Journal of System Simulation

Volume 32 | Issue 7 Article 5

7-15-2020

A Survey on Visual SLAM based on Deep Learning

Ruijun Liu
Beijing Key Laboratory of Big Data Technology for Food Safety, School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China;

Xiangshang Wang
Beijing Key Laboratory of Big Data Technology for Food Safety, School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China;

Zhang Chen
Beijing Key Laboratory of Big Data Technology for Food Safety, School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China;

Bohua Zhang
Beijing Key Laboratory of Big Data Technology for Food Safety, School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China;

A Survey on Visual SLAM based on Deep Learning

Abstract

Abstract: Following the development of computer vision and robotics, visual Simultaneous Localization and Mapping becomes a research focus in the field of unmanned systems. The powerful advantages of deep learning in the image processing offer a huge opportunity to the wide combination of the two fields. T*he outstanding research achievements of deep learning combined with visual odometry, loop closure detection and semantic Simultaneous Localization and Mapping are summarized. A comparison between the traditional algorithm and method based on deep learning is carried out. The development direction of visual Simultaneous Localization and Mapping based on deep learning is forecasted.*

Keywords

deep learning, Visual Simultaneous Localization and Mapping, visual odometry, loop closure detection, semantic Simultaneous Localization and Mapping

Recommended Citation

Liu Ruijun, Wang Xiangshang, Zhang Chen, Zhang Bohua. A Survey on Visual SLAM based on Deep Learning[J]. Journal of System Simulation, 2020, 32(7): 1244-1256.

第 32 卷第 7 期 系统仿真学报© Vol. 32 No. 7
2020 年 7 月 Journal of System Simulation Jul., 2020

基于深度学习的视觉 SLAM 综述

刘瑞军*，王向上，张晨，章博华
(北京工商大学计算机与信息工程学院食品安全大数据技术北京市重点实验室，北京 100048)

摘要：随着计算机视觉和机器人技术的发展，视觉同时定位与地图创建已成为无人系统领域的研究焦点，深度学习在图像处理方面展现出的强大优势，为二者的广泛结合创造了机会。*总结了深度学习与视觉里程计、闭环检测和语义同时定位与地图创建结合的突出研究成果，对传统算法与基于深度学习的方法做了对比，展望了基于深度学习的视觉同时定位与地图创建发展方向。*

关键词：视觉同步定位与地图创建；深度学习；视觉里程计；闭环检测；语义同步定位与地图创建
中图分类号：TP391.9 文献标识码：B 文章编号：1004-731X (2020) 07-1244-13
DOI: 10.16182/j.issn1004731x.joss.19-VR0466

A Survey on Visual SLAM based on Deep Learning

*Liu Ruijun**, *Wang Xiangshang*, *Zhang Chen*, *Zhang Bohua*
(Beijing Key Laboratory of Big Data Technology for Food Safety, School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China)

Abstract: Following the development of computer vision and robotics, visual Simultaneous Localization and Mapping becomes a research focus in the field of unmanned systems. The powerful advantages of deep learning in the image processing offer a huge opportunity to the wide combination of the two fields. T*he outstanding research achievements of deep learning combined with visual odometry, loop closure detection and semantic Simultaneous Localization and Mapping are summarized. A comparison between the traditional algorithm and method based on deep learning is carried out. The development direction of visual Simultaneous Localization and Mapping based on deep learning is forecasted.*

Keywords: deep learning; Visual Simultaneous Localization and Mapping; visual odometry; loop closure detection; semantic Simultaneous Localization and Mapping

引言

同时定位与地图创建(Simultaneous Localization and Mapping，SLAM)，是指在陌生环境中，机器实现环境感知、理解和完成自身定位，以及路径规划[1-2]。在某些特殊场合可以代替人工，比如军事，交通，服务业等领域。长期以来，定位是实现路径规划的前提，在定位时，机器的首要任务是对周围环境的感知，然后对其刻画。现有许多针对已知先验环境信息的机器自主定位和地图创建的解决方法[1]。但是在大多情况下，事先获取环境先验信息很困难，因此，需要机器在陌生环境中，移动时一边计算自身位置，一边创建环境地图[3]。这也促进了 SLAM 的研究，随着算法和传感器的发展，SLAM 研究近些年来取得了巨大突破。

SLAM 本质上是一个状态估计问题，根据传感器做划分，主要是激光、视觉两大类。激光 SLAM

收稿日期：2019-08-30 修回日期：2019-12-01；
基金项目：北京市自然科学基金(4202016，9192008)，教师队伍建设-创新团队(IDHT20180507)；
作者简介：刘瑞军(通讯作者 1982-)，男，北京，博士，副教授，研究方向为虚拟现实、图像处理；王向上(1995-)，男，河南，硕士生，研究方向为虚拟现实、图像处理。

的研究在理论和工程上都比较成熟，现有的很多行业已经开始使用激光 SLAM 完成工业工作；而视觉 SLAM (Visual SLAM，VSLAM)是将图像作为主要环境感知信息源的 SLAM 系统，VSLAM 以计算相机位姿为主要目标，通过多视几何方法构建 3D 地图[4]，视觉 SLAM 还处于实验室研究阶段，实际应用较少。SLAM 系统的处理过程一般都是分为 2 个阶段：帧间估计和后端优化[5]，这种处理方式是由 PTAM[6]首先提出并实现的，它区分出前后端完成特征点跟踪和建图的并行化，前端跟踪需要实时响应图像数据，地图优化放在后端进行，后续许多视觉 SLAM 系统设计也采取类似的结构[5]。另外，PTAM 也是第一个在后端优化使用非线性优化的系统，提出了关键帧(keyframes)机制，不用精细处理每一幅图像，而是把几个关键图像串起来优化其轨迹和地图。在整个 SLAM 系统中，帧间估计是根据相邻两帧间的传感器信息获取该时间间隔内的运动估计，后端优化指对之前帧间估计产生的路径累积漂移误差做优化，解决机器检测到路径闭环后历史轨迹的优化问题。与激光 SLAM 相比，视觉 SLAM 对色彩和纹理等信息更敏感，在提高帧间的估计精度和闭环检测方面有巨大潜力。

传统的 VSLAM 分为特征点法和直接法。特征点法从每帧图片中提取分布均匀的角点和周围的描述子，通过这些角点周围的描述子的不变性完成帧间匹配，接下来使用对极几何恢复相机姿态并确定地图坐标，最终根据最小化重投影误差完成相机位姿和地图的微调[7]。而直接法是根据光度误差确定相机位姿和地图的，不用提取角点和描述子，正因为这样，直接法不能表征一张图像的全局特征，直接法的闭环检测面临的累积漂移的消除问题一直没有得到很好的解决[8]。

传统的 VSLAM 方法仍面对以下几个问题：

(1) 对光照较为敏感，在光照条件恶劣或者光照情况复杂的环境中鲁棒性不高；

(2) 相机运动幅度较大时，传统方法的特征点追踪容易丢失；

(3) 对于场景中的动态对象的处理不够理想；

(4) 计算量大，系统响应较慢。

近些年，计算机视觉与深度学习相互结合，促使视觉相关任务的准确率、执行效率以及鲁棒性等实际表现得到了巨大提升，比如实例分类[9]、对象检测[10]、行为识别[11]等领域的表现。VSLAM 系统以计算机视觉为基础，这为神经网络在该领域的应用提供了很大的发挥空间。将深度学习与 VSLAM 结合，有以下优势：

