 SLAM 算法是基于直接方法,如 LSD-SLAM[6]、DSO[7],可以生成半稠密或稠密点云。直接视觉里程计方法通过直接在图像强度上优化几何形状和光度误差,使得图像的所有信息可以被利用。不仅提供高精确度和鲁棒性,还可以在没有关键点的环境中,提供更多关于环境的几何信息。

三维语义重建不同于传统的 SLAM 方法,更侧重识别三维环境中的语义信息。它需要智能机器对人类的抽象概念进行学习,如场景、物体和形状。Guan 等[8]提出一种语义可视化 SLAM 方法,该方法结合二维对象检测和 ORB 特征点,引入对象间相对位置不变性的语义约束,提供精确的位姿和丰富的语义信息。Yu 等[9]提出的 DS-SLAM 系统是一种结合 ORB-SLAM 的系统,采用语义分割方法从 RGB 图像检测物体所在的像素点,通过运动一致性检查物体描述子是否为动态,最后利用深度图像的静态像素构建八叉树地图。Wang 等[10]针对动态环境下智能机器定位精度差、语义理解环境能力低下等问题,提出一种高效率的 SLAM 解决方案,可以生成具有特定物体高精度语义信息的语义地图。

上述三维语义重建的 SLAM 方法大多建立在基于特征的方法或直接方法上。在基于特征的方法中[11],每个帧的显著图像被提取并匹配于姿态估计、优化和循环闭合,但是这种基于特征方法性能容易受到时间限制。而基于直接法虽然更快,这是因为可以利用图像的原始光度信息恢复相机姿态和结构,但这种方法受到的限制很显而易见,即对光照变化和循环闭合检测很敏感。同时无论是直接法还是特征点法,环境均设置为静态场景,且由于摄像头运动造成场景部分变化[12],进而影响动态环境下的鲁棒性,出现无法计算较多动态物体的情况。

为了解决以上问题,本文采用一种新的半直接 SLAM 算法,与语义分割网络算法相结合,设计构建三维语义重建算法,对三维环境的语义信息关联,实现三维语义重建环境。

2　SLAM 算法与语义分割算法

2.1　复杂环境下的半直接单目视觉 SLAM 算法

为了保持直接方法的快速性能和基于特征方法的高精度及闭环能力,本文使用一种新的视觉同步定位与映射的半直接法,可以有效地处理复杂环境中的低

纹理、运动目标和感知混叠等问题。同时为了提高动态环境下的定位精度,采用对光照变化具有鲁棒性的运动检测模块,可以动态实时实现复杂室内场景三维重建。

本文算法由五个并行线程组成,包括跟踪、特征提取、局部映射、循环检测、全局优化(图 1)。

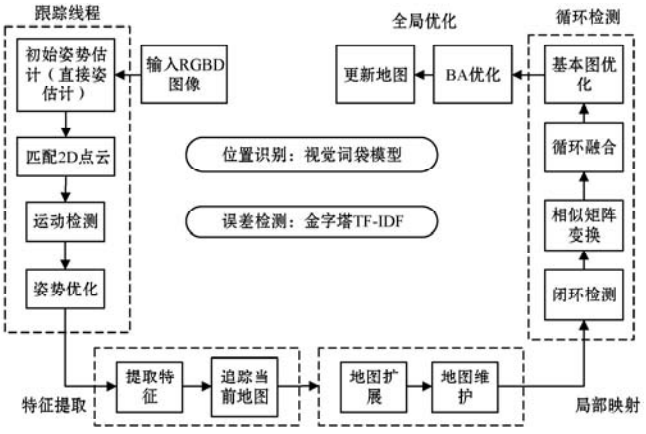


图 1 系统框架

该方法是在 ORB-SLAM 的基础上,进行改进并提出的一种处理深度图像的 SLAM 方法。相比较其他三维重建 SLAM 算法,该算法主要有以下几点优势。

1)结合基于直接方法和基于特征方法的优点,保持快速的跟踪速度,同时又有较高的重建精度。

2)采用鲁棒性的循环检测模块,提高动态环境下的定位精度,进而达到全局优化。

3)将直接法和特征法的空间信息融合到视觉词袋模型中(BoVW)[13],减少感知混叠。

4)改进金字塔项频率逆文档频率(TF-IDF)[14],提高闭环检测的精度比。

2.1.1 单目视觉初始化

单目相机无法从一帧图像中恢复出观测到的目标点的尺度,所以单目模型通常采用三角测量的方式通过连续几帧图像计算帧之间的相对姿态,来恢复尺度信息。导致视觉里程计在开始工作之前需要有一个初始化的阶段,为之后的位姿估计提供尺度信息。

