基于全卷积神经网络的多尺度人脸检测

储慧敏[1],杨会成[1],张　丽[2],潘　玥[1]
(1. 安徽工程大学 电气工程学院,安徽 芜湖 241000;
2. 安徽华东广电技术研究所有限公司,安徽 芜湖 241000)

摘　　要:如何快速而准确地定位到人脸,针对这个问题,提出了一种基于全卷积神经网络的多尺度人脸检测方法. 首先用全卷积层替换 VGG 网络中的全连接层,然后用二分类代替分类层,最后进行该算法下的人脸检测,通过对待检测的图片进行多尺度变换并将其输入到全卷积神经网络中,得到相应的概率矩阵,人脸图框通过非极大值抑制法获取. 试验结果表明,该算法的准确率较高,检测时间短,性能较好.

关　键　词:卷积神经网络;人脸检测;VGG;多尺度变换
中图分类号:TP391.41;TP183　　**文献标识码**:A　　**文章编号**:1673 - 1670(2019)05 - 0048 - 06

0　引言

判别给定图像上是否存在人脸是人脸检测的目的,若有人脸,就将人脸框出来. 在现实运用中,人脸检测意义重大. 在 20 世纪 70 年代就早已有人开始着手研究,但那时候科学技术比较落后以及需求有限,所以直到 20 世纪 90 年代,人脸检测技术才开始加速发展. 早期的人脸检测主要基于模型匹配[1],即模板图像与目标图像相同位置的像素灰度值或颜色值进行比较,并计算各灰度差或颜色差的总和. 基于特征的检测往往需要先验知识,因此特征的选取与界定需要研究者来决定. 人脸检测发展的中期奠定了许多优秀检测方法的思想基础,如主成分分析(PCA)[2]、统计学习与神经网络在人脸检测中的应用. 统计学习[3]的基本思想是从大量给定的正向的例子和反向的例子集合中通过推导总结得到接受最大范围正例并排斥最大范围反例,具体到人脸检测的问题上,就是将样本分为人脸样本和非人脸样本训练,进而得到分类算法. 这种方法和基于特征的人脸检测方法相比较,优势在于能够对大量的人脸样本进行学习和测试,然后将人脸的内在联系弄清楚,而这并不需要依托于先验知识,可有效减少由知识不准确或不完整带来的检测错误,从而使检测效果得到提升. 但该方法受样本空间的影响比较大.

基于统计的人脸检测方法存在着不少难以解决的问题,而卷积神经网络可以自动提取人脸特征并能够迅速而准确地将背景信息等一些干扰信息剔除掉,一些研究者利用这个特点对图像进行检测,取得了良好的效果. 尽管基于卷积神经网络的人脸检测方法取得了不错的效果,但是卷积神经网络的模型过于单一,所以检测过程中仍存在着不少问题. 因为单个的卷积神经网络模型需要同时具有特征提取、降维和分类这三项功能,因此所需要设计的网络结构模型就会比较复杂,这也就导致检测所需要的时间比较长. 近年来以 RCNN、Fast RCNN、Faster RCNN 和 YOLO 等为代表的卷积神经网络算法表现出了很好的检测性能,能够快速而准确地输出物体的类别和物体的回归框,对人脸检测的发展起了很大的推动作用[4-7].

1　卷积神经网络

卷积神经网络[8]是深度学习模型中的一种多层人工神经网络,包括输入层、卷积层(convolution layer)、池化层(pooling layer)、全连接层(FC)和输出层(图 1). 卷积和池化是卷积神经网络中特别重

收稿日期:2019 - 06 - 24
基金项目:安徽省高校自然科学研究重点项目(KJ2018A0122)
作者简介:储慧敏(1995—　),女,安徽省潜山市人,安徽工程大学电气工程学院硕士研究生,主要从事图像处理与模式识别方面的研究.

要的两个步骤.

