中国科学: 信息科学 2020 年 第 50 卷 第 7 期: 1110–1120
SCIENTIA SINICA Informationis
论文

全天候自然场景下的人脸佩戴口罩识别技术

张修宝[1*], 林子原[1], 田万鑫[1,2], 王艳[1], 沈海峰[1], 叶杰平[1]

1. 北京嘀嘀无限科技发展有限公司 AI Labs, 北京 100094
2. 北京邮电大学信息与通信工程学院, 北京 100876
* 通信作者. E-mail: zhangxiubao@didiglobal.com

收稿日期: 2020–03–06; 修回日期: 2020–03–20; 接受日期: 2020–04–16; 网络出版日期: 2020–06–23

摘要 日常生活中, 面对经呼吸道传播的传染性疾病或厂矿生产过程中产生的扬尘沙土, 人们佩戴口罩进行防护可保护身体健康和生命安全. 人脸佩戴口罩的自动化识别可以有效监督人们佩戴口罩, 是抑制疾病快速传播和保护身体健康的重要技术手段. 对于生活和生产中的口罩佩戴识别的需求, 本文提出了基于深度学习的人脸检测和口罩佩戴识别相结合的方法. 该方法在人脸检测中利用特征融合金字塔, 结合空间和通道注意力学习, 以及分割分支进行神经网络弱监督学习. 另外针对检测后的人脸子图像, 采用图像分类的方法实现快速识别, 并加入注意力学习机制, 增强模型对口罩区域特征的学习. 利用近 20 万的公开和企业自有数据, 并采用数据增强等方法, 在全天候自然场景下取得了 99.50% 的识别准确率. 该技术已广泛应用于滴滴出行实际业务中, 日均处理百万数量级的请求. 该服务已对外开放, 关键算法已开源, 从而使其发挥更大的应用价值和社会价值.

关键词 口罩佩戴识别, 人脸检测, 人脸属性识别, 特征金字塔, 注意力学习

1 引言

为有效预防以呼吸道传播为主要传播方式的疾病大范围流行, 疾控专家建议人们尽量佩戴口罩, 降低感染风险. 在工业生产中, 对于不可避免产生的扬尘、沙土, 佩戴具有颗粒物防护功能的口罩也是一项极关键的劳动保护措施. 考虑到一部分人由于工作、生活的特殊需要, 可能长期暴露在相对高危的环境中, 如何以尽可能快速高效的方法督促相关人员佩戴口罩并保证落实效果, 保护人们的生命健康与安全, 便成为一项重大考验. 不仅出行运输行业面临车辆地域分布广泛、活动频率较高、活动时间跨度较长、司机数量庞大等现实问题, 其他行业也存在人力资源紧张, 监督成本高昂等情况, 完全依赖人力进行检查不可避免地存在工作强度大、效率低、覆盖面窄、时效性较差等弊端. 因此, 利用计算机视觉技术替代人工进行生产安全检查工作无疑是具有重大积极意义的.

引用格式: 张修宝, 林子原, 田万鑫, 等. 全天候自然场景下的人脸佩戴口罩识别技术. 中国科学: 信息科学, 2020, 50: 1110–1120, doi: 10.1360/SSI-2020-0046
Zhang X B, Lin Z Y, Tian W X, et al. Mask-wearing recognition in the wild (in Chinese). Sci Sin Inform, 2020, 50: 1110–1120, doi: 10.1360/SSI-2020-0046

本文提出一种结合注意力学习对人脸戴口罩区域进行判别的口罩佩戴识别技术方案. 这一方案由基于特征融合和分割监督的人脸检测 (DFS – detection with feature fusion and segmentation supervision) 和口罩佩戴识别两大技术模块组成. 前者实现在图像中准确定位人脸区域的功能, 后者在单个人脸区域的基础上利用注意力学习进一步分析人脸属性, 判断人脸是否佩戴口罩. 测试结果表明, 上述技术方案有效解决了 24 小时复杂光照、多种类型遮挡、人脸姿态变化、不同距离人脸尺度、口罩款式类型多样化等实际应用中的难题. 在人脸佩戴口罩识别任务中, 图像级准确率达到 99.50%, 可快速定位未按要求佩戴口罩的重点人员并且可以灵活适应不同场景环境. 配合出行服务运营平台的教育、服务管控等其他手段, 可进一步督促平台用户做好个人健康防护; 结合社会各企事业单位的实际情况和防控管理措施, 可切实有效做好监督工作, 为各行各业的安全生产提供强大科技助力.

2 相关研究

人脸佩戴口罩识别是一种人脸属性识别, 它包括人脸检测技术和口罩佩戴分类技术两部分. 由于人脸的特殊性, 在目标检测技术的基础上发展了一系列的人脸检测算法. 而口罩佩戴识别可通过目标分类技术来实现. 因此人脸佩戴口罩识别涉及目标检测技术和分类技术.

