# 大规模复杂场景下基于 ResNet 的回环检测技术研究

王红君[1] 郝金龙[1] 赵 辉[1,2] 岳有军[1]

[1](天津市复杂系统控制理论及应用重点实验室 天津 300384)

[2](天津农学院 天津 300384)

**摘 要** 回环检测(LCD)是同步定位与地图构建(SLAM)中的重要环节,对 SLAM 的精度和鲁棒性具有显著影响。由于大规模复杂场景下光照、摄像机视角、存在移动物体、气候、地貌特征等条件的大幅变化,使得回环检测的精度和鲁棒性受限。为解决此问题,提出一种基于深度残差网络(Deep Residual Network,ResNet),利用信息熵(Information Entropy)改进的局部聚合描述符向量(Vector of Locally Aggregated Descriptors,VLAD)的回环检测方法 RIV-LCD。采用弱监督迁移训练算法训练 ResNet 来提取图像特征;使用信息熵加权的 VLAD 对图像特征进行处理;通过词袋法进行匹配,得到匹配结果。在 Nordlandsbanen 数据集上进行的验证和对比实验表明:在大规模复杂场景中剧烈环境条件变化下,RIV-LCD 具有良好的精度和鲁棒性。

**关键词** 回环检测 同步定位与地图构建 深度残差网络 信息熵 局部聚合描述符向量 词袋法

**中图分类号** TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.07.020

# RESNET-BASED LOOP CLOSURE DETECTION TECHNOLOGY IN LARGE-SCALE COMPLEX SCENE

Wang Hongjun[1] Hao Jinlong[1] Zhao Hui[1,2] Yue Youjun[1]

[1](*Tianjin Key Laboratory of Complex System Control Theory and Applications*, *Tianjin 300384*, *China*)

[2](*Tianjin Agricultural College*, *Tianjin 300384*, *China*)

**Abstract** Loop closure detection(LCD) is an important part of simultaneous localization and mapping(SLAM), which has a significant impact on the accuracy and robustness of SLAM. Due to the large changes of illumination, camera angle, moving objects, climate and geomorphic features and other conditions in large-scale complex scenes, the accuracy and robustness of LCD are limited. In order to solve this problem, this paper proposes a loop closure detection method RIV-LCD based on deep residual network(ResNet) and vector of locally aggregated descriptors(VLAD) improved by information entropy. Weak supervised migration training was used to train ResNet to extract the features of the image; we used the information entropy weighted VLAD to process the image features; the matching result was obtained by the bag-of-words matching. Verification and comparison experiments on the Nordlandsbanen data set show that the RIV-LCD has good accuracy and robustness in the large-scale complex scene with dramatic changes in environmental conditions.

**Keywords** LCD SLAM Deep residual network Information entropy VLAD BoW

## 0 引 言

近年来,同时定位与地图构建技术(Simultaneous Localization and Mapping,SLAM)在多个领域被推广应用,但仍有一些问题亟待解决。在大规模复杂场景下,SLAM 系统面对复杂环境变化时精度和鲁棒性差是需要解决的主要问题,集中表现在关键帧提取困难、回环检测过程中回环位置难以确定、跟踪性能差等。其中,回环检测是移动机器人抑制累计误差的关键,通过对

收稿日期:2019-06-05。天津市重点研发计划科技支撑重点项目(18YFZCNC01120)。**王红君**,教授,主研领域:智能控制理论及应用,农业信息化与精准农业智能监控理论与技术。**郝金龙**,硕士生。**赵辉**,教授。**岳有军**,教授。

同一位置的重识别,SLAM 系统可以对姿态和全局地图进行优化,提高系统的精度和稳定性。

在大规模复杂场景下,往往存在着摄像机视角大幅改变、地貌特征改变、大量移动物体、光照剧烈改变和天气季节改变等[1]环境条件的改变,限制了回环检测算法的使用。2013 年 Milford 等提出的 SeqSLAM 是第一个大规模复杂场景下有一定效果的视觉定位系统。SeqSLAM 通过当前图像序列来匹配已有最相近的图像序列,它关注的是图像序列的整体特征而不是单个图像的特征,其对季节变化拥有良好的应对能力。但 SeqSLAM 的准确性与图像序列的拍摄角度的一致性紧密相关[2],采用图像序列匹配也会使尾部图像序列无效[1],采用暴力匹配方式进行图像匹配使得计算成本随着场景规模激增。

