计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(3):164−170 [doi: 10.15888/j.cnki.csa.007835]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于人体姿态估计的手机使用状态监控①

刘　军[1], 范长军[2], 瞿崇晓[2]

[1](中国人民解放军 63650 部队, 乌鲁木齐 841700)
[2](中国电子科技集团公司第五十二研究所, 杭州 310012)
通讯作者: 范长军, E-mail: fcjun@126.com

摘　要: 智能手机的日益普及给人们带来便捷的同时也带来了不少的隐患, 在一些特定的场景下需要对手机的使用进行监控和限制. 本文设计了一套手机使用状态监控系统, 先采用 YOLOv3 检测图像中的人体, 然后通过 OpenPose 姿态估计算法获得人体关节点, 再通过 YOLOv3 判断手部区域是否有手机, 最后通过神经网络分类器识别当前的手机使用状态. 系统的应用测试表明该方案具有良好的检测与识别效果, 能够满足相关场景的应用需求.

关键词: 人体检测; 人体姿态估计; 手机检测; 手机使用状态监控

引用格式: 刘军,范长军,瞿崇晓.基于人体姿态估计的手机使用状态监控.计算机系统应用,2021,30(3):164–170. http://www.c-s-a.org.cn/1003-3254/7835.html

Phone Usage Monitoring Based on Human Pose Estimation

LIU Jun[1], FAN Chang-Jun[2], QU Chong-Xiao[2]

[1](Chinese People's Liberation Army No. 63650, Urumqi 841700, China)
[2](The 52nd Research Institute of China Electronics Technology Group Corporation, Hangzhou 310012, China)

Abstract: The increasing popularity of smartphones brings not only convenience to people but also a lot of hidden dangers. Thus, it is necessary to monitor and restrict the use of phones in some specific situations. In this paper, we design a system to monitor phone usage. First, YOLOv3 is used to detect human bodies in an image. Then, the joints for each person are obtained by the OpenPose pose estimation algorithm. Furthermore, YOLOv3 is employed to judge whether there is a mobile phone in the hands. Finally, the current phone usage status is recognized by a neural network classifier. The experimental results show that the proposed scheme has good detection and recognition performance and can meet the application requirements in relevant scenarios.

Key words: human detection; human pose estimation; phone detection; phone usage monitoring

随着信息技术的快速发展, 手机的使用越来越普遍, 人对手机的依赖程度越来越高, 带来方便快捷的同时也给人们的生活和工作带来了一系列问题和挑战. 比如, 驾驶员在开车时因为打电话或玩手机而手离开方向盘, 或行人在过马路时低头看手机, 从而导致发生车祸的事件时有发生; 在学校里, 学生低头"刷手机"的行为也给学习效果和课堂纪律带来了不良的影响. 此外, 一些特殊场所, 如部队驻地或者涉及信息安全的重要资料档案室, 手机的违规使用容易造成失泄密事件, 对国家安全造成不必要的损失. 在上述场景下手机的使用是被严格控制的, 需要对相关人员的一些违规使用手机行为进行实时预警, 其中就包括通过摄像头检测他们是否在打电话、玩手机或用手机拍照等. 因此, 研究对手机使用状态的监控具有重要的应用价值和现实意义.

当前, 已有一些关于手机使用状态检测与监控的

① 收稿时间: 2020-06-10; 修改时间: 2020-07-07; 采用时间: 2020-09-01; csa 在线出版时间: 2021-03-03

© 中国科学院软件研究所 http://www.c-s-a.org.cn

研究工作, 主要集中于安全驾驶领域针对驾驶员打电话行为的检测. 文献 [1] 首先采用渐进校准网络算法进行人脸的检测与实时跟踪, 确定打电话检测候选区域, 然后通过基于卷积神经网络的算法在候选区域实现驾驶员打电话行为的检测. 文献 [2] 先对监控图像中目标车辆的车窗、驾驶员候选区域进行定位, 获得驾驶员的头肩区域后, 再采用卷积神经网络进行接打电话的检测与识别. 除了此类基于计算机视觉的方法[3], 还可以通过传感器进行驾驶员打电话的识别, 如文献 [4] 采用 WiFi 和手机传感器相结合的方式来检测与识别危险驾驶动作, 此类应用受设备和场景的限制较大. 与上述驾驶员打电话相关的研究工作相比, 针对玩手机等行为进行识别的研究工作较少. 文献 [5] 从采集的图象中截取包含人体的周围区域, 判断人是否拿着手机或者人体周围是否有手机, 计算人脸的朝向, 然后判断人的状态是否为“玩手机”.