2.1 目标检测技术

目标检测技术是指从一幅场景中找出所有感兴趣的目标, 确定它们的位置和类别, 包括了检测和分类两个过程. 近几年来深度学习技术飞速发展, 目前基于深度学习技术的目标检测算法分为两类 [1]: 一类是双阶段目标检测算法, 另外一类是单阶段目标检测算法. 常见的双阶段目标检测算法包括 Faster R-CNN [2], R-FCN [3] 和 FPN [4] 等. 该类算法在基于特征提取的基础上, 由独立的网络分支生成大量的候选区域, 然后对这些候选区域进行分类和回归, 确定目标的准确位置框和类别. 对于单阶段目标检测算法, 常见的有 YOLO V3 [5], SSD [6] 和 RetinaNet [7] 等. 该类算法直接在生成候选区域的同时进行分类和回归. 通常情况下, 双阶段目标检测算法精度更高, 而单阶段的目标检测算法速度更快.

人脸检测是目标检测的一种特例. 对于任意一幅输入的图像, 采用一定的算法或策略对其进行搜索以确定其中是否含有人脸, 如果存在人脸, 则返回所有人脸的位置等信息. 目标检测一般会检测多个类别, 而人脸检测是二类问题, 只检测人脸和背景两类. 针对人脸类别的单一性, 以及人脸的五官特殊性, 在目标检测算法的基础上, 发展出了大量的人脸检测算法. MTCNN [8] 以 3 个级联网络实现快速人脸检测, 并利用图像金字塔实现不同尺度人脸的检测; Face R-CNN [9] 基于 Faster R-CNN [2] 框架进行人脸检测; SSH [10] 提出了对不同深度的卷积层分别进行检测以实现多尺度; FAN [11] 提出了基于锚点级的注意力机制; PyramidBox [12] 利用人脸的上下文信息提高遮挡人脸检测, 即结合人头、身体等信息. 上述算法主要解决不同于其他领域的人脸多尺度、遮挡等问题 [13,14].

2.2 目标分类技术

目标分类技术按照目标的性质、用途等进行归类. 基于深度学习的分类技术包括骨干网络和损失函数两部分. 骨干网络有 AlexNet [15] 使用层叠的卷积层以及 Dropout, Relu 等; VGG-Net [16] 使用更深的网络结构, 探索了深度与性能的关系; GoogLeNet [17] 在增加网络深度和宽度的同时减少参数, 在多个尺寸上同时进行卷积再聚合, 并使用 1×1 的卷积来进行升降维; ResNet [18] 设计了一种短连接的结构, 解决了神经网络加深后的退化问题. 损失函数常见有 sigmoid 交叉熵和 softmax 损失函数等.

