控制与决策
Control and Decision
ISSN 1001-0920,CN 21-1124/TP

《控制与决策》网络首发论文

题目：一种基于视觉特征区域建议的目标检测方法
作者：李会军，王瀚洋，李杨，叶宾
DOI：10.13195/j.kzyjc.2018.1299
收稿日期：2018-09-24
网络首发日期：2019-02-02
引用格式：李会军，王瀚洋，李杨，叶宾．一种基于视觉特征区域建议的目标检测方法[J/OL]．控制与决策. https://doi.org/10.13195/j.kzyjc.2018.1299

DOI: 10.13195/j.kzyjc.2018.1299

一种基于视觉特征区域建议的目标检测方法

李会军[1]，王瀚洋[1†]，李杨[1]，叶宾[1]

(1. 中国矿业大学 信息与控制工程学院，江苏 徐州 221116)

摘 要: 虽然基于深度学习的目标检测器具有较高的检测精度,但是大多数检测器的检测速度不能满足实时性要求.此外,目前主流的实时检测算法如SSD(Single Shot multibox Detector)和YOLO(You Only Look Once),对小目标的检测精度不高.因此,本文提出了一种基于视觉特征区域建议的目标检测算法,能够综合平衡检测精度和检测速度.算法分为区域建议和网络分类两部分,区域建议即根据目标的特征信息提取候选区域ROI(Region Of Interest);网络分类使用CNN(Convolutional Neural Network)对区域建议中提取的ROI进行处理,计算每个ROI类别的置信度,置信度大于设定阈值的ROI即为目标检测结果.实验结果表明,算法的检测精度明显高于Faster R-CNN、SSD和YOLO,并且具有接近SSD和YOLO的检测速度.

关键词: 目标检测；区域建议；卷积神经网络分类；视觉特征提取

中图分类号: TP399 **文献标志码:** A

An Object Detector based on Visual Feature Region Proposal

LI Hui-jun[1]，*WANG Han-yang*[1†]，*LI Yang*[1]，*YE Bin*[1]

(1. School of Information and Control Engineering，China University of Mining and Technology，Xuzhou 221116，China)

Abstract: Although the detector that based on Deep Learning can achieve high detection accuracy, but most of their speed cannot meet the real-time requirements. For the moment, the popular real-time detector, such as SSD(Single Shot multibox Detector) and YOLO(You Only Look Once). When detecting small objects, their accuracy is not high. Therefore, we propose a detector that based on visual features region proposal in this paper, which can balance the detection accuracy and speed. This detector is divided into two parts: region proposal and network classification. In the region proposal stage, we extract ROI(Region Of Interest) according to the feature information of the objects, ROI is also called candidate region; in the network classification stage, we use CNN(Convolutional Neural Network) to process the ROI, then calculate class confidence of each ROI, and get the final candidates whose confidence is greater than the threshold value. Experimental results show that the detection accuracy of this detector is significantly higher than Faster R-CNN, SSD and YOLO, and its speed is close to the speed of SSD and YOLO.

Keywords: target detection；region proposal；Convolution Neural Network classification；visual features extraction

0 引 言

随着研究的不断深入,深度学习已经成为目标检测和分类的常用工具[1-5].目前,目标检测领域的深度学习方法主要分为两类: 一类是基于回归的一阶检测器,如YOLO、SSD和FPN(Feature Pyramid Networks)等[6-8];另一类是基于区域建议的二阶检测器,如R-CNN(Regions with CNN feature)、SPP-Net(Spatial Pyramid Pooling Networks)、Fast R-CNN和Faster R-CNN等[9-12].

一阶检测器虽然检测速度较快,但是在模型训练时会出现前景和背景类别不均衡[13](如YOLO、SSD),同时难以提取小目标特征信息(如YOLO),从而导致目标检测效果不理想[14].二阶检测器虽然检测精度较高,但由于较为原始的候选区域生成方法(如R-CNN、Fast R-CNN)和相对复杂的网络结构(如Faster R-CNN),导致检测速度难以满足特殊场景中实时性要求[15].

