第 58 卷 第 22 期/2021 年 11 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于机器视觉的石化场景人员危险行为识别

杨斌，云霄*，董锴文，刘西想，黄瀚

中国矿业大学信息与控制工程学院，江苏 徐州 221116

摘要 针对石油化工场景下传统的人体行为识别算法只关注人员自身行为，无法识别打手机、抽烟等属于人-物交互危险行为的问题，在基于骨骼点的人体行为识别任务中引入目标检测机制，提出基于深度学习的人-物交互行为识别算法。首先，采用 OpenPose 算法进行姿态估计，进而利用行为识别方法获取初始行为类别；其次，针对传统方法丢失背景和语义信息的问题，使用 YOLOv3 算法检测感兴趣物体，获得类别和位置信息；然后，通过判断人与物体的空间位置关系来表征人-物交互关系；最后，提出决策融合策略，对人的初始行为类别、物体信息、人-物交互关系进行决策融合，得到最终的行为识别结果。以打手机和抽烟行为为例对所提算法进行验证分析，结果表明，所提算法可以对石化场景下人员的危险行为进行准确识别。

关键词 机器视觉；姿态估计；行为识别；目标检测；决策融合

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202158.2215001

Human's Dangerous Action Recognition in Petrochemical Scene Using Machine Vision

Yang Bin, Yun Xiao*, Dong Kaiwen, Liu Xixiang, Huang Han

School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China

Abstract Traditional human action recognition algorithms in petrochemical scenarios focus only on human behaviors and cannot recognize other dangerous behaviors prompted by human-object interactions, such as cell phone calls and smoking. To solve this problem, this paper introduces the object detection mechanism in skeleton-based human action recognition task and proposes a recognition algorithm for human-object interaction using deep learning. First, we used the OpenPose algorithm for pose estimation and then employed the action recognition method to obtain the initial action label. Second, to solve the problems of losing background and semantic informations in traditional methods, the YOLOv3 algorithm was used to detect the objects of interest and obtain their category and location informations. Then, we characterized the human-object interaction relationship by determining the spatial relationship between humans and objects. Finally, a decision-making fusion strategy was proposed, merging the initial action categories of the human, object information, and human-object interaction relationship, to obtain the final action recognition result. Cell phone calls and smoking behaviors were used as examples to verify and analyze the proposed algorithm. Results show that the proposed algorithm can accurately identify dangerous personnel behaviors in a petrochemical scene.

Key words machine vision; pose estimation; action recognition; object detection; decision fusion

OCIS codes 150.0155; 100.4996; 100.2960

收稿日期：2020-12-14；**修回日期**：2021-01-12；**录用日期**：2021-01-21

基金项目：江苏省自然科学基金青年项目(BK20180640)、国家自然科学基金青年项目(61902404，51734009，51504255，51734009，61771417)、国家重点研发计划(2016YFC0801403)、江苏省重点研发计划(BE2015040)

通信作者：*yx. tong@163. com

1 引 言

石油化工(以下简称“石化”)属于高危行业,该场景下严禁使用明火、打手机、抽烟等危险行为发生。与传统人工分析视频的方法相比,智能视频监控技术能及时对视频中的危险行为进行识别和预警[1],无需人工干涉,大大提高监控效率和有效性。因此,需要研究基于机器视觉的石化场景人员危险行为识别方法,从而实现智能视频监控系统。

一方面,目前国内外大都针对驾驶员打手机行为的检测方法进行研究,检测方法主要分为基于手机信号检测和基于机器视觉检测两大类。Yang等[2-3]采用手机信号分析方法,对驾驶员是否存在打手机行为进行检测。但该类方法无法区分具体行为人,误检率高。基于机器视觉的检测算法通过视频图像对打手机行为进行检测。文献[4]使用支持向量机模型识别开车打手机行为。文献[5]通过分析手部的姿势和面部图像进行识别,采用快速区域卷积神经网络进行目标检测。文献[6]利用监督下降法定位脸部特征,使用左右两个边界框来判断是否打手机。在这类问题中,由于驾驶室空间狭小,驾驶员位置及其与摄像头之间的角度和距离基本保持不变,人脸一般保持正面,上述方法可以通过有效定位人脸和手部获得较好的识别结果。但石化场景(如加油站)人员杂乱,行人存在移动和侧对、背对摄像头等复杂情况,导致上述驾驶员打手机识别方法不适用于石化场景,目前国内外也缺少对该场景下打手机行为识别的相关研究。

抽烟行为检测方法可分为利用传感器等硬件设备和机器视觉技术两大类。利用烟雾传感器等硬件设备进行烟雾检测的方法通过提高传感器的敏感程度来提升检测性能。例如,文献[7]通过采用同质半导体传感器来提升烟雾粒子探测器的性能;文献[8]采用含氖硅胶片,研发出一种非基于 γ 射线检测烟雾粒子和电离子的检测器。但改进硬件所需成本较高,且对于石化场景这样的室外场所,空气流动、灰尘、水蒸气等干扰因素导致抽烟行为产生的有限烟雾量可能无法达到传感器的检测敏感度,因此该类检测方法局限性较大。基于机器视觉对视频烟雾检测的研究主要集中于烟雾的颜色、形状、纹理等静态特性和烟雾的运动、扩散等动态特性。Millan-Garcia 等[9]利用烟雾的颜色空间特征对视频图像进行处理,排除了非烟雾区域。该方法的不足之处是颜色信息对阈值的设定较为敏感。Yuan[10]利用块运动方向模型,通过累积烟雾的主运动方向进行烟雾检测。李鹏等[11]提出了一种将高斯混合模型与卷积神经网络相结合的视频烟雾检测方法。但此类方法在训练前必须对提取的烟雾特征进行人工整合,否则难以达到准确性和实时性要求,且获取训练样本难度大,没有代表性。因此,研究者开始利用如抽烟手势等其他特征对抽烟行为进行检测,Davis[12]建立了不同的手势模型,对采样图像中手指尖的运动轨迹与各个手势模型进行匹配,进而进行手势检测。但该类方法只能检测人的手势,忽视物体(此处指香烟)信息,难以区分与抽烟行为手势相似的其他行为。

另一方面,基于 RGB 视频的行为识别方法易受背景、光照等无关因素的影响,导致识别精度不高、鲁棒性较差[13-14]。人体姿态估计算法的发展使得基于骨骼点的人体行为识别方法成为研究热点,该类方法克服基于 RGB 视频的行为识别方法的缺陷,对背景、光照、人体外观形变等具有较好的鲁棒性[15-16]。但是在石化场景中,打手机和抽烟等危险行为属于人-物交互行为,仅仅依靠人体骨骼信息难以区分相似动作(比如打手机和摸耳朵),丢失某些必要的物体和语义信息[17-18]。

