基于卷积神经网络的教室人脸检测算法

王萌, 苏寒松, 刘高华*, 李燊
天津大学电气自动化与信息工程学院, 天津 300072

摘要 针对教室场景下后排学生人脸微小难以检测的情况,提出一种基于卷积神经网络的教室人脸检测算法。采用两阶段检测形式,运用残差神经网络的结构对教室人脸进行特征提取,同时构建特征金字塔,并将 Softmax 损失函数与中心特征损失函数结合,运用合适的激活函数进行训练。此算法在教室场景下获得 95.2%的准确率,且在通用数据集 Wider Face 的三个等级验证集上分别获得 93.0%,87.3%,58.3%的平均精度均值。

关键词 机器视觉; 人脸检测; 教室考勤; 卷积神经网络; 深度学习
中图分类号 TP391 **文献标识码** A **doi**: 10.3788/LOP56.211501

Classroom Face Detection Algorithm Based on Convolutional Neural Network

Wang Meng, Su Hansong, Liu Gaohua*, Li Shen
College of Electrical Automation and Information Engineering, Tianjin University, Tianjin, 300072, China

Abstract This study proposes a face detection algorithm based on a convolutional neural network considering the scenario of a classroom, where the faces of students sitting in the back rows might not be visible. First, the algorithm extracts face features in two stages using a residual neural network. Then, it builds a feature pyramid and combines the Softmax loss function with center loss function to train a face recognition model based on a proper activation function. Upon applying the algorithm to the Wider Face dataset, it achieves an accuracy of 95.2% and mean average precision values of 93.0%, 87.3%, and 58.3% for three levels of validation sets, respectively.

Key words machine vision; face detection; classroom attendance; convolutional neural networks; deep learning
OCIS codes 150.1135; 100.2000; 100.3008

1 引 言

在教室场景下的人脸检测是指在课堂图片中检测出人脸所在位置的检测过程,是机器视觉领域至关重要的内容。人工清点课堂人数费时费力,在此条件下运用计算机视觉进行课堂人脸检测的技术应运而生。早期的传统人脸检测算法对人脸特征提取并不完全,检测准确率很低。随着卷积神经网络的不断发展,人脸检测技术在精度上大幅度超越传统检测方法。但对于教室这一特殊场景,依然存在着小的人脸难以被检测或易被误检等问题。

本文提出一种针对教室场景的人脸检测算法,主要工作为:采用深度残差神经网络[1]对待测图像进行特征提取,并通过构建特征金字塔结构[2],对由深度残差神经网络提取到的不同层的分块特征进行融合,有利于检测小的人脸;之后将特征送入区域候选网络,得到人脸的区域提议和区域得分,并对其进行非极大值抑制[3]操作,将较高得分的区域提议进行感兴趣区域池化的操作;最后经过全连接层,得到检测到的人脸的置信度及坐标。此外,本文将采用的 Softmax 损失函数与中心特征损失函数[4]结合,增加了类间距离,减小了类内距离,使检测的结果更利于分类;另外,采用 ELU 函数[5]作为激活函数,避免训练过程中负半轴信息丢失的问题;最后,对选择通用人脸数据和教室场景人脸数据结合,进行网络训练,使训练结果更适合于教室场景,且不至于对

收稿日期: 2019-03-28; **修回日期**: 2019-04-10; **录用日期**: 2019-04-26
基金项目: 广州市科技计划项目(201802020008)
* **E-mail**: suppig@126.com

教室场景过拟合。

通过实验验证，本文算法不仅在教室场景下对于小的人脸检测任务获得了优秀的效果，而且在人脸检测通用数据集 Wider Face[6] 上也取得很好的效果。

2 基本原理

2.1 两阶段人脸检测

人脸检测是目标检测领域中的一个重要分支。近年来，基于卷积神经网络的目标检测算法可分为两种形式，单阶段检测形式和两阶段检测形式。这两种形式的区别在于：单阶段的人脸检测形式中不包含对区域进行提议这一阶段，而是直接产生检测到人脸的坐标信息及置信度。通常而言，此种单阶段人脸检测形式的精度远不如两阶段的人脸检测形式[7]，尤其不利于对后排小的人脸的检测。对教室场景而言，尤其是小的人脸检测这一问题，将运用两阶段人脸检测形式的检测算法。

