DOI: 10.11992/tis.201707018
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180116.1749.002.html

一种多层特征融合的人脸检测方法

王成济[1,2], 罗志明[1,2], 钟准[1,2], 李绍滋[1,2]
（1. 厦门大学 智能科学与技术系, 福建 厦门 361005; 2. 厦门大学 福建省类脑计算技术及应用重点实验室, 福建 厦门 361005）

摘　要: 由于姿态、光照、尺度等原因, 卷积神经网络需要学习出具有强判别力的特征才能应对复杂场景下的人脸检测问题。受卷积神经网络中特定特征层感受野大小限制, 单独一层的特征无法应对多姿态多尺度的人脸, 为此提出了串联不同大小感受野的多层特征融合方法用于检测多元化的人脸; 同时, 通过引入加权降低得分的方法, 改进了目前常用的非极大值抑制算法, 用于处理由于遮挡造成的相邻人脸的漏检问题。在 FDDB 和 WiderFace 两个数据集上的实验结果显示, 文中提出的多层特征融合方法能显著提升检测结果, 改进后的非极大值抑制算法能够提升相邻人脸之间的检测准确率。

关键词: 人脸检测; 多姿态; 多尺度; 遮挡; 复杂场景; 卷积神经网络; 特征融合; 非极大值抑制
中图分类号: TP391.41　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)01−0138−09

中文引用格式: 王成济, 罗志明, 钟准, 等. 一种多层特征融合的人脸检测方法[J]. 智能系统学报, 2018, 13(1): 138–146.
英文引用格式: WANG Chengji, LUO Zhiming, ZHONG Zhun, et al. Face detection method fusing multi-layer features[J]. CAAI transactions on intelligent systems, 2018, 13(1): 138–146.

Face detection method fusing multi-layer features

WANG Chengji[1,2], LUO Zhiming[1,2], ZHONG Zhun[1,2], LI Shaozi[1,2]
(1. Intelligent Science & Technology Department, Xiamen University, Xiamen 361005, China; 2. Fujian Key Laboratory of Brain-inspired Computing Technique and Applications, Xiamen University, Xiamen 361005, China)

Abstract: To address the issues of pose, lighting variation, and scales, convolutional neural networks (CNNs) need to learn features with strong discrimination handle the face detection problem in complex scenes. Owing to the size limitations of the specific feature layer's receptive field in convolutional neural networks, the features computed from a single layer of the CNNs are incapable of dealing with faces in multi poses and multi scales. Therefore, a multi-layer feature fusion method that is realized by fusing the different sizes of receptive fields is proposed to detect diversified faces. Moreover, via introducing the method of weighted score decrease, the present usual non-maximum suppression algorithm was improved to deal with the detection omission of neighboring faces caused by shielding. The experiment results with the FDDB and WiderFace datasets demonstrated that the fusion method proposed in this study can significantly boost detection performance, while the improved non-maximum suppression algorithm can increase the detection accuracy between neighboring faces.

Keywords: face detection; multi pose; multi scale; occlude; complex scenes; convolutional neural network; feature fusion; non-maximum suppression

人脸识别技术作为智能视频分析的一个关键环节, 在视频监控、网上追逃、银行身份验证等方面有着广泛的应用。人脸检测是人脸识别的基础关键环节之一, 在智能相机、人机交互等领域也有着广泛的应用。人脸检测是在输入图像中判断是否存在人脸, 同时确定人脸的具体大小、位置和姿态的过程。作为早期计算机视觉的应用之一, 人脸检测的

收稿日期: 2017−07−10.　**网络出版日期:** 2018−01−18.
基金项目: 国家自然科学基金项目 (61572409, 61402386, 81230087, 61571188).
通信作者: 李绍滋. E-mail: szlig@xmu.edu.cn.

相关研究可以追溯到 1970 年[1]。由于真实场景中人脸的复杂性和背景的多样性, 人脸检测技术在复杂场景下还存在着许多挑战。

近年来深度卷积神经网络 (CNN) 使图像识别、目标检测等计算机视觉任务取得长足进步[2-4]。目标检测问题可以看作两个子问题的组合: 目标定位问题和目标分类问题。目标定位问题主要确定物体在图像中的具体位置, 目标分类问题将确定目标相应的类别。受 ren 等[4]提出的区域候选框提取网络 (region proposal network, RPN) 的启发, Huang 等[5]和 Yu 等[6]认为用于解决图像分割问题的框架同样适用于目标检测问题, 它们对于图片中的每一个像素点都判断该像素是否属于人脸区域以及当属于人脸区域时相对于人脸区域边界坐标的偏移量 (当前像素点与人脸边界在空间坐标上的相对偏移)。UnitBox[6]将用于图像分类的 VGG16[7]网络改造为全卷积神经网络 (FCN)[8], 在 pool4 特征层的基础上预测像素点的分类得分, 在 pool5 特征层的基础上预测人脸区域内像素点坐标的偏移量。UnitBox[6]首次使用重叠率评价人脸区域内像素点坐标偏移量回归的好坏, 重叠率损失函数将人脸区域内每个像素点的上下左右 4 个偏移量当作一个整体, 利用了这 4 个偏移量之间的关联性。Yu[6]认为用于预测人脸区域内像素点坐标偏移量的特征需要比预测人脸分类的特征有更大的感受野, 所以他们仅利用了 pool5 层特征预测坐标偏移量, 在预测每一个像素点的分类得分时 UnitBox 使用椭圆形的人脸区域的标注, 在测试时在分类得到的得分图上做椭圆检测, 然后提取检测出的椭圆的中心点对应的矩形框作为最终检测结果。在实验中我们发现在使用椭圆标注训练得到的得分图像无法拟合出标准的椭圆, 尤其当多个人脸区域有重叠时, 无法分开多个人脸区域。实验中还发现, 使用 pool5 层的特征虽然有很好的感受野但在处理小人脸时会因为感受野过大造成小人脸区域内坐标偏移量回归不准确, 影响最终检测结果。

