计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(12):262−267 [doi: 10.15888/j.cnki.csa.006102]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Faster R-CNN 的人脸检测方法①

董兰芳, 张军挺

(中国科学技术大学 计算机科学与技术学院, 合肥 230027)

摘　要: 近年来, 基于候选区域的快速卷积神经网络 (Faster R-CNN) 算法, 在多个目标检测数据集上有出色的表现, 吸引了广泛的研究兴趣. Faster R-CNN 框架本来是用做通用目标检测的, 本文将它应用到人脸检测上, 分别使用 ZF 和 VGG16 卷积神经网络, 在 WIDER 人脸数据集上训练 Faster R-CNN 模型, 并在 FDDB 人脸数据库上测试. 实验结果表明, 该方法对复杂光照、部分遮挡、人脸姿态变化具有鲁棒性, 在非限制性条件下具有出色的人脸检测效果. 这两种网络结构, 在检测效率和准确性上各有优势, 可以根据实际应用需求, 选择使用合适的网络模型.

关键词: 人脸检测; 候选区域; 卷积神经网络; 非限制性条件

引用格式: 董兰芳,张军挺.基于 Faster R-CNN 的人脸检测方法.计算机系统应用,2017,26(12):262–267. http://www.c-s-a.org.cn/1003-3254/6102.html

Face Detection Using the Faster R-CNN Method

DONG Lan-Fang, ZHANG Jun-Ting

(College of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China)

Abstract: Recently, the faster R-CNN has demonstrated impressive performance on various object detection benchmarks, and it has attracted extensive research interests. We train a faster R-CNN model on the WIDER face dataset with the ZF and VGG16 convolutional neural network respectively, and then we test the trained model on the FDDB face benchmark. Experimental results demonstrate that the method is robust to complex illumination, partial occlusions and facial pose variations. It achieves excellent performance in detecting unconstrained faces. The two kinds of network have their own advantages in detection accuracy and efficiency, so we can choose to use an appropriate network model according to the actual application requirements.

Key words: face detection; candidate region; convolutional neural network; unconstrained condition

1　引言

传统的人脸检测方法, 主要在 Viola 和 Jone[1]的工作基础上进行改进, 研究者们关注手工设计的图像特征, 以及不同的级联结构. 各种复杂的特征, 如文献[2-6]等特征被设计出来, 代替 Haar-like 特征. 这些手工设计的特征在一定程度上可以改进人脸检测, 但是它们大部分维度较高, 增加了计算复杂度. 文献[7]对稀疏表示算法进行改进, 使用度量学习方法, 有助于挖掘出人脸特征在组间和组内之间的关系. 另外一种思路就是学习不同的级联结构, 检测多角度人脸, 比如平行级联结构[8], 金字塔框架[9], 宽度优先搜索树[10]等. 这些方法对于特定的人脸角度, 都需要学习一个级联分类器, 而准确标记每个人脸的角度比较困难, 工作量比较大.

近年来, 深度卷积神经网络 (CNN)[11]不断改进与发展, 在计算机视觉领域中占据着越来越重要的地位. 相较于传统手工设计的特征, CNN 特征能够克服复杂光照、部分遮挡、角度旋转的影响.

在人脸检测领域中, 也有很多基于深度学习的检

① 收稿时间: 2017-03-15; 修改时间: 2017-04-05; 采用时间: 2017-04-07

测算法被提出. 文献[12]使用深度卷积神经网络, 对大量的人脸与非人脸图像进行二进制分类训练. 训练好模型之后, 将网络的最后一层全连接层转换为卷积层, 前向网络, 可以得到热力图, 用于定位人脸. 文献[13]构建了 5 个共享权值的 CNN 网络, 这几个网络分别用来提取头发、眼睛、鼻子、嘴巴、胡子的特征. 根据人脸的空间结构, 器官的相对位置, 来定位人脸. 即使图像存在严重遮挡的情况, 该方法仍能检测出人脸. 文献[14]提出了基于 CNN 的级联网络, 滑动窗口先通过较小的 12-net 卷积神经网络, 快速排除大部分背景区域, 然后使用矫正网络微调检测窗口, 接着级联 24-net 卷积神经网络, 使检测结果更加精确. 该方法在单个 CPU 上的速度是 14 fps, 检测准确性较高. 文献[15]提出基于候选区域的卷积神经网络 (R-CNN), R-CNN 框架经过科研人员近几年的改进, 共发展了三个版本. 其中 Faster R-CNN[16]是最新的一个版本, 通过 RPN 网络回归计算出高质量的候选区域, 提高了检测效率, 在目标检测数据集上取得出色的成绩.