针对以上问题,本文将人体行为识别与目标检测相结合,提出基于机器视觉的石化场景人员危险行为识别方法。针对基于骨骼点的人体行为识别方法丢失背景和语义信息的问题,引入目标检测算法,获得感兴趣物体的类别和位置信息;根据人-物交互思想,确定人-物交互关系;最后对各部分结果进行决策融合,得到最终行为识别结果,提升打手机、抽烟等人-物交互行为的识别准确度。

2 基于机器视觉的石化场景人员危险行为识别方法

2.1 方法概述

所提基于机器视觉的石化场景人员危险行为识别方法流程如图 1 所示。对于由摄像头捕获的视频流,先用人体姿态估计算法获取人体骨骼信息,进而利用基于骨骼点的人体行为识别方法获得人员的初始行为类别;利用目标检测算法检测感兴趣物体,获得其类别和位置信息;然后通过判断人与物体的空间位置关系来确定人-物交互关系;最后对初始行为类别、物体信息、人-物交互关系进行决策融合,得到最终的行为识别结果。

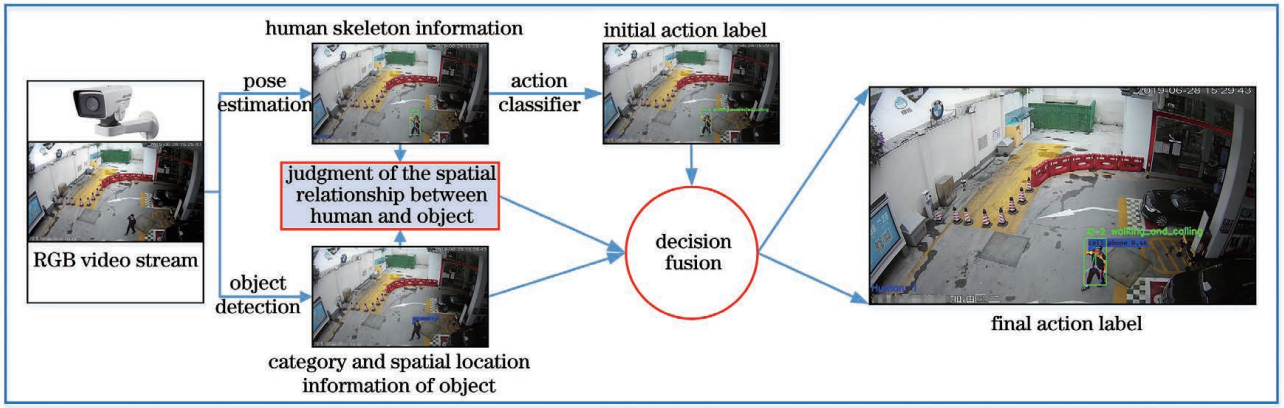

图 1　基于机器视觉的石化场景人员危险行为识别方法

Fig. 1　Method for identifying dangerous actions of personnel in petrochemical scenes based on machine vision

2.2　基于骨骼点的人体行为识别

基于骨骼点的人体行为识别流程如图 2 所示，先由人体姿态估计算法 OpenPose[19] 获取视频中的人体骨骼信息，进而利用行为识别模块得到人员的初始行为类别。

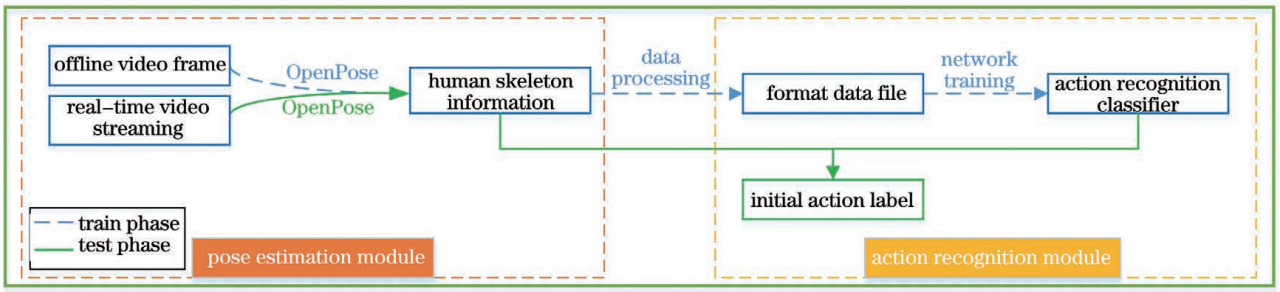

图 2　基于骨骼点的人体行为识别流程

Fig. 2　Process of skeleton-based human action recognition

2.2.1　人体姿态估计

人体姿态估计是对图片中人体的骨骼点（如头部、手部等）进行位置估计。OpenPose 采用 Bottom-Up 方法，是首个基于深度学习的实时性姿态识别模型。Bottom-Up 方法的网络推理速度不会随着场景中人的数量增多而变慢，适用于有实时性要求或多人的场景。综合考虑速度和精度，选用速度快、精度高的 OpenPose 作为人体姿态估计算法。

OpenPose 通过 VGG-19 卷积神经网络结构提取图像特征，然后将图像特征分为两个分支分别进行 T 个阶段的迭代训练。分支 1 输出人体骨骼点的置信度集合 $S=(S_1,S_2,\cdots,S_j,\cdots,S_J)$，$S_j$ 表示人体的第 j 个骨骼点的置信度，J 为骨骼点总数；分支 2 输出其亲和度集合 $L=(L_1,L_2,\cdots,L_c,\cdots,L_C)$，$L_c$ 表示人体骨骼点中第 c 个连接的亲和度，C 为连接总数。检测出骨骼点之后，再基于亲和度集合 L 对其进行高准确性聚类，最终识别出人体姿态。

训练阶段 t 的置信度（用 S 表示）分支的损失函数 f_S^t 定义为

$$f_S^t=\sum_{j=1}^{J}\sum_{\boldsymbol{p}}W(\boldsymbol{p})\cdot\left\|S_j^t(\boldsymbol{p})-S_j^*(\boldsymbol{p})\right\|_2^2,\tag{1}$$

式中：W 是二元掩码，当图像中 $\boldsymbol{p}$ 点的骨骼点标注信息丢失时，$W(\boldsymbol{p})$ 为 0；$S_j^*(\boldsymbol{p})$ 表示置信度 S 的真实值。