(1) 基于深度学习的 VSLAM 系统有很强的泛化能力，可以在光线复杂的环境中工作；

(2) 对于动态物体的识别和处理更加有效；

(3) 采用数据驱动的方式，对模型进行训练，更符合人类与环境交互的规律。有很大的研究和发展空间；

(4) 采用神经网络可以更好地将图像信息与其他传感器地数据融合，优化帧间漂移；

(5) 更高效地提取语义信息，有益于语义 SLAM[12]的构建和场景语义的理解；

(6) 端到端的 VSLAM，舍去前端点跟踪、后端优化求解的一系列过程，直接输入图像给出位姿估计。

深度学习一般用在 VSLAM 系统的一个或多个环节，基于前述分析，本文对基于深度学习的 VSLAM 方法做了广泛调研。主要针对基于深度学习的视觉里程计[13-14]、闭环检测[15-16]和语义 SLAM 做出了综述，并讨论了基于深度学习的 VSLAM 的研究方向和发展趋势。

1 VSLAM 与深度学习的相关结合

VSLAM 可以构建周围环境的 3D 地图，并计算相机的位置和方向。深度学习和 SLAM 的结合是近几年比较热的一个研究方向，常用深度学习方法替换传统 SLAM 中的一个或几个模块。

DOI: 10.16182/j.issn1004731x.joss.19-VR0466

第 32 卷第 7 期 系统仿真学报 Vol. 32 No. 7
2020 年 7 月 Journal of System Simulation Jul., 2020

1.1 深度学习与视觉里程计

移动机器人完成自主导航，首先需要确定自身的位置和姿态，即定位。视觉里程计(Visual Odometry，VO)通过跟踪相邻图像帧间的特征点估计相机的运动，并对环境进行重建。VO 大多借助计算帧间的运动估计当前帧的位姿。基于深度学习的视觉里程计，无需复杂的几何运算，端到端的运算形式使得基于深度学习的方法更简洁。

Daniel 和 Malisiewicz[17]提出的网络架构完成点跟踪，得到相邻帧间的单应性。如图 1 所示，该模型主要特点是利用 2 个 CNN[18]的协作生成单应性矩阵，完成相机位姿的估计。第一个称为 MagicPoint 网络，提取单张图像的显著点，网络采用类似 VGG[9]风格的结构，对灰度图进行运算，并为每个像素输出一个“点”概率，在最后生成的张量中每个空间位置代表分布在一个局部 8×8 区域的概率，加上一个没有被检测点的垃圾通道，再通过上采样生成带特征点标记的热图。第二个名为 MagicWarp，对 MagicPoint 的输出进行操作，结构与第一个网络类似，将卷积后的两帧图片对应的张量连接起来，通过特征点的匹配来获取单应性矩阵，然后估计跟输入有关的单应性。系统速度快且规模小，可以在单核 CPU 达到每秒处理 30 帧的速度。模型在训练之前，手工标定了一个包含各种特征点的数据集，比如：角，边，以及几何特征不明显的曲线等特征，用于训练 MagicPoint，为了 MagicWarp 参数的学习，设计了一个基于点云的三维相机运动轨迹的数据集，实现从三维向二维空间的映射，找到相邻帧间相机位姿的变化矩阵，发现特征点的单应性，这样也更符合人类的感知。但是该模型不得不面对的一个问题是它只能追踪相邻两帧图像的低级特征点，完成位姿估计，由于对上下文环境信息没有记忆，所以得到结果精度还有待提高。

虽然深度学习已经成为解决很多计算机视觉问题的主流方法，并取得了不错的效果，但是关于 VO 的研究非常有限，尤其在三维几何问题上。因为现有的神经网络结构和预训练模型大多是为了解决识别和分类问题而设计的，这也就驱使深度卷积神经网络(CNNs)从图像中提取更高层次的外观信息。学习图像的表象特征，限制了 VO 只在经过训练的环境中发挥作用，这也是传统 VO 算法严重依赖几何特征而不是外观特征的原因。同时，运动是一个连续的变化过程，理想的 VO 算法应该针对一系列图像的变化和连接来建模，而不是处理单个图像，这也意味着我们要对图像序列学习。

相较于 MagicPoint 和 MagicWarp 仅针对两帧图像的基础几何特征完成单应性估计，Wang 等[19]针对图像序列提出了一种基于深度学习的单目 VO 的 DeepVO 算法，如图 2 所示，直接从原始 RGB 图像提取实例表征完成姿态估计。证明了单目 VO 问题可以通过基于深度学习端到端的方式高效解决，提出了一个 RCNN[20]架构，利用 CNN 学习到的实例表征表示，使基于深度学习的 VO 算法能够适应全新的环境，对于图像序列的序列依赖性和复杂运动，利用深度递归神经网络实现隐式封装和自动学习。DeepVO 由基于 CNN 的特征提取和基于 RNN[21]的特征序列模型两部分组成。提出端到端的 VO 系统架构，如图 2 所示，以视频剪辑或单目图像序列作为输入，每个时间戳内，通过 RGB 图像帧减去训练集的平均 RGB 值进行预处理，将 2 幅连续的图像叠加在一起，形成一个张量，训练深度 RCNN 提取运动信息和估计姿态。采用长短时记忆(LSTM[22])作为 RNN，它通过引入记忆门和记忆单元来学习上下文依赖关系。具体来说，将图像张量输入 CNN，生成一个有效的单目 VO 特征，然后通过 LSTM 进行序列学习，每个图像对在通过网络的每个时间戳产生一个位姿估计。CNN 部分使用非线性单元(ReLU)激活，获取显著特征，使用 0-padding 保留卷积后张量的空间维数。每个时刻使用 LSTM 更新状态状态，为了学习神经网络的超参 θ，损失函数由位置向量 p 和角度变量 φ 组成的均方误差构成：

$$\theta^{*}=\arg\min_{\theta}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{t}\left\|\hat{p}_k-p_k\right\|_2^2+\kappa\left\|\hat{\varphi}_k-\varphi_k\right\|_2^2 \quad (1)$$

式中：$\|\cdot\|$为 2 范数；κ 是一个比例因子用来平衡位置和角度的权重；N 为样本数量；角度 φ 用欧拉角定义。他们采用的这种 VO 方法它不依赖于位姿估计的传统 VO，通过将 CNN 和 RNN 相结合，实现了 VO 的特征提取和序列建模的同时进行，无需对 VO 系统的参数进行仔细调整。

针对 DeepVO 这类数据驱动的网络，常对于隐藏在数据背后的规律比较敏感，所以对于类似标签误差这种信息干扰，DeepVO 具备良好的鲁棒性及拟合能力，体现了数据驱动模型的一个显著优势。虽然基于深度学习的 VO 方法在相机的位姿估计方面得到了一些结果，但是目前还不能取代基于几何的方法，深度学习方法是一个可行的补充，把几何用深度神经网络的特征表示、知识以及模型做有机结合，进一步提高 VO 的准确性和鲁棒性是目前可以预见的发展方向。