本文分别使用 ZF[17]和 VGG16[18]深度卷积神经网络, 在 WIDER[19]人脸数据集上训练 Faster R-CNN 模型, 并在 FDDB[20]人脸数据库上测试. 实验结果表明, 使用 Faster R-CNN 框架可以快速、准确地检测非限制性条件下的人脸. 其中使用 ZF 网络模型检测的速度更快, 而 VGG16 网络模型则更加精确.

2 基于候选区域的卷积神经网络发展概览

2.1 RCNN(Region based CNN)

基于候选区域的卷积神经网络目标检测算法, 由 Girshick 等人[15]首次提出. 图 1 展示了使用 RCNN 进行目标检测的流程, 主要分为三个步骤:

① 预先找出图像中目标可能出现的位置, 即候选区域. 常用的方法有 selective search[21]和 edge boxes[22].

② 将每个候选区域缩放到固定尺寸, 并输入到卷积神经网络中, 将 CNN 的全连接层输出作为特征.

③ 使用 SVM 进行分类, 并对提取到的窗口进行纠正的边框回归计算, 使检测得到的窗口跟目标真实窗口更加吻合.

该方法最大的优点, 是可以解决特征鲁棒性问题, 使用 CNN 特征进行分类有较高的准确率. 然而 RCNN 存在一些不可忽视的问题:

① 训练过程是多阶段的. 首先对卷积神经网络微调训练; 然后提取全连接层特征输作为 SVM 的输入, 训练得到目标检测器; 最后训练边框回归器.

② 训练过程需要耗费较多的空间与时间. 如图 1 所示, 提取候选区域使用 selective search 算法, 假设一张图像提取 2000 个候选区域, 就需要进行 2000 次卷积神经网络运算, 而每次提取的特征都需要写入磁盘空间, 这一过程需要耗费大量计算与存储资源.

③ 目标检测速度慢. 在 GPU 上使用 VGG16 网络进行目标检测, 需要花费 47 秒.

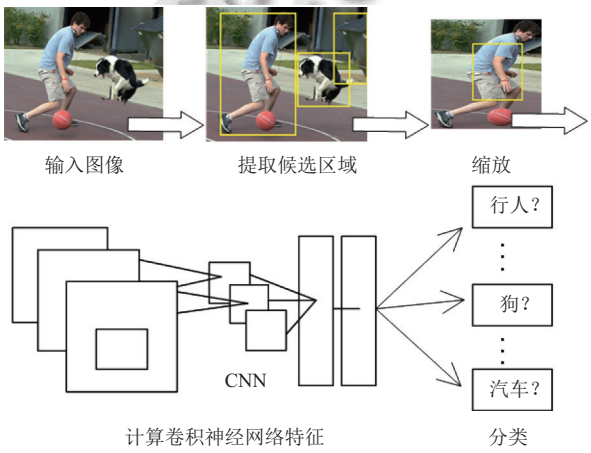

图 1 RCNN 目标检测流程框图

2.2 SPP-net(Spatial pyramid pooling)

2.1 中的 RCNN 方法, 每个候选区域都需要前向卷积神经网络来提取特征, 而每个候选区域都是整个图像的一部分. 因此完全可以对图像提取一次卷积特征, 然后只需将候选区域在原图的位置映射到卷积层特征图上, 这样对于一张图像, 只需要前向一次网络.

传统 CNN 网络固定了输入图像的尺寸, 在实际使用中, 需要对图像进行裁剪或者缩放预处理. 无论裁剪还是缩放, 都无法保证图像不失真. 比如裁剪可能会截断检测目标, 而缩放会拉伸物体, 失去“原形”. SPP-net[23]和 Fast R-CNN[24]先后被提出来解决上述问题.

如图 2 所示, SPP-net 最后一个卷积层的顶部, 连接到空间域金字塔池化层, 该池化层产生固定长度的输出, 用来输入到全连接层. 可以这样理解, SPP-net 将原先固定大小的池化窗口, 改成自适应大小, 每个候选区域使用不同大小的金字塔映射. 即使输入图像尺寸发生改变, 经过池化后的特征长度仍然保持一致. 例如当输入图像尺寸是 224×224 时, conv5 输出尺寸为 13×13×256, 对于 13×13 的激活图, 将它分别池化成

4×4、2×2、1×1 三张子图, 得到 (16+4+1)×256 维的特征向量. 当输入图像大小发生改变时, 池化窗口尺寸相应发生改变, 与激活图尺寸成正比, 使得池化后的特征向量长度不变.