训练阶段 t 的亲和度（用 L 表示）分支的损失函数 f_L^t 定义为

$$f_L^t=\sum_{c=1}^{C}\sum_{\boldsymbol{p}}W(\boldsymbol{p})\cdot\left\|L_c^t(\boldsymbol{p})-L_c^*(\boldsymbol{p})\right\|_2^2,\tag{2}$$

式中：$L_c^*(\boldsymbol{p})$ 表示亲和度 L 的真实值。

可由 f_S^t 和 f_L^t 计算得到全局损失函数 f，表达式为

$$f=\sum_{t=1}^{T}f_S^t+f_L^t。\tag{3}$$

2.2.2　人体行为识别

人体行为识别技术利用模式识别、深度学习等方法自动分析图像中人体的行为特征，自动识别其

行为类别。考虑到实际场景需求，使用图 3 所示的基于深度学习的人体行为识别网络（https://github.com/LZQthePlane/Online-Realtime-Action-Recognition-based-on-OpenPose）。该网络对输入的人体骨骼数据进行特征提取，进而完成行为识别的任务。

如图 3 所示，4 个 block 都具有相似的（神经元个数不同）组成结构，以 block 1 为例进行说明。首先，输入层后加入全连接层（FC）可以更好地提取特征进而提高识别精度；然后，FC 层后使用修正线性单元（ReLU）作为激活函数；最后，激活层后面加入批量归一化层（BN）来降低网络对初始化权重的不敏感性，从而加快网络训练速度，使其快速收敛。在 block 4 后面再次加入 FC，通过整合前端具有类别区分性的归一化特征信息以完成分类任务，使用 softmax 作为行为分类器。

使用交叉熵损失函数，交叉熵 $H(p,q)$ 定义为

$$H(p,q)=-\sum_{i=1}^{I}p(x_i)\log\left[q(x_i)\right], \tag{4}$$

式中：I 表示样本总数；$p(x_i)$ 表示真实概率分布；$q(x_i)$ 表示预测概率分布。模型中的参数 w 使用梯度下降法进行更新，表达式为

$$w\leftarrow w-\alpha\times\frac{\partial l_{\text{loss}}}{\partial w}, \tag{5}$$

式中：α 是学习率。当 l_{loss} 低于某个阈值或迭代次数达到设定值时，迭代终止。

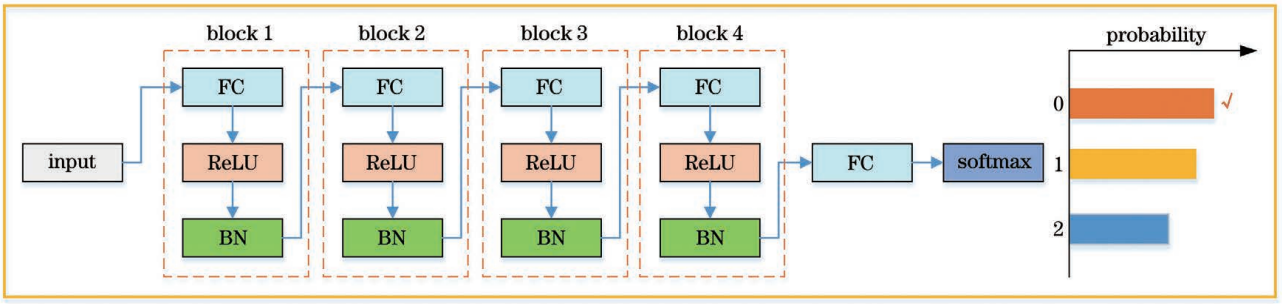

图 3　基于骨骼点的人体行为识别网络

Fig. 3　Skeleton-based human action recognition network

Adam 优化器通过引入梯度的二阶矩估计进行参数优化[20]，因此选用 Adam 优化器进行参数优化。

2.3　目标检测

目标检测是对图像中的目标进行分类并确定其位置的过程，是行为识别等其他高级计算机视觉任务的基础。

YOLOv3 算法[21]在特征提取网络中采用 Darknet-53，引入残差网络结构，提高特征提取能力。YOLOv3 算法借鉴 FPN[22]多尺度特征融合思路，输出 3 个尺度的特征图，改善小目标检测效果，在不影响精度的前提下有很大速度优势，成为研究热点[23-24]。

YOLOv3 的损失函数主要分为 3 部分[25]：坐标损失 l_1、类别损失 l_2 以及置信度损失 l_3。坐标损失定义为

$$l_1=\frac{1}{2}\sum_{n=1}^{N}\lambda_{\text{obj}}\times(2-w_{\text{truth}}\times h_{\text{truth}})\times\sum_{r\in(x,y,w,h)}(r_{\text{truth}}-r_{\text{predict}})^2, \tag{6}$$

式中：N 为训练样本个数；r 表示 Bounding Box 的四个属性值 (x,y,w,h)，(x,y) 为初始点坐标，w 和 h 分别表示宽和高；λ_{obj} 为二值变量，当某个预测单元格存在物体时，λ_{obj} 为 1，否则为 0。类别损失定义为

$$l_2=\frac{1}{2}\sum_{n=1}^{N}\lambda_{\text{obj}}\times\sum_{k=0}^{K-1}\left[\begin{matrix}(k==c_{\text{class, truth}})\\ ?\ 1:0-c_{\text{class, predict}}\end{matrix}\right]^2, \tag{7}$$

式中：k 表示物体类别，取值为 0 到 $K-1$，K 为类别总数；$c_{\text{class, truth}}$ 为物体类别的真实值；$c_{\text{class, predict}}$ 为物体类别的预测值。置信度损失定义为

$$l_3=(c_{\text{truth}}-c_{\text{predict}})^2, \tag{8}$$

式中：c_{truth} 为置信度的真实值；c_{predict} 为置信度的预测值。总损失函数 l 可表示为

$$l=l_1+l_2+l_3。 \tag{9}$$

综合考虑速度、精度以及对小目标的检测能力，选用 YOLOv3 算法对感兴趣物体进行检测。针对抽烟和打手机行为，定义“cell phone”和“cigarette”两种物体类别，初始化类别标签为“nothing”。