2.2 特征提取网络

近年来，用于特征提取的卷积神经网络结构得到了充分发展，从最初 LeCun 等[8]、Krizhevsky 等[9]简单的网络，不断向更深层次网络发展。但随着网络的逐层加深，出现了准确率下降的问题，并且更深层次的网络在训练时常常伴随着梯度消失或梯度爆炸的问题，极易出现信息丢失，从而导致训练结果很差。而“残差模块结构”[1]能避免上述问题，这种结构的特点是，存在一个可直接将输入的图像信息传到输出端的支路，这样不仅可以确保信息的完整性，避免信息丢失，而且在网络进行学习的过程中只需学习输入和输出间差别的部分。残差模块的结构如图 1 所示。其中，x 为卷积层的输入，$f(x)$ 为经过激活函数的输出。

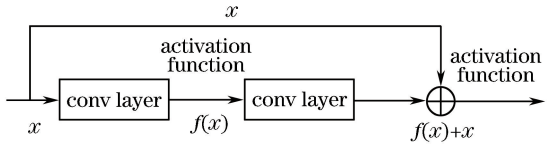

图 1 残差模块结构

Fig. 1 Structure of residual module

特征提取网络采用 50 层卷积神经网络，以残差模块的连接方式相连。这样使训练过程变得简易，并且可以更充分地提取人脸特征。具体而言，按照具体参数不同总共分为 5 个模块，其参数如表 1 所示。表中卷积层的参数分别是卷积核的 kernel size、通道数及 stride。用 Conv3_x 的参数举例，表示经过 4 组相同的操作，其中每一组包含 3 个卷积核，其参数分别是 kernel size 为 1×1，通道数为 128，kernel size 为 3×3，通道数为 128，kernel size 为 1×1，通道数为 512。

表 1 特征提取网络的参数

Table 1 Parameters of feature extraction network

Block	Conv1	Conv2_x	Conv3_x	Conv4_x	Conv5_x
Parameter	7×7,64,stride 2	3×3 max pool, stride 2 $\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$

2.3 构建特征金字塔

通常，两阶段的目标检测算法运用特征提取网络最后一层提取出的特征进行特征预测。在特征提取的过程中，特征图尺寸越来越小。而对于教室人脸检测任务而言，最重要的是对后排小的人脸的检测。小的人脸在经过前向传播的特征逐层提取后的信息所剩无几，且其信息大多分布在底层，因此若直接使用网络提取的最后一层特征进行预测，会丢失很多重要的信息。

特征金字塔[2]是一种可以移植到特征提取网络中的结构。其特征融合与前向传播的方向相反，将较深层次的特征图进行上采样操作，然后将上层与底层的特征横向逐层相连。与惯常的只用一层特征进行特征预测所不同的是，采用特征金字塔结构特征融合的每一层特征均单独进行特征预测。

深层次的特征图拥有更加抽象的特征，低层次的特征图拥有更加具体的特征，这样的相连方式可以将不同强度的特征相融合，不至于将底层的重要信息丢失。整体网络构成如图 2 所示，其中 FC 表示全连接层，RoI Pooling 表示感兴趣区域池化操作，P2-P6 表示特征金字塔的构成部分。

2.4 分类损失函数

人脸检测的最终目标是得到人脸、非人脸分类结果，及人脸所在位置的坐标。区分人脸与非人脸部分这一分类任务依赖于分类损失函数。常用的分类损失函数是 Softmax 损失函数[10]，其表达式为