在应用场景中,如果需要快速准确地检测视觉特征较强的小目标(例如处理速度＜40ms),现有的一阶和二阶检测器均无法满足检测要求.因此,本文提

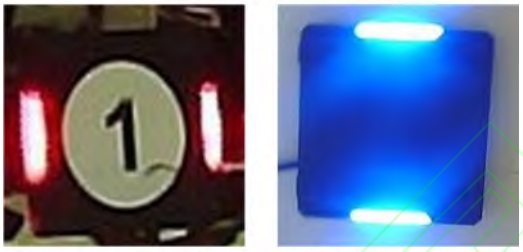
收稿日期: 2018-09-24；**修回日期:** 2018-11-26.
基金项目: 徐州市应用基础研究项目(KC18069)；中国矿业大学研究生教育教学改革研究与实践课题
†通讯作者. E-mail: ts16060151a3@cumt.edu.cn

出了一种基于视觉特征区域建议的二阶检测器RM R-CNN(Improved R-CNN in RoboMaster),并成功应用于RoboMaster全国机器人大赛中的装甲检测.

1 应用场景

RoboMaster机器人大赛采用红蓝双方机器人对抗形式,以敌方机器人上的装甲为击打目标.在比赛过程中,为了提高击打效率,需要机器人自动完成装甲识别.装甲与相机镜头的距离一般在0.5米至2.5米之间,此时装甲在视野中是一个小目标.红蓝色机器人和装甲外观如图1所示.

(a) 红色机器人和蓝色机器人



(b) 红色装甲和蓝色装甲

图 1 红蓝机器人和红蓝装甲

比赛场地位于整体光线较暗的全封闭体育馆内,场地周围存在复杂的灯光干扰,如图2所示.

图 2 RoboMaster机器人大赛场地

为了保证机器人在对抗时的击打效果,目标检测器的处理时间必须低于40ms,同时需要尽可能提高识别准确率.实验结果表明, Faster R-CNN的处理时间约为160ms, 不能满足实时性要求; YOLO和SSD虽然处理速度较快,但在小目标出现频率较高的场景中检测精度将会大大降低.针对上述问题,本文提出的RM R-CNN算法,具有较好的检测效果.算法框架如图3所示.

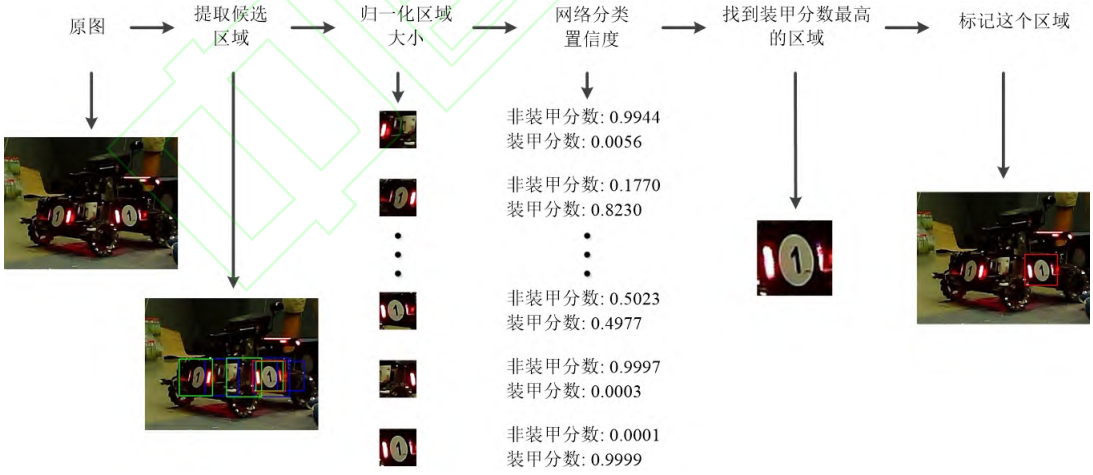


图 3 RM R-CNN算法流程

2 区域建议

RoboMaster机器人大赛中的目标装甲两侧有两个发光灯条,可以呈现红、蓝两种颜色,具有明显的视觉特征. RM R-CNN在区域建议阶段,可以根据目标装甲的视觉特征生成候选区域,大大减少了候选区域的数量,使得目标检测速度和精度均优于主流的二阶检测器,区域建议的算法流程如表1所示.