上述的研究工作均取得了不错的效果, 但是目前所监控的手机使用状态种类少且单一, 针对一些特殊场景下的复合需求仍缺少相应的研究工作. 比如, 在重要资料档案室等敏感场所打电话、拍照或玩手机等各类手机使用行为具有不同的影响, 需要同时对这些行为进行识别与监控. 此外, 现有的方法对异物遮挡、图像旋转、光照变化等的适应性也各有不足. 近年来, 基于深度学习的人体姿态估计得到快速发展, 给人体行为分析提供了良好的技术支撑. 本文提出了一种基于人体姿态估计的手机使用状态监控系统, 实现了对打电话、玩手机、手机拍照等行为的识别.

1 算法设计与实现

1.1 算法总体框架

基于人体姿态估计的手机使用状态监控系统的落地应用着重考虑两个方面: 手机使用状态检测的准确率和运行效率. 准确高效的系统实现面临着诸多的挑战, 比如, 人体姿态的变化容易遮挡手机, 现有的人体姿态估计算法计算量大耗时较长. 为了应对上述挑战, 设计整个的算法框架与流程如图 1 所示.

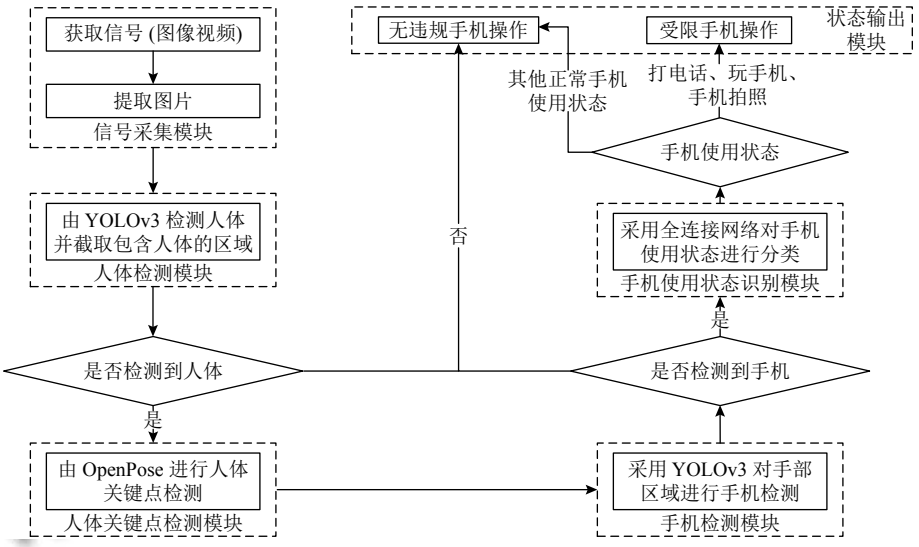

图 1 算法整体框架与流程图

总的算法框架涉及到 3 个关键的功能组件, 分别是目标检测、人体姿态估计以及手机使用状态识别. 其中, 目标检测包括两部分, 分别是人体检测与手机检测. 人体检测用于判断图像中是否有人, 手机检测进而判断人的手中是否握有手机; 人体姿态估计则在中间环节主要用于对检测出的人体进行姿态估计, 以便于得到手部的位置, 方便后续的手机检测; 在检测到手机并获取人体姿态骨架后, 基于人体姿态关键点及其与手机的空间位置关系进行手机使用状态的分类识别.

