Computer Science and Application 计算机科学与应用, 2023, 13(12), 2230-2235
Published Online December 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.1312223

Hans汉斯

视觉语义SLAM中关键帧选取策略的研究

徐　畅，邝　坚

北京邮电大学计算机学院，北京

收稿日期：2023年11月15日；录用日期：2023年12月14日；发布日期：2023年12月21日

摘　要

基于视觉的同步定位与地图构建(视觉SLAM)是目前计算机科学中重要的研究领域，是无人驾驶、环境感知、机器人等领域的重要技术。近些年，随着深度学习的迅猛发展，语义分割作为其核心衍生技术之一，拓展出了非常广泛的应用场景，为人类提供了像素级别的图像理解。为了结合语义分割与视觉SLAM，探索语义分割在视觉SLAM中的应用，本文基于ORBSLAM2与SegNet语义分割网络，探讨并提出一种在语义SLAM中，满足实时语义信息获取要求的关键帧选择策略。并通过语义延迟性能测试，结果表明，改进后的选择策略能保证使用的关键帧的语义信息与其他线程使用的帧是较为接近的，并且延迟性能优于传统的顺序关键帧选取策略。

关键词

软件工程，视觉SLAM，语义SLAM，关键帧，语义分割

Research on Key Frame Selection Strategy in Visual Semantic SLAM

Chang Xu, Jian Kuang

School of Computer Science, Beijing University of Posts and Telecommunications, Beijing

Received: Nov. 15th, 2023; accepted: Dec. 14th, 2023; published: Dec. 21st, 2023

Abstract

Visual-based simultaneous localization and map Construction (Visual SLAM) is an important research field in computer science, and it is an important technology in the fields of unmanned driving, environmental perception, robotics, etc. In recent years, with the rapid development of deep learning, semantic segmentation, as one of its core derivative technologies, has expanded a very wide range of application scenarios, providing pixel-level image understanding for human

文章引用: 徐畅, 邝坚. 视觉语义 SLAM 中关键帧选取策略的研究[J]. 计算机科学与应用, 2023, 13(12): 2230-2235.
DOI: 10.12677/csa.2023.1312223

beings. In order to combine semantic segmentation with visual SLAM and explore the application of semantic segmentation in visual SLAM, based on ORBSLAM2 and SegNet semantic segmentation networks, this paper discusses and proposes a key frame selection strategy that can meet the requirements of real-time semantic information acquisition in semantic SLAM. Through the semantic delay performance test, the results show that the improved selection strategy can ensure that the semantic key frame information used is close to the current tracking frame, and the delay performance is better than the traditional sequential key frame selection strategy.

Keywords

Software Engineering, Visual Slam, Semantic Slam, Key Frame, Semantic Segmentation

Copyright © 2023 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

同步定位与建图算法(SLAM)是机器人利用相机、激光和里程计等传感器来构建未知环境地图并实现自我定位的过程，它在机器人自主完成任务中起着关键作用。传统的 SLAM 根据传感器设备获取环境信息的不同，可以主要分为两类，一类是基于雷达的激光 SLAM。激光 SLAM 主要使用激光雷达测距原理来感知环境，获得机器人与周边环境的距离信息。这类方法获取信息单一，构建地图较为简单，硬件成本高。但其优势是结果精确。另一类是基于摄像头的视觉 SLAM，视觉 SLAM 主要是通过单目、双目或 RGBD 相机传感器获取周边环境的图像信息。其成本低，性价比高，获取的信息量大[1]。

传统的 SLAM 技术只能构建关于环境的几何结构地图，帮助机器人执行一些低层次的定位和导航任务。以 ORBSLAM [2]为例，机器人无法通过以 ORB 特征为路标点的稀疏点云地图理解更高级的物体意义，同样难以满足智能环境交互的要求。但随着近年来深度学习的发展，语义分割作为一项重要的分支在计算机视觉领域崛起。语义分割通过卷积神经网络获取图像中像素级别的信息，理解空间中目标物体的坐标点和语义属性，语义信息同时包含空间环境的几何信息和高层次信息。语义信息与视觉 SLAM 相结合，这让机器人拥有了在不同的环境中帮助人类完成复杂工作的能力，使得机器人能够感知和理解所处环境的信息。这对于人机交互、无人驾驶的实现有着非常重大的意义。

