DOI: 10.3969/j.issn.1674-1579.2019.02.001

深度学习在视觉 SLAM 中应用综述

李少朋[1,2]，张　涛[1]

摘　要：视觉 SLAM 一直是近年来火热的研究方向，其处理对象为视觉图像；深度学习在图像处理中展现出的愈加突出的优势，使二者的广泛结合成为了可能. 总结了传统 SLAM 与基于深度学习的 SLAM 的特点、性质，重点介绍和总结了深度学习在视觉里程计、回环检测中的研究成果，展望了基于深度学习的视觉 SLAM 的研究发展方向.

关键词：深度学习；同时定位与建图；视觉里程计；回环检测

中图分类号：TP751　　**文献标志码**：A　　**文章编号**：1674-1579(2019)02-0001-10

A Survey of Deep Learning Application in Visual SLAM

LI Shaopeng[1,2], ZHANG Tao[1]

Abstract: Visual SLAM has been the hot research topic in recent years, which treats visual image as processing objects. Deep learning shows the prominent advantages in image processing, which makes it possible to combine the visual SLAM and deep learning. The characteristics and properties of traditional SLAM and SLAM based on deep learning are summarized. The prominent achievements on visual odometry and loop closure detection incorporated with deep learning are introduced. The future research directions of advanced SLAM based on deep learning are discussed.

Keywords: deep learning; visual simultaneous localization and mapping (SLAM); visual odometry; loop closure detection

0　引　言

同时定位与地图构建(SLAM)是智能体携带其传感器在运动过程中对自身进行定位，同时以合适的方式描述周围的环境[1]. SLAM 能够比传统的文字、图像和视频等方式更高效、直观地呈现信息；在 GPS 不能正常使用的环境中，SLAM 也可以作为一种有效的替代方案实现在未知环境中的实时导航. SLAM 技术在服务机器人、无人驾驶汽车、增强现实等诸多领域发挥着越来越重要的作用.

如图 1 所示，一个完整的 SLAM 框架由以下 4 个方面组成：前端跟踪、后端优化、回环检测、地图重建. 跟踪前端即视觉里程计负责初步估计相机帧间位姿状态及地图点的位置；后端优化负责接收视觉里程计前端测量的位姿信息并计算最大后验概率估计；回环检测负责判断机器人是否回到了原来的位置，并进行回环闭合修正估计误差；地图重建负责根据相机位姿和图像，构建与任务要求相适应的地图.

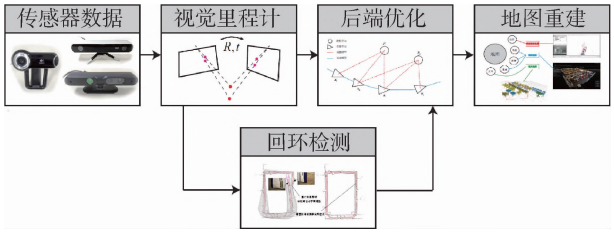

图 1　SLAM 流程示意图

Fig. 1　Workflow of SLAM

近年来，视觉 SLAM 技术得到了广泛的研究和长足的发展. 目前，较为先进的视觉 SLAM 方案在公开数据集及实际物理实验中均取得了较高的精度. 传统的视觉 SLAM 方案分为特征点法和直接法两类. 特征点法从每帧图片中提取稳定的特征点，通过

收稿日期：2019-02-23；录用日期：2019-03-20.
Manuscript received Feb. 23, 2019; accepted Mar. 20, 2018.

1. 清华大学自动化系，北京 100084；2. 火箭军工程大学，西安 710025.
1. Department of Automation, Tsinghua University, Beijing 100084, China;
2. Rocket Force University of Engineering, Xi'an 710025, China.

这些特征点具有不变性的描述子完成相邻帧的匹配;然后通过对极几何较为鲁棒地恢复相机的姿态和地图点坐标,最后通过最小化投影误差完成相机位姿和地图结构的微调,每帧所提取的特征点通过聚类等操作进行回环检测或重定位. 但是特征点的提取及匹配是较为耗时的工作,使得经典的特征点法比直接法运行速度要慢. PTAM[2] 是早期较为典型的特征点法,该方法基于非线性优化后端采用了基于关键帧的 Bundle Adjustment (BA)[3] 来求解位姿与地图结构,之后的很多的特征点法都是该方法的改进版本,其中最为成功的一个就是 ORB-SLAM[4],这也是目前效果最好的特征点法. 直接法不再提取特征点,直接通过光度误差来恢复相机的姿态和地图结构,不用计算关键点和描述子. 由于直接法中未提取特征点,没有能够表征一帧图像的全局特征,直接法的回环检测还是一个开放的话题,所以长时间导航的漂移问题是直接法所面临的主要问题. LSD-SLAM[5] 是典型的直接法 SLAM,该方法能够在无 GPU 加速的情况下实时运行. DSO-SLAM[6] 以 LSD-SLAM 为基础,通过光度修正在一定程度上弥补了未提取稳定特征点的不足,滑动窗口优化及边缘化策略充分利用的各帧图像的信息. DSO-SLAM 无论在估计精度还是运行效率上都有非常优异的表现.

但自 2017 年以来,传统的视觉 SLAM 方案再没有取得实质性的进展,传统视觉 SLAM 方法有以下几个方面的问题还没有较为完备的解决方案:

(1) 在光照条件恶劣或光照变化较大等不利条件下,算法的鲁棒性还不是很高;

(2) 在相机运动较大的情况,传统算法容易出现“跟丢”的情况;

(3) 传统算法不能识别前景物体,即对场景中运动的物体只能当作“坏点”来处理,没有较好的解决方案.