2017 年 Siam 等[3]提出的 Fast-SeqSLAM 是一种高效的 SeqSLAM 版本。Fast-SeqSLAM 的核心是通过近似最近邻算法代替了 SeqSLAM 中暴力匹配的方式,从而在不降低精度的情况下降低了时间复杂度。

随着 CNN 的发展,AlexNet[4]、VGG[5]、GoogLeNet[6]和 ResNet[7]等被用来进行图像特征的提取,解决了对象分类、场景识别和物体检测等识别问题。2017 年国防科技大学的 Bai 等[8]提出了一种融合 CNN 与 SeqSLAM 的回环检测算法 SeqCNNSLAM。该算法使用先前训练好的 Places-CNN[9]的第 3 卷积层或第 5 池化层来进行图像特征的提取,再通过 SeqSLAM 来进行图像的序列。在 Nordland 和 Gardens point 等数据集上验证了 SeqCNNSLAM 不但有 SeqSLAM 应对环境季节变化的能力,还对视角变化具有鲁棒性。2018 年徐建鹏等[10]提出基于 Faster-RCNN 神经网络回环检测的优化算法,该算法使用 Faster-RCNN 神经网络对图像序列进行检测,将获得的图像语义特征、像素位置及特征图等构建成二维语义特征向量图,根据二维语义特征向量图之间的相似度匹配得到初始回环,再经位姿验证获得最终回环结果。徐建鹏的基于 Faster-RCNN 神经网络回环检测的优化算法和 SeqCNNSLAM 的成功表明,将 CNN 与回环检测算法融合能够改善回环检测算法的精度和鲁棒性。

不同于 SeqSLAM 使用图像像素值当作图像特征,也不同于 SeqCNNSLAM 延续 SeqSLAM 使用图像序列进行匹配,本文采用 ResNet 对关键帧进行特征提取,使用词袋法进行单幅图片的特征匹配,采用弱监督的迁移训练方法来训练 ResNet,提出一种基于深度残差网络和利用信息熵改进的局部聚合描述符向量的回环检测方法 RIV-LCD。

# 1 相关理论

## 1.1 深度残差网络

深度残差网络(Deep Residual Network,ResNet)由 He 等[7]提出,解决了随着 CNN 网络层数加深,准确率下降的问题。ResNet 在 ILSVRC 和 COCO 2015 上取得了五项第一,优于其他各种 CNN 模型在 ImageNet 数据集上的表现,TOP5 误差仅为 3.57%[7]。ResNet 由若干个 building block 或 bottleneck 组成,其结构如图 1 所示。不同数量的 building block 或 bottleneck 组成了不同深度的 ResNet。本文使用 50 层的 ResNet 来对图像进行特征提取。

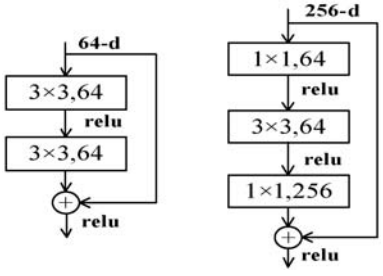


图 1 building block 与 bottleneck 的结构图

## 1.2 局部聚合描述符向量

局部聚合描述符向量(VLAD)[12]通过计算图像特征描述子与其所属的聚类中心的差矢量来聚合图像特征。

如果给定 $N$ 个 $D$ 维的本地特征描述子 $\{X_i\}$ 作为输入,$K$ 个聚类中心 $\{C_k\}$ 作为 VLAD 的元素,VLAD 的输出是一个 $D\times K$ 维的矩阵 $\boldsymbol{V}$。位置元素 $\boldsymbol{V}(j,k)$ 的计算公式如下:

$$\boldsymbol{V}(j,k) = \sum_{i=1}^{N}(x_i(j) - C_k(j)) \tag{1}$$

式中:$x_i(j)$ 和 $C_k(j)$ 分别是第 $i$ 个本地特征描述子和第 $k$ 个聚类中心的第 $j$ 维元素。

## 1.3 词袋法

词袋法(Bag-of-Words,BoW)[13]最早出现在自然语言处理和信息检索领域。该模型忽略文本的语法和语序等要素,将其仅仅看作是若干个词汇的集合。BoW 使用一组无序的 words 来表达一段文字或一个文档。近年来,BoW 模型也被广泛应用于图像检索。

