文章编号：1005-5630(2020)04-0033-08 **DOI**: 10.3969/j.issn.1005-5630.2020.04.006

基于深度卷积神经网络的视觉里程计研究

苏健鹏，黄影平，赵柏淦，胡 兴
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘要：视觉里程计利用视频信息来估计相机运动的位姿参数，实现对智能体的定位。传统视觉里程计方法需要特征提取、特征匹配/跟踪、外点剔除、运动估计、优化等流程，解算非常复杂，因此，提出了基于卷积神经网络的方法来实现端对端的单目视觉里程计。借助卷积神经网络对彩色图片自动学习提取图像帧间变化的全局特征，将用于分类的卷积神经网络转化为帧间时序特征网络，通过三层全连接层输出相机的帧间相对位姿参数。在 KITTI 数据集上的实验结果表明，提出的 Deep-CNN-VO 模型可以较准确地估计车辆的运动轨迹，证明了方法的可行性。在简化了复杂模型的基础上，与传统的视觉里程计系统相比，该模型的精度也有所提高。

关键词：视觉里程计；自主定位；深度学习；卷积神经网络
中图分类号：TP 391 **文献标志码**：A

Research on visual odometry using deep convolution neural network

SU Jianpeng，HUANG Yingping，ZHAO Bogan，HU Xing
（School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China）

Abstract: The visual odometry uses visual cues to estimate the pose parameters of the camera motion and localize an agent. Existing visual odometry employs a complex process including feature extraction, feature matching/tracking, and motion estimation. The processing is complicated. This paper presents an end-to-end monocular visual odometry by using convolution neural network (CNN). The method modifies a classification CNN into a sequential inter-frame variation CNN. In virtue of the deep learning technique, the method extracts the global inter-frame variation feature of video images, and outputs pose parameters through three full-connection convolution layers. It has been tested in the public KITTI database. The experimental results show the proposed Deep-CNN-VO model can estimate the motion trajectory of the camera and the feasibility of the method is proved. On the basis of simplifying the complex model, the accuracy is improved compared with the traditional visual odometry system.

收稿日期：2019-10-24
基金项目：国家自然科学基金（61374197）
作者简介：苏健鹏 (1994—)，男，硕士研究生，研究方向为深度学习在无人驾驶中的应用。E-mail：sujianpeng12345@163.com
通信作者：黄影平 (1966—)，男，教授，研究方向为无人驾驶。E-mail：huangyingping@usst.edu.cn

Keywords: visual odometry; self-localization and navigation; deep-learning; convolutional neural network

引　言

通过单目或多目摄像机实现车辆定位的方法称之为视觉里程计[1]。定位就是要获取车辆在三维空间中的位置和姿态信息。位置是世界坐标系下的三维坐标，姿态则是车辆运动方向与 3 个坐标轴的夹角，即俯仰角、航向角和侧滚角。位置和姿态 6 个自由度的信息可以唯一确定车辆在指定坐标系中的空间状态。视觉里程计被广泛用于无人驾驶、机器人、潜航器等，是继全球卫星定位系统，惯性导航，车轮里程计等定位技术后的一种新的导航技术，其性价比及可靠性较高。

视觉里程计通过相机获取图片序列，经过特征提取、特征匹配/跟踪，外点剔除和运动估计等处理模块得到车辆的位姿更新，进而推算运动轨迹，实现定位导航[1]。Nister 等[2] 首次实现实时视觉里程计系统，最先采用基于匹配的方法代替基于跟踪的方法进行特征关联以避免基于互相关的跟踪而引起的特征漂移，采用随机采样一致性（RANSAC）算法消除外点，并给出了单目和双目视觉里程计的实现途径和方法，在单目视觉里程计系统中，提出了被后人广泛使用的 5 点算法。在双目视觉里程计系统中，他们提出了采用 3D 到 2D 重投影误差代替 3D 点之间的欧拉距离的误差的运动估计方法。这些工作为视觉程计的研究奠定了基础，当前大多数视觉定位导航系统都遵循这种框架。在实际场景中，单纯的角点并不能满足需求，于是研究人员设计更加稳定的图像特征如 SIFT、SUFT 等。虽然 SIFT和 SUFT 考虑到图像变换过程中的许多问题，但是计算量较大。2011 年 Rublee 等[3] 提出了 ORB（oriented FAST and rotated BRIEF）算法，该算法提取的图片特征不仅保留了 SIFT 和 SUFT特征的优点，且速度是 SIFT 算法的 30 多倍。2011 年 Geiger 等[4]将图片的稀疏特征运用到视觉里程计中，提出了实时单目视觉里程计的 VISO2-M 算法，该算法是当时最好的 SLAM 算法之一。2015 年 Mur-Artal 等[5] 在 ORB 算法上进行研究，提出了 ORB-SLAM 算法，并在精度上取得了良好的效果。Mur-Artal 等[6] 继续在 ORB-SLAM的基础上进行优化并推出了 ORB-SLAM2，是目前定位较精确的视觉定位系统。