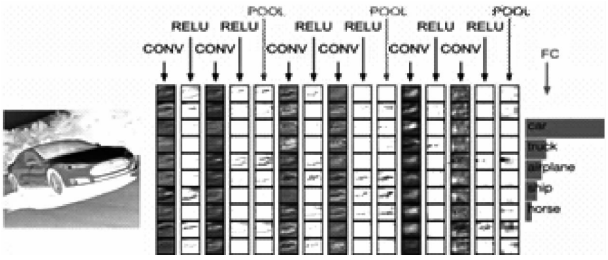

图 1　卷积神经网络结构

卷积是数学分析中的一种重要运算,在图像处理中一般用一维卷积或二维卷积,笔者采用的是二维卷积. 给定一个图像 $X \in R_{M\times N}$,和滤波器 $W \in R_{M\times N}$,一般 $m \ll M, n \ll N$,其卷积为

$$y_{ij} = \sum_{u=1}^{m}\sum_{v=1}^{n} w_{uv} \cdot x_{i-u+1,j-v+1}. \quad (1)$$

在卷积层中,对卷积核与特征图进行卷积操作,然后将前面一层中不同的特征图中位置相同的局部特征提取出来,再将卷积后的局部进行加权和加偏置,最后通过非线性激活函数进行映射,继而得到与之相对应的映射值. 提取人脸图像的特征是进行卷积操作的本质,卷积的特点就是可以局部连接和权值共享. 其中,卷积核[22]是一个可训练的权值矩阵,若要提取不同的特征,就设置不同的卷积核参数. 在实际应用中,卷积后将会获得大量的特征映射图,如果将所有的特征都输入到分类器中,这个计算量是非常巨大的. 所以,对卷积层进行池化操作非常有必要. 最大值池化(max pooling)、均值池化(mean pooling)和随机池化(stochastic pooling)是人脸检测中常用的池化操作.

2　VGGNet 网络

2.1　VGG 的网络结构

VGGNet[9-10]是牛津大学视觉几何组 VGG(Visual Geometry Group)小组开发的一个著名的应用于图片分类的卷积神经网络模型. 在 2014 年 ImageNet 大规模视觉识别挑战赛 ILSVRC(ImageNet Large Scale Visual Recognition Challenge)上打败了众多强劲的对手,最终取得第 2 名的好成绩. 该模型图片特征提取能力非常优秀,在多个迁移任务中的表现甚至比当年的第 1 名都要优秀,所以,要想从图像中提取特征,首先想到的就是 VGGNet 模型. 该网络是从 Alexnet 网络[11]发展而来的,网络结构主要的变化是采用了更小的卷积核以及更多的网络层数. 具体模型如图 2 所示.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
Input(224*224 RGB image)					
Conv3-64	Conv3-64 LRN	Conv3-64 Conv3-64	Conv3-64 Conv3-64	Conv3-64 Conv3-64	Conv3-64 Conv3-64
maxpool					
Conv3-128	Conv3-128	Conv3-128 Conv3-128	Conv3-128 Conv3-128	Conv3-128 Conv3-128	Conv3-128 Conv3-128
maxpool					
Conv3-256 Conv3-256	Conv3-256 Conv3-256	Conv3-256 Conv3-256	Conv3-256 Conv3-256 Conv1-256	Conv3-256 Conv3-256 Conv3-256	Conv3-256 Conv3-256 Conv3-256 Conv3-256
maxpool					
Conv3-512 Conv3-512	Conv3-512 Conv3-512	Conv3-512 Conv3-512	Conv3-512 Conv3-512 Conv1-512	Conv3-512 Conv3-512 Conv3-512	Conv3-512 Conv3-512 Conv3-512 Conv3-512
maxpool					
Conv3-512 Conv3-512	Conv3-512 Conv3-512	Conv3-512 Conv3-512	Conv3-512 Conv3-512 Conv1-512	Conv3-512 Conv3-512 Conv3-512	Conv3-512 Conv3-512 Conv3-512 Conv3-512
maxpool					
FC-4096					
FC-4096					
FC-4096					
Soft-max					

图 2　VGGNet 各级别网络结构

Karen Simonyan 和 Andrew Zisserman 共设计了 A、A – LRN、B、C、D、E 6 种网络模型,对比实验表明,模型 D 效果最好,笔者采用的 VGGNet 网络模型就是 D 模型,共有 16 层. D 模型主要分为 13 层卷积层和 3 层全连接层,为了强化特征提取效果,每个卷积核都选用 3 * 3 大小,由图 1 可知,首先为 13 个卷积层:第 1、2 层卷积层各有 64 个特征图,第 3、4 层卷积层各有 128 个特征图,第 5 至第 7 卷积层各有 25 个特征图,第 8 至 13 层卷积层各有 512 个特征图. 5 个子采样层分别位于第 2 层与第 3 层卷积、第 4 层与第 5 层卷积、第 7 层与第 8 层卷积、第 10 层卷积与第 11 层卷积、第 13 层卷积与第 1 层全连接层之间,缩放因子都为 2. 最后是 3 个全连接层,第 1、2 层分别有 4 096 个神经元,第 3 层的神经元个数由分类数目决定,这里是 1 000 个图像类别,因此设置 1 000 个神经元.

