doi:10.3772/j.issn.1002-0470.2018.04.004

基于混合卷积神经网络的人头检测方法①

吉训生② 吴 凡

(江南大学物联网工程学院 无锡 214122)

摘 要 考虑到行人检测是视频监控领域的一项重要技术,其检测效果易受遮挡严重、光照不均等因素的影响,而人头检测是行人检测的重要研究内容,本文提出了一种基于混合卷积神经网络的人头检测方法。该方法将快速区域卷积神经网络(CNN)架构引入到局部模型的构建中,可以更好地获取图像的上下文信息,以得到更好的检测效果。通过全局模型预测头部的位置和尺度,利用成对模型确定待测目标间的成对关系。最后将局部、全局和成对模型融合成一个混合卷积神经网络框架,进行人头检测。研究结果表明,网络结构优化后的模型比多卷积神经网络方法在实时性显著提高 52.3 倍的同时,还可以将检测精度提高 1.8%,计算复杂度和内存消耗也大大降低。

关键词 图像处理,行人检测,人头检测,上下文,卷积神经网络(CNN),迁移学习

0 引 言

视频监控是获取图像信息的比较直接的方式,行人检测是视频监控系统的关键技术之一,目前已被广泛应用于商场、学校、银行、车站等公共设施。受光照条件、摄像头位置、背景变化、视角、行人相互遮挡等因素的影响,检测效果受到一定的限制。人头检测是行人检测的一个研究内容,检测人头相较于检测全身,受到这些因素的干扰较小。

目前已有很多人头检测方法。其中有顾炯[1]使用的以方向梯度直方图(HOG)作为重点描述特征构建基于统计学习方法的人头分类器,该方法将线性支持向量机(support vector machine,SVM)作为每个 HOG 特征的基础分类模式,通过 AdaBoost 算法提高 SVM 的分类效果。景文博[2]等优化的基于深度信息的人头检测方法,该方法通过双目摄像头获取整个图像的深度信息和强纹理点,确定人头区域。Aziz Kheireddin[3]等通过标记并更新每个检测到的骨骼轮廓的方式来确定人头的方法,在拥挤环境下该方法有很好的鲁棒性,但计算量较大;基于深度信息和骨骼轮廓的方法,它需要获得 RGB-D(Depth)图,实际应用有一定难度。近年来,基于 RGB 图的方法通过引入上下文信息,以获得更好的特征表示。上下文信息是利用目标内部及邻域、目标间及所处场景等各种类型的上下文信息,丰富目标的信息表达,一定程度上提高了目标检测的准确性[4]。如获得 2011 年 VOC 目标检测算法的第 1 名的利用局部上下文的(deformation part model,DPM)[5]方法,获得 2014ILSVRC 第 2 名的基于深度模型的 DeepID-Net[6]。上下文信息可以进行全局场景建模,如使用上下文预测目标位置,使用全局场景信息加速目标检测。

众所周知,在引入深度学习后,人脸检测效果得到了大幅提高。卷积神经网络(convolution renral network, CNN)近年来已广泛用于目标检测。其中,Girshick 在 2014 年提出的 R-CNN 在目标检测中已取得巨大成功[7]。fast R-CNN[8]、faster R-CNN[9]、

① 国家自然科学基金(61771223)和江苏省前瞻性联合研究(BY2016022 - 28)资助项目。
② 男,1969 年生,博士,教授;研究方向:信号处理等;联系人,E-mail: jixunsheng@163.com
(收稿日期:2017-11-16)

mask R-CNN[10] 等方法的提出，又把卷积神经网络在目标检测中的应用推向一个新的高度，也给目标检测的研究提供了新的思路。传统方法多是采用人工设计特征，卷积神经网络方法在输入图像后直接获得卷积特征，卷积特征相比人工设计特征有更好的检测效果。人头检测属于目标检测的一类，Vu[11] 等提出结合上下文信息和卷积神经网络进行人头检测的方法，通过将 RCNN 与 CNN 获得得分参数交叉验证，构建关联得分函数，相比 RCNN 单模型，使检测精度提升。