上述方法都是基于几何原理，当更换场景后需要对代码中的参数进行大量调整以适应新的场景需求，而使用深度学习的算法完全不同于上述思路。2008 年 Roberts 等[7] 尝试使用光流和机器学习相结合的方法预测车辆运动轨迹，提出了一个基于神经网络视觉里程计的模型，模型由 160 个 KNN（K-Nearest-Neighbors）学习机组成，这种模型使用 K 近邻的方法计算每个单元的光流，再将光流转化为位姿参数。2015 年 Kishore 等[8] 首次提出了使用卷积神经网络（Convolutional Neural Network）来学习视觉里程计，其模型将传统方法与深度学习的方法相结合，先使用双目图片估计出深度，再利用两个不同的卷积神经网络分别学习图片特征得到车辆的角度和速度。2015 年 Kendall 等[9] 提出了一种端对端的视觉里程计模型 PoesNet，该模型首次将视觉里程计设计为端对端的网络，即输入图片经过神经网络后直接输出位姿，不过由于每次输入的都是单张图片，无法建立视觉里程计的时序性，导致其鲁棒性及泛化能力较差，在新场景的应用中定位偏移较大。为了解决时序性的问题，Ronald 等[10] 提出 DeepVO 的网络模型，该模型将长短期记忆网络（Long Short-Term Memory, LSTM）加入到整个神经网络中，通过 LSTM 网络可以很好的构建图片之间的联系，形成图片之间的时序性。2018 年 Mccormac 等[11] 用 Mask-RCNN 网络对距离函数进行重构，设计出一个在线的 SLAM 系统并具有很高的内存效率。相对于传统的视觉里程计，基于机器学习的视觉里程计无需建立复杂的物体运动的几何模型，甚至无需考虑相机的校准参数以及相对尺度问题，运动估计的准确性与鲁棒性依赖于神经网络估计器的设计和用于训练的图像库是否涵盖待测场景的变化。

2015 年 Dosovitskiy 等[12] 提出了一种用于估计帧间光流的卷积神经网络 Flownet-s，取得了很好的效果。考虑到光流是相机位姿参数的最好

的体现，本文借助 Flownet-s 的结构提出了一种新的基于卷积神经网络单目视觉里程计模型，我们称之为 CNN-VO（Convolutional Neural Network Visual Odometry）。在此基础上引入 GoogLeNet[13] 的 Inception（细胞）结构对模型进行改进，增加了模型的深度。由于 Inception 结构中的多个 1×1 卷积的串联，使得图片在相同的运算量下可以获取更多的细微特征，弥补了单一卷积核特征提取不全面的缺陷。改进后的模型称为 Deep-CNN-VO（Deep Convolutional Neural Network Visual Odometry）。在预处理阶段建立两张图片的相对关系，每次输入连续的两帧图片，得到后一帧相对于前一帧的 6 个相对位姿参数，将两种模型进行对比实验，结果表明 Deep-CNN-VO 相较于 CNN-VO 在性能上具有明显提升，和现有的视觉里程计系统相比，本文所提方法也不逊色。