随着深度学习在计算机视觉领域的发展,越来越多的视觉问题都通过深度学习的方式取得了更高的突破. 目前深度学习在图像分类、识别、物体检测、图像分割等几大领域的表现都远远高于传统人工设计的算法. 视觉 SLAM 同样以图像为处理对象,这为神经网络的学习能力在该领域的应用提供了很大的可能. 深度学习与 SLAM 的结合在改善了视觉里程计和场景识别等由于手工设计特征而带来的应用局限性,潜在提高了机器人的学习能力和智能化水平. 采用深度学习方式处理 SLAM 问题,有以下几个研究层面的优势:

(1) 基于深度学习的 SLAM 方案对光照有较好的不变性,能够在光照条件较为恶劣的条件下工作;

(2) 基于深度学习的 SLAM 方案能够识别并提取环境中移动的物体,可以进行动态环境下的 SLAM 建模;

(3) 通过深度学习的方式可以提取高层语义信息,为语义 SLAM 的构建以及场景语义信息的理解及使用提供了更大的帮助;

(4) 采用深度学习的方式更有利于信息及知识的压缩保存,更有益于机器人知识库的构建;

(5) 基于深度学习的 SLAM 方案更符合人类认知及环境交互的规律,有更大的研究及发展的潜力.

基于前述分析,本文对基于深度学习的 SLAM 方案做了广泛调研. 重点在基于深度学习的视觉里程计、回环检测方法两个方面做了综述. 并指出了未来基于深度学习的视觉 SLAM 方案的研究趋势与发展方向.

1 基于深度学习的视觉里程计

视觉里程计(VO)是通过分析关联图像之间的多视几何关系确定机器人位置与姿态过程. 相较于传统的视觉里程计方法,基于深度学习的方法无需特征提取,也无需特征匹配和复杂几何运算,使得整个计算过程更加直观简洁. 根据训练方法和数据集标签化程度的不同,将基于深度学习的视觉里程计方法分为监督学习,无监督学习,半监督学习三类分别进行讨论.

1.1 监督学习方法

监督学习方法的基本思路为通过图片帧的输入来映射出该帧的位置和姿态,自卷积神经网络(CNN)得到大规模应用以来,基于监督学习的视觉里程计设计方法就得到了学者的广泛关注和研究,最早的研究通过分类网络末端用 Softmax 层来输出各帧的速度大小和方向,虽然效果不理想,但是说明了深度学习在该领域应用的可行性[7].

PoseNet[8] 为早期监督学习方法的典型代表,其通过 SFM 对所采集图像进行批处理,计算出其对应的位姿作为数据集的标签. 然后建立由图片到六自由度位姿的回归模型,模型的神经网络结构借鉴了 GoogleLeNet[9] 的网络结构及参数,并在此基础上做

了相应的修改和在训练. PoseNet 通过迁移学习,在无大量标签数据集的支持下,得到了精度较高的位姿定位. 然而这种以图片帧为输入,以绝对位姿为输出模式在泛化能力上有一定的不足. 文献[10]采用 Siamese 网络分别从相邻图像提取特征估计了图像间的相对位姿. 文献[11]在原有研究成果的基础上,以图像帧之间的稠密光流作为输入,以图像帧之间的相对位姿进行训练. 首先将对稠密光流图像进行降采样学习其较“粗略”的全局特征,同时将原图像进行分割,学习其较“精细”的局部特征,之后通过局部特征与全局特征相结合的方式来表征整个图像,之后以图像特征为输入,相对位姿为输出训练整个网络. 网络的训练是分步进行的,该方法与 PoseNet 相比无论在精度还是在泛化能力上都有了一定的提高. 这种局部特征与全局特征共同学习的方式为该领域的研究提供了很好的思路.

在基于监督学习的视觉里程计方法中,目前效果最好且应用较为广泛的为 DeepVO[12],DeepVO 能够从序列原始图像直接映射出其对应的位姿,它不仅能够通过卷积神经网络(CNN)学习图像的特征,而且能够通过深度递归神经网络学习(RNN)隐式地学习图像间的动力学关系及内在联系. 在特征提取方面,相邻两帧图像在通道上进行组合,堆叠成六通道的图像(每个图像有 RGB 三通道),然后通过多层 CNN 网络对图像进行提取,将提取的特征输入到 RNN 网络中,Long short-term memory(LSTM)在一定的滑动窗口同时训练连续的图像帧最后输出图像位姿. 这种 CNN + RNN 的结构充分利用了当前帧图像和之前某一区域帧的信息,符合人类认知及信息处理的流程,整个过程不涉及任何几何计算,可以行端到端的学习,由于模型学习的是各帧之间的位姿关系,该模型也有较好的泛化能力,可以在陌生的环境下使用. DeepVO 与经过精细设计和优化的传统方法相比在精度上没有绝对的优势,但因其巨大的研究价值得到了广泛的关注,VINet[13]采用类似的结构将通过 CNN 提取的图像间的特征与 IMU 数据同时输入到 RNN 网络进行训练并输出位姿. Deep EndoVO[14]将该结构应用在了内镜胶囊机器人上进行定位,取得了较好的效果并证明该方法的实用价值.