本文基于 Vu[11] 的方法，在使用较新的卷积神经网络模型的局部模型完成初始检测任务后，结合上下文信息进行扩展。通过引入全局模型和成对模型，降低了头部目标检测受闭塞因素的影响，最后将三者融合为混合模型。为了提高检测效果，论文还为图像中的多个目标假设建立联合得分函数。得分函数的所有参数通过优化后的结构化输出损失函数进行学习，对 CNN 模型的网络构建进行优化，使用参数化修正线性单元（parametric rectified linear units，PReLU）作为激活函数[12]。

本文中的训练集和测试集采用 Vu[11] 等在 Hollywood 数据集基础上构建的 HollywoodHeads 数据集和 Casablanca 数据集。

1 模型的构建

混合卷积神经网络主要包括局部、全局和成对模型。将数据直接输入到局部和全局模型，成对模型的输入为局部模型输出的通过非极大值抑制（non-maximumsuppression，NMS）的候选分数，最后将三者得到的分数进行组合。其结构如图 1 所示。

1.1 局部模型

局部模型使用选择性搜索提议（selective search）方法[13] 来获取头部的对象提案，对提案适当扩展以获取头部附近的局部上下文信息。为节省训练和参数优化时间，CNN 模型采用迁移学习的方法，即初始化已在大型数据集训练过的预训练网络，在其基础上进行网络优化。

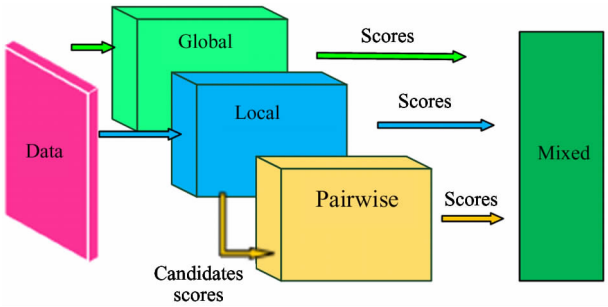

图 1 混合模型示意图

在对比了 MatConvnet[14] 提供的 ImageNet 上的 AlexNet、VGG-verydeep-16、VGG-verydeep-19 网络和由 Vu[11] 提供的 Oquab 网络后发现，VGG-verydeep-19 网络的训练效果最好，但该网络过深，速度很慢，本文将在 Oquab 对 AlexNet 网络微调的基础上，还将对该网络进行进一步调整。

在构建局部模型的过程中，在初始化 Oquab 的预训练模型后，用 RoI（region of interest）Pooling[8] 替代原先的第五层的 Pooling，加上 2048 个节点的全连接层，最后利用 Softmax 损失层对头部和背景进行分类，随机初始化后加上边界框回归（bounding box regression）层进行候选区域位置调整。网络结构如图 2 所示。

其中 RoI 池化层构建的目的是提取固定维度的特征表示，并将图像中 RoI 对应于特征图的权值[15]。感兴趣区域池化层如图 3 所示。若感兴趣区域池化层固定空间幅度 $H \times W$ 为 7×7，输入候选用 (r,c,h,w) 表示，(r,c) 代表感兴趣区域左上角坐标，(h,w) 为高度和宽度。感兴趣区域池化层将 $h \times w$ 感兴趣区域窗口分割成 $r \times c \times (h/7) \times (w/7)$ 块，然后对每块进行 maxpooling，便可输出同一位 7×7 的特征映射。

在该网络中，激活函数采用 PReLU，其对应的函数形式如下式所示：

$$f(y_i) = \begin{cases} y_i & (y_i > 0) \\ a_i y_i & (y_i < 0) \end{cases} \tag{1}$$