基于以上工作, 本文使用矩形的人脸区域标注, 摈弃了 UnitBox[6]后处理中的椭圆检测的部分, 转而使用非极大值抑制算法过滤大量重复的矩形框; 当两个人脸区域重叠率超过非极大值抑制算法的阈值时, 以前的非极大值抑制算法只能够保留一个人脸会造成漏检, 为了避免这个问题, 本文根据矩形框的重叠率对预测矩形框的得分加权降低非最大矩形框的置信度, 然后使用置信度阈值来过滤矩形框, 这样当两个人脸检测出的矩形框重叠率大于制定阈值时也不会直接过滤掉, 避免漏检。在特征的感受野过大的问题上, 本文重新探索了不同卷积层在人脸检测任务中的重要性, 同比较不同大小感受野的特征组合方法对准确率的影响, 发现结合 pool4 层的特征和 pool5 层的特征能同时处理大人脸和小人脸。

1　相关工作

人脸检测大致可以分为 3 个部分: 候选框提取、图像分类、边框坐标回归。传统方法采用滑动窗口提取候选框, 然后使用 Harr_[9]、SIFT[10]、HOG[11]等手工提取的特征结合 SVM[12]、boosting[9, 13]等机器学习算法对候选框进行分类。这种穷举的策略虽然包含了目标所有可能出现的位置, 但是缺点也是明显的: 1) 基于滑动窗口的区域选择策略没有针对性, 时间复杂度高, 窗口冗余; 2) 手工设计的特征对于多样性的变化并没有很好的鲁棒性。

为了解决滑动窗口计算复杂度高的问题, 出现了利用图像中的纹理、边缘、颜色等信息的基于区域候选框的解决方案[14-15], 这种方案可以保证在选取较少窗口的情况下保持较高的召回率。这大大降低了后续操作的时间复杂度, 并且获取的候选窗口要比滑动窗口的质量更高。Ross B. Girshick 等[2]提出的 RCNN 框架, 使得目标检测的准确率取得极大提升, 并开启了基于深度学习目标检测的热潮。Fast RCNN[3]方法利用特征图提取候选框极大地降低了基于深度学习目标检测方法的时间复杂度。Faster R-CNN[4]方法更进一步, 首次提出了自动提取图片中区域候选框的 RPN 网络, 并将传统的提取候选框的操作集成到特征学习网络中, 使得目标检测问题可以达到 end-to-end。CascadeCNN[16]使用 3 个独立的卷积神经网络分级过滤候选框。DDFD[17]首次将全卷积神经网络[8]成功地应用于人脸检测问题中。

2014 年 J. Long 等[8]提出全卷积神经网络 (fully convolution network, FCN) 并成功地应用在图像分割任务中, 直到现在 FCN 依然是图像分割的主流框架。全卷积神经网络 (FCN) 与卷积神经网络 (convolution neural network, CNN) 的主要不同是 FCN 将 CNN 中的全连接层通过卷积层实现, 并使用反卷积操作得到与输入同样大小的输出, 因此网络的输出由原始 CNN 的关于整张图像上的分类结果变成了 FCN 中关于整张图像的像素级的分类, 也就是输入图像的每一个像素点都对应有一个分类的输出结果。FCN 是直接对像素点进行操作, 在经过一系列的卷积和反卷积的操作后得到与原始输入图像同样大小的中间结果, 最后经过 softmax 操作输出类

别概率。FCN 的主要网络是在现有的 AlexNet[18]、VGGNet[7]和 ResNet[19]等用于图像分类的 CNN 网络模型上增加反卷积操作来实现的。DenseBox[5]在文献[15]基础上将人脸区域坐标回归问题视为在特征图的每一个像素位置预测这个像素坐标相对于人脸区域边界坐标的偏移量的问题，然后使用类似图像分割的方法来处理，并采用了 l_2 损失函数作为坐标回归的损失函数，UnitBox[6]认为同一个像素的 4 个偏移量之间是相互关联的，为了体现这种关联性提出了使用重叠率损失函数，通过不断优化预测人脸矩形框与真实人脸矩形框的重叠率，使得最终预测的矩形框与真实矩形框的重叠率不断增加。