1　网络架构及方法

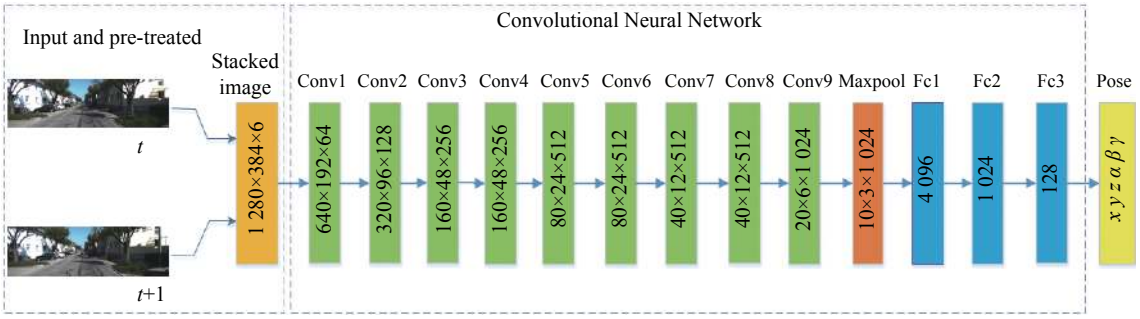
网络结构如图 1 所示。图 1（a）是 CNN-VO 网络架构图，图 1（b）是改进后 Deep-CNN-VO 网络架构图。它们都包括图片输入和预处理模块，卷积神经网络模块和位姿输出模块。其工作原理为：每次将连续的两帧图片进行叠加处理后输入到卷积神经网络模块提取图片的全局帧间变化特征，将高维的特征图输入到三层全连接层降低特征维度，最后输出车辆相对于前一帧的平移和旋转坐标。

1.1　图片输入和预处理模块

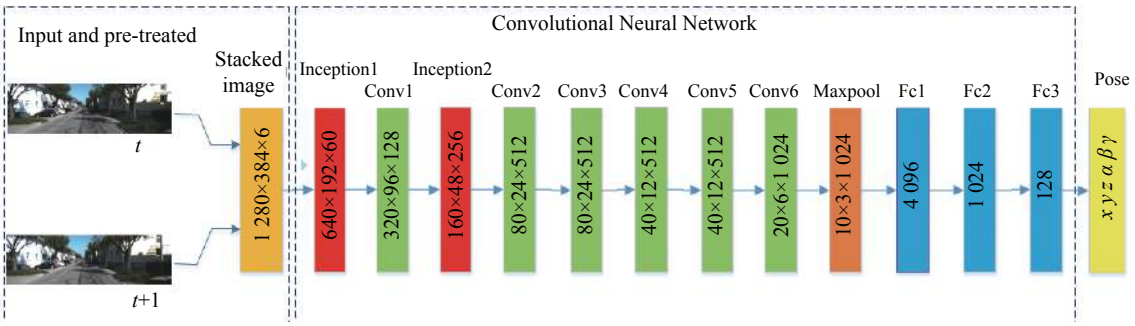
输入图像的分辨率为 1 280×384。输入时每次输入连续两帧图片，叠加后产生一个 6 通道组成的（土黄色框）图像输入到卷积神经网络。

1.2　卷积神经网络模块

卷积神经网络模块由卷积层，池化层和全连接层组成。CNN-VO 网络的卷积层借鉴了 Dosovitskiy 等[12] 提出的 Flownet-s 结构，该网络通过卷积网络学习图片特征预测光流。如图 1（a）所示，CNN-VO 卷积层由 9 个独立的子卷积层（绿色矩形）组成，其参数如表 1 所示。每个子



(a) CNN-VO 网络架构图



(b) Deep-CNN-VO 网络架构图

图 1　视觉里程计端到端网络数据流图

Fig. 1　Flow diagram of visual odometry end-to-end network

卷积层后面引入一个非线性激活函数即 ReLU（Rectified Linear Unit）[14]，其数学表达为:

$$ReLU(x)=\begin{cases}0, & x<0\\ x, & x\geqslant 0\end{cases} \tag{1}$$

该函数能够有效地进行梯度下降以及反向传播，避免了梯度爆炸和梯度消失问题。子卷积核的大小由最初的 7×7 逐步减小为 5×5 再到 3×3。图 1(b)中，Deep-CNN-VO 网络将 CNN-VO 网络中的三层子卷积更改为 Inception 结构(红色矩形)，子卷积层数由 9 层降低为 6 层。保留下来的子卷积层结构的卷积单元数目及核心数与 CNN-VO 网络一致。