总体的系统算法流程如下: 首先, 采用 YOLOv3 检测图片中的人体; 其次, 对检测出的人体, 通过 OpenPose 进行人体关键点的检测, 获取手部的位置; 然后, 通过标签为“手”的关键点坐标来获取手部区域, 并采用 YOLOv3 对这些区域进行手机检测, 判断手机是否存

© 中国科学院软件研究所 http://www.c-s-a.org.cn

在; 最后, 根据手机的存在情况设计神经网络分类器, 将人体骨架中与手机操作强相关的若干关节点以及手机的位置作为输入, 进行手机使用行为的识别. 具体的算法流程请参见图 1.

原则上, 人体检测对于采用 OpenPose 进行的人体姿态估计不是必须的, 但是对于一个实际的应用系统而言, 监控场景中并不总是存在“人”, 预先检测和截取人体区域, 可减少不必要的人体姿态估计计算资源和时间, 以提高效率. 此外, 人体关键点检测和手机检测易受身体姿态、遮挡、光照等的影响, 前置步骤估算出目标的预期位置, 便于有针对性地对局部区域进行处理, 提高相应的检测准确度.

1.2 人体检测与手机检测算法

在数据采集之后, 首先要对获取的图像进行人体检测. 在通过 OpenPose 推理得到关键点为“手”的坐标后, 针对手部区域再进行手机检测, 以判断其是否携带手机. 此两类检测的功能不同, 但是检测的原理类似, 此处选用 YOLO v3 来作为算法基线, 以实现相应的功能.

YOLO 最早是由 Redom 等在 2016 年提出的一个端到端的深度卷积神经网络模型, 相比于以 RCNN[6] 系列算法为代表的两步检测网络, 它能够兼顾速度和检测精度[7]. 经过 Redmon 等的持续研究, YOLO 随后发展出 v2、v3 等版本[8,9]. 相比于前两个版本, YOLOv3 采用了特征融合以及多尺度检测的方法, 目标检测的精度和速度都得到了很大提升.

YOLOv3 的网络架构为 darknet-53, 它去掉了 v2 中的池化层和全连接层, 并在前向传播过程中通过改变卷积核的步长来实现张量尺寸的变换; 它采用了残差的设计思想, 用简化的残差块来加深网络结构, 以提升网络的速度; 针对手机等小目标漏检率高的问题, YOLOv3 借鉴了特征图金字塔网络, 增加了从上至下的多级预测, 采用多尺度来对不同大小的目标进行检测, 可解决远距离目标图象过小的问题, 具体参见图 2. YOLOv3 的损失函数主要由 3 部分组成: 目标置信度损失, 目标分类损失, 以及目标定位偏移量损失, 三者之间通过加权系数进行平衡. 针对前两者, 不同于 YOLOv2 采用 Softmax+交叉熵来处理, YOLOv3 采用 n 个二值交叉熵来实现. 交叉熵越小, 代表两个概率分布越接近, 可较好地刻画两个概率分布之间的距离. 针对后者, 采用的是真实偏差值与预测值之差的平方和.

