文章编号：2095-2163(2021)03-0172-05 中图分类号：TP391 文献标志码：A

一种改进的多任务级联网络人脸检测算法研究

周 航，蔡茂国，吴 涛，沈冲冲
（深圳大学 电子与信息工程学院，广东 深圳 518060）

摘 要： 传统人脸检测算法在复杂环境背景下一直存在着检测准确率及效率低等问题。近年来，得益于人脸数据集的增长以及计算机硬件的极速发展，使用深度神经网络的人脸检测算法在准确度方面已有很大提升，但使用的模型结构越来越复杂，检测速度也相对变慢。本文提出一种改进的多任务卷积神经网络（Multi-task convolutional neural networks，MTCNN）算法。在制造数据集时更改 *IOU* 阈值参数，来获取更多、更精确的人脸样本；对与置信度损失有关的交叉熵损失函数和与偏移量损失有关的均方差函数求均值，使得整个网络收敛得更加平稳。经在 AFW、PASCAL 以及 FDDB 数据集上实验，与传统算法相比，该算法在保证实时性的同时提升了检测准确率，可应用于追求更高准确率的人脸检测系统。

关键词： 深度神经网络；MTCNN；人脸检测

Research on an improved Multi-task Cascade Network face detection algorithm

ZHOU Hang, CAI Maoguo, WU Tao, SHEN Chongchong
(College of Electronics and Information Engineering, Shenzhen University, Shenzhen Guangdong 518060, China)

【Abstract】 Traditional face detection algorithms have always had problems such as low detection accuracy and low efficiency in the context of complex environments. In recent years, it has benefited from the growth of face data sets and the rapid development of computer hardware. Face detection algorithms using deep neural networks are in accuracy. This aspect has been greatly improved, but the structure of the model used is becoming more and more complex. The detection speed is relatively slow. Therefore, it is very important to design a detection model that takes into account both accuracy and real-time performance. This paper proposes an improved multi-task convolutional neural networks (MTCNN) algorithm. The first in the manufacturing data sets changes the *IOU* threshold parameter to obtain a more accurate face samples. Secondly, the cross emotion loss function related to the confidence loss and the average of the mean square error function related to the offset loss make the convergence of the entire network more stable. After experiments on the AFW, PASCAL and FDDB data sets, compared with traditional algorithms, this algorithm improves the detection accuracy while ensuring real-time performance, which could be applied to face detection systems pursuing higher accuracy.

【Key words】 deep neural network; MTCNN; face detection

0 引 言

人脸检测已广泛应用于人们的日常生活中，如公共刑侦追逃[1]、考勤打卡、金融机构的门禁控制、自动驾驶以及机器人等都用到了人脸检测技术。人脸检测是计算机视觉和模式识别的重要应用方向之一，是解决与面部相关工作的前提。例如，人脸识别、人脸表情识别、活体检测[2]等。将检测到的人脸图像提取出来，作为后续工作的输入信息，可以大大减少计算量。面部检测算法可分为两大类：基于手工特征提取的人脸检测算法和基于深度学习的人脸检测算法。Viola 等人[3]提出的级联人脸检测算法，利用 Haar-Like 特征和 AdaBoost 级联分类器，实现了良好的实时性能。然而，文献[4]表明，即使检测系统提取更高维的特征和分类器，在人脸图像发生较大变化的情况下，也可能严重退化。除此之外，文献[5]中介绍了一种用于面部检测的可变形部分模型，但其计算量较为复杂，并且需要在训练阶段对图片进行复杂的标注。

近年来，卷积神经网络（CNN）在各种计算机视觉任务（如，图像分类[6]、人脸识别[7]）中都取得了重大进展。受深度学习方法在计算机视觉任务中取得巨大成功的启发，一些研究开始采用深度卷积神经网络进行人脸检测。如：Yang 等人[8]训练了深度卷积网络，来进行面部属性识别，以获取面部区域的人脸特征，从而进一步生成面部候选窗口。但由于其复杂的 CNN 结构，该方法在实践中非常耗时。Li 等人[9]使用级联 CNN 进行人脸检测，但需要从人脸