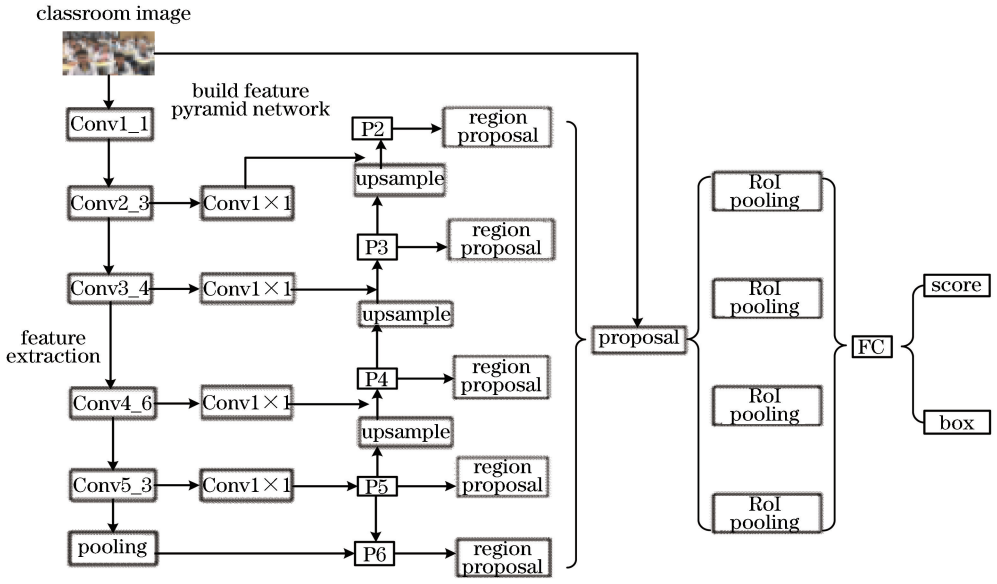

图 2 教室人脸检测算法的整体网络结构

Fig. 2 Overall network structure of classroom face detection algorithm

$$L_s=-\sum_{i=1}^{m}\mathrm{lb}\left(\frac{\exp(\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{b}_{y_i})}{\sum_{j=1}^{n}\exp(\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{b}_j)}\right),\tag{1}$$

式中：$\boldsymbol{x}_i$ 为网络所提取的特征组成的矩阵；$\boldsymbol{W}_j$ 为学习到的权重组成的矩阵；$\boldsymbol{b}_j$ 为偏置组成的矩阵；i 为 1 到 m 所有整数取值，表示某一批次(batch)，m 为总共的 mini-batch 数；j 为 1 到 n 所有整数取值，表示某一分类，n 为总体分类数；y_i 为某一种类别，$\boldsymbol{b}_{y_i}$ 为属于 y_i 这一类的偏置组成的矩阵。Softmax 损失函数表示实际输出概率与期望输出概率的距离，Softmax 损失函数的值越小，两个概率分布就越接近。此损失函数仅描述不同类别之间的类间距，即仅区分了目标为人脸和非人脸部分，没有考虑到将人脸这一种类的类内距离缩小。

为使模型学到的特征判别度更高，不仅要增大类间距，也需减小类内距离。因此引入中心特征损失函数[10]，其表达式为

$$L_c=\frac{1}{2}\sum_{i=1}^{m}\|\boldsymbol{x}_i-\boldsymbol{c}_{y_i}\|_2^2,\tag{2}$$

式中：$\boldsymbol{c}_{y_i}$ 为第 y_i 个类别的特征中心组成的矩阵。则整体的损失函数可以理解为所提取到的特征和特征中心之间的距离。因此当令 L_c 取全局最小值时，每个样本的特征与特征中心之间的距离取得最小值，也就是说，此时的类内间距最小。使 L_c 逐渐变小的过程，也正是每一类特征逐渐聚合的过程。图 3 中使用中心特征损失函之后两种特征类间距离加大，类内距离缩小，从而更利于分类。

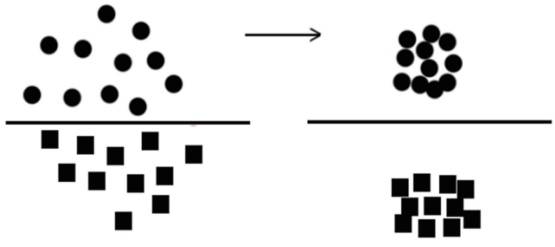

图 3 利用中心特征损失函数使类内距离缩小

Fig. 3 Reducing inner distance based on center feature loss function

将中心特征损失函数与 Softmax 损失函数一起使用，在迭代过程中，不仅能加大类间距，更能减小类内距离，因此可以获得更高的特征判别度，从而获得更好的分类效果。函数表达式为

$$L=L_s+L_c=-\sum_{i=1}^{m}\mathrm{lb}\left[\frac{\exp(\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{b}_{y_i})}{\sum_{j=1}^{n}\exp(\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{b}_j)}\right]+\frac{1}{2}\sum_{i=1}^{m}\|\boldsymbol{x}_i-\boldsymbol{c}_{y_i}\|_2^2。\tag{3}$$

