基于多特征信息融合的目标检测方法研究

丁 哲 陆文总* 闫芬婷
(西安工业大学电子信息工程学院 陕西 西安 710021)

摘 要 针对目标检测中小目标漏检、准确率较低和容易受到相似目标干扰等问题,在 SSD(Single Shot MultiBox Detector)模型基础上,提出一种融合低层手工特征和深层网络特征的目标检测方法。通过对图像提取融合多种目标特征,获取目标大概位置和类别。基于非极大值抑制(NMS)将交并比小于 0.7 的窗口消除,解决行人部分重叠与小目标的漏检问题,提升目标检测的准确性和目标检测模型的泛化能力。该模型在 VOC2007 公开数据集上的平均检测精度较 SSD 算法提升了 4%,NMS 机制的加入有效提升了目标检测速度和稳定性。

关键词 多特征融合 目标检测 SSD 算法 非极大值抑制

中图分类号 TP391.9 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.11.021

TARGET DETECTION METHOD BASED ON MULTI-FEATURE INFORMATION FUSION

Ding Zhe Lu Wenzong* Yan Fenting
(*School of Electronic Information Engineering, Xi'an Technology University, Xi'an 710021, Shaanxi, China*)

Abstract Aiming at the problems of small target missed detection, low accuracy and vulnerable to similar target interference, based on single shot multi-Box detector(SSD) model, we propose a target detection method combining low-level manual features and method for deep network features. The variety of target features were integrated through the image extraction, and the approximate location and category of the target were obtained. We eliminated the window with less than 0.7 ratio based on non-maximum value suppression(NMS), which solved the problem of pedestrian overlap and small target omission, and improved the accuracy of target detection and generalization of target detection model. The average detection accuracy of our model on VOC2007 public dataset is 4% higher than that of SSD algorithm, and NMS mechanism effectively improves the speed and stability of target detection.

Keywords Feature fusion Target detection SSD algorithm Non-maximum suppression

0 引 言

随着计算机视觉技术的蓬勃发展,目标检测已然成为当前的研究热点,人们对目标检测技术也提出了更高的需求。虽然该领域已研究了数十年,但仍然存在行人目标背景复杂、形态不一、目标相互遮挡和检测精度有待提高等问题。因此,研发一种准确率高且鲁棒性强的行人检测算法具有重要意义。

2005 年 Dalal 等[1]将 HOG 特征图像局部变化与 HOG-LBP 特征相结合,并采用 SVM 分类器配合对目标进行检测。近年来基于深度学习的目标检测方法层出不穷,2014 年 Girshick 等[2]提出了基于区域的卷积神经网络 R-CNN,但计算候选框的耗时较大,实时性难以满足。2015 年 Redmon 等[3]提出了 YOLO 算法,采用一个单独的卷积神经网络模型实现端到端的目标检测,检测速度有所提升,但对小目标检测效果不好。2016 年 Liu 等[4]提出了 SSD(Single Shot MultiBox Detector)算法,其结合了 Faster R-CNN 算法[5]和 SSD 算法的优势,在检测精度和实时性方面均有一定的突破。

本文采用 SSD 算法作为基础检测框架,为了弥补卷积神经网络中难以学习到图像统计特征、边缘约

收稿日期:2019-06-25。陕西省自然科学基金项目(2016JM8095)。丁哲,硕士生,主研领域:目标检测与跟踪 。陆文总,副教授。闫芬婷,硕士生。

束弱等不足,针对卷积神经网络深层输出特征对目标分类不准确现象,提出一种将卷积神经网络浅层提取的特征与深度特征[6]融合的 SSD 检测方法,建立融合多特征的网络模型,有效地提高了行人检测的准确率。

1　多特征信息融合的目标检测模型

1.1　目标检测模型

多特征信息融合[7]的目标检测模型主要由基础网络部分、特征提取层部分、原始包围框生成部分和卷积预测部分组成。融入多特征的检测模型是在 VGG16[8]网络结构的基础上只增加了两层 3×3×256 的卷积层以满足目标尺度变化,而且较原 SSD 模型提升了实时性。本文在神经网络第二个卷积层后提取图像的方向梯度直方图(HOG)、RGB 颜色加权直方图和 LBP 纹理[9]加权直方图三种手工特征;同时在多个特征图上利用 Softmax 分类与位置回归,得到一系列固定大小的边界框和目标类别[10]的得分;最后根据非极大值抑制得到检测识别的结果。图 1 为 SSD 算法特征融合框图。