表 1 区域建议的算法流程

算法1 区域建议

输入: 一张彩色图片
输出: ROI的像素数据和ROI的坐标点
当该帧不为空时:
1. 将彩色图转换成灰度图;
2. 二值化灰度图, 得到二值图;
3. 膨胀处理二值图以消除小的光斑, 使得灯条更加明显;
4. 在处理后的二值图提取轮廓, 然后找出所有轮廓的最小包围旋转矩形, 并计算出矩形的面积;
5. 保留符合面积、长宽比、角度要求的旋转矩形;
6. 获取所有符合要求的旋转矩形的坐标 x_c, y_c, w 和 h;
7. 根据旋转矩形的长宽和角度, 在其左右侧分别画框;
8. 每两个旋转矩形间画框;
9. 保留符合长宽比要求的自定义矩形框;
10. 获取所有符合要求的自定义矩形框(ROI)的坐标, 包括 x_{lu}, y_{lu}, w 和 h;
结束
返回ROI的像素数据和ROI的坐标点

首先对二值化后的图像进行阈值分割再提取轮廓, 然后找出符合灯条几何要求的旋转矩形. 算法1中, 在第5步定义了满足灯条旋转矩形的参数: 面积＞30(像素面积), 1.6＜长宽比＜10, 0°＜旋转角度＜20°或者70°＜旋转角度＜90°. x_c, y_c 为旋转矩形的中心点坐标, x_{lu}, y_{lu} 为候选区域的矩形左上角坐标, w 和 h 为所有矩形的宽和高. 在第9步中经过反复实验后, 仅保留长宽比小于1.8并大于0.56的矩形. 图4为本阶段候选区域提取流程效果图, 图中红色、绿色及蓝色框分别是算法流程中第7步和第8步中不同方法生成的候选区域.

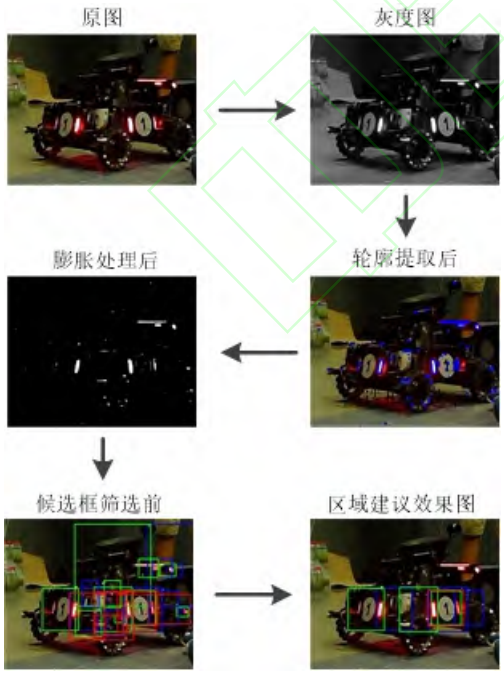


图4 候选区域提取效果图

通过几何特征等限定条件筛选后, 每帧剩余数十个左右的候选区域. 如果视野中高亮物体较多, 候选区域的数量将略微增加.

3 参考文献

3.1 网络结构

LeNet-5[16] 和 ZFNet[17] 对于小尺寸数据集(如MNIST数据集)具有较高的分类精度, 这类浅层网络结构和小规模的全连接层能够有效减少计算量、加快分类速度. 因此, 本文对VGG-16[18] 网络进行修改和调整后, 得到如图5所示的9层卷积神经网络.