2.5 激活函数

在神经网络进行特征提取的过程中，激活函数的运用起着至关重要的作用。如果不使用激活函数，则整个网络的输入与输出完全呈线性关系，即使增加网

络的层数，网络也不具备很好的拟合能力。

通常运用 ReLu[11] 函数来提高网络的非线性能力。ReLu 函数的正半轴函数值与输入相同，负半轴函数值取 0。ReLu 函数作为激活函数存在着许多问题，在反向传播的过程中需要进行求导操作，此时若负半轴信号的梯度为 0，则会导致神经元不能更新参数，也就是神经元不再学习。为解决 ReLu 函数这个问题，在整体网络中引入 ELU[5] 函数。此函数在输入为负数的情况下，有一定的输出，并且具有一定的抗干扰能力，这样既修正了数据分布，又保留了一些负轴的值，使得负轴信息不完全丢失。ELU 激活函数的数学表达式为

$$f(x)=\begin{cases}x, & x>0\\ \alpha[\exp(x)-1], & x\leqslant 0\end{cases}, \tag{4}$$

式中：x 为激活函数的输入；$f(x)$ 为 ELU 激活函数；α 为常数，取 0.25。

3 实　　验

3.1 训练数据库

在神经网络的训练过程中，训练数据库对训练结果起着至关重要的作用，尤其是对于教室这一特殊场景，选择合适的数据集才能更好地完成教室人脸检测这一特殊任务。

Wider Face[6] 数据库于 2016 年推出，是目前最有难度的人脸检测数据库。数据集中涵盖了丰富场景下不同大小、姿态、角度的人脸，也包含着很多密集的小的人脸。

但对于教室场景，Wider Face 所提供的数据过于复杂，有很多混乱场景的人脸图片没有必要对其进行训练。因此，选择 Wider Face 数据库中较为简单的场景图，并且对于教室场景，引入教室场景下的人脸图片，筛选出不同教室背景、不同姿态、不同光照下的图片，并对图中人脸进行标注，对整体卷积神经网络进行训练。这种结合教室场景的训练方法对检测教室人脸更加有利，同时运用 Wider Face 中的图片，使整个网络的训练过程中不至于对教室场景产生过拟合。

3.2 实验结果

在 Nvidia GTX 1080Ti 显卡的 GPU 下，利用 Tensorflow[12] 这一深度学习框架，在 Ubuntu16.04 系统下，采用 Python 语言进行编程，训练完的模型在教室场景下取得了很好的检测效果。定义准确率为教室场景下检测正确的人脸占图片中全部人脸的比例。整体而言，在 2000 张不同视角、不同分辨率、不同光照、多种背景的教室场景图片中，达到了 95.2%的准确率。检测效果如图 4 所示。

图 4　教室场景检测实例。(a)不同背景；(b)不同视角；(c)不同光照

Fig. 4　Examples of classroom face detection. (a) Different backgrounds; (b) different perspectives; (c) different lighting

本文方法在训练时仅使用 Wider Face 训练集中的部分图片与自制的教室人脸数据图片。本文算法不仅在教室场景下获得了良好的检测结果，还具有很好的泛化能力，在 Wider Face 验证集上超越了很多现有的人脸检测算法，例如 LDCF+[13]，faceness-WIDER[14]，multi-task Cascade CNN[15]，two-stage CNN[6] 及 ACF-WIDER[16]。Wider Face 验证集按照检测难度分为简单、中等、困难 3 个类别。检测方法在 3 个类别的检测结果的 Precision-Recall 曲线如图 5 所示。

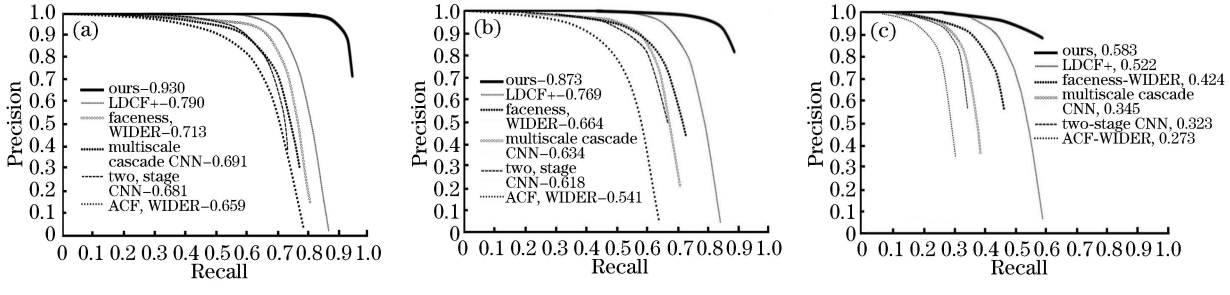

图 5　在 Wider Face 验证集上的结果。(a)简单；(b)中等；(c)困难

Fig. 5　Results on verification set of Wider Face. (a) Easy; (b) medium; (c) hard

曲线横坐标为召回率(Recall),纵坐标为精确率(Precision),图中图例数字为平均精度均值(mAP),本算法在 WiderFace 数据集的三个等级验证集上分别获得 0.930,0.873,0.583 的平均精度均值。如果一种算法的 Precision-Recall 曲线被另一种算法的 Precision-Recall 曲线包住,则后者的性能优于前者,从图中的平均精度均值也可看出,本文算法优于其他算法。