2　算法框架

本节主要介绍整体算法流程，如图 1 所示。在训练阶段有 3 个输入：RGB 的训练图片、单通道的区域像素分类标签和四通道的人脸区域内像素点坐标偏移标签。经过 FCN 网络后有两个输出：第一个是像素级分类得分的概率图，判断该像素点是否属于某个人脸区域；另一个是 1 个 4 通道的像素点坐标偏移图，4 通道的像素点坐标偏移图中的 4 个通道分别对应每一个像素值与离它最近的人脸区域的上下左右 4 个边框坐标的偏移量。最后使用交叉熵损失函数和重叠率损失函数指导网络训练，我们使用联合训练。标签形式见 2.1 节，网络的具体细节见 2.2 节。每一个像素都需要计算交叉熵损失，但仅仅对包含在标注的人脸区域内的像素点计算重叠率损失。在测试阶段输入图片经过训练好的 FCN 模型输出每一个像素点的分类得分和人脸区域内像素点坐标偏移量，对每一个得分大于阈值的像素点我们从对应四通道坐标偏移图取出该像素点相对于离它最近的人脸区域边界坐标的偏移量，假设像素点 $p(x_i,y_i)$ 的预测得分 s_i 大于阈值且预测的坐标偏移为 $(\mathrm{d}x_i^t,\mathrm{d}y_i^t,\mathrm{d}x_i^b,\mathrm{d}y_i^b)$，则像素点 $p(x_i,y_i)$ 的预测矩形框坐标为 $(x_i-\mathrm{d}x_i^t,y_i-\mathrm{d}y_i^t,x_i+\mathrm{d}x_i^b,y_i+\mathrm{d}y_i^b)$，使用 NMS 算法过滤重复检测的矩形框，得到最终检测结果。

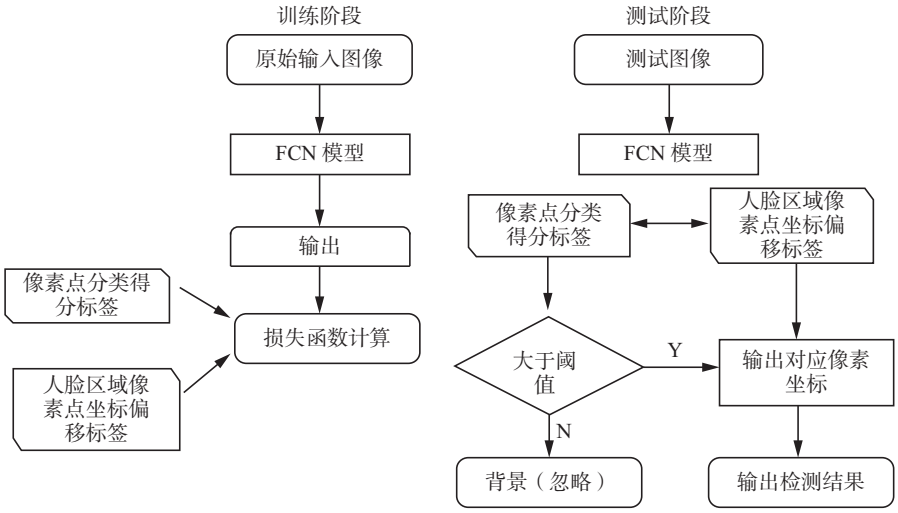
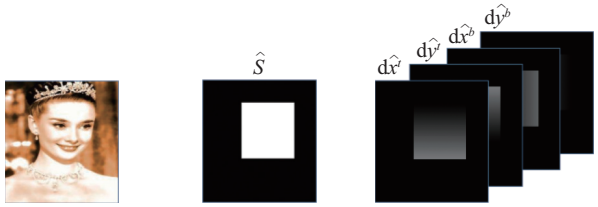


图 1　算法流程

Fig. 1　Algorithm procedure

2.1　训练标签的制作

训练标签如图 2 所示。



(a) 原始图像　(b) 类别分类标签　(c) 坐标偏移标签

图 2　训练标签

Fig. 2　Ground truth

对于每一张训练的图像，将图像上每一个人脸标注的矩形区域，以 1 填充，其他区域填充 0，作为每一个像素点的人脸置信度得分 $\hat{s}$。假设像素点 $p(x_i,y_i)$ 包含在某个人脸区域中，假设这个人脸区域左上角坐标为 $p_t(x_t,y_t)$，右下角坐标为 $p_b(x_b,y_b)$，则像素点 $p(x_i,y_i)$ 的标签向量形式：$\hat{t}_i=\{\hat{s}_i,\widehat{\mathrm{d}x^t}=x_i-x_t,\widehat{\mathrm{d}y^t}=y_i-y_t,\widehat{\mathrm{d}x^b}=x_i-x_b,\widehat{\mathrm{d}y^b}=y_i-y_b\}$。