在初始化阶段,假设特征点的像素坐标为(u,v),由于在计算内参时使用的是在相机坐标系下的坐标,对于在世界坐标下为$\boldsymbol{P}_W$,相机的位姿与旋转矩阵$\boldsymbol{R}$和平移向量$\boldsymbol{t}$的关系式为:

$$ZP_{uv}=Z\begin{pmatrix}u\\v\\1\end{pmatrix}=\boldsymbol{K}(\boldsymbol{R}\boldsymbol{P}_W+\boldsymbol{t})=\boldsymbol{K}\boldsymbol{T}\boldsymbol{P}_W \tag{1}$$

式中:变换矩阵$\boldsymbol{T}$为相机外参;$\boldsymbol{K}$为相机内参。根据相机的运动,可求解得到相机坐标系下的三维坐标(X,Y,Z),得到尺度信息,为单目视觉里程计提供一个良好的初始值。

2.1.2 初始姿态估计和特征对准

在视觉里程计通过单目相机获取到上一帧的图像之后,将使用直接法估计当前帧的姿态。通过最小化对应于两个相邻帧中相同 3D 点的图像块光度差异,可以获得上一帧到当前帧的相对姿态$\boldsymbol{T}_{k,k-1}$。

$$\boldsymbol{T}_{k,k-1}=\operatorname{argmin}\frac{1}{2}\sum_{u_i\in R}\|\delta I(T_{k,k-1},u_i)\|^2 \tag{2}$$

式中:R表示前一帧图像中检测到特征点的区域和对应空间点投影到当前帧中的区域;u_i是前一帧图像中特征点的像素坐标。

光度误差δI可以表示为:

$$\delta I(\boldsymbol{T},\boldsymbol{u})=I_k(\pi(\boldsymbol{T}\cdot\boldsymbol{P}_{k-1}))-I_{k-1}(\boldsymbol{u}) \tag{3}$$

式中:I表示图像光度;π是相机投影模型的函数;$\boldsymbol{P}_{k-1}$表示对应u的 3D 点位置。

在获得初始姿态之后,为了优化新帧相应 2D 点u'_i的位置,提出的算法最小化新帧与可协方差关键帧之间的光度差异。

$$\boldsymbol{u}'_i=\operatorname{argmin}\frac{1}{2}\sum_{j\in N}\|I_k(\boldsymbol{u}'_i)_j-I_r(\pi(\boldsymbol{T}_{r,w}\cdot\boldsymbol{p}_{wi})_j)\|^2 \tag{4}$$

式中:N代表剩余模式;$\boldsymbol{p}_{wi}$是世界坐标中协方差关键帧跟踪地图点。

当直接法无法估计初始相机姿态时,跟踪线程将提取新帧的 ORB 特征,并将这些特征与参考关键帧特征进行匹配,以此完成初始姿态估计。

2.1.3 循环检测与优化

循环检测和优化主要解决位姿估计随时间出现偏差、循环校正和基本图优化问题。当智能设备运作时由于误差产生漂移问题,导致它的位姿估值并没有回到初始点,我们可以通过计算图像间的相似性来实现回环检测。循环校正是融合重复的映射点,用相似变换矩阵校正关键帧姿态,进行连接转换达到循环两端对齐,所有匹配的地图点在相似变换矩阵计算中融合内联的地图点。

视觉词袋模型(BoVW)以较少时间可以显示良好的效果,但场景内部局部特征之间的空间关系无法反映,而且易遭到感知混叠。为解决这一问题,先用一个分层的 Kmean ++ 算法构建一个视觉词汇树,将直接法和特征法的空间信息融入其中,改进相似度得分函数,最终减少感知混叠,实现优化。

2.2 语义分割算法

三维重建是智能机器环境感知的核心技术问题,而场景语义信息则有助于智能机器更好地理解环境。语义分割方法可以分割场景信息,是构建语义地图的重要手段。图像数据可以利用语义信息进行像素级别