4 结 论

针对教室这一特殊场景,提出一种基于卷积神经网络的教室人脸检测算法。采用残差神经网络的结构对教室人脸进行特征提取,并通过构建特征金字塔,检测后排的小的人脸;同时,将 Softmax 损失函数与中心特征损失函数结合,使检测结果更利于分类,并且用合适的激活函数增强了整体网络的非线性。本文方法在教室场景下获得了优秀的检测效果,同时在 Wider Face 数据集上也获得了很好的效果。

参 考 文 献

[1] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition [C] // The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 26-July 1, 2016, Las Vegas, Nevada. New York: IEEE, 2016: 770-778.

[2] Seferbekov S, Iglovikov V, Buslaev A, *et al*. Feature pyramid network for multi-class land segmentation [C] // The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 18-22, 2018, Salt Lake City, Utah. New York: IEEE, 2018: 272-275.

[3] Neubeck A, van Gool L. Efficient non-maximum suppression [C] // 18th International Conference on Pattern Recognition (ICPR'06), August 20-24, 2006, Hong Kong, China. New York: IEEE, 2006: 9210072.

[4] Qi C, Su F. Contrastive-center loss for deep neural networks [C] // 2017 IEEE International Conference on Image Processing (ICIP), September 17-20, 2017, Beijing, China. New York: IEEE, 2017: 2851-2855.

[5] Clevert D A, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (ELUs) [J/OL]. (2016-02-22) [2019-03-02]. https://arxiv.org/abs/1511.07289.

[6] Yang S, Luo P, Loy C C, *et al*. WIDER FACE: a face detection benchmark [C] // The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 26-July 1, 2016, Las Vegas, Nevada. New York: IEEE, 2016: 5525-5533.

[7] Ou P, Zhang Z, Lu K, *et al*. Object detection in of remote sensing images based on convolutional neural networks [J]. Laser & Optoelectronics Progress, 2019, 56(5): 051002.
欧攀, 张正, 路奎, 等. 基于卷积神经网络的遥感图像目标检测 [J]. 激光与光电子学进展, 2019, 56(5): 051002.

[8] 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.

[9] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C] // Advances in Neural Information Processing Systems 25 (NIPS 2012), December 3-8, 2012, Harrahs and Harveys, Lake Tahoe. New York: NIPS, 2012: 1097-1105.

[10] Long X, Su H S, Liu G H, *et al*. A face recognition algorithm based on angular distance loss function and convolutional neural network [J]. Laser & Optoelectronics Progress, 2018, 55(12): 121505.
龙鑫, 苏寒松, 刘高华, 等. 一种基于角度距离损失函数和卷积神经网络的人脸识别算法 [J]. 激光与光电子学进展, 2018, 55(12): 121505.

[11] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks [C] // Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS), April 11-13, 2011, Fort Lauderdale, USA. [S.l.: s.n.], 2011: 315-323.

[12] Liu F, Liu P Y, Li B, *et al*. Deep learning model design of video target tracking based on TensorFlow platform [J]. Laser & Optoelectronics Progress, 2017, 54(9): 091501.
刘帆, 刘鹏远, 李兵, 等. TensorFlow 平台下的视频目标跟踪深度学习模型设计 [J]. 激光与光电子学进展, 2017, 54(9): 091501.

[13] Ohn-Bar E, Trivedi M M. To boost or not to boost? On the limits of boosted trees for object detection [C] // 2016 23rd International Conference on Pattern Recognition (ICPR), December 4-8, 2016, Cancun, Mexico. New York: IEEE, 2016: 3350-3355.

[14] Yang S, Luo P, Loy C C, *et al*. From facial parts responses to face detection: a deep learning

approach[C]// 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 3676-3684.

[15] Zhang K P, Zhang Z P, Li Z F, *et al*. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[16] Yang B, Yan J J, Lei Z, *et al*. Aggregate channel features for multi-view face detection[C]// IEEE International Joint Conference on Biometrics, September 29-October 2, 2014, Clearwater, FL, USA. New York: IEEE, 2014: 14838106.