doi:10.3772/j.issn.1002-0470.2020.01.005

基于深度学习的动态场景相机姿态估计方法①

路 昊②* 石 敏③* 李 昊* 朱登明** ***

(*华北电力大学控制与计算机工程学院 北京 102200)

(**中国科学院计算技术研究所前瞻研究实验室 北京 100190)

(***太仓中科信息技术研究院 太仓 215400)

摘 要 针对现有增强现实技术中应用较为广泛的基于标识物的定位注册方法的不足,提出了一种在不断变化运动的复杂动态场景下估计相机连续运动的3维姿态的方法。基于深度神经网络对输入图像序列建立端到端的学习模型,将卷积神经网络(CNN)作为高层特征提取器,同时利用长短期记忆神经网络(LSTM)建立视频连续帧之间的时序关系,完成相机连续运动的3维姿态估计,从而避免了相机快速运动及场景不断运动变换导致图像特征提取效果不好的情况。另一方面,通过迁移学习的方法来预测未知视频序列的相机3维姿态信息,解决了原始数据量不够的问题。在公共数据集上的实验结果表明,相对于 PoseNet,基于连续视频序列的输入,其预测精度得到一定的提升。

关键词 姿态估计;卷积神经网络(CNN);长短期记忆神经网络(LSTM);动态场景;迁移学习

0 引 言

近20年的时间里,增强现实无论在技术层面还是应用层面都得到了前所未有的发展。对于现实中普遍应用的单目视觉而言,连续运动的相机的位姿估计是3维姿态估计的核心技术。传统的3维姿态估计方法存在许多的局限,GPS 无法应用于室内定位且定位精度较低;高精度的惯性导航单元价格过于昂贵,低廉的精度损失又相对过大;基于人工标识的定位方法需要预先设定场景,导致应用场景无法扩展。而反观人类,依靠双眼获取的视觉信息就可实现对周围环境的感知,因此基于连续视频序列的相机运动姿态估计,即通过摄像头采集的视频图像信息就可以实现在位置环境下的自身定位,同时恢复周围环境的3维结构,这对于增强现实设备实现自身与虚拟物体的相对运动映射关系尤为重要。

由于基于特征点匹配估计相机姿态的方法受目标场景特征提取的难易程度影响较大,尤其是在复杂动态场景下,由于相机的快速运动或者目标场景的快速变化导致特征点提取效果不理想。基于这种情况,利用端到端的深度学习方法来解决这一问题成了研究的重点。

Kendall 等人[1]提出了一种具有鲁棒性和实时性的单目相机6自由度重定位方法。通过卷积神经网络(convolution neural network, CNN)对输入的单张 RGB 图像进行6自由度相机姿态的回归预测。

Wu 等人[2]提出了一种基于卷积神经网络(CNN)的相机姿态重定位方法。利用卷积神经网络(CNN)对图片中采样得到的像素点的世界坐标进行预测,建立预测点与对应相机坐标点的映射关系。Kendall 和 Cipolla[3]在另一篇文章中提出了建

① 国家重大科技专项(2017ZX05019005)资助项目。

② 男,1997 年生,硕士生;研究方向:虚拟现实与图形图像处理等;E-mail: 2535916@qq.com

③ 通信作者,E-mail: shi_min@ncepu.edu.cn

(收稿日期:2019-02-21)

立基于不确定性的贝叶斯卷积神经网络来对单张 RGB 图像进行相机位姿的回归预测，提高预测的置信度。

Costante 等人[4]提出了一种帧间运动学习估计方法。利用卷积神经网络对连续视频图像中稠密光流进行视觉特征提取，通过全连接神经网络再对新学习的视觉特征进行帧间的相机运动估计。

Zhou 等人[5]提出了一种基于无监督学习框架估计单目相机运动及深度信息。通过利用单视角深度网络以及多视角位姿网络对非结构化的视频序列进行端到端的联合学习。

Walch 等人[6]提出了利用长短记忆神经网络（long short-term memory neural network，LSTM）[7]对单帧图像的特征进行降维，并与 PoseNet[1]网络结构相结合，提高了定位精度，然而并没有利用多个连续图像之间的关系。

