

人脸检测算法的优化

龚格格¹, 吴 珊¹, 郭湘南²

(1. 武汉邮电科学研究院, 湖北 武汉 430000;

2. 武汉烽视威科技有限公司, 湖北 武汉 430000)

摘 要:面部特征被广泛应用于一系列视频监控系统,其中公安系统中人脸检测模块尤为突出。由于人脸的巨大视觉变化,如遮挡、光照、大的姿态变化问题使人脸检测一直存在着瓶颈,在实际应用中这些问题依旧很常见。对此,文中通过简要介绍候选框生成算法,同时结合 Faster RCNN、联合人脸检测和对齐的级联卷积神经网络框架的优缺点进行分析和改进,提出了快速级联卷积神经网络模型。由于候选框网络和 RoI 检测网络共享卷积层,在候选框网络中使用多层卷积层信息,采用 RoI 池化和 L2 归一化将身体信息与面部信息进行融合,实现结合身体上下文信息来处理较小的人脸区域,并对数据集进行测试来验证模型的有效性,弥补因视觉变化导致人脸检测中的不足,提高人脸检测网络性能。

关键词:人脸检测;候选框生成算法;Faster RCNN;快速级联卷积神经网络模型;网络性能

中图分类号: TP301.6

文献标识码: A

文章编号: 1673-629X(2019)06-0047-05

doi:10.3969/j.issn.1673-629X.2019.06.010

Optimization of Face Detection Algorithm

GONG Ge-ge¹, WU Shan¹, GUO Xiang-nan²

(1. Wuhan Institute of Posts and Telecommunications Science, Wuhan 430000, China;

2. Wuhan Feng Visual Technology Co., Ltd., Wuhan 430000, China)

Abstract: Facial features are widely used in a series of video monitoring systems, among which the face detection module in the public security system is particularly prominent. Due to the huge visual changes of face, such as occlusion, illumination and large posture changes, human detection has always been a bottleneck, and these problems are still very common in practical applications. For this, through the brief introduction of the candidate generation algorithm, at the same time, combined with Faster RCNN, analysis and improvement of the advantages and disadvantages of face detection and alignment cascaded convolutional neural network framework, we propose a fast cascaded convolution neural network model. Since the candidate box and RoI detection networks share the convolution layer, multi-layer convolutional layer information is used in the candidate box network. The RoI pooling and L2 normalized body and facial information are used to fuse, dealing with the smaller face region with the physical context information, and testing data sets to verify the validity of the model, which makes up for the deficiency of face detection caused by visual changes and improves the performance of face detection network.

Key words: face detection; candidate frame generation algorithm; Faster RCNN; fast cascade convolution neural network model; network performance

0 引言

人脸检测是人脸识别的基础,然而,人脸的巨大视觉变化在现实应用中给这些任务带来了巨大的挑战。十多年前,Viola 和 Jones^[1]提出的 cascade face detector 采用 Haar-Like feature 和 AdaBoost 对 cascade classifier 进行训练,具有良好的实时效率,然而该人脸检测算法由于采用人工设计,对光照变化的鲁棒性并不是很好,

对头部姿态比较大的人脸检出率非常低。相当多的工作表明^[2-4],该算法在应对大量视觉影响的人脸图片检测时性能急剧下降。基于此,文中提出了一种快速级联卷积神经网络模型,以提高算法性能。

1 候选框生成算法

候选框是人脸检测网络的基础。检测任务与分类

任务并不相同,检测任务需首先对单个样本进行分类并生成候选框,其后对所生成候选框的区域进行分类,而分类任务是直接对单个样本进行分类。

2 Faster RCNN 人脸检测网络框架

考虑到区域候选框是检测网络最大的瓶颈,该系统中 Faster RCNN 使用 RPN(region proposal network)来有效解决计算候选框耗时的问题。