# 2 ResNet 的弱监督迁移训练

## 2.1 预训练

对于训练一个已知结构的卷积神经网络,最核心的问题是数据集的获取和损失函数的确定。由于回环

检测数据集规模偏小,而 ResNet 的层数深权值参数多,大规模重复训练时,容易出现参数过拟合问题。采用关联数据集进行预训练十分重要,可以有效规避过拟合问题。在第一阶段,采用关联性强的场景识别大型数据集 Places2[14] 进行预训练;在 Place2 中,选取适合应用环境的图片对 ResNet 进行场景识别训练,使得 ResNet 可以获得提取特定环境图像特征的能力。对于第一阶段场景识别的训练,可以使用 Softmax 分类器和交叉熵损失函数。

$$L_1 = \sum_{i=1}^{N} y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \quad (2)$$

式中:$y$ 为标签值,$\hat{y}$为预测值,$N$ 为总的类别数。Softmax 分类器的作用是输出分类结果的概率分布,可以做到多类别输出,而交叉熵损失函数可以刻画两个概率分布之间的差异。两者配合,可以计算出 CNN 预测的概率分布与实际标签概率分布之间的差异。

## 2.2　弱监督迁移训练

第二阶段弱监督迁移训练需要使用回环检测数据集,Nordlandsbanen 数据集[15] 比较适用。Nordlandsbanen 数据集记录了特隆赫姆和博德之间 729 km 的铁路,在四个不同季节的同一条铁路线路一共拍摄了四次。如图 2 所示,图中依次是春夏秋冬四个季节在同一位置拍摄的图片。由于季节不同,四次拍摄拥有不同的光照(白天和夜晚)、地貌特征(植被雨雪覆盖等)和运动物体(乘客列车等)。可以有针对性地应用于训练 CNN,使其在不同光照、气候和地表外貌条件下,获得对同一地点图像的共同特征提取能力。

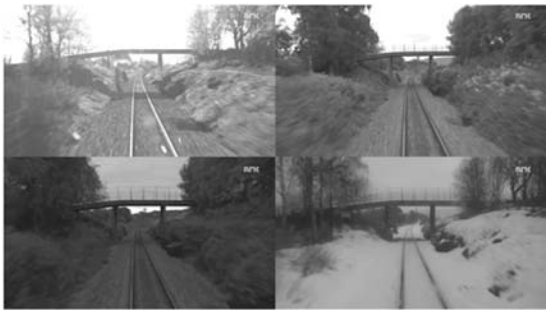

图 2　Nordlandsbanen 数据集

由于单输入网络不能直接输入两幅图片进行比较,无法进行图像匹配任务训练,所以第二阶段采用弱监督的迁移训练方式进行训练。其中,选取损失函数的关键是要体现出图片之间的差异,用来监督训练,这种关联变量因数据集的特征而定,可以是 GPS 坐标,也可以是图像在序列中的位置。在 Nordlandsbanen 数据集上,用于弱监督训练更合适的变量是图像在序列中的位置,因为该数据集四个视频流的每一帧都经过对齐,可以使用 Triplet 损失函数[16] 进行弱监督迁移训练。Triplet 损失函数如下:

$$L_\theta = \sum_j l(\min_i d_\theta^2(q, p_i^q) + m - d_\theta^2(q, n_j^q)) \quad (3)$$

式中:对于给定的训练图像 $q$,把数据集中地理距离相近的图像记作 $p_i^q$,即潜在阳性(Potential Positives);而对于数据集中地理相距较远的图像记作 $n_j^q$,即确定的阴性(Definite Negatives);$d_\theta^2(x,y)$为图像拍摄地点间的距离,可以用序列间距来表示;$l(X)$为铰链损失函数。