Wang 等人[8]提出了对连续的 2 帧图片叠加之后进行特征提取，提取的特征经过循环神经网络进行时序建模估计位姿，鲁棒性得到增强，但是选用的卷积网络层数较少，整体定位精度一般。

本文提出了一种基于由深度卷积神经网络（CNN）和长短期记忆神经网络（LSTM）组成的端到端的结构，使用深度卷积网络提取特征并传进长短期记忆神经网络，对连续视频序列中相机位置之间的相对运动关系建模，进行相机姿态的回归预测，提高了模型收敛的速度，并且提升了定位精度。

1　深度相机位姿估计

整体网络结构（图 1）属于端到端的模型，每一帧图片对应一个表示相机位姿的 7 维向量 $\boldsymbol{p}=[\boldsymbol{x},\boldsymbol{q}]$，其中 $\boldsymbol{x}=(x,y,z)^{\mathrm{T}}\in\mathbb{R}^3$ 表示位置向量，使用 4 元数 $\boldsymbol{q}$ 表示方向，是考虑到标准化后的 4 元数同样能映射到旋转矩阵，并且标准化的数据更容易进行网络训练。

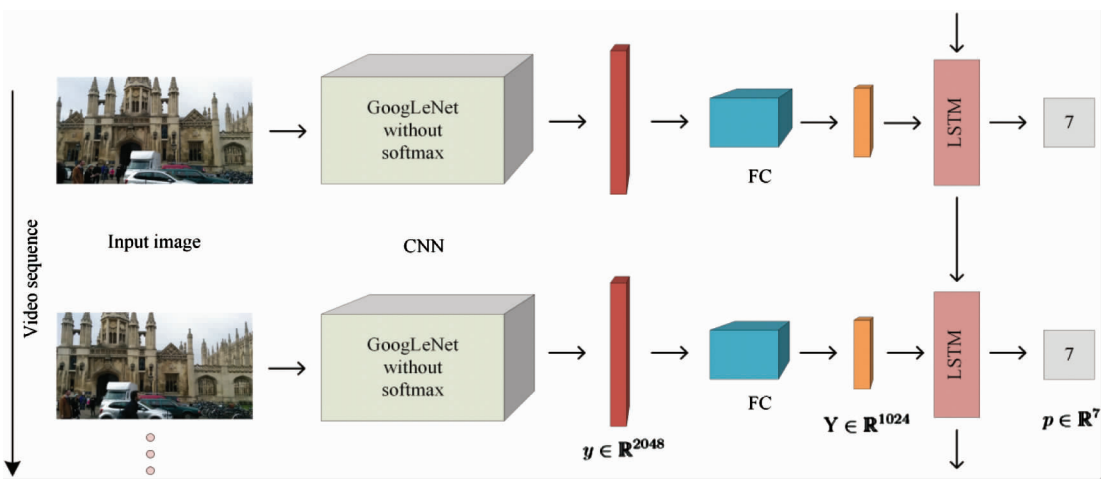

图 1　网络结构图

整个网络的损失函数如式（1）所示：

$$loss(Seq_i)=\sum_{t=1}^{n}\|\boldsymbol{x}'_t-\boldsymbol{x}_t\|_2+\beta\sum_{t=1}^{n}\|\boldsymbol{q}'_t-\boldsymbol{q}_t\|_2 \quad (1)$$

其中，seq_i 是网络输入的第 i 序列，(x_t,q_t)，(x'_t,q'_t) 则是对应的训练数据的真实值与估计值，前者 $\boldsymbol{x}_t$ 表示第 t 帧时刻的对应 3 维坐标位置，后者 $\boldsymbol{q}_t$ 表示第 t 帧时刻的对应相机的姿态角度。β 是根据数据集设定的比例因子，以保持位置和姿态误差的估计值近似相等。