为了加快 VGG 网络训练时的收敛速度,先训练网络结构较简单的级别 A,然后用 A 网络的权重对后面的复杂模型进行初始化. VGG 采用 Multi – Scale 方法[12]对其进行预测,将图像的尺寸调整到 Q,并将图片输入到卷积网络对其进行计算. 分类时采用滑窗的方式,对各组分类结果求其平均值,再对调整到尺寸 Q 的结果进行平均值计算,这就

是最后所需要的结果,图片数据的利用率和预测准确率得到提高. 在训练中,数据增强使用的也是 Multi – Scale 方法,将原始图像的尺寸调整到 S,然后再随机裁切 224 * 224 的图片,这样做可以有效地防止模型过拟合. 试验中,S 的取值范围为 256 ~ 512,使用 Multi – Scale 得到多组数据,并将这些数据集合在一起进行训练. 表 1 是训练时得到的结果,训练时 D 和 E 的错误率为 7.5%. 最终作者参赛的网络结构是融合了 Multi – Scale 的 D 网络和使用Single – Scale 的 6 个不同等级的网络,错误率有所下降,为 7.3%. 但是后来他们发现有别的方法可以降低错误率,所以该团队反复进行试验测试,最低错误率为 6.8% 左右. 经过反复试验的对比,他们做出了如下总结:1) LRN 层在该网络结构中起不了多大的作用,可以去除;2)网络越深其检测效果越好;3) 1 * 1 的卷积也有效果,但是没有 3 * 3的卷积好,卷积核越大学习的特征空间越大.

表 1 各级别 VGGNet 训练时的错误率

ConvNet config	Smallest image side		top – 1 val error/%	top – 5 val error/%
	Train(S)	Train(Q)		
B	256	224,256,288	28.2	9.6
C	256	224,256,288	27.7	9.2
	384	352,384,416	27.8	9.2
	[256;512]	256,384,512	26.3	8.2
D	256	224,256,288	26.6	8.6
	384	352,384,416	26.5	8.6
	[256;512]	256,384,512	24.8	7.5
E	256	224,256,288	26.9	8.7
	384	352,384,416	26.7	8.6
	[256;512]	256,384,512	24.8	7.5

2.2 VGGNet 网络的特点

VGG 网络结构较简洁,由图 1 和表 1 可以看出,max pooling 将相邻层隔开,而且激活单元都用的是 ReLU 函数[13];VGG 的一个重要特点是小卷积核和小池化核,使用小卷积核一方面可以减少参数,另一方面也可以减少过拟合,通过增加卷积子层的数目来防止性能降低;同 AlexNet 网络相比,VGG 网络的通道数有所增加,更多的信息将会被提取出来.

3 多尺度变换人脸检测

传统目标检测算法的窗口大小是固定的,图片保持不动,窗口滑动,然后提取各个窗口的特征,最后将这些特征输入分类器. 若待检测目标较简单,则传统的目标检测算法检测效果会很好. 但是实际应用时,图片往往很复杂,传统的目标算法容易造成检错或漏检. 所以笔者进行人脸检测时采用卷积神经网络,先训练分类网络,再进行多尺度变换,本文的多尺度变换通过构造高斯金字塔模型来实现.

3.1 金字塔模型

图像金字塔[14-16],顾名思义,是由一组图像构成的形似金字塔的一种结构,而这组图像的分辨率和尺寸有规律地变化,逐渐变大或逐渐减小,高斯金字塔和拉普拉斯金字塔是最常用的模型. 高斯金字塔在达到终止条件前一直向下降,对图像进行采样,而拉普拉斯金字塔与之相反.

采样时,对图像进行低通滤波后再作隔行隔列降采样,这样所得到的图像就是下一级图像. 因此,可以通过以下方式来构造第 k 层图像 G_k:先将 G_{k-1} 和窗函数 $W(m,n)$ 进行卷积运算,再将得到的结果进行采样,即:

$$G_k = \sum_m \sum_n W(m,n) G_{k-1}(2x+m, 2y+n) . \tag{2}$$

其中,x 要小于金字塔图像中第 k 层行数,而 y 要小于其列数.