2.2　多级特征串联

网络模型结构如图 3 所示，使用的是去掉了全连接层和 softmax 层的 VGG16 网络[7]作为模型共享的特征提取网络。在共享的特征提取网络的基础上，在 pool4 特征层后添加了两个独立的卷积层 sc_conv4 和 bbx_conv4，每一个卷积层包括 32 个 3×3 的卷积核，并保持特征图分辨率大小不变，在 pool5

特征层后同样添加了含有 32 个 3×3 的卷积核的卷积层 bbx_conv5。因为 pool4 特征层的分辨率是输入的 1/16, 为了得到与输入同样大小的输出, 对 sc_conv4 和 bbx_conv4 分别做了步长为 16 的反卷积操作, 将 sc_conv4 和 bbx_conv4两个特征层的分辨率放大 16 倍并保持特征维度不变, 对 bbx_conv5 使用反卷积放大 32 倍使分辨率与输入相同。sc_conv4 层输出的特征首先被放大 16 倍, 输入到含有 32 个 3×3 卷积核的卷积层和 1 个卷积核大小为 1×1 的卷积层, 最后输入到 sigmoid 激活函数得到每一个像素点的类别分类得分。为了得到预测的 4 维坐标偏移图, 将反卷积后的 bbx_conv4 和 bbx_conv5 两个特征层串联后经过连续两层含有 32 个 3×3 卷积核的卷积层得到 4 维人脸区域内的坐标偏移图。

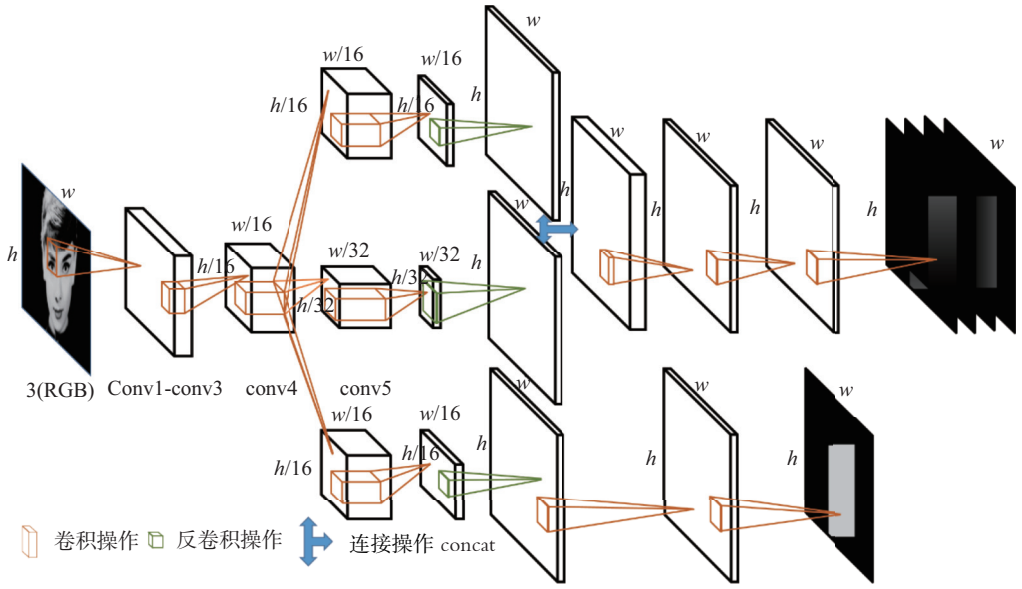


图 3　模型结构

Fig. 3　Model structure

在卷积神经网络中 pooling 层主要起降低分辨率的作用, 越往后特征层的分辨率会越小, 也越能够提取出抽象的语义信息, 但越抽象的特征细节信息丢失越多, 在处理像素级分类任务时仅使用高层抽象的特征会导致边缘部分分类不准确。但是若完全依靠前面层的特征, 虽然能够提高对人脸区域边缘的像素点的分类能力, 但是由于浅层特征的抽象能力不够使得整体上分类结果不准确。文献[8, 20]的研究表明通过融合不同的特征层能够显著提升网络的效果, FCN[8]中的实验也证明融合不同特征层特征的有效性, 主要融合方式有 FCN-32、FCN-16、FCN-8。UnitBox[6]认为人脸区域边框回归需要抽象的语义信息, 所以仅使用了 pool5 层的特征用于处理边框回归任务, 但实际实验中表明融合 pool5 和 pool4 两个特征层的特征能显著提升结果。

本文的模型共享特征层后对于不同的任务添加了多个 3×3 的独立卷积操作, 像素级分类得分的标签是[0, 1], 而人脸区域内坐标偏移量的标签是[0, +w](这里的 w 代表所有标注人脸区域的宽或高的最大值), pool5 特征层的分辨率是输入的 1/32, pool4 是输入的 1/16, 使用与输出同样数量的卷积操作会丢失大量信息, 不仅不会帮助模型训练反而会将前面学习到的错误结果放大降低网络的性能, 而使用更多的卷积操作虽然会增加模型的表达能力但也会增加模型的时间复杂度。

