

DOI: 10.7652/xjtuxb201905008

面向图像的有效目标区域提取方法

崔云博, 杜友田, 王航

(西安交通大学智能网络与网络安全教育部重点实验室, 710049, 西安)

摘要: 针对图像标注、目标识别等实际应用中图像的前景目标定位不够准确的问题,提出了一种图像中的有效目标区域提取方法。该方法以提取图像的前景目标为目的,将目标区域提取问题转化为二分类问题,实现了对图像中有效目标区域的提取,主要包括 4 个步骤:利用选择性搜索算法生成图像中的候选目标区域;通过对像素值的差值化处理来进行图像区域的特征增强;基于深度学习实现对候选目标区域进行分类;区域选择与融合。在 MSCOCO 数据集上进行实验,结果表明,该方法在保证较高召回率的基础上,达到了比现有多种算法更加准确的目标区域定位结果。

关键词: 目标区域提取;选择性搜索;特征增强;深度学习

中图分类号: TP391.4 **文献标志码:** A **文章编号:** 0253-987X(2019)05-0052-06

Extraction Method for Key Object Region in Images

CUI Yunbo, DU Youtian, WANG Hang

(Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: To improve the performance of object region extraction, which is usually used in some practical applications such as image annotation and object detection, we present a new approach to the key object region extraction from images. The method ignores the category information of various objects and transforms the task into a binary classification problem. Our method includes four steps: First, a selective search algorithm is used to generate candidate object regions in the image; Second, the feature enhancement of image regions is performed by difference processing of the pixel values; Then, we classify the candidate object regions based on deep learning; Finally, region selection and fusion are finished. Experimental results over MSCOCO dataset demonstrate that the proposed approach achieves more accurate results than several related methods with high recall rates.

Keywords: object region extraction; selective search; feature enhancement; deep learning

随着计算机视觉技术及深度学习的迅速发展,面向图像的目标检测任务重新成为了备受关注的研究问题。目标检测是将图像中的有效目标从背景中分离出来并进行识别,为图像内容的进一步分析提供基础,对于图像分类^[1]、图像标注^[2]、图像标题与

描述生成^[3]、图像搜索^[4]等应用具有重要意义。

Freund 等于 1995 年提出了基于 Adaboost 和矩形 Harr 特征相结合的人脸检测方法^[5]。2005 年, Dalal 等提出利用方向梯度直方图(HOG)^[6]来进行目标的特征表示。近年来,基于深度学习的目

收稿日期: 2018-09-05。 作者简介: 崔云博(1994—),男,硕士生;杜友田(通信作者),男,副教授。 基金项目: 国家自然科学基金资助项目(61572397);北京字节跳动科技有限公司合作项目(20180509);中央高校基本科研业务费专项资金项目(xjj2017063)。

网络出版时间: 2019-02-27

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1069.T.20190226.1355.018.html>

标检测研究也被广泛关注。2014 年,Girshick 等提出了 R-CNN 算法,并将其应用于目标检测任务中^[7],在此基础上,2015 年提出一种改进的 Fast R-CNN 算法,提高了算法的效率^[8]。2015 年,Ren 等提出了 Faster R-CNN 算法,通过引入自主设计的候选区域网络 RPN(region proposal networks),提高了目标检测速度^[9]。2017 年,He 等提出了 Mask R-CNN 算法,用于目标语义识别并取得了良好的识别效果^[10]。一般地,目标检测算法包括两步:首先是对目标进行定位,即找出目标的候选区域;然后在此基础上进行目标识别或分类。提取目标候选区域的效果在很大程度上影响了后续的目标识别或分类结果,传统的目标候选区域产生方法是将矩形窗口在不同尺度的图像上进行滑动,其缺点是计算量巨大、计算速度慢。此后,研究者提出了 Bing、Edge-Boxes 等算法^[11-12]来挑选目标候选区域,在保证召回率的基础上提高了计算速度,但是在实际应用中,该类方法仍然会存在大量冗余的区域建议。

为了解决以往算法的图像中前景目标位置不准确、精确度不高等问题,本文提出一种高效的有效目标区域提取方法,主要包括基于选择性搜索^[13]的候选目标区域(object proposal)生成、目标区域特征增强、基于深度网络的分类、区域筛选与融合 4 个步骤。实验结果表明,在保证较高召回率的基础上,本文方法得到了更加准确的有效目标区域提取结果。