作者简介：周 航（1997-），男，硕士研究生，主要研究方向：图像处理、活体检测；蔡茂国（1965-），男，博士，教授，主要研究方向：图像处理；吴 涛（1995-），男，硕士研究生，主要研究方向：图像处理、人脸识别；沈冲冲（1995-），男，硕士研究生，主要研究方向：目标检测、深度学习。

收稿日期：2021-01-19

检测中进行边界框校准，需要额外的计算量，并忽略了人脸界标定位与边界框回归之间的内在关联。与此同时，面部对齐也吸引了研究者的广泛兴趣。该领域的研究工作大致可分为 2 类：基于回归方法[10]和模板匹配方法[11]。Zhang 等人[12]提出，利用深度 CNN 作为辅助特征来增强面部对齐性能。然而，以往大多数人脸检测和对齐方法都忽略了这 2 个任务之间的内在联系。例如，在文献[13]中，将像素差分特征与随机森林相结合，用于对齐和检测，但这些手动设计的特征在一定程度上限制了其性能。Zhang 等人[14]使用多任务 CNN 来提高多视角人脸检测的准确性，但检测回归受到弱人脸检测器产生的初始检测窗口的限制。另一方面，在训练过程中，提取样本对于提高检测器的性能非常重要。期望设计自动提取高级语义特征用于人脸检测的方法，能够自动适应当前的训练状态，并针对不同环境下不同类型的人脸进行增强。

基于以上研究，本文对多任务级联网络 MTCNN 做出相应改进。首先，在生成数据集时，通过更改 *IOU* 阈值来获取更多、更精确的人脸样本，为后续级联网络的有效训练提供数据基础，其次，对与置信度损失有关的交叉熵损失函数和与偏移量损失有关的均方差函数求均值，使得整个网络收敛的更加平稳。

1　相关工作

1.1　图像金字塔

人脸检测是一种单类型的多目标检测。检测到的图片类型为 2 种。一种图片仅包含一张脸，一种图片包含多张脸。本文的深度学习基于仿生学、计算机视觉和人眼。同样，通过扫描感受野（Human field of vision）来寻找目标，当感受野特别大而目标特别小时，人和机器将无法识别目标区域，此时需要放大图片或缩小感受区域。人脸检测是从图像中找到人脸，为了检测图片中的人脸，需要通过卷积神经网络来实现。目前，有 2 种方法可以检测所有脸部。一种是使用不同的感受野，从小到大扫描大小不变的图片；另一种是保持恒定大小的感受野来扫描多张不同大小的图片。本文使用后一种作为人脸检测的算法，即图像金字塔，如图 1 所示。图像金字塔只在侦测 PNet 网络时应用，通过使用缩放比率为 0.709 的参数，不断缩放图片，最终使得图片的最小边长大于等于 12。因为 PNet 最小输入图片的 *size* 为(3,12,12)。

图 1　图像金字塔

Fig. 1　Image pyramid

1.2　网络结构

本文采用改进的多任务卷积神经网络（Multi-task convolutional neural networks，MTCNN）进行人脸检测和提取，这是中国科学院深圳研究院在 2016 年提出的人脸检测任务的多任务神经网络模型。该模型的网络结构主要采用 3 级级联网络，并采用候选框加分类器的思想实现快速高效的人脸检测。这 3 个级联的网络分别是：快速生成候选窗口的 PNet、进行高精度候选窗口过滤选择的 RNet 和生成最终边界框的 ONet。网络结构如图 2 所示。

