基于单一神经网络的多尺度人脸检测

刘宏哲[①] 杨少鹏*[①] 袁家政[②] 王雪峤[③] 薛建明[①]

[①](北京联合大学北京市信息服务工程重点实验室 北京 100101)

[②](北京开放大学 北京 100081)

[③](北京联合大学计算机技术研究所 北京 100101)

摘 要：人脸检测是指检测并定位输入图像中所有的人脸，并返回精确的人脸位置和大小，是目标检测的重要方向。为了解决人脸尺度多样性给人脸检测造成的困难，该文提出一种新的基于单一神经网络的特征图融合多尺度人脸检测算法。该算法在不同大小的卷积层上预测人脸，实现实时多尺度人脸检测，并通过将浅层的特征图融合引入上下文信息提高小尺寸人脸检测精度。在数据集FDDB和WIDERFACE测试结果表明，所提方法达到了先进人脸检测的水平，并且该方法去掉了框推荐过程，因此检测速度更快。在WIDERFACE难、适中、简单3个子数据集上测试结果分别为87.9%, 93.2%, 93.4% MAP，检测速度为35 fps。所提算法与目前效果较好的极小人脸检测方法相比，在保证精度的同时提高了人脸检测速度。

关键词：多尺度人脸检测；上下文信息；特征图融合；卷积神经网络

中图分类号：TP391.4 **文献标识码**：A **文章编号**：1009-5896(2018)11-2598-08

DOI: 10.11999/JEIT180163

Multi-scale Face Detection Based on Single Neural Network

LIU Hongzhe[①] YANG Shaopeng[①] YUAN Jiazheng[②]

WANG Xueqiao[③] XUE Jianming[①]

[①](*Beijing Key Laboratory of Information Service Engineering*, *Beijing Union University*, *Beijing* 100101, *China*)

[②](*Beijing Open University*, *Beijing* 100081, *China*)

[③](*Institute of Computer Technology*, *Beijing Union University*, *Beijing* 100101, *China*)

Abstract: Face detection is finding and locating all faces in the input image, and then returning the position and size of the faces. It is an important direction of target detection. In order to solve the problem which is caused by the diversity of face size, a new single shot multiscale face algorithm is presented based on feature fusion. This method combines predictions from multiple feature maps with different resolutions to handle faces of various sizes, and the fusion of the feature maps in the shallow layers can improve the detection accuracy of the small size face by introducing the contextual information. Experimental results on the FDDB and WIDERFACE datasets confirm that the proposed method has competitive accuracy. Additionally, the object proposal step is removed, which makes the method fast. The proposed model achieves 87.9%, 93.2% and 93.4% Mean Average Precision (MAP) on the WIDERFACE sub-datasets respectively, at 35 fps. The proposed method outperforms a comparable state-of-the-art HR model, and at the same time improves the speed while ensuring the accuracy.

Key words: Multi-scale face detection; Contextual information; Feature map fusion; Convolution neural network

收稿日期：2018-02-07；改回日期：2018-07-05；网络出版：2018-07-23

*通信作者： 杨少鹏 shaopeng568@163.com

基金项目：国家自然科学基金(61571045)，北京市属高校高水平教师队伍建设支持计划项目(IDHT20170511)，国家科技支撑项目(2015BAH55F03)，北京联合大学新起点项目(Zk10201703)，北京市教委科技计划一般项目(KM201811417002)

Foundation Items: The National Natural Science Foundation of China (61571045), The Supporting Plan for Cultivating High Level Teachers in Colleges and Universities in Beijing (IDHT20170511), The National Science and Technology Support Project (2015BAH55F03), The Foundation of Beijing Union University (Zk10201703), The Foundation of Beijing Municipal Education Commission (KM201811417002)

1　引言

人脸检测可分为约束环境下的人脸检测和非约束环境下的人脸检测[1]。约束环境为直立正脸、背景单一、人脸数量单一等相对理想的场景，非约束环境为人脸尺度多样性、姿态多样性、遮挡、化妆等恶劣条件[2]。实际应用中主要是非约束场景下的人脸检测。人脸检测是人脸相关研究的前置步骤，如人脸验证[3,4]、人脸识别[5]、人脸跟踪[6]、人脸属性分析等。人脸识别和人脸验证技术已经广泛应用在大规模安防布控领域，人脸检测作为识别或验证的第1步，其检测的效果会直接影响到后续应用的准确率，因此其具有重要的意义。