为了充分发挥卷积神经网络提取特征的能力以及长短期记忆神经网络的时序相关性，同时考虑到数据量规模，本文将 CNN 与 LSTM 相结合，提出了一种端到端的模型：

（1）首先通过迁移学习将分类数据集上训练好的权重参数作为 CNN 部分的初始化参数，将针对图像分类任务的预训练模型作为本文目标场景的初始化模型。

(2)利用初始化后的卷积神经网络对目标训练集进行特征提取,从视频序列中提取特征,同时对特征进行降维处理并送入 LSTM 单元中。

(3)利用长短期记忆神经网络对 CNN 输出进行连续时序上的学习训练,最终输出一个由 3 维位置向量以及代表方向的 4 元数组成的 7 维姿态向量。

2　基于 CNN 和 LSTM 的网络模型

2.1　卷积神经网络

考虑动态场景的运动模糊和光照的鲁棒性,本文选用卷积神经网络提取图像特征。本文深度卷积网络结构参考了 PoseNet 的思想,选择使用在分类任务上表现优秀的深度卷积网络。表 1 在 GoogLeNet 的网络结构上进行修改,去掉了原本用于分类任务而设定的 softmax 层,取而代之的是在最后增加了一层全连接层对高维特征进行降维,得到的结果序列传递给 LSTM 进行时序建模。如表 1 所示,为本文深度卷积网络各层输出,本文输入统一为 224×224 分辨率的 3 通道彩色图像,中间的网络层相比 GoogLeNet,去掉了 2 个原本用于中间输出的分支结构,网络的最终输出为 2 048 维特征向量。

表 1　卷积神经网络各层输出

层类别	卷积核尺寸	步长	填充	输出尺寸
convolution	7×7	2	3	112×112×64
max pool	3×3	2	1	56×56×64
convolution	3×3	1	1	56×56×192
max pool	3×3	2	1	28×28×192
2×Inception				28×28×480
max pool	3×3	2	1	14×14×480
5×Inception				14×14×832
max pool	3×3	2	1	7×7×832
2×Inception				7×7×1 024
avg pool	7×7	1	1	1×1×1 024
linear				1×1×2 048

本文采用的这种卷积网络为深度卷积网络,涉及到参数计算的共 13 层,深层网络可以得到更深层次的特征图;另一方面,这种卷积网络结构采取了 9 个 Inception 模块,这种模块采用不同尺度卷积核提取特征并聚合,增加了网络模型的宽度。这种网络从深度和宽度上的设计保证了网络的性能,更适用于动态场景的复杂特征提取。

其中卷积层在卷积神经网络中用于特征提取,卷积层输出大小为

$$O = (W - K + 2P)/S + 1 \tag{2}$$

其中,O 为输出尺寸,W 为长度或宽度,K 为过滤器尺寸,P 是填充,S 是步幅。

为避免神经网络退化为最原始的感知机,本文的每个卷积层之后都选用 ReLU 作为激活函数,使得输入输出之间保持高度非线性的关系。

在这种卷积神经网络中,还使用了 2 种池化操作:最大池化和平均池化。池化最直接的作用就是引入了不变性,这种不变性包括了平移不变性、旋转不变性以及尺度不变性。同时池化只保留了主要的特征,减少了参数,提高了模型的泛化能力,避免了过早地出现过拟合现象。

池化层输出的大小为

$$O = (W - F)/S + 1 \tag{3}$$

其中,O 为输出尺寸,W 为长度或宽度,F 为池化单元尺寸,S 为步幅。

在本文卷积网络最后部分设置了全连接层,该层的作用是将网络学习到的特征映射到样本的标记空间中,即把卷积输出的 2 维特征图转化为一个 1 维的向量,将由 1 维向量组成的序列输入到 LSTM 中进行时序建模。

2.2　长短期记忆网络

在估计连续视频序列中的相机运动问题上,考虑到视频前后的连续相关性,上一状态的相机运动参数信息同样会影响下一状态的位置信息,所以在 CNN 之后,还设计了一个 LSTM 进行时序学习。视频序列通过深度卷积网络得到的特征序列作为 LSTM 的输入,学习连续图片特征之间的时序相关性,并使得每一个图片特征对应一个 7 维的输出,包括 3 维的位置向量和代表方向的 4 元数。