1.2　无监督学习方法

无监督学习在视觉里程计中应用较早,起初无监督学习在该领域的应用是提取稳定的特征点,通过特征点的匹配来求解相对位姿. 近年来,随着深度学习技术的发展,研究者逐步把侧重点放在了直接的位姿估计上. 文献[15]较早地采用自编码的方式同时估计图像的深度及图像间的运动. 文献[16]通过无监督学习的方式进行单一图像的深度估计,该方法采用双目数据集,通过多重目标损失训练网络产生视差图. 如图 2 所示,以训练双目数据集左侧图像为例,左侧图像通过网络训练分别产生左侧视差图和右侧视差图,左侧视差图与右侧图像通过几何计算产生右侧左侧图像,右侧视差图和左侧图像通过几何计算产生右侧图像,则训练误差由 3 个部分组成:(1) 重建误差,即重建出的图像与真实图像的差;(2) 视差图平滑误差;(3) 左右视差图一致性误差. 通过模型的训练,该网络对单个图像深度估计达到了非常高精度,超过了最先进的监督学习方法. 这种估计深度的无监督学习方法为基于无监督学习的视觉里程计设计提供了很好的思路.

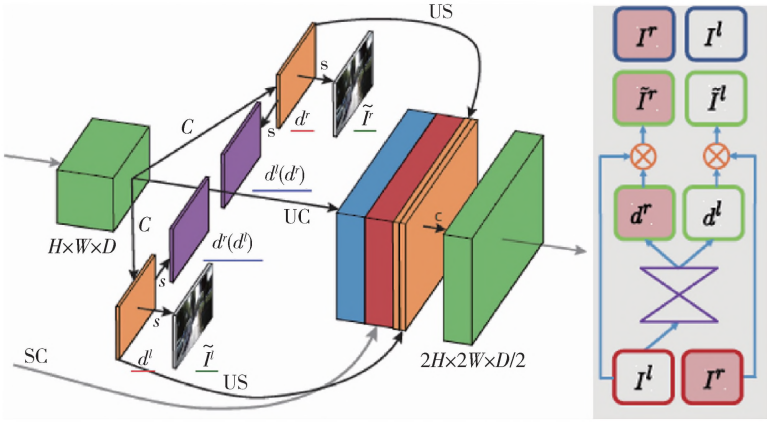

图 2　网络训练示意图(左侧为训练损失、右侧为数据传输流程)[27]

Fig. 2　Network training diagram of Literature[27]

参照文献[17]中的思路与方法,文献[18]通过无监督学习的方式同时估计出了图像的深度、图像间的位姿状态以及图像中的动态物体,该方法以单帧图像通过 Depth CNN 网络生成深度图,相邻两帧图像通过 Pose CNN 生成图像间的位姿,根据深度图与位姿将原图像投射到目标图像上,最后通过真实目标图像与投射产生目标图像的重建误差来训练网络. 在重建误差计算的过程中,会出现一些"坏点",这些"坏点"会被给予一定的权值来减小"坏点"对整个系统的影响. 然而这些被给予权值的"坏点"就是图像中动态物体所代表的点. 该方法在网络结构设计、初值设定和训练方法上都采用了较为合适策略,是目前效果最好的无监督学习方法. 之后也有基于该方法的改进算法,其中 UnDeepVO 提出了一种基于双目数据集的无监督学习方法[19],其在对双目相机左右侧图像进行位姿和深度学习的同时,又针对双目相机某一侧的图像进行帧间位姿的学习,与单目的学习方案相比,该方法能够真实地恢复位姿的尺度.

GeoNet 在文献[20]的基础上做了较大的改进[21-22],其在计算图像重构误差的过程中分两步进行,首先假设图像之间是刚性变换,不存在动态的前景物体,通过生成图像与真实图像间结构误差[23]及像素误差的加权和来进行光流的预测. 然后以此为初值,通过 FlowNet2.0[24]进行场景中运动物体的光流计算. 除此之外 GeoNet 在进行图像投影计算时,将多个图像在通道上融合,通过求解多帧之间的位姿,提升了位姿估计的精度.

1.3　半监督学习方法

半监督学习方法除了应用图像进行自我监督外,还应用了部分标签作为监督信息. SFM-Net[19]采用了与文献[18]类似思路,不过 SFM-Net 更加侧重于对场景的描述,注重场景中前景物体的运动状态,其能够根据输入的序列图像,输出单个图像的深度、图像间的位姿及运动光流、图像中运动的前景物体以及其相对于背景的运动状态. 该方法在模型构建及训练上做了很大的改进. SFM-Net 首先训练的单个图像通过卷积和反卷积操作生成对应深度图像,然后融合成深度点云;以相邻帧的图像对为输入经过神经网络计算输出图像间的位姿关系,识别出场景中的运动物体并进行图像分割,输出前景物体掩码(Mask)及其相对于背景的运动状态,当前帧 I_t 的点云根据图像间的位姿以及前景物体的运动状态进行点云位姿变换并投射到下一帧 I_{t+1} 生成图像间的运动光流. 整个网络根据数据集提供标签的不同可以进行无监督和半监督学习,其中无监督学习的损失函数包括:(1) 光度误差,即根据图像 I_t 以及 I_t 与 I_{t+1} 之间的光流场,生成图像 $\tilde{I}_{t+1}$,$\tilde{I}_{t+1}$ 与 I_{t+1} 之间的光度差;(2) 深度图、光流场以及推断运动图的平滑误差;(3) 前向和反向投影中深度图的一致性误差. 除此之外,在数据集部分标签可用时可以进行监督学习,其损失函数包括:(1) 在深度图像可用时可进行生成深度图像的监督;(2) 在图像间相对位姿可用时可进行位姿监督;(3) 在当前一些合成数据集中包含一些真实的光流场及前景物体信息,可用以监督学习. 该方法充分利用了数据集的信息,可以同时进行不同程度的学习,不过该方法的侧重点并不是计算位姿间的关系,其中前景物体运动状态的估计相对于位姿估计浪费了一定的计算量.