的分割处理[15],获得的类别标签与视觉 SLAM 图像中每个像素互相匹配。随着深度学习卷积网络的兴起,图像的视觉特征可以通过多卷积层的分层抽象进行提取,这为实现三维语义重建奠定了基础。

2.2.1　传统的语义分割 FCN 模型

FCN[16]作为语义分割的经典网络,它将分类网络转换成卷积网络。卷积网络是一种可视化模型,没有限制输入图片数据的大小,经过卷积网络的学习与计算生成限定大小的输出。通过产生层次化的结构特征,实现端对端的网络训练。基于 FCN 体系结构的 SegNet[17]和 U-Net[18]网络模型,使用全网络模型而且都在此基础上对编码解码网络有了进一步改善。其中 SegNet 编码器当对图像功能点完全存储时,效率最高且保存完整性,可以存储压缩形式的编码器特征映射(降维、最大池索引),降低内存提升性能。U-net 网络结构具有捕捉上下文和精度定位的特点,可以实现跳跃连接,达到精确的语义融合效果。

SegNet 和 U-Net 基于传统语义分割 FCN 网络模型结构,都是先卷积后上采样。虽然语义丰富但空间信息损失严重,易形成局部语义丢失,导致图像的分辨率下降。为了提高分辨率的特征,同时减少下采样率保证语义丰富和精细,本文采用 DilatedFCN 结构的 DeepLab v3 + 网络模型,在其网络结构中引入空洞卷积,替代下采样层,如图 2 所示。

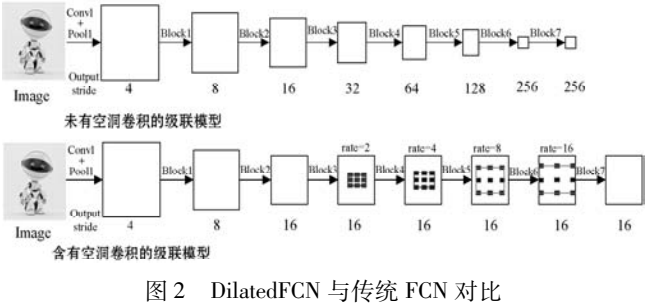


图 2　DilatedFCN 与传统 FCN 对比

2.2.2　DeepLab V3 + 模型

DeeLab[19]网络结构主要有两个模块:深度卷积神经网络(DCNN)和概率图模型(Dense CRFs)。DeepLab 系列不断发展,DeepLab v1 使用全连接 CRF 提高模型捕获细节的能力。而 DeepLab v2 在 v1 基础上用 ResNet 残差网络替代 VGGNet 网络,采用空间金字塔池化(ASPP)获取多尺度信息。随后 DeepLab v3 则在 ASPP 模块进行扩充,可以进一步提高编码图像及全文的上下级图像特征。最终 DeepLab v3 + 模型使用改进的 Xception 网络替代 DeepLab v3 模型中采用的 ResNet 残差网络,得到更好的网络分割效果,这种改进的网络模型主要体现三点:

1) 参考可变形网络的修改,增加了更多的层,有效学习密集空间变换。

2) 所有的最大池化层使用 stride = 2 的深度可分离卷积替换为空洞卷积。

3) 在深度卷积网络后增加 BN 和 ReLU。

本文选用包含空洞卷积和空间金字塔池化模型的 DeepLab v3 + 作为此次的语义分割方法模型(图 3)。

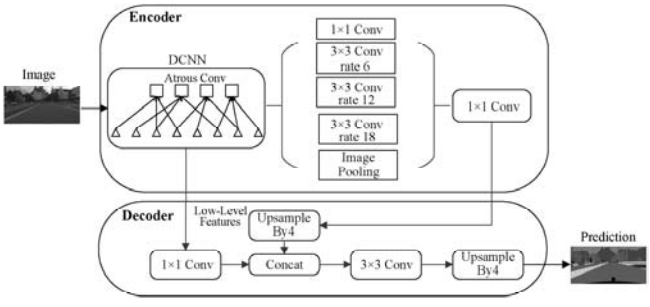


图 3　DeepLab v3 + 网络结构

3　三维语义地图构建

本文的三维重建设计框架如图 4 所示。首先输入深度相机获取的原始 RGBD 图像分别输入两个不同的模块,并附带颜色和深度信息。一个是 SLAM 系统模块,它可以提供相机位姿与帧之间的对应关系,以及融合几何信息的三维重建地图;另外是语义分割模块,预测标签概率,得到语义信息。通过跟踪每个表面的类概率分布达到语义标签融合,根据 SLAM 系统提供的对应关系融合语义信息来更新地图。最后使用 Octo-map 模块让获得的相机位姿信息和语义点云信息建立三维语义建模,并利用 Rviz 可视化工具显示地图。