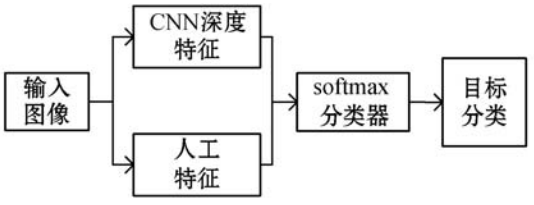

图 1　SSD 算法特征融合框图

1.2　目标检测先验框设置方法

融合多特征的 SSD 算法为每个单元设置长宽比和尺度不同的先验框,然后以这些先验框为基准预测边界框位置,降低模型训练的难度。在训练过程中,图像中的目标会根据先验框匹配原则在设置的多个先验框中挑选出最适合它们的先验框进行训练。先验框的设置主要为长宽比和大小两个方面,先验框的尺度呈线性递增,特征图的先验框和大小尺度也呈线性关系。

$$s_h = s_{\min} + \frac{s_{\max} - s_{\min}}{n-1}(h-1) \quad h = 1,2,\cdots,n \tag{1}$$

式中:n 为特征图的个数;s_h 为先验框大小相对图片的比例;$s_{\max}$为比例的最大值;$s_{\min}$为比例最小值。对于特定的长宽比,先验框实际的宽和高为:

$$x_h^a = s_h\sqrt{a_r} \quad y_h^a = \frac{s_h}{\sqrt{a_r}} \tag{2}$$

式中:a_r 为常数。

1.3　目标检测方法的损失函数

本文多特征信息融合的目标检测方法总的损失函数定义为定位损失(loc)和置信损失(conf)的加权和,即:

$$L(x,c,l,g) = \frac{1}{N}(L_{\text{conf}}(x,c) + \alpha L_{\text{loc}}(x,l,g)) \tag{3}$$

式中:N 为先验框的正样本数量;x 可表示为 $x_{i,j}^p = \{0,1\}$;c 为置信度;l 为预测框;g 为真实框;α 为权重系数;$L_{\text{conf}}(x,c)$为置信损失;$L_{\text{loc}}(x,l,g)$为定位损失。

$$L_{\text{conf}}(x,c) = -\sum_{i\in\text{pos}}^{N} x_{i,j}^p \log(\hat{c}_i^p) - \sum_{i\in\text{neg}} \log(\hat{c}_i^0) \tag{4}$$

$$\hat{c}_i^p = \frac{\exp(\hat{c}_i^p)}{\sum_p \exp(\hat{c}_i^p)}$$

$$L_{\text{loc}}(x,l,g) = \sum_{i\in\text{pos}}^{N}\sum_{m\in\{cx,xy,w,h\}} x_{ij}^k smooth_{L1}(l_i^m - \hat{g}_j^m) \tag{5}$$

$$\begin{cases}\hat{g}_j^{cx} = (g_j^{cx} - d_i^{cx})/d_i^w \\ \hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^h\end{cases} \qquad \begin{cases}\hat{g}_j^h = \log\left(\dfrac{g_j^h}{d_i^h}\right) \\ \hat{g}_j^w = \log\left(\dfrac{g_j^w}{d_i^w}\right)\end{cases}$$

式中:$smooth_{L1}(\cdot)$函数从两个方面限制梯度,当预测框与 ground truth 差别过大时,梯度不至于过大,当预测框与 ground truth 差别很小时,梯度值足够小;$(g_{cx}, g_{cy}, g_w, g_h)$表示预测包围框;$(d_{cx}, d_{cy}, d_w, d_h)$表示错误包围框;$(l_{cx}, l_{cy}, l_w, l_h)$表示预测的包围框相对于错误包围框的偏移量。

1.4　非极大值抑制(NMS)

对目标检测过程中,大量的候选框会在同一目标的位置产生,但候选框之间有大量重叠,从最大概率候选矩形框开始,分别判断候选框与目标真实包围框的交并比是否大于某一固定阈值,选择概率最大的目标边界框,将其他概率低的边界框消除掉。不断重复,找到所有被保留下来的包围框。检测窗口的重叠率 $p(\delta_1,\delta_2)$可表示为:

$$p(\delta_1,\delta_2) = \frac{\delta_1 \cap \delta_2}{\delta_1 \cup \delta_2} > \varphi \tag{6}$$

式中:δ_1和δ_2为两个检测窗口;将阈值 φ 设定为 0.7,将重叠率低于 0.7 的窗口消除,从而提高检测速度。

2　多特征融合方法

卷积神经网络中,基础网络用来提取输入图像的浅层特征和深层特征。其中,浅层特征直接用于目标检测与包围边框回归。考虑到卷积神经网络在迭代过

程中易出现梯度流失现象,损失网络提取到的有效特征信息会影响目标检测的准确性。本文利用 SSD 卷积神经网络将提取深层特征和浅层特征信息在网络中的 Flatten 层将其转化成一维向量进行融合,在浅层卷积加入 RGB 颜色特征、方向梯度直方图(HOG)和局部二值模式(LBP)三种人工特征,图 2 为多特征信息融合框架。