Faster RCNN^[5] 采用 RPN 生成候选框,RPN 的结构属于深度卷积神经网络^[6-8]。通过将候选框送入 Faster RCNN^[9] 中进行分类,产生检测结果。Faster RCNN 的网络结构如图 1 所示,整个 Faster RCNN 网络由两部分组成,分别为 RPN 和 Faster RCNN 系统。其中 RPN 称为全卷积网络^[7],将单张图像作为输入数据。卷积层 1 至 ReLU5 是 ZF (Zeilerhe 和 Fergus 模型)通用结构^[8]中包含的部分。ZF 包含了多个不同层,其中激活函数采用了 ReLU,以此来提高网络检测的性能。归一化层采用 LRN 策略(local response normalization,局部响应归一化层),对数据进行归一化操作。由图可知网络层为 RPN 的核心层。网络层包含了推荐层 1、候选框回归层、1×1 卷积层、候选框概率层等。

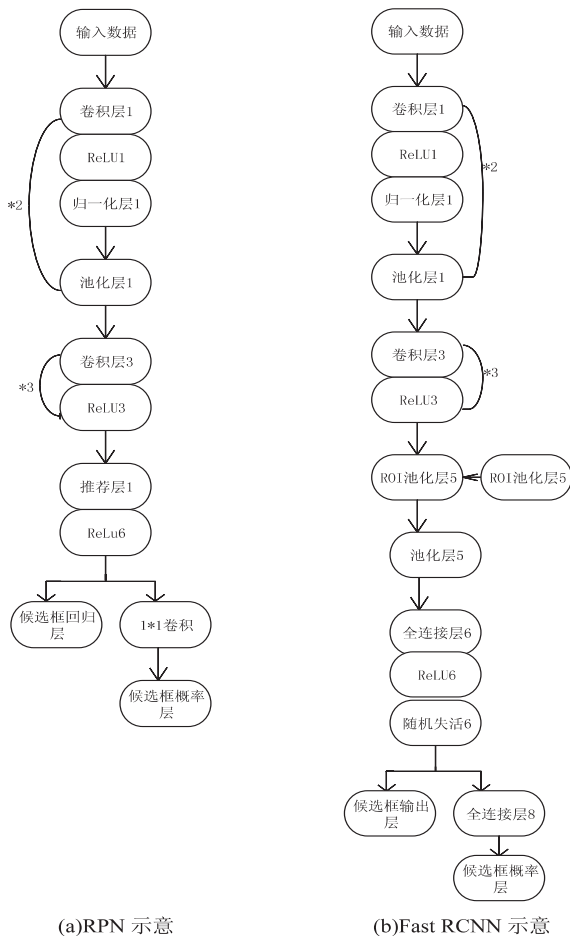


图 1 Faster RCNN 网络框架

图 1(b)是 Faster RCNN 的详细网络结构,利用单张图像作为输入数据,其中候选框位置与输入的候选框的坐标相同。卷积 1 至 ReLU5 同 RPN 一样,池化层 5 到随机失活 6 和 softmax 损失函数即为 ZF 结构的完整模型。

2.1 Faster RCNN 训练与测试

通过对每个锚进行二分类(是目标或者不是)来训练 RPN。若锚与真实框有最高的交并比或者锚与任意真实框的交并比大于 0.7,则作为正样本,将交并比小于 0.3 的锚作为负样本,其余的均直接丢弃。

在 Faster RCNN 中,将一个锚 i 的损失函数定义为:

$$L(p_i, t_i) = L_{\text{cls}}(p_i, p_i^*) + r p_i * L_{\text{reg}}(t_i, t_i^*)$$

其中, p_i 为锚 i 的预测概率。如果锚是正样本,则 p_i^* 为 1,负样本时 p_i^* 为 0。 $t_i = \{t_x, t_y, t_w, t_h\}_i$ 表示预测候选框坐标的 4 个参数。 $t_i^* = \{t_x^*, t_y^*, t_w^*, t_h^*\}_i$ 对应锚的正样本的真实值。候选框分类损失 L_{cls} 是两个类别 Softmax 损失。候选框概率层和候选框回输出分别包含 $\{p_i, t_i\}$ 。

$$\text{Smooth}_{\text{L1}}(x) = 0.5x^2 (|x| < 1)$$

$$\text{Smooth}_{\text{L1}}(x) = |x| - 0.5 (x < -1 \text{ 或 } x > 1)$$

选取的四个坐标参数为:

$$t_x = (x - x_a)/w_a, t_y = (y - y_a)/h_a, t_w = \log(w/w_a), t_h = \log(h/h_a)$$

$$t_x^* = (x^* - x_a)/w_a, t_y^* = (y^* - y_a)/h_a, t_w^* = \log(w^*/w_a), t_h^* = \log(h^*/h_a)$$