2. 相关工作

2.1. 语义分割

随着计算机视觉的迅速发展，以卷积神经网络为基础实现的语义分割已经成为计算机视觉领域的重要分支。语义信息的提取主要通过神经网络来完成，其前置任务是目标检测，不仅要将目标检测出来，还要与其他物体做出准确的类别分割。剑桥大学提出的 SegNet [3]网络使用最大池化索引在解码器中进行上采样，提高了输出准确率。U-Net [4]网络允许解码器通过跳转连接体系结构汇集编辑器丢失的特征，从而解决了信息丢失的问题。PSPNet [5]模型设计了金字塔池化模块，并通过引入空洞卷积和扩张策略来修改 ResNet [6]架构，在多个数据集上表现出良好的性能。

本文在实验期间，主要以 SegNet 为例，在此基础上，对比不同关键帧的优化选取方法的效果。SegNet 是一种用于语义分割的深度全卷积神经网络结构，其核心由一个编码器网络和一个对应的解码器网络以

及一个像素级分类层组成。解码器使用在对应编码器的最大池化步骤中计算的池化索引来执行非线性上采样，这与反卷积相比，减少了参数量和运算量，而且消除了学习上采样的需要。其网络结构如图 1 所示。

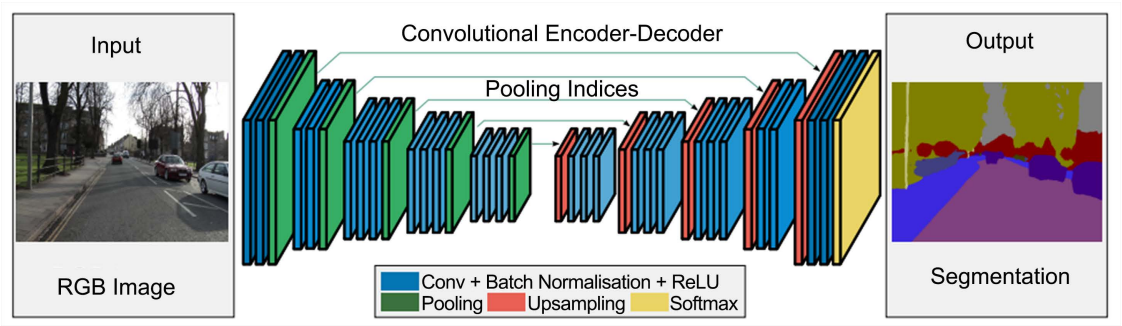

Figure 1. SegNet architecture [3]
图 1. SegNet 结构[3]

SegNet 的应用包括计算机视觉领域的各种任务，如物体识别、图像分割、自动驾驶、医学图像处理等。它在许多视觉任务中表现出色彩，但由于其复杂性，通常需要计算资源来进行训练和推理。一般来说，在普通的桌面或笔记本电脑上，如不通过 GPU 加速，通常一帧图像的处理时间在 200~500 ms 范围内。

2.2. 关键帧序列以及语义处理线程

在视觉 SLAM 中，关键帧是一个重要的概念。关键帧指的是那些对于 SLAM 系统来说特别重要的图像帧，它们通常包含着足够多的信息，能够用于地图构建和相机位姿估计。这些帧通常选自整个图像序列，以确保系统可以有效地定位相机并构建地图。关键帧的选择和使用在视觉 SLAM 系统的设计和优化中非常重要，因为它们直接影响系统的稳定性、鲁棒性和实时性。

关键帧的选择通常基于一些标准，以确保它们对 SLAM 系统的性能有贡献。这些标准可以包括图像质量、视角变化、特征点分布等。一般来说，选择具有良好视角覆盖和区分度高的帧作为关键帧。以 ORBSLAM2 [7]为例，选取规则有最基本的几条：