在单目相机深度预测领域,文献[21]采用一种半监督学习方法,在无深度图像作为监督数据时,采用了与文献[16]类似的方法,以图像的重构误差以及估计深度图像的平滑误差作为损失函数,在深度图像存在时,可以再加上真实深度图像与估计深度图像的误差作为损失函数. 这种半监督方式加速了模型的收敛,并且解决了无监督学习的不适定问题.

1.4　方法总结及应用分析

监督学习和无(半)监督学习在该领域均取得了一定的成果,从现有的成果来看,无(半)监督学习在数据集采集、网路训练的可操作性以及最终的估计精度和泛化能力方面均有一定的优势. 重要的是,无(半)监督学习以位姿为输入量,通过位姿变换后的图像与实际图像的吻合程度为监督信息进行学习,这更符合我们人类认知的规律,具有较大的发展潜力. 由于半监督学习通过部分人工标定或测量的信息,能够较好的恢复场景的尺度,与无监督相比有较大优势. 然而在特定任务的限制环境中,可以采用监督学习的方式,采集特定环境下的数据集进行训练,与无(半)监督方法相比会得到更好的效果.

2　深度学习与回环检测

回环检测是判断机器人回到了原来的位置并将累计误差合理的分配到回环的轨迹上,图像之间的描述和匹配是回环检测的关键技术. 在传统方法中,研究者们通常涉及人工的特征(hand-crafted fea-

tures)来描述一幅图像. 人工特征分为局部特征和全局特征,局部特征包括 ORB,SIFT,SURF 等,词袋法(bag of-visual-words)[25]通过局部特征的统计数据来描述整个图像. 全局特征包括 GIST[26]、Vector of locally aggregated descriptors (VLAD)[27]、Fisher vector[28]等分别以不同的计算方式描述整幅图像的特征.

随着深度学习技术的发展,越来越多的研究倾向于采用深度神经网络特征来描述一幅图像进行回环检测;近两年也有不少研究者通过对三维点云学习方式提取其特征,为基于三维点云的回环检测提供了可能.

2.1　二维图像深度回环检测

随着 Places 数据集[29]的提出与应用,深度神经网络具备了强大的场景描述和识别能力,而这正是回环检测所需要的. 国防科技大学[30]较早地将深度学习应用在回环检测中,其将 AlexNet[17]迁移到回环检测问题中,用其中间层的输出作为特征来描述整幅图像,通过二范数进行特征匹配来确定是否存在回环. 之后研究者还通过 LSH 数据压缩、图像帧的管理[31]或主成分分析(principal component analysis, PCA)[32-33]来增强匹配的效率. 清华大学高翔等[34]提出通过无监督学习的方式,采用堆叠去噪自动编码器(stacked denoising auto-encoder,SDA)的方式描述整幅图像来进行图像的匹配实现回环检测,并取得了较好的效果. 仿真结果表明,初期的基于深度学习的回环检测方法与传统方法相比具有较强的鲁棒性,但是这些方法并未有针对性设计网络结构,也未进一步地进行网络的训练,在回环检测的精度及效率上没有明显的提高.

之后有不少研究者针对回环检测问题在网络训练和数据处理方面都进行了相应的改进性设计. 文献[37]没有直接对整幅图像进行特征提取,而是根据预训练的结构提取图像中的路标区域,之后通过 ConvNet 计算每块区域的特征并将特征进行压缩. 通过特征匹配完成路标区域的匹配,通过各个路标区域的相似性来计算整幅图像之间的相似性. 除此之外,该方法还将路标区域框的大小作为监督条件以减小出现假阳性的概率. 方法设计显着地提高了对场景中视点变化或局部遮挡的鲁棒性. 文献[38]用 Places 数据集训练了网络,通过局部敏感 Hashing 变换在精度损失较小的情况下将匹配速度加快了两个数量级;并将特征输出编码成语义信息来划分搜索空间大大减小场景了搜索的时间. 这种根据两层特征划分搜搜空间实现“由粗到细”的场景匹配的方法,可以很好的应用在图像库的搜索过程中.

同时也有研究者采用了网络学习特征与人工设计特征相结合的方式进行场景识别[39-40],其中 NetVLAD[40]是目前在场景识别领域识别效果最好的网络,其通过 CNN 特征与局部聚合描述向量(vector of locally aggregated descriptors, VLAD)相结合的方式构建神经网络,并通基于 Google Street View Time Machine 数据集[35-36]对网络进行弱监督学习. 图像输入后经过一系列卷积操作生成 $W \times H \times D$ 维的特征,这些特征可看作 $N(N = W \times H)$ 个 D 维的局部特征,这些局部特征通过 NetVLAD 层进行类似于 VLAD 操作,产生整个图像的特征向量. NetVLAD 层对传统 VLAD 算法做了相应的改进,传统 VLAD 算法如式(1)所示,其中权重 $a_k(\boldsymbol{x}_i)$ 是与局部向量和聚合簇(cluster) $\boldsymbol{c}_k$ 相关的量.

