同济大学

毕业设计(论文)开题报告

课题名称	基于深度学习的语义 SLAM 研究与实现
副 标 题	
学院（系）	计算机科学与技术系
专　　业	计算机科学与技术
学生姓名	姚福飞　　学　号　1452275

2020　年 2 月 24 日

一、毕业设计课题背景

1 课题背景

在机器人领域，智能服务机器人逐渐走上了行业的风口浪尖。我们的生活中也出现了越来越多移动机器人的身影。随着传感器技术、人工智能技术和计算技术等的不断提高，智能移动机器人逐渐的融入了人们的学习、生活和工作。无论是服务型移动机器人还是工厂里流水线上的移动装配机器人、物流行业的自动分拣机器人，都离不开 SLAM 实时定位于地图构建这一环节。如何让机器人清晰、高效的知道自己在哪里、周围有什么就成为了制约移动机器人工作能力和运行效率的一个重要条件。

目前，SLAM 技术已经广泛的应用于机器人定位导航、VR/AR、无人机、无人驾驶等领域。如 SLAM 可以辅助机器人执行路径规划、自主探索、导航等任务。智能家用扫地机器人都可以通过用 SLAM 算法高效绘制室内地图、智能分析和规划扫地环境。在 VR/AR 领域，SLAM 技术用于辅助增强视觉效果，构建视觉效果更为真实的地图，从而针对当前视角渲染虚拟物体的叠加效果，使之更真实没有违和感。在无人机领域， SLAM 可以快速构建局部 3D 地图，并与地理信息系统（GIS）、视觉对象识别技术相结合，可以辅助无人机识别路障并自动避障规划路径。在无人驾驶领域，SLAM 技术可以提供视觉里程计功能，并与 GPS 等其他定位方式相融合，从而满足无人驾驶精准定位的需求。

近年来由于深度学习(Deep Learning)的快速发展，尤其是全卷积神经网络(Fully Convolutional Network，FCN)的出现，实现了从图像像素到像素类别的变换，结合语义的 SLAM 技术得到了快速的发展。

构建语义地图都是一个大家都一致认同的发展方向，主要是因为目前视觉 SLAM 方案中所采用的图像特征的语义级别太低，造成特征的可区别性太弱。另外，采用当前方法构建出来的点云地图对不同的物体并未进行区分。这样的点云地图因为包含信息不足，再利用性十分有限。现阶段，语义 SLAM 的难点在于怎样设计损失函数，将深度学习的检测或者分割结果作为一个观测，融入 SLAM 的优化问题中一起联合优化，同时还要尽可能提高处理速度增强实时性。

在复杂环境下，怎样利用视频或时序图像中的语义信息，来增强移动机器人 SLAM 系统的实时定位和姿态确认一直是语义 SLAM 技术上一个亟待解决的问题。这一问题不仅涉及到复杂环境下的目标检测和语义识别，还涉及到因时序图像的帧冗余性和不确定性引发的语义分割标签的不稳定性。

本选题是希望在这样的背景下，通过分析研究当前各类算法框架的特点，探索出一种基于深度学习的 SLAM 算法，以实现提高定位准确度的目的。

2 文献综述

2.1 图像语义分割

语义分割是一项重要的密集预测任务，其中预测的目标是图像中每个像素的已知类的后验分布。目前出现的所有方法可大致分为传统方法和深度学习方法两大类。

传统图像语义分割根据灰度、色彩、空间纹理、几何形状等特征把图像划分成若干个互不相交的区域，使得目标与背景分离，一般采用马尔科夫随机场(Markov Random Fields, MRFs)[1]和条件随机场(Conditional Random Fields, CRFs)[2]来构建概率图模型，并使用图论的方法来求解。