其中 CNN-VO 的卷积层参数如表 1 所示。Deep-CNN-VO 剩余的 6 个卷积层的参数如表 2 所示。由于 CNN-VO 中的子卷积层中的 5 层 3×3 结构对图片小特征捕捉较好，因此选取原结构中的 7×7 及 5×5 卷积核进行改善。Conv1 改为 Inception1, Conv3 和 Conv4 改为 Inception2。

Deep-CNN-VO 中的两层 Inception 结构借鉴了 GoogLeNet[13] 中的 Inception 结构并进行调整，其结构如图 2 所示。图 2(a)是 Inception1 的结构，图 2(b)是 Inception2 的结构。相较于 CNN-VO，图 2(a)在原有 7×7 卷积层上引入多个 1×1 的卷积，降低了计算的复杂度，而且在相同的感知野中能够获得更多的图片特征。在 7×7 卷积作用的同时加入 5×5 卷积获取图片特征，这样可以在一个 Inception 层中获取多种图片特征，解决了仅用单一卷积核获取图片特征不足的情况。两种改进型结构都在原有的卷积层前串联了多个 1×1 卷积层，可以得到图片同样位置更多的非线性特征。最后再将 4 种卷积后得到的特征合并送入到下一个卷积层进行处理。该模块通过多个卷积核的串联叠加，在保持原有计算量的情况下获取了更多的细微特征，对提高视觉里程计的精度有帮助。这个改进实际上是增加了子卷积层的深度。

在卷积层的最后一层后使用池化层(橙色矩形)进行降维处理，在获得较好图片特征的同时降低了数据计算量。经过 6 个卷积层，2 个细胞层和池化层处理后，图像从最初的三维特征变为 1 024 维特征，引入全连接层(蓝色矩形)进行高维特征的降维处理。共包含 3 个全连接层，与卷积层一样，每个全连接层后有 1 个非线性激活函数。在降维到 128 层以后输出位姿，得到帧间相对位姿的 6 个参数。

1.3　位姿输出模块

经过三层全连接层后得到车辆的 6 个帧间相对位姿，分别是平移 $T=[x,y,z]$ 和旋转 $\emptyset=[\alpha,\beta,\gamma]$，在已知第一帧位姿的情况下，通过两帧之间的相对位姿逆变换得到每一帧的绝对位姿。

1.4　损失函数设计

位姿包含平移量和旋转量两种不同的尺度变

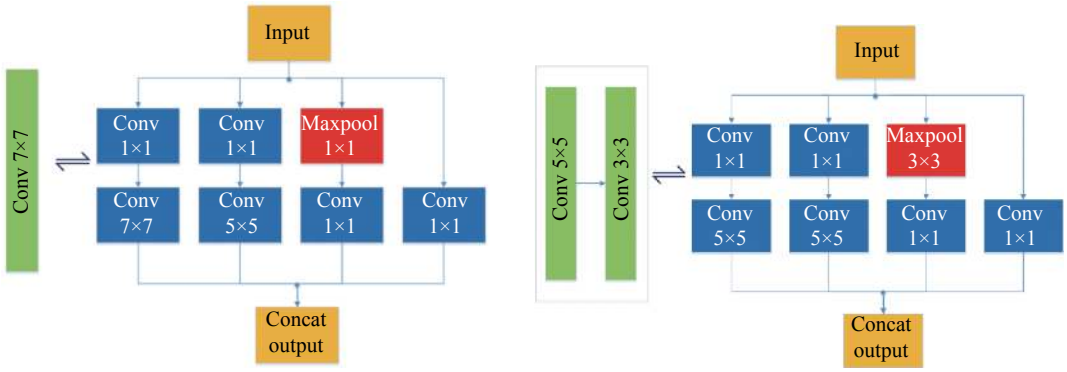
表 1　CNN-VO 子卷积参数
Tab. 1　CNN-VO subconvolution parameter