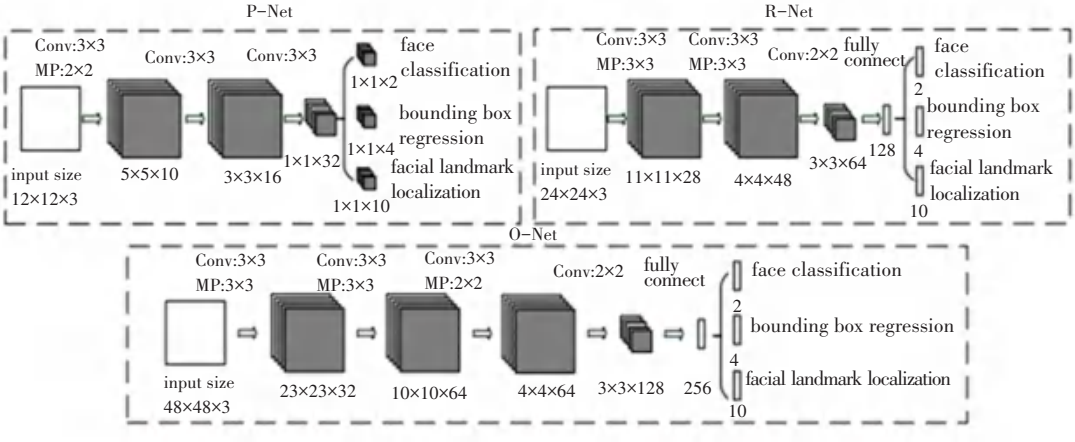

图 2　MTCNN 网络结构图

Fig. 2　MTCNN network structure diagram

MTCNN 的 PNet 全称为 Proposal Network，其基本构造是一个全卷积网络，对图像金字塔通过 FCN 进行初步特征提取与标定边框，并进行 Bounding-Box Regression 调整窗口与非极大值抑制（Non-Maximun Suppression，NMS）进行大部分窗口的过滤，该部分最终输出很多张可能存在人脸的区域，并将这些区域输入 RNet 加以进一步处理。

RNet 全称为 Refine Network，其基本的构造是一个卷积神经网络。因为 PNet 的输出只是具有一定可信度的人脸区域，在此网络中将对输入进行细化选择，并且舍去大部分的错误输入。在此，使用边框回归和面部关键点定位器进行人脸区域的边框回归，此后将输出较为可信的人脸区域，提供给 ONet 使用。与 PNet 使用全卷积输出的 1×1×32 的特征对比，RNet 在最末端的一个卷积层后使用了一个 128 的全连接层，在保留更多图像特征的同时，准确性也优于 PNet。

ONet 全称为 Output Network，基本结构是一个较为复杂的卷积神经网络。该网络的输入特征更多，在网络结构的最末端也同样是一个更大的 256 的全连接层。其中保留了更多的图像特征，同时进行人脸判别和人脸区域边框回归，基于此将输出人脸区域的左上角和右下角的坐标。ONet 具有更多的特征输入、更复杂的网络结构和更好的性能，这一层的输出将作为最终的网络模型输出。

1.3 数据集和标签制作

训练样本采用的是以 CelebA 数据集为基准进行偏移而成。CelebA 是香港中文大学的开放数据集，其中包含 10 177 个名人的 202 599 张照片。通过对CelebA数据集样本的人脸坐标框进行随机偏移，可以获得不同类型的数据样本。为了更好地区分样本，本文使用图像重合度（Intersection over Union，*IOU*）来区分样本。*IOU* 为 2 个框交集面积与并集面积的比值，*IOU* 小则重合程度低，*IOU* 大则重合程度高。Base-MTCNN 设置 $IOU > 0.65$ 为正样本、$0.45 < IOU < 0.65$ 为部分样本、$IOU < 0.3$ 为负样本。针对 CelebA 标签的特异性，实验验证：当 $IOU > 0.4$ 时为正样本（图 3 中绿色框区域），$0.15 < IOU < 0.4$ 为部分样本（图 3 中黄色框区域），$IOU < 0.15$ 时为负样本（图 3 中蓝色框区域），蓝色框获取的样本数量更准确一些。