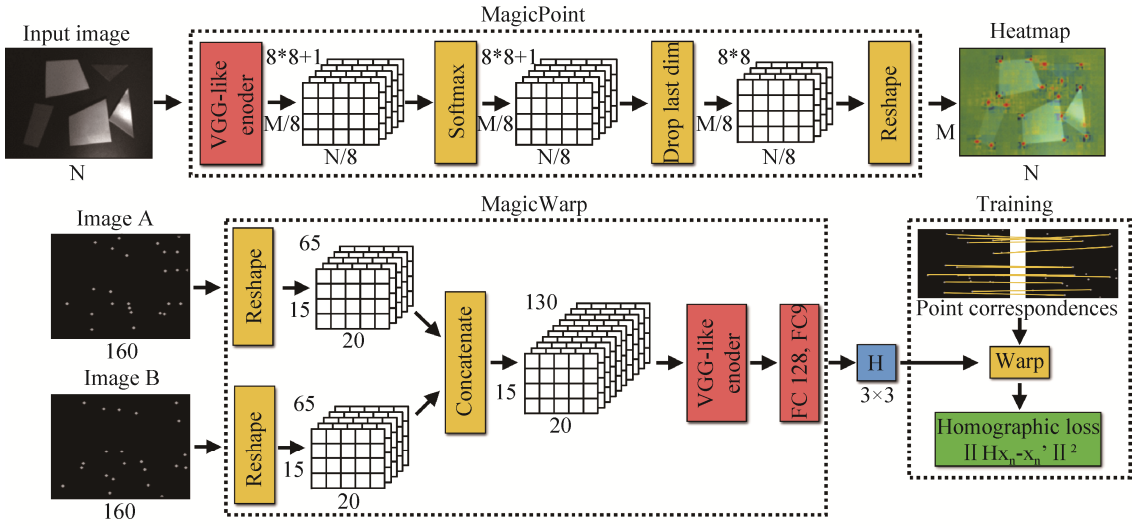

图 1 MagicPoint 和 MagicWarp 结构[17]
Fig. 1 MagicPoint and MagicWarp[17]

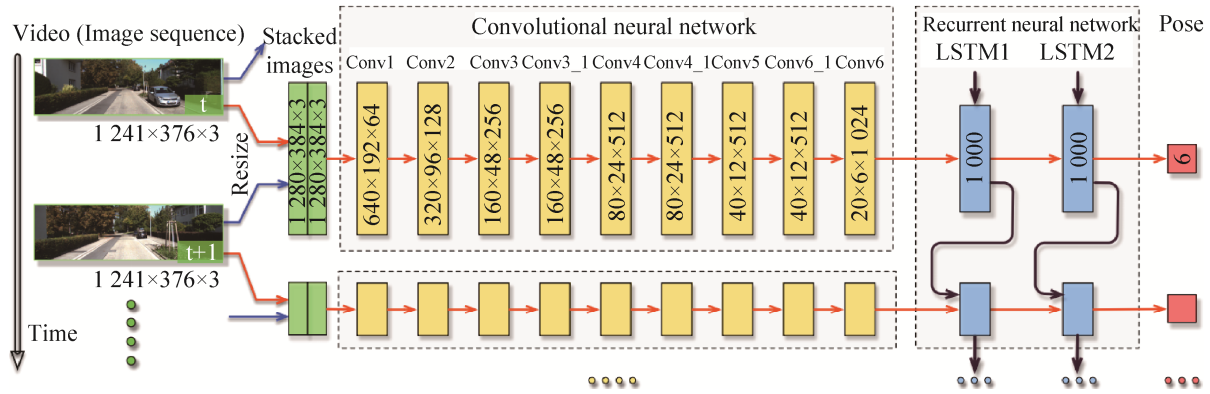

图 2 DeepVO 网络结构[19]
Fig. 2 Architecture of DeepVO[19]

1.2 深度学习与闭环检测

闭环检测应用在机器人建图环节，新采集到一张图像，判断它是否在图像序列中出现过，即确定机器人是否进入某同一历史地点，或者在特征点配准丢失后重新获取一个初始位置。高效的闭环检测是 SLAM 精确求解的基础，帧间匹配主要集中在

第 32 卷第 7 期 系统仿真学报 Vol. 32 No. 7
2020 年 7 月 Journal of System Simulation Jul., 2020

误差累积的消除，图优化算法能够有效地降低累计误差。闭环检测实质上是场景的识别问题，传统 SLAM 的闭环检测通过手工提取的稀疏特征或者像素稠密的特征完成匹配，深度学习则采用神经网络学习图片深层次特征的方法，场景识别率的表现更好。所以，基于深度学习的场景识别方法能够有效提升闭环检测的准确率[23]。

许多研究人员针对闭环检测问题在网络训练和数据处理方面做了部分优化。比如文献[24]提出的方法不是直接对整幅图像做特征提取，而是根据预训练的网络提取图像中的路标区域，然后通过 ConvNet[27]计算每块区域的特征并将其压缩，完成路标区域的特征匹配之后，利用各个路标区域的相似性来计算全局图像之间的相似性，为了降低假阳性出现的概率，模型把路标区域框的范围作为监督条件。该方法显著提升了面对场景中视点变化或局部遮挡的鲁棒性。

为了提高闭环检测的准确率和效率，Yi 等[25]提出了一种 BoCNF 的特征词袋匹配方法，该方法以视觉词袋法为基础，将 CNN 提取到的特征建立视觉词袋，通过 Hash 随机映射[26]将降维的视觉词和词袋特征关联，实现快速准确的场景识别。

BoCNF 基于 BoW (Bag of Visual Word)的想法将提取的视觉特征量化为离线的视觉词汇的视觉词，然后构建用于在线检索匹配图像的倒排索引，如图 3 所示。与其他基于 BoW 的算法类似，他们的框架包括离线阶段和在线阶段，每个阶段包括 2 个部分。在离线和在线阶段的开始，数据库图像或查询图像的 ConvNet 功能分两步提取：地标检测和 ConvNet 特征提取[23]。离线阶段，将数据库图像的 ConvNet 特征提取并把它们量化为视觉词之后，执行倒排索引构建，将视觉词链接到数据库图像。在线阶段，在将查询图像的 ConvNet 特征提取并将其量化为视觉词之后，在称为粗略匹配的阶段中，首先使用倒排索引检索前 K 候选数据库图像，随后精细匹配执行通过基于 Hash 的投票方案在前 K 个候选者中找到与查询图像的最终匹配。

深度 VSLAM 采用学习的方式，以后的发展与人类的感知和思考方式会更加相似。文献[28]构建的模型的输入为图像序列，第一步根据 Local Pose Estimation Network[29]计算图像间的相对位姿，然后通过位姿聚合方法压缩相对位姿信息，把计算结果输入到 Neural Graph Optimization 网络，输出相对位姿的全局绝对位姿信息，通过 Soft Attention 模型提取关键信息，生成各帧之间的相似性矩阵，利用相似性矩阵完成 SLAM 中的闭环检测，输出整个的路径地图，把模型的输出与真实路径的差异作为损失函数。该方法在模拟环境中表现优秀，证明了 Soft Attention[30]模型在闭环检测环节的作用。根据观察该模型的实验表现，在实际环境中没有达到预期的效果，但是端到端的学习模式，和整个模型的数据处理过程，符合人类的感知过程，未来有很大的发展空间。

1.3 深度学习与语义 SLAM

语义 SLAM 在建图过程中获取环境几何信息的同时，识别环境里的独立对象，获取其位置、姿态和个体轮廓等语义信息扩展了传统 SLAM 问题的研究内容，将一些语义信息集成到 SLAM 的研究中，以应对复杂场景的要求[31]。其实语义特征，本质上把局部特征进一步归纳，达到人类可以理解的分类层面。