3　损失函数设计

人脸检测问题可以看作两个子问题的组合: 人脸区域定位问题和图像分类问题。图像分类是对整张输入图像分一个类别, 而图像分割是标注图片每一个像素到对应类别的任务, 本文将人脸检测问题中的图像分类问题看成人脸区域分割问题。当将图像中的每一个像素都分配一个对应的候选框, 那么人脸检测问题可以分解为图像分割问题和候选框回归问题两个子问题, 分别对应候选框得分和候选框回归。每一个像素的分类得分也是这个像素对应预测矩形框的得分。本文使用多任务联合训练, 主要包括人脸区域分割任务和人脸区域内像素点坐标偏移回归任务。针对分类任务我们使用的是交叉熵损失函数L_{ce}, 人脸区域的坐标偏移量回归使用重叠率损失函数L_{iou}, 为了使两个损失函数在训练的过程中的梯度保持在同一个量级上, 我们引入了一个权值 λ, 使得最终的损失函数 L 为

$$L = \lambda L_{ce} + L_{iou} \tag{1}$$

3.1 交叉熵损失函数

像素级分类问题是要得到每一个像素输入属于每个类别的概率，人脸检测问题是二分类问题，即人脸与非人脸。本文使用 sigmoid 激活函数实现从特征空间到[0, 1]概率空间的映射，得到每一个像素分类得分的概率，然后使用交叉熵损失函数指导网络训练。sigmoid 激活函数为

$$f_i(x)=\frac{1}{1+\exp^{-(w_j x+b_j)}} \tag{2}$$

式中的$w_j x+b_j$表示在激活函数前的卷积核大小为 1×1 的卷积层。假设像素点$p(x_i,y_i)$被预测为人脸的概率为p_{f_i}，则非人脸的概率为$1-p_{f_i}$，若该像素点在人脸区域内该像素点的标签$g_i=1$，否则$g_i=0$。具体的交叉熵损失函数为

$$L_{\text{ce}}=-\sum_i g_i \ln(p_{f_i})+(1-g_i)\ln(1-p_{f_i}) \tag{3}$$

3.2 重叠率损失函数

之前的候选框坐标回归算法中常用的损失函数是$L_{2\text{loss}}$，他们认为候选框的 4 个坐标是 4 个独立变量可以分开处理，实际上候选框的 4 个坐标之间是有相互关联的，在训练过程中能够相互影响，提升最终的检测结果，所以 UnitBox[6]引入了重叠率损失函数，使候选框坐标间的关联性体现在损失函数中指导网络训练。本文在训练候选框坐标偏移时同样也是使用了 UnitBox 中提出的 IOU 损失函数。如图 4 所示，假设像素点$p(x_i,y_i)$人脸区域边框和预测矩形框分别是$g_i=(x_i,y_i,w_i,h_i)$，$\widetilde{g_i}=(\widetilde{x_i},\widetilde{y_i},\widetilde{w_i},\widetilde{h_i})$，则$g_i$与$\widetilde{g_i}$的重叠率 IoU 为

$$\text{IoU}_i=\frac{g_i\cap\widetilde{g_i}}{g_i\cup\widetilde{g_i}} \tag{4}$$

l_2 损失函数为

$$L_{2\text{loss}}=-\sum_i \|g_i-\widetilde{g_i}\|^2 \tag{5}$$

重叠率损失函数为

$$L_{\text{iou}}=-\sum_i \ln(\text{IoU}_i) \tag{6}$$

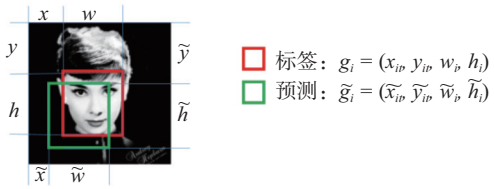


图 4　重叠率

Fig. 4　Intersection-over-union

4 基于加权得分的非极大值抑制方法

非极大值抑制方法 (non-maximum suppression, NMS) 是目标检测中常用的后处理方法，当算法对同一个目标检测出多个重叠率较高的框，需要使用 NMS 来选取重叠区域里分数最高的矩形框 (人脸的概率最大)，非极大值抑制方法采用的是排序—遍历—消除的过程，在这个过程中检测出来的矩形框的得分不变，在一定程度上会影响算法性能。N.Bodla 等[21]发现在排序阶段对重叠率高于阈值且得分较低的预测框的得分进行加权，再过滤掉得分低的矩形框能有效解决非极大值抑制算法导致的漏检问题。

受文献[21]的启发，我们在非极大值抑制的过程中使用两次遍历和消除过程，在第一次遍历过程中，当两个框的重叠率大于α时，将得分较低的窗口的得分乘以一个权值，然后根据加权后的得分过滤掉低于φ的窗口，完成后再次使用没有加权的非极大值抑制方法得到最终检测结果。在实验过程中，测试了两种不同的加权方法：线性加权和高斯加权。两种加权方法的具体计算：当两个窗口交并比小于α，则得分低的窗口的得分要乘以权值 weight。