人脸检测方法在过去的几十年里如雨后春笋般飞速地发展，越来越多的研究人员开始研究人脸检测算法，并取得了较好的效果。目前有很多基于视频或图像的人脸检测方法[7]。传统的人脸检测方法如Viola和Jones等人[8]在2001年提出的基于Adaboost的方法。随后出现一些基于Adaboost的改进算法如HeadHunter和SURF-Adaboost[9,10]算法。然而传统方法在遇到尺度、姿态多样性问题时，主要以牺牲速度为代价，将人脸细分，然后分类训练模型，这导致这些方法在进行人脸检测时，需要分类去检测人脸，时间复杂度较高。同时也受限于人工提取特征的表达能力和分类器的分类能力，传统的方法还无法达到较好的检测效果。

在神经网络快速发展的今天，研究人员开始使用模型表达能力更好的卷积神经网络(CNN)做分类器。例如：级联卷积神经网络人脸检测方法(CascadeCNN)[11]是经典的Adaboost人脸检测方法的深度卷积网络实现，和传统的先提取人工设计的特征再分类结果相比，CascadeCNN具有更好的检测精度和更快的检测速度。Wu等人[12]针对人脸姿态多样性问题设计出一种使用漏斗型级联结构的人脸检测算法，这种方法在开源项目Seetaface中得到了很好应用。FacenessNet[13]是一种通过分析头发、眼睛、鼻子等人脸关键点是否存在，再判断是否是人脸的方法，该方法对遮挡有较强的鲁棒性。

人脸检测是通用物体检测方法在具体场景下的一个应用。通用物体检测方法Fast R-CNN[14]，Faster R-CNN[15]，单一神经网络物体检测器(Single Shot multibox Detector, SSD)[16]、R-FCN[17]的出现，给人脸检测方法带来了很大的启示。Jiang等人[1]设计出基于Fast R-CNN的人脸检测方法，达到了较理想的效果。CMS-RCNN[18]也是基于Fast R-CNN的方法，不同点是该算法结合了身体上下文信息，提高了小尺寸人脸的检测精度。Hu等人[19]通过扩大检测框的感受野的方式更充分地利用了上下文信息，进一步解决现有检测算法检测小尺寸人脸精度较低的问题。

为了提高人脸检测精度，本文做了以下两点改进：(1)基于单一神经网络物体检测SSD模型结构，提出了一种新型多尺度人脸检测模型；(2)提出一种适用于人脸检测的特征融合模型，有效地引入了上下文信息。本文方法有效地提高了人脸检测精度，特别是小尺寸人脸的检测精度。在FDDB和WIDERFACE数据集上测试和验证都达到了先进的效果。

本文的第2节介绍了单一神经网络物体检测模型SSD；第3节首先通过分析现有引入上下文方法的优缺点，然后提出适合本实验的特征融合方法；第4节详细介绍本文提出的基于特征融合的多尺度人脸检测模型和训练模型的方法；第5节详细介绍本文方法在FDDB和WIDERFACE数据集上的实验结果，实验结果验证了本文方法的合理性与有效性；第6节总结本文工作内容并展望未来工作计划。

2　SSD多尺度检测框架

2.1　多尺度检测模型网络结构

单一神经网络物体检测器SSD是一个1阶段通用物体检测方法，模型结构如图1所示，SSD结构可分为基础网络结构和额外的卷积层结构[16]。基础网络结构是VGG16网络的前5个卷积层，这是用于图像分类的标准结构。在基础网络后面，首先将

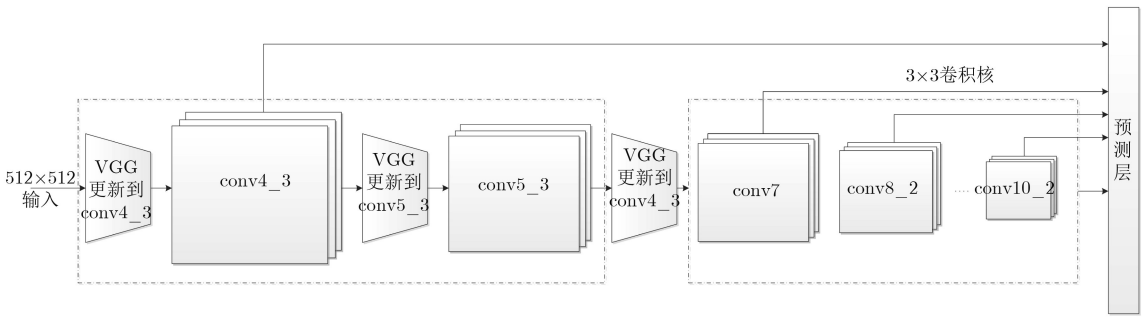


图 1 SSD网络结构

VGG16的两个全连接层改成卷积层，再添加几个卷积层构成额外的卷积层结构。输出是一系列离散检测框的位置信息和相应类别的得分，最后经过非极大值抑制得到最终结果。相比较两阶段检测器Faster R-CNN, SSD去掉了最耗时的人脸框推荐阶段，在保证精度的同时，提升了检测速度。