$W(m,n) = h(m) \cdot h(n)$ 为 5×5 窗口函数,h 为高斯密度分布函数,窗口函数 $W(m,n)$ 为:

$$W(m,n) = \frac{1}{256}\begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix} . \tag{3}$$

由此可以看出,金字塔分解得到的各层图像 $G_0, G_1, \cdots, G_N$ 构成图像的多尺度空间,可以任意选取其中的 G_k, G_{k+1}, G_N 来构成所需要的金字塔.

本文算法采用的是高斯金字塔,先将图像通过一个低通滤波器对其进行平滑处理,然后对平滑处理后的图像进行抽样或者插值,就会获得尺度变换后的图片. 对高斯函数与数字图像进行卷积计算,该结果可以将其定义成尺度空间,即:

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y), \quad (4)$$

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2}e . \quad (5)$$

其中,* 是卷积运算,(x,y) 是像素点位置,σ 是尺度因子.

3.2　微调 VGGNet 网络

将 VGGNet 网络 softmax 改为 2 层,即分类层变为二分类. 因此将损失函数改写为:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)}\lg(h_\theta(x^{(i)})) + (1-y^{(i)})\lg(1-h_\theta(x^{(i)})) . \quad (6)$$

其中,m 为训练时批次大小.

将 VGGNet 的全连接层改为全卷积层是因为实际检测中图片的大小是不能确定的,而使用全连接层的条件是图片大小必须是确定的,这是因为全连接层的神经元数目是确定的,故而替换成全卷积层可以不受图片尺寸的限制.

3.3　人脸检测

将已经训练好的 VGGNet 网络改成全卷积神经网络,然后对目标图像进行多尺度变换,并构建高斯金字塔. 将图像输入到该网络中,对其进行多尺度变换,并将其输入到全卷积神经网络中,得到相应的概率矩阵,接着使用非极大值抑制算法,得到的概率最大的区域即为人脸区域,最后标出人脸.

4　实验结果与分析

4.1　训练及检测

笔者从数据集 Wider Face 上采集训练数据,该数据集有 32 203 个图像和 392 703 个人脸图像,这些图像的尺寸、光照、姿态等都不尽相同. 随机选取其中的 16 800 张相片和 16 688 张图片分别作为正、负样本,再从数据集中随机选取 8 384 张相片作为测试所要的集合. 在相同数据集下分别对 VGGNet、CaffeNet[17]、GoogleNet[18] 进行训练测试. 每个网络各训练 30 次,随着迭代次数的增加,模型的准确率逐渐稳定. 由表 2 可以看出,在损失值上,VGGNet、CaffeNet、GoogleNet 的训练损失值分别在0.004和0.14、0.006 和 0.7、0.004 和 0.65 之间波动,测试损失值分别降低到 0.032、0.033、0.041. 在检测率上,VGGNet 网络的训练准确率为 99.22%,CaffeNet 网络和 GoogleNet 网络的训练准确率分别为 98.99% 和 98.81%,VGGNet 网络的准确率要高于 CaffeNet 网络和 GoogleNet 网络. 综上所述,VGGNet 网络准确率高、结构复杂度一般、泛化能力强、损失值低,相同条件下训练时间将近减少一半,所以,笔者采用 VGG 网络.

表 2　3 种算法效果对比

网络	损失值		检测率/%
	训练	检测	训练
VGGNet	0.004 ~ 0.14	0.032	99.22
CaffeNet	0.006 ~ 0.7	0.033	98.99
GoogleNet	0.004 ~ 0.65	0.041	98.81

4.2　检测结果

笔者测试用的 500 张图片均来自 AFLW 数据集,这些照片各方面都有所差异. 为了验证本文算法的有效性,将其与基于面部特征的检测算法[19]、基于模板匹配算法[20]、本文算法无多尺度[21] 变换时进行比对,表 3 中的检测率、误检率、漏检率及检测时间为本次检测结果. 检测率等于检测所得到的有人脸的样本数除以测试集样本数,误检率等于非人脸检测为人脸的样本数除以测试集样本数,漏检率等于没有检测到的人脸的样本数除以测试集样本数. 从表 3 可知本文算法检测率更高,误检率更低,检测时间更短,泛化能力更好. 与本文算法无尺度变换的情况相比,检测率得到很大提高,漏检率降低,表明多尺度变化在算法中起着很重要的作用. 本文算法只需要运行一次,检测速度得到大大加快. 图 3 为本文所提的算法检测效果.