在深度学习算法流行之前，物体识别方向评价最好的 Bag of Visual Word 算法将之前数据处理提取的 SIFT[32]等特征融合成复杂且全面的特征，再送到分类器进行分类。其实对于深度学习方法，图像的语义表征也是一层层抽象的，根据对神经网络各层输出进行可视化的结果，我们发现底层网络中提取出的特征大部分是点线等的低层语义，中间层的网络将图像特征抽象为一些标识物的局部部件，而在顶层的图像特征上升到了物体的级别，逐层特征提取抽象，这是神经网络的又一突出优势。

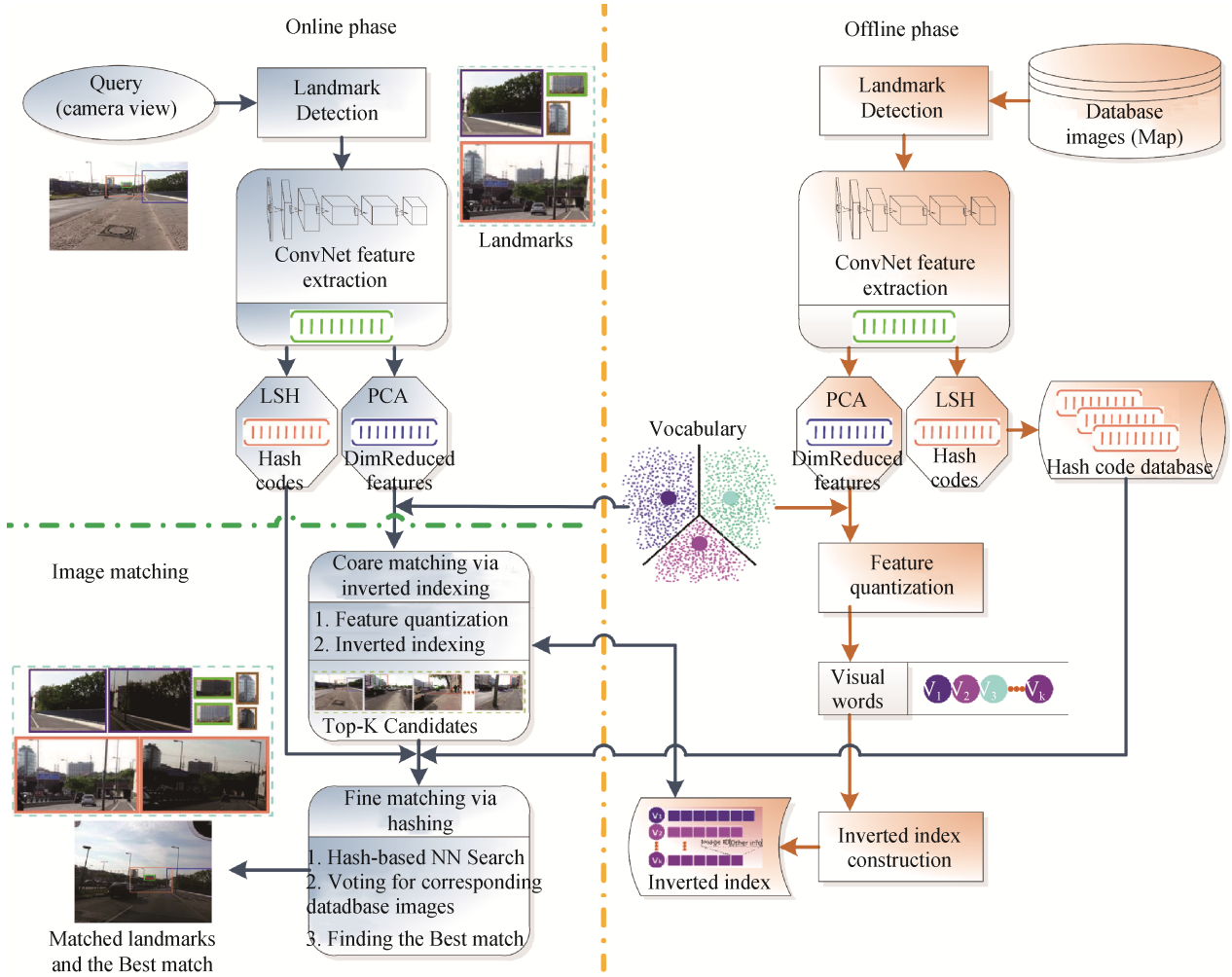

图 3 BoCNF 特征匹配流程[25]

Fig. 3 Generation process of node-link graph[25]

Lei 等[33]以对象分类、语义分割为基准，提出了基于八叉树的球形卷积核 CNN，用于处理 3D 点云数据。将图片转换成原始点云，并基于八叉树的结构做空间划分，把球形卷积核作用到网络的每一层，对特征实现分层下采样，球形卷积核把点 x_i 附近的空间划分为多个体积小块。针对第 j 个邻接点 x_j，卷积核首先确定其对应的体积小块，并通过该小块的权重矩阵 W_k 来计算激活值。该模型对特征分层下采样并根据空间分区构建 3D 邻域，代替常用的 K-NN[34]范围搜索，一定程度上减少了计算和存储成本，适合高分辨率输入。通过神经元和 3D 点空间位置的对应关系确定该点需要使用的球形卷积核，根据这种关联避免训练过程中生成动态卷积核，达到高效、高分辨率的点云学习。

目前的语义 SLAM 研究还处于初级阶段，但其前景广阔。语义 SLAM 的难点在于误差函数的设定，将深度学习的检测或分割结果作为一个观测值，融入 SLAM 的优化问题中一起联合优化，同时还需要做到 GPU 的实时[35]。

Girisha 等[36]针对无人机航拍视频提出了一种语义解析的方法。如图 4 所示，镜头边界检测算法首先用于识别关键帧，随后，通过使用 U-Net[37]，对这些关键帧执行语义分割，航拍数据频率为 29 fps，每个连续帧之间变化是微小的，因此，使用镜头边界检测算法来识别关键帧，以便于单帧和帧与帧之间的分析。而关键帧识别的过程，目的是从连续帧中识别出镜头边界，并且将整个模块用关键帧表示。通过将每一帧划分为 16×16 大小的非重叠网格

https://dc-china-simulation.researchcommons.org/journal/vol32/iss7/5
DOI: 10.16182/j.issn1004731x.joss.19-VR0466

来识别每一帧的镜头边界。采用卡方距离计算相邻两帧之间相应的网格直方图差：

$$d(H_i;H_{i+1})=\sum_i\frac{(H_i(I)-H_{i+1}(I))^2}{H_i(I)} \tag{2}$$

式中：H_i 为第 i 帧直方图；H_{i+1} 为第(i+1+帧直方图；I 为两帧中同一位置的图像块。连续两帧之间的直方图平均差计算如下：

$$D=\frac{1}{N}\sum_{k=1}^{N}d_k(H_i,H_{i+1}) \tag{3}$$

式中：D 为连续两帧的平均直方图差；d_k 为第 k 个图像块之间的卡方差；N 为图像中图像块的总数。在直方图差异大于阈值 T_{shot} 的帧上识别镜头边界，利用语义分割算法对识别出的关键帧进行进一步处理，识别出场景中出现的各种对象(绿化、道路)。U-Net 利用 contracting path 中提取的特征进行反卷积完成区域定位，在 contracting path 实现卷积运算，然后再通过 ReLU 激活函数提取特征。改进后的网络可以处理 256×256 大小的彩色图像(RGB)，而不仅仅是灰度图像，这是通过在每一层使用 3D 卷积操作来实现的。除了最大池化操作，每个层还考虑填充，以保留最相关的特性，以便进一步处理。另外，Bowman 等[38]提出了一种将尺度信息和语义信息融合的理论框架，在语义 SLAM 的概率数据融合上做的工作具有创新性，引入 EM 估计[39]把语义 SLAM 转换成概率问题，优化目标仍然是常用的重投影误差。