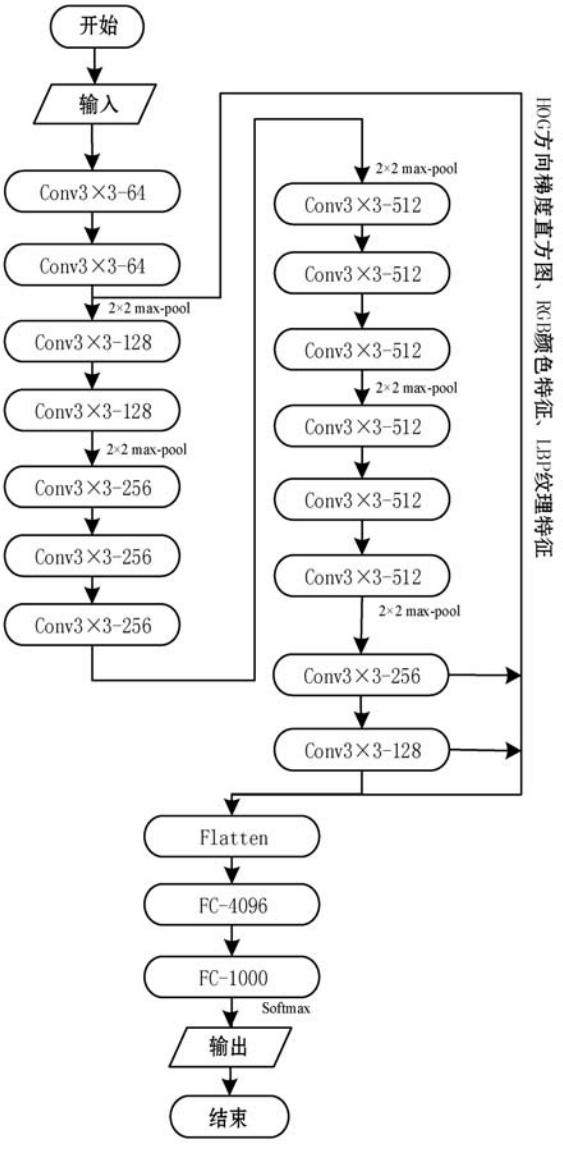

图 2 多特征信息融合框架

该多特征信息融合框架以 VGG16 为基础网络模型,是神经网络由浅到深不断迭代的过程,整个框架共 17 层,在第二个卷积层之后提取图像目标的三种人工特征,最后在网络的 Flatten 层融合人工特征和深度特征并对其分类、回归。

2.1 浅层特征提取

通过计算和统计图像局部区域的梯度方向直方图来提取图像的方向梯度直方图(HOG)特征。HOG 对图像局部进行方格单元操作,因此对图像的几何形变具有良好的不变性,并且在较强的局部光学归一化、精细的方向抽样与粗的空域抽样条件下,只需要行人保持直立的姿势,行人微小的肢体动作不会影响检测效果,能够很好地对运动行人目标进行描述。行人目标提取 HOG 特征如图 3 所示。

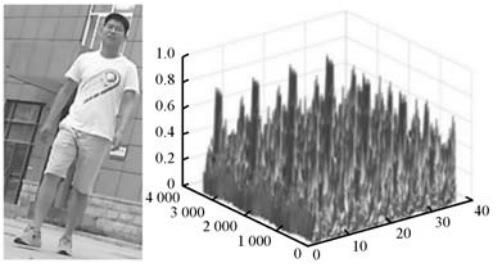

图 3 HOG 特征图

颜色特征是目标最直观的特征,提取该特征的速度快,而且有很好的区分度。RGB 颜色直方图是用来描述不同色彩在整幅图像中所占的比例,本文考虑到模板内颜色的分布情况,因此设目标区域有 n 个像素,其中心像素坐标为 u_0,像素集为 $\{x_1,x_2,\cdots,x_n\}$,目标模板在 R 分量的特征是 $\{\lambda=0,1,\cdots,L-1\}$,G 分量与 B 分量的特征均为 $\{\lambda=0,1,\cdots,(L-4)/4\}$,然后对每个 bin 内像素点统计,得到该区间像素点占整幅图像像素的比例。在统计图像过程中,离跟踪框中心越近的点对识别跟踪的贡献越大。图 4 为行人目标所对应的 RGB 颜色直方图。

$$P(\lambda)=\begin{cases}\sum_{u=1}^{n}K\left(\dfrac{|u_i-u_0|}{d}\right)^2 & b(u_i)=\lambda\\ 0 & \text{其他}\end{cases} \tag{7}$$