在 LSTM 单元当中被放置了 3 个门结构,分别叫做输入门、遗忘门和输出门。输入门的作用对象

是细胞状态,能够将新的信息选择性地记录到细胞状态中,如式(4)所示:

$$i_t = sigmoid(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (4)$$

其中,h_{t-1} 为上一时刻输出,x_t 为当前输入,W 和 b 分别为权值和偏置。

遗忘门 f_t 的作用对象是细胞状态,能够将历史信息选择性遗忘,如式(5)所示:

$$f_t = sigmoid(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (5)$$

输出门 o_t 控制当前信息流出,由上一时刻输出 h_{t-1} 和当前输入 x_t 共同决定,如式(6)所示:

$$o_t = sigmoid(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (6)$$

2.3　CNN + LSTM 网络模型

因为基于深度神经网络训练学习模型,需要大量的训练样本,但是本文任务场景的当前数据集还不足以支撑完整的神经网络的训练,很容易导致过拟合现象的发生。在训练之前,本文利用了在 Places 数据集上训练好的 GoogLeNet[9] 的权重参数作为 CNN 部分的初始化参数,即将用作图像识别分类的预训练模型用作本文目标场景的初始化模型,这一方法在 Kendall 等人[1]的工作中得到证明。

在动态场景下,由于相机的快速运动以及目标场景存在变化,所以场景中的特征点不能很好地通过特征提取及匹配算法计算出相应的 3 维空间位置,从而导致无法计算连续视频序列之间的相机相对运动关系。因此本文将 CNN 和 LSTM 相结合,摒弃了提取特征点的步骤,直接利用深度卷积网络作为特征提取器,在视频序列中的每一帧提取一个 2 048维的特征,本文将输出的 2 048 维特征向量作为序列传送到 LSTM 中进行时序建模。

实践中直接将 2 048 维特征序列输入 LSTM 并没有取得很好的效果,这是因为 2 048 维特征对于 LSTM 太长以至于无法很好地关联前后特征。为了减少 LSTM 输入的特征维度,并且尽量少丢失特征信息,在 CNN 网络输出后添加了一个全连接层,将 2 048 维的特征降为 1 024 维,并将 1 024 维特征序列传递给 LSTM 进行时序建模,构建图像特征的时序关系,从而保证视频中每帧图像的位姿状态并不是完全独立的,而是存在前后依赖。实验表明,这种做法增强了模型的学习能力,并且一定程度上提高了精度。

3　实　验

3.1　实验环境

本文实验环境的机器配置参数如表 2 所示。

表 2　实验环境配置

硬件	参数
处理器	Intel Corei7-2600 CPU @ 3.40 GHz
内存	DDR3 1 600 MHz,8 G
硬盘	三星 SL,6 729 GB,7 200 r/min
显存	GTX1 070,8 G

3.2　实验结果

在微软公共数据集 7-Scenes[10] 上选取了其中 3 个室内场景数据集 Chess、Office、Heads 以及 PoseNet 文章所发布的 CambridgeLandmarks 数据集中室外场景数据集 King's College(图 2)。这 2 种数据集的差异很大,7-Scenes 数据集的特点是空间范围非常小,而图像非常多,覆盖密集;而 CambridgeLandmarks 数据集的特点是空间范围很大,图片相对较少,覆盖很稀疏。

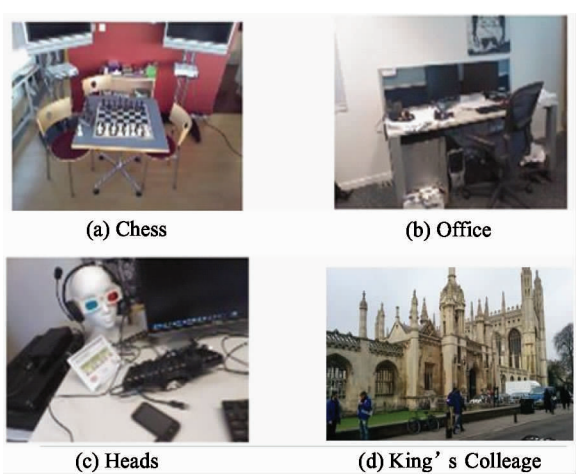

图 2　测试数据集

实验设定 batchsize 为 75,学习率为 0.0001。为了方便进行对比,每个数据集的训练集、测试集选取相同,且损失函数式(1)中的 β 同 PoseNet 保持一