1) 若仍在定位，不选择；
2) 若局部地图处于全局闭环情况下，不选择；
3) 若距离上一次回环较近，不选择；
4) 跟上次选择相比已经过去 N 帧，N 为最大指定值，保证不超过上限；
5) 跟上次选择相比至少过去 M 帧，且建图线程处于空闲状态，保证下限；
6) 关键帧序列的关键帧数量不超过 3 个、

按上述规则确定了当前帧是关键帧，那么就需要将其输入其他线程使用，例如建图线程以及回环检测线程。但是在输入其他模块之前，需要将关键帧的观测信息进行更新，关键帧信息更新这里不再赘述。

本文的重点在于关键帧序列确定后，如何选择更有效的关键帧获取最新的语义信息，以保证其他线程使用的语义信息的较新的。在 ORBSLAM2 中，基本的线程模型与架构主要由三个线程组成，分别是追踪线程、建图线程、回环检测线程。

在视觉语义 SLAM 中，通常的解决方案是前端通过卷积神经网络语义分割获得语义信息。本文在 ORBSLAM2 的基础上，新增了一个语义处理线程，这个线程主要负责产生语义信息保存。具体来说，语

义线程会通过一定的策略挑选出关键帧，作为模型处理得到语义信息的数据基础，随后调用语义模型处理关键帧，并将得到的语义信息保存并返回这两个关键帧的语义结果。图 2 展示了加入语义线程后的系统框架。

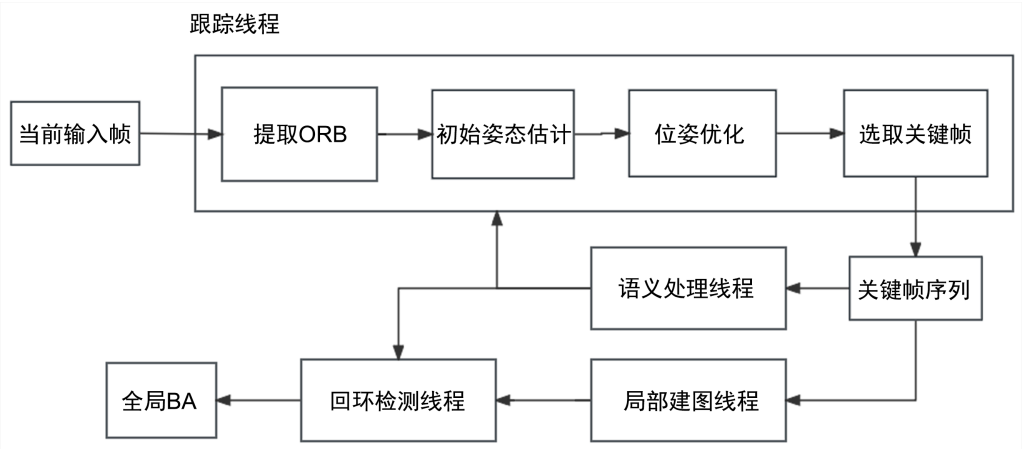

Figure 2. New system framework for semantic processing threads
图 2. 新增语义处理线程的系统框架

3. 关键帧序列选取策略的研究

3.1. 英文缩写

语义处理线程在关键帧序列中挑选关键帧后，需要生成语义掩膜。生成语义掩膜的过程是十分耗时的，以 SegNet 为例，在普通 CPU 平台上，单帧图像处理生成语义掩膜要耗时几百毫秒。ORBSLAM2 在跟踪线程中获取当前帧的速度约为几十毫秒，在并行的情况下，其他线程也会使用语义处理线程得到的语义信息。在处理速度相差较大的情况下，如果关键帧选择策略是按照图像顺序依次挑选的，则跟踪中的当前帧可能无法获得最新的语义信息。可以大概计算出，每次单帧图像语义信息处理后，已有约 10 倍左右当前帧输入。如图 3 所示，考虑到处理速度范围误差，两者之间的延迟会呈现增长趋势，导致其他线程使用的关键帧语义信息与当前帧的误差越来越大。