层数	卷积核大小	步长	通道数
卷积层1	7×7	2	64
卷积层2	5×5	2	128
卷积层3	5×5	2	256
卷积层4	3×3	1	256
卷积层5	3×3	2	512
卷积层6	3×3	1	512
卷积层7	3×3	2	512
卷积层8	3×3	1	512
卷积层9	3×3	2	1 024

表 2　Deep-CNN-VO 子卷积参数

Tab. 2　Deep-CNN-VO subconvolution parameter

层数	卷积核大小	步长	通道数
卷积层1	5×5	2	128
卷积层2	3×3	2	512
卷积层3	3×3	1	512
卷积层4	3×3	2	512
卷积层5	3×3	1	512
卷积层6	3×3	2	1 024



(a) 7×7 卷积转化为 Inception 结构图　　(b) 3×3 及 5×5 卷积转化为 Inception 结构图

图 2　细胞结构图

Fig. 2　Inception structure

量，所以本文的损失函数为两者的加权求和。设第 i 至 i+1 帧图片的相对平移为 T_i，相对旋转为 $\emptyset_i$，使用均方误差作为损失函数：

$$\text{loss} = \text{argmax}\frac{1}{N}\sum_{i=1}^{N}\left\| T_i - T_g\right\|_2^2 + \delta\left\| \emptyset_i - \emptyset_g\right\|_2^2 \tag{2}$$

式中：T_g 和 $\emptyset_g$ 分别表示平移量和旋转量的真实位姿；δ 是调节平衡旋转量和平移量的一个参数，实验中使用 10，50，100 分别进行实验，发现当 δ 取 50 时可以获得最好效果，$\|\ldots\|_2$ 表示 2 范数。

2　实验结果

2.1　实验平台及训练设置

实验采用 KITTI[15] 公共数据集中的 Visual Odometry/SLAM 视频图像进行实验。它提供了从公路、农村和城市场景中的 22 个经过校正的双目图像序列，每个序列的范围从 500～5 000 m 长度不等，帧速率为 10 帧/s，图像分辨率为 1 241×376 及 1 226×370。其中前 11 个序列提供了从激光雷达和 GPS 获得的各帧位姿参数的地面真实值。采用 00,01,02,08,09 序列作为训练集，这些场景相对于其他序列图片数量较多，行驶距离较长且场景更丰富。使用 03,04,05,07,10 序列作为测试。

系统工作站采用 NVDIA GTX 1080Ti GPU 进行训练，配备 32G 内存以及 Intel Core Xeon 3.4 GHz CPU, 测试用的笔记本配备 NVDIA GTX 1 060 以及 Intel Core i7 2.7 GHz CPU。

网络采用连续两帧图片的叠加作为网络的输入，求解的是连续两帧间的相对位姿态，因此需要使用相邻两帧的相对位姿作为真值进行训练。然而，KITTI 提供的每帧的位姿参数是相对于起始点的绝对位姿，因此需要进行如下转换。KITTI 提供的第 n 帧图片的绝对位姿为齐次矩阵为 T_n，第 n+1 帧图片为齐次矩阵为 T_{n+1}，则两

帧间的相对位姿矩阵为：

$$\boldsymbol{T}_{rn}=\boldsymbol{T}_n^{-1}\boldsymbol{T}_{n+1}=\begin{bmatrix}R_{n1}&R_{n2}&R_{n3}&x_n\\R_{n4}&R_{n5}&R_{n6}&y_n\\R_{n7}&R_{n8}&R_{n9}&z_n\\0&0&0&1\end{bmatrix}^{-1}\times\begin{bmatrix}R_{(n+1,1)}&R_{(n+1,2)}&R_{(n+1,3)}&x_{(n+1)}\\R_{(n+1,4)}&R_{(n+1,5)}&R_{(n+1,6)}&y_{(n+1)}\\R_{(n+1,7)}&R_{(n+1,8)}&R_{(n+1,9)}&z_{(n+1)}\\0&0&0&1\end{bmatrix}=\begin{bmatrix}R_{rn1}&R_{rn2}&R_{rn3}&x_{rn}\\R_{rn4}&R_{rn5}&R_{rn6}&y_{rn}\\R_{rn7}&R_{rn8}&R_{rn9}&z_{rn}\\0&0&0&1\end{bmatrix}\tag{3}$$