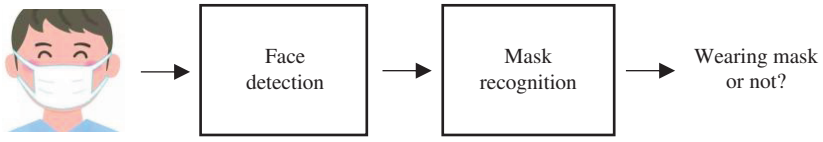

图 1 (网络版彩图) 人脸佩戴口罩识别整体框图

Figure 1 (Color online) Block diagram of face mask recognition

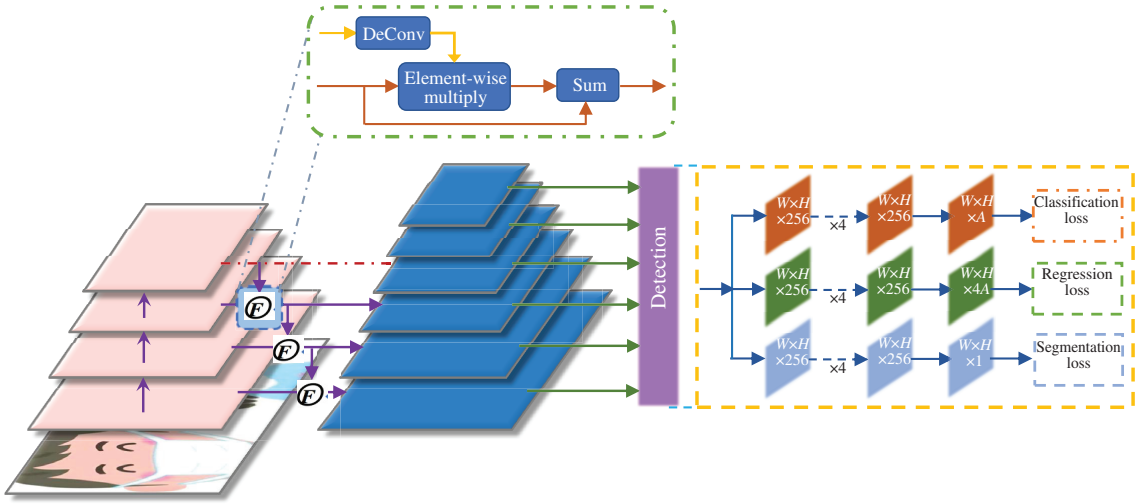

图 2 (网络版彩图) DFS 人脸检测算法整体架构图

Figure 2 (Color online) Architecture of DFS face detection algorithm

3 基于深度学习的人脸佩戴口罩识别

本文提出了一种基于 CNN 的人脸佩戴口罩识别算法, 包括人脸检测和佩戴口罩识别两部分. 如图 1 所示, 人脸检测模块首先对输入的图像进行人脸检测; 佩戴口罩识别模块对检测到的人脸区域按照一定比例扩展, 并裁剪出人脸区域子图像, 采用图像分类的方法, 对整个人脸区域子图像进行二分类, 从而得到是否佩戴口罩的识别结果.

3.1 人脸检测

人脸检测模块使用了 DFS [19] 算法. 该算法以卷积神经网络中的特征融合为基础, 利用特征融合金字塔 [4] 结构同时以空间和通道注意力学习 (attention) 的方式融合高低层特征, 以防止高层特征图上的语义信息覆盖低层特征图上的细节信息. 这使得语义和细节相互补充, 在不失细节信息的同时将语义信息作为上下文线索从而能够增强低层特征. 此外, 为使检测网络以一种弱监督的方式实现注意力学习以便学习到更显著的特征, 该算法还提出了一种辅助训练单阶段检测器的语义分割分支. 具体地说, DFS 算法独特的语义分割分支能够分层地利用更强的语义分割监督信息监督训练网络, 使得用于预测人脸的各级特征图专注于各自最适宜检测到的不同尺寸人脸. 如图 2 所示, 输入图像经过网络逐层提取特征后, 自最高层向下, 相邻层间通过特征融合模块进行特征融合后, 再和低一层的特征进行融合, 依此类推. 融合后的各层特征被 Detection 模块用于对应尺度人脸的检测, 其损失函数由分类、回归和分割 3 个部分构成.

(1) 特征融合. 从图 2 中可以看到模型中的特征金字塔结构, 以及实现从上到下特征融合的模块 F, 特征融合算法可以表示为

$$\widetilde{\phi_i} = F(\phi_{i+1}, \phi_i; \theta), \tag{1}$$

展开上述公式, 具体的实现方式如下:

$$\widetilde{\phi_i} = \phi_i \cdot \psi(\phi_{i+1}; \theta) + \phi_i, \tag{2}$$

其中, ϕ_i 和 ϕ_{i+1} 分别表示浅层特征和更深层的特征, $\widetilde{\phi_i}$ 表示更新后的特征图, ψ 表示针对高层特征图的反卷积操作, θ 表示反卷积操作的参数.

采用特征融合的方式, 可以有效结合高层的语义信息和低层的细节信息, 从而提升模型的检测能力, 而加法操作则是为了强化低层的细节信息, 这些细节信息有助于检测一些难样例. DFS 算法采用反卷积的方法对高层特征进行转换, 而不是采用上采样加 1×1 卷积的方式, 有两个方面的优势: 一是采用上采样后将使后续卷积操作的参数量加倍, 从而影响推理速度, 二是开始阶段利用 1×1 卷积对最高层特征降低通道数会不可避免的损失最高层的语义信息, 最终损害融合的效果.

(2) 语义分割. DFS 算法以标注的人脸框作为人脸检测候选框的弱监督信息, 具体到每一层则是标注框在相应层所对应的感受野, 这样使得人脸检测框和分割框在每一层具有相同的尺度. 分割层有效帮助网络快速学习到人脸区域的特征信息, 同时让分类和回归更容易收敛. 在推理时候, 分割分支会被去除, 这样不会增加模型的参数和计算量. 另外直接采用人脸标注框作为分割信息, 无需额外标注, 虽然人脸标注框中会含有一定的背景信息, 但相对于其中占主导的人脸区域, 其影响可以忽略不计. 相对于其他算法利用分割预测图在主网络中针对特征信息进行引导, 弱监督方式不会引入多余的参数和运算.

(3) 注意力学习. DFS 算法中引入了注意力学习, 从而实现对关键特征信息的关注. 如式 (2) 所示, 像素级的乘积操作可以看作是空间和通道注意力学习的相互结合, 以实现层级间互信息的最大化, 从而引导模型在迭代训练中学习到更显著的人脸相关信息.