$$V(j,k) = \sum_{i=1}^{N} a_k(\boldsymbol{x}_i)(x_i(j) - c_k(j)),$$
$$a_k(\boldsymbol{x}_i) = \frac{e^{-\alpha \| x_i - c_k \|^2}}{\sum_{k'} e^{-\alpha \| x_i - c_{k'} \|^2}} \tag{1}$$

在 NetVLAD 层中将权重的相关参数 $\boldsymbol{w}_k$、$\boldsymbol{b}_k$ 和聚合簇向量 $\boldsymbol{c}_k$ 均设为需要学习的参数,如式(2)所示,这种设计保证了计算的平滑性,能够顺利地计算梯度,保证了模型能够进行端到端的学习.

$$V(j,k) = \sum_{i=1}^{N} a_k(\boldsymbol{x}_i)(x_i(j) - c_k(j)),$$
$$a_k(\boldsymbol{x}_i) = \frac{e^{\boldsymbol{w}_k^{\mathrm{T}} x_i + b_k}}{\sum_{k'} e^{\boldsymbol{w}_k^{\mathrm{T}} x_i + b_{k'}}} \tag{2}$$

传统 VLAD 是无监督的分类学习,由于 NetVLAD 中存在监督数据(即两幅图像是否来自同一位置是已知的),这样在学习聚合簇向量 $\boldsymbol{c}_k$ 位置时,比传统 VLAD 方法更有优势. 最后经过正则化操作 NetVLAD 层输出了 $D \times K$ 维向量,用以描述图像的全局特征. 模型的训练采用的谷歌街景数据集,该数据集用 GPS 标注了图像对应的位置,由于在同一位置由于视角或方向的不同,场景也是不同的,但可以肯定的是不在同一地方(GPS 位置相差较远)描述的肯定不是同一场景,在同一地方描述可能是同一场景. 所以,模型采用了弱监督学习的方式:

$$L_\theta = \sum_j l(\min_i d_\theta^2(q, p_i^q) + m - d_\theta^2(q, n_i^q)) \tag{3}$$

式中 $l(x) = \max(x,0)$,p_i^q 为与图像 q 的 GPS 位置相近的图像,n_i^q 为较远的图像. m 是可调节参

数,总体意思是让同一位置场景匹配度最高的图像的分值要低于 GPS 相距较远的图像. NetVLAD 巧妙地结合了 CNN 网络提取和人工特征提取算法,在弱数据集上实现了端到端的训练.

2.2 三维点云深度回环检测

由于三维点云的无序性、坐标系统难统一等问题一直没有得到很好的解决,对于基于三维点云的回环检测问题仅停留于理论研究的阶段. 其中文献[41]以三维体素(voxel)网格模型为对象用于判断点云是否来自同一位置为目标,首先提取体素网格模型中的人工统计特征,然后将人工特征编码为向量输入到神经网络中以数据集中点云是否匹配为监督信息进行学习训练. 该方法是该领域为数不多的方法中的一个,其在点云匹配的精度及效率上都有待进一步的提高. 但其基于深度学习的三维点云位姿配准及分类方法对该领域提供了很多可供借鉴的方法.

在基于三维点云的位姿配准方面,传统的基于点云人工特征的配准方法已取得不俗的效果[42-44],在深度学习领域,文献[45]通过对三维点云特征的学习实现了位姿的配准,该方法首先通过随机球面覆盖集(random sphere cover set, RSCS)算法将点云划分为独立的小块点云,划分后的局部点云命名为超点(super-point);然后为超点选择归一化的局部坐标系,将超点数据投影到当前 2D 深度图像中并做显著性检验和过滤;之后通过深度学习自编码的方式将提取超点特征,通过超点的特征匹配完成当前帧的初定位,最后通过 ICP 算法对位姿进行微调得到相机的位姿. 该方法对三维点云回环检测在点云划分管理以及检测回环后的位姿矫正方面有很大的借鉴意义.

在点云分类及特征提取方面,3Dmatch[46] 首先将点云转换为三维体素网格模型,体素网格模型较为有序的数据为输入通过 Siamese 网络判断点云是否匹配,根据监督数据集完成训练. 由于体素网格模型中网格的有序性,大部分对于三维点云模型的处理也都是基于此模型,然而向三维体素网格模型的转换会造成一定程度的失真,而且付出较大的计算代价. 2017 年 PointNet[47] 有针对性地设计了网络直接用欧式空间中的点云作为输入进行学习,欧式空间中的点云具有无序性、单位点之间相互联系、对旋转有不变性等特点, PointNet 进行了有针对性的设计,以点云 n 个点云为输入(每个点包含三通道的位置信息),首先通过空间变换网络进行训练出一个空间变换矩阵(T-Net)(通过正则化项使其保持正交性),在空间位置上对点云进行规范化,然后每个点通过多层感知机(multi-layer perceptron, MLP)进行特征学习是每个点赋予了 64 维的特征,之后再通过特征空间变换(64 维)和多层感知机学习是每个点具有 1024 维的特征,然后通过最大池(Max-Pooling)完成特征压缩,之后再通过多层感知机完成类别的输出;在语义分割方面,由于语义信息与局部信息和全局信息都有关,所以通过局部特征和全局特征的融合完成语义识别. 该方法利用 MaxPooling 函数与参数顺序无关的特点成功的处理了点云的无序性,并且通过空间变换网络(T-Net)规范化了空间点云的空间一致性问题. 最后结果表明 PointNet 能够有效的识别出点云中信息丰富的物体的骨架(skeleton)信息,达到了很好的效果. 在 PointNet 作者之后的研究中,PointNet + +[48] 采用 PointNet 的基本思路进行点云的局部特征提取,首先通过最远点法分割点云(点云区域可重复),然后对局部点云进行特征提取生成规模更小的点云,以此类推直至点云规模小于一定值时通过全连接层(FCN)输出点云的类别;当以点云分割为目标是则需要恢复点云规模给出语义信息. PointNet 的出现一定程度上改变了针对三维点云研究的格局,使得模型的研究不再依赖于体素网格模型,加快了三维点云特征提取、识别等技术的发展.