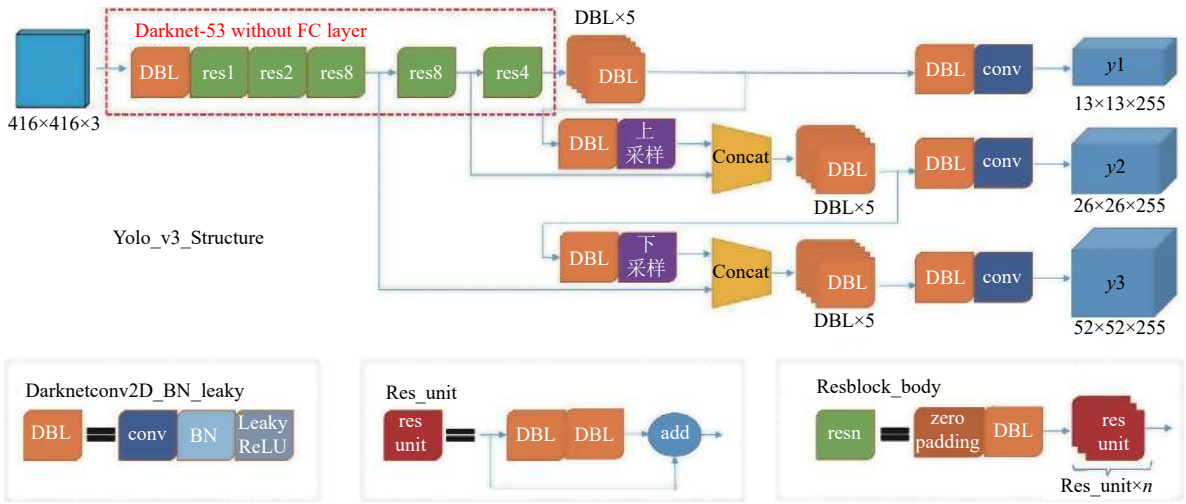

图 2 YOLOv3 的网络结构

人体检测可以直接采用 YOLOv3 在 COCO 数据集上的预训练模型来推理[9], 手机检测对应的网络模型是在 YOLOv3 预训练模型的基础上用采集的数据集重新训练得到的. 在通过人体姿态估计算法获取手部的关节点后, 以此点为中心将手部及其附近区域划出一片固定大小的区域 (如 208×208), 并调整为统一的大小 (如 416×416), 再由采集的数据集对 YOLOv3 模型进行训练.

如图 2 所示, YOLOv3 分别输出 13×13、26×26、52×52 三种不同尺寸的特征图, 并且在回归预测部分每一个单元格借助 3 个锚点框 (anchor box) 预测 3 个边框, 即每个输出张量中的任一网格会输出 3 个预测框.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

以包含 80 种类别的 COCO 数据集为例, 输出张量的维度为 (5+80)×3=255. 其中, 5 表示每个预测框的置信度以及坐标信息, 即 (c, x, y, w, h), 3 则表示每个网格预测的模版框个数. 针对手机检测场景, 仅有一类待检测目标, 故将输出维度变为 (5+1)×3=18, 以减少计算量并提高检测精度和速度.

2 人体姿态估计算法

在人体行为监控中, 人体关键点检测与分析是重中之重, 此类问题往往又被统一归为人体姿态估计问题. 近年来多种人体姿态估计方法被研发出来, 早期的方法只用于单人关键点检测, 先识别出人身体的各个部位, 然后再连接各部分来获得姿态. 近年来多人姿态估计也取得了较快的发展. 多人姿态估计主要分为两类, 第一类是自顶向下 (Top-down) 的方法, 即先检测出图像中的所有人, 再对每一个人进行姿态估计, 这种方法具有较高的准确率但是处理速度不高, 如 AlphaPose; 第二类是自底向上 (down-top) 的方法, 即先检测出所有的关节点, 再判断每一个关节属于哪一个人, 这种方法可以做到实时检测人体关键点, 如 OpenPose.

OpenPose 由卡耐基梅隆大学的研究人员于 2017 年提出[10], 它是一个实时的、多人骨骼关节点检测的二维姿态估计开源库, 可以在单目摄像头的基础上获得实时且准确率高的二维人体骨骼关节点坐标. OpenPose 借鉴了卷积姿态机[11] 中采用大卷积核获得大感受野的思想, 使得 OpenPose 算法可以较好地处理遮挡情况下的人体姿态估计问题. 其网络模型如图 3 所示.