(4) 损失函数. 模型的整体损失函数由 3 部分构成, 即分类损失函数、回归损失函数和分割损失函数, 如下所示:

$$L = \sum_k \frac{1}{N_k^c} \sum_{i \in A_k} L_c(p_i, p_i^*) + \lambda_1 \sum_k \frac{1}{N_k^T} \sum_{i \in A_k} I(p_i^*) L_r(t_i, t_i^*) + \lambda_2 \sum_k L_s(m_k, m_k^*), \tag{3}$$

其中, k 为特征融合金字塔的层级, A_k 表示特征层 p_k 的锚点集, 当锚点为正时, p_i^* 为 1, 否则为 0, p_i 为模型分类的结果, t_i 为预测边界框的 4 个参数构成的向量, t_i^* 为正锚点所对应的真实边界框. 分类损失 $L_c(p_i, p_i^*)$ 为针对人脸和背景的二分类 focal loss 函数, N_k^c 为参与分类损失计算的锚点的数量. 回归损失 $L_r(t_i, t_i^*)$ 为 smooth L_1 损失, $I(p_i^*)$ 为标志函数, 用于限制损失函数仅计算正锚点的回归, 满足 $N_k^T = \sum_{i \in A_k} I(p_i^*)$. 分割损失 $L_s(m_i, m_i^*)$ 为像素级的 sigmoid 交叉熵, m_k 为每一特征层的分割预测图, m_i^* 为每一特征层弱分割的真实图. λ_1 和 λ_2 为损失函数的权重, 其中 λ_1 为 1, λ_2 为超参数, 用于平衡检测损失和分割损失.

(5) WIDER FACE 评测. 2019 年 4 月 DFS 人脸检测算法在 WIDER FACE 官方验证集和测试集上进行了评测[1)]. WIDER FACE 是目前业界公开的数据规模最大、检测难度最高的人脸检测数据集之一, 由香港中文大学于 2016 年建立, 共包含 32203 张图像和 393703 个人脸标注. 其中, 40% 的数据

1) http://mmlab.ie.cuhk.edu.hk/projects/WIDERFace/WiderFace_Results.html.

表 1 DFS 与其他算法在 WIDER FACE 验证集上的对比

Table 1 Comparison of DFS and other algorithms on WIDER FACE validation set

Algorithms	Easy	Medium	Hard
MTCNN	84.8	82.5	59.8
Face R-CNN	93.7	92.1	83.1
SSH	93.1	92.1	84.5
FAN	95.3	94.2	88.8
PyramidBox	96.1	95.0	88.9
DFS	96.9	95.9	91.2

为训练集 (training), 10% 的数据为验证集 (validation), 50% 的数据为测试集 (testing). 每个集合中的数据根据人脸检测的难易程度分为 easy, medium, hard, 如 WIDER FACE[2)]所示, 由于汇集了人脸尺寸大小变化、拍照角度引起的人脸姿态变化、不同程度的人脸遮挡、表情变化、光照强弱差异, 以及化妆等多种影响因素, 该数据集在全球人脸检测领域极具挑战性, 吸引了多家国内外科技巨头及高校院所在这个数据集上进行算法效果的验证. 如表 1 所示 [19], 在 easy, medium 和 hard 3 个验证子集中, DFS 算法性能分别达到 96.9%AP (AP: average precision), 95.9%AP 和 91.2%AP. 在 easy, medium 和 hard 3 个测试子集中, 性能分别达到 96.3%AP, 95.4%AP 和 90.7%AP. DFS 算法在 6 项评估结果中取得 5 项第一和 1 项第二的成绩[3)], 因此它可以很好地解决实际环境下人脸存在各种姿态变化问题, 如侧脸、低头、仰头, 以及人脸遮挡和不完整等, 同时可有效解决不同光照条件的影响, 如强光、弱光、反光等, 另外对不同尺度人脸的检测有一定的鲁棒性.

3.2 佩戴口罩识别

(1) 识别方法. 通过人脸检测获得图像中的人脸区域后, 需要识别当前人脸区域是否佩戴口罩. 佩戴口罩识别可以采用检测和识别两种方法. 口罩检测需要对人脸区域中的口罩目标进行框标注, 标注工作量大, 且训练过程中产生大量的候选框, 需要对这些候选框分别进行分类和回归, 因此整个训练和推理过程较为复杂, 计算量大, 速度较慢. 采用目标分类方法, 即对人类区域进行分类, 仅需标注类别, 标注工作量小. 另外目标分类整个训练和推理过程相对简单, 计算量小, 速度快. 由于口罩面积较大, 在人脸占比大, 因此采用目标分类方法可以取得非常好的效果. 综合考虑实际中业务需求的紧急性以及应用中的计算成本, 本文选择了目标分类的方法, 这样在保证精度的同时, 提高算法的速度, 从而能够应对日均千万级的海量检测请求, 节约计算资源和成本, 真正实现更准、更快、更省的产品应用需求.