线性加权为

$$\text{weight}_i=1-\text{IoU}_i \tag{7}$$

高斯加权为

$$\text{weight}_i=\exp(-\frac{\text{IoU}_i^2}{\text{sigma}}) \tag{8}$$

5 实验与结果分析

为了验证方法的有效性，我们使用 Wider Face 数据集[22]的训练集训练，并在 FDDB 数据集[23]和 Wider Face 数据集[22]的验证集上评测结果，并与当前领先的算法进行比较，此外本文还比较了使用不同加权方式的非极大值抑制方法的性能。

5.1 实验数据

FDDB 人脸评测[23]平台的测试集有 2 845 张图片，共有 5 171 张标注人脸，范围包括不同姿态、不同分辨率、不同遮挡情况的图像。评测指标是检测出的矩形区域和标注区域的重叠率，重叠率大于等于 0.5 表示检测正确。

Wider Face 数据集[22]是由香港中文大学公开发布的人脸检测基准数据集，包含训练集、验证集和测试集 3 部分，是现有 FDDB 数据集中标注的图像数量的 10 倍。共包含 3.2 万张图像，39.3 万张手工标注的人脸，平均每张图像有 12 个标注的人脸。Wider Face 数据集中的人脸姿态、大小、遮挡情况变化多样，数据集以小人脸为主且人脸区域的分辨率偏低。整个 Wider Face 数据集中的图像分为 61 个事件类别，根据标注人脸的大小，数据集中的人脸检测任务分为 3 个难度等级 Easy、Medium、Hard，所以有 3 条评测曲线。

5.2　实验设置与结果分析

本文使用的训练数据来自 Wider Face[22]的训练集, 总共有 12 880 张图像, 统一将训练图像的宽和高用 ImageNet[24]上的图像均值填充为 32 的倍数, 测试时同样对图像填充为 32 的倍数。训练是以标注的人脸区域中心周围占整个人脸区域 3/5 的区域为正样本, 该区域关于标注的人脸区域中心对称。其他像素点设为负样本。由于原始的 UnitBox[6]论文没有公布测试模型和源代码, 在本文中我们复现了 UnitBox[6]代码作为比较对象。在使用多任务联合训练, 由于人脸区域分类的损失和人脸区域边框回归的损失函数不在同一个数量级上, 本文对分类损失赋权 0.001。训练是在 WiderFace训练集上训练, 每次使用一张图像, 使用 Adam 算法[25]在整个数据集上迭代训练 30 轮, 本文使用加权的非极大值抑制算法做后处理。

图 5 中比较了本文的算法与原始 UnitBox[6]算法在 FDDB 数据集上的性能, 同时对比了另外 7 个经典的人脸检测算法: DDFD[17]、CascadeCNN[16]、ACF-multiscale[26]、Pico[27]、HeadHunter[28]、Joint-Cascade[29]、Viola-Jones[9], 实验表明本文的多级特征串联能明显提升算法性能。本文的方法在共享的卷积层和串联的特征层后都添加了卷积层, 同时本文单独对 pool5 层的特征添加同样的卷积层作为对比实验 (UnitBox-refine)。从图 5 中可以看出, 仅仅在 pool5 层输出的特征后添加卷积操作的结果为 0.859, 而在结合 pool4 和 pool5 层特征后再添加卷积操作的结果为 0.906, 说明仅仅对单层特征进行多次卷积和池化操作不能有效提升检测结果。

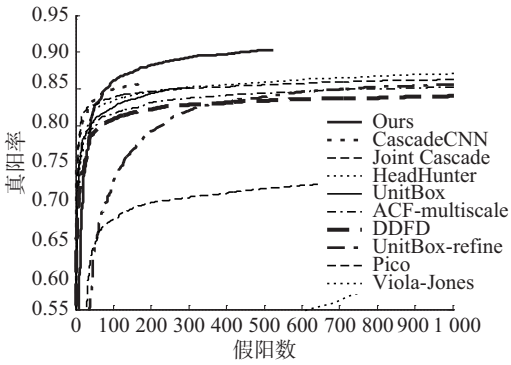


图 5　FDDB 数据集 ROC 曲线

Fig. 5　ROC Curve on FDDB dataset

同样的, 在 WiderFace 数据集的验证集上测试比较了本文算法与其他领先算法的性能。图 6 展示了本文算法在 WiderFace 验证集的 Easy、Medium 和 Hard 三个难易程度上的性能曲线。还对比了多个先进的人脸检测算法: LDCF+[30]、Multiscale Cascade CNN[22]、Faceness-WIDER[31]、ACF-WIDER[26], 在 Easy 难度上本文算法比 LDCF+[30]高 0.5 个百分点, 在 UnitBox[6]的基础上提高了 9 个百分点, 在 Medium 难度上取得了 0.737 的检测结果, 在 Hard 难度上比 UnitBox[6]提升了 9.8 个百分点。图 7 展示了本文算法的部分检测结果。