传统图像语义分割方法利用浅层视觉特征进行图像目标分割，然后使用人工标注语义信息，来完成图像理解任务。随着深度学习技术的兴起，研究者们开始将深度学习模型引入到传统的语义分割方法中，即在利用传统方法分割出目标区域的基础上，进一步采用卷积神经网络等方法学习目标特征并训练分类器，对目标区域进行分类，从而实现目标区域的自动语义标注。1998 年，Lecun 最早提出了 LeNet 网络[3]，并设计了卷积神经网络的 3 层结构:卷积层、池化层、非线性层。2012 年，Hinton 研究组提出了 AlexNet[4]，首创了深度卷积神经网络模型。该网络在 LeNet 的基础上调整了网络架构并加深了网络深度。2015 年， Karen Simonyan 等提出无监督积神经网络 VGG[5]， VGG 网络在第一层使用了感受域更小的卷积层，使得模型的参数更少，非线性更强，也因此使得决策函数更具区分度，模型更好训练。Jonathan Long 等提出的全卷积网络(Fully Convolutional Networks, FCN)[6]在神经网络的全连接层的位置使用卷积层，并使用了反卷积(deconv)和跳级(skip)结构，使得网络在直接输出语义预测结果的同时拥有更好的鲁棒性和精确性。Liang-Chieh Chen 等使用深度神经网络(Deep Convolutional Nets)和全连接条件随机场(Fully Connected CRFs)[7]获得比较精细的分割结果。Kaiming He 等提出的 Mask R-CNN[8]在 Faster R-CNN[9]的基础上加入了语义分割结构，并取得较好的实验效果。

2.2 视觉 SLAM 技术

SLAM(Simultaneous Localization And Mapping)的任务是在构建环境模型的同时估计在其中运动的机器人的运动状态。SLAM 是移动机器人实现自主化的一项基本任务，通常使用激光雷达、声呐等距离传感器来感知外部环境，使用普通相机作为传感器的 SLAM 系统叫做视觉 SLAM，常用的传感器还有 RGB-D 相机、惯性传感器，仅使用视觉相机作为外部传感器来构建 SLAM 系统叫做视觉 SLAM。应用于视觉 SLAM 的计算机视觉技术包括显著特征的检测、描述以及匹配，图像的识别与检索等。视觉 SLAM 问题的解决方案通常分为概率滤波器法、运动恢复结构法(Structure from Motion, SfM)、生物启发技术三种。

目前为止大多数的 SLAM 解决方案都是基于概率技术的。例如：扩展卡尔曼滤波(Extended Kalman Filter，EKF)[10]，极大似然（Maximum Likelihood，ML）和最大期望(Expectation-Maximization，EM)[11]。当要联合考虑机器人和地图不确定性的时候，扩展卡尔曼滤波和最大期望(Expectation-Maximization，EM)是最常用的，因为它们在这种情况下取得最好的效果。但在大场景中的导航或者回环检测中添加信息的能力有限。以增量的方式构建地图的方法最先由 Guivant 提出[12]，它还给出了随机地图的概念，并且使用扩展卡尔曼滤波器的方法提升了 SLAM 问题的精度。

Pollefeys 在 2004 年提出了从运动恢复结构(Structure from Motion，SfM)技术从图像序列中计算相机位姿和场景的三维结构[13]。SfM 起源于图形学和计算机视觉，标准的处理过程是从输入的图像中提取显著性特征，进行匹配，使用即最小化重投影误差（Bundle Adjustment，BA）非线性优化技术来最小化重投影误差[14, 15]。Nistér 在 2004 年提出了视觉里程计（Visual Odometry，VO）这一概念[16]。视觉里程计可以同时确定每一帧下相机的位姿和特征点在三维世界坐标下的位置。视觉里程计在每一帧上能够处理比概率的方法多得多的特征点。

受到生物启发而构建的 SLAM 系统也是一类常见的视觉 SLAM 解决方法。Milford 模拟鼠类感知环境创建了名为 RatSLAM 的系统[17]。RatSLAM 能够使用单个相机生成复杂环境的一致稳定表示,实验结果显示在室内和室外环境中的实时任务中都有很好的性能[18, 19]。Collett 研究了蚂蚁在沙漠中的行为，以分