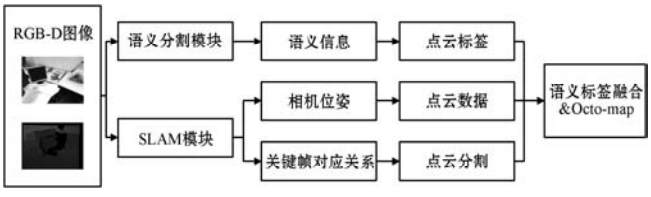


图 4　三维语义重建框架

其中语义标签融合是三维语义重建的关键[20],本文采用的是贝叶斯融合方法。该方法通过物体置于单个帧之上,在一个像素处对可能性进行语义标记,并对物体进行归一化以产生有效的概率分布实现,多应用于多视图语义融合。

首先在三维重建中每个表面(索引)储存了在类别标签集合上的离散概率分布,$P(L_s = l_i)$。每一个新生成的表面元素都是用语义类均匀分布来初始化。通过 CNN 架构可以将包括 RGB、深度或法线任意组合的图像来向前传递,然后输出以简化的方式解释为类标签 $P(Q_u = l_i \mid I_k)$ 上每个像素独立概率分布,u 表示像素坐标。使用摄像机投影 $u(s,k) = \Pi(\boldsymbol{T}_{CW}(k)_W x(s))$,

将给定位姿 $W_{x(s)}$ 处每个表面与像素坐标 u 关联，最后齐次变换矩阵 $\boldsymbol{T}_{CW}(k)=\boldsymbol{T}_{WC}^{-1}(k)$，可以使得通过递归贝叶斯更新相应的概率。其应用每个表面的标签概率，最后用常数 Z 归一化产生适当的分布。正是 SLAM 对应关系可以使我们准确将多幅图像中标签假设联系开来，并以贝叶斯方法进行融合。

$$P(l_i \mid I_{1,2,\cdots,k})=\frac{1}{Z}P(l_i \mid I_{1,2,\cdots,k-1})P(O_{u(s,k)}=l_i \mid I_k) \quad (5)$$

图 5 是各个模块的节点关系，详细展现了输入图像数据信息后，在三维语义建模系统各个模块中运行实现的步骤，介绍了三维语义框架内部各个模块的联系，是三维语义框架（图 4）技术层面的实现。

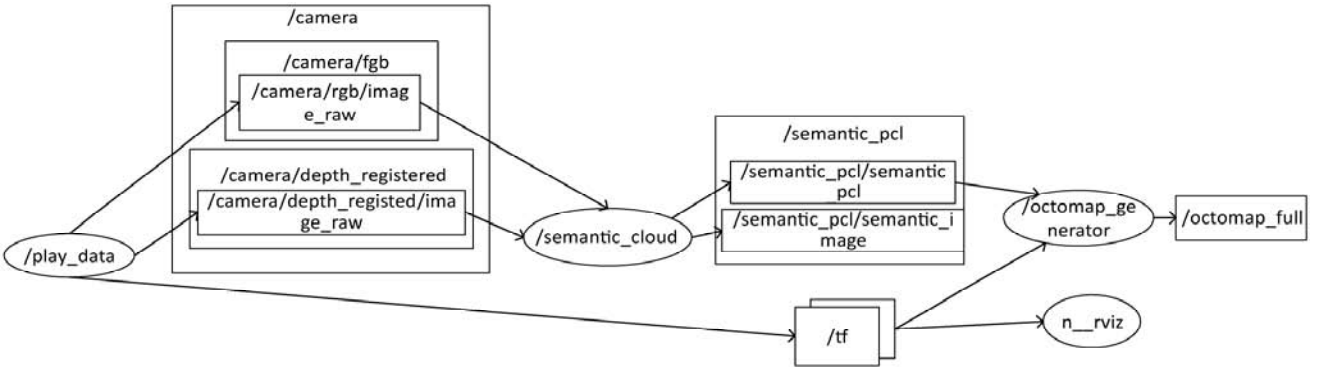


图 5 系统节点

4 实验结果与分析

4.1 实验平台与数据集

实验基于 Intel i7 CPU，16 GB 内存和 NVIDIA 2080 Ti 显卡环境搭建。选择在 Python 2.7，Ubuntu 18.04 系统下进行操作，实现语义分割算法和 SLAM 算法。