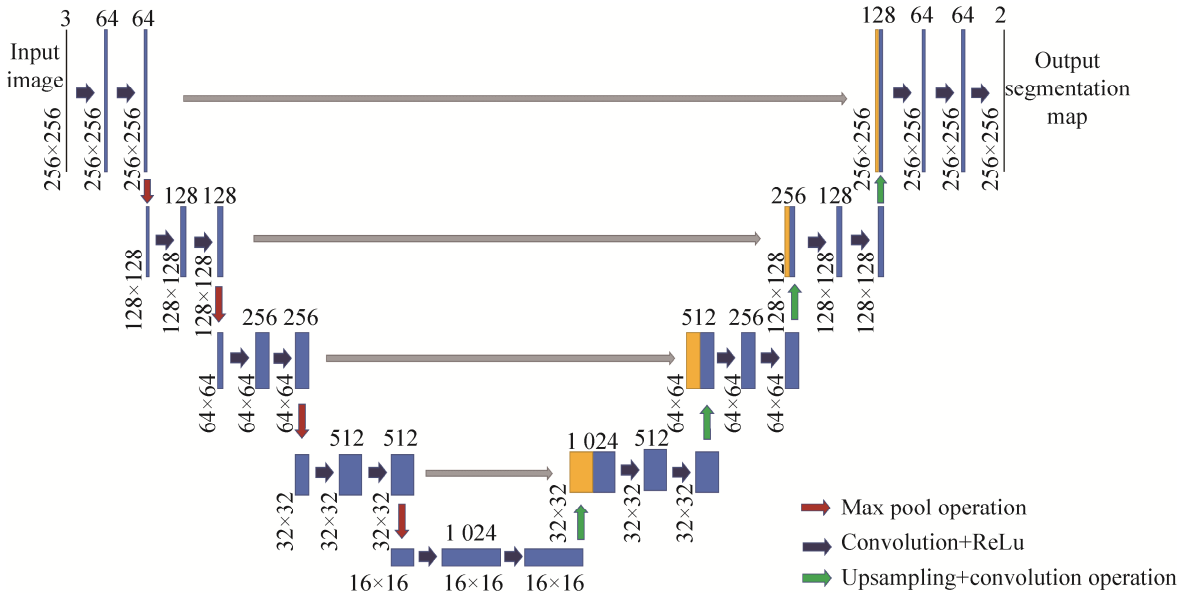

图 4 修改后的 U-Net 结构[37]

Fig. 4 Modified U-Net structure[37]

2 深度学习方法与传统 VSLAM 对比

用深度学习做端到端的 VSLAM 非常直接，能够绕开许多传统 VSLAM 系统中极为麻烦的环节，如外参标定、多传感器频率的匹配，同时可以避开前后端算法中一些棘手的问题。结合深度学习作为一种新的 VSLAM 实现方法有很强的理论意义，但是端到端 VSLAM 的问题也非常明显，如表 1 所示。

VSLAM 这样一个包含很多几何模型数学的问题，通过深度学习去端到端解决，使用数据驱动的形式去学习，在原理上是完全没有依据的，而且也没理由能得到高精度解。另一个很大的问题是模型的泛化性很难得到保证，传统的 VSLAM 系统通常是一个很复杂的结构，从前端到后端每一步操作都有明确的目的，传统方法每个环节都有详细的数学理论作支撑，具备很强的可解释性，但是需要仔细挑选参数。而用高度依赖数据的深度学习去近似 VSLAM 系统，对于某些数据集可能产生不错的效

果，换个场景可能就不那么敏感了，但是如果数据集足够大的话，神经网络还是能够展现其在数据上极强的适应性，所以数据集的体积对神经网络的准确率是一个重要的影响因子。

传统的 VSLAM 仍然面临着对环境的适应性问题，深度学习有望在这方面发挥较大的作用。目前，深度学习已经在语义地图、重定位、回环检测、特征点提取与匹配以及端到端的视觉里程计等问题上完成了相关工作，但对于某些特殊场景的应用需求，还需要进一步的发展。

为了更详细的表现深度 VSLAM 对数据的适应性，这里我们使用 KITTI 数据集的单目序列对几种现有计算 VSLAM 的 VO 解的结果做了简要对比，如表 2 所示。

表 1 VSLAM 算法与基于深度学习的 VSLAM 方法的对比

Tab. 1 Comparison between traditional VSLAM algorithm and VSLAM method based on deep learning

指标	主要参数传统 VSLAM 方法	基于深度学习的 VSLAM 方法
模型的训练周期	+小规模数据集，周期短	-大规模数据集，训练周期长
实现方式	+主要使用数学公式实现，每一步目标明确	-使用数据驱动的形式，数学公式较少，中间变量的意义不易理解
模型泛化能力	Node-针对小规模数据集，模型参数少，泛化能力不强	+大规模数据参与训练，参数多，泛化能力强
鲁棒性	-鲁棒性弱	+鲁棒性强
设计流程	-特征点的提取方式与分类器的结构分开设计	+分类器的结构和特征提取的结构是一个整体
语义识别能力	传统对于语义的识别主要基于几何特征，能级较差	+语义分类、分割能力强

表 2 传统 VSLAM 算法与基于深度学习 VSLAM 方法关于 VO 的准确度在 KITTI 数据集上的对比

Tab. 2 Comparison of accuracy of VO in KITTI dataset between traditional VSLAM algorithm and VSLAM method based on deep learning

方法	标准	00	01	05	07
ORB-SLAM2	平移均方差/%	11.43	107.57	9.04	9.77
	旋转均方差/°	0.58	0.89	0.26	0.36
	绝对轨迹误差	40.65	502.20	29.95	16.04
	相对位姿误差/m	0.169	2.970	0.140	0.105
	相对位姿误差/°	0.079	0.098	0.058	0.047
DSO[40]	绝对轨迹误差	113.18	116.81	47.46	16.72
VISO2[41]	平移均方差/%	12.66	41.93	15.10	10.80
	旋转均方差/°	2.73	7.68	3.65	4.67
	绝对轨迹误差	78.40	127.59	65.73	17.88
	相对位姿误差/m	0.243	1.423	0.245	0.195
	相对位姿误差/°	0.141	0.432	0.131	0.176
CNN-SVO[42]	绝对轨迹误差	17.53	10.77	2.44	11.51
DF-VO[43]	平移均方差/%	2.25	66.98	1.15	0.93
	旋转均方差/°	0.58	17.04	0.30	0.29
	绝对轨迹误差	12.64	6.872	3.75	1.74
	相对位姿误差/m	0.040	1.281	0.025	0.023
	相对位姿误差/°	0.056	0.725	0.038	0.030
SfM-Learner[44]	平移均方差/%	21.32	22.41	12.99	12.61
	旋转均方差/°	6.19	2.79	4.66	6.31
	绝对轨迹误差	104.87	109.51	60.89	20.12
	相对位姿误差/m	0.282	0.660	0.158	0.081
	相对位姿误差/°	0.227	0.133	0.153	0.181

https://dc-china-simulation.researchcommons.org/journal/vol32/iss7/5
DOI: 10.16182/j.issn1004731x.joss.19-VR0466