式中:$K(\cdot)$表示每个像素权重大小的核函数,使得目标中心区域范围的权重较大;$b(u_i)$表示像素点 u_i 处的特征值;d 为检测窗口的带宽。

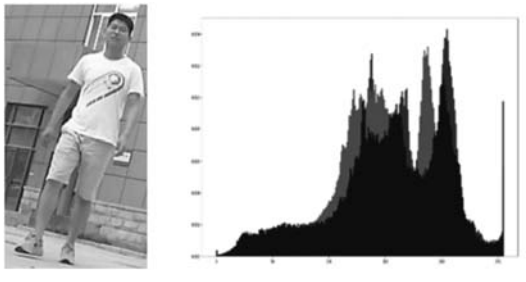

图 4 RGB 颜色特征直方图

当目标基本特征区分度较低的情况下,采用 LBP 纹理特征对目标进行区分,LBP 纹理特征是描述图像局部特征的方法,可以反映目标表面的固有特征。本文为适应不同尺度的纹理特征,实现灰度和旋转的不变性,利用圆形领域代替方形领域。改进的 LBP 算子在半径为 R 的圆形领域内可以有任意多个像素点,改进前后对比如图 5 所示,图 6 为行人目标 LBP 纹理加

权直方图。

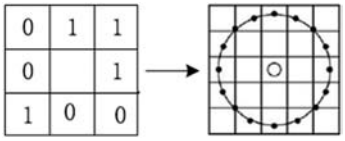

图 5　LBP 算子改进前后对比图

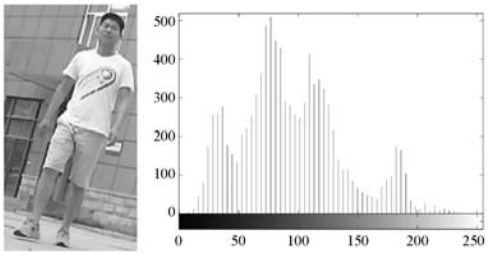

图 6　LBP 纹理加权直方图

2.2　多特征融合层

改进后的 SSD 模型的融合特征层由浅层人工特征与深层网络特征组成,浅层特征通过下采样以减小维度,深层特征通过上采样增加维度。上采样方法为直接填充,即用原特征图上某点的值填充上采样后该点对应区域的所有值。将提取的所有在 Flatten 层的特征值转化为一维向量并融合,然后训练卷积神经网络模型。特征融合方式如图 7 所示。

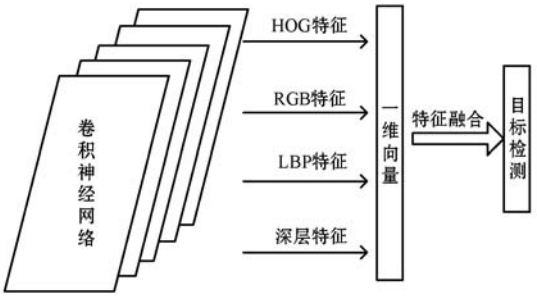

图 7　特征融合方式

3　实验与结果分析

本文实验环境如下:操作系统为 Ubuntu 16.04,CPU 环境为:2 × Intel Xeon Gold 6128 CPU @ 3.40 GHz,内存 32 GB,1T 7200 SATA3.5 + 512 GB SSD,GPU 环境为:2 × NVIDIA Quadro P2000 5 GB 显卡。本文选用 VOC2007 数据集进行模型测试,数据集中共包含 20 个种类,其中训练集有 5 011 幅,测试集有 4 952 幅。本文算法模型搭建环境为 TensorFlow 1.12.0 + Keras 2.2.2 + Python 3.4.0,VOC 函数利用准确率(P)、召回率(R)和平均精度(mAP)评估识别效果。本文以校园采集视频序列作为测试数据,对模型进行评估测试。为了验证融合多特征目标检测算法的性能,本文又在 VOC2007 数据集对模型进行测试。表 1 为几种常见的目标检测算法在 VOC2007 数据集上的性能对比结果。准确率及召回率计算公式如下:

$$P = \frac{N_C}{N_D} \times 100\% \tag{8}$$

$$R = \frac{N_C}{N_R} \times 100\% \tag{9}$$

式中:N_R 为当前帧真正目标的像素集合;N_D 为通过检测算法检测到的目标的像素集合;N_C 为检测到的目标和真实目标的交集部分的像素集合。

表 1　算法检测性能对比表

网络模型	均值平均精度(mAP)/%	帧频(FPS)/(帧 · s^{-1})
Faster R-CNN	73.2	7
YOLO	66.4	21
SSD 300	74.3	46
融合多特征的 SSD	77.1	27
使用 NMS 的 SSD	75.6	51
本文算法	78.6	32