本文主要采用 ADE20K[21]、TUM-RGBD[22] 和 SUN-RGBD[23] 数据集。这三种数据集场景比较丰富，数据来源不同，可以用于验证本文模型的有效性。ADE20K 数据集是 SUN-RGBD 数据集的扩展，包含室内和室外 150 个对象，有超过 20 000 幅图片组成的训练集，本文的语义分割模型选用 ADE20K 数据集的室内对象进行训练与评估。SUN-RGBD 数据集由 37 个室内对象组成，其中 5 284 幅图像用于训练，5 051 幅图像用于测试。本文选用 SUN-RGBD 数据集用于语义分割模型的微调。TUM-RGBD 数据集的原始数据由德国慕尼黑工业大学发布，提供大量的 RGB-D 数据和标签信息，选用 TUM-RGBD 数据集作为本文系统的测试集，用作室内三维语义建模的测试。

4.2 实验分析

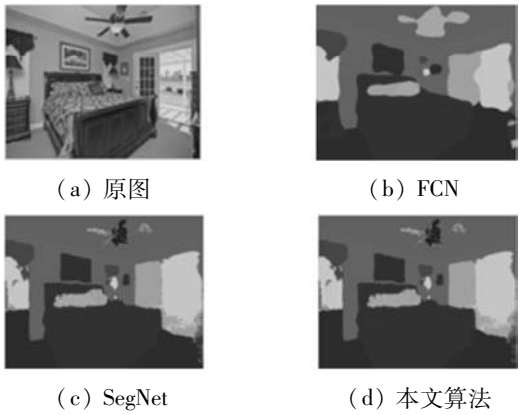
4.2.1 语义分割模块验证

本文设计的 SLAM 系统利用深度相机采集的图片，获取位姿与深度信息。语义分割线程利用关键帧的深度信息，获得语义分割结果。在 ASPP 模块进行扩充，得到高编码上下文图像特征。最后使用改进的 Xception 网络模型可以得到更好的分割效果。使得地图上每个点都与一种特定的颜色信息相关联。

对于场景分析的基准，本文训练了三个语义分割网络：SegNet、FCN-8S、本文使用的 DeepLab V3 +。为了保证公平比较和准确测试，扩展网络结构采用填充流和对象流。在语义分割网络框架中，梯度下降算法采用 AdamW，学习率设为 0.001，权重衰减设为 0.01，Batchsize 设为 8，迭代 1 000 次。通过表 1 可知，本文算法在像素精度、平均精度、平均 IoU 和加权 IoU 指标下略有提升，性能效果更好。图 6 是使用 ADE20K 验证集得到的语义结果示例，与其他两种网络相比，本文采用的 DilatedNet 模型结构效果更加详细，可以具有更少的训练数据但更高的视觉复杂度的优点。

表 1 语义分割算法在 ADE20K 的性能

对比工作	像素精度/%	平均精度/%	平均 IoU	加权 IoU
FCN-8S	70.62	39.89	0.283 9	0.573 3
SegNet	71.50	32.24	0.206 3	0.538 4
本文算法	73.46	43.62	0.313 4	0.601 4



（a）原图 （b）FCN （c）SegNet （d）本文算法

图 6 语义分割结果对比

4.2.2　系统性能评估与分析

为了评价 SLAM 算法的能力,我们选用 TUM 的公开数据集测试序列作为本文系统的测试集,同时增加 fr2_desk 和 fr2_360_hemisphere 序列来评估系统实时性能。其中测试序列包含了两种相机不同的运行轨迹视角,一种是相机沿着 X-Y-Z 轴进行移动,简称 XYZ;另外一种相机则是按照翻滚、俯仰和偏航轴进行移动,简称 RPY。在本次实验中,我们选用绝对轨迹误差(ATE)作为 SLAM 系统的性能评价指标,ATE 是通过直接计算相机位姿的真实值与 SLAM 系统的估计值之间的差,具有良好的反映算法精度和轨迹全局一致性的直观效果。通过使用关键帧姿态的绝对轨迹误差和每帧的时间来评估本文的 SLAM 算法准确性和实时性。