式中最后一列是平移参数，$T=[x,y,z]=[x_{rn},y_{rn},z_{rn}]$，矩阵$\begin{bmatrix}R_{rn1}&R_{rn2}&R_{rn3}\\R_{rn4}&R_{rn5}&R_{rn6}\\R_{rn7}&R_{rn8}&R_{rn9}\end{bmatrix}$代表旋转欧拉角$R=[\alpha,\beta,\gamma]$。其转换关系为：$\alpha=\text{atan2}\left(-R_{rn7},\sqrt{R_{rn1}^2+R_{rn4}^2}\right)$，$\beta=\text{atan2}(R_{rn4},R_{rn1})$，$\gamma=\text{atan2}(R_{rn8},R_{rn9})$。

网络使用批量梯度下降法作为优化算法，选择 Adam（Adaptive Moment Estimation）作为优化器，进行 150 000 次迭代，其中 Adam 优化器参数设置 $\beta=0.9$，$\beta_2=0.999$，学习率初始设置为 0.000 1，学习率将随着迭代次数的增加适当变小，以便找到优化函数的最优解。batch size 设为 32，每经过一轮 batch size，训练样本将会打乱，以保证误差不会产生突变，误差曲线稳定下降。

训练过程中引入验证机制，在每经过 4 轮训练以后，从验证数据集中随机选取一张图片，对当前训练产生的模型进行测试，若误差持续下降，则证明训练模型结构有效。验证机制的引入可以有效防止训练过程中的过拟合情况。

2.2　运动估计结果及与其他方法的比较

使用 KITTI 数据集中的图像序列对本文两种方法（Deep-CNN-VO 和 CNN-VO）进行测试评估，并与 VISO2-M[4] 以及 DeepVO[10] 两种单目视觉里程计方法进行比较，其中 VISO2-M 是传统单目视觉里程计方法中比较有代表性的算法，也是开源的。DeepVO 采用了长短期记忆网络进行轨迹估计，是目前使用深度学习进行端到端单目视觉里程计研究定位较为精确的算法。完整序列的运动估计误差分别用平均平移误差（ATE）和平均旋转误差（ARE）表示[15]，误差越小代表与真实轨迹越相近。表 3 展示了用 4 种方法对 KITTI 数据库中 03、04、05、07、10图像序列的平均平移误差和平均旋转误差。

从表中可以看出，Deep-CNN -VO 模型相较于 CNN-VO 模型在 5 个测试序列上，无论是 ATE 还是 ARE 都有显著提升，说明卷积神经网络中 Inception 结构起作用。与传统的视觉里程计方法 VISO2-M 相比，除了序列 04，Deep-CNN-VO 在其他 4 个序列，以及总体上 ATE 和 ARE 都有提升。与目前比较著名的基于深度学习的视觉里程计 DeepVO 相比，ATE 的表现差一些，ARE 基本持平，主要原因是 DeepVO 引入了包含时序，特性的长短时记忆网络（LSTM）进行优化而本文方法则完全采用相对简单的卷积神经网络。在计算效率方面，本文方法计算量较小，对硬件要求相对较低，更容易移植到实时的系统中。

利用视觉里程计估计得到的每一帧相对位姿参数推算车载相机的运动轨迹。图 3 展示了以上 4 种算法对 KITTI 数据库中 04,05,07,10 序列的轨迹构建以及和真实轨迹的比较。从图中可以看出，序列 04 的直线轨迹表现相对较好，其余序列中由于车辆行驶过程中会有转弯等因素的影响，轨迹出现偏移。结果表明，VISO2-M 算法和 CNN-VO 模型产生的轨迹精度偏离真值较大，DeepVO 及本文 Deep-CNN-VO 算法相对比较接近真实轨迹。