表 3　算法性能效果对比

算法	检测率/%	误检率/%	漏检率/%	单张检测时间/s
文献[11]	95.69	8.23	4.31	2.367
文献[12]	96.40	11.18	3.64	1.975
无尺度变换	89.77	6.56	10.21	0.572
本文算法	98.28	5.68	1.75	0.972

5　结论

笔者基于 VGGNet 网络结构对人脸进行检测,该模型结构简单、训练容易且检测效果好. 将 VGGNet 的全连接层改为全卷积层是因为实际检测中图片的大小是不能确定的,而使用全连接层的条件是图片大小必须是确定的,这是因为全连接层的神

经元数目是确定的,故而替换成全卷积层可以不受图片尺寸的限制. 实验结果显示,本文算法准确率高达 99.37%,检测时间短、检测效果良好,可用于实际检测. 笔者所提出的算法不仅在人脸识别系统中起着重要作用,还在别的方面有着重要的应用价值.

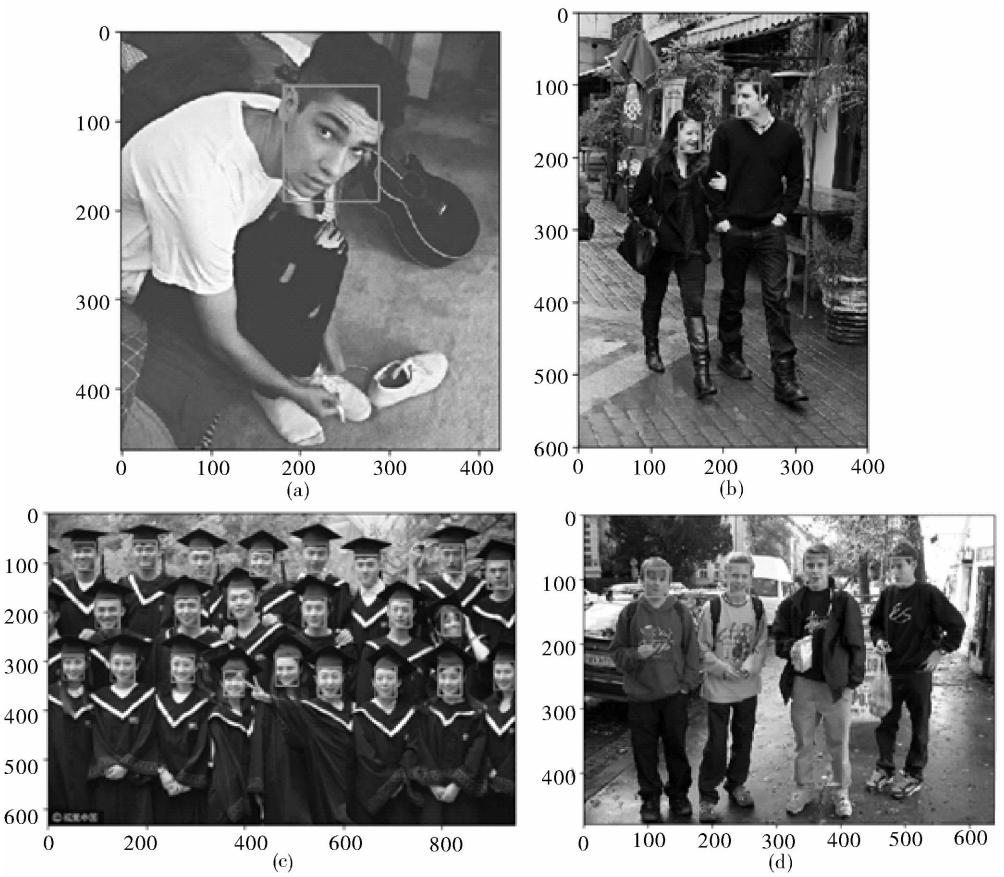

图 3 人脸检测效果展示

参考文献:

[1] ZHENG C H, LIU B, ZHOU Y. Face detection algorithm based on scale – independent cascade convolution neural network [J]. Application research of Computers, 2019, 36(3):55 – 60.

[2] CAO J J, KWONG S, WANG R. A Noise – detection based Ada – Boost algorithm for mislabeled data[J]. Pattern Recognition, 2012, 45(12):4451 – 4465.