## 2.3　训练流程

两个阶段的训练流程如图 3 所示。

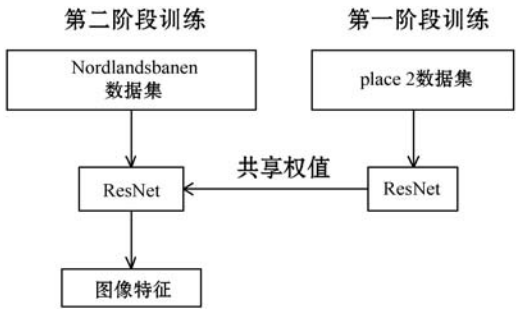


图 3　训练流程示意图

# 3　信息熵加权 VLAD

由于属于每一个聚类中心的本地特征描述子所包含的信息量不同,可以设置一个权重参数 $a_k(x_i)$ 当作 $(x_i(j) - C_k(j))$ 的权值[17] 来描述每一个类本地特征描述子间的关系:

$$\boldsymbol{V}(j,k) = \sum_{i=1}^{N} a_k(x_i)(x_i(j) - C_k(j)) \quad (4)$$

可以用信息熵度量属于每一个聚类中心的本地特征描述子所包含的信息量。根据香农给出的信息熵定义公式[18],对于任意一个随机变量 $X$,其信息熵定义如下(单位为比特(bit)):

$$H(X) = -\sum_{x \in X} P(x) \log P(x) \quad (5)$$

仿照式(5),本地特征描述子的信息熵定义如下:

$$Entropy(X) = \sum_{k=1}^{c} -p_k \log(p_k) \quad (6)$$

式中:$c$ 为本地特征描述子聚类中心个数;$p_k$为第 $k$ 类本地特征描述子在所有本地特征描述子中所占的比例,即第 $k$ 类的先验概率。该信息熵反映了集合 $X$ 中的本地特征描述子平衡分布的期望,也可以度量 $X$ 中包含信息量的大小。

将特征矩阵 $\boldsymbol{V}$ 信息熵的值 $Entropy(X)$ 赋给 $a_k(x_i)$,可以得到:

$$\boldsymbol{V}(j,k) = \sum_{i=1}^{N} Entropy(X)(x_i(j) - C_k(j)) \quad (7)$$

ResNet 提取图像本地特征和信息熵加权 VLAD，产生本地特征描述子的流程示意图如图 4 所示。

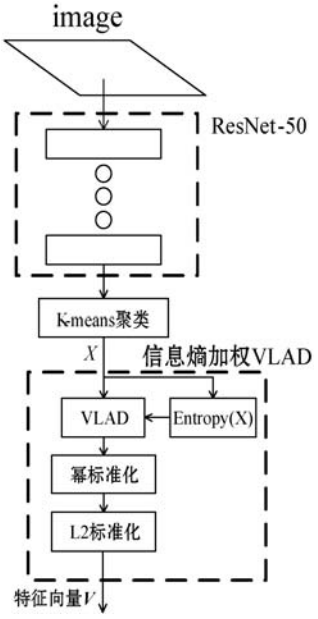


图 4 ResNet 与信息熵加权 VLAD 提取特征流程示意图

## 4.1 基于 BoW 的特征向量匹配

使用 BoW 时，要通过 K-means 聚类算法聚类出本地特征描述子的 K 聚类中心，得到一个字典。匹配时，先将离线数据库图片以及在线数据库图片的本地特征描述子投射到字典的空间中，得到图片本地特征描述子相对应的 words 分布；然后再以同样的方法，得到待匹配目标图片的本地特征描述子相对应的 words 分布；最后通过比较这些 words 分布的相似性，得到 TOP1 候选结果。所使用的 BoW 的算法流程如图 5 所示。