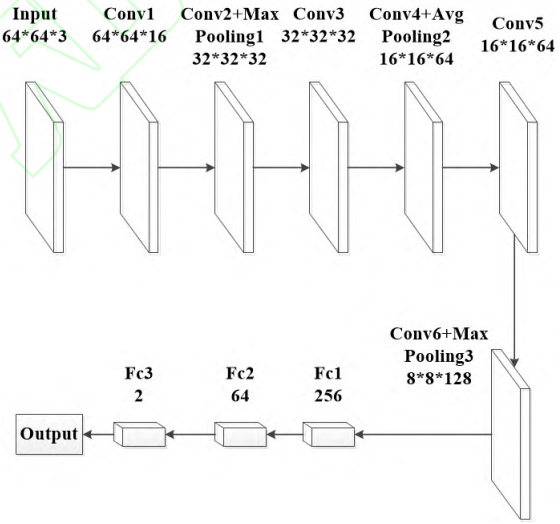


图5 卷积神经网络

将ROI尺寸归一化为64×64像素大小, 作为卷积神经网络的输入, 然后根据网络输出值判断ROI是否为装甲. 每个卷积层后的激活函数均为ReLU[19], 卷积核大小均为3 × 3. 因为该网络要判断ROI是否为装甲, 属于二分类问题, 所以Fc3层的输出维度为2.

池化层中特征提取的误差主要来自两个方面: 邻域大小受限造成的估计值方差增大; 卷积层参数误差造成估计均值偏移. 最大值池化能减小第一种误差, 保留更多图像背景信息; 均值池化能减小第二种误差, 保留更多纹理信息[20]. 池化层输入的Feature Map为 F, 采样的池化区域为 $c \times c$, 偏置为 b, 均值池化层的输出 S 为:

$$S_{ij}=\frac{1}{c}(\sum_{i=1}^{c}\sum_{j=1}^{c}F_{ij})+b \tag{1}$$

最大值池化层的输出S为:

$$S_{ij}=\max_{i=m,j=n}^{c*c}F_{ij}+b \tag{2}$$

式中$\max_{i=m,j=n}^{c*c}(F_{ij})$表示最大值池化层从输入$F$提取区域为$c\times c$中最大元素.

最大值池化层和均值池化层的采样区域均为2×2.第三次池化后输出$8\times 8\times 128$的Feature Map,扁平成1维向量后接上3个全连接层,最后的全连接层Fc3输出2个值,通过Softmax[21]计算每个候选区域属于装甲类或非装甲类的概率. Softmax公式如下所示:

$$S_i=\frac{e^{V_i}}{e^{V_1}+e^{V_2}} \tag{3}$$

式中V_i代表第i个元素值,表示该元素在这2个元素中的概率值.

AdamOptimizer实现简单、计算高效、能够自动调整步长[22].因此,分类网络训练中使用AdamOptimizer优化器.

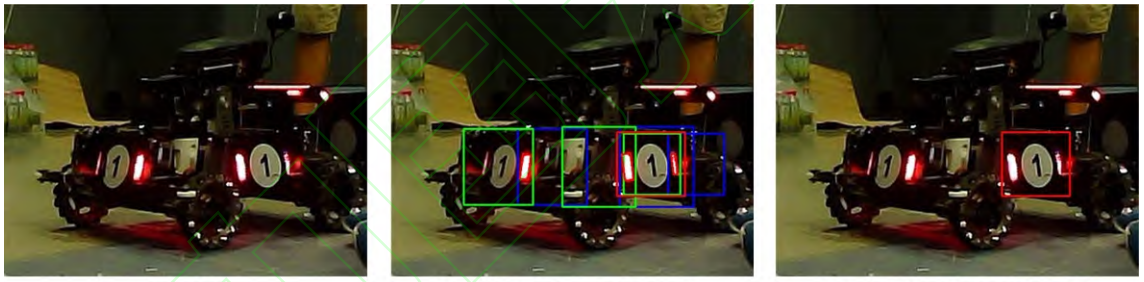
3.2 最优ROI搜索

每帧图片经过网络处理完毕后,按顺序保存ROI的坐标及其置信度,然后搜索置信度最高的ROI,搜索算法如表2所示.

表 2 最优ROI搜索的算法流程

算法2 最优ROI搜索
输入: 每帧图片ROI的坐标及其对应的置信度 输出: 每帧图片最高装甲类别的置信度及对应的ROI 当候选区域数量不为0时: 1. 找出装甲类别置信度最高值且大于阈值0.5,以及其对应的索引.若该值不存在,则没有输出结果; 2. 根据索引,找出对应的候选区域坐标; 3. 依据坐标在图片画框; 结束或跳转至下一帧

图6是本阶段的效果图: (a)是原图, (b)是提取候选区域的效果图, (c)是找出最高装甲类别置信度的最终图.尽管(b)中有数个候选区域是标注到装甲并具有较高的置信度,但它们不是装甲类别置信度最高的候选区域.