2.4　人与物体空间位置关系判断

以打手机行为为例，给出了人与物体空间位置关系的判断方法。

图 4(a)和图 4(b)所示为真实的打手机行为。但是如图 4(c)和图 4(d)所示，尽管同时检测到打手

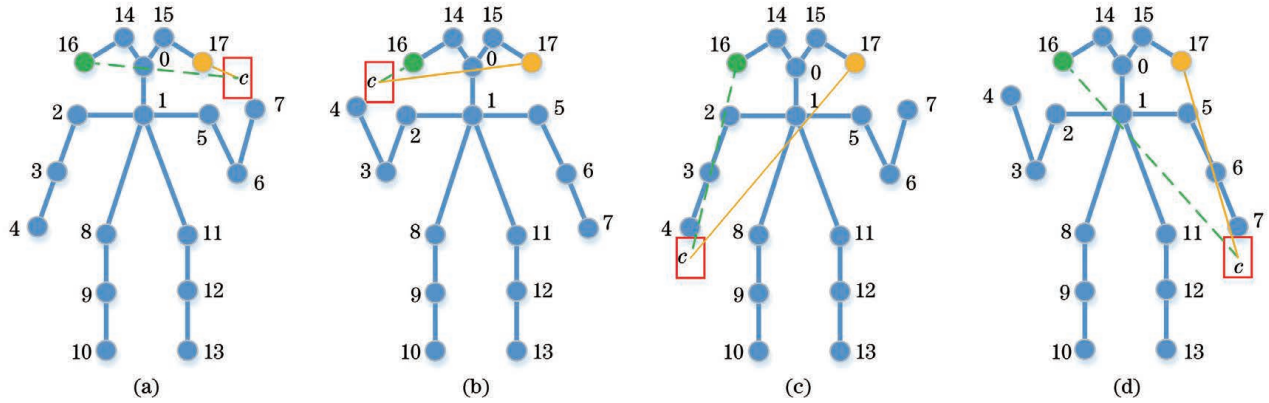

图 4　人与手机的空间位置关系判断方法

Fig. 4　Judgment method of the spatial position relationship between human and cell phone

机的姿态和手机(方框),但显然行为人此时并没有在打手机。因此,只根据打手机时的姿态和检测到的手机还不足以确定某人是否有打手机的行为,还需要判断人与手机的交互关系。

为了解决上述问题,分别计算图 4 中方框的中心点 c 与人体左耳(标号 17)和人体右耳(标号 16)的欧氏距离 d_1(细实线)和 d_2(虚线)。如图 4(a)、(b)所示,只有 d_1、d_2 与距离阈值 $d_{\text{threshold}}$ 的关系满足

$$d_1 < d_{\text{threshold}} \text{ or } d_2 < d_{\text{threshold}}, \tag{10}$$

该行为才能最终被判定为打手机。经充分实验,最终将距离阈值 $d_{\text{threshold}}$ 设定为 0.06。

为了更直观地说明人与手机的位置关系判断的必要性和有效性,图 5 给出了部分可视化结果。其中,图 5(1, 2, 3, 4)-a 是仅仅基于人体骨骼信息的行为识别结果,图 5(1, 2, 3, 4)-b 是对人体骨骼信息、手机信息、人与手机位置关系进行融合后的行为识别结果。

以图 5(2-b)为例,标注的红色(蓝色)直线表示手机的中心坐标到人体左耳(右耳)坐标的欧氏距离 d_1(d_2),显然 d_1 和 d_2 都大于设定的距离阈值 $d_{\text{threshold}}$。可以观察到,图 5(2-b)已将图 5(2-a)的错误结果“walking_and_calling”修正为“walking_suspected_calling”,同理可分析图 5(1, 3, 4)-b。

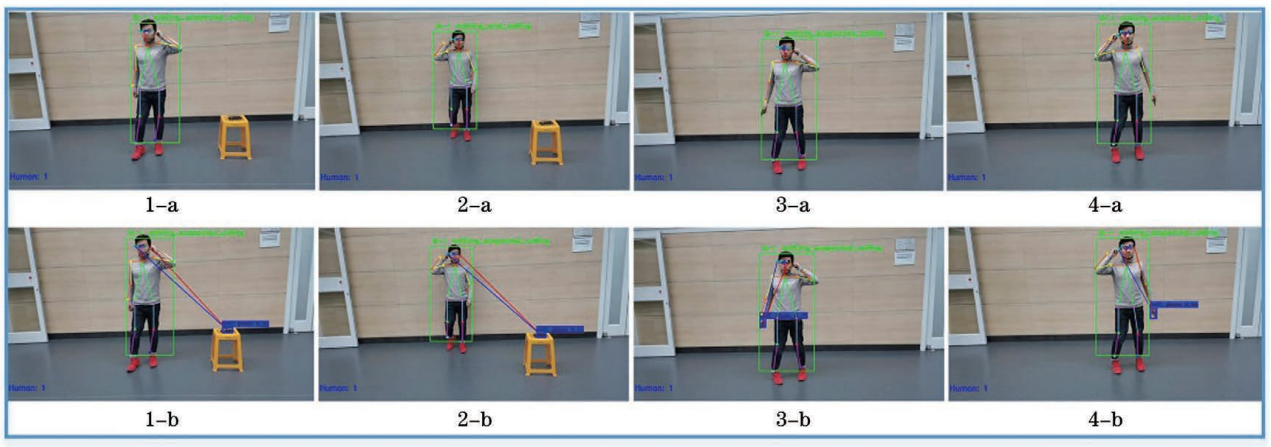

图 5　人与手机的空间位置关系判断结果

Fig. 5　Judgment result of the spatial position relationship between human and cell phone

2.5　决策融合

根据一定的准则,采用决策融合模块对人员的初始行为类别、物体信息、人-物交互关系进行融合判断,获得最终的人体行为类别。以打手机行为为例,决策融合策略可以表述如下。

1) 当初始行为的类别编号为 0 时,表明初始行为类别无需修正。

2) 当初始行为类别编号为 1 或 2 时,目标检测的结果才会影响最终的行为类别:

a. 如果物体类别为“cell phone”,且 d_1 和 d_2 满足设定的阈值 $d_{\text{threshold}}$,则将初始行为类别修正为“walking_and_calling”,否则修正为“walking_

suspected_calling”，从而降低打手机行为的漏识别率；

b. 如果物体类别为“nothing”（表明图片中未检测出“cell phone”），则将最终的行为类别确定为“walking_suspected_calling”，以此降低打手机行为的误识别率。

3　实验结果与分析

3.1　打手机行为识别

3.1.1　数据集

研究的石化场景人员危险行为识别属于具体场景应用，通用行为识别数据集[26-27]不适合用来验证所提方法。为验证所提方法的有效性，构建了打手机行为数据集，图 6 为该数据集的部分示例。该数据集来源于三部分：模拟石化场景（贴有“禁止打手机”标志）、徐州市某镇某加油站实地拍摄、深圳市某区某加油站实际监控视频。考虑到摄像头视角、尺度大小、类间差异等因素对行为识别结果的影响，数据集中包含多个行为人、多种角度和尺度的视频帧，以此增强模型的泛化性。视频的分辨率为 1280×720，包含如表 1 所示的 3 种初始行为类别。