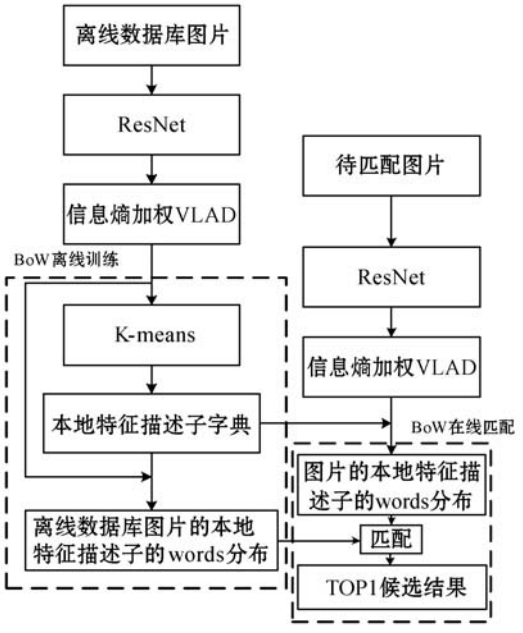


图 5 BoW 的离线训练与在线匹配流程示意图

# 5 实验与分析

## 5.1 实验平台与测试集

实验在一台图像处理服务器上进行，该服务器配备为 64 GB 运行内存、48 个 2.20 GHz 的 Intel Xeon CPU 以及 2 张 12 GB 显存的 GeForce GTX 1080Ti 显卡。在该实验平台上搭建深度学习环境 Anaconda3 以及 Tensorflow 进行实验。在 Nordlandsbanen 数据集上等间距按顺序在每个季节图像序列中抽取出 33 626 幅图片，四个季节共 134 504 幅图片作为测试集。

## 5.2 可行性验证实验

在 Nordlandsbanen 数据集的测试集上进行测试，通过春天图片匹配夏天相同拍摄位置的图片，得到可行性验证准确率召回率曲线，如图 6 所示。其中："ResNet + 信息熵加权 VLAD + BoW"为 RIV-LCD 算法得到的准确率召回率曲线；"ResNet + 无权值 VLAD + BoW"为使用没有权值原始的 VLAD 得到的准确率召回率曲线；"ResNet + BoW"为没有使用 VLAD 得到的准确率召回率曲线。

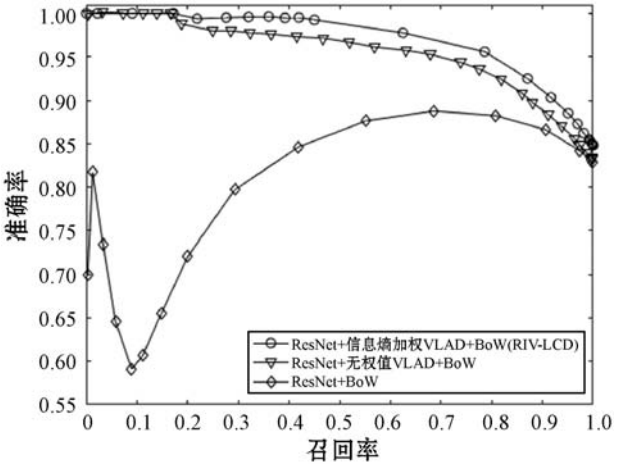


图 6 可行性验证 precision-recall 曲线

可以看出，使用 VLAD 处理后的图像特征匹配时，可以明显提高召回图像的匹配准确率，经过信息熵加权 VLAD 处理后的图像特征匹配时，有更高的准确率。

在 Nordlandsbanen 数据集的测试集上进行测试，通过春天的图片匹配其他季节相同拍摄位置的图片，结果如图 7 所示。可以看出在光照、气候、地表外貌大幅变化时，RIV-LCD 仍可以完成回环检测，证明该算法对环境条件剧烈变化具有良好的鲁棒性。而错误匹配主要存在于两种情况：地貌特征十分相似和隧道内弱光照的情况，分别如图 8、图 9 所示。