可以看出,仅融合多特征的 SSD 目标检测算法的平均精度比其他几种算法都略高一些,帧频为 27 帧/s^{-1}。选取 VGG16 为基础网络模型,并在网络中加入手工特征后,对目标的分类能力比原始 SSD 模型有所提升,而且稳定性也较好。仅采用非极大值抑制算法(NMS)能够有效消除多余的重叠边框。本文将阈值设置为 0.7,可以看出,使用 NMS 的 SSD 目标检测、识别算法的均值平均精度 mAP 比原始 SSD 算法提升 1.3%。本文算法在融合多特征的 SSD 算法基础上加入非极大值抑制(NMS),mAP 比原始 SSD 算法提升 4.3%。

为定量对实验结果进行分析,本文采用中心位置误差和覆盖率 2 个指标评价本文算法的定位性。中心位置误差是指图像定位候选框的中心位置与原图像目标的真实位置之间的欧式距离。覆盖率是指目标定位框与目标真实位置的重叠部分所占的比重。对 VOC2007 数据集中的 9 963 幅图像进行实验,本文算法与其他几种算法的中心位置误差和覆盖率对比如表 2 所示。

表 2　中心位置误差和覆盖率

算法	中心位置误差	覆盖率/%
原始 SSD 算法	15.1	79.4
仅特征融合 SSD	11.5	85.7
本文算法	6.8	90.3

通过本文算法与其他几种算法的对比可以得出，本文算法的目标检测定位精度明显优于原 SSD 算法。

对融合多特征的 SSD 算法与原始 SSD 算法进行实验对比，对校园内采集的一组图像序列进行实验，改进前后实验结果对比图如图 8 所示。

（a）原始 SSD 算法　　（b）融合多特征的 SSD 算法

图 8　改进前后 SSD 算法检测结果对比

可以看出，融合前的算法对不完整的行人目标、目标大面积遮挡和较小目标无法准确检测，而通过多特征融合后的 SSD 算法能识别出大面积遮挡的目标和不完整的目标。与原始 SSD 等方法相比，改进后的 SSD 方法在 VOC2007 公开数据集上具有更低的平均误检率，平均准确率较传统 SSD 算法提升 4% 左右，其融合了行人浅层和深层特征，提高了目标预测的稳定性和鲁棒性，采用非极大值抑制算法，能够有效提升检测速度，而且对小目标和大面积遮挡目标有更好的检测效果。

4　结　语

基于 SSD 检测网络框架，在卷积神经网络的浅层提取目标的手工特征，并且与卷积网络中深度特征进行融合，通过非极大值抑制（NMS）算法消除重叠得分较低的窗口。不仅可以降低计算成本，提高检测速度，而且提高了检测准确率。通过多特征信息融合后的 SSD 网络模型在 VOC2007 公开数据集上进行验证，结果表明，本文方法较原 SSD 检测方法在小目标检测的准确率和稳定性方面有明显优势。

参　考　文　献

[1] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005: 886 – 893.

[2] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision, 2015: 1440 – 1448.

[3] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779 – 788.

[4] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]//European Conference on Computer Vision. Springer, 2016: 21 – 37.

[5] 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, 2015, 39(6): 1137 – 1149.

[6] Dollar P, Appel R, Belongie S, et al. Fast feature pyramids for object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(8): 1532 – 1545.

[7] Cao J L, Pang Y W, Li X L. Learning multilayer channel features for pedestrian detection[J]. IEEE Transactions on Image Processing, 2017, 26(7): 3210 – 3220.

[8] Xie X M, Han X, Liao Q, et al. Visualization and pruning of SSD with the base network VGG16[C]//International Conference on Deep Learning Technologies. ACM, 2017: 90 – 94.

[9] Li T M, Hou W J, Lyu F, et al. Face detection based on depth information using HOG-LBP[C]//2016 6th International Conference on Instrumentation & Measurement, Computer, Communication and Control, 2016: 779 – 784.

[10] Parate M R, Sinha S, Bhurchandi K M. Integral channel feature based arbitrary object tracking[C]//2016 22nd National Conference on Communication (NCC), 2016: 1 – 6.