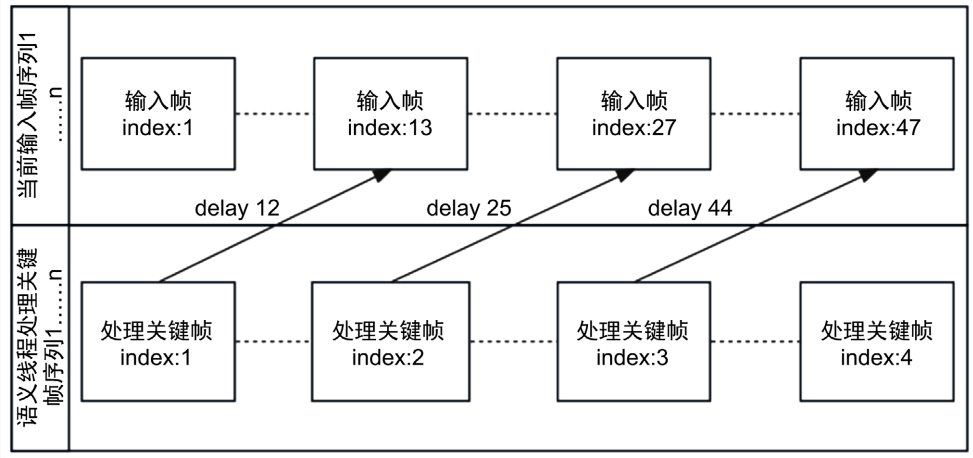

Figure 3. The delay of input frames and semantic processing keyframes
图 3. 输入帧与语义处理关键帧的延迟

在传统的解决方案中，通常是按照图像顺序选定每一帧关键帧，将其进入语义处理线程并得到语义信息。若语义处理线程耗时较大，则这种方式会导致一定误差。因此，研究一种新的关键帧序列选取策略，可以使其他线程得到尽可能新的语义信息是有必要的。

正如图 3 所示，当语义处理线程完成第 1 个关键帧的处理时，当前输入帧序列可能已经到达了第 13 帧。当语义处理线程完成第 3 个关键帧的处理时，当前输入帧序列可能已经到达了第 47 帧。当前输入帧与当前处理完语义信息的关键帧之间的帧数差可以作为衡量延迟的一种方法，比较明显的是关键帧 id 与当前帧 id 的差值，即定义：

$$Framediff = \Delta\left(Input_i, KeyFrame_j\right) \tag{1}$$

公式中的结果即可衡量当前获取语义信息的关键帧与当前输入帧之间的帧数延迟。在此基础上可知，只需使当前处理的关键帧尽量最新即可。而在 ORBSLAM2 中，关键帧序列通过一个动态数组容器存储，新的关键帧会加入到动态数组尾部。通过 C++ STL 类的方法，可以很容易的获取和管理动态数组存储的关键帧。

不同于一般的思路，在关键帧序列选择时，不按照时间顺序选择关键帧，而是分为序列首尾双向选择关键帧。具体做法是，首先选择当前关键帧序列中首部第一个未使用的关键帧，其次选择关键帧序列尾部的最后一帧，选择这两帧图像加入语义处理线程获得当前的语义信息，赋值给关键帧队列中的存在帧作为平均语义信息。在实际实验中，关键帧序列分割一次，约赋给 5-6 帧关键帧语义信息，这样的程度在准确率上也是可以接受的。序列首部的关键帧选择能保证序列按顺序将所有关键帧做语义处理，序列尾部的关键帧选择能保证序列所用的语义信息是与当前输入帧距离最近的。即可以保证公式中的延迟减小。

3.2. 测试与验证

本文将提出的选择策略在 TUM 数据集[8]上进行了测试和验证，该数据集被广泛应用于评估视觉 SLAM。结果显示本策略能有效降低帧数差所带来的延迟。在语义线程中，每处理完一次语义信息，记录当前关键帧与当前输入帧的序数差。得出如表 1 所示数据，对比了顺序选择策略与首尾选择策略随当前帧变化的延迟。数据表明使用首尾策略后，语义帧延迟增长速度减少非常明显，而顺序策略下，语义帧延迟会以接近线性的形式增长。