其中, x, x_a, x^* 分别代表候选框、锚、真实框。此方法实现 bounding box 回归。

RPN 由一个全卷积网络实现,使用 SGD (stochastic gradient descent) 算法^[9]进行端到端的训练。训练时采用与图像中心采样策略^[6]一致的方式来训练网络。实际中通过对每张图片随机采样 512 个锚使得正负锚的比例为 1:1,最后对采样后的锚计算小批的损失函数。

所有层权重的初始化均采用零均值方差为 0.02 的高斯分布。将所有被初始化后的层在 CelebA 人脸检测^[10]上进行初始化预训练,与练习实验^[11]中的操作相同。采用 5 000 个 CelebA 的图像用 0.001 的学习率进行训练。采用 0.7 的动量和 0.003 的权重,整个网络通过 Caffe^[12]实现。

在训练 RPN 网络时由于对候选框欠缺考虑,所以联合优化将成为一个重点。该系统中采用四部训练算法来完成 RPN 和检测网络共享卷积层的优化算法。第一步,训练 RPN。RPN 用 CelebA 预训练模型进行初始化并且端到端地微调候选框任务;第二步,用已经

生成的候选框训练 Faster RCNN 并检测网络;第三步,用检测初始化 RPN,固定共享的卷积层,指微调 RPN 中的其他层;最后,固定共享的卷积层,并微调全连接层。

由于生成的候选框有很大部分重叠,该系统采用 NMS (non-maximum suppression)^[13] 算法来减小冗余。在 NMS 中固定交并比阈值为 0.7,这样将会筛选出大概 3 500 个候选框。NMS 在大幅度减少候选框数量的

情况下对最后的检测精度并无影响。NMS 之后,将前 N 个候选框输入到检测网络。

2.2 联合人脸检测和对齐的网络框架

该系统通过对联合人脸检测和对齐的网络进一步研究,找出人脸检测性能较高的原因,并将其与 Faster RCNN 进行相关对比,从不同框架结构出发,结合自身的优势提出有利于人脸检测的优化模型。