2.1.1 多尺度检测 SSD[16]方法的核心是在不同特征图上进行检测，模型选择的特征图包括：38×38 (conv4_3), 19×19 (conv7), 10×10 (conv8_2), 5×5 (conv9_2), 3×3 (conv10_2), 1×1 (conv11_2)。这些特征图是逐层递减的，类似于图像金字塔结构，因此可以实现多尺度检测。

2.1.2 用于预测的卷积核 在每一个待检测的特征图上使用3×3大小的卷积核去预测，对于k个通道的大小为$m\times n$的特征图，使用$3\times3\times k$卷积核做卷积操作，预测出物体类别和相对于默认框的偏移量。

2.1.3 不同宽高比的默认离散检测框 每一个默认离散检测框和原始图像的相对位置是固定的。所以，需要预测每一个默认框相对原始图像的位移和相应类别的得分。默认检测框如图2。第1幅图像为带有标签的原图像，第2幅和第3幅图像是和标签匹配的不同的特征图，分别是8×8和4×4大小的特征图。其中虚线框为不同宽高比的默认检测框，假设一个$m\times n$大小特征图上的检测框有4种不同比例，那么一个特征图上有$m\times n\times4$个默认检测框，每个默认检测框需要被预测出4个位置偏移和相应类别的得分。

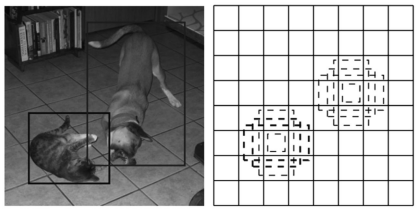


图 2 默认检测框[16]

2.2 损失函数

训练时首先要对标定的真实位置信息和相应类别进行预处理，把这些信息对应到默认的检测框上。根据默认检测框和真实标定的框做Jaccard相似度计算，阈值大于0.5的默认框为正样本，其它为负样本[16]。

SSD采用的损失函数是通过把MultiBox方法[20]扩展成多分类任务得到的，是位置回归和类别置信度的加权求和。

$$L(x,c,l,g)=\frac{1}{N}\left(L_{\mathrm{c}}(x,c)+\alpha L_{\mathrm{l}}(x,l,g)\right) \tag{1}$$

其中，N为匹配到的默认框个数，如果N=0，则设损失函数值为0。位置损失函数采用Smooth L1函数，具体定义详见文献[17]。置信度损失函数采用softmax函数处理多分类问题：

$$L_{\mathrm{c}}(x,c)=-\sum_{i\in P}^{N}x_{ij}\ln\hat{c}_i-\sum\ln\hat{c}_i \tag{2}$$

$$\hat{c}_i=\exp(c_i)\Big/\sum_{j=1}^{k}\exp(c_j) \tag{3}$$

式(3)中的k为样本中类别个数。

3 引入上下文信息的方法

由于受到特征信息不足和分辨率低的影响，现有的检测算法很难从图片中找到小尺寸人脸。而添加身体、头发等上下文信息有助于找到小尺寸人脸[21−24]。人脸检测中引入上下文信息的主要方法有：扩大检测框的感受野和反卷积操作。

扩大感受野的方式：文献[19]通过扩大感受野的方式添加上下文信息，如图3所示，从左到右依次是：(1)不借助上下文信息；(2)使用3倍原尺寸人脸感受野大小检测框添加合适的下文信息；(3)使用固定300像素的检测框添加上下文信息。由图3可以看出借助上下文信息可以很容易找到小尺寸人脸。但是，不合理地引入上下文信息也会增加背景噪声，影响检测的准确性。

反卷积方式：通过反卷积操作可以有效地整合浅层的特征信息[24]，图4为反卷积融合模块，图中的反卷积层是由检测模型的浅层特征图通过反卷积操作得到的，反卷积层包含了丰富的上下文信息，将其与卷积层融合在一起，可以有效地引入上下文信息，提高小目标的检测精度。如图4，所有卷积

图 3 增加上下文信息[19]