第 32 卷第 7 期 系统仿真学报 Vol. 32 No. 7
2020 年 7 月 Journal of System Simulation Jul., 2020

VO 作为 VSLAM 系统中必不可少的一环，目的是为了获取局部稳定的运动轨迹，从而减轻后端优化以及闭环检测的压力。我们将深度学习的方法，如 DF-VO、SfM-Learner、CNN-SVO 和基于几何的方法，如 ORB-SLAM2、DSO、VISO2 进行对比，然而，对于深度学习方法，一个有趣的结果是，除了深度模型外，参与测试的基于几何方法在序列 01 上特征点跟踪的表现并不好，这表明深度模型可以作为几何方法的补充，在未来的设计中，可以通过深度学习与几何方法的结合来克服几何方法的失败情形。此外，与几何方法相比，深度 VSLAM 方法具有更好的相对姿态估计。在长序列评价中，针对尺度漂移问题，ORB-SLAM2 表现出较少的旋转漂移，但产生较高的平移漂移，尺度漂移问题，有时这个问题可以通过闭环检测来解决。使用尺度一致的深度预测进行尺度恢复，这在大多数单目 VO/SLAM 系统中缓解了这个问题。结果表明，该方法在长序列上具有较小的平移漂移。更重要的是，深度学习 VO 表现出更小的相对位姿误差，这使深度 VO 方法成为帧到帧跟踪的健壮模块。

传统 VSLAM 方法过于依赖低级别几何特征，例如点、边和平面，仅凭低级特征很难对环境中观察到的标志物作语义表征。而深度学习方法的输入可以是原始的 RGB 图像，不像光流或者深度图像这种经过预处理的数据输入，因为网络可以学习一种有效的特征表示方法，这种学习后的特征表示不仅将原始的高维图像压缩成一个紧凑的表征，还促进了后续环节的计算过程。另外，低级特征往往使闭环检测过于依赖相机视角，这就导致在模糊或重复性较高的纹理环境中失败率较高。基于深度学习的目标识别能够估计出标志物的大小，生成一组便于区分的语义表征，适和与视角无关的闭环检测。在特征地图中识别到多个同类物体时，需要进行关键数据的关联。数据关联和识别一般采用离散方法解决离散性问题，但是传统 VSLAM 解决的是尺度信息的连续优化问题，相比之下，基于深度学习的 VSLAM 可以把传感器数据和语义表征的位置信息融合为一个优化问题，结合尺度信息，语义信息和数据关联。随后再把它拆分成两个相互关联的问题：首先是离散数据的关联和语义表征的种类估计，另外是尺度信息的连续优化问题。推测出的语义表征和传感器姿态影响着关联数据和表征种类的分布，而这反过来又影响传感器-标志物姿态的优化[37]，这也是结合深度学习的 VSLAM 较传统方法的一个明显的优势。

3 展望

就人类的感知方式来说，在面对场景中的对象时，除了可以获取位置信息(三维)外，还可以确定颜色数据(三维)，此外，能够获取语义信息比如表面硬度、实例轮廓、是否可以触摸等信息。但是，若仅凭深度 VSLAM 构建三维点云，还是远远不够的，因此，需要在更高维度上构建内容更为丰富的高阶地图，从而满足各种需求。定位与感知不是 VSLAM 的最终目标，VSLAM 是以精确的定位和感知为前提，完成复杂的任务。这对深度学习和 VSLAM 的结合提出了更高的要求，对深度 VSLAM 训练时，以任务的完成情况为标准进行训练。

3.1 分布式的 VSLAM 建图

分布式的 VSLAM 可以在绝对定位不可用的情况下，适用于多视觉传感器应用的强大工具。在传感器分散情况下，它不依赖于与中央实体的通信。可以将分布式深度学习 SLAM 集成到一个完整的 VSLAM 系统中。为了实现数据关联和优化，现有的分布式 VSLAM 系统在所有传感器之间交换完整的地图数据，从而以与传感器数量平方成正比的复杂性进行大规模数据传输。与之相比，文献[45]提出的方法在两个阶段中实现有效的数据关联：首先，将密集的全局描述符定向地发送到一个传感器，只有当这一阶段成功后，才会把实现相对位姿估计需要的数据再次发送给传感器。所以，数据关联可以通过传感器计数扩展，针对紧凑的场所

表现更突出。使用分散式的位姿图优化方法，交换最小量的数据，这些数据与轨迹重叠是线性的，最后对整个系统的输出做描述，并确定每个组件中的瓶颈。但是，二维图像、三维点云数据随着场景规模的增大，信息规模也会越来越大，存储大量的图片或者点云是不可取的，因此对数据的压缩和存储方式的选择是必要的，比如八叉树的存储方式和网格数据划分等。人类对于场景的记忆是基于城市、街道等先验知识的掌握，通过对关键信息有选择的记忆，不是存储见到的每一帧图像，只用记住去过哪个地点、哪个街区，当再次经过同一地点时，就能够做到场景的对应，这也是未来分布式深度 VSLAM 所面临的问题，即固定存储空间下对非关键信息的剔除，对不同传感器间的联合节点的识别，传感器数据的存储，实现快速的匹配问题。

3.2 高维多传感器数据处理与融合

深度学习的发展为传感器大数据的特征提取与操作提供了新方法。激光测距传感器是传统 SLAM 的传感器，具有高精度，数据采集不受时间限制等优势。Li 等[46]提出 Recurrent-OctoMap，从长期的 3D 激光雷达数据中学习，对语义建图实现 3D 细化，是一种融合语义特征的学习方法，不仅仅是简单地融合分类器地预测。在他们的方法中，将创建的 3D 地图用八叉树[47]表示，并参与后面的计算，将每个节点建模为 RNN，从而获得 Recurrent-OctoMap。在这种情形下，语义建图过程被表达成序列到序列的编码-解码问题。另外，为了延长 Recurrent-OctoMap 观察到的数据持续周期，他们开发了一个强大的 3D 定位和建图的 SLAM 系统，并实现对两周以上的雷达动态数据持续建图。通常用于 3D 语义地图细化广泛的方法是贝叶斯估计，其融合了马尔可夫链之后连续预测概率，但是传统的贝叶斯方法被证实不如 Recurrent-OctoMap 的实验室表现。

机器人在环境变化比较复杂的场景中实现建图时，单一传感器实现的特征检测往往不够全面，可能会有漏检的情况发生，并且单一类型传感器在复杂的环境中实现数据关联成本过高，准确性较低，不能满足 SLAM 在部分特殊场景下的应用需求。VSLAM 对数据关联非常敏感，数据关联可以建立多传感器数据与其他测量数据之间的关系，以确定它们是否有一个公共源。VSLAM 中数据关联用来确定测量数据与地图特征的关系，机器人位姿的不确定性、特征密度的变化、环境中动态特征的干扰以及观测误差的存在使得数据关联的数据处理过程变得很复杂。错误的关联不但会使机器人的定位产生偏差，还会影响到已创建的地图，导致算法发散。Zhang 和 Singh[48]提出的一个利用 3D 激光扫描仪数据、影像数据和 IMU 数据进行运动估计和地图创建的方法，使用一个有序多层的从粗滤配准到精确优化的处理流程。首先使用 IMU 数据做运动预测，然后使用 IMU 和视觉结合的方法估计运动，之后再使用激光雷达进行帧与帧的匹配做更深一步的优化和地图创建，这样，VSLAM 就可以在高动态的运动环境中使用，也可以在黑暗、无纹理、无显著结构的复杂环境里运行。多传感器的数据融合能够确定各传感器测量数据和特征源的对应关系，并确保这些对应关系在复杂环境中能发挥最优性能，而深度神经网络在匹配多类型数据，处理不同频率的数据方面表现出强大的能力，多传感器融合的 VSLAM 是未来的一个重要发展方向。