其中，i 代表不同的通道，PReLU 在负数区域不是连续的，且是可学习的，PReLU 数学表达为

$$y_i = \max(0, x_i) + a_i \times \min(0, x_i) \tag{2}$$

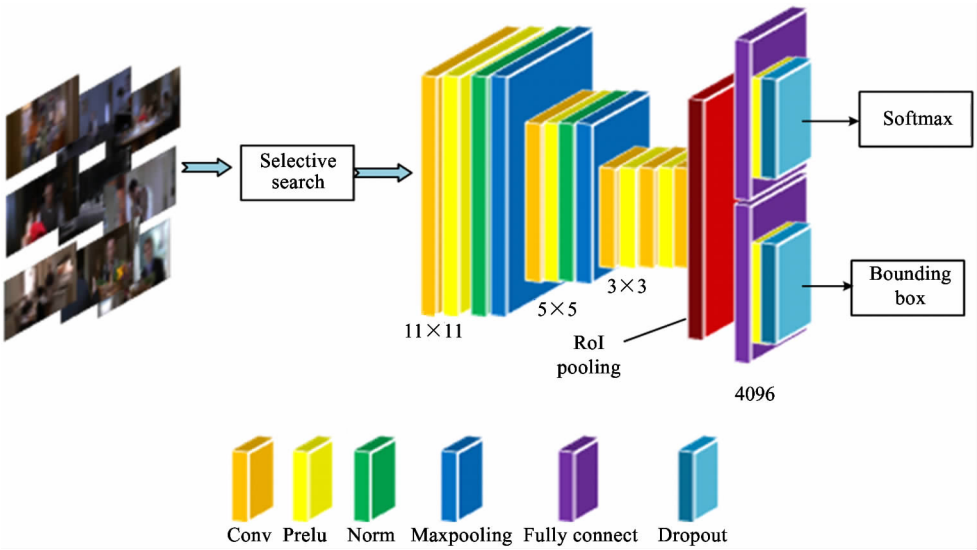

图 2　局部模型的网络结构

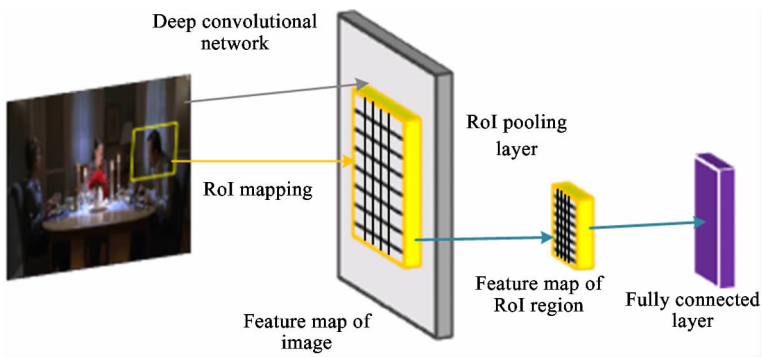

图 3　感兴趣区域池化层

其中 $a_i = 0$ 时,即为常见的 ReLU。a_i 赋值为固定值,即为 Leaky ReLU。而在局部模型网络中 a_i 的更新方式采用下式:

$$\Delta a_i = \mu \Delta a_i + \varepsilon \frac{\partial \varepsilon}{\partial a_i} \tag{3}$$

式中,μ 为动量,ε 为学习率,a_i 初始化为 0.2。因为检测目标为头部,边界框长宽比确定为 $R \in [\frac{2}{3}, \frac{3}{2}]$,将它们作为候选。其中,候选边界框与真实值边界框的交叉联合(intersection over union,IoU)重叠率大于 0.6 为正(头部),小于 0.5 为负(背景)。

在网络训练过程中,考虑到网络中有较多的连续,且梯度较小。为了减小每次梯度计算过程中噪声的影响,使用动量为 0.9、学习率为 0.01、权重衰减为 0.005 的带动量的随机梯度下降(stochastic gradient descent, SGD)来最小化独立对数损失的和值,以优化网络参数。

1.2　全局模型

全局 CNN 模型是在输入上给出完整的低分辨率图像以及训练预测头部的粗略位置和尺度后,使用图像的所有像素进行预测。为了使整个模型可以提供精确头部的位置和尺度的定位,之后引入一个成对的 CNN 模型,建立头部目标间的位置和尺度关系。