致:在 King's College 数据集上设为 500,在 Heads、Chess 和 Office 室内场景数据集上设定为 120、500、250。

本文实时记录了在训练过程中,损失(loss)值随迭代次数的变化情况,并与 PoseNet 的损失函数变化情况进行了对比。以 Chess 数据集的实验结果为例,图 3 展示了本文网络结构与 PoseNet 的 loss 值均随着迭代次数增加的变化,在 2 种网络结构上都会趋于收敛,但是本文网络结构在初始阶段 loss 值明显低于 PoseNet,并且最终收敛时 loss 值低于 PoseNet。在不同数据集上均取得了类似的情况,整体的 loss 明显低于 PoseNet,证明了本文提出的网络更容易训练。

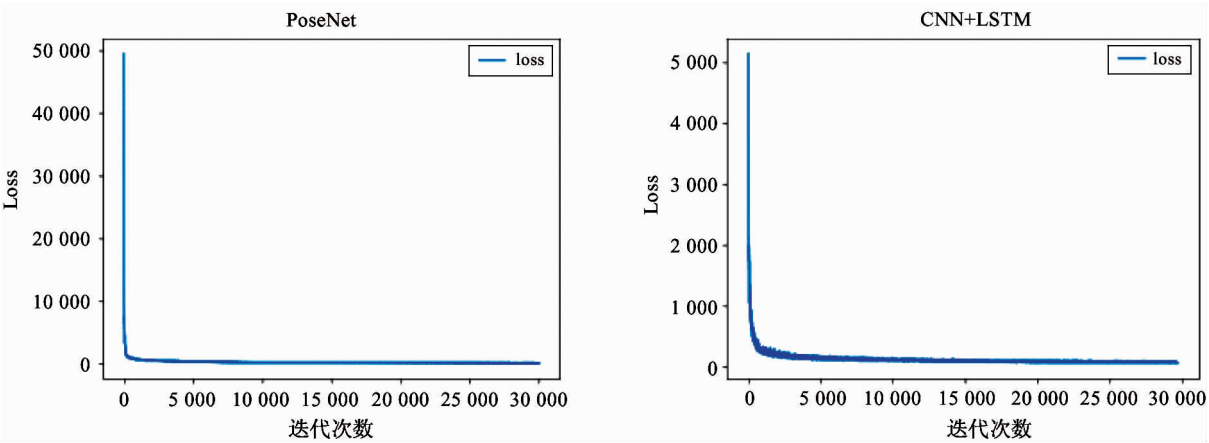

图 3　Chess 数据集 loss 收敛情况

如表 3 所示,在多个公共数据集上,本文采用的网络结构和 PoseNet 网络结构预测姿态结果的误差中值进行了对比,其中位置误差 E_x 以单位米(m)衡量,角度误差 E_q 以单位度(°)衡量。

表 3　各数据集误差表

数据集	PoseNet		CNN + LSTM	
	E_x	E_q	E_x	E_q
Heads	0.35 m	8.68 °	0.28 m	6.02 °
Chess	0.46 m	6.08 °	0.34 m	4.12 °
Office	0.51 m	7.12 °	0.48 m	6.32 °
King's College	2.17 m	3.07 °	1.96 m	3.01 °

Heads 数据集是 4 个数据集中场景空间范围最小的,仅 1 m^3。其中连续的帧差异较小,移动速度较小。在该数据集上,将 1 000 张连续图片作为训练集,另外 1 000 张连续图片作为测试集,实验结果的位置误差 E_x 在 4 个数据集中最小,预测的相机位置误差中值仅 0.28 m,并且相比 PoseNet,本文的位置精度提高了 20.0%,方向精度提高了 30.6%。

Chess 数据集的空间范围为 6 m^3,数据集图片较多几乎覆盖整个空间。其中图片中含有棋盘格纹理特征丰富,并且连续图像的旋转较慢。选取 4 000 张图片作为训练集,另外 2 000 张图片作为测试集,实验结果中方向误差 E_q 在室内数据集中最小。相比 PoseNet,本文方法位置精度提高了 26.1%,方向精度提高了 32.2%。