1 有效目标区域提取框架

问题定义:给定测试图像集以及已被标注的表示有效目标区域或背景区域的正、负样本 $\{x_1^+, \dots, x_{N_1}^+, x_1^-, \dots, x_{N_2}^-\}$,提取测试图像中的有效目标区域。本文中的有效目标区域指的是图像中准确包含典型目标的区域。

由于在很多情况下难以获取到大量具有目标类别标注的样本,而且在一些应用中也不需要进行端对端的目标类别识别。因此,本文忽略了目标类别

信息,提出一种提取图像中有效目标区域的方法,将目标区域提取过程看作一个二分类问题:目标区域和背景区域,实现了对图片中前景目标和背景的区分,最终获取到有效目标的区域位置。整体方案框架如图 1 所示,主要包括 4 个环节:①基于选择性搜索算法对图像进行搜索和计算,生成一定数量的候选目标区域;②对候选区域进行目标特征增强处理,增强不同目标之间的共性视觉特征,使不同类别目标的颜色特征聚合更为紧凑;③构建基于深度学习的分类网络,并利用该分类网络对图像中的目标区域和背景区域进行分类,判断其是否属于目标区域;④将被识别为有效目标的候选区域进行区域筛选和融合,最终实现有效目标区域的提取过程。

2 有效目标区域提取算法

2.1 选择性搜索

在生成候选目标区域时,传统方法利用不同尺度的滑动窗口对图像进行扫描,其缺点是会产生大量冗余的候选区域,增加不必要的计算量。因此,本文基于选择性搜索算法^[13]进行候选目标区域的生成。选择性搜索算法是将图像作为输入,首先产生大量的初始候选区域,然后根据相邻区域之间的相似度,计算出相似区域的层次分组,并不断进行区域融合,最终输出一定数量的候选区域。相邻区域之间的相似度计算基于颜色、纹理、尺寸和空间交叠这 4 种属性进行。对于通过选择性搜索算法所生成的目标区域建议,本文去掉了一些面积过小的候选区域(实验中设定像素点数小于 2 000)。

2.2 目标区域特征增强

本文忽略了目标的类别信息,没有对图像中的目标进行类别区分,而是将提取目标的所在区域作为算法的核心内容,该过程也可以看作是将不同类别的目标作为同一类别进行提取。由于多种类别的目标之间有较大区别,使得目标在特征空间中的位置较为分散,增大了目标区域提取的难度,因此本文

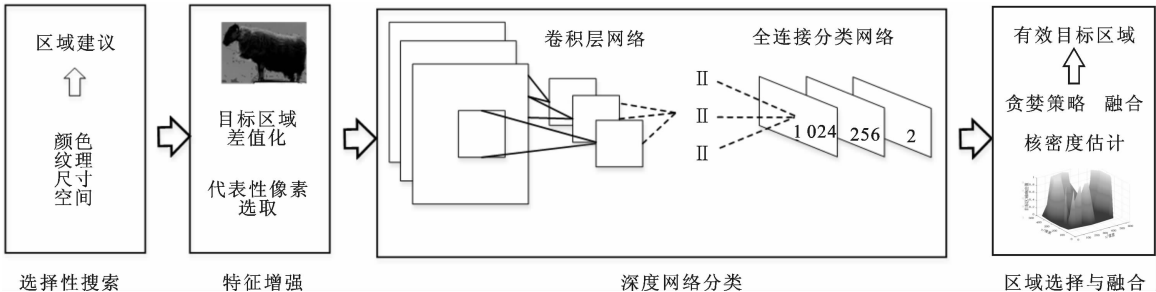


图 1 有效目标区域提取框架

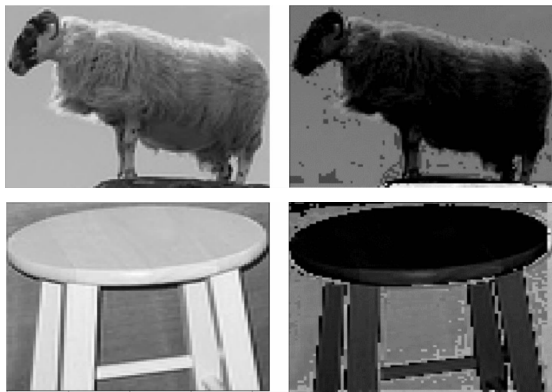
提出采用目标区域增强的方法来增大目标区域样本与背景区域样本之间的区别。

目标区域内容具有较为连续的视觉感受、明显的边缘区域、较为复杂的细节结构等共性特点,而对于不同的目标,通常具有不同的颜色、纹理、亮度等。因此,目标区域特征增强的目的是突出目标区域的共性特点,增强过程可以看作将目标区域差值化,即