表 1　初始行为类别

Table 1　Initial action categories

Action No.	Action label
0	walking
1	walking_suspected_calling
2	walking_and_calling

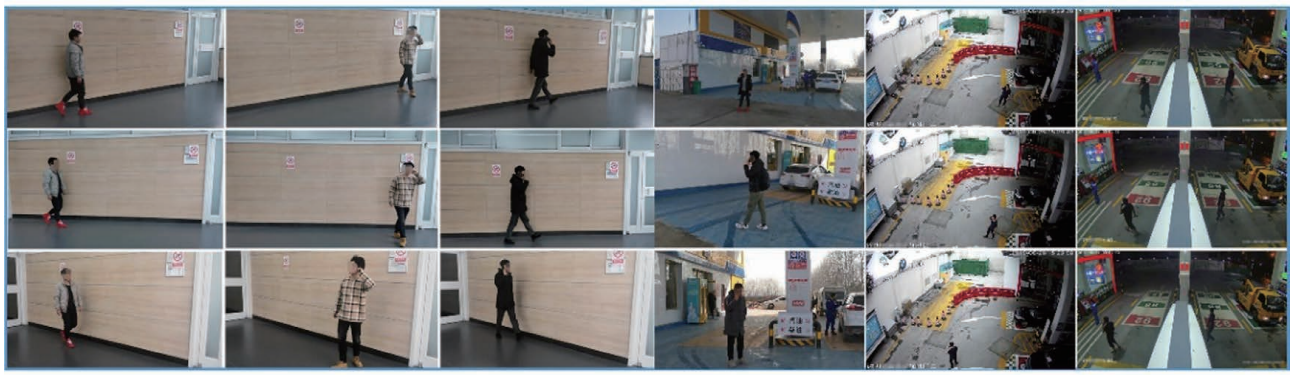

图 6　打手机行为数据集示例

Fig. 6　Example of cell phone call action data set

按照一定比例，将数据集随机划分为训练集、交叉验证集和测试集。训练集和交叉验证集共 45 个片段，每个片段 300 帧；测试集共 3914 帧。其中，行为 0 有 519 帧，行为 1 有 1763 帧，行为 2 有 1632 帧。

3.1.2　实验步骤

1）使用 OpenPose 算法提取人体骨骼点信息，构建人体行为识别网络的训练集。

2）使用步骤 1）构建的数据集对图 3 所示的网络进行训练，得到人体行为分类器。将学习率设为 0.0001，batch_size 设为 32，epoch 设为 100。训练和交叉验证过程的损失值和准确率如图 7 所示，测试结果如图 8 所示。

3）搜集、标定图片，制作目标检测数据集，训练 YOLOv3 网络，得到目标检测模型。

4）使用测试集对所提方法进行测试，得出实验结果。

3.1.3　实验结果

测试集共 3914 帧，行为 0 有 519 帧，行为 1 有 1763 帧，行为 2 有 1632 帧。图 9 和图 10 分别为加入检测模块前、后的识别结果的混淆矩阵。

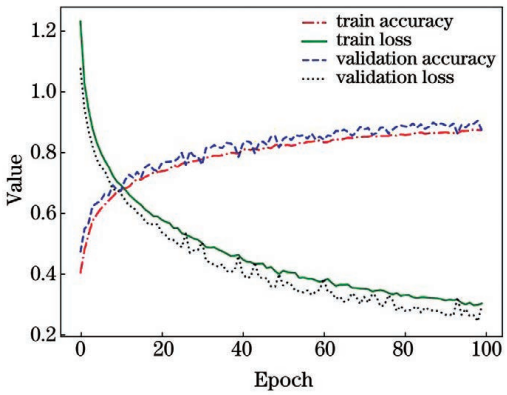

图 7　损失值和准确率

Fig. 7　Loss value and accuracy rate

采用视频帧的识别精确率作为评价指标，可以计算出每种行为的识别精确率和召回率。表 2 和表 3 分别为加入目标检测模块前、后的行为识别结果。

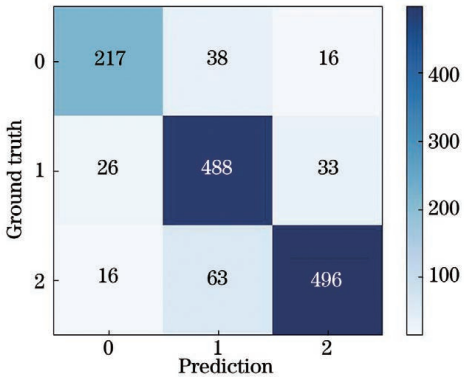

图 8　测试结果

Fig. 8　Test result

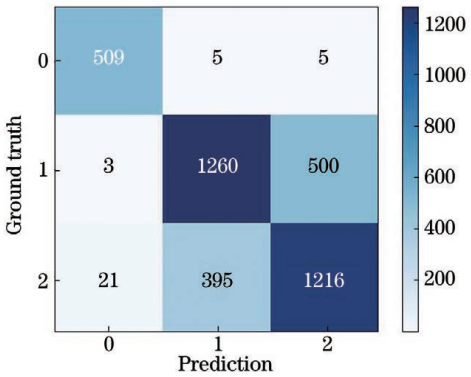

图 9　混淆矩阵(仅骨骼信息)

Fig. 9　Confusion matrix (only skeleton information)

为了更直观地表明所提方法的有效性,图 11 和图 12 给出了行为人背对或侧对镜头、尺度变化时的实验结果。

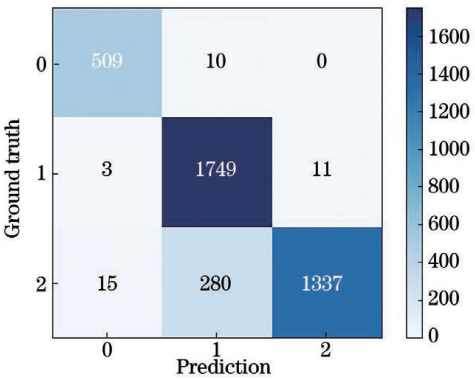

图 10　混淆矩阵(加入目标检测)

Fig. 10　Confusion matrix (adding object detection)

表 2　行为识别结果(加入目标检测模块前)

Table 2　Action recognition results (before adding the object detection module)