析它们如何通过视觉路标点进行引导，而不是信息素轨迹[20]。这类视觉路标点引导的方式被证明是可行的且容易在机器人系统上实现的。

2.3 语义 SLAM 和深度学习

很久之前，研究者们就试图将物体信息结合到 SLAM 中。例如[21, 22] 就曾把物体识别与视觉 SLAM 结合起来，构建带语义的地图。另一方面，把标签信息引入到 BA(Bundle Adjustment)或优化端的目标函数和约束中，可以结合特征点的位置与标签信息进行优化[23]。Flint 等人提出了应用于室内场景的在线 SLAM 模型[24]，该模型利用曼哈顿世界假设进行主要平面的分割。2015 年，斯坦福的 Vineet 等人首次实现了一个接近实时的系统[25]，能同时进行建图和语义分割，展示了把语义 SLAM 推向实用的可能性。但是这些方法采用的物体识别或者语义分割的方法都是基于传统工具的。Bowman 引入了最大期望(Expectation Maximization, EM)[26]估计来把语义 SLAM 转换成概率问题，其优化目标仍是重投影误差。

随着深度学习的发展，研究者们开始将神经网络应用于 SLAM 领域，目前深度学习在 SLAM 中的应用有位姿估计、重定位、回环检测、图像深度计算以及 SLAM 的语义生成等方面[27， 28, 29，30]。Yasin Almalioglu 等使用神经网络实现了位姿估计和深度计算[31]。Keisuke Tateno 等将深度学习方法与传统方法融合[32]，使用神经网络预测图像深度和像素类别，实现了一个实时单目 SLAM。Michael Strecke 等用 SDF 表示将多目标跟踪表示为 RGB-D 图像的直接对准[33]，并使用概率方法直接进行数据关联和遮挡处理，并使用期望最大化(Expectation Maximization, EM)框架将 Mask R-CNN 的识别和分割结果融合到 SLAM 里，该方法在动态的室内场景取得较好效果。

3 基于深度学习的语义 SLAM 可行性

SLAM 实际应用进程中的一大障碍就是算力限制，传统的基于特征的方法由于有特征的提取、匹配以及检索等操作，导致语义和实时性不可兼得。现有的实时 SLAM 多采用特征点方法，这样构建的稀疏地图不能存储语义，无法用于机器人与环境交互。

深度学习方法在许多领域的应用中都有比较好的表现。相对于传统方法，深度学习的优点就是不需要提取特征，且有非常突出的非线性能力。深度学习在计算机视觉领域的应用已经取得很大的成就，比如在物体检测、语义分割、图片分类等中表现都很好，也有学者尝试将深度学习应用于 SLAM 领域，比如位姿估计、重定位、回环检测、深度预测等任务[27， 28, 29，30]，深度学习现阶段已经发展出许多模型压缩算法，这些研究工作为深度学习在 SLAM 领域的应用提供了基础，这里将尝试进行深度学习在语义 SLAM 中应用的研究。

二、毕业设计方案介绍（主要内容）

1 课题拟研究内容

本课题拟从现阶段有的 SALM 和深度学习应用研究出发，参考一些理论基础方法，探索将深度学习应用在语义 SLAM 各个任务的方法，并结合实际应用场景和现有移动平台算力，研究在不同应用场景中各个任务的均衡方法。

2 初步方案介绍

2.1 研究深度学习用于深度估计可行性

由于语义地图需要使用稠密地图，使用传统方法计算深度往往是 SLAM 系统中占据 CPU 时间最多的任务，这里尝试使用深度学习方法计算深度，并与传统方法进行比较，选取其中一种比较好的方法用于构建 SLAM 系统。