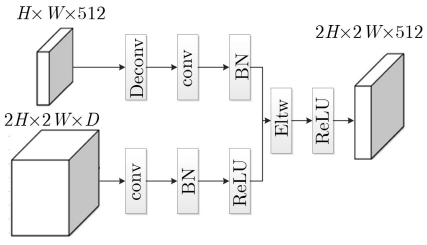


图 4 反卷积融合模块

核的大小是3×3，维数为512，反卷积核大小为2×2，维数为512。

为了有效地引入与人脸相关的上下文信息，本文根据人脸检测的特点，设计了一个新的特征融合模型，本模型类似于编码解码的结构，网络的卷积层为编码部分，网络的反卷积层为解码部分，最后经过特征图融合操作，实现了上下文信息的扩充，最终提高人脸检测精度，特别是小尺寸人脸的检测精度。模型详细设计方法将在本文的4.3节介绍。

4　多尺度人脸检测模型

SSD检测模型的不同检测分支是相互独立的，提高任意分支的检测精度就可以提高整个模型的检测精度。本文提出的基于特征图融合的检测模型即通过提高小尺寸人脸的精度来提高整个网络的检测性能。

本方法采用单一神经网络结构，如图5所示。模型的输入是640×640像素的待检测图像，然后在不同大小的特征图上使用一系列3×3大小的卷积核分别预测人脸的得分和相应位置，其中深层的特征图需要先和浅层的特征图融合引入上下文信息之后再检测(详见4.3节)，网络的输出是不同大小的人脸检测框和相应的位置信息，最后经过非极大值抑制算法得到最终结果。

4.1　基于特征图融合的多尺度人脸检测框架

本文借鉴Faster R-CNN算法基于检测框(anchors)[15]的结构，提出一种新的多尺度人脸检测模型。和两阶段的Faster R-CNN不同的是本文将检测框用到不同的特征图实现多尺度检测。为了弥补小尺寸人脸特征信息少，分辨率低的问题，本文在conv3_3和conv4_3中分别做了特征融合，利用结合上下文信息的方式提高小尺寸人脸的检测精度。模型结构如图5。

由图5可以看出该结构类似于SSD的网络结构，开始使用VGG16[16](本文称作基础网络结构)，在基础网络之后添加了辅助的卷积层。用于检测的特征图是逐层递减的，所以可以实现检测不同尺寸下的人脸。由于深层的特征图感受野小而包含的特征信息多，所以适合检测尺寸相对较小的人脸；浅层的特征图感受野大，适合用于检测尺寸相对较大的人脸。

本文选择conv3_3, conv4_3, conv5_3, conv7, conv8_2, conv9_2用作待检测的特征图，SSD中选择conv4_3用作开始的检测层，该特征图检测框(anchors)步长为8，即检测框中心每移动一次相当于在原图像上滑过8个像素，不适合小尺寸人脸检测。因此本文选择conv3_3做为开始的检测层，该特征图检测框步长为4，步长大小比较利于小尺寸人脸检测。conv3_3到conv9_2的检测框步长由4开始以2为倍数增长，最大的为conv9_2的检测框步长为128像素。

在每个特征图中，本文使用3×3的卷积核做卷积操作，预测默认检测框的4个位置偏移信息和人脸的置信度，类似于SSD预测方法。这使得特征图中每个位置需要$k(c+4)$个3×3的卷积核，具体到本方法，k为不同宽高比的检测框个数，因为只有是人脸或者不是，所以c为2。例如conv3_3中k为1，需要预测出默认框的4个位置信息和2个类别得分。

4.2　检测框的设定

神经网络主要通过加深网络层数来减小特征图的大小，减少计算量和内存的消耗。为了达到多尺度检测，通用的方法是在最后一个特征图使用不同大小的检测框(anchors)，最后采用最大抑制算法得到最后结果。这种方法会增加模型的计算量。本文通过在单一网络的不同特征图使用不同大小的检测框可以达到相同的多尺度检测效果。同时，还可以共享参数，减小计算量，提高了检测速度。

理论上的检测框的感受野是均匀的，该范围内的任意输入都会影响到输出。实际上，越是中间的输入对输出的影响越重，类似一种中心高斯分布[25]。根据上述理论和SSD的检测框设计方法，本文设置出符合多尺度人脸检测的检测框如表1。首先由于