3　结　论

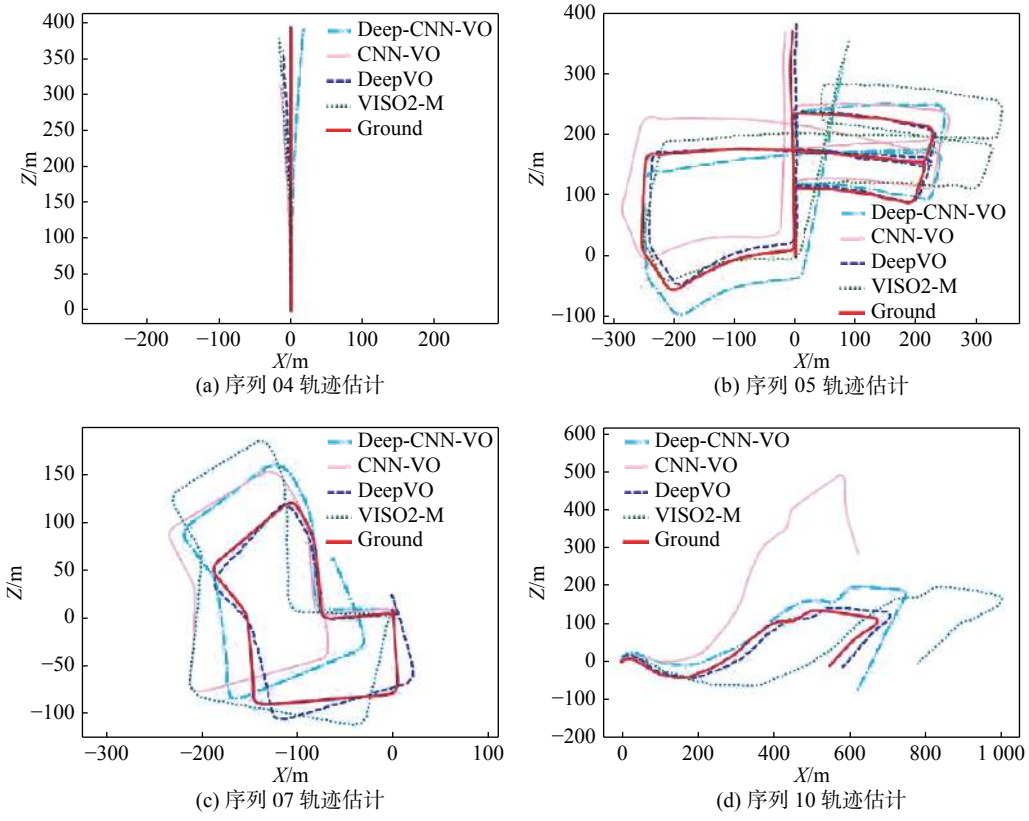
本文尝试将用于分类的卷积神经网络转化获取图像帧间时序变化特征的网络，实现了采用基于深度学习卷积神经网络技术的车辆自主定位方法。相较于传统的视觉里程计方法，本文方法采用端到端的方式，无需根据场景建立复杂的几何模型，是未来视觉里程计技术的一个发展方向。本文工作证明了其可行性，是一个有益的尝试和探索。

本文方法借助卷积神经网络对彩色图片自动学习提取图像帧间变化的全局特征，采用改进型 Inception 结构代替单一的卷积层，多个 1×1 卷积核串联使用可以在相同运算量的基础上提取出更多细微的图片特征，通过三层全连接层

表 3　测试序列 03、04、05、07、10 实验结果对比

Tab. 3　The comparison experimental results of 03, 04, 05, 07 and 10 test sequence

序列	Deep-CNN-VO		CNN-VO		VISO2-M[4]		DeepVO[10]	
	ATE/%	ARE/((°)/m)	ATE/%	ARE/((°)/m)	ATE/%	ARE/((°)/m)	ATE/%	ARE/((°)/m)
03	8.79	0.046 7	15.53	0.065 2	8.47	0.088 2	8.49	0.068 9
04	11.87	0.068 2	18.05	0.025 8	4.69	0.044 9	7.19	0.069 7
05	6.98	0.037 1	7.92	0.052 9	19.22	0.035 4	2.62	0.036 1
07	9.81	0.093 3	13.94	0.083 1	23.61	0.041 1	3.91	0.046 0
10	17.75	0.045 8	21.45	0.105 9	41.56	0.329 9	8.11	0.088 3