2.2 实验研究视觉里程计方法

SLAM 中一个比较重要的任务就是视觉里程计，现有方法既有基于深度学习的方法又有传统的基于特征的方法，基于特征的方法比较成熟，在双目的立体 SLAM 中已经又比较高效可靠的方法，这里需要从鲁棒性、是否便于优化等方面进一步研究是否需要使用深度学习方法。

2.3 构建基本 SLAM 系统

完成上述任务后构建基本 SLAM 的最后一步是需要寻找一种高效、便于优化的地图表示方法，目前主要的地图表示方法有点云地图、八叉树地图、TSDF(Truncated Signed Distance Function)等，在构建 SLAM 的过程中需要进一步研究使用哪一种表示方法。

2.4 地图语义的添加

这一步需要研究一种比较高效的语义分割方法，并探索将语义融合到地图中。现有较好的图像语义生成方法是深度学习方法，这里在地图中融入语义需要找一个合理的优化方法，待进一步研究。

2.5 其他功能

语义 SLAM 除了上述功能外还需要有丢失重定位和回环检测，在实现上述基本任务后视情况决定是否需要探索实现这些功能。

三、 毕业设计的主要参考文献

[1] Stan Z. Li. “Markov random field models in computer vision.” European conference on computer vision. Heidelberg: Springer,1994:361-370.

[2] John Lafferty，Andrew Mccallum，Fernando C.N. Pereira. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.” International Conference on MaG chine Learning. Wiliamstown: Morgan Kaufmann,2001:282-289.

[3] Lecun Y, Bottou L, Bengio Y, et al. “Gradient based learning applied to document recognition.” Procedings of the IEEE,1998,86(11):2278-2324.

[4] Alex Krizhevsky，I Sutskever，G Hinton.“Imagenet classification with deep convolutional neural networks.” Advances in neural information processing systems. Nevada:ACM,2012:1097-1105.

[5] Karen Simonyan, Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” arXiv preprint arXiv:1409.1556,2014.

[6] Jonathan Long, Evan Shelhamer, Trevor Darrell. “Fully Convolutional Networks for Semantic Segmentation.” arXiv preprint arXiv:1411.4038v2.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L.Yuille. “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs.” arXiv preprint arXiv:1412.7062v4.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick. “Mask R-CNN.” arXiv preprint arXiv:1703.06870v3.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun.“Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” arXiv preprint arXiv:1506.01497v3.

[10] S. Thrun, W. Burgard, D. Fox. “Probabilistic Robotics.” MIT Press, 2005.

[11] Montemerlo M, Thrun S, Koller D, et al. “FastSLAM: a factored solution to the simultaneous localization and mapping problem.” In: Proceedings of the AAAI National Conference on Artificial Intelligence, pp. 593-598.

[12] Guivant J.“Efficient simultaneous localization and mapping in large environments.” Dissertation, University of Sydney, Australia.

[13] Pollefeys M, Van L, Vergauwen M, et al.“Visual modeling with a hand-held camera.” Int J Comput Vis, 59(3): 207-232.

[14] Triggs B, Mclauchlan P, Hartley R, Fitzgibbon A.“Bundle adjustment – a modern synthesis.” In: Proceedings of the International Workshop on Vision Algorithms: Theory and Practice, pp. 298-375.

[15] Engels C, Stewénius H, Nistér D.“Bundle adjustment rules.” In: Photogrammetric Computer Vision.

[16] Nistér D, Naroditsky O, Bergen J. “Visual Odometry.” In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1:652-659.