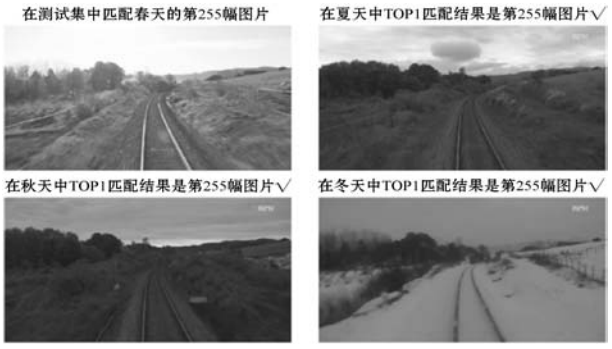


图 7　RIV-LCD 准确匹配到的图像

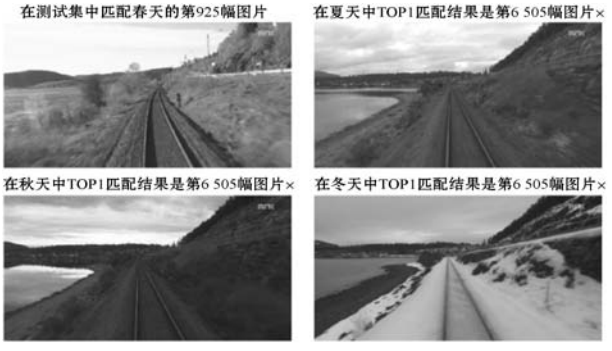


图 8　地貌相似时 RIV-LCD 匹配到的错误图像

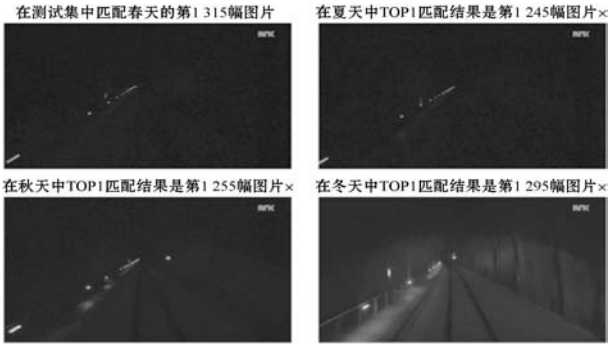


图 9　隧道内弱光照时 RIV-LCD 匹配到的错误图像

### 5.3　对比实验

在 Nordlandsbanen 数据集上进行测试,使用秋天的图像来匹配夏天的图像,测试结果如图 10 所示。“RIV-LCD”为 RIV-LCD 算法得到的准确率召回率曲线,同样“SeqSLAM”、“Fast-SeqSlam”和“CNN-SeqSlam”分别为 SeqSLAM、Fast-SeqSlam 和 CNN-SeqSlam 算法得到的准确率召回率曲线。

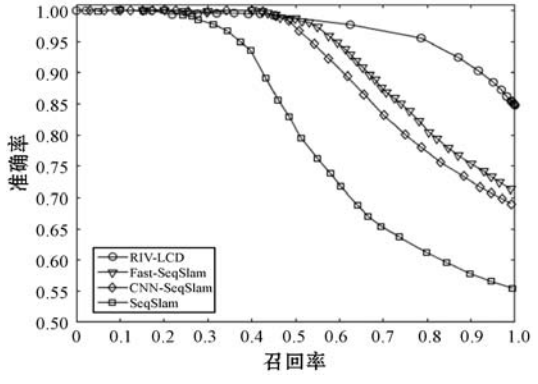


图 10　在 Nordlandsbanen 数据集上得到的 precision-recall 曲线

对比图 10 中四种回环检测算法准确率召回率曲线可以看出,随着召回率的升高,RIV-LCD 的准确率下降最缓慢,在获得最大召回率时准确率为 84.8%。相比之下,SeqSLAM 的准确率下降最快,在获得最大召回率时准确率为 55.5%。同样 Fast-SeqSlam 和 CNN-SeqSlam 的准确率下降比 RIV-LCD 快,在获得最大召回率时准确率分别为 69.0% 和 71.3%。这些都说明本文的 RIV-LCD 在大规模复杂场景下精确度和鲁棒性更高。

## 6　结　语

在大规模复杂场景下,为解决现有部分回环检测算法无法使用的问题,本文提出了一种新的回环检测算法 RIV-LCD。实验表明:大规模复杂场景下 RIV-LCD 在面对光照、气候、地表外貌大幅变化时,仍然可以准确地进行回环检测;在同样的大规模复杂场景下,RIV-LCD 比 SeqSLAM、Fast-SeqSlam 和 CNN-SeqSlam 拥有更高的准确率和鲁棒性。

## 参　考　文　献

[1] Milford M J, Wyeth G F. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights[C]//2012 IEEE International Conference on Robotics and Automation. IEEE, 2012.

[2] Sünderhauf N, Neubert P, Protzel P. Are we there yet? challenging SeqSLAM on a 3000 km journey across all four seasons[C]//Proceedings of the Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation(ICRA), Chemnitz, Germany, 2013.

[3] Siam S M, Zhang H. Fast-SeqSLAM: A fast appearance based place recognition algorithm[C]//2017 IEEE International Conference on Robotics and Automation(ICRA). IEEE, 2017.

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097 - 1105.

[5] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB]. arXiv:1409.1556, 2014.

[6] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2015.

[7] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2016.