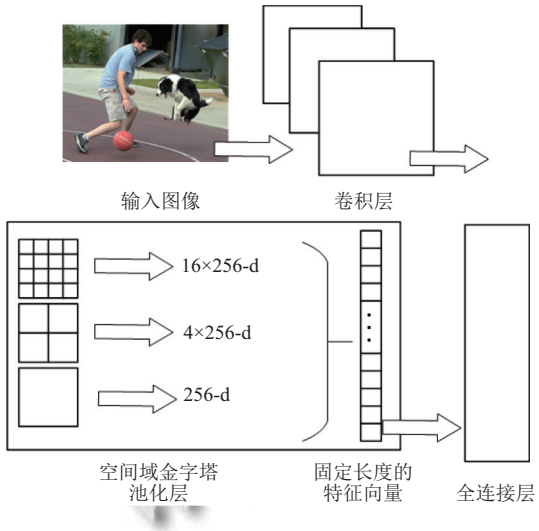

图 2 SPP-net 示意图

相对于 R-CNN 来说, SPP-net 可以大大加快目标检检测速度, 但是仍然存在一些问题:

① 训练分为多个阶段, 如微调网络, 训练 SVM, 训练边框回归器, 步骤较繁琐.

② SPP-net 微调网络时, 固定住卷积层参数, 只对全连接层进行微调训练. 对于新的分类或检测任务, 有必要对卷积层进行微调.

2.3 Fast RCNN

Fast R-CNN 融合了 R-CNN 和 SPP-NET 的精髓, 该方法目标检测流程如图 3 所示. 首先输入一张图像到 CNN 中, 得到卷积特征图; 接着对每个候选区域, 使用感兴趣区域 (ROI) 池化, 从卷积特征图中提取固定长度的特征向量; 然后将该特征向量输入到全连接层中, 这里有两个分支网络: 其中一个网络使用 softmax 对目标进行分类, 另一网络对坐标进行回归计算, 矫正边框位置.

这里 ROI 是指卷积特征图中的一个窗口, ROI 池化, 是将 ROI 卷积特征图转化为固定空间尺寸 (H×W), 该尺寸参数跟任何一个 ROI 都相互独立. 例如有一个 ROI 窗口尺寸为 h×w, 则需要划分 h/H×w/W 个子窗口, 然后在每个子窗口中使用最大池化方法, 得到对应的输出. ROI 池化层, 其实是 2.2 中空间域金字塔池化层的一种特殊情况, 只含有一个金字塔层. 在特征图的每个通道上, 池化操作之间都是相互独立的.

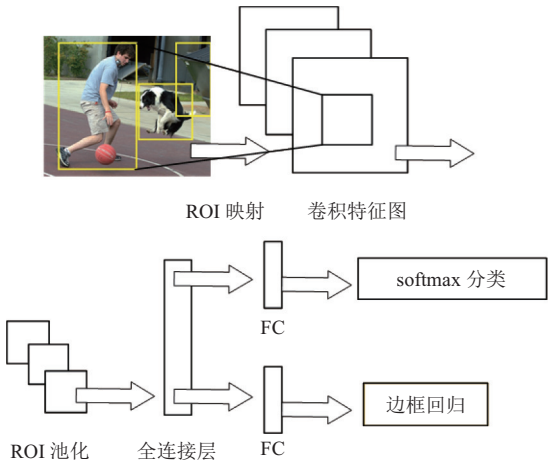

图 3 Fast R-CNN 目标检测流程框图

Fast R-CNN 网络含有 ROI 池化层, 用来提取固定长度的特征向量, 使用了 softmax 替代 SVM 分类, 同时将边框回归计算也加入到了网络中, 主要有以下几处优点:

① 训练使用多任务损失函数, 除了候选区域的提取, 其它训练过程都是端到端的.

② 在网络训练过程中, 将卷积层进行了微调, 取得更好的检测结果.

③ 不需要额外磁盘空间来存储特征向量.

然而 Fast R-CNN 方法仍然存在一个性能瓶颈, 检测时间大多消耗在候选区域的提取上, 无法满足实时应用. 比如使用 selective search 算法提取候选区域, 需要花费约 2 秒时间, 而特征分类只需要 0.3 秒.