[17] Milford M, Wyeth G, Prasser D. “RatSLAM: a hippocampal model for simultaneous localization and mapping.” In: Proceeding of the IEEE International Conference on Robotics

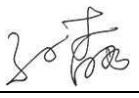
and Automation, 1:403-408.
[18] Milford M, Wyeth G.“Mapping a suburb with a single camera using a biologically inspired SLAM system.” IEEE Trans Robot, 24(5):1038-1053.
[19] Glover A, Maddern W, Milford M, et al. “FAB-MAP + RatSLAM: appearance-based slam for multiple times of day.” In: Proceedings of the IEEE International Conference on Robotics and Automation.
[20] Collett M.“How desert ants use a visual landmark for guidance along a habitual route.” In: Psychological and Cognitive Sciences, 107(25):11638-11643.
[21] A. Nüchter, J. Hertzberg. “Towards semantic maps for mobile robots.” Robotics and Autonomous Systems, vol. 56, no. 11, pp. 915–926, 2008.
[22] J. Civera, D. Gálvez-López, L. Riazuelo, J. D. Tardós, J. Montiel.“Towards semantic slam using a monocular camera.” in Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on, pp. 1277–1284, IEEE, 2011.
[23] N. Fioraio and L. Di Stefano.“Joint detection, tracking and mapping by semantic bundle adjustment.” 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1538–45, 2013.
[24] Flint, D. Murray, I. D. Reid. “Manhattan Scene Understanding Using Monocular, Stereo, and 3D Features.” In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 2228– 2235. IEEE, 2011.
[25] Vineet, O. Miksik, M. Lidegaard, M. Niessner, S. Golodetz, V. A. Prisacariu, O. Kahler, D. W. Murray, S. Izadi, P. Peerez, and P. H. S. Torr. “Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction.” In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pages 75–82. IEEE, 2015.
[26] Sean L. Bowman，Nikolay Atanasov，Kostas Daniilidis. “Probabilistic data association for semantic SLAM.” DOI：10.1109/ICRA.2017.7989203
[27] K.Konda, R.Memisevic.“Learning visual odometry with a convolutional network.”in International Conference on Computer Vision Theory and Applications, 2015.
[28] A.Kendall,M.Grimes,R. Cipolla.“Posenet: A convolutional network for real-time 6-dof camer a relocalization.” in Proceedings of the IEEE International Conference on Computer Vision, pp. 2938–2946, 2015.
[29] Y.Hou, H.Zhang, S.Zhou. “Convolutional neural network-based image representation for vi sual loop closure detection.” arXiv preprint arXiv:1504.05241.
[30] Jamie Watson, Michael Firman, Gabriel J. Brostow, Daniyar Turmukhambetov. “Self-Supervised Monocular Depth Hints.” arXiv preprint arXiv:1909.09051v1 .
[31] Yasin Almalioglu, Muhamad Risqi U. Saputra, Pedro P. B. de Gusmo, Andrew Markham, Niki Trigoni. “GANVO: Unsupervised Deep Monocular Visual Odometry and Depth Estimation with Generative Adversarial Networks.” arXiv preprint arXiv:1809.05786v3.
[32] Keisuke Tateno, Federico Tombari, Iro Laina1, Nassir Navab. “CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction.”
[33] Michael Strecke, Jörg Stückler. “EM-Fusion: Dynamic Object-Level SLAM with Probabilistic Data Association.” arXiv preprint arXiv:1904.11781v1.
[34] 高翔,张涛等. 《视觉 SLAM 十四讲：从理论到实践》[M]. 北京:电子工业出版社. 2017:1-377。
[35] Richard Hartley,Andrew Zisserman. “Multiple View Geometry in Computer Vision.” Cambridge University Press. 2003.
[36] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. J. Kelly, and A. J. Davison. SLAM++: Simultaneous localisation and mapping at the level of objects[J]. In 2013 IEEE Conference on Computer Vision and Pattern Recognition, pages 1352–1359, June 2013.
[37] Raul Mur-Artal,Juan D. Tardos. “ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras.” IEEE Transactions on Robotics[J]. 2017(33): 1255 – 1262.

四、审核意见

指导教师审核意见：（针对选题的价值及可行性作出具体评价）

该选题具有较好的应用研究价值，研究方案也具备较好可行性。

同意开题。

指导教师签名 

2020 年 2 月 24 日

专业审核意见：

负责人签名_____

_____年_____月_____日