全局模型的网络架构和局部模型基本相同,但模型的输入为整个图像,输出为多尺度图的每个单元格的分数。全局模型的网络结构如图 4 所示。

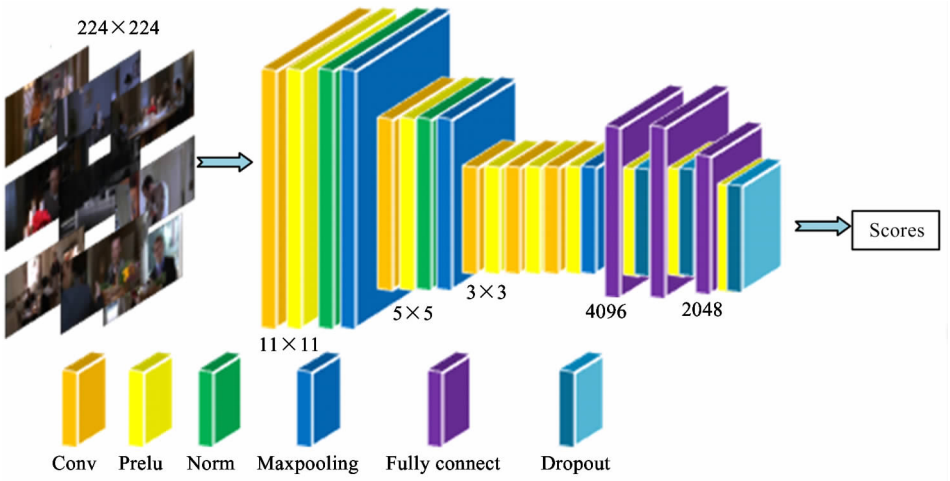

图 4　全局模型网络结构

由于网络的输入要求为 224 ×224 的完整图像，输入图像需进行各向同性缩放。网络的输出为一个多尺度的分数网格，主要为图像中目标假设的粗离散位置和尺度。目标假设的形式为 284 个方形单元格 C 构成的网格，方形单元格包含 28 ×28、56 ×56、112 ×112 和 224 ×224 四个尺寸，目标假设步幅为单元格大小的一半。

全局模型经过 SGD 训练，学习率0.0001，动量 0.9，权重衰减0.005。最小化网格单元 C 对数损失函数的和定义为

$$l(f_c(x), y_c) = \sum_{y\in\{0,1\}} \log(1 + \exp((-1)^{y_c+y+1} f_{c,y}(x))) \quad c \in \{1,\cdots,C\} \tag{4}$$

其中，$f_c(x) \in R^2$ 是输入图像 x 的网格单元网络 c 的输出，$y_c \in \{0,1\}$ 是网格单元格 c 的标签：背景或头。如果图像 x 中的单元格和真实值边界框的 IoU 重叠率大于 0.3，则网格单元的标签将被标记为头部，反之为背景。

因为网格单元解析度不高，全局模型不能准确定位，全局模型用来重新确定局部和成对模型的候选。

1.3　成对模型

成对模型的构建目标是提取多个头部候选。通过局部模型中选择性搜索从输入图像中获得候选区域，根据局部模型提供的分数，使用 NMS 选择候选，NMS 的阈值为 0.3。在重组候选特征后，通过单点网络和成对点网络计算其单点势能和成对点势能，最后使用结构化损失训练参数。其主要流程如图 5 所示。