2.4 Faster RCNN

Faster R-CNN 减少了提取候选区域的计算压力, 它主要分为两部分, 如图 4 所示. 第一部分是全卷积神经网络, 也称作 RPN(Region proposal network), 该网络用来产生候选区域; 第二部分是 Fast R-CNN 检测器, 使用第一部分网络产生的候选区域进行分类与边框回归计算. 整个系统共享卷积特征图, 将这两部分连接起来, 成为单一、统一的网络.

RPN 网络结构如图 5 所示, 任意尺寸图像输入到 RPN 中, 可以输出高质量的矩形候选区域集. 使用 RPN 网络在卷积特征图上滑动, 将特征图上 $n \times n$ 大小的窗口作为输入 (在文献[16]中, n=3), 后面分别连接两

个 1×1 的同级卷积层, 这两个全连接卷积层分别用作分类和回归.

对于每一个滑动窗口, 可以同时预测多个区域是否存在目标. 将滑动窗口的中心点作为锚点, 文献[16]分别使用 3 种缩放比和 3 种长宽比, 在每一个滑动窗口位置, 可以得到 k=9 个锚矩形框. 对于一个尺寸为 W×H 的特征图, 总共有 W×H×k 个锚矩形框. RPN 网络输出 2k 个是否存在目标的概率, 以及 4k 个回归坐标值.

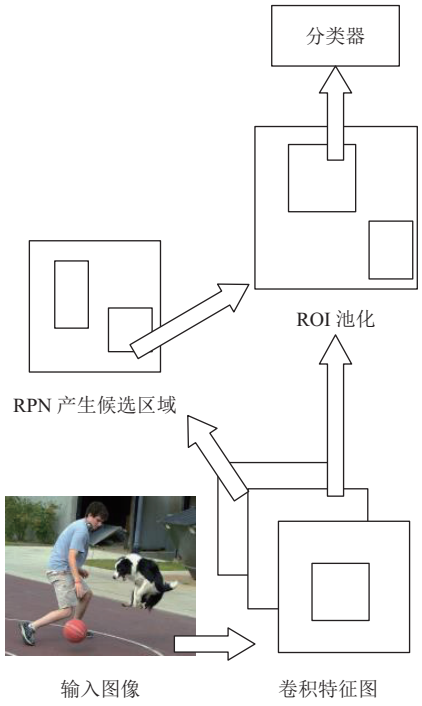

图 4 Faster R-CNN 目标检测流程框图

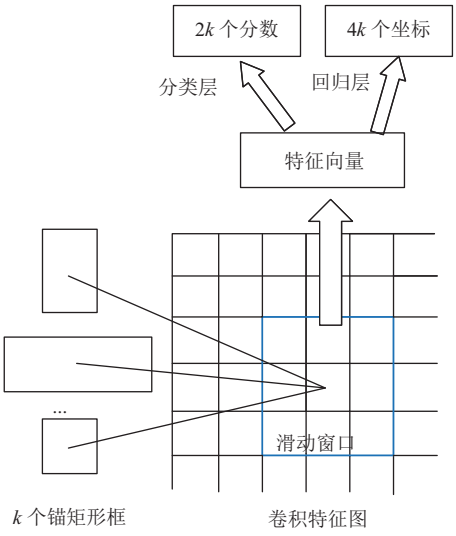

图 5 RPN 网络结构

Faster R-CNN 通过共享卷积层的方法, 将候选区域的提取和 CNN 分类结合在一起, 使用端到端的网络进行训练和测试, 速度和精度都得到不错的提高.

3 实验与分析

3.1 实验数据

本文在 WIDER 人脸数据库上训练 Faster R-CNN 模型, 该数据库有 12880 张图像, 共 15 多万张人脸, 均是在自然场景下拍摄. 如图 6 所示, 随机选取了 4 张图像, 图像中的人脸尺寸、姿态、光照等存在较大变化, 使该数据库更加具有挑战性.

图 6 WIDER 数据库测试图像

在 FDDB 人脸数据库上对训练好的模型进行评估. FDDB 是世界上最权威的人脸检测评估数据库之一, 该数据库包含 2845 张图像, 共含有 5171 个人脸, 每张图像均有人脸坐标的详细标签.

3.2 实验结果与分析

实验环境, 在 Inter® Core™ i7-2600 CPU 和 GTX1080 GPU 配置下实现. 我们分别使用两种卷积神经网络进行训练, 它们是 VGG16 和 ZF 网络. 本文使用在 ImageNet 上预训练得到的模型, 对网络权值进行初始化. 使用 SGD 算法 (随机梯度下降) 更新权值, 将初始学习率设置为 0.001, 经过 5 万次迭代后, 减少学习率为 0.0001, 继续迭代 3 万次结束训练. 图 6 中的矩形框, 是在 WIDER 测试集上进行人脸检测的结果.