$$p'_{ij} = p_{ij} - \frac{1}{K} \sum_{k=1}^K \hat{p}_k \quad (1)$$

式中: p_{ij} 为区域中任一像素点的值; p'_{ij} 为特征增强后的像素值; $\hat{p}_k$ 为目标区域中的代表性像素点。本文选择矩形目标区域的水平和竖直两条中线上的 1/4、1/2 和 3/4 位置上共 5 个像素点作为代表性像素点,即 $K=5$,并对于 R、G、B 这 3 个通道分别进行计算。

在实际的图像中,目标区域特征增强前后的对比结果如图 2 所示。由图 2 可知,对目标区域进行特征增强之后,使得目标之间的差异程度更小,同时该操作依旧保留了目标的形状信息,有利于区分生成的有效目标区域和其他区域。



(a) 特征增强前

(b) 特征增强后

图 2 目标区域特征增强前后比较

2.3 基于深度网络的目标区域检测

本文采用的网络结构基于 VGG-16 深度网络^[7],保留了该网络的前 13 层结构作为提取图片特征的部分,在此基础上添加了 3 个全连接层作为分类网络,最终输出一个 2 维向量,以此判断输入图像区域是有效目标区域还是背景区域(非有效目标)。其中对输入到网络中的图像区域(image patch)要首先进行缩放处理,调整至 224×224 大小,其目的是对于不同输入都可以保证在进入全连接层之前,网络所产生的向量长度是一致的。另外,在具体的网络结构中,卷积层所使用的激活函数是 ReLU 函数;池化层所采用的池化方式为最大池化;3 个全

连接层的神经元个数则分别设置为 1 024、256 和 2;为了消除过拟合现象,层与层之间会随机丢掉一些连接权重;最后,使用 softmax 函数作为激活函数,保证输出向量的元素值大小属于区间 $[0, 1]$ 。

在训练网络时,首先需要构建训练样本集,包括正样本和负样本两部分。本文将图像集提供的真实目标边界框所标记的图像区域作为正样本;在选择负样本时,则是基于 2.1 节选择性搜索算法得到的目标候选区域,并选择与真实目标区域交叠度小于一定阈值的图像区域作为负样本。本文采用交并比 U 来度量图像区域之间的交叠度,定义为

$$U = \frac{A(S^+ \cap S)}{A(S^+ \cup S)} \quad (2)$$

式中: S^+ 、 S 分别为正样本 x^+ 代表的图像区域和其他任意图像区域; $A(\cdot)$ 为图像区域的面积。在挑选负样本时将交叠度阈值 ξ_1 设为 0.3,即若 $U < 0.3$,则将 S 对应的图像区域定义为负样本 x^- 。对于样本标签,定义向量 $y = (1, 0)^T$ 为正样本标签,向量 $y = (0, 1)^T$ 为负样本标签。

2.4 基于目标响应的区域选择与融合

通过基于深度网络的分类,每个候选区域 x 会对应一个 2 维的预测向量 $\hat{y}$ 。该预测向量第 1 维分量,记为 $\hat{y}$,反映了 x 被判别为候选目标区域的置信度,也可以看作 x 输入到上述深度网络系统后得到的响应。因此,对于每幅图像我们可以得到二元组集合 $\{(x, \hat{y})\}$ 。为了更好地考虑上述预测结果在空间上的关联性,基于核密度估计方法对上述预测置信度进行重新估计,得到结果 $\{(x, \bar{y})\}$

$$\bar{y} = \frac{1}{\Delta} \sum_i \hat{y}_i G\left(\frac{\|x^c - x_i^c\|_2}{h}\right) \quad (3)$$