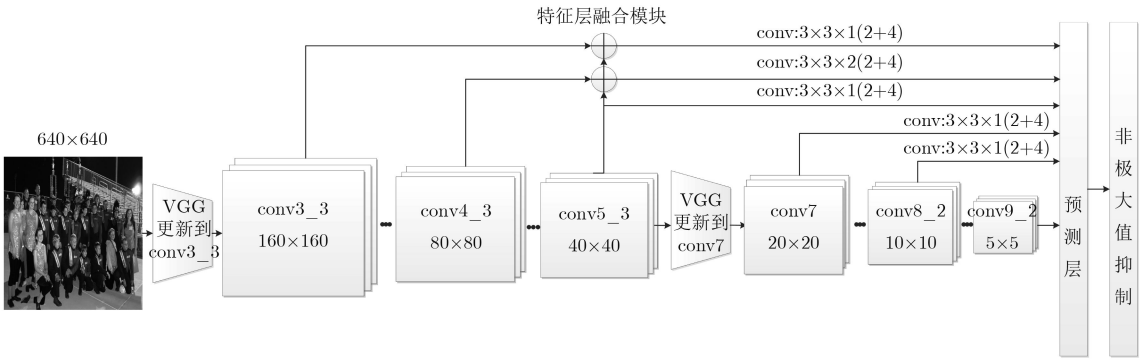


图 5 基于特征图融合的多尺度人脸检测网络结构

人脸形状的特点，本文把检测框的宽高比设置为1。本文定义A_{d}为检测框密度，如式(4)所示，其中A_{s}为检测框尺寸，n为检测框的步长。本文通过设置检测框大小保证每个待检测特征图的A_{d}值为4，主要目的是为保证不同大小的人脸匹配到的检测框密度是相同的。通过合理地设置每个特征图的检测框大小，有效地提高了多尺度人脸的检测效果。

$$A_{\mathrm{d}} = A_{\mathrm{s}}/n \tag{4}$$

表 1 检测框参数

特征层	步长n	检测框大小	宽高比
conv3_3	4	16	1
conv4_3	8	32	1
conv5_3	16	64	1
conv7	32	128	1
conv8_2	64	256	1
conv9_2	128	512	1

4.3 特征图融合模型

SSD结构中不同检测分支是相互独立的，因此容易出现相同物体被不同大小的检测框同时检测出来的问题。本文利用特征图融合的方式可以增加不同层之间的联系，减少重复框的出现；另一方面合理地引入上下文信息可以提高小尺寸人脸的检测精度。综上所述，本文提出一种新的特征图融合模型，如图6所示：图6(a)将conv4_3与conv5_3经过特定方式融合构成一个待检测特征图conv4&5；图6(b)将conv4&5经过特定方式融合构成一个待检测特征图conv3&4。图6中，没有标注的卷积核conv的大小为3×3，维数为512，卷积核conv1的大小是1×1，维数为256，反卷积核Deconv的大小为2×2，维数为512。

4.4 训练

4.4.1 数据集扩充 本文使用的WIDERFACE数据集训练模型。数据集中有3.2×10^4张图片，标定了39.3×10^4张人脸[2]。为了使得训练的模型对不同尺寸的人脸都有较好的鲁棒性，采用如下数据增广方法：

(1) 采用文献[26]中的方法对图像进行色彩失真；

(2) 由于WIDERFACE中的人脸，80%分布在40～140像素。所以分别对原图像进行0.5和2.0分辨率采样，平衡人脸尺寸的分布。

4.4.2 检测框匹配策略 在训练阶段，本方法需要建立标定的真实值和默认检测框的对应关系。本文设定检测框和真实值的重叠率大于一定阈值为正样本(本实验为0.35)。由于人脸尺寸变化是连续的，默认的检测框大小是离散的，所以某些尺寸的人脸和预设的检测框匹配效果不佳。本方法采用以下两个步骤解决这个问题：

(1) 增加检测框的数量，除了conv3_3外的每个待检测特征图增加一个检测框，检测框的大小设为原检测框的0.5倍；

(2) 采用SSD的匹配策略并把阈值从0.50降到0.35，增加匹配到的检测框数量。

经过匹配后，大多数的默认检测框为负样本。这导致供训练的正样本和负样本不平衡。为了解决这个问题，训练时不使用所有的负样本，采用困难样本挖掘方法[26]保持正负样本为1:3的比例。

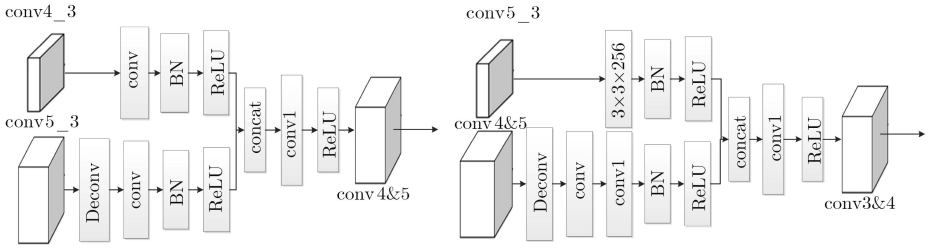
4.4.3 损失函数 本方法将SSD多分类任务的损失函数优化成适用于人脸检测任务的目标函数，如式(5)所示：

$$L\left(x,c,l,g\right)=\frac{1}{N}\left(L_{\mathrm{c}}\left(x,c\right)+\alpha L_{\mathrm{l}}\left(x,l,g\right)\right) \tag{5}$$