(2) 模型结构. 佩戴口罩识别模型是基于 ResNet50 改进的, 加入了注意力学习机制, 进一步强化口罩区域, 并针对损失函数进行了优化等. 图 3 所示为佩戴口罩识别的模型结构示意图, 图像经过输入层的数据预处理后, 依次经过模型的 Block1, Block2, Block3, 并和 attention 层的权重相乘后经过 Block4 和 Block5 提取特征后输出分类结果. 在输入层的数据预处理中, 提出了人脸区域扩展的方法, 即针对人脸检测获得的边界框进行一定比例的扩展. 这是因为实际人脸检测中, 由于受到多种复杂条件的影响, 检测框少数情况下会存在一定程度的坐标误差. 通过边界框扩展不但可以消除检测误差, 还能够包含更多的人脸佩戴口罩区域, 如耳带等, 从而提高模型的分类能力. 实验中采用了宽高等比

2) http://shuoyang1213.me/WIDERFACE/index.html.

3) https://www.sohu.com/a/310461017_100092059.

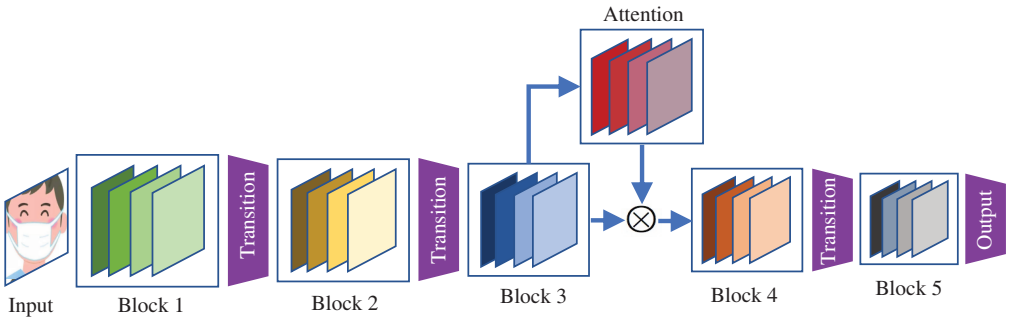

图 3 (网络版彩图) 基于 **Resnet50** 的戴口罩识别模型结构示意图

Figure 3 (Color online) Structure diagram of mask recognition model based on Resnet50

表 2 实验数据来源及数量分布

Table 2 Experimental data sources and quantity distributions

Data source	Mask	NoMask	Total (thousand)
Collected online	5	30	35
In-car	40	65	105
Phone	40	10	50
Recorded	2	3	5
Total	87	108	195

例扩展, 扩展的比例是 20%. 另外, 针对口罩在人脸中的佩戴位置固定的特性, 模型中加入注意力学习机制, 从而提高针对困难样本的识别能力. 如图 3 所示, 在 ResNet50 网络模型的 Block3 层后加入了 attention 网络层, 引导模型重点关注人脸中的口罩区域, 从而对特征层中的对应区域进行加权, 弱化非口罩区域的特征表达, 提高口罩区域特征的显著性. 本文的 attention 网络采用了卷积的形式, 尺度与 Block3 层特征图相对应, 利用点对点乘积的方式对该层特征图进行加权, 使 Block3 层特征更关注相应尺度下的口罩区域, 从而实现注意力学习.

4 实验

4.1 实验数据

由于应用环境是全天候自然环境, 人脸佩戴口罩识别面临很多问题, 包括各种人脸姿态, 各种人脸尺度, 各种拍摄条件, 各种口罩类型、颜色、样式和材质. 本文通过不同类别数据的搜集、增强处理等实现训练样本的多样化. 实验数据有 4 个来源: 网上搜集的公开数据 (collected online)、车内车载设备数据 (in-car)、用于质检的手机图像数据 (phone), 以及部门同学录制的数据 (recorded), 其中后 3 个为企业私有数据, 具体的类别和数量分布如表 2 所示. 为增强数据的多样性, 除了在训练数据的搜集中尽可能多的获取不同类型的数据外, 还在模型训练过程中动态地进行了数据增强, 包括翻转、平移、缩放、裁剪等, 从而提高模型针对不同视角、位置、尺度、缺损情况下的检测和分类能力.

4.2 参数设置

模型训练基于 NVIDIA Tesla P40 的 4GPU, 训练过程中采用了随机梯度下降的方法, 其 mo-

表 3　基于质检手机图像的人脸佩戴口罩识别结果

Table 3　Face mask recognition results on mobile phone image

Test set	Target	Result	
		NoMask	Mask
Designated drive (100 Mask + 105 NoMask)	Precision (%)	100.00	98.99
	Recall (%)	99.06	100.00
	Accuracy (%)	99.51	
Designated drive (3247 Mask + 1111 NoMask)	Precision (%)	93.33	98.63
	Recall (%)	96.03	97.66
	Accuracy (%)	97.24	
Car hailing driver (1455 Mask + 25 NoMask)	Precision (%)	96.15	100.00
	Recall (%)	100.00	99.73
	Accuracy (%)	99.73	

mentum 为 0.9, weight_decay 为 0.0005, base_lr 为 0.001, 采用 multistep 的学习策略, gamma 为 0.1, batch_size 为 64, max_iter 为 10 万次.