联合人脸检测和对齐的网络框架如图 2 所示。

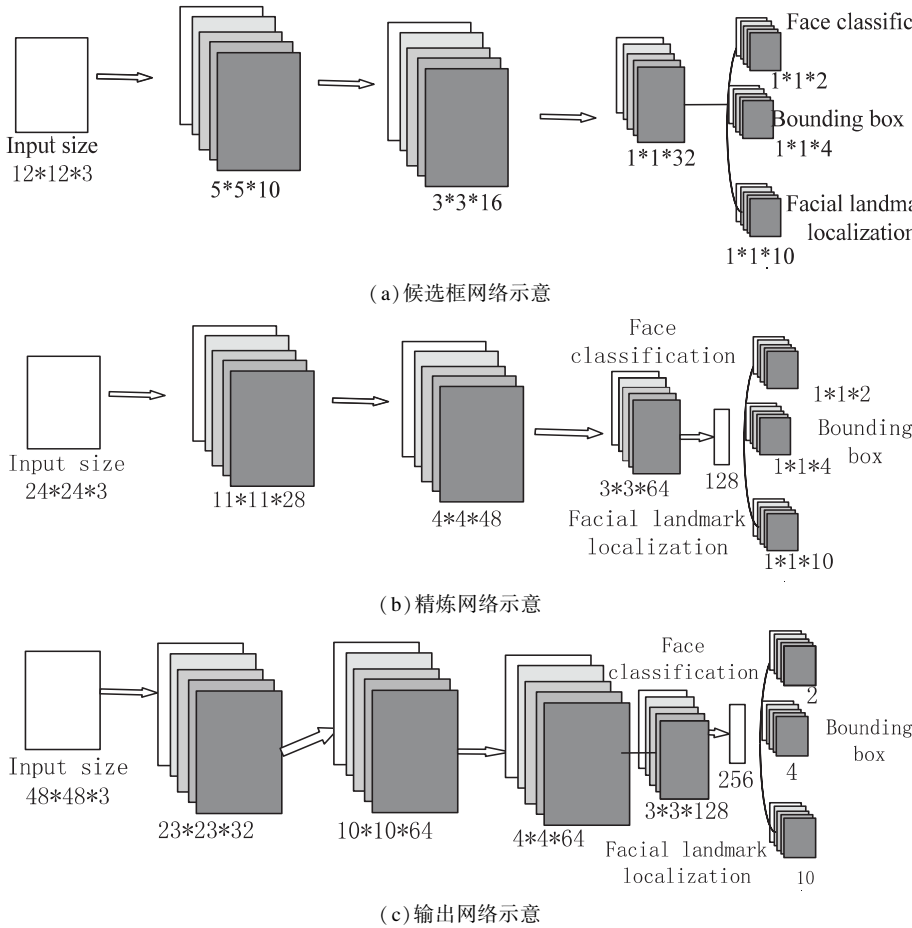


图 2 联合人脸检测和对齐的网络框架

由网络框架图可看出,此网络包括三个阶段。第一阶段为候选框网络,即快速产生候选框窗口。第二阶段为精炼网络,即优化产生的候选框网络。第三阶段为输出网络,优化输出结果,并将面部关键点位置输出。与多任务的学习框架相比,此算法的性能得到了显著性的提高。

给定一个图像,依据图像的不同尺度,来构建一个图像金字塔,其中三个级联框架的输入分为如下三个阶段:

第一阶段:利用全卷积网络,即 P-Net 来获取候选人脸窗口。然后利用估计的边界框回归向量对候选对象进行校准。在此之后,使用非最大抑制 (NMS) 将高度重叠的候选项进行合并。

Net),通过精准排除大量的错误候选对象,利用边界框回归进行校准,并进行 NMS 去除重叠窗体。

第三阶段:这一阶段类似于第二阶段,只是在去除重叠候选窗口的同时,显示五个人脸关键点定位。

2.3 人脸检测和对齐的联合训练

由图 2 可知,联合人脸检测和对齐的网络主要包含 3 个部分:人脸/非人脸分类器、边界框回归、人脸关键点定位。

人脸分类:学习目标是两类分类问题。对于每个样本,使用交叉熵损失函数。 p_i 是人脸的概率,表示样本是一张脸,符号 $y_i^{\text{det}} \in \{0,1\}$ 表示背景的真实标签。

$L_i^{\text{det}} = - (y_i^{\text{det}} \log p_i + (1 - y_i^{\text{det}}) (1 - \log p_i))$, $y_i^{\text{det}} \in \{0,1\}$

边界框回归:对于每个候选窗口,预测它与最近的

ground truth 之间的偏移量(即,边框的左、上、高、宽)。学习目标是一个回归问题,采用欧氏距离损失函数为每个样本。 y_i^{box} 为通过网络预测得到, y_i^{box} 为实际的真实的背景坐标,其中 y 为一个(左上角 x , 左上角 y , 长,宽)组成的四元组,因此 $y_i^{box} \in R^4$ 。

$$L_i^{box} = \|y_i^{box} - y_i^{box}\|_2^2$$

人脸关键点定位:与边界盒回归任务相似,计算网络预测的地标位置和实际真实地标的欧氏距离,并最小化该距离。 $y_i^{landmark}$ 通过网络预测得到, $y_i^{landmark}$ 是第 i 个样本的地面真实坐标,由于一共 5 个点,每个点 2 个坐标,因此 $y_i^{landmark} \in R^{10}$ 。

$$L_i^{landmark} = \|y_i^{landmark} - y_i^{landmark}\|_2^2$$

2.4 实验结果与对比分析

在检测和分类问题上有很多评价指标。其中召回率和准确率则是检测、分类、识别问题的两个重要评价指标。召回率:在所有样本中检测出相关样本的概率;准确率:在系统所检出的样本中,真正相关的概率。一个好的检测系统希望准确率和召回率都尽可能高,一般情况下准确率和召回率呈负相关。为了有效地评价系统的检测能力,这里引入 AP(平均精度)来对系统的检测能力进行衡量。