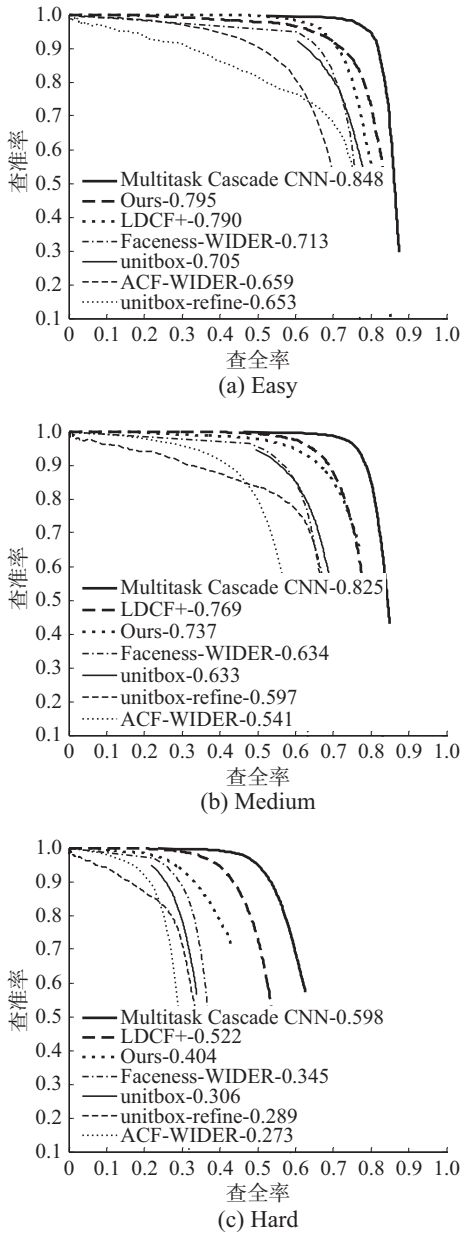


图 6　WiderFace 验证集上的准确率-召回率曲线

Fig. 6　Percision-recall curve on Wider Face Val set

表 1 比较了加权得分的非极大值抑制方法和不加权的极大值抑制方法的后处理结果, 这里高斯加权中使用的方差 sigma=0.5。可以看出在 FDDB 数据集中使用高斯加权和线性加权获得的提升一样, 在 WiderFace 数据中使用高斯加权的提升明显大于线性加权, 说明高斯加权的方法更适合于小人脸检测问题。在图 8 中我们展示了部分不同的 NMS 方法的处理结果。

(a) WiderFace数据集部分检测结果　(b) FDDB 数据集部分检测结果

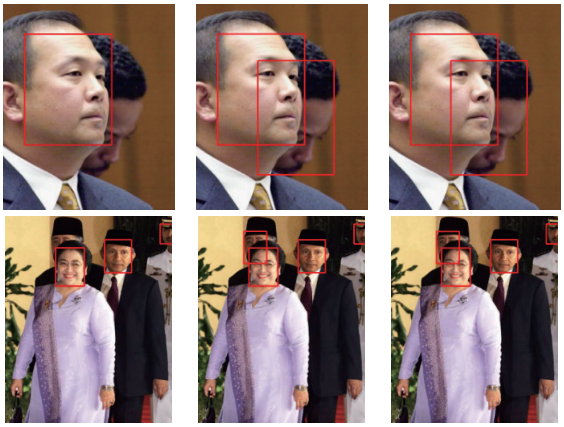
图 7　检测结果

Fig. 7　Detection results

表 1　NMS 对比实验准确率

Table 1　The accuracy of contrast experiment

数据集	方法	准确率/%
FDDB	NMS	90.29
	NMS-gaussian	90.62
	NMS-linear	90.62
WiderFace-Easy	NMS	79.50
	NMS-gaussian	79.90
	NMS-linear	79.80
WiderFace-Medium	NMS	73.70
	NMS-gaussian	74.10
	NMS-linear	73.90
WiderFace-Hard	NMS	40.40
	NMS-gaussian	40.60
	NMS-linear	40.50



(a) NMS　(b) NMS-gaussian　(c) NMS-linear

图 8　不同 NMS 的后处理结果对比

Fig. 8　The comparesion of NMS methods

6　结束语

目标检测和图像分割问题是计算机视觉中两个重要的基本问题, 本文的人脸检测方法试图将解决

图像分割问题的算法框架尝试应用于人脸检测问题。在前人的基础上本文探索了不同的特征串联方法对人脸区域坐标回归的影响, 通过实验发现并不是特征组合得越多结果越好, 本文使用 pool4 和 pool5 两个特征层的特征取得了很大的提升。在后处理阶段, 本文通过比较分析不同的非极大值抑制策略的性能, 发现通常使用的不加权的非极大值抑制方法虽然高效, 但会在一定程度上影响目标检测方法的性能。本文在人脸区域分类问题和人脸区域内像素点坐标偏移量回归两个问题实际上是分开处理, 在今后的研究中如何发现并使用这两个问题之间的关联性是一个很重要的研究思路。本文虽然使用加权得分的方法在一定程度上缓解了非极大值抑制方法检测算法的影响, 但没有得出一般性的结论, 这个问题同样值得深入研究。

参考文献:

[1] ZAFEIRIOU S, ZHANG Cha, ZHANG Zhengyou. A survey on face detection in the wild: past, present and future [J]. Computer vision and image understanding, 2015, 138: 1–24.