式中: $G(\cdot)$ 为零均值、单位协方差矩阵的高斯核函数; $\Delta = \sum_i G(\|x^c - x_i^c\|_2/h)$, h 为核函数带宽; x^c 为目标区域中心坐标处的特征值。置信度较低的区域对应有效目标的概率较小,若 $\bar{y} > \delta$, x 则被保留为候选目标区域,本文实验中选择 $\delta = 0.7$,此时候选目标区域依然存在大量相互交叠,这些区域代表了同一个目标。为了得到准确、稀疏的目标区域提取结果,采用贪婪策略对其进行合并:对于每个候选的目标区域,计算它与其他候选区域的 U 值;如果搜索到与其交叠度大于一定阈值 $\xi_2 = 0.7$ 的候选区域,则比较两者的置信度,并保留置信度更大的候选区域作为新的目标区域;迭代上述过程,直至找不到可以合并的目标区域。

3 实验结果与分析

3.1 实验数据集

本文选择 MSCOCO2014 数据集作为实验数据来源,该数据集分成训练集、验证集和测试集 3 个部分,包含了图像文件以及其对应的目标区域标注结果,其中图像内容共包含 80 类目标概念,标注结果提供了图像中目标的类别及位置。在实验过程中,本文随机选择了训练集中的 3 万张图像来训练网络,其中涵盖了所有的 80 类目标概念,为了保证训练样本平衡性与学习性能,本文基于这些图像共生成 10 万个正样本和 10 万个负样本用于训练深度网络。由于本工作不涉及不同目标类别的判别,因此在使用标注结果时,利用了图像中目标的位置信息而忽略了其类别信息。在测试过程中,为了评估本文方法,采用了具有标注结果的验证集。从验证集中选取涵盖所有的 80 类目标概念的 1 000 张图像来评估最终的实验结果。

3.2 目标区域特征增强的结果

图 3 展示了进行特征增强前后,目标区域中所有像素值在特征空间中的分布变化,本文随机选取了 20 个目标区域并对其增强前后像素值的分布在 RGB 空间进行了展示,图 3 中 x 、 y 、 z 分别表示 R、G、B 空间。由图 3 可知:在特征增强前,目标区域的多样性导致了像素值分布非常分散;对目标区域进行特征增强之后,目标区域在特征空间中的分布更加集中,且分布在原点附近的较小范围内,这对于提高目标区域分类的性能有很大帮助。

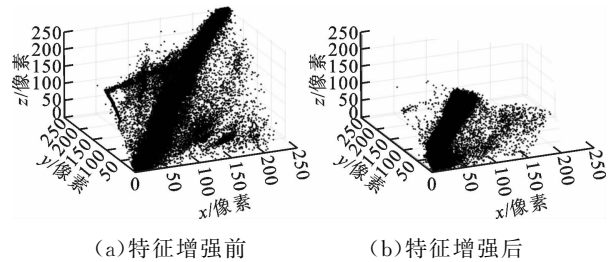


图 3 特征增强前后区域像素值在空间中的分布

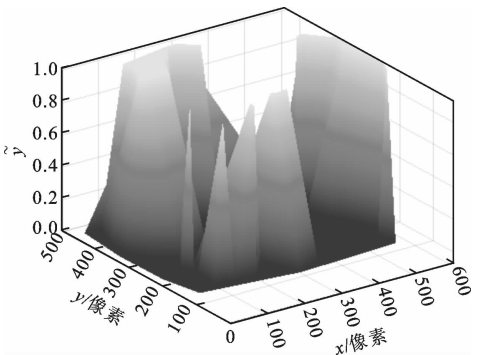
3.3 目标区域的融合结果

针对深度网络输出的候选目标区域,基于目标响应方法,将提取结果进行进一步融合,得到最终的目标区域,目标区域的融合过程如图 4 所示。由图 4 可知,通过基于目标响应的区域融合过程,最终得到的目标区域不仅保留了网络输出结果中置信度高的区域,保证了有效目标可以达到较高的召回率,同时也过滤掉了大量对应于同一目标的重叠区域,在

一定程度上提高了目标检测的准确率。



(a)原始图像上的候选目标区域



(b)目标区域响应



(c)区域融合后的目标区域

图 4 目标区域的选择

3.4 目标区域提取召回率比较

本文工作的目的是提取出目标候选区域,为图像分类、目标识别等应用提供帮助。召回率是判断方法有效性的一个重要指标,召回率低意味着检测结果会漏掉一些目标区域,影响后续研究任务的性能,以往工作多采用召回率作为性能评估指标^[14]。