采用的数据集为 WIDER FACE^[14] 数据集。根据人脸检测的难易程度,WIDER FACE 将数据集分为简单、中等、困难三个级别,以此来检验测试效果,如表 1 所示。

表 1 不同网络模型的人脸检测结果

网络模型	Faster RCNN	联合人脸检测和对齐的级联卷积神经网络
AP(简单)	0.889	0.910
AP(中等)	0.860	0.873
AP(困难)	0.626	0.640
检测时间(ms/im)	71	42

根据在 WIDER FACE 数据集上的结果可知,联合人脸检测和对齐的级联卷积神经网络检测效果好于 Faster RCNN。

3 快速级联卷积神经网络结构

Faster RCNN 在具有挑战性的 WIDER FACE 数据集上测试时,由于图像包含更多较小、遮挡和不完整的对象,导致性能下降很多。基于现阶段存在的问题,文中提出了一种基于 RPN 方法的快速级联卷积神经网络,来解决如遮挡、光照、大的姿态变化等问题。首先,该系统的突出之处在于将多尺度区域候选网络和上下文多尺度神经网络引入到快速级联卷积神经网络,且支持同时观察多尺度特征,对人脸候选区域进行推断。

然后,通过计算置信度得分和边界框回归,人脸检测系统能够在给定的人脸图像中通过对这些生成的候选框的置信分数进行阈值化来决定检测结果。

3.1 快速级联卷积神经网络模型

快速级联卷积神经网络框架如图 3 所示。卷积层 1 至卷积层 5 具有与 RPN 同样的结构。系统通过采用 RoI 池化层来进行下采样,方便连接组合,沿通道进行 L2 归一化来连接组合,然后通过 1x1 的卷积层进行单层特征值的融合,卷积后的特征图上的候选框采用与 RPN 相同的方法。最后将产生的候选框进行 bounding box 回归,同时将人脸的候选框映射至卷积层 3 至卷积层 5,卷积层 3 至 5 用来产生候选框。从整体网络看,由于共享卷积层 1 至 5,实现了只需一次计算就可以更新所有参数的目的,节约了时间成本,提高了效率。此外,该系统采用多个层的特征图进行连接组合分别来实现联合人脸检测和对齐神经网络的三个阶段。

由于在 Faster RCNN 中采用的区域候选框特征均是通过单个深层卷积特征图产生的,因此存在很难检测到小面孔的问题。为解决此问题,文中采用多个卷积层特征的组合,有助于获得微小面部的信息。同时,由于面部检测需要定位面部以及确认面部信息,该系统将具有定位能力的低级特征和具有语义信息的高级特征融合在一起^[15]。

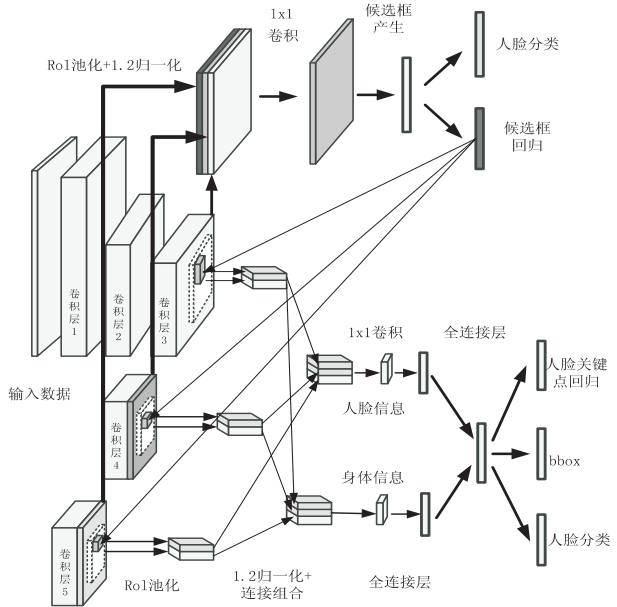


图 3 快速级联卷积神经网络框架

3.2 身体上下文信息的整合

对于人脸检测,由于各种挑战使得脸部被遮挡,在这种情况下,Faceness^[12] 考虑通过对面部部分响应的空间结构和排列进行评分来找到面部,但当面部部分由于遮挡而缺失时,人脸检测将变得难以执行。先前的研究表明^[14],利用上下文可以达到减少整体误差的