4.3　评价指标

为评价模型的性能, 对戴口罩和不戴口罩两类均利用 Precision (精确率)、Recall (召回率) 进行了评价, 并利用 Accuracy (准确率) 对两者综合性能进行了评价, 具体如下 [20]:

$$\text{Presion} = \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100\%, \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100\%, \tag{5}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \times 100\%, \tag{6}$$

其中, 戴口罩样本被正确分到戴口罩类别, 记这一类样本为 TP; 不戴口罩类别被错误分到戴口罩类别, 记这一类别为 FP; 戴口罩样本被错误分到不戴口罩类别, 记这一样本为 FN; 不戴口罩样本被正确分到不戴口罩类别, 记这一样本为 TN. 精确率表示模型检测到的所有样本中该类正确样本数量的占比; 召回率表示模型检测的所有样本中该类正确样本数量占集合中所有该类样本数量的比例; 准确率表示模型检测的所有类别中各自正确样本数量的和占集合所有样本数量的比例. 精确度和召回率反映的是模型针对一类样本的性能指标, 而准确率反映的是模型针对所有类别的整体性能指标.

4.4　实验分析

(1) 验证实验. 为了验证人脸佩戴口罩识别的效果, 在基于质检手机图像和车载设备图像的不同数据集上进行了测试. 两种不同图像的区别是, 手机拍照一般正对人脸, 成像质量更高、更清晰, 代驾质检手机图像中一般司机佩戴头盔, 且绝大部分是在晚上拍摄, 光线较差, 网约车质检手机图像白天和晚上均有, 没有佩戴头盔; 而车载设备安装在车内后视镜位置, 即人脸右前上方, 存在人脸姿态变化较大, 车内环境复杂, 即光照、遮挡、表情、不完整等问题, 实验结果如表 3 和 4 所示.

从表 3 和 4 中可以看出, 测试数据集覆盖均衡情况、不均衡情况、白天时间、夜晚时间、常规样本、易错样例、难样例等多种情况, 因此具有充分性. 另外由于车载图像较质检手机图像的成像环境更复杂, 且无法要求驾驶员的配合, 因此其整体佩戴口罩指标低于后者, 与预期和实际情况相符.

表 4 基于车载图像的人脸佩戴口罩识别结果

Table 4 Face mask recognition results based on vehicle monitoring image

Test set	Target	Result	
		NoMask	Mask
Test 2k (1k Mask + 1k NoMask)	Precision (%)	99.10	100.00
	Recall (%)	100.00	99.11
	Accuracy (%)	99.55	
Test 1.5k (little Mask)	Precision (%)	99.71	89.83
	Recall (%)	98.86	97.25
	Accuracy (%)	98.71	
Night bad case (263 Mask)	Precision (%)	NAN	100.00
	Recall (%)	NAN	98.46
	Accuracy (%)	98.46	
Day 1.1k (117 Mask + 1037 NoMask)	Precision (%)	96.72	70.34
	Recall (%)	96.62	70.94
	Accuracy (%)	94.02	
Night 3.6k (419 Mask + 3246 NoMask)	Precision (%)	96.27	76.15
	Recall (%)	97.13	70.88
	Accuracy (%)	94.13	

表 5 基于质检手机图像的对比实验结果

Table 5 Comparative experiment results on mobile phone image

Test set	Target	Data collected online		In-car data		Attention mechanism	
		NoMask	Mask	NoMask	Mask	NoMask	Mask
Designated drive (100 Mask + 105 NoMask)	Precision (%)	91.70	100.00	100.00	98.99	100.00	98.99
	Recall (%)	73.30	65.00	99.06	100.00	99.06	100.00
	Accuracy (%)	96.60		99.51		99.51	
Designated drive (3247 Mask + 1111 NoMask)	Precision (%)	77.49	99.46	91.26	99.02	93.33	98.63
	Recall (%)	98.56	90.21	97.20	96.83	96.03	97.66
	Accuracy (%)	92.34		96.92		97.24	

(2) 对比实验. 针对人脸佩戴口罩识别算法中加入垂直领域的数据, 增加注意力学习机制的前后效果, 进行了对比实验, 其中表 5 给出了在质检手机数据集上的对比实验结果, 表 6 给出了在车载数据集上的对比实验结果, 表 7 给出了针对车载难样例的对比实验. 从表 5 和 6 中可以看到, 加入车载图像数据后, 在所有测试集上准确率均具有不同程度的提升, 特别是易错测试集上提升较大. 加入注意力学习机制后, 除了均匀集上已经达到较高的准确率并保持不变外, 其他测试集上也有不同程度的提升. 可以看出加入车载图像和注意力学习机制, 提升了模型的整体识别能力.