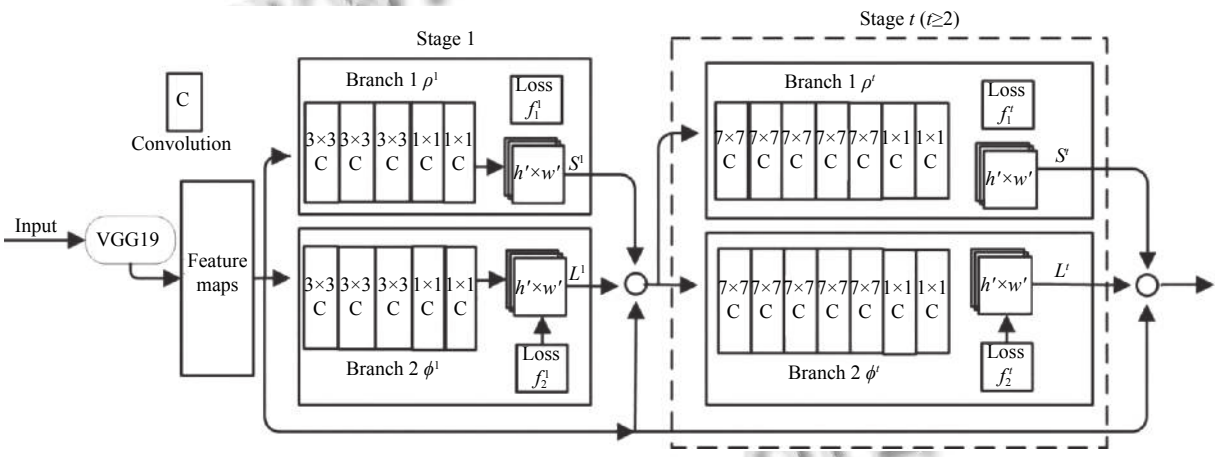

图 3 YOLOv3 整体网络结构图

具体地, OpenPose 模型使用 VGG-19 深度神经网络提取图像的原始特征图 (feature map), 然后再分成两个分支, 第 1 个支路中的每一阶段使用卷积神经网络预测身体关键点的热度图, 第 2 个支路中的每一阶段使用卷积神经网络预测部分亲和字段 (Part Affinity Fields, PAF). 部分亲和字段是记录肢体位置和方向的 2 维向量, 它表示身体各部分之间的关联程度. 关键点热度图和部分亲和字段在每一个阶段下与输入特征层的关系映射视为 S_t 和 L_t ($t\in[1,2,\cdots,T]$), 输入层除第一个阶段为 VGG-19 网络输出的特征层外, 其余阶段 (即 $t\geq2$) 的输入层均为前一个阶段的两个输出向量与 VGG-19 输出层的连接组合, 如式 (1) 所示, 其中, F 是通过 VGGNet 提取出的原始图像的特征, δ^t 和 ξ^t 分别表示在阶段 t 时 L 和 S 的卷积神经网络. 在通过各层网络模型计算之后, 通过贪心推理分析置信度图 S_T 和部分亲和字段 L_T, 为图像中的所有人输出二维身体关节点. 综上, 模型采用尺寸为 $W\times H$ 的彩色图像作为输入, 生成图像中每个人的骨骼关键点的二维图像位置作为输出.

$$\begin{cases} S^t=\delta^t\left(F,S^{t-1},L^{t-1}\right), \forall t\geq 2 \\ L^t=\xi^t\left(F,S^{t-1},L^{t-1}\right), \forall t\geq 2 \end{cases} \tag{1}$$

OpenPose 是一种自下而上的方法, 在无人的背景图区域可能会误检出关节点, 影响整个系统的性能. 上一步骤通过 YOLO 算法检出图像中的人的位置, 可有效缓解此问题. 本文采用了基于微软 COCO 数据集预训练的可检测 18 个身体关键点的模型[12], 其以检测出

© 中国科学院软件研究所 http://www.c-s-a.org.cn

的人体区域彩色图像作为输入, 可输出图像中人体的 18 个骨架关节点的二维坐标及置信度, 如图 4 所示.