目的^[16],同时上下文推理也是物体识别难题的关键部分,基于此在该系统设计中参考身体上下文信息。

3.3 L2 归一化

由于特征来自不同层的特征图,因此可能会有不同的尺度和数值范围,采用 L2 归一化保证来自每个层的特征值大致保持在相同的范围。

4 结束语

为了避免人脸的巨大视觉变化,如遮挡、光照、大的姿态变化问题,基于人脸检测的一些瓶颈问题做了相关研究。基于现有深度卷积神经网络模型,并对其进行进一步的试验、分析、改进,提出了一种快速级联卷积神经网络模型。通过研究 Faster RCNN 和多级联联合人脸检测的卷积神经网络模型分析,抽取出适合于人脸检测的候选框与检测网络模型,采用级联卷积神经网络卷积共享策略,简化神经网络和检测网络的训练和测试,节约了算法所需要的时间^[13]。同时,由于卷积层的共享,实现了级联卷积神经网络端到端的优化,对于较小的人脸区域采用候选框网络和 RoI 检测网络共享卷积层进行处理,激活函数使用 Leaky ReLU,从而提高了神经网络的性能。

参考文献:

- [1] VIOLA P, JONES M J. Robust real-time face detection[J]. International Journal of Computer Vision, 2004, 57(2): 137-154.
- [2] ZHU Xiangxin, RAMANAN D. Face detection, pose estimation, and landmark localization in the wild[C]//IEEE conference on computer vision and pattern recognition. Providence, RI, USA: IEEE, 2012: 2879-2886.
- [3] MATHAIS M, BENENSON R, PEDERSOLI M, et al. Face detection without bells and whistles[C]//European conference on computer vision. Zurich, Switzerland: [s. n.], 2014: 720-735.
- [4] YAN Junjie, LEI Zhen, WEN Longyin, et al. The fastest deformable part model for object detection[C]//IEEE conference on computer vision and pattern recognition. Columbus, OH, USA: IEEE, 2014: 2497-2504.
- [5] 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: MIT Press, 2015: 91-99.
- [6] 施徐敢. 基于深度学习的人脸表情识别[D]. 杭州: 浙江理工大学, 2015.
- [7] 赵志国. 基于深度学习的低分辨率多姿态人脸识别[D]. 大连: 大连理工大学, 2015.
- [8] 池燕岭. 基于深度学习的人脸识别方法的研究[D]. 福州: 福建师范大学, 2015.
- [9] GIRSHICK R. Fast R-CNN[C]//IEEE international conference on computer vision and pattern recognition. [s. l.]: IEEE, 2015: 1440-1448.
- [10] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Computer vision and pattern recognition. Gordon Christie: IEEE, 2015: 3431-3440.
- [11] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//European conference on computer vision. [s. l.]: Springer International Publishing, 2014: 818-833.
- [12] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Computer vision and pattern recognition. Columbus, OH, USA: [s. n.], 2013: 580-587.
- [13] 彭辉, 张长水, 荣钢, 等. 基于 K-L 变换的人脸自动识别方法[J]. 清华大学学报: 自然科学版, 1997(3): 67-70.
- [14] 陈金辉. 静态图像行人检测算法研究[D]. 上海: 华东理工大学, 2015.
- [15] JIA Yangqing, SHELHAMER E, DONAHUE J, et al. Caffe: convolutional architecture for fast feature embedding[C]//ACM international conference on multimedia. Orlando, Florida, USA: ACM, 2014: 675-678.
- [16] YANG Shuo, LUO Ping, LOY C C, et al. WIDER face: a face detection benchmark[C]//IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE, 2016: 5525-5533.