Action label	Recall /%	Precision /%
walking	98.07	95.50
walking_suspected_calling	71.47	75.90
walking_and_calling	74.51	70.66

表 3　行为识别结果(加入目标检测模块后)

Table 3　Action recognition results (after adding the object detection module)

Action label	Recall /%	Precision /%
walking	98.07	96.58
walking_suspected_calling	99.21	85.78
walking_and_calling	81.92	99.18

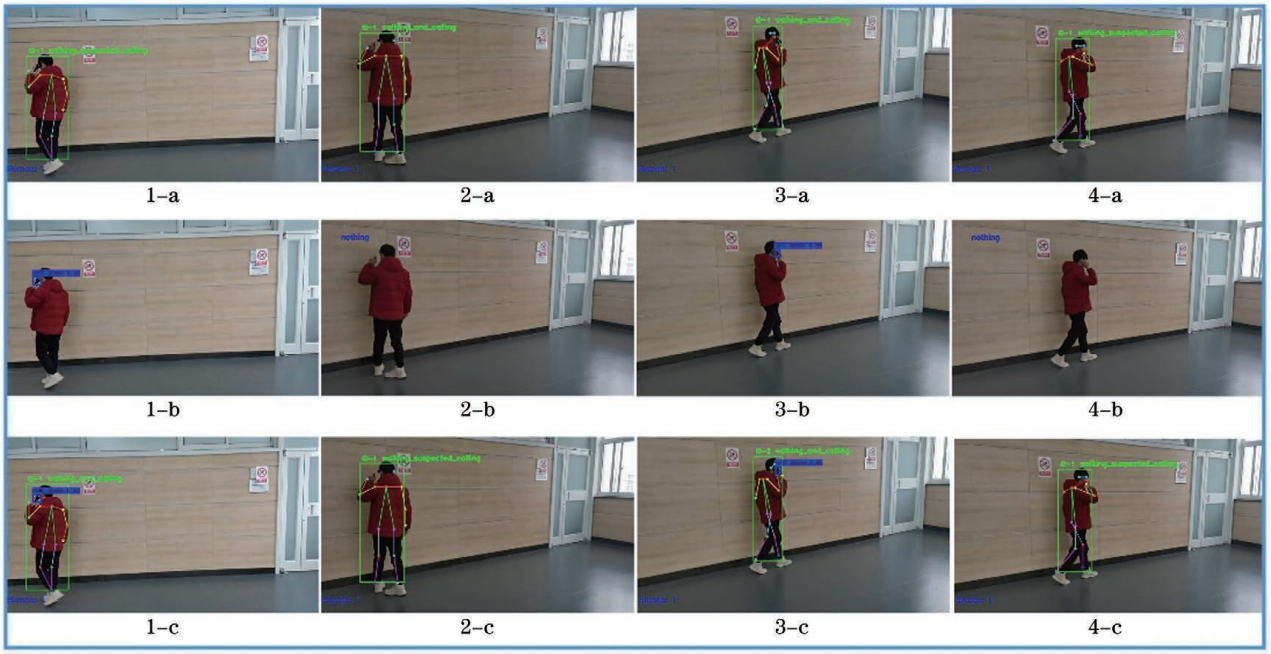

图 11　模拟石化场景的实验结果

Fig. 11　Experimental results of simulated petrochemical scene

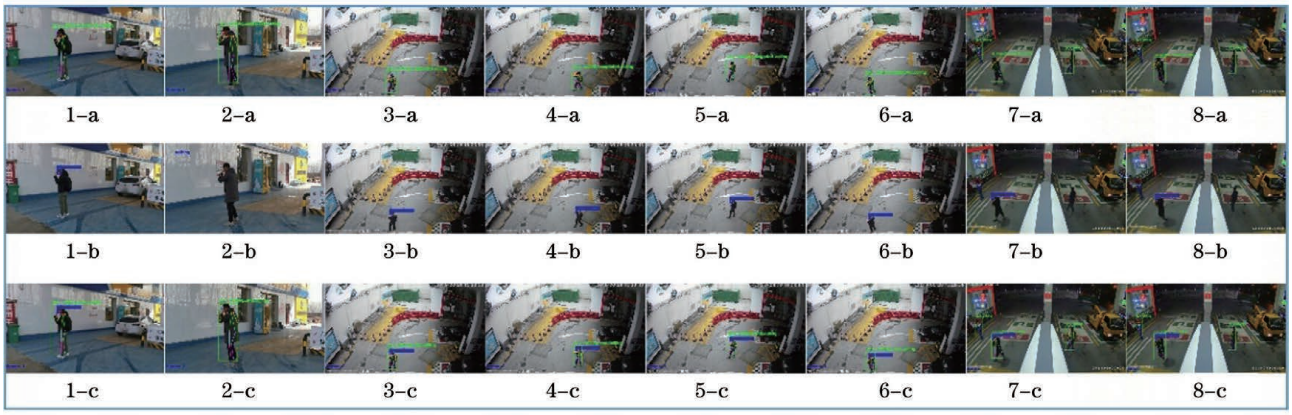

图 12　实际石化场景的实验结果

Fig. 12　Experimental results of actual petrochemical scene

图 11 为加油站场景模拟的实验结果，图 11(1，2，3，4)-a 为初始行为类别，图 11(1，2，3，4)-b 为相应的目标检测结果，图 11(1，2，3，4)-c 为最终的行为识别结果。图 11(1-a)的真实类别为“walking_and_calling”，未加入目标检测模块前被误识别成“walking_suspected_calling”(漏警)；加入目标检测模块并经过决策融合后，误识别得到修正，如图 11(1-c)所示。图 11(2-a)的真实类别为“walking_suspected_calling”，初始被误识别成“walking_and_calling”(虚警)；后来误识别得到修正，如图 11(2-c)所示。由于图 11(3-a)和图 11(4-a)的初始识别结果准确，因此加入目标检测模块并经过决策融合后识别结果并不发生改变，如图 11(3-c)和图 11(4-c)所示。

图 12 为实际石化场景的实验结果。其中，前两列为徐州市某镇某加油站实地拍摄，剩余列是深圳市某区某加油站的监控视频。经分析可得：在拍摄角度、尺度等发生变化时，所提方法可以对人-物交互的危险行为进行准确识别，进一步证明了所提方法的有效性。

3.2　抽烟行为识别

所提方法不仅能提升石化场景下打手机行为的识别准确率，并且适用于该场景下抽烟等其他危险行为。采用和 3.1.1 节相同的数据集构建方法和 3.1.2 节相同的实验步骤，继续对石化场景下的抽烟行为进行实验研究，结果表明所提方法具有良好的扩展性。