我们在 FDDB 数据库上评估训练的模型, 共有两种评价指标, 分别是离散分数和连续分数. 离散分数, 是当检测的人脸区域与对应的人脸标签区域重叠部分超过 50% 时, 得分为 1, 否则为 0; 连续分数则是上述重叠的比率. 本文采用离散分数作为评价指标, 实验结果如图 7 所示. 使用 VGG16 网络训练的 Faster R-CNN

模型, 人脸检测的真阳率较高, 误检率相对 ZF 网络要低. 对于人脸检测效率, 使用 ZF 模型检测一张图像的平均时间为 0.094 秒, VGG 模型则需要 0.23 秒. 两种网络各有优势, 可以根据具体应用选择合适的网络模型.

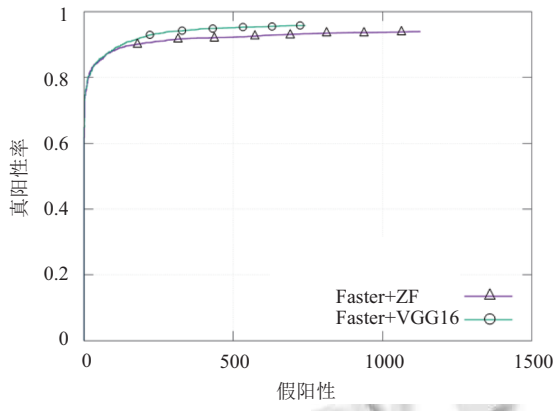

图 7 两种网络模型在 FDDB 上的测试结果比较

将本文实验结果与其它一些人脸检测方法进行比较, 如图 8 所示, 记录了各个人脸检测方法在 FDDB 上测试评估的 ROC 曲线. 除了 Viola-Jones 的经典人脸检测方法之外, 其它几种方法都是近几年提出来的, 其中 DDFD, Cascade CNN 以及 Joint Cascade 这三种方法, 都是使用卷积神经网络提取特征. 从图中可以看出, 当假阳性数量大于 200 左右时, 使用 Faster R-CNN 方法做人脸检测比其他方法表现得更好.

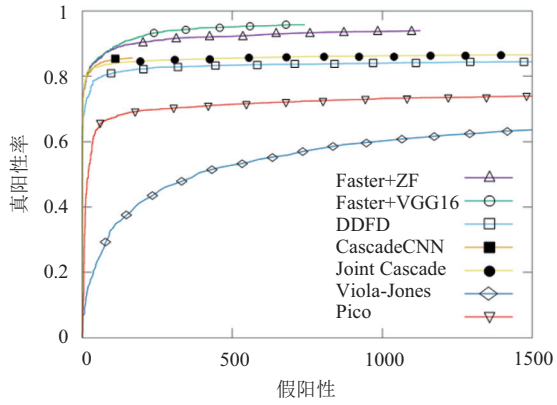

图 8 与其它人脸检测方法比较

随机从 FDDB 数据库中选取几张图像, 使用 Faster R-CNN 模型进行人脸检测, 效果如图 9 所示.

4 结语

Faster R-CNN 本来是用作通用目标检测的, 本文在 WIDER 人脸数据库上分别训练了 ZF 和 VGG16 网络模型, 并在 FDDB 数据集上测试. 实验结果表明, 该模型可以快速、准确地检测自然条件下的人脸, 对复杂光照、部分遮挡、人脸姿态变化具有鲁棒性. Faster R-CNN 最大的优点, 是共享了 RPN 和 Fast R-CNN 网络的卷积层, 大量减少了计算候选区域消耗的时间.