系统实时性能的比较结果如表 2 所示,在没有移动对象的序列中,本文系统的实时性能均优于 ORB-SLAM2 算法。而系统误差定性分析的结果如表 3 和表 4 所示,通过计算两种算法的四个指标:均方根差、均值、中值误差和标准差,可以看出本文算法所得数据都比 ORB-SLAM2 算法数据小,鲁棒性更好。图 7 是轨迹对比图,虚线是真实轨迹,两条实线分别是本文算法的轨迹路线和 ORB-SLAM2 的轨迹路线。从图 7 中可见本文算法轨迹路线更靠近虚线的真实轨迹,进一步证明了本文算法精度有所提高。

表 2　实时性能对比 Meantime　　单位:s

TUM 序列	ORB-SLAM2	本文算法
fr1_xyz	0.025 9	0.021 4
fr1_rpy	0.024 1	0.020 3
fr2_desk	0.033 2	0.027 6
fr2_360_hemisphere	0.024 7	0.018 0

表 3　XYZ 运动序列 ATE

算法指标	ORB-SLAM2	本文算法
RMSE	0.235 4	0.056 2
Mean	0197 1	0.046 7
Median	0.136 8	0.047 0
STD	0.033 1	0.009 4

表 4　RPY 运动序列 ATE

算法指标	ORB-SLAM2	本文算法
RMSE	0.117 6	0.046 1
Mean	0.116 4	0.042 0
Median	0.113 9	0.042 1
STD	0.016 7	0.008 9

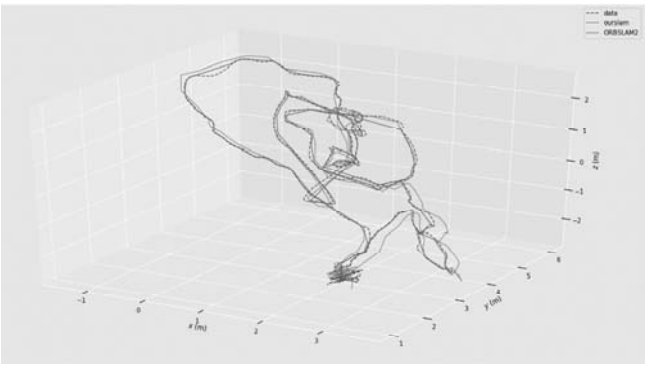


图 7　轨迹对比

4.3　语义地图构建

本文实现选取 TUM 序列,在 ROS 系统下,运行三维语义系统。记录视频序列数据并同时执行 SLAM 以获得摄像机轨迹,使用视频序列和计算的摄像机轨迹执行语义重建。系统可以很好地分割场景,在此序列场景中,出现椅子、墙面和沙发等 14 种对象进行语义分割。其中左下角是我们训练好的语义标签,动态展现语义分割结果。语义信息和几何信息通过贝叶斯融合产生一致的结果,结合语义分割结果使用 Octo-map 模块建立三维语义重建地图,利用 Rviz 可视化工具显示其地图,如图 8 所示。