(下转第 135 页)

## 参 考 文 献


[1] Liu B. Sentiment analysis and opinion mining[M]//Synthesis Lectures on Human Language Technologies. Morgan & Claypool, 2012.

[2] Shen C L, Sun C L, Wang J J, et al. Sentiment classification towards question-answering with hierarchical matching network[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018:3654 – 3663.

[3] Wang J Y, Bao B K, Xu C S. Sentiment-aware multi-modal recommendation on tourist attractions [C]//International Conference on Multimedia Modeling, 2019:3 – 16.

[4] 陈自岩,黄宇,王洋,等. 一种利用语义相似特征提升细粒度情感分析方法[J]. 计算机应用与软件,2017,34(3):33 – 36,86.

[5] 卿勇,刘梦娟,薛浩,等. OPEN:一个基于评论的商品特征抽取及情感分析框架[J]. 计算机应用与软件,2018,35(1):65 – 71.

[6] 李佳丽,封化民,潘扬,等. 基于卷积神经网络的情感分析算法[J]. 计算机应用与软件,2018,35(4):287 – 292.

[7] Jiang L, Yu M, Zhou M,et al. Target-dependent twitter sentiment classification[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011.

[8] Pérez-Rosas V. Learning sentiment lexicons in Spanish[C]//Eighth International Conference on Language Resources and Evaluation(LREC-2012). 2012.

[9] Vo D T, Zhang Y. Target-dependent twitter sentiment classification with rich automatic features[C]//Proceedings of the 24th International Conference on Artificial Intelligence. AAAI Press, 2015:1347 – 1353.

[10] Wang Y Q, Huang M L, Zhu X Y, et al. Attention-based LSTM for aspect-level sentiment classification[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016.

[11] Tang D Y, Qin B, Liu T. Aspect level sentiment classification with deep memory network [C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 214 – 224.

[12] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques[C]//Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, 2002:79 – 86.

[13] Mohammad S M, Kiritchenko S, Zhu X D. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets [EB]. arXiv preprint arXiv:1308.6242, 2013.

[14] Qian Q, Tian B, Huang M, et al. Learning tag embeddings and tag-specific composition functions in recursive neural network[C]//Meeting of the Association for Computational Linguistics & International Joint Conference on Natural Language Processing. 2015.

[15] Tang D Y, Qin B, Feng X C, et al. Effective lstms for target-dependent sentiment classification [C]//International Conference on Computational Linguistics, 2016:3298 – 3307.

[16] Ruder S, Ghaffari P, Breslin J G. A hierarchical model of reviews for aspect-based sentiment analysis[EB]. arXiv preprint arXiv:1609.02745, 2016.


(上接第 129 页)


[8] Bai D D, Wang C Q, Zhang B, et al. CNN feature boosted SeqSLAM for real-time loop closure detection[J]. Chinese Journal of Electronics, 2018, 27(3):488 – 499.

[9] Zhou B L, Lapedriza A, Xiao J X, et al. Learning deep features for scene recognition using places database[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 1, 2014:487 – 495.

[10] 徐建鹏,卜凡亮. 基于 Faster-RCNN 神经网络的回环检测优化算法[J]. 计算机应用研究,2019,36(12):3628 – 3631.

[11] Korrapati H, Mezouar Y. Multi-resolution map building and loop closure with omnidirectional images[J]. Autonomous Robots, 2016, 41(4):967 – 987.

[12] Jégou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation [C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010.

[13] Angeli A, Filliat D, Doncieux S, et al. Fast and incremental method for loop-closure detection using bags of visual words[J]. IEEE Transactions on Robotics, 2008, 24(5): 1027 – 1037.

[14] Zhou B, Khosla A, Lapedriza A, et al. Places: An image database for deep scene understanding[J]. Journal of Vision, 2017, 17(10):296.

[15] The Norwegian Broadcasting Corporation: The NordlandsbanenDataset[OL]. http://nrkbeta. no/2013/01/15/nordlandsbanen-minute-by-minute-season-by-season/.

[16] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2015:815 – 823.

[17] Arandjelović R, Gronat P, Torii A, et al. NetVLAD: CNN architecture for weakly supervised place recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018,40(6):1437 – 1451.

[18] Shannon C E. A mathematical theory of communication[J]. Bell Labs Technical Journal, 1948, 27(4):379 – 423.