3 深度学习与三维重建

传统的三维重建方法中,比较典型的方法为 KinectFusion[49] 和 ElasticFusion[50],深度学习对于三维重建的贡献主要集中正在单目图像的深度点云估计上[51-52]. 文献[53]将深度学习应用到了 Structure from Motion 领域,进行了初步探索. CodeSLAM[54] 为首个基于深度学习的实时三维重建方案,也是目前仅有的基于深度学习的完整方案. 其以单目光度图像的深度估计为基础,提出了一种紧凑、密集的几何场景表示方案. CodeSLAM 根据每帧图像生成可进行参数优化的点云表示,结合每帧图像对应的位姿对场景统一优化以实现全局的一致性.

深度学习在实时三维重建中的应用处于初步探索阶段,随着 CodeSLAM 的提出与实现,深度学习在该领域应用的可行性得到了验证. 深度三维重建将

会凭借其巨大的研究价值引来更多的关注,深度三维重建方案也会得到进一步的提升和改进.

4　未来展望

4.1　高层级地图构建

在人类的认知中,我们看到场景中的事物时,除了知道其位置信息外(3 通道),还知道其颜色信息(3 通道),除此之外还知道其语义信息以及是否可触碰、柔软坚硬等一系列的信息. 然而我们深度 SLAM 方法仅仅构建的三维点云信息,这是不够的,所以需要在更高的维度上构建更为丰富的更高阶的地图从而适应多样化的任务,也反过来帮助机器人的自我导航. SLAM 创始人 Andrew J. Davison 在他近期的综述 Future Mapping[55] 中也有了类似的设想.

4.2　类似人类的感知与定位

由于深度 SLAM 采用了智能的方式,其之后的发展方向会越来越接近人类的感知和思考模式,其中文献[56]做了类似的探索,其构建了一个完全端到端的模型,该模型以序列图像为输入,首先根据 Local Pose Estimation Network 求图像间的相对位姿,之后通过 Pose Aggregation 对相对位姿信息进行压缩,然后将处理后的相对位姿信息传入 Neural Graph Optimization 网络,该网络根据输入的相对位姿信息输出全局的绝对位姿信息,并通过大脑的 Soft Attention 模型提取路径关键信息并通过信息搜索生成各帧之间的相似性矩阵,通过相似性矩阵完成 SLAM 中的回环检测功能,最后输出了整个行走的路径,该路径与真实路径的差异作为损失函数对网络的进行训练. 最后该网络在游戏的模拟环境中达到了较好的效果,并且验证了 Soft Attention 模型执行回环检测对全局位姿估计的作用. 根据现有的深度学习技术发展程度,该方法并没有在真实环境中,达到较好的效果,但是这种端对端的训练模式,以及整个网络的信息处理的过程,符合我们人类认知的流程,具有很大的发展潜力.

4.3　主动 SLAM 方法

人类到了陌生的环境,会主动的去环顾四周来更好地完成自我的定位和环境的感知. 当我们迷路是会主动地去寻找自己记得的标志物或者退回到原来的地方从而确定自己的位置,未来智能机器人也应该有类似的能力. 其中,文献[57]做了初步的探索,其通过深度学习的方式进行了 Active SLAM 初步的尝试,该模型在进行学习的过程中除了输出相机当前位姿外还输出相机运动的策略,该策略用以辅助相机的下一步更好的定位,该网络模型在模拟的环境中进行了验证,并取得了一定的效果.

4.4　与任务要求相融合

定位与感知不是最终目的,最终目的是通过精确的定位及感知完成多样化的任务. 这对深度 SLAM 提出了更高的要求,在对深度 SLAM 网络进行学习训练时,需要以任务的完成情况为指标进行训练. Google 的 DeepMind 做了类似的尝试[58],其利用谷歌街景数据集,在采集到图像输出行动策略以智能机器人是否能够到达预定位置为目标进行学习训练,从而完成机器人在无地图等先验信息下的导航.

4.5　记忆的存储与提取

在回环检测方面,无论二维图像或是三维点云随着场景规模的增加其数据量也会越来越大,一直保存大量的图片或点云显然是不行的. 涉及到知识的压缩与提取,人类不会记住自己看到过的每一帧图象或点云,但是也能够在自己到了之前到过的地方后完成识别或回环检测,因为人类有更高级的知识、有城市、景点、区域等概念可以帮助我们完成区域划分,我们能够记住到过哪个城市、哪条街就足够了,未来深度 SLAM 也需要将感知信息压缩,划分搜索空间完成回环检测. 另外,人类随着记忆量的增加也会忘记具体信息甚至到了自己曾经到过的地方也不会察觉,在固定存储空间条件下需要进行非关键信息的剔除. 另外,类似人类用谷歌地图等,长期定位可以借助云存储的方式完成大量信息的储存.