图 13 给出了抽烟行为数据集示例，图 14 为抽烟行为的识别结果，分析方法同图 11。

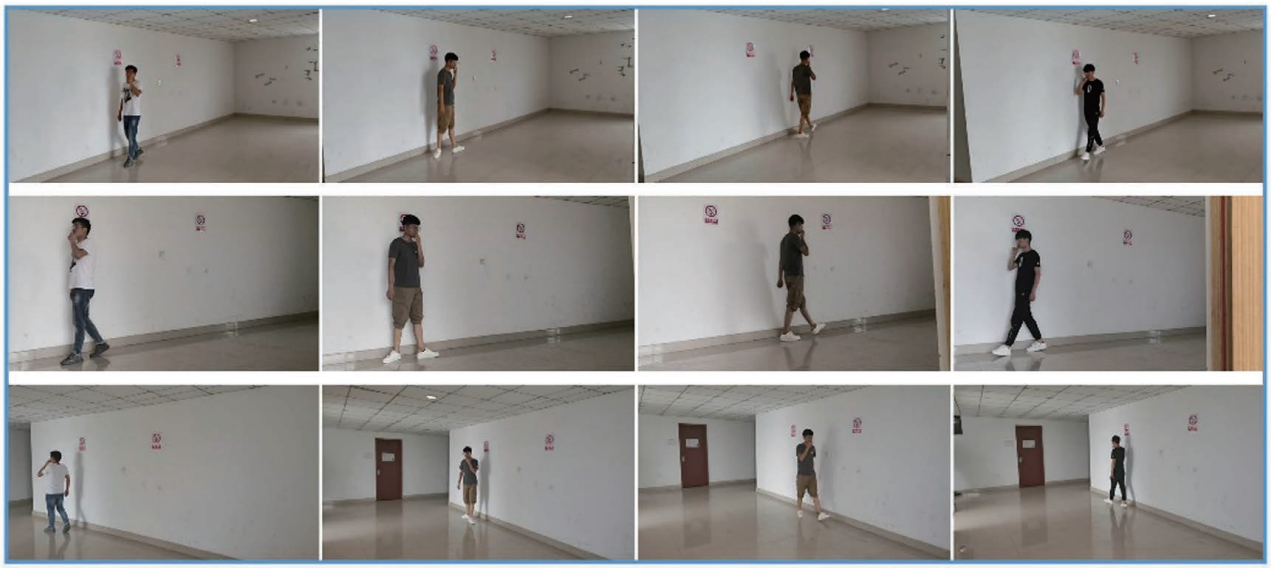

图 13　抽烟行为数据集示例

Fig. 13　Examples of smoking action data set

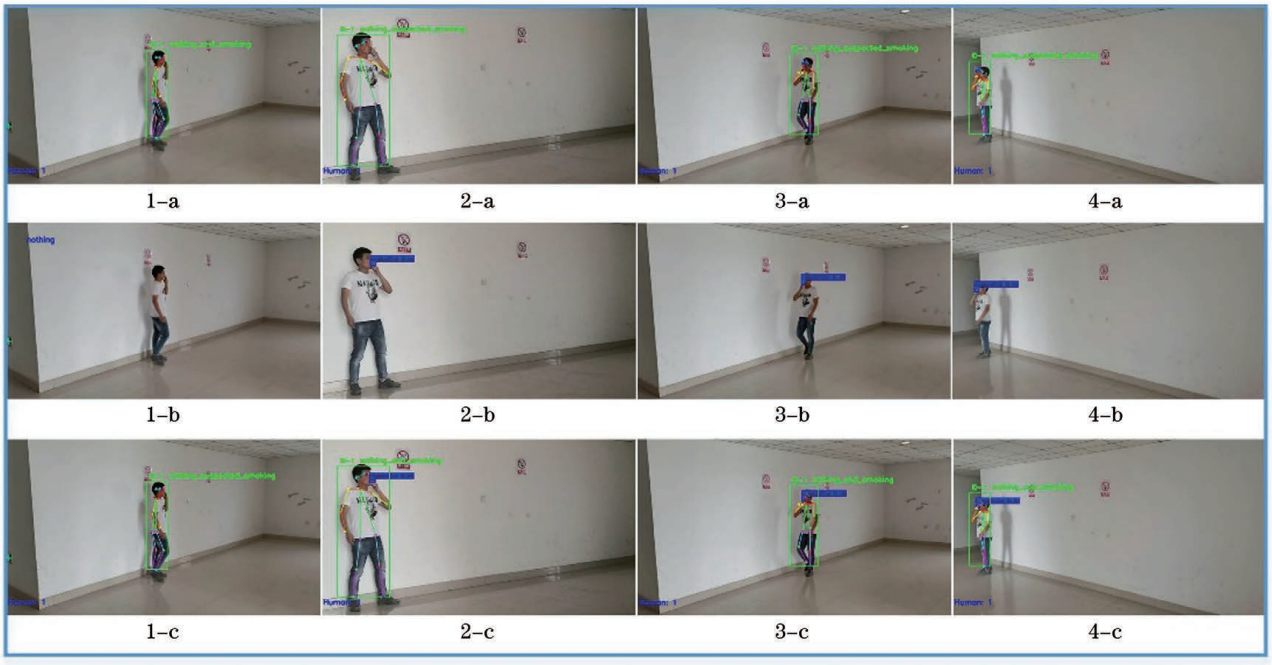

图 14　抽烟行为识别结果

Fig. 14　Results of smoking action recognition

4　结　　论

将目标检测与基于骨骼点的人体行为识别相结合，提出基于机器视觉的石化场景人员危险行为识别方法。采用 OpenPose 算法进行人体姿态估计，继而利用基于骨骼点的人体行为识别方法获取人员的初始行为类别，克服了基于 RGB 视频的行为识别方法的鲁棒性差、速度慢等问题；使用 YOLOv3 算法获得手机和香烟的类别和位置信息，解决了传统行为识别方法丢失背景和语义信息的问题；通过人与物体之间的空间位置关系来表征人-物交互关系；最后对各模块结果进行决策融合，提升人-物交互行为的识别准确度。在自建数据集上的实验结果表明，所提方法的识别精度优于基于骨骼点的人体行为识别方法，可以对石化场景下打手机、抽烟等属于人-物交互的危险行为进行准确识别。但是，当目标检测算法未能检测出手机、香烟等小物体时，会影响识别效果。因此，如何改善目标检测算法对小目标的检测能力、如何对其与基于骨骼点的人体行为识别方法进行深度融合，将是下一步研究方向。

致谢　感谢纪贺、李朋峰、吴佳佳、杨俊秋、张德保、周亚旭同学对本课题研究的帮助。

参　考　文　献

[1]　Huang K Q, Chen X T, Kang Y F, et al. Intelligent visual surveillance: a review[J]. Chinese Journal of Computers, 2015, 38(6): 1093-1118.
黄凯奇，陈晓棠，康运锋，等. 智能视频监控技术综述[J]. 计算机学报，2015，38(6)：1093-1118.