Table 1. Comparision of delays
表 1. 延迟对比

策略	第 13 帧延迟	第 35 帧延迟	第 47 帧延迟	第 71 帧延迟
首尾策略	11	13	15	15
顺序策略	10	28	37	59

作为机器人状态估计的基本组成部分，SLAM 的速度直接影响到更高级任务的顺利执行。因此，本课题在 TUM 数据集的 fr2/desk 序列上测试了不同系统运行时处理关键帧的平均时间成本，即图像关键帧经过语义处理线程后，获取语义信息处于可使用状态的时间成本。并与其他系统进行了比较。实验耗时结果和硬件平台如表 2 所示。由于 DynaSLAM 系统使用像素级语义分割网络，因此其每帧的平均时间成本较高。YOLO-SLAM 由于系统架构优化和硬件性能等限制，速度非常慢。ORB-SLAM2 与 MaskRCNN 结合如果不使用关键帧优化策略，主要耗时时间在语义分割上，其他线程要等待语义分割的结果。通过关键帧策略优化，提高了关键帧图像获得语义信息的速度，进而提升了处理图像所花费的平均时间。

Table 2. Comparison of average cost of processing images
表 2. 处理图像所花费平均成本对比

方法	处理图像花费平均时间(ms)	硬件平台
YOLO-SLAM	685.33	Intel i5-4228U CPU
Dyna-SLAM	210.33	Nvidia Tesla M40 GPU
ORB-SLAM2	30.12	Intel i7-12650h + RTX 3060Ti
ORB-SLAM2 + SegNet	74.38	Intel i7-12650h + RTX 3060Ti
ORB-SLAM2 + SegNet 策略	58.77	Intel i7-12650h + RTX 3060Ti

4. 结论

本文提出了一种关键帧选择策略，旨在结合语义分割与视觉 SLAM，以满足实时语义信息获取的需求。该策略基于 ORBSLAM2 与 SegNet 语义分割网络，其中神经网络对语义信息的处理相对较耗时，可能导致当前其他线程使用的语义信息不是最新的情况。改进后的关键帧选择策略能够确保所使用的语义信息与其他线程处理的帧接近，从而有效降低了延迟性能。与传统的顺序关键帧选择策略相比，这一改进带来了更好的实时性，更好地满足了语义信息获取的需求。通过在 TUM 数据集上的测试和验证，进一步证明了其有效性，结果显示，该策略能够降低由于帧数差引起的延迟。

参考文献

[1] 王妍. 基于语义分割的室内动态视觉 SLAM 与回环检测研究[D]: [硕士学位论文]. 西安: 西安理工大学, 2023. https://doi.org/10.27398/d.cnki.gxalu.2023.000286

[2] Mur-Artal, R., Montiel, J.M.M. and Tardos, J.D. (2015) ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, **31**, 1147-1163. https://doi.org/10.1109/TRO.2015.2463671

[3] Badrinarayanan, V., Kendall, A. and Cipolla, R. (2017) Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 2481-2495. https://doi.org/10.1109/TPAMI.2016.2644615

[4] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention-MICCAI* 2015: 18*th International Conference*, Munich, October 5-9 2015, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[5] Zhao, H., Shi, J., Qi, X., *et al*. (2017) Pyramid Scene Parsing Network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 2881-2890. https://doi.org/10.1109/CVPR.2017.660

[6] He, K., Zhang, X., Ren, S., *et al*. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[7] Mur-Artal, R. and Tardós, J.D. (2017) Orb-Slam2: An Open-Source Slam System for Monocular, Stereo, and RGB-D Cameras. *IEEE Transactions on Robotics*, **33**, 1255-1262. https://doi.org/10.1109/TRO.2017.2705103

[8] Jürgen, S., Nikolas, E., Felix, E., *et al*. (2012) A Benchmark for the Evaluation of RGB-D SLAM Systems. 2012 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Portugal, 7-12 October 2012, 573-580. https://doi.org/10.1109/IROS.2012.6385773