[3] DAS J, ROY H. Human face detection in color images using HSV color histogram and WLD[C]. Tetovo:6th International Conference on Computational Intelligence, Communication Systems and Networks, 2014:198 – 202.

[4] VIOLA P, JONES M J. Robust real – time face detection [J]. International Journal of Computer Vision, 2004, 57(2):137 – 154.

[5] LIENHART R, MAYDT J. An extended set of Haar – like features for rapid object detection[C]. Rochester:International Conference on Image Processing, 2002:900 – 903.

[6] LI S Z, ZHU L, ZHANG Z Q, et al. Statistical learning of multi – view face detection[C]. Copenhagen:7th European Conference on Computer Vision, 2002:67 – 81.

[7] WANG Y Q. An analysis of the Viola – Jones face detection algorithm[J]. Image Processing Online, 2014, 4: 128 – 148.

[8] MUKHERJEE S, SAHA S, LAHIRI S, et al. Convolutional Neural Network based face detection[C]. Kolkata: 1st International Conference on Electronics, Materials Engineering and Nano – Technology, 2017:1 – 5.

[9] JONES M J, VIOLA P. Fast multi – view face detection [EB/OL]. https://www.researchgate.net/publication/228362107_Fast_multi – view_face_detection.

[10] GIRSHICK R. Fast R – CNN[C]. Santiago:2015 IEEE International Conference on Computer Vision, 2015:1440 – 1448.

[11] REN S Q, HE K M, GIRSHICK R, et al. Faster R –

CNN:towards real-time object detection with region proposal networks[C]. Montreal:29th Conference on Neural Information Processing Systems,2015:91 – 99.

[12]SALAH A, NADIF M. Social regularized von Mises-Fisher mixture model for item recommendation[J]. Data Mining and Knowledge Discovery, 2017,31(5):1218 – 1241.

[13]SEO Y D, KIM Y G, LEE E J, et al. Personalized recommender system based on friendship strength in social network services[J]. Expert Systems with Applications, 2017, 69:135 – 148.

[14]MA H, JIA M, ZHANG D, et al. Combining tag correlation and user social relation for microblog recommendation [J]. Information Sciences, 2017, 385:325 – 337.

[15]ZHANG Z J, LIU H. Social recommendation model combining trust propagation and sequential behaviors[J]. Applied Intelligence, 2015, 43(3):695 – 706.

[16]SHI M, LIU J X, ZHOU D. A hybrid approach for automatic mashup tag recommendation[J]. Journal of Web Engineering, 2017, 16(7/8):676 – 692.

[17]BEDI P, VASHISTH P. Empowering recommender systems using trust and argumentation[J]. Information Sciences,2014,279:569 – 586.

[18]LI W M, YE Z B, XIN M J, et al. Social recommendation based on trust and influence in SNS environments[J]. Multimedia Tools and Applications, 2017, 76(9): 11585 – 11602.

[19]CUI L Z, DONG L Y, FU X H, et al. A video recommendation algorithm based on the combination of video content and social network[J]. Concurrency and Computation:Practice and Experience,2017, 29(14):101 – 108.

[20]LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient – based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278 – 2324.

[21]BELDJOUDI S, SERIDI H, ZUCKER C F. Personalizing and improving resource recommendation by analyzing users preferences in social tagging activities[J]. Computing and Informatics,2017, 36(1):223 – 256.

(责任编辑:王彦江)

Multi-Scale Face Detection Based on Full Convolution Neural Network

CHU Huimin[1], YANG Huicheng[1], ZHANG Li[2], PAN Yue[1]

(1. School of Electrical Engineering, Anhui Polytechnic University, Wuhu, Anhui 241000, China;
2. Anhui Huadong Photoelectric Technology Institute Co. Ltd., Wuhu, Anhui 241000, China)

Abstract: In the report, aimed at detecting faces quickly and accurately, a multi-scale face detection method based on full Convolutional Neural Network(CNN) is proposed. Firstly, the full connectivity layer of the convolutional neural network model VGG is changed to full convolution layer. Secondly, the layer is divided into two categories of face and non-face. Finally, when the trained classification model is used for face detection, the image to be detected is input to the full convolutional network through multi-scale transformation to obtain the probability characteristic figure, and the most accurate face frame is obtained by the inhibition of non-maximal value. The experimental results show that the proposed algorithm has high detection accuracy, short detection time and good performance in face detection.

Key words: convolutional neural network, face detection, VGG, multi-scale transformation