为了更充分地测试加入注意力学习机制的效果, 对白天和晚上的难样例进行了进一步的对比实验, 其中的难样例包括人脸不完整、侧脸角度大、低头角度大、模糊、部分遮挡、存在反光、夜晚光线暗、佩戴口罩不规范、少见的防毒防尘口罩等. 如表 7 中所示, 显然, 加入了注意力学习机制后, 两者准确率均有近 2 个百分点的提升, 显示了该方法在针对难样例上的有效性.

表 6 基于车载图像的对比实验结果

Table 6 Comparative experiment results on vehicle monitoring image

Test set	Target	Data collected online		In-car data		Attention mechanism	
		NoMask	Mask	NoMask	Mask	NoMask	Mask
Test 2k (1k Mask + 1k NoMask)	Precision (%)	98.31	100.00	99.10	100.00	99.10	100.00
	Recall (%)	100.00	98.31	100.00	99.11	100.00	99.11
	Accuracy (%)	99.15		99.55		99.55	
Test 1.5k (little Mask)	Precision (%)	99.14	84.03	99.62	88.98	99.71	89.83
	Recall (%)	98.20	91.74	98.77	96.33	98.86	97.25
	Accuracy (%)	97.59		98.54		98.71	
Night bad case (263 Mask)	Precision (%)	NAN	100.00	NAN	100.00	NAN	100.00
	Recall (%)	NAN	61.24	NAN	98.06	NAN	98.46
	Accuracy (%)	61.24		98.06		98.46	

表 7 基于车载难样例的对比实验结果

Table 7 Comparative experiment results on difficult vehicle monitoring samples

Test set	Target	In-car data		Attention mechanism	
		NoMask	Mask	NoMask	Mask
Day 1.1k (117 Mask + 1037 NoMask)	Precision (%)	96.76	62.69	96.72	70.34
	Recall (%)	95.18	71.79	96.62	70.94
	Accuracy (%)	92.18		94.02	
Night 3.6k (419 Mask + 3246 NoMask)	Precision (%)	96.63	63.93	96.27	76.15
	Recall (%)	94.58	74.46	97.13	70.88
	Accuracy (%)	92.28		94.13	

综上所述, 本文提出的人脸佩戴口罩识别的方法, 在不同测试集上进行了充分的验证, 具有的高准确率性能可满足行业的实际应用要求.

5 结语与展望

本文提出的人脸口罩佩戴识别技术方案是立足于活用技术解决生产生活中的紧迫问题这一目标, 针对性地在预处理、算法、模型、数据、训练等维度做优化改进, 将人脸检测和人脸属性识别技术结合, 并积极利用损失函数权重策略和数据增强等方法, 最终产生的一套综合性解决方案. 2020 年 1 月末起, 滴滴将这一技术应用于出车前的智能出车质检系统以及车载设备中, 能在出车前和行程中基于这一技术自动分析平台网约车司机是否佩戴口罩以及佩戴是否规范, 以进一步督促司机做好个人防护. 截至 4 月 13 日, 智能出车质检系统已覆盖全国 351 个城市. 自 2020 年 2 月 19 日起, 这套口罩佩戴识别技术方案已经对全社会开放和开源[4], 企业和个人开发者均可快速便捷地获取和部署, 配合各种云、边设备, 依照各级公共设施、出入口、检查站、生产经营场所的安全生产相关规定, 满足对人员佩戴口罩情况的全天候实时检测需求. 这不仅是计算机视觉技术助力公共卫生建设, 保障人民群众健康和生命安全的又一实例, 更深刻体现了科学技术服务社会, 推动经济发展的重大社会价值.

4) http://github.com/didi/maskdetection.

展望未来, 我们在钻研先进技术的基础上, 更需要持续深耕应用场景, 主动挖掘更多现实应用难题, 比如: 各类公共场所低分辨率监控图像的识别问题, 各行业从业人员佩戴口罩是否规范的问题, 是否按标准实施其他防护工作的动作理解问题等, 并尝试寻找新的技术突破口. 期待计算机视觉技术在学术研究和实践应用两方面都能取得长足发展, 更期待社会生产的方方面面都能持续受惠于科技进步.