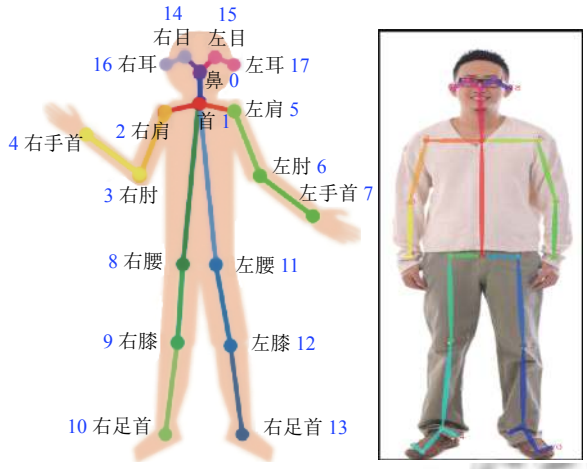

图 4 OpenPose 人体关键点示意图

3 手机使用状态识别算法

经过人体检测、人体姿态估计与手机检测, 能够得到人本身的姿态以及是否携带手机的状态, 此时, 需要进一步判断当前是何种手机使用状态, 因为不同的场景下对手机使用的限制是不同的, 即便在同一场景下不同的手机使用行为造成的潜在影响也是不同的. 手机使用状态的识别处于最后的环节, 受前面诸多环节的影响. 由于受光照变化、遮挡等的影响, 人体检测以及手机检测的效果不一定理想. 比如, 人体检测时目标的置信度过低, 或者手机被遮挡而无法检测出等. 为了兼容上述各类异常情况并保证主要场景下的识别准确度, 本文采用了可配置的规则, 并且设计了阈值配置方案来针对不同情况进行处理.

具体地, 人体检测、手机检测与人体姿态估计都存在着相应的置信度, 当置信度阈值设定过高时, 有可能遗漏待检测的目标, 当置信度阈值设定过低时, 误检测的目标将会很多, 占用大量计算时间. 简而言之, 置信度阈值设定是一个对检测准确度和效率进行折中的过程, 因此要针对具体应用场景对其合理设置. 通常情况下, 针对具体的应用场景可多次试验择优选择对应阈值. 在本文中, 由 OpenPose 采用默认配置实现人体姿态的估计, 并规定人体检测的置信度大于 50% 且手机检测的置信度大于 30% 时才触发下一环节的手机使用状态识别.

为设计神经网络分类器, 首先预设手机的使用状态为 4 类: 打电话、玩手机、手机拍照以及其他活动, 因此分类网络最后的输出层的节点数为 4. 神经网络分类器 (图 5) 的实现具体如下: 前三层神经网络的激活函数采用 tanh 函数, 最后一层网络的激活函数采用 Softmax, 对应的损失函数采用交叉熵, 以将多个神经元的输出映射到 (0, 1) 区间内, 并且这些神经元的输出满足累和为一的性质, 可以将其理解为概率, 也即该图片样本中的人的行为被划分为各类手机使用状态的置信度. 当检测得到手机的置信度大于预设的阈值时, 训练一个神经网络模型, 网络的输入采用人体姿态关键点、手机以及它们的位置坐标. 当检测的置信度无法满足阈值门限时, 将对应的样本自动归类为其他活动.

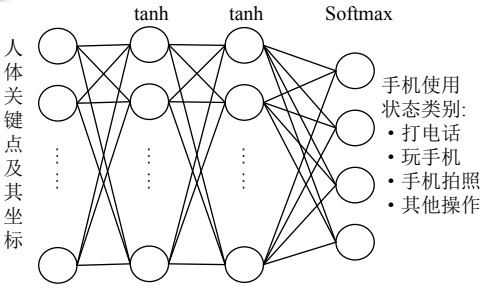

图 5 分类网络结构

4 实验及其分析

为对提出的方法进行验证, 搜集了大量的手机使用状态的照片, 包括打电话、玩手机、手机拍照三类图片的数据各 5000 张, 以及其他手机使用场景 7000 余张, 比如手持手机行走等, 并使用 LabelImg 工具进行标注, 生成训练所需要的 XML 文件.

在实验过程中, 涉及到 4 个神经网络模型, 分别是人体检测模型、人体姿态估计模型、手机检测模型以及手机使用状态分类模型. 其中, 人体检测和人体姿态估计模型分别是在对应的 COCO 数据集上训练得到的开源 YOLOv3 和 OpenPose 预训练模型, 可以直接使用; 手机检测模型则是采用 darknet53.conv.74 预训练权重并修改 yolov3.cfg 配置文件后, 基于收集的数据进一步训练得到的; 手机使用状态分类模型是根据手机的存在情况通过将对应的人体关键点、手机及其坐标输入设计的神经网络分类器中训练得到的. 人体检测、手机检测、人体姿态估计与手机使用状态识别的效果如图 6 所示, 检测结果会以检测框的形式显示, 同时, 也会给出对应的置信度. 从测试的结果可以看出本

© 中国科学院软件研究所 http://www.c-s-a.org.cn

文方案具有良好的识别效果, 能满足相关场景的应用需求.