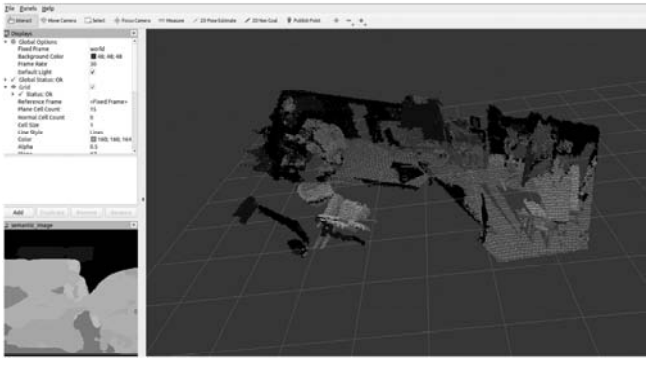


图 8　三维点云地图

5　结　语

本文的主要工作是在传统的 ORB-SLAM2 算法基础上提出新颖的一种半直接法单目视觉 SLAM 方法,结合基于深度神经网络的语义分割算法,来构建实体场景的三维语义建模。与 ORB-SLAM2 算法相比,跟踪线程采用直接方法计算初始姿态估计和解决特征,以此避免特征提取带来的时间消耗。同时采用运动检测模块去除运动物体,特征提取作为单独的线程来执行。然后利用视觉之间的空间信息和改进的金字塔 TF-IDF 匹配方案获得更好的循环闭合检测性能。实验结果表明,本文算法在保持良好实时性的

同时,取得了较好的语义分割网络准确率,而且提高了 SLAM 算法的精度与动态实体场景下系统建模的鲁棒性。

本文研究除了可以用于图像数据采集和构建三维语义地图之外,还可以进行智能导航。三维语义重建架构中的 SLAM 模块,可以提供智能机器的自身位姿以及三维坐标点信息;语义分割网络模块可以通过建立周围环境的三维语义点云信息,从而使得像人类一样了解环境存在的物体信息;最终让智能机器选择合适的路径规划算法。

参 考 文 献

[1] Wolf D, Prankl J, Vincze M. Fast semantic segmentation of 3D point clouds using a dense CRF with learned parameters [C]//2015 IEEE International Conference on Robotics and Automation,2015:4867 - 4873.

[2] Whelan T, Leutenegger S, Renato F. ElasticFusion: Dense SLAM without a pose graph [C]//Robotics: Science and Systems,2015.

[3] 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.

[4] Davison A, Reid I, Molton N. MonoSLAM: Real-time single camera SLAM[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2007,29(6):1052 - 1067.

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

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

[7] Engel J, Koltun V, Cremers D. Direct sparse odometry[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2016,40(3):611 - 625.

[8] Guan P, Cao Z, Chen E, et al. A real-time semantic visual SLAM approach with points and objects[J]. International Journal of Advanced Robotic Systems,2020,17(1):211563099.

[9] Yu C, Liu Z, Liu X, et al. DS-SLAM: A semantic visual slam towards dynamic environments[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2018:1168 - 1174.

[10] Wang Z, Zhang Q, Li J, et al. A computationally efficient semantic SLAM solution for dynamic scenes[J]. Remote Sensing,2019,11(11):1363.

[11] Spla B, Tao Z A, Xiang G C, et al. Semi-direct monocular visual and visual-inertial SLAM with loop closure detection [J]. Robotics and Autonomous Systems,2019,112:201 - 210.

[12] 房立金,刘博,万应才. 基于深度学习的动态场景语义 SLAM[J]. 华中科技大学学报(自然科学版),2020,48(1):121 - 126.

[13] Galvez-López D, Tardos J. Bags of binary words for fast place recognition in image sequences[J]. IEEE Transactions on Robotics,2012,28(5):1188 - 1197.

[14] 李博,杨丹,邓林. 移动机器人闭环检测的视觉字典树金字塔 TF-IDF 得分匹配方法[J]. 自动化学报,2011,37(6):665 - 673.

[15] Takos G. A survey on deep learning methods for semantic image segmentation in real-time[EB]. arXiv:2009.12942, 2020.

[16] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,39(4):640 - 651.

[17] Badrinarayanan V, Handa A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling[EB]. arXiv:1505.07293,2015.

[18] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation[EB]. arXiv: 1505.04597,2015.

[19] Chen L, Papandreou G, Kokkinos I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,40(4):834 - 848.

[20] McCormac J, Handa A, Davison A, et al. SemanticFusion: Dense 3D semantic mapping with convolutional neural networks[C]//2017 IEEE International Conference on Robotics and Automation,2017:4628 - 4635.

[21] Zhou B, Zhao H, Puig X, et al. Scene parsing through ADE20K dataset[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition,2017:5122 - 5130.

[22] Handa A, Whelan T, McDonald J, et al. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM[C]// 2014 IEEE International Conference on Robotics and Automation,2014:1524 - 1531.

[23] Song S, Lichtenberg S, Xiao J. SUN RGB-D: A RGB-D scene understanding benchmark suite[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition,2015: 567 - 576.