对相关算法以及不同的深度网络模型与本文方法在 MSCOCO 数据集上进行了性能比较,结果如图 5 所示。平均召回率是指对不同 U 阈值情况下的召回率取平均值,在本实验中, U 阈值取 0.5~1,步长为 0.1。本文选取的 3 种对比算法都是目前候

选目标区域提取任务中性能较好的算法,为了分析不同网络结构对本文方法性能的影响,将 Alex-Net^[15]作为深度网络模型进行对比。由图 5 可知,相比于其他算法,本文方法在较少候选区域的条件下就可以达到更高的召回率,同时本文方法选取的 VGG-16 网络也较 AlexNet 表现出更好的性能。

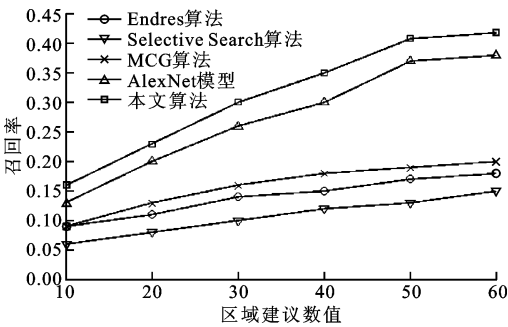


图 5 平均召回率比较($U \in [0.5, 1]$)

为了进一步分析本文方法在特定类别的目标区域提取中的性能差别,统计了不同类别目标的目标区域提取结果,并选取了 MSCOCO 数据集中部分类别的结果,如表 1 所示,其中 $\delta=0.7$,即把与正样本的交叠度大于 0.7 的目标区域作为正例进行统计,同时在实验中比较了选择性搜索算法和本文方法的性能。为了保证算法比较的公平性,考虑到算

表 1 不同类别目标的召回率对比

目标类别	召回率/%	
	选择性搜索 ^[13]	本文方法
person	52.63	53.04
bicycle	48.33	50.00
car	37.93	62.50
truck	43.85	45.15
sheep	37.50	62.50
zebra	36.36	72.73
bottle	21.05	47.37
cup	37.50	50.00
bowl	57.14	85.71
banana	27.27	81.82
sandwich	40.00	80.00
broccoli	28.57	64.29
carrot	31.24	76.92
chair	47.37	57.89
couch	68.75	87.50
dining table	75.86	58.62
oven	58.82	70.59

法之间存在差异性,对于选择性搜索算法,每张原始图像的候选区域数量平均选择了 100 个左右,而对于本文方法则平均选取了 40 个左右的目标区域。

由表 1 可知,相比于选择性搜索算法,本文方法在目标区域更少的条件下,对有效目标得到了更高的召回率,表现出了更好的性能,对不同类别的目标,本文方法在性能上表现出了一定的差异性。对 truck、bowl 这 2 类目标,最终的召回率统计结果有比较大的差别,这可能因为对于图像面积较大的某些类目标(如 truck),其图像内容也会变得更加复杂,这就在一定程度上对分类结果的准确性造成了影响,并且在利用选择性搜索策略进行候选区域生成时,面积较大、内容较复杂的区域也无法完整找到,相比于面积较小、内容较单一的某些类目标(如 bowl),最终性能上存在一定的差异。

图 6 展示了有效目标区域提取的部分样例。由图 6 可知,本文方法对于图像中大部分的有效目标区域实现了较好提取,并且最终的提取结果在保证较高召回率的基础上,有较高的准确性。