(a) 序列 04 轨迹估计　(b) 序列 05 轨迹估计

(c) 序列 07 轨迹估计　(d) 序列 10 轨迹估计

图 3　序列 04、05、07、10 轨迹估计

Fig. 3　The trajectory estimation of 04,05,07 and 10 sequence

输出相机的帧间相对位姿参数，并据此推算相机的运动轨迹。在共用 KITTI 数据集上进行测试，实验结果表明本文提出的 Deep-CNN-VO 模型可以较准确地估计车辆的运动轨迹，与目前比较著名的基于深度学习的视觉里程计 DeepVO 相比，虽然 ATE 效果差一点，其主要原因是 DeepVO 引入了包含时序特性的长短时记忆网络（LSTM）进行优化，而本文方法则完全采用相对简单的卷积神经网络。由于仅使用 CNN 网路作为结构，神经单元相对于 DeepVO 要小，因此本文方法计算量较小，对硬件要求相对较低，更容易移植到实时的系统中。下一步的工作将考虑结合残差方法及双目图片共同优化网络，达到更好的运动估计精度。

参考文献：

［1］ 慈文彦, 黄影平, 胡兴. 视觉里程计算法研究综述 [J].

计算机应用研究, 2019, 36(9): 2561 – 2567.

[2] NISTER D, NARODITSKY O, BERGEN J. Visual odometry[C]//Proceedings of 2004 IEEE computer society conference on computer vision and pattern recognition. Washington, DC, USA: IEEE, 2004: I − I.

[3] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: an efficient alternative to SIFT or SURF[C]//Proceedings of 2011 international conference on computer vision. Barcelona, Spain: IEEE, 2011: 2564 − 2571.

[4] GEIGER A, ZIEGLER J, STILLER C. StereoScan: dense 3d reconstruction in real − time[C]//Proceedings of 2011 IEEE intelligent vehicles symposium (IV). Baden-Baden, Germany: IEEE, 2011: 963 − 968.

[5] MUR-ARTAL R, MONTIEL J M M, TARDÓS J D. ORB-SLAM: a versatile and accurate monocular SLAM system[J]. IEEE Transactions on Robotics, 2015, 31(5): 1147 – 1163.

[6] 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.

[7] ROBERTS R, NGUYEN H, KRISHNAMURTHI N, et al. Memory − based learning for visual odometry[C]// Proceedings of 2008 IEEE international conference on robotics and automation. Pasadena, CA, USA: IEEE, 2008: 47 − 52.

[8] KONDA K, MEMISEVIC R. Learning visual odometry with a convolutional network[C]//Proceedings of the 10th international conference on computer vision theory and applications − volume 2. Berlin, Germany, 2015: 486 − 490.

[9] KENDALL A, GRIMES M, CIPOLLA R. PoseNet: a convolutional network for real − time 6 − DOF camera relocalization[C]//Proceedings of 2015 IEEE international conference on computer vision. Santiago, Chile: IEEE, 2015: 2938 − 2946.

[10] WANG S, CLARK R, WEN H K, et al. DeepVO: towards end-to-end visual odometry with deep recurrent convolutional neural networks[C]//Proceedings of 2017 IEEE international conference on robotics and automation. Singapore: IEEE, 2017: 2043 − 2050.

[11] MCCORMAC J, CLARK R, BLOESCH M, et al. Fusion++: volumetric object-level SLAM[C]//Proceedings of 2018 IEEE international conference on 3D vision. Verona, Italy: IEEE, 2018: 32 − 41.

[12] DOSOVITSKIY A, FISCHER P, ILG E, et al. FlowNet: learning optical flow with convolutional networks[C]//Proceedings of 2015 IEEE international conference on computer vision. Santiago, Chile: IEEE, 2015: 2758 − 2766.

[13] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//Proceedings of 2015 IEEE conference on computer vision and pattern recognition (CVPR). Boston, MA, USA: IEEE, 2015: 1− 9.

[14] XU B, WANG N Y, CHEN T Q, et al. Empirical evaluation of rectified activations in convolutional network[J]. arXiv: 1505.00853, 2015.

[15] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//Proceedings of 2012 IEEE conference on computer vision and pattern recognition (CVPR). Providence, RI, USA: IEEE, 2012: 3354− 3361.

（编辑：张　磊）