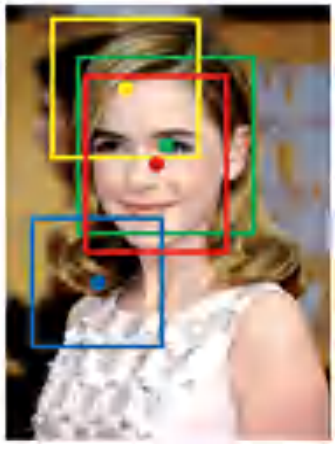

图 3 建议框的选择

Fig. 3 Selection of the suggestion box

经删选处理后，输入 PNet 尺寸为 12×12 的样本 1 793 714张，输入 RNet 尺寸为 24×24 的样本 1 775 208 张，输入 ONet 尺寸 48×48 的样本 1 720 519张，共计5 280 441张样本，样本视例则如图4~图 6 所示。

图 4 正样本

Fig. 4 Positive samples

图 5 部分样本

Fig. 5 Part of the samples

图 6 负样本

Fig. 6 Negative samples

1.4 损失函数

多任务级联网络的损失由以下 2 部分组成：人脸/非人脸判别与生成的备选框回归。其中，人脸/非人脸判别采用 cross-entropy 损失函数，如式（1）：

$$L_i^{det} = -\left(y_i^{det}\log(p_i) + (1 - y_i^{det})(1 - \log(p_i))\right), \quad (1)$$

其中，p_i 为网络预测判断是人脸的概率，$y_i^{det} \in \{0,1\}$。

人脸回归采用欧式距离损失函数，如式(2)、式(3)所示：

$$L_i^{box} = \| \hat{y}_i^{box} - y_i^{box} \|_2^2, \quad (2)$$

其中，$y_i^{box} \in R^4$。

$$\min \sum_{i=1}^{N} \sum_{j \in \{det,box\}} \alpha_j \beta_i^j L_i^j. \quad (3)$$

其中，N 为训练样本的个数；α 表示人物的重要性；β 表示样本的类型。在 PNet、RNet 与 ONet 训练时将分别采用不同的系数。

2　实验与分析

研究中将使用深度卷积神经网络改进的多任务级联网络 MTCNN 作为模型来进行实验，实验将 CelebA 人脸公开数据集作为基准数据，通过更改 *IOU* 阈值参数，共产生5 289 441个正、负部分样本。并在 PASCAL、AFW 和 FDDB 3 个公开数据集上进行测试。模型训练的硬件环境为 Intel Xeon E3-1225 处理器，搭载 Tesla K40c 显卡，在 Centos7.0 系统中搭建 Pytorch 的 Python3.6 环境进行训练；权重初始化为 0；级联网络 PNet、RNet、ONet 分别训练 500、500、140 个 *epochs*。具体实验环境配置见表 1。

表 1　本文环境配置

Tab. 1　Environment configuration in this article

设备	名称
CPU	Intel Xeon E3-1225
GPU	GeForce Tesla K40c
实验环境	CUDA9.0+Pytorch1.1.0

采用随机抽样方法，分别从 AFW、FDDB、PASCAL 公开人脸数据库中抽取数据集做对比实验，用来分析本方法在人脸检测项目中的表现。

AFW 数据集共包含 205 张图片，其中含有 473 张人脸。研究中得到的级联网络 PNet 改进前后损失函数下降对比结果分别如图 7、图 8 所示。

PNet 共迭代 500 次。可以看出，相比以往算法，本文改进算法损失函数下降得更加平稳，使得人脸检测的鲁棒性进一步加强。图 9 为 AFW 数据集部分测试结果展示。