Office 数据集空间范围 7.5 m^3,图片较多但是很多张图片存在运动模糊的情况。Walch 等人[6]的工作结果表明,在 Office 数据集上,传统方法无法对其中的一些图片进行有效识别,无法预测位姿;而本文方法在该数据集上,每一张图能给出一个合理的位姿预测结果,这证明了深度方法在动态场景下相比传统方法的优势和价值。本文方法训练集选取 6 000张序列帧,测试集选取 4 000 张的序列帧,相比 PoseNet,位置精度提高了 5.9%,方向精度提高了 11.2%。

King's College 数据集为大型室外场景,涉及 5 600 m^2的区域。场景面积大,但该数据集图片相对较少,训练容易产生过拟合。训练集选取1 220帧

图片,测试集选取 343 帧图片。位置误差 E_x 明显大于其他 3 个数据集,达到 1.96 m,但是相比 PoseNet,本文方法的位置精度提高了 9.7%,方向精度基本一致。

为了验证采用迁移学习的重要性,本实验在 King's College 数据集上增设了 1 组对比实验,表 4 展示了采用迁移学习前后的误差对比。在 PoseNet 及本文模型上,迁移学习都展示了良好的效果。引入初始化参数后,本文模型的位置精度提高了 23.1%,方向精度提高了 15.4%。

表 4　迁移学习前后误差对比

	PoseNet		CNN + LSTM	
	E_x	E_q	E_x	E_q
采用迁移学习	2.17 m	3.07 °	1.96 m	3.01 °
不用迁移学习	2.75 m	4.13 °	2.55 m	3.56 °

由于采用了 LSTM,当前帧会对同序列帧预测结果产生影响。如图 4 所示,当出现大面积遮挡的帧,甚至无效帧时,对该帧图片能够得到一个相对合理的值,但相邻帧的预测精度下降。本实验对不包含该类帧的序列进行测试,位置误差 E_x 达到 1.65 m。

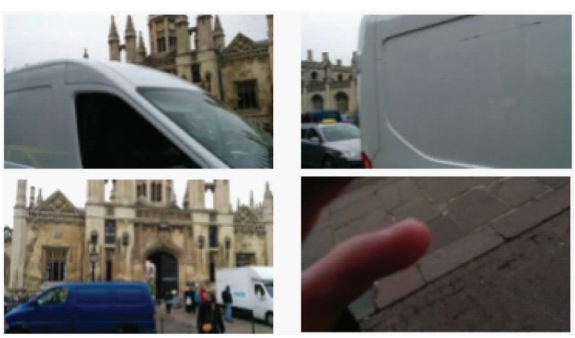

图 4　King's College 部分测试集

针对大型室外场景的实验结果位置误差明显大于小型室内场景的结果,可能原因一方面是空间范围大,另一方面是相机移动速度快。所以增设了 1 组实验对相机速度和位置误差的关系进行分析。在 King's College 测试集中选取了数据集中相对低速和高速的 40 帧序列进行测试。如图 5、图 6 所示,横轴均为选定的测试序列,纵轴为每一帧图片对应的位置误差大小,单位为 m。