5　结束语

除以上深度学习在视觉里程计、回环检测方面的应用外,其在语义 SLAM[59]、图像局部特征提取及匹配[60]、配准尺度学习[61]等方面均取得可观的实验结果. 目前,传统的 SLAM 方案研究较为成熟,基于深度学习的 SLAM 技术处于一个刚刚起步、渐有起色的阶段. 随着人工智能技术的发展,未来视觉 SLAM 中的各个关键技术将部分或全部被深度学习所取代. 基于深度学习的视觉 SLAM 方法具有巨大的研究空间,也将会在工程中发挥越来越大的作用.

参　考　文　献

[1] DAVISON A J, REID I D, MOLTON N D, et al. MonoSLAM: Real-time single camera SLAM[J]. IEEE

Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6): 1052-1067.

[2] KLEIN G, MURRAY D. Parallel tracking and mapping for small AR workspaces[C]// The 6th IEEE and ACM International Symposium. New York:IEEE, 2007: 225-234.

[3] STRASDAT H, MONTIEL J M M, DAVISON A J. Real-time monocular SLAM: Why filter? [C] // 2010 IEEE International Conference. New York: IEEE, 2010: 2657-2664.

[4] MUR-ARTAL R, TARDÓS J D. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras[J]. IEEE Transactions on Robotics, 2017, 33(5): 1255-1262.

[5] ENGEL J, SCHÖPS T, CREMERS D. LSD-SLAM: Large-scale direct monocular SLAM [C] // European Conference on Computer Vision. Springer:Cham, 2014: 834-849.

[6] ENGEL J, KOLTUN V, CREMERS D. Direct sparse odometry[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(3): 611-625.

[7] KONDA K R, MEMISEVIC R. Learning visual odometry with a convolutional network[J]. VISAPP, 2015(1): 486-490.

[8] KENDALL A, GRIMES M, CIPOLLA R. Posenet: A convolutional network for real-time 6-dof camera relocalization[C]// Computer Vision (ICCV), 2015 IEEE International Conference. New York:IEEE, 2015: 2938-2946.

[9] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]// Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition. New York:IEEE, 2015: 1-9.

[10] MELEKHOV I, YLIOINAS J, KANNALA J, et al. Relative camera pose estimation using convolutional neural networks[C] // International Conference on Advanced Concepts for Intelligent Vision Systems. Springer, Cham, 2017: 675-687.

[11] 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, 2016, 1(1): 18-25.

[12] 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. New York: IEEE, 2017: 2043-2050.

[13] CLARK R, WANG S, WEN H, et al. VINet: Visual-inertial odometry as a sequence-to-sequence learning problem[C] // Thirty-First AAAI Conference on Artificial Intelligence. Washington D. C.: AIAA, 2017.

[14] TURAN M, ALMALIOGLU Y, ARAUJO H, et al. Deep endovo: A recurrent convolutional neural network (rcnn) based visual odometry approach for endoscopic capsule robots[J]. Neurocomputing, 2018, 275: 1861-1870.

[15] KONDA K, MEMISEVIC R. Unsupervised learning of depth and motion[J]. arXiv preprint arXiv:1312.3429, 2013.

[16] GODARD C, MAC AODHA O, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[J]. Computer Vision and Pattern Recognition (CVPR), 2017, 2(6): 7.

[17] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C] // Advances in Neural Information Processing Systems, 2012: 1097-1105.

[18] ZHOU T, BROWN M, SNAVELY N, et al. Unsupervised learning of depth and ego-motion from video[J]. Computer Vision and Pattern Recognition (CVPR), 2017, 2(6): 7.

[19] VIJAYANARASIMHAN S, RICCO S, SCHMID C, et al. Sfm-net: Learning of structure and motion from video [J]. arXiv preprint arXiv:1704.07804, 2017.

[20] LI R, WANG S, LONG Z, et al. UnDeepVO: Monocular visual odometry through unsupervised deep learning [J]. arXiv preprint arXiv:1709.06841, 2017.

[21] KUZNIETSOV Y, STÜCKLER J, LEIBE B. Semi-supervised deep learning for monocular depth map prediction[C] // IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 6647-6655.

[22] YIN Z, SHI J. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose[J]. arXiv preprint arXiv:1803.02276, 2018.

[23] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[24] ILG E, MAYER N, SAIKIA T, et al. Flownet 2.0: Evolution of optical flow estimation with deep networks [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York:IEEE, 2017.

[25] PENG T, LI F. Bag of visual word model based on binary hashing and space pyramid[C]//Eighth International Conference on Digital Image Processing (ICDIP 2016).

International Society for Optics and Photonics, 2016, 10033: 100335T.

[26] OLIVA A, TORRALBA A. Modeling the shape of the scene: A holistic representation of the spatial envelope [J]. International Journal of Computer Vision, 2001, 42(3): 145-175.

[27] ARANDJELOVIC R, ZISSERMAN A. All about VLAD [C]//2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York:IEEE, 2013: 1578-1585.

[28] PERRONNIN F, DANCE C. Fisher kernels on visual vocabularies for image categorization[C]//Computer Vision and Pattern Recognition, IEEE Conference on CVPR IEEE. New York:IEEE,2007: 1-8.

[29] ZHOU B, LAPEDRIZA A, XIAO J, et al. Learning deep features for scene recognition using places database [C]//Advances in Neural Information Processing Systems. 2014: 487-495.

[30] HOU Y, ZHANG H, ZHOU S. Convolutional neural network-based image representation for visual loop closure detection[C]//IEEE International Conference on Information and Automation. New York: IEEE, 2015: 2238-2245.