PASCAL 数据集包含 851 张图片，共有 1 335 张人脸。图 10 为级联网络 RNet 改进前后损失函数下降对比结果。

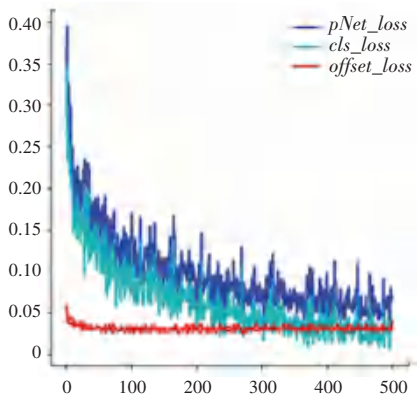

图 7　loss 函数

Fig. 7　loss function

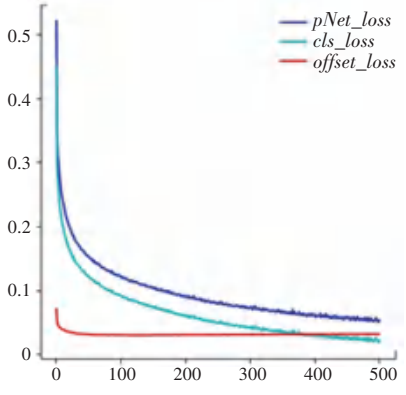

图 8　Proposed_loss 函数

Fig. 8　Proposed_loss function

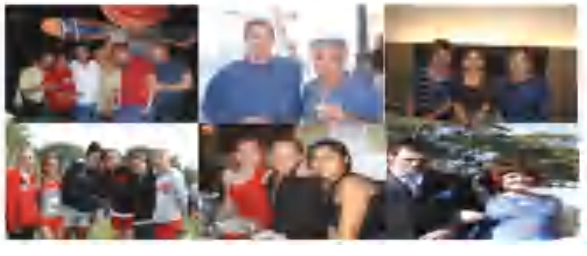

图 9　AFW 测试集效果图

Fig. 9　AFW tests set renderings

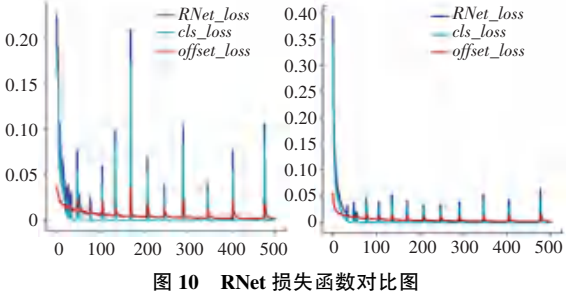

图 10　RNet 损失函数对比图

Fig. 10　Comparison diagram of RNet loss function

图 11 为 PASCAL 数据集的部分测试结果展示。

FDDB 数据集来源于 Yahoo 新闻，包含2 845张图片和5 171张标注人脸，其特点是低像素人脸较多，环境较前两种数据集更复杂。由此得到的级联

网络 ONet 改进前后损失函数下降对比图结果见图 12。

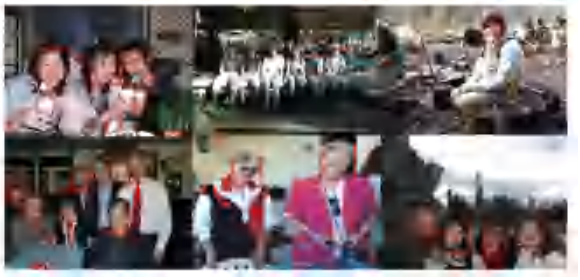

图 11 PASCAL 测试集效果图

Fig. 11 PASCAL test sets renderings

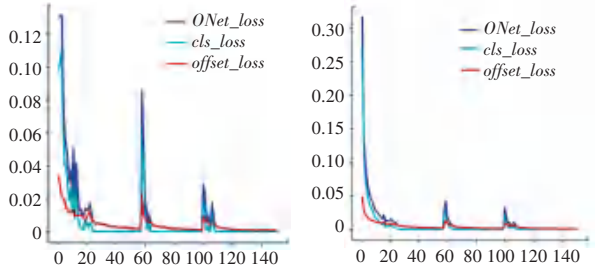

图 12 ONet 损失函数对比图

Fig. 12 Comparison diagram of ONet loss function

ONet 共迭代 140 次。可以看出,本文损失函数下降得更加平稳,使得人脸检测的鲁棒性进一步加强。图 13 为 FDDB 数据集部分测试结果展示。

图 13 FDDB 测试集效果图

Fig. 13 FDD Btest sets renderings

3 结束语

本文通过调整 *IOU* 阈值参数,对已获取的数据集样本进行准确性增强,并对级联网络的损失函数取均值,使得整个损失函数下降得更加平稳,收敛得更加迅速,增加了人脸检测的鲁棒性。其次,本文模型的参数仅有 2.1 M,具有计算量小、容易实现、实时性强等特点,可以应用到配置不高、但对时延敏感的设备上。

参考文献

[1] 梁爽. 基于人脸检测识别技术的网上追逃系统设计与实现[D]. 上海:上海交通大学,2016.

[2] LI Xiaobai, KOMULAINEN J, ZHAO Guoying, et al. Generalized face anti-spoofing by detecting pulse from face videos [C]// 2016 23rd International Conference on Pattern Recognition (ICPR). Cancun, Mexico: IEEE Press, 2016: 4244-4249.

[3] VIOLA P A, JONES M J. 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(CVPR 2001). Kauai, HI, USA:IEEE, 2001:511-518.

[4] PHAM M T, GAO Yang, HOANG V D D, et al. Fast polygonal integration and its application in extending haar-like features to improve object detection [C]// 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA:IEEE, 2010:942-949.