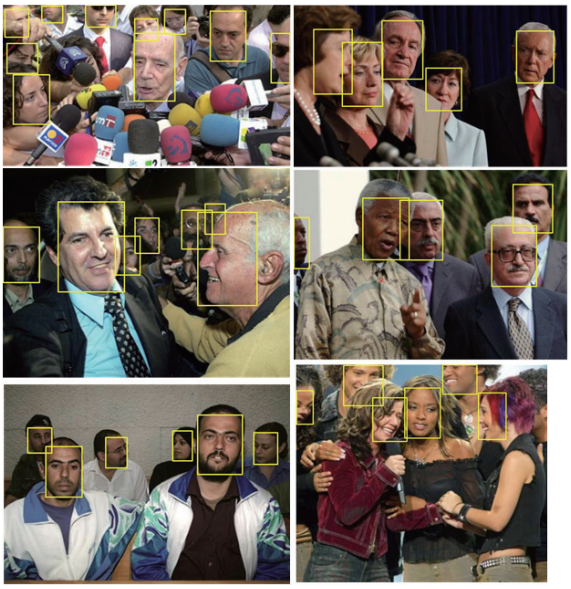

图 9 随机选取图像做人脸检测

展望未来, 发现有一些应用, 需要进行人脸检测, 并识别人脸生物特征, 比如性别、年龄、表情等. 可以借鉴 Faster R-CNN 共享卷积层的思想, 使用一个网络, 对大量人脸生物特征进行识别, 来提高效率.

参考文献

1 Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, HI, USA. 2001, 1. I-511–I-518.

2 Lienhart R, Maydt J. An extended set of haar-like features for rapid object detection. Proc. of IEEE International Conference on Image Processing. Rochester, NY, USA. 2002, 1. I-900–I-903.

3 Li SZ, Zhu L, Zhang ZQ, *et al*. Statistical learning of multi-view face detection. Proc. of the 7th European Conference on Computer Vision. London, UK. 2002. 67–81.

4 Jones M, Viola P. Fast multi-view face detection. Mitsubishi Electric Research Lab. TR-20003-96. 2003, 3. 14.

5 Froba B, Ernst A. Face detection with the modified census transform. Proc. of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition. Seoul, South

Korea. 2004. 91–96.
6 Jin HL, Liu QS, Lu HQ, *et al*. Face detection using improved LBP under Bayesian framework. Proc. of the Third International Conference on Image and Graphics (ICIG'04). Hong Kong, China. 2004. 306–309.
7 Shao H, Chen S, Zhao JY, *et al*. Face recognition based on subset selection via metric learning on manifold. Frontiers of Information Technology & Electronic Engineering, 2015, 16(12): 1046–1058.
8 Wu B, Ai HZ, Huang C, *et al*. Fast rotation invariant multi-view face detection based on real adaboost. Proc. of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition. Seoul, Korea. 2004. 79–84.
9 Li SZ, Zhang ZQ. Floatboost learning and statistical face detection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2004, 26(9): 1112–1123. [doi: 10.1109/TPAMI.2004.68]
10 Huang C, Ai HZ, Li YZ, *et al*. High-performance rotation invariant multiview face detection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2007, 29(4): 671–686. [doi: 10.1109/TPAMI.2007.1011]
11 Li QF, Zhou XF, Gu AH, *et al*. Nuclear norm regularized convolutional Max Pos@Top machine. Neural Computing and Applications, 2016: 1–10.
12 Farfade SS, Saberian MJ, Li LJ. Multi-view face detection using deep convolutional neural networks. Proc. of the 5th ACM on International Conference on Multimedia Retrieval. Shanghai, China. 2015. 643–650.
13 Yang S, Luo P, Loy CC, *et al*. From facial parts responses to face detection: A deep learning approach. Proc. of the IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 3676–3684.
14 Li HX, Lin Z, Shen XH, *et al*. A convolutional neural network cascade for face detection. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 5325–5334.
15 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 580–587.
16 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems. Montréal, Canada. 2015. 91–99.
17 Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. Proc. of the 13th European Conference on Computer Vision. Cham, Germany. 2014. 818–833.
18 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014.
19 Yang S, Luo P, Loy CC, *et al*. Wider face: A face detection benchmark. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 5525–5533.
20 Jain V, Learned-Miller EG. FDDB: A benchmark for face detection in unconstrained settings. UMass Amherst Technical Report. Amherst: University of Massachusetts, 2010.
21 Uijlings JRR, van de Sande KEA, Gevers T, *et al*. Selective search for object recognition. International Journal of Computer Vision, 2013, 104(2): 154–171. [doi: 10.1007/s11263-013-0620-5]
22 Zitnick CL, Dollar P. Edge boxes: Locating object proposals from edges. Proc. of the 13th European Conference on Computer Vision. Cham, Germany. 2014. 391–405.
23 He KM, Zhang XY, Ren SQ, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition. Proc. of the 13th European Conference on Computer Vision. Cham, Germany. 2014. 346–361.
24 Girshick R. Fast R-CNN. Proc. of the IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 1440–1448.