[2] 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, OH, USA, 2014: 580–587.

[3] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1440–1448.

[4] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada, 2015, 1: 91–99.

[5] HUANG Lichao, YANG Yi, DENG Yafeng, et al. DenseBox: unifying landmark localization with end to end object detection[J]. arXiv preprint arXiv: 1509.04874, 2015.

[6] YU Jiahui, JIANG Yuning, WANG Zhangyang, et al. UnitBox: An advanced object detection network[C]//Proceedings of the 2016 ACM on Multimedia Conference. Amsterdam, The Netherlands, 2016: 516–520.

[7] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the International Conference on Learning Representations. Oxford, USA, 2015.

[8] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA, 2015: 3431–3440.

[9] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features[C]//Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, HI, USA, 2001, 1: I-511–I-518.

[10] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91–110.

[11] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA, 2005, 1: 886–893.

[12] OSUNA E, FREUND R, GIROSIT F. Training support vector machines: an application to face detection[C]//Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Juan, Argentina, 1997: 130–136.

[13] FRIEDMAN J, HASTIE T, TIBSHIRANI R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)[J]. The annals of statistics, 2000, 29(5): 337–407.

[14] ZITNICK C L, DOLLÁR P. Edge boxes: locating object proposals from edges[C]//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland, 2014: 391–405.

[15] UIJLINGS J R R, VAN DE SANDE K E A, GEVERS T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104(2): 154–171.

[16] LI Haoxiang, LIN Zhe, SHEN Xiaohui, et al. A convolutional neural network cascade for face detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA, 2015: 5325–5334.

[17] FARFADE S S, SABERIAN M J, LI Lijia. Multi-view face detection using deep convolutional neural networks[C]//Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. Shanghai, China, 2015: 643–650.

[18] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Proceedings of the 26th Annual Conference on Neural Information Processing Systems 2012. Lake Tahoe, Nevada, USA, 2012: 1097–1105.

[19] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 2016: 770–778.

[20] HARIHARAN B, ARBELÁEZ P, GIRSHICK R, et al. Hypercolumns for object segmentation and fine-grained local-

ization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA, 2015: 447–456.

[21] BODLA N, SINGH B, CHELLAPPA R, et al. Improving object detection with one line of code[J]. arXiv preprint arXiv: 1704.04503, 2017.

[22] YANG Shuo, LUO Ping, LOY C C, et al. Wider Face: A face detection benchmark[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 2016: 5525–5533.

[23] JAIN V, LEARNED-MILLER E. FDDB: A benchmark for face detection in unconstrained settings[R]. UMass Amherst Technical Report UMCS-2010-009, 2010.

[24] DENG Jia, DONG Wei, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA, 2009: 248–255.

[25] KINGMA D P, BA J L. Adam: A method for stochastic optimization[C]//Proceedings of International Conference on Learning Representations. Toronto, Canada, 2015.

[26] YANG Bin, YAN Junjie, LEI Zhen, et al. Aggregate channel features for multi-view face detection[C]//Proceedings of the 2014 IEEE International Joint Conference on Biometrics (IJCB). Clearwater, FL, USA, 2014: 1–8.

[27] MARKUS N, FRLJAK M, PANDZIC I S, et al. A method for object detection based on pixel intensity comparisons organized in decision trees[J]. CoRR, 2014.

[28] MATHIAS M, BENENSON R, PEDERSOLI M, et al. Face detection without bells and whistles[C]//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland, 2014: 720–735.

[29] CHEN Dong, REN Shaoqing, WEI Yichen, et al. Joint cascade face detection and alignment[C]//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland, 2014: 109–122.

[30] OHN-BAR E, TRIVEDI M M. To boost or not to boost? On the limits of boosted trees for object detection[C]//Proceedings of the 23rd International Conference on Pattern Recognition (ICPR). Cancun, Mexico, 2016: 3350–3355.

[31] YANG Shuo, LUO Ping, LOY C C, et al. From facial parts responses to face detection: A deep learning approach[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 3676–3684.

作者简介：

王成济，男，1993 年生，硕士研究生，主要研究方向为视频目标检测和图像分割。

罗志明，男，1989 年生，博士研究生，主要研究方向为图像分割、目标检测、医学图像分析。发表学术论文 8 篇。

李绍滋，男，1963 年生，教授，博士生导师，主要研究方向为计算机视觉、机器学习和数据挖掘。先后主持或参加过多项国家 863 项目、国家自然科学基金项目、教育部博士点基金项目、省科技重点项目等多个项目的研究，发表学术论文 300 多篇。