3.3 自适应的 VSLAM

随着机器人技术的高速发展，VSLAM 也具有更多的实际应用意义。VSLAM 需要相机视野的三维环境信息和相应的轨迹信息，所以 VSLAM 对相机定位的实时以及精度的依赖性较高[49]。随着稀疏矩阵和非线性优化理论在 VSLAM 中广泛的应用，逐渐提出了许多 VSLAM 实现方案，比如传统的 LSD-SLAM[50]、ORB-SLAM、RGBD-SLAM[51]等方案，基于深度学习的 VSLAM 如 DeepVO、SFM-Net[52]等算法，然而，现存大部分 VSLAM 的实现方案在视觉里程计环节精度不够，或者过于依

DOI: 10.16182/j.issn1004731x.joss.19-VR0466

赖硬件性能。因此，以帧到地图的特征匹配为基础，面对特征地图数据体积大、计算资源消耗过多等问题，实现特征地图的自适应是必要的。张峻宁等[53]提出了一种自适应特征地图匹配的 VSLAM 方法，首先进行数据初始化，将当前帧转化成对应点云，把特征地图划分为多个子区域作为计算单位，利用角点的响应程度提取少量显著的特征点，然后进行各帧特征点匹配。接下来，为解决局部地图角点匹配消失问题，提出子区域特征点补充和局部地图扩建的方法，实现当前帧特征点的快速再匹配。最后，为了进一步提高 VO 环节相机位姿估计精度，增加了特征地图局部优化环节，提出帧到帧、帧到特征地图的局部地图优化模型，并通过加入 g2o[54]算法实现了相机位姿和地图特征点的同时优化。通过子区域分块、特征点补充与地图扩建的方式自适应维护特征地图规模，使得帧到特征地图的位姿估计兼顾了实时性和精度，另一方面提出的帧到帧、帧到模型的 g2o 特征地图更新方式，该方法在位姿估计的精度、累计误差的消除等方面表现显著[53]。鉴于深度学习在前端的优异表现，可以将深度 VSLAM 的结果与自适应优化结合，可以有效的降低 VO 相邻帧间的漂移误差，在保证实时性前提下，达到较好的定位精度和建图能力。

4 结论

本文以 VSLAM 和深度学习的结合为线索展开论述，描述了深度学习与视觉里程计、闭环检测和语义 SLAM 的结合现状，叙述了算法的可行性和高效性。接下来把传统 VSLAM 和深度 VSLAM 做了简要对比，如鲁棒性、训练周期、泛化能力等方面。最后根据二者的发展现状以及实际应用需求，从分布式的 VLSAM 建图、多传感器数据融合以及自适应 3 个方面做了展望。

自从深度学习在许多领域展现出强大的优势后，很多研究人员试图将深度学习端到端的理念整个应用到 VSLAM 中。但是，截止到目前，效果不够理想，大多深度学习的方法用来代替 VSLAM 部分环节，传统的基于几何的方法现在仍是主流。但是随着深度学习和多传感器的发展，VSLAM 会逐渐吸收深度学习带来的优势，提升其准确性和泛化能力。相信在不远的将来，VSLAM 的整个系统都会被深度学习取代，而不仅仅作为其中某个环节的实现方法，实现精确的基于深度学习的 VSLAM 方法。

参考文献：

[1] 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 (S1552-3098), 2016, 32(6): 1309-1332.

[2] Fuentes-Pacheco J, Ruiz-Ascencio J, Rendón-Mancha J M, et al. Visual simultaneous localization and mapping: A survey[J]. Artificial Intelligence Review (S0269-2821), 2015, 43(1): 55-81.

[3] 刘浩敏, 章国峰, 鲍虎军. 基于单目视觉的同时定位与地图构建方法综述[J]. 计算机辅助设计与图形学学报, 2016, 28(6): 855-868.
Liu Haomin, Zhang Guofeng, Bao Hujun. A survey of monocular simiultaneous localization and mapping[J]. Journal of Computer-Aided Design & Computer Graphics, 2016, 28(6): 855-868.

[4] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2012: 3354-3361.

[5] Kummerle R, Grisetti G, Strasdat H, et al. g2o: A general framework for graph optimization[C]// IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2011: 3607-3613.

[6] Belter D, Skrzypczyński P. Precise self-localization of a walking robot on rough terrain using ptam[M]. Baltimore, USA: Adaptive Mobile Robotics, 2012: 89-96.

[7] Mur-artal R, Tardos J D. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras[J]. IEEE Transactions on Robotics (S1552-3098), 2017, 23(5): 1255-1262.

[8] Engel J, Koltunk V, Cremers D. Direct sparse odometry [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence (S0162-8828), 2018, 40(3): 611-625.

[9] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.

LAS VEGAS: IEEE, 2016: 779-788.

[10] Ren S, He K, Girshick R B, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (S0162-8828), 2017, 39(6): 1137-1149.

[11] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3128-3137.

[12] Sünderhauf N, Pham T T, Latif Y, et al. Meaningful maps with object-oriented semantic mapping[C]// 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vancouver, Canada: IEEE, 2017: 5079-5085.

[13] Zhou Y, Li H, Kneip L. Canny-vo: Visual odometry with rgb-d cameras based on geometric 3-d–2-d edge alignment[J]. IEEE Transactions on Robotics (S1552-3098), 2018, 35(1): 184-199.

[14] Costante G, Mancini M, Valigi P, et al. Exploring representation learning with CNNs for frame-to-frame ego-motion estimation[J]. IEEE Robotics and Automation Letters (S2377-3766), 2016, 1(1): 18-25.

[15] Shahid M, Naseer T, Burgard W. DTLC: Deeply trained loop closure detections for lifelong visual SLAM[C]// Proceedings, Workshop on Visual Place Recognition, Conference on Robotics: Science and Systems (RSS). Ann Arbor, USA: RSS, 2016: 1-8.

[16] Hou Y, Zhang H, Zhou S L. Convolutional neural networkbased image representation for visual loop closure detection[C]// IEEE International Conference on Information and Automation. Piscataway, USA: IEEE, 2015: 2238-2245.

[17] Daniel D, Malisiewicz T, Rabinovich A. Toward geometric deep SLAM[EB/OL]. (2017-07-24) [2019-08-20], https://arxiv.org /pdf/1707.07410.pdf.

[18] Sharif Razavian A, Azizpour H, Sullivan J, et al. CNN features off-the-shelf: an astounding baseline for recognition[C]// Proceedings of the IEEE conference on computer vision and pattern recognition workshops. Columbus, Ohio: IEEE, 2014: 806-813.

[19] Wang S, Clark R, Wen H, et al. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks[C]// 2017 IEEE International Conference on Robotics and Automation (ICRA). Singapore: IEEE, 2017: 2043-2050.

[20] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015: 2625-2634.