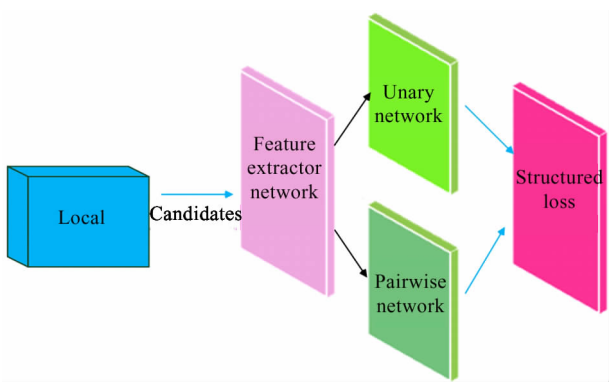

图 5　成对模型的构建过程

1.3.1　成对模型构建

成对模型构建参考了 Vu[11] 模型，构建中所使用的联合得分函数通过随机梯度下降算法最小化结构化代理损失对模型中的参数进行训练。联合得分函数 $S(y;\omega)$ 其目的是将相同图像中的候选标签建立如下式所示的关系：

$$S(y;\omega) = \sum_{i\in V} \theta_i^P(y_i;\omega) + \sum_{(i,j)\in\varepsilon} \theta_{ij}^P(y_i, y_j; k_{ij};\omega) \tag{5}$$

式(5)中，V 是从图像中提取的一组候选边界框，定义每个边界框为一个二进制变量 y_i，$i \in V$。目标类标签为 1，背景类标签为 0。$\hat{y}_i$ 是在训练图像中所有候选的真实值标签，ε 表示定向的成对候选，$k_{ij} \in \{1,\cdots,K\}$ 表示边缘 $(i,j) \in \varepsilon$ 的聚类指数，ω 表示可训练参数，θ_i^U 和 θ_{ij}^P 是取决于 ω 的单点和成对

点势能,$y = (y_i)(i \in \nu)$ 是所有二进制变量的向量。

1.3.2 成对模型训练

模型中参数更新的步骤包括:

(1)使用 NMS 方法,用局部模型生成的分数来选择一组候选;

(2)计算联合评分函数的势能,对模型中的参数进行正向传递;

(3)计算结构化损失,进而计算出梯度;

(4)在模型中对梯度进行逆向传递。

其中结构化代理损失就是将参数的当前值、图像数据 $x = (x_i)_{i \in V}$ 和真实值标签 $\hat{y} = (\hat{y}_i) i \in V$ 映射为实数。这个损失函数为

$$l(w,\hat{y},x) = \sum_{i:\hat{y}_i=1} v(s_i(w,x)) + \sum_{i:\hat{y}_i=0} v(-s_i(w,x)) \tag{6}$$

其中,v 是有上边界的任意非递增函数,使用 $v(t) = \log(1+\exp(-t))$ 使其更接近于传统检测器训练的 Softmax 函数。结合卷积神经网络的检测任务,类别判断更倾向于根据得分的概率分布。在每个得分计算得到的概率基础上,使概率分布更接近标准。

使用随机梯度下降算法最小化结构化代理损失来训练成对模型的参数,使用标准差为 0.01 的零均值高斯初始化单点网络和成对点网络的权重,SGD 的动量设定为0.9,权重0.0005,学习率0.00001。另外,考虑到计算机显存容量,本文中的每个图像使用 16 个候选。

最后先将局部模型和成对模型所得到的分数 s_l、s_p 组合,$s_{lp} = as_{\text{local}} + (1-a)s_{\text{pairwise}} + b$,其中 $a \in [0,1]$ 和 $b \in [-10,10]$。混合 CNNs 的分数由全局模型得分和 s_{lp} 组合而成 $s_{\text{mixed}} = cs_{lp} + (1-c)s_{\text{global}}$,$c \in [0,1]$。参数 a、b、c 是通过在验证集上最大化平均精度选择。

2 实 验

本文使用的数据集是 Vu[11] 等制作的 HollywoodHeads 头部检测数据集[9] 和 Casablanca 数据集。HollywoodHeads 数据集为来自于 21 部好莱坞电影中 224740 帧的 369846 个头部注释。训练集使用 HollywoodHeads 数据集其中 15 部电影的 216719 帧图像,验证集为 3 部电影中的 6719 帧,测试集是另外 3 部电影的 1302 帧。Casablanca 数据集为黑白电影 Casablanca 的 1466 个视频帧。