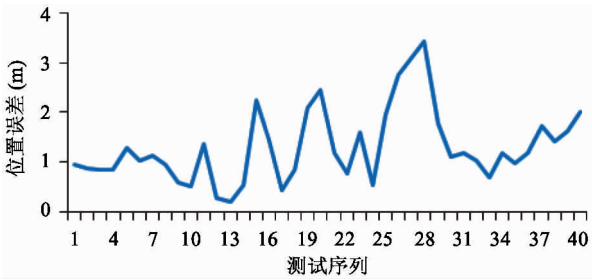

图 5　低速序列位置误差

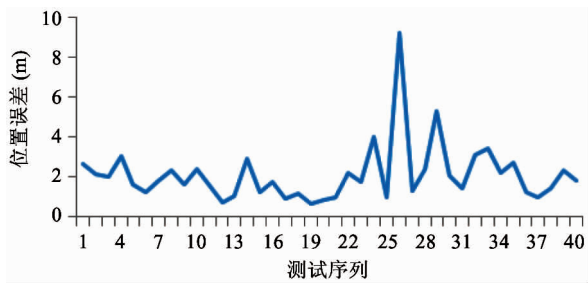

图 6　高速序列位置误差

低速序列中,相邻帧的相机平均位移为 0.745 m,测试结果的平均位置误差为 1.290 m,误差中值为 1.138 m;高速序列相邻帧的相机平均位移为 1.363 m,测试结果的平均位置误差为 2.126 m,误差中值为 1.794 m。所以移动速度较快的序列所估计的位置误差较大,且明显高于测试集整体的误差中值,与推测相符。

4　结　论

针对基于自然特征的视觉 3 维注册方法对于纹理特征较稀疏的场景以及复杂动态场景表现不理想的情况,本文提出了一种基于 CNN + LSTM 网络的端到端模型,以深度 CNN 作为特征提取器,以 LSTM 进行时序建模,从而建立连续视频序列中相机位置之间的相对运动关系。同 PoseNet 相比,获得了更快的收敛速度,更低的损失值,更高的精度。未来的工作将针对高速移动相机的位姿估计展开,以便提高在高速情况下估计的位置精度。

参考文献

[1] Kendall A, Grimes M, Cipolla R. PoseNet: a convolutional network for real-time 6-DOF camera relocalization [C]//2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015:2938-2946

[2] Wu J, Ma L, Hu X. Predicting world coordinates of pixels in RGB images using Convolutional Neural Network for camer are localization[C]//7th International Conference on Intelligent Control and Information Processing, Siem Reap, Combodia, 2017:161-166

[3] Kendall A, Cipolla R. Modelling uncertaintyin deep learning for camera relocalization[J]. *Designing Engineering and Analyzing Reliable and Efficient Software*, 2015, 31:4762-4769

[4] Costante G, Mancini M, Valigi P, et al. Exploring representation learning with CNNs for frame-to-frame ego-motion estimation[J]. *IEEE Robotics & Automation Letters*, 2015, 1(1):18-25

[5] Zhou T, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motionfrom video[C]//Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6612-6619

[6] Walch F, Hazirbas C, Leal-Taixe L, et al. Image-based localization using LSTMS for structured feature correlation [C]//Proceedings of the IEEE International Conferenceon Computer Vision, Venice, Italy, 2017: 627-637

[7] Hochreiter S, Schmidhuber J. Long short-term memory [J]. *Neural Computation*, 1997, 9(8):1735-1780

[8] Wang S, Clark R, Wen H, et al. DeepVo: towards end-to-end visual odometry with deep recurrentcon volutional neural networks[C]//2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 2017: 2043-2050

[9] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015:1-9

[10] Shotton J, Glocker B, Zach C, et al. Scene coordinate regression forests for camera relocalization in RGB-D images[C]//IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013:2930-2937

A method of estimating camera pose in dynamic scene based on DNN

Lu Hao*, Shi Min*, Li Hao*, Zhu Dengming**,***

(*Control and Computer Engineering Institute, North China Electric Power University, Beijing 102200)

(**Prospective Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

(***Taicang Institute of Information Technology, Taicang 215400)

Abstract

Aiming at the shortcomings of the widely used marker-based location registration, a method for estimating the three-dimensional(3D) pose of the camera continuous motion is proposed, which can be used under the complex dynamic scene. An end-to-end learning model for the input image sequence is established based on deep neural network. Convolutional neural network (CNN) is used as a high-level feature extractor, while long short-term memory neural network (LSTM) is used to establish the timing correlation between consecutive video frames. The proposed method is used to estimate the 3D pose of the camera's continuous motion, poor effect of extracting image feature due to the rapid movement of the camera and the continuous motion change can be avoided. Moreover, the method of migration learning is used to predict the camera 3D pose information of the unknown video sequence, which solves the problem of insufficient original data volume. Experimental results on public datasets show that, compared to PoseNet, the prediction accuracy is improved based on the input of continuous video sequences.

Key words: pose estimation, convolutional neural network(CNN), long short-term memory neural network (LSTM), dynamic scene, transfer learning