[2]　Yang J, Simon S, Gayathri C, et al. Detecting driver phone use leveraging car speakers[C]//Proceedings of the 17th Annual International Conference on Mobile Computing and Networking, September 19-23, 2011, Las Vegas, Nevada, USA. New York: ACM, 2011: 97-108.

[3]　Rodríguez-Ascariz J M, Boquete L, Cantos J, et al. Automatic system for detecting driver use of mobile phones[J]. Transportation Research Part C: Emerging Technologies, 2011, 19(4): 673-681.

[4]　Berri R A, Silva A G, Parpinelli R S, et al. A pattern recognition system for detecting use of mobile phones while driving[C]//2014 International Conference on Computer Vision Theory and Applications (VISAPP), January 5-8, 2014, Lisbon, Portugal. New York: IEEE Press, 2014: 411-418.

[5]　le T H N, Zheng Y T, Zhu C C, et al. Multiple scale faster-RCNN approach to driver's cell-phone usage and hands on steering wheel detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 26-July 1, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 46-53.

[6]　Seshadri K, Juefei-Xu F, Pal D K, et al. Driver cell phone usage detection on strategic highway research

program (SHRP2) face view videos[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 35-43.

[7] Kohl D, Eberheim A, Schieberle P. Detection mechanisms of smoke compounds on homogenous semiconductor sensor films[J]. Thin Solid Films, 2005, 490(1): 1-6.

[8] Liu B J, Alvarez-Ossa D, Kherani N P, et al. Gamma-free smoke and particle detector using tritiated foils[J]. IEEE Sensors Journal, 2007, 7(6): 917-918.

[9] Millan-Garcia L, Sanchez-Perez G, Nakano M, et al. An early fire detection algorithm using IP cameras[J]. Sensors (Basel, Switzerland), 2012, 12(5): 5670-5686.

[10] Yuan F N. A fast accumulative motion orientation model based on integral image for video smoke detection[J]. Pattern Recognition Letters, 2008, 29(7): 925-932.

[11] Li P, Zhang Y. Video smoke detection based on Gaussian mixture model and convolutional neural network[J]. Laser & Optoelectronics Progress, 2019, 56(21): 211502.
李鹏, 张炎. 基于高斯混合模型和卷积神经网络的视频烟雾检测[J]. 激光与光电子学进展, 2019, 56(21): 211502.

[12] Davis J. Visual gesture recognition[J]. IEEE Proceedings-Vision, Image, and Signal Processing, 1994, 141(2): 101-106.

[13] Hu J F, Wang X H, Zheng W S, et al. RGB-D action recognition: recent advances and future perspectives[J]. Acta Automatica Sinica, 2019, 45(5): 829-840.
胡建芳, 王熊辉, 郑伟诗, 等. RGB-D行为识别研究进展及展望[J]. 自动化学报, 2019, 45(5): 829-840.

[14] Guo F Z, Kong J, Jiang M. Action recognition based on adaptive fusion of RGB and skeleton features[J]. Laser & Optoelectronics Progress, 2020, 57(20): 201506.
郭伏正, 孔军, 蒋敏. 自适应融合RGB和骨骼特征的行为识别[J]. 激光与光电子学进展, 2020, 57(20): 201506.

[15] Li C, Zhong Q Y, Xie D, et al. Skeleton-based action recognition with convolutional neural networks[C]//2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), July 10-14, 2017, Hong Kong, China. New York: IEEE Press, 2017: 597-600.

[16] Yan S J, Xiong Y J, Lin D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//2018 the Association for the Advancement of Artificial Intelligence, February 2-7, 2018, New Orleans, Louisiana, USA. Menlo Park: AAAI Press, 2018: 7444-7452.

[17] Rosenfeld A, Ullman S. Hand-object interaction and precise localization in transitive action recognition[C]//2016 13th Conference on Computer and Robot Vision (CRV), June 1-3, 2016, Victoria, BC, Canada. New York: IEEE Press, 2016: 148-155.

[18] Kim S, Yun K, Park J, et al. Skeleton-based action recognition of people handling objects[C]//2019 IEEE Winter Conference on Applications of Computer Vision (WACV), January 7-11, 2019, Waikoloa, HI, USA. New York: IEEE Press, 2019: 61-70.

[19] Cao Z, Simon T, Wei S H, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1302-1310.

[20] Tian W H, Zeng K M, Mo Z Q, et al. Recognition of unsafe driving behaviors based on convolutional neural network[J]. Journal of University of Electronic Science and Technology of China, 2019, 48(3): 381-387.
田文洪, 曾柯铭, 莫中勤, 等. 基于卷积神经网络的驾驶员不安全行为识别[J]. 电子科技大学学报, 2019, 48(3): 381-387.

[21] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2020-12-14]. https://arxiv.org/abs/1804.02767.

[22] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[23] Sun Y C, Pan S G, Zhao T, et al. Traffic light detection based on optimized YOLOv3 algorithm[J]. Acta Optica Sinica, 2020, 40(12): 1215001.
孙迎春, 潘树国, 赵涛, 等. 基于优化YOLOv3算法的交通灯检测[J]. 光学学报, 2020, 40(12): 1215001.

[24] Zhao Q, Li B Q, Li T W. Target detection algorithm based on improved YOLO v3[J]. Laser & Optoelectronics Progress, 2020, 57(12): 121502.
赵琼, 李宝清, 李唐薇. 基于改进YOLO v3的目标检测算法[J]. 激光与光电子学进展, 2020, 57(12): 121502.

[25] Lyu S, Cai X, Feng R. YOLOv3 network based on improved loss function[J]. Computer Systems & Applications, 2019, 28(2): 1-7.
吕铄, 蔡烜, 冯瑞. 基于改进损失函数的 YOLOv3 网络[J]. 计算机系统应用, 2019, 28(2): 1-7.

[26] Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset[EB/OL]. (2017-05-19) [2020-12-14]. https://arxiv.org/abs/1705.06950.

[27] Shahroudy A, Liu J, Ng T T, et al. NTU RGB+D: a large scale dataset for 3D human activity analysis[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 1010-1019.