为了评估检测的性能,使用基于精度召回率(precision-recall, PR)曲线的标准平均精度(average precision, AP)度量[16]。与真实值有高重叠率(IoU >0.5)的检测被认为是正确的。若多次检测到同一个真实值,则被认为是错误的。

本文对比了 Vu[11] 使用的方法、基于 R-CNN 的目标检测器[7](R-CNN)、基于 DPM 的面检测器[17](DPM Face)和 VJ-CRF[18] 的基准。R-CNN 使用 HollywoodHeads 数据集的训练子集训练人头上的 R-CNN 目标检测器。CNN 模式首先针对用于培训局部模式的所有区域提案,并进行微调。受内存限制,R-CNN 训练的 SVM 阶段是在一组训练图像上完成。对于基于 DPM 的面部检测器,使用香草 DPM 模型[17]。测试了不同方法在 HollywoodHeads 和 Casablanca 数据集的表现,结果如图 6 和图 7 所示。图 6 为 HollywoodHeads 数据集,图 7 为 Casablanca 数据集。

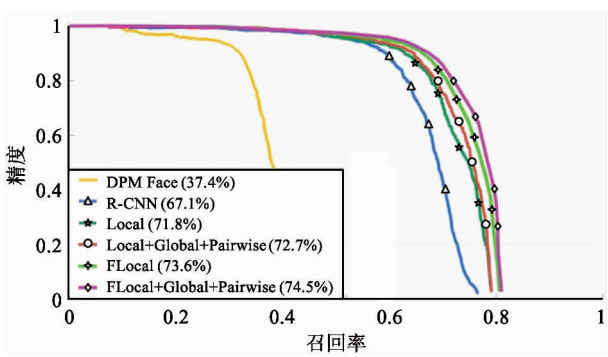

图 6 HollywoodHeads 数据集对比结果

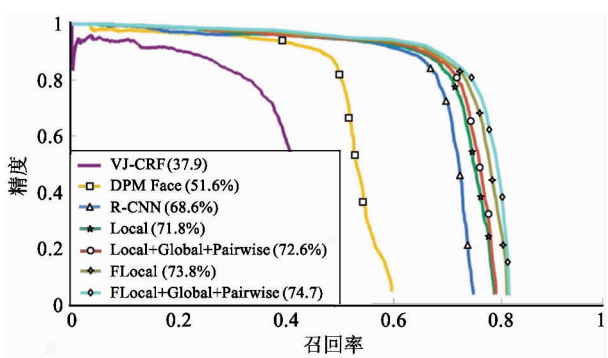

图 7 Casablanca 数据集对比结果

可以看出,基于 CNN 的方法相比于传统方法有着明显的优势。由于 Casablanca 数据集相比于 HollywoodHeads 数据集面部较多,DPM 面部检测方法召回率较高。本文方法在 HollywoodHeads 数据集和 Casablanca 数据集相比于 Vu[11] 方法的召回率明显提升。不同模型组合的 AP 如表 1 所示。另外,本文使用的 PReLU 比 ReLU 在精度上仅提高0.1%。

表 1 不同混合模型在 HollywoodHeads 和 Casablanca 上的 AP

Test set	FLocal	Pairwise	FLocal + Global	FLocal + Pairwise	FLocal + Global + Pairwise
HollywoodHeads	73.6	72.4	74.3	73.9	74.5
Casablanca	73.8	73.5	74.1	74.5	74.7

本文方法相较于 Vu[11] 在局部模型上训练速度上有 4.51 倍的提高,而在测试速度上有着 52.33 倍的提高。其实时性如表 2 所示。通过将分类和位置调整放入网络中实现,相比于 Vu[11] 的模型产生近 39GB 的数据存储,本文改进后的数据存储只有 8GB。

表 2 改进前后的实时性对比

平均值	局部模型 (Local)	改进后局部的模型 (FLocal)
训练速度 (images/s)	292	1316
测试速度 (patches/s)	1120	58610

本文实验基于 Matlab 的 MatConvnet[14] 框架进行。硬件配置:NVIDIA GPU 1060(6G)。由于显存限制,训练过程对训练数据进行了裁剪,在成对模型中运行得分函数限制簇(Patch)最高为 16。