其中，N代表匹配到的默认框，当N为0时，本文设置损失为0。c为预测的置信度，l为预测得到的框坐标，g为标定的真实值。位置损失函数采用Smooth L1函数[12]，置信度损失函数采用softmax函数，具体定义见文献[16]。

4.4.4 其它细节 本实验设备为一台装有2个英伟达GTX泰坦x GPU(12G显存)的服务器，本文设置权重衰减为0.0005，动量为0.9。在前8×10^4迭代中使用的学习率为10^{-3}，之后分别使用10^{-4}和10^{-5}再各训练2×10^4次。

训练分两个阶段进行，首先不加特征融合模块，在WIDERFACE数据集上训练12×10^4次，得



(a) conv4_3 和 conv5_3 融合模块　　(b) conv3_3 和 conv4&5 融合模块

图 6 特征图融合模型

到训练好的模型，之后冻结训练好的模型参数，加上特征融合模块，再训练12×10⁴次得到最后的模型。

5　实验结果与分析

5.1　WIDERFACE数据集测试结果分析

WIDERFACE测试集数据集被准确地分为难，适中，简单3种类别，其中“难”子集包含尺寸为10到50像素的人脸，“适中”子集包含尺寸为50到300像素的人脸，“简单”子集包含尺寸大于300像素的人脸，所以该测试集适合测试本文提出的模型。

5.1.1 融合模型合理性验证　首先，为了验证图6模型的合理性，本文还使用相同的融合方法，选择不同的层进行融合，完成如下两个实验：(1)为了增加引入的上下文信息，从更浅的特征图fc6开始进行反卷积得到待融合层，本文将该模型命名为对比模型1；(2)为了减少引入上下文信息，将特征图conv3_3与conv4_3融合形成一个检测分支，将特征图conv4_3与conv5_3融合形成一个检测分支，本文将该模型命名为对比模型2。最后，在WIDERFACE“难”子集上验证融合模型的合理性，验证结果如表2所示。

表 2　不同融合方式的MAP对比结果

模型名称	数据集	MAP
本文的融合型	WIDER	0.879
对比模型1	FACE	0.823
对比模型2	(Hard)	0.836

5.1.2 与其它方法的对比　由于本算法是文献[16,19]方法的改进方法，为了验证本文所提出的改进算法的有效性，本文在相同的设备和相同的数据集上，从检测的精度和速度上与文献[16,19]方法进行了对比，另外本文还与研究人员比较关注的基于Faster-rcnn的人脸检测方法进行了对比，对比结果如表3所示，在WIDERFACE简单、适中和难的子数据集上的精度和召回率曲线如图7所示。可以看出本文方法较其它方法有一定提高。

表 3　实验结果MAP对比

方法	难	适中	简单	检测速度(fps)
Faster-rcnn	0.712	0.845	0.897	<10
SSD-face	0.737	0.882	0.910	<43
HR	0.831	0.914	0.925	<5
本文方法	0.879	0.932	0.934	<35

5.2　FDDB数据集测试结果分析

本文还将在WIDERFACE数据集上训练好的模型在FDDB数据集上进行验证，由于FDDB数据集用的是椭圆标注，而本文训练的模型输出的是矩形框，所以，模型被测试之前需要训练一个将矩形框转换成椭圆框的线性回归模型。在FDDB上测试结果如图8所示，测试结果显示在离散和连续情况下的精度分别是98.4%, 86.3%(MAP)，都达到了较先进的检测效果。

5.3　实验效果

图9为本方法检测效果展示，图中矩形框表示检测器检测出的人脸位置。由图9可以看出本方法还可以在表情夸张、姿态多样化、遮挡和光照恶劣等不理想条件下实现多尺度人脸检测。

6　结束语

本文提出了基于特征图融合的多尺度人脸检测方法，采用单一神经网络结构，在训练时更容易优化。预测阶段，融合后的特征图和其它不同大小的特征图构成6个独立的检测分支，然后分别用一系列3×3大小的卷积核预测人脸得分和相应的位置，实现实时多尺度人脸检测。对比其它优秀的人脸检测方法，本方法在保证精度的同时，提高了检测速度。

本方法是在SSD方法基础上添加了上下文信息，然而上下文信息并不是任何时候都有助于人脸检测，有时额外的上下文信息会引入很多不必要的背景噪声。如何更好地结合上下文信息进一步提高人脸检测精度值得进一步探索。