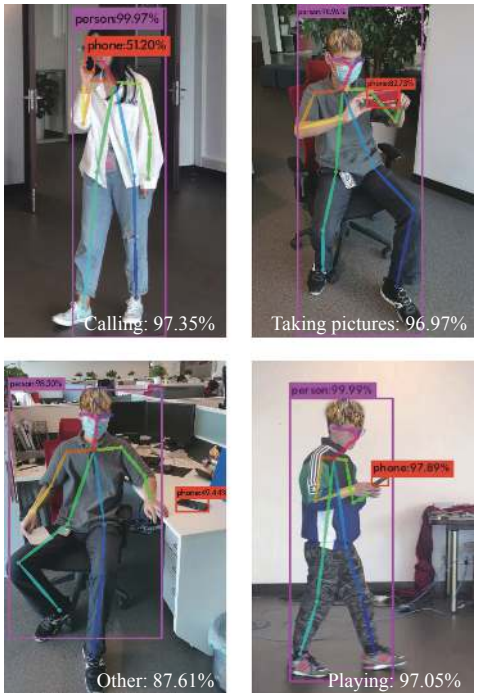

图 6 手机使用状态识别效果图

为方便进行试验, 每一次将采集的数据集随机打散并按照比例进行分配, 训练集 90%, 测试集 10%, 并进行一次试验, 如此循环往复 5 次得到的实验结果如下如表 1 所示. 这里选择分类算法中常用的准确率 (*Precision*)、召回率 (*Recall*) 和 *F*1 值 (*F*1-*score*) 来进行评估, 它们的定义如下:

$$\begin{cases} Precision = \dfrac{TP}{TP+FP} \\ Recall = \dfrac{TP}{TP+FN} \\ F1-score = \dfrac{2\times Precision\times Recall}{Precision+Recall} \end{cases}$$

其中, *TP*、*FP*、*FN* 分别表示将正样本预测为正样本, 将负样本预测为正样本, 以及将正样本预测为负样本的样本数.

表 1 手机使用状态分类结果 (%)

状态	*Precision*	*Recall*	*F*1-*score*
1	91.72	89.23	90.46
2	92.08	90.41	91.24
3	89.98	88.06	89.01
4	90.63	87.37	88.97
5	90.32	88.41	89.35
平均值	90.95	88.70	89.81

从表 1 中可以看出, 本文方法经过 5 次测试得到的平均准确率达 90.95%, 平均召回率达 88.70%, 平均 *F*1 值达 89.81%. 本文提出的方法既能对手机使用状态进行准确地识别, 又能做到比较全面地检测, 在 4 种手机使用状态识别的各类指标上均取得了比较好的效果, 并且检测的结果相对稳定.

目前, 已有一些采用 OpenPose 人体骨架进行行为识别的研究工作[13], 其中一些涉及到手机使用行为识别, 如 NTU RGB+D 动作分析数据集就包含了“打电话”和“玩手机”两类行为[14]. 为验证本文总体方案 (以 M0 表示) 的效果, 在其基础上设计以下 3 类方法: M1—去掉 M0 的人体检测步骤, 根据人体骨架截取人体区域; M2—去掉 M0 的手机检测步骤, 采用 YOLO 在人体区域直接检测手机; M3—去掉 M0 的手机检测步骤, 分类阶段网络的输入不采用手机及其位置信息. 针对方法 M0~M3, 将采集的数据集随机打散, 分割训练集与测试集的比例为 8:2, 分别训练并测试, 得到 *F*1-*score* 的值并比较, 结果如图 7 所示. 从图中可以看出, 当不限定手部位置进行手机的检测或完全不检测手机时, 手机使用状态识别的性能明显下降, 而本文方法考虑了诸多方面的因素, 整合并发挥了 YOLOv3 和 OpenPose 的优势, 取得了较好的识别效果.