[5] MATHIAS M, BENENSON R, PEDERSOLI M, et al. Face detection without Bells and Whistles[M]//FLEET D, PAJDLA T, SCHIELE B, et al. Computer Vision-ECCV 2014. ECCV 2014. Lecture Notes in Computer Science. Cham:Springer, 2014, 8692:720-735.

[6] KRIZHEVSKYA, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [C]// Advances in Neural Information Processing Systems (NIPS). Nevada, USA:NIPS, 2012:1097-1105.

[7] SUN Y, WANG X, TANG X. Deep learning face representation by joint identification-verification[J]. CoRR, 2014: abs/1406.4773.

[8] YANG S, LUO P, 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. Washington DC, USA:IEEE, 2015: 3676-3684.

[9] LI H, LIN Z, SHEN X, et al. A convolutional neural network cascade for face detection [C]// Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition. Boston, USA:IEEE, 2015:5325-5334.

[10] ZHANG Jie, SHAN Shiguang, KAN Meina, et al. Coarse-to-Fine Auto-encoder Networks (CFAN) for real-time face alignment[C]// Computer vision-ECCV 2014, part 2: 13th European conference on computer vision (ECCV 2014). Zurich, Switzerland:dblp, 2014:1-16.

[11] YU Xiang, HUANG Junzhou, ZHANG Shaoting, et al. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model[C]//2013 IEEE International Conference on Computer Vision. Sydney, NSW, Australia: IEEE, 2013: 1944-1951.

[12] ZHANG Z, LUO P, LOY C C, et al. Facial landmark detection by beep multi-task learning [J]. European Conference on Computer Vision. Cham:Springer, 2014: 94-108.

[13] CHEN Dong, REN Shaoqing, WEI Yichen et al. Joint cascade face detection and alignment [M]//FLEET D, PAJDLA T, SCHIELE B, et al. Computer Vision-ECCV 2014. Lecture Notes in Computer Science. Cham:Springer, 2014, 8694:109-122.

[14] ZHANG Cha, ZHANG Zhenyou. Improving multiview face detection with multi-task deep convolutional neural networks [C]// 2014 IEEE Winter Conference on Applications of Computer Vision (WACV). Steamboat Springs, CO, USA:IEEE, 2014: 1036-1041.