一种端到端的考场多目标行为识别算法

姚　捃,郭志林
(成都理工大学 工程技术学院,四川 乐山 614000)

摘　要:人工监考存在监考人员容易疲惫、监考行为缺乏客观的执行准则、违规行为证据无法留存等问题,因此越来越多的高校建设了智能化教室,并在教室开始实施利用行为识别进行自动化的监考任务,以期在监考工作中解放人工劳动的同时提供公平公正客观的监考程序。在实际考场监控的边缘设备中利用 TSN 双流、3DCNN 等结合时空特征的网络很难实现实时的、相对准确的监控任务。提出一种端到端的考场多目标行为识别算法。相对于以提取空间、时序特征并进行融合为主流思想的行为识别算法,利用视频帧以多目标检测和多目标行为识别相结合的行为识别算法在考场环境中更加快速准确。算法借助了多标签学习、注意力机制和特征金字塔等策略来改进任务,同时利用迁移学习对本地采集的考场行为视频数据集进行再训练,得到最终的考场行为识别模型,结果表明达到了主流数据集中上水平,并在考场环境中具有良好的高效性与准确性。

关键词:多目标检测;端到端;行为识别;多标签学习;特征金字塔
中图分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1673-629X(2022)09-0174-06
doi:10.3969/j.issn.1673-629X.2022.09.027

An End-to-end Multi-objective Behavior Recognition Algorithm for Examination Room

YAO Jun,GUO Zhi-lin
(School of Engineering & Technique,Chengdu University of Technology,Leshan 614000,China)

Abstract: There are some problems in manual invigilation, such as invigilator fatigue, invigilator behavior lack of objective implementation criteria, violation evidence cannot be retained. Therefore, more and more colleges and universities have built intelligent classrooms and began to implement the task of automatic invigilation by using behavior recognition in the classrooms, so as to liberate manual labor in the invigilation work and provide a fair and objective invigilation procedures. In the edge equipment of actual examination room monitoring, it is difficult to realize real-time and relatively accurate monitoring tasks by using TSN dual stream, 3DCNN and other networks combined with temporal and spatial characteristics. An end-to-end multi-objective behavior recognition algorithm for examination room is proposed. Compared with the behavior recognition algorithm which takes extracting spatial and temporal features and fusing as the mainstream idea, the behavior recognition algorithm based on the combination of multi-target detection and multi-target behavior recognition in video frames is faster and more accurate in the examination room environment. The algorithm improves the task with the help of multi label learning, attention mechanism and feature pyramid. At the same time, the locally collected examination room behavior video data set is retrained by transfer learning to obtain the final examination room behavior recognition model. The results show that it has reached the upper level of the mainstream data set, and has high efficiency and accuracy in the examination room environment.

Key words: multi target detection; end-to-end; behavior recognition; multi label learning; feature pyramid

0　引　言

行为识别是人工智能和计算机视觉领域最热门的研究话题之一,应用于安防安检、火灾预警、考场监控、婴幼儿看护、病床报警等重要场合。

早期的行为识别基于手工提取特征的方法,利用特征描述符包括轨迹特征、梯度直方图特征(HOG)、光流直方图特征(HOF)、混合动态纹理、光流场、人体关节点等构建特征空间,进而使用机器学习的方法进行行为的识别。张恒鑫等[1]借助 OpenPose 得到人体区域中关节点的二维坐标构建骨架模型,提取基于关

收稿日期:2021-10-19　　**修回日期**:2022-02-23
基金项目:教育部产学合作协同育人项目(201802047024);成都理工大学工程技术学院青年科学基金(C122019008)
作者简介:姚　捃(1982-),男,硕士,讲师,研究方向为计算机视觉。

节向量的多种动作时空特征,采用 kNN 模型进行行为识别,在 MIT 及 Weizmann 数据集上取得了不错的效果。

由于视频场景的复杂性、光线变化、人体遮挡等问题,手工提取特征的方法具有一定的局限性,渐渐被深度学习方法所替代,如卷积神经网络、快慢信息网络、TSN 双流网络、3D 卷积神经网络、双流与 LSTM 结合的混合网络等。这些网络利用深度学习强大的特征提取能力,从各个角度提取出行为识别所需要的人体外观和人体运动的高级时空特征(如光流特征、3D 时空特征),能更有效地进行行为识别。

何嘉宇等[2]构建了特征金字塔层次结构以增强网络检测不同持续时长的行为片段的能力,在公开数据集上获得了先进的识别性能。董琪琪、王昊飞等[3-6]利用 3D 卷积神经网络,有效提取行为识别所需要的时序信息与空间信息,展现了较好的性能。窦雪婷、陈京荣等[7-10]利用 LSTM 长短记忆网络,对数据之间的联系进行有选择的记忆与遗忘,仅保留对识别结果有益的数据,取得了良好的识别率。

基于深度学习提取时空特征的行为识别算法存在硬件需求过高、时效性不足、复杂场景准确率偏低等问题。叶黎伟、窦刚等[11-15]利用 YOLOv3 网络在没有提取时序信息的基础上,通过改进的网络进行了行为识别,解决了推理速度过慢的问题,取得了不错的效果。

Gharahdaghi、Gutoski 等[16-19]利用 LSTM、TSN、3DCNN 等提取高级时空信息进行行为识别,在可穿戴设备、教育教学、电子广告、物联网等领域取得了良好的应用。