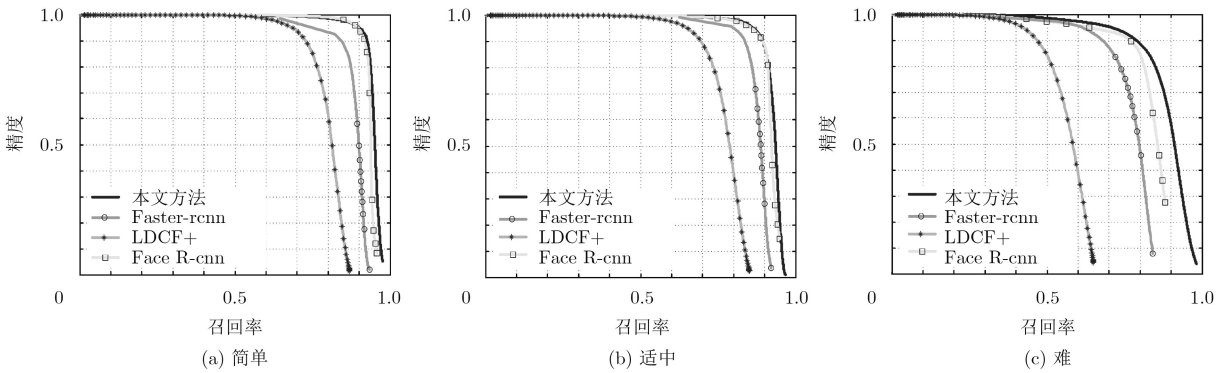
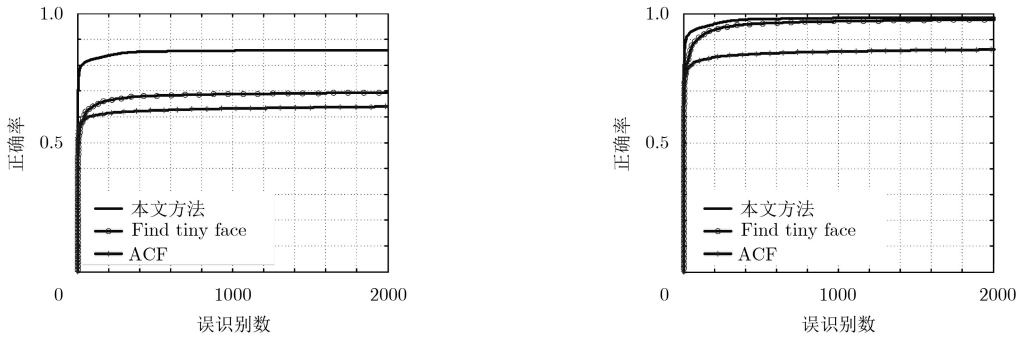
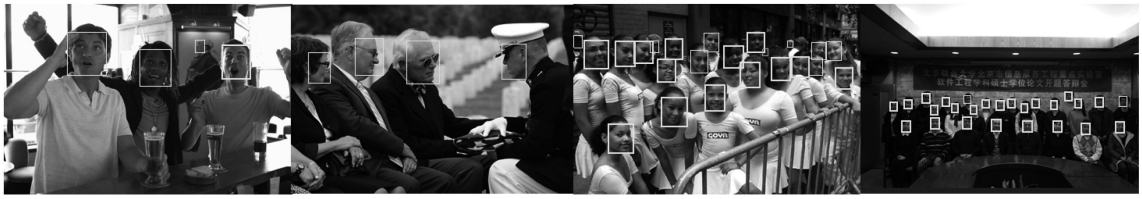


图 7 测试结果曲线



(a) 连续 ROC 曲线　　(b) 离散 ROC 曲线

图 8 FDDB上测试ROC曲线



(a) 表情夸张　(b) 姿态多样　(c) 遮挡　(d) 光照恶劣

图 9 实验效果图

参 考 文 献

[1] JIANG Huaizu and LEARNED M E. Face detection with the faster r-cnn[C]. IEEE International Conference on Automatic Face & Gesture Recognition, Washington, D.C., USA, 2017: 650–657. doi: 10.1109/FG.2017.82.

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

[3] CROSSWHITE N, BYRNE J, STAUFFER C, *et al*. Template adaptation for face verification and identification[C]. IEEE International Conference on Automatic Face & Gesture Recognition, Washington, D.C., USA, 2017: 1–8. doi: 10.1109/FG.2017.11.

[4] MAJUMDAR A, SINGH R, and VATSA M. Face verification via class sparsity based supervised encoding[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1273–1280. doi: 10.1109/TPAMI.2016.2569436.