[31] BAI D, WANG C, ZHANG B, et al. Matching-range-constrained real-time loop closure detection with CNNs features[J]. Robotics and Biomimetics, 2016, 3(1): 15.

[32] ZHANG X, SU Y, ZHU X. Loop closure detection for visual SLAM systems using convolutional neural network [C]//The 23rd International Conference on Automation and Computing (ICAC). New York:IEEE, 2017: 1-6.

[33] XIA Y, LI J, QI L, et al. Loop closure detection for visual SLAM using PCANet features[C]//2016 International Joint Conference on Neural Networks (IJCNN). New York:IEEE, 2016: 2274-2281.

[34] GAO X, ZHANG T. Unsupervised learning to detect loops using deep neural networks for visual SLAM system[J]. Autonomous Robots, 2017, 41(1): 1-18.

[35] TORII A, ARANDJELOVIĆ R, SIVIC J, et al. 24/7 place recognition by view synthesis[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York:IEEE,2015: 1808-1817.

[36] TORII A, SIVIC J, PAJDLA T, et al. Visual place recognition with repetitive structures[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York:IEEE,2013: 883-890.

[37] SÜNDERHAUF N, SHIRAZI S, DAYOUB F, et al. On the performance of convnet features for place recognition [C]//2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). New York:IEEE, 2015: 4297-4304.

[38] SÜNDERHAUF N, SHIRAZI S, JACOBSON A, et al. Place recognition with convnet landmarks: viewpoint-robust, condition-robust, training-free[J]. Proceedings of Robotics: Science and Systems XII, 2015.

[39] 薛昆南, 薛月菊, 毛亮, 等. 基于卷积词袋网络的视觉识别[J]. 计算机工程与应用, 2016, 52(21): 180-187.

[40] ARANDJELOVIC R, GRONAT P, TORII A, et al. NetVLAD: CNN architecture for weakly supervised place recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE,2016: 5297-5307.

[41] GRANSTRÖM K, SCHÖN T B. Learning to close the loop from 3D point clouds[C]//2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). New York:IEEE, 2010: 2089-2095.

[42] SATTLER T, LEIBE B, KOBBELT L. Fast image-based localization using direct 2d-to-3d matching[C]//2011 IEEE International Conference on Computer Vision (ICCV). New York:IEEE, 2011: 667-674.

[43] SATTLER T, HAVLENA M, RADENOVIC F, et al. Hyperpoints and fine vocabularies for large-scale location recognition[C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2015: 2102-2110.

[44] LI Y, SNAVELY N, HUTTENLOCHER D P, et al. Worldwide pose estimation using 3d point clouds[M]. Large-Scale Visual Geo-Localization. Springer, Cham, 2016: 147-163.

[45] ELBAZ G, AVRAHAM T, FISCHER A. 3d Point cloud registration for localization using a deep neural network auto-encoder[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2017: 2472-2481.

[46] ZENG A, SONG S, NIEßNER M, et al. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2017: 199-208.

[47] CHARLES R Q, SU H, KAICHUN M, ET al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2017: 77-85.

[48] QI C R, YI L, SU H, et al. Pointnet: Deep hierarchi-

cal feature learning on point sets in a metric space[C]//Advances in Neural Information Processing Systems. 2017: 5105-5114.

[49] IZADI S, KIM D, HILLIGES O, et al. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera[C]//Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology. ACM, 2011: 559-568.

[50] WHELAN T, LEUTENEGGER S, SALAS-MORENo R, et al. ElasticFusion: dense SLAM without a pose graph[M]. Robotics: Science and Systems, 2015.

[51] ZHOU H, UMMENHOFER B, BROX T. Deeptam: Deep tracking and mapping[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 822-838.

[52] ALEOTTI F, TOSI F, POGGI M, et al. Generative adversarial networks for unsupervised monocular depth prediction[C]//European Conference on Computer Vision. Springer. Cham, 2018: 337-354.

[53] REZENDE D J, ESLAMI S M A, MOHAMED S, et al. Unsupervised learning of 3d structure from images[C]//Advances in Neural Information Processing Systems. 2016: 4996-5004.

[54] BLOESCH M, CZARNOWSKI J, CLARK R, et al. CodeSLAM—learning a compact, optimisable representation for dense visual SLAM[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 2560-2568.

[55] DAVISON A J. FutureMapping: The computational structure of spatial AI systems[J]. arXiv preprint arXiv: 1803.11288, 2018.

[56] PARISOTTO E, CHAPLOT D S, ZHANG J, et al. Global pose estimation with an attention-based recurrent network[J]. arXiv preprint arXiv: 1802.06857, 2018.

[57] CHAPLOT D S, PARISOTTO E, SALAKHUTDINOV R. Active neural localization[J]. arXiv preprint arXiv: 1801.08214, 2018.

[58] MIROWSKI P, GRIMES M K, MALINOWSKI M, et al. Learning to navigate in cities without a map[J]. arXiv preprint arXiv: 1804.00168, 2018.

[59] 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.

[60] HAN X, LEUNG T, JIA Y, et al. Matchnet: Unifying feature and metric learning for patch-based matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2015: 3279-3286.

[61] LI S, ZHANG T, ZHANG D, et al. Metric learning for patch-based 3-D image registration[J]. IEEE Transactions on Automation Science and Engineering, 2019.

作者简介：李少朋(1992—)，男，博士研究生，研究方向为人工智能、视觉 SLAM；**张　涛**(1969—)，男，博士研究生，研究方向为机器人、人工智能、控制理论和飞行器控制.