3 结 论

在人头检测工作中,本文优化了混合 CNN 模型,改进了局部模型和成对模型的损失函数,引入新的激活函数。研究表明,改进后的局部模型在一定程度上实现了端到端的目标检测。在大型 HollywoodHeads 数据集上的仿真结果表明:改进后的算法可以将处理过程的实时性提高52.3倍,检测精度提高了1.8%,显著降低计算复杂度和存储消耗。这项工作表明,将优秀的目标分类模型迁移学习至人头检测,并结合上下文信息,可取得良好的检测效果。

参考文献

[1] 顾炯. 基于头肩轮廓特征的人头检测系统的研究:[硕士学位论文][D]. 上海:东华大学信息科学与技术学院,2012

[2] 张姗姗, 景文博, 刘学. 一种基于深度信息的人头检测方法[J]. 长春理工大学学报(自然科学版), 2016, 39(2):107-111

[3] Aziz K, Merad D, Iguernaissi R, et al. Head detection based on skeleton graph method for counting people in crowded environments [J]. *Journal of Electronic Imaging*, 2016, 25(1):013012

[4] 李涛. 基于上下文的目标检测研究:[博士学位论文][D]. 成都:电子科技大学机器人研究中心, 2016

[5] Felzenszwalb P F, Girshick R B, Mc Allester D, et al. Object detection with discriminatively trained part-based models[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2010, 32(9): 1627-1645

[6] Ouyang W, Wang X, Zeng X, et al. DeepID-net: deformable deep convolutional neural networks for object detection[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015. 2403-2412

[7] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014. 580-587

[8] Girshick R. Fast R-CNN[C]. In: Proceedings of the

IEEE International Conference on Computer Vision, Santiago, Chile, 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, Dollár P, et al. Mask R-CNN[J]. arXiv preprint arXiv:1703.06870, 2017

[11] Vu T H, Osokin A, Laptev I. Context-aware CNNs for person head detection[C]. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 2893-2901

[12] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015. 1026-1034

[13] Van de Sande K E A, Uijlings J R R, Gevers T, et al. Segmentation as selective search for object recognition [C]. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 2011. 1879-1886

[14] Vedaldi A, Lenc K. Matconvnet: convolutional neural networks for matlab[C]. In: Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australlia, ACM, 2015. 689-692

[15] 叶国林, 孙韶媛, 高凯珺. 基于加速区域卷积神经网络的夜间行人检测研究[J]. 激光与光电子学进展, 2017, 54(8):117-123

[16] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (VOC) challenge[J]. *International Journal of Computer Vision*, 2010, 88(2): 303-338

[17] Mathias M, Benenson R, Pedersoli M, et al. Face detection without bells and whistles[C]. In: Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, 2014. 720-735

[18] Viola P, Jones M J. Robust real-time face detection[J]. *International Journal of Computer Vision*, 2004, 57(2): 137-154

Head detection using hybrid convolution neural networks

Ji Xunsheng, Wu Fan

(School of Internet of Things Engineering, Jiangnan University, Wuxi 214122)

Abstract

Based on the consideration that pedestrian detection is on important technique in video surveillance, and its detection effect is easy to be affected by serious occlusion, uneven illumination, etc., while human head detection is an important part of pedestrian detection, a human head detection method based on hybrid convolution neural networks (CNNs) is proposed. The method introduces fast regional convolutional neural network architecture into the construction of the local model for obtaining context image information for detecting person better. The global model is built to predict the position and scale of the head. The pairwise model is used to get the pairwise from objectives. At last, the local, global and pairwise models are fused into a federated CNN framework in order to detect head. Compared with the context CNNs, the research achievement shows that the hybrid CNNs can improve the real-time and average precision 52.3 times and 1.8 percent individually. Computational complexity and memory consumption can be reduced significantly.

Key words: image processing, pedestrian detection, head detection, context, convolution neural network (CNN), transfer learning