(a) 原图 (b) 提取候选区域的效果图 (c) 装甲检测的最终效果图

图 6 装甲检测效果图

4 实验验证及分析

4.1 实验数据和计算平台

为了避免网络训练时正负样本不均衡,在收集实验数据时,已将所有候选区域手动分类、并保证正负样本数量基本一致.训练数据集共选取2450个样本(正样本1225个、负样本1225个).图7所示为红色装甲类的正样本和非红色装甲类的负样本,其中负样本可以是蓝色装甲、其他光源及高亮物体.

(a) 红色装甲正样本

(b) 红色装甲负样本

图 7 红色装甲类正样本和非红色装甲类负样本

计算平台: CPU为Inter Core i7-7700HQ@2.80GHz,四核八线程,16G内存, GPU为NVIDIAGTX 1050,相机为200万像素基于OV2710方案的CMOS相机,深度学习框架使用Tensorflow.

4.2 实验结果

分类模型在2450个训练样本上的分类准确率为99.7%,在930个测试样本上的分类准确率为99.3%.测试完成后,使用三段800*600的视频验证,视频录制环境如图8 所示: (a)为蓝色装甲快速运动的视频,(b)为蓝色装甲静止的视频, (c)为红色装甲快速运动的视频.其中视频(a)和(b)的环境亮度较高,视频(a)和(b)因为运动较快且倾角较大,使得装甲在视野中很小,检测难度较大.

(a) 蓝色装甲快速运动 (b) 蓝色装甲静止 (c) 红色装甲快速运动

图 8 视频录制环境

为了验证RM R-CNN目标检测方法的有效性,在同样实验硬件平台和相同的主网络结构下,与FasterR-CNN、YOLO和SSD进行横向对比.为了分析RM R-CNN网络分类的准确性,与RM R-CNN+VGG16进行对比,即在同样的区域建议方法上,使用了两个不同的网络进行分类.实验中预测框与实际框的交并比低于0.5视为误识别帧,表3、表4、表5分别为视频(a)、(b)、(c)的实验结果数据表.

表 3 装甲检测实验数据表(视频a,共2610帧)

算法	RM R-CNN	Faster R-CNN	SSD	YOLO	RM R-CNN+VGG16
每帧处理速度	35.0ms	162.2ms	30.6ms	24.6ms	69.4ms
准确率	89.5%	87.1%	88.9%	79.0%	93.2%
召回率	88.1%	86.8%	80.4%	52.5%	89.5%

表 4 装甲检测实验数据表(视频b,共1151帧)

算法	RM R-CNN	Faster R-CNN	SSD	YOLO	RM R-CNN+VGG16
每帧处理速度	33.6ms	157.9ms	29.5ms	23.9ms	65.2ms
准确率	100%	97.9%	94.6%	95.9%	100%
召回率	100%	95.2%	96.9%	87.3%	100%

表 5 装甲检测实验数据表(视频c,共12209帧)

算法	RM R-CNN	Faster R-CNN	SSD	YOLO	RM R-CNN+VGG16
每帧处理时间	36.3ms	164.6ms	30.7ms	25.4ms	74.7ms
准确率	98.5%	90.8%	85.6%	79.7%	99.1%
召回率	96.9%	87.9%	83.9%	61.8%	97.3%

综合三种场景下的实验数据分析,在相同的主网络结构和测试硬件中比较各检测器性能.RM RCNN速度略慢于SSD和YOLO,速度分别约为SSD的0.8倍和YOLO的0.6倍; 其速度明显快于Faster RCNN,约为后者的5倍. RM R-CNN的召回率和准确率均高于其他三种常见的检测器,较Faster RCNN的准确率和召回率分别约提升7%和8%,较SSD的准确率和召回率分别约提升12%和14%,较YOLO的准确率和召回率分别约提升20%和54%.尽管RM R-CNN的速度并非最优,但是也可以满足每帧处理时间低于40ms的需求.同时其准确率和对小目标的召回率都要优于其他三种检测器,因此能更好的适应于本场景.