[5] GAO Yuan, MA Jiayi, and YUILLE A L. Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples[J]. *IEEE Transactions on Image Processing*, 2017, 26(5): 2545–2560. doi: 10.1109/TIP.2017.2675341.

[6] HARIS KHAN M, MCDONAGH J, and TZIMIROPOULOS G. Synergy between face alignment and tracking via discriminative global consensus optimization[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA, 2017: 3791–3799. doi: 10.1109/ICCV.2017.409.

[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, Columbus, USA, 2014: 580–587. doi: 10.1109/CVPR.2014.81.

[8] VIOLA P and JONES M. Rapid object detection using a boosted cascade of simple features[C]. IEEE Computer Society Conference on Computer Vision & Pattern Recognition, Kauai, USA, 2001: 511. doi: 10.1109/CVPR.2001.990517.

[9] LI Jianguo, WANG Tao, and ZHANG Yimin. Face detection using SURF cascade[C]. IEEE International Conference on Computer Vision Workshops, Ontario, Canada, 2012: 2183–2190. doi: 10.1109/ICCVW.2011.6130518.

[10] MATHIAS M, BENENSON R, PEDERSOLI M, *et al*. Face detection without bells and whistles[C]. European Conference on Computer Vision, Zurich, Switzerland, 2014: 720–735.

[11] LI Haoxiang, LIN Zhe, SHEN Xiaohui, *et al*. A convolutional neural network cascade for face detection[C]. Computer Vision and Pattern Recognition. Boston, USA, 2015: 5325–5334. doi: 10.1109/CVPR.2015.7299170.

[12] WU Shuzhe, KAN M, SHAN Shiguang, *et al*. Funnel-structured cascade for multi-view face detection with alignment-awareness[J]. *Neurocomputing*, 2016, 221(C): 138–145.

[13] YANG Shuo, LUO Ping, CHEN C L, *et al*. Faceness-Net: Face detection through deep facial part responses[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 40(8): 1845–1859. doi: 10.1109/TPAMI.2017.2738644.

[14] GIRSHICK R. Fast r-cnn[C]. Proceedings of The IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 1440–1448.

[15] REN Shaoqing, HE Kaiming, 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. doi: 10.1109/TPAMI.2016.2577031.

[16] LIU Wei, ANGUELOV D, ERHAN D, *et al*. SSD: Single shot multibox detector[C]. European Conference on Computer Vision, Amsterdam, Netherlands, 2016: 21–37.

[17] DAI Jifeng, LI Yi, HE Kaiming, *et al*. R-fcn: Object detection via region based fully convolutional networks[C]. Advances in Neural Information Processing Systems, Barcelona, Spain, 2016: 379–387.

[18] ZHU Chenchen, ZHENG Yutong, LUU K, *et al*. CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection[OL]. arXiv preprint arXiv:1606.05413, 2016.

[19] HU Peiyun and RAMANAN D. Finding tiny faces[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii, USA, 2017: 1522–1530. doi: 10.1109/CVPR.2017.166.

[20] ERHAN D, SZEGEDY C, TOSHEV A, *et al*. Scalable object detection using deep neural networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 2147–2154. doi: 10.1109/CVPR.2014.276.

[21] CHEN Chenyi, LIU Mingyu, TUZEL O, *et al*. R-cnn for small object detection[C]. Asian Conference on Computer Vision, Taipei, China, 2016: 214–230.

[22] BELL S, LAWRENCE ZITNICK C, BALA K, *et al*. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2874–2883. doi: 10.1109/CVPR.2016.314.

[23] WONG R Y and HALL E L. Sequential hierarchical scene matching[J]. *IEEE Transactions on Computers*, 1978, 27(4): 359–366. doi: 10.1109/TC.1978.1675108.

[24] FU C Y, LIU Wei, RANGA A, *et al*. DSSD: Deconvolutional single shot detector[OL]. arXiv preprint arXiv:1701.06659, 2017.

[25] WEI Xiang, ZHANG Dongqing, YU H, *et al*. Context-aware single-shot detector[C]. IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, USA, 2018: 1784–1793. doi: 10.1109/WACV.2018.00198.

[26] HOWARD A G. Some improvements on deep convolutional neural network based image classification[OL]. arXiv preprint arXiv:1312.5402, 2013.

刘宏哲：女，1971年生，教授，硕士生导师，研究方向为数字图像处理、旅游信息化.

杨少鹏：男，1990年生，硕士生，研究方向为模式识别.

袁家政：男，1971年生，教授，博士生导师，研究方向为数字图像处理、视觉计算与定位技术.

王雪峤：女，1986年生，讲师，研究方向为模式识别.

薛建明：男，1992年生，硕士生，研究方向为模式识别.