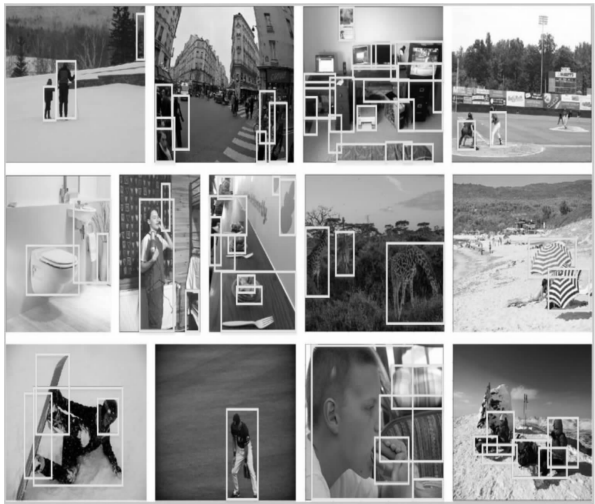


图 6 本文方法在 MSCOCO 验证集上的部分结果

4 结 论

本文提出了一种提取图像中有效目标区域的方法以解决目标检测等任务中有效目标定位不准确的问题。以选择性搜索策略为基础,通过目标区域特征增强、深度网络分类、目标区域融合等环节,完成了对区域中是否存在有效目标的准确判别,并实现了对目标区域的提取。本文的目标区域提取结果是独立于目标类别的,不需要目标类别标注,该方案复杂度较低,具有较好的性能。实验结果表明:本文方法对选择性搜索算法在性能上有较大的提高;相较

于其他的目标候选区域提取算法,本文方法也可以使对有效目标的提取达到更高的召回率。

参考文献:

- [1] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2016: 770-778.
- [2] ZHOU Z H, ZHANG M L, HUANG S J, et al. Multi-instance multi-label learning [J]. Artificial Intelligence, 2012, 176(1): 2291-2320.
- [3] WU J, YU Y, HUANG C, et al. Deep multiple instance learning for image classification and auto-annotation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2015: 3460-3469.
- [4] LIN K, YANG H F, HSIAO J H, et al. Deep learning of binary hash codes for fast image retrieval [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2015: 27-35.
- [5] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2001: 511.
- [6] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2005: 886-893.
- [7] 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. Piscataway, NJ, USA: IEEE, 2014: 580-587.
- [8] GIRSHICK R. Fast R-CNN [C]//Proceedings of 2015 IEEE International Conference on Computer Vision. Piscataway, NJ, USA: IEEE, 2015: 1440-1448.
- [9] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.
- [10] HE K, GKIOXARI G, DOLLAR P, et al. Mask R-CNN [C]//International Conference on Computer

Vision. Piscataway, NJ, USA: IEEE, 2017: 2980-2988.

- [11] CHENG M M, ZHANG Z, LIN W Y. Bing: binarized normed gradients for objectness estimation at 300fps [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2014: 3286-3293.
- [12] ZITNICK C L, DOLLAR P. Edge boxes: locating object proposals from edges [C]//European Conference on Computer Vision. Berlin, Germany: Springer, 2014: 391-405.
- [13] UIJLINGS J R P, SANDE K E A, GEVERS T, et al. Selective search for object recognition [J]. International Journal of Computer Vision, 2013, 104(2): 154-171.
- [14] HOSANG J, BENENSON R, DOLLAR P, et al. What makes for effective detection proposals [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(4): 814-830.
- [15] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[本刊相关文献链接]

- 赵振冲,王晓丹.引入拒识的最小风险弹道目标识别.2018,52(4):132-138.[doi:10.7652/xjtuxb201804019]
- 余星达,陈文杰,王鼎,等.非接触式身份识别的深度学习算法.2019,53(4):122-127.[doi:10.7652/xjtuxb201904018]
- 袁帅,秦贵和,晏婕.应用残差生成对抗网络的路况视频帧预测模型.2018,52(10):146-152.[doi:10.7652/xjtuxb201810020]
- 钟德星,杨元,刘瑞玲,等.基于单目视觉的装配机器人研究及应用.2018,52(5):81-87.[doi:10.7652/xjtuxb201805012]
- 王博,张为,刘艳艳,等.采用机器学习的火焰前景提取算法.2017,51(8):26-32.[doi:10.7652/xjtuxb201708005]
- 刘小勇,郑琨.光纤缺陷实时检测与分类方法研究.2014,48(9):1-5.[doi:10.7652/xjtuxb201409001]
- 孟自强,李亚超,李浩林,等.双/多基地合成孔径雷达前视三维分辨力研究及运动参数设计.2014,48(8):29-35.[doi:10.7652/xjtuxb201408006]
- 屠礼芬,仲思东,彭祺.自然场景下运动目标检测与阴影剔除方法.2013,47(12):26-31.[doi:10.7652/xjtuxb201312005]
- 杜友田,辛刚,郑庆华.融合异构信息的网络视频在线半监督分类方法.2013,47(7):96-101.[doi:10.7652/xjtuxb201307018]

(编辑 赵炜)