RM R-CNN和RM R-CNN+VGG16相比,后者的准确率和召回率分别约提升1%和0.5%.尽管使用VGG16作为分类网络的精度更高,但是RM R-CNN的速度约为其速度的2倍,因此不能满足应用场景的实时性要求.通过对比实验分析,RM R-CNN比其他方法更能满足检测要求.

5 结论

本文实现了一种有效的目标检测方法,能够快速准确地检测具有较强视觉特征的小目标,在较暗的场景中的检测效果将更加优异.算法在区域建议中,通过特征信息提取候选区域,使得区域建议更具有目的性,有效控制了候选区域的数量; 同时改进VGG-16的网络结构,在保证分类精度高的前提下加快了分类速度.实验结果表明,相比于Faster R-CNN, SSD和YOLO, RM R-CNN在速度和精度上的综合性能更好. RM R-CNN可以用于解决其他特殊场合下的目标检测问题,如夜晚或阴暗场景下车辆和车牌的检测、行人检测、城市安防等.

参考文献(References)

[1] Xie X, Wang C, Chen S, et al. Real-Time Illegal Parking Detection System Based on Deep Learning[C]. Proceedings of the 2017 International Conference on Deep Learning Technologies. Chengdu: ACM, 2017: 23-27.

[2] Tsehay Y K, Lay N S, Roth H R, et al. Convolutional neuralnetwork based deep-learning architecture for prostate cancer detection on multiparametric magnetic resonanceimages[C]. Medical Imaging 2017. Orlando: Computer-Aided Diagnosis. International Society for Optics and Photonics, 2017: 1013405.

[3] Zhang X, Chen G, Saruta K, et al. Deep Convolutional Neural Networks for All-Day PedestrianDetection[C]. International Conference on Information Science and Applications. Singapore: Springer, 2017: 171-178.

[4] Sun X, Wu P, Hoi S C H. Face detection using deep learning: An improved faster RCNN approach[J]. Neurocomputing, 2018, 299: 42-50.

[5] Jiang H, Learned-Miller E. Face detection with the faster R-CNN[C]. Automatic Face Gesture Recognition. Washington: IEEE, 2017: 650-657.

[6] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas: IEEE, 2016: 779-788.

[7] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[C]. European Conference on Computer Vision. Amsterdam: Springer, 2016: 21-37.

[8] Lin T Y, Dollár P, Girshick R B, et al. Feature Pyramid Networks for Object Detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE, 2017, 1(2): 3.

[9] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. Columbus: IEEE, 2014: 580-587.

[10] He K, Zhang X, Ren S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis Machine Intelligence, 2015, 37(9): 1904-1916.

[11] Girshick R. Fast r-cnn[C]. Proceedings of the IEEE international conference on computer vision. Santiago: IEEE, 2015: 1440-1448.

[12] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]. Advances in neural information processing systems. Montreal: IEEE, 2015: 91-99.

[13] Lin T Y, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection[J]. IEEE Transactions on Pattern Analysis Machine Intelligence, 2017, 99: 2999-3007.

[14] Zhang S, Zhu X, Lei Z, et al. S^3FD: Single Shot Scale-Invariant Face Detector[C]. Proceedings of the IEEE international conference on computer vision. Venice: IEEE, 2017: 192-201.

[15] Ren Y, Zhu C, Xiao S. Object Detection Based on Fast/Faster RCNN Employing Fully Convolutional Architectures[J]. Mathematical Problems in Engineering, 2018, 2018.

[16] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[17] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]. European conference on computer vision. Zurich: Springer, 2014: 818-833.

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

[19] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines[C]. Proceedings of the 27th international conference on machine learning. Haifa: IMLS, 2010: 807-814.

[20] Lee C Y, Gallagher P W, Tu Z. Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree[C]. Artificial Intelligence and Statistics. Cadiz: Computer Science, 2016: 464-472.

[21] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[22] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv: 1412.6980, 2014.

作者简介

李会军 (1980–), 男, 副教授, 博士, 从事计算机视觉、计算机控制等研究, E-mail:plutoli@163.com.

王瀚洋 (1994–), 男, 硕士生, 从事深度学习、计算机视觉的研究, E-mail:ts16060151a3@cumt.edu.cn.