[21] Elman J L. Finding structure in time[J]. Cognitive science (S0364-0213), 1990, 14(2): 179-211.

[22] Graves A. Supervised Sequence Labeling with Recurrent Neural Networks[M]. Heidelberg: Springer, 2012: 5-13.

[23] Chen Z, Jacobson A, Sünderhauf N, et al. Deep learning features at scale for visual place recognition[C]// 2017 IEEE International Conference on Robotics and Automation (ICRA). Singapore: IEEE, 2017: 3223-3230.

[24] Sünderhauf N, Dayoub F, Shirazi S, et al. On the Performance of ConvNet Features for Place Recognition[C]// International Conference on Intelligent Robots and Systems (IROS). Hamburg: IEEE, 2015: 4297-4304.

[25] Yi H, Hong Z, Zhou S. BoCNF: efficient image matching with Bag of ConvNet features for scalable and robust visual place recognition[J]. Autonomous Robots (S0929-5593), 2017, 42(9): 1-17.

[26] Lin K, Yang H F, Hsiao J H, et al. Deep learning of binary hash codes for fast image retrieval[C]// Proceedings of the IEEE conference on computer vision and pattern recognition workshops. Boston, USA: IEEE, 2015: 27-35.

[27] Sünderhauf N, Shirazi S, Jacobson A, et al. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free[C]// Proceedings of Robotics: Science and Systems XI. Michigan, USA: RSS, 2015: 1-10.

[28] Parisotto E, Singh Chaplot D, Zhang J, et al. Global pose estimation with an attention-based recurrent network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City: IEEE, 2018: 237-246.

[29] Hwang J, Park S, Kwak N. Athlete pose estimation by a global-local network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, Hawaii: IEEE, 2017: 58-65.

[30] Southall C, Stables R, Hockman J. Automatic Drum Transcription for Polyphonic Recordings Using Soft Attention Mechanisms and Convolutional Neural Networks[C]// The 18th International Society for Music Information Retrieval Conference. Suzhou: ISMIR, 2017: 606-612.

[31] Sünderhauf N, Pham T T, Latif Y, et al. Meaningful Maps with Object-Oriented Semantic Mapping[C]// 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). New York: IEEE, 2017: 5079-5085.

https://dc-china-simulation.researchcommons.org/journal/vol32/iss7/5
DOI: 10.16182/j.issn1004731x.joss.19-VR0466

[32] Ng P C, Henikoff S. SIFT: Predicting amino acid changes that affect protein function[J]. Nucleic Acids Research (S0305-1048), 2003, 31(13): 3812-3814.

[33] Lei H, Akhtar N, Mian A. Octree guided CNN with Spherical Kernels for 3D Point Clouds[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA: IEEE, 2019: 9631-9640.

[34] Mani I, Zhang I. KNN approach to unbalanced data distributions: a case study involving information extraction[C]// Proceedings of workshop on learning from imbalanced datasets. Washington: ICML, 2003: 126.

[35] Radwan N, Valada A, Burgard W. VLocNet++: Deep Multitask Learning for Semantic Visual Localization and Odometry[J]. IEEE Robotics and Automation Letters (S2377-3766), 2018, 3(4): 4407-4414.

[36] Girisha S, Manohara P, Ujjwal V, et al. Semantic Segmentation of UAV Aerial Videos using Convolutional Neural Networks[C]// 2019 IEEE Knowledge Engineering (AIKE). Sardinia, Italy: IEEE, 2019: 21-27.

[37] Han Y, Ye J C. Framing U-Net via deep convolutional framelets: Application to sparse-view CT[J]. IEEE Transactions on Medical Imaging (S0278-0062), 2018, 37(6): 1418-1429.

[38] Bowman S L, Atanasov N, Daniilidis K, et al. Probabilistic data association for semantic slam[C]// 2017 IEEE International Conference on Robotics and Automation (ICRA). Singapore: IEEE, 2017: 1722-1729.

[39] Jordan M I, Jacobs R A. Hierarchical Mixtures of Experts and the EM Algorithm[J]. Neural Computation (S0899-7667), 1994, 6(2): 181-214.

[40] Engel J, Koltun V, Cremers D. Direct sparse odometry[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (S0162-8828), 2017, 40(3): 611-625.

[41] Geiger A, Ziegler J, Stiller C. Stereoscan: Dense 3d reconstruction in real-time[C]// 2011 IEEE Intelligent Vehicles Symposium (IV). Baden-Baden, Germany: IEEE, 2011: 963-968.

[42] Loo S Y, Amiri A J, Mashohor S, et al. CNN-SVO: Improving the mapping in semi-direct visual odometry using single-image depth prediction[C]// 2019 International Conference on Robotics and Automation (ICRA). Montreal, Canada: IEEE, 2019: 5218-5223.

[43] Zhan H, Weerasekera C S, Bian J, et al. Visual Odometry Revisited: What Should Be Learnt?[EB/OL]. (2019/09/21) [2019/10/05], https://arxiv.org/abs/1909.09803.pdf.

[44] Zhou T, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii: IEEE , 2017: 1851-1858.

[45] Cieslewski T, Choudhary S, Scaramuzza D. Data-efficient decentralized visual SLAM[C]// 2018 IEEE International Conference on Robotics and Automation (ICRA). Prague, Czech Republic,: IEEE, 2018: 2466-2473.

[46] Li S, Zhi Y, Anestis Z, et al. Recurrent-OctoMap: Learning State-based Map Refinement for Long-Term Semantic Mapping with 3D-Lidar Data[J]. IEEE Robotics and Automation Letters (S2377-3766), 2018, 3(4): 3749-3756.

[47] Hornung A, Kai M W, Bennewitz M, et al. OctoMap: An efficient probabilistic 3D mapping framework based on octrees[J]. Autonomous Robots (S0929-5593), 2013, 34(3): 189-206.

[48] Zhang J, Singh S. Laser–visual–inertial odometry and mapping with high robustness and low drift[J]. Journal of Field Robotics (S1556-4967), 2018, 35(8): 1242-1264.

[49] Garcia-Fidalgo E, Ortiz A. Vision-based topological mapping and localization methods: a survey[J]. Robotics and Autonomous Systems (S0921-8890), 2015, 64: 1-20.

[50] Engel J, Schöps T, Cremers D. LSD-SLAM: Large-Scale Direct Monocular SLAM[M]. Munich: Computer Vision – ECCV 2014. 2014: 834-849.

[51] Scherer S A, Zell A. Efficient onbard RGBD-SLAM for autonomous MAVs[C]// 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. Tokyo, Japan: IEEE, 2013: 1062-1068.

[52] Vijayanarasimhan S, Ricco S, Schmid C, et al. Sfm-net: Learning of structure and motion from video[EB/OL]. (2017/04/25) [2019/08/25], https://arxiv.org/abs/1704.07804.pdf.

[53] 张峻宁, 苏群星, 刘鹏远, 等. 一种自适应特征地图匹配的改进 VSLAM 算法[J]. 自动化学报, 2019, 45(3): 553-565.
Zhang Junning, Su Qunxing, Liu Pengyuan, et al. An Improved VSLAM Algorithm Based on Adaptive Feature Map[J]. Acta Automatica Sinica, 2019, 45(3): 553-565.

[54] Grisetti G, Kümmerle R, Strasdat H, et al. g2o: A general framework for (hyper) graph optimization[C]// 2011 IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China: IEEE, 2011: 3607-3613.