参考文献

1 Zou Z X, Shi Z W, Guo Y H, et al. Object detection in 20 years: a survey. 2019. ArXiv:1905.05055

2 Ren S Q, He K M, Girshick R B, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 1137–1149

3 Dai J F, Li Y, He K M, et al. R-FCN: object detection via region-based fully convolutional networks. In: Proceedings of Conference on Advances in Nerual Information Processing Systems, Barcelona, 2016. 379–387

4 Lin T, Dollar P, Girshick R B, et al. Feature pyramid networks for object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 2017. 936–944

5 Redmon J, Farhadi A. YOLOv3: an incremental improvement. 2018. ArXiv:1804.02767

6 Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016. 21–37

7 Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell, 2020, 42: 318–327

8 Zhang K P, Zhang Z P, Li Z F, et al. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett, 2016, 23: 1499–1503

9 Wang H, Li Z F, Ji X, et al. Face R-CNN. 2017. ArXiv:1706.01061

10 Najibi M, Samangouei P, Chellappa R, et al. SSH: single stage headless face detector. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), Venice, 2017. 4885–4894

11 Wang J F, Yuan Y, Yu G. Face attention network: an effective face detector for occluded faces. 2017. ArXiv:1711.07246

12 Tang X, Du D K, He Z, et al. Pyramidbox: a context-assisted single shot face detector. In: Proceedings of European Conference on Computer Vision (ECCV), Munich, 2018. 797–813

13 Pang Y W, Xie J, Khan M H, et al. Mask-guided attention network for occluded pedestrian detection. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, 2019. 4966–4974

14 Xie J, Pang Y W, Cholakkal H, et al. PSC-Net: learning part spatial co-occurence for occluded pedestrian detection. 2020. ArXiv:2001.09252

15 Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. In: Proceedings of International Conference on Neural Information Processing System, Lake Tahoe, 2012. 1097–1105

16 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of International Conference on Learning Representations, San Diego, 2015

17 Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, 2015

18 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016. 770–778

19 Tian W X, Wang Z X, Shen H F, et al. Learning better features for face detection with feature fusion and segmentation supervision. 2018. ArXiv:1811.08557

20 Qian Y, Ding X, Liu T, et al. Identification method of user's travel consumption intention in chatting robot. Sci Sin Inform, 2017, 47: 997–1007 [钱岳, 丁效, 刘挺, 等. 聊天机器人中用户出行消费意图识别方法. 中国科学: 信息科学, 2017, 47: 997–1007]

Mask-wearing recognition in the wild

Xiubao ZHANG[1*], Ziyuan LIN[1], Wanxin TIAN[1,2], Yan WANG[1], Haifeng SHEN[1] & Jieping YE[1]

1. *AI Labs, Beijing DiDi Infinity Technology and Development Co., Ltd, Beijing* 100094, *China*;
2. *School of Information and Communication Engineering, Beijing University of Post and Telecommunication, Beijing* 100876, *China*
* Corresponding author. E-mail: zhangxiubao@didiglobal.com

Abstract For public health and safety, wearing of masks is one of the most significant means to prevent infections. Additionally, masks protect employees of heavy industry from certain diseases during manufacture. To meet the demand of automatic mask-wearing recognition in scenes of life, we propose a recognition algorithm based on face detection and face attribute recognition. The face detection model not only adopted a fused feature pyramid and a spatial and channel attention mechanism but also a segmentation branch for weak supervision learning. Then for the detected face, we used classification for fast recognition. Moreover, we employed nearly 200000 images, attention mechanisms, data augmentation, and other techniques to enhance the robustness. Besides, this technology has been widely used in Didi Chuxing's inspection systems and achieves 99.50% accuracy. Importantly, both the service and key algorithms have been opened to the public to maximize their social and application value.

Keywords mask-wearing recognition, face detection, face attribute recognition, feature pyramid, attention

Xiubao ZHANG was born in 1981. He received his Ph.D. degree in instrument science and technology from Beijing University of Aeronautics and Astronautics, Beijing, in 2012. Currently, he is an expert algorithm engineer in the AI Labs, Didi Chuxing. Before that, he worked at Leshi Internet Information and Technology Corp., etc. His research interests include face detection and recognition, pedestrian detection, and person ReID.

Ziyuan LIN was born in 1991. She received her B.S. degree from the School of Humanities, Beijing Language and Culture University, China, in 2014. She works as a senior project manager in AI Labs, Didi Chuxing, focusing on the field of computer vision, at present.

Haifeng SHEN was born in 1977. He received his Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications, Beijing, in 2006. Currently, he is a principal algorithm engineer and leads the computer vision group in AI Labs, Didi Chuxing. Before that, he worked at several companies like Panasonic, Baidu, and Microsoft. His research interests include computer vision and speech recognition.

Jieping YE was born in 1975. He received his Ph.D. degree in computer science from the University of Minnesota, Twin Cities, in 2005. Currently, he is the head of DiDi AI Labs, a VP of Didi Chuxing, and a DiDi Fellow. He is also a professor at the University of Michigan, Ann Arbor. His research interests include big data, machine learning, and data mining with applications in transportation and biomedicine.