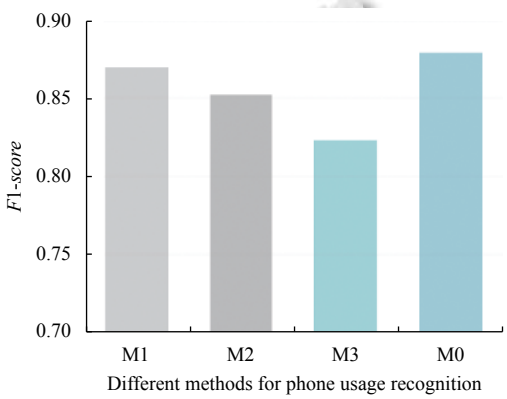

图 7 不同方案的性能对比

此外, 从表 1 中还可以看出, 在一些测试数据集下召回率会偏低, 经过仔细对比发现, 在这些数据集中出现了人体被部分遮挡或手机被完全遮挡的样本, 此时人体检测或手机检测的置信度小于阈值, 导致了此类样本被直接认定为其他活动, 而没有识别出相应的手机使用状态. 这是本文方法待改进之处, 是笔者未来算法优化的重点方向.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

5 结论与展望

本文给出了基于人体姿态估计的手机使用状态监控方案, 用于监控社会生活中的一些手机违规使用场景, 以避免违规使用手机带来的负面影响, 具有一定的现实意义. 该方案整合了 YOLOv3 目标检测算法和 OpenPose 人体姿态估计算法, 先通过人体检测获取人的前景图, 再通过人体姿态估计获取手部的位置, 以提高手机检测的准确度, 并由手机的位置结合人体姿态来判定当前的手机使用状态. 系统测试结果表明, 本文提出的方案应用效果良好, 可以满足相关的应用需求. 针对由于遮挡等因素导致人体或手机检测的置信度过低, 并造成漏检或误识别的情况, 笔者在未来的工作中将优化现有方法, 并着力加以解决.

参考文献

1 熊群芳, 林军, 岳伟, 等. 基于深度学习的驾驶员打电话行为检测方法. 控制与信息技术, 2019, (6): 53–56, 62.

2 陈鑫嘉, 张震. 驾驶员开车接打电话的状态识别方法及装置: 中国, 201611185468.8. 2016-12-20.

3 王丹. 基于机器视觉的驾驶员打电话行为检测 [硕士学位论文]. 北京: 北京理工大学, 2015.

4 杨蕾. WiFi 和手机传感器相结合的危险驾驶动作检测研究 [硕士学位论文]. 西安: 西北大学, 2018.

5 游忍, 邵延华, 刘明华. 基于图像视频检测人“玩手机”的方法及系统. 中国, 201910570793.3. 2019-09-02.

6 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 580–587.

7 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 779–788.

8 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 6517–6525.

9 Redmon J, Farhadi A. YOLO: Real-time object detection. https://pjreddie.com/darknet/yolo/. [2020-03-28]

10 Cao Z, Martinez HG, Simon T, *et al*. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019. [doi: 10.1109/TPAMI.2019.2929257]

11 Wei SE, Ramakrishna V, Kanade T, *et al*. Convolutional Pose Machines. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 4724–4732.

12 Hidalgo G, Cao Z, Simon T, *et al*. OpenPose: Real-time multi-person keypoint detection library for body, face, hands, and foot estimation. https://github.com/w1ld3/openpose. [2020-03-28]

13 吕洁, 李洪奇, 赵艳红, 等. 基于关节空时特征融合的人体行为识别. 计算机工程与设计, 2020, 41(1): 246–252.

14 陈星宏. 基于 3D 骨架的肢体动作识别研究 [硕士学位论文]. 成都: 电子科技大学, 2019.

© 中国科学院软件研究所 http://www.c-s-a.org.cn