在很多应用场景下,不能实现实时识别是行为识别算法并未得到有效应用的主要原因。

因此在考场的环境下,提出一种利用视频帧以多目标检测和多目标行为判别结合的行为识别算法,算法基于两点前提:

(1)在考场这种特殊场景下,没有明显光照变化、场景单一、摄像头角度固定、人体没有明显遮挡,因而可以采集到相对高清的视频图像。

(2)建立目标检测与行为判别两个本地数据集,利用非常成熟的 YOLOv3 算法,可以实现实时的多目标检测;同时将行为判别任务作为 YOLOv3 的分类标签输出,完成端到端的多目标行为识别。

实验结果表明,算法在特定场景具有良好的鲁棒性、高效性和准确率。

1　算法设计

YOLOv4 和 YOLOv5 虽然是在 YOLOv3 基础上进一步改进,但都是在细枝末节上进行的优化,反而丢失了 YOLOv3 在工业界的普遍适用性。所以总体思路是利用 YOLOv3 进行多目标检测,同时将行为判别作为各个目标的分类标签,并根据实际任务对输出网络加以修正。

1.1　数据准备

要实现考场任务的多目标检测,并将各目标行为作为分类标签在 YOLOv3 同时输出,必须提前设计好本地任务的目标检测和行为判别数据集。

利用部署在考场的高清摄像头采集本地视频数据,并进行视频筛选滤除长时的正常视频。对视频数据中典型的正常行为与违规行为进行人体框标注,打上对应行为标签。考场主要行为标签如表 1 所示。

表 1　行为标签

人体框	行为	判定
有	出现纸条	违规
有	出现书本	违规
有	出现手机	违规
有	扭头	违规
有	斜视	违规
有	说话	违规
有	埋头	违规
有	探头	违规
有	-	正常
无	-	正常

由于多目标检测任务和行为判别任务需要 YOLOv3 网络同时输出结果,因此要对两个任务的数据集进行整体性标注,将各种考场行为设计为目标的类别标签。整体性标注过程如图 1 所示。

图 1　数据集整体标注

1.2　模型训练

训练阶段,神经网络模型主要采用了以特征金字塔和 Darknet53 相结合作为 backbone 的 YOLOv3 算法,并基于考场监考的任务特点在 YOLOv3 网络结构基础上进行了如下 5 点改善:

(1)考场只检测人体目标,同时考虑到坐姿下人

体目标大小基本一致,因此 YOLOv3 只需设定 1 个 anchor box,对这个 anchor box 进行行为类别的分类输出,极大地提升了检测效率。

(2)修改任务输出,将 YOLOv3 输出的 80 个类别的分类置信度向量改为本任务所需的 8 种行为标签向量,每个人体目标框后携带 8 个行为标签的置信度(出现纸条、出现书本、出现手机、扭头、斜视、说话、埋头、探头),修改后的输出与 YOLOv3 输出对比如图 2 所示。

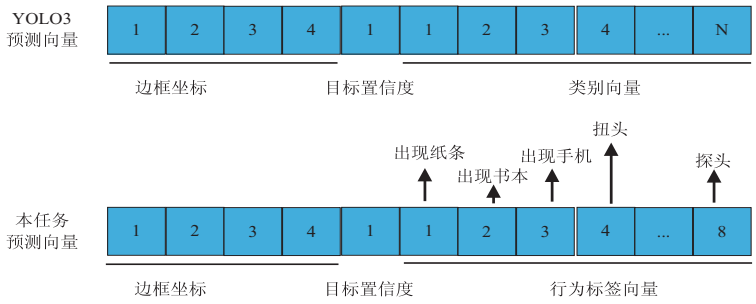

图 2 修改后的输出向量

(3)加入注意力机制,行为判别时主要行为都发生在人体上半身区域,因此利用注意力机制能很好地提升考场监考任务准确率。

CBAM(Convolutional Block Attention Module)模块,给定中间特征图,CBAM 按顺序推导出沿通道和空间两个独立维度的注意力图,然后将注意图相乘到输入特征图进行自适应特征细化。通道注意力聚焦在“什么”是有意义的输入图像,为了有效计算通道注意力,需要对输入特征图的空间维度进行压缩,对于空间信息的聚合,常用的方法是平均池化。空间注意力聚焦在“哪里”是最具信息量的部分,这是对通道注意力的补充。为了计算空间注意力,沿着通道轴应用平均池化和最大池化操作,然后将它们连接起来生成一个有效的特征描述符,如图 3 所示。

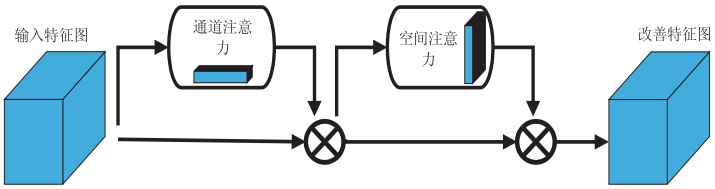

图 3 CBAM 模块

CBAM 可以无缝集成到任何 CNN 架构中,开销可以忽略不计,并且可以与基础 CNN 一起进行端到端训练。对于考场特殊环境下的考试人员,违规行为集中在上半身区域,因此加入 CBAM 模块,尤其是 CBAM 中的空间注意力图将有效地提升目标检测与行为判别的准确率,在残差网络中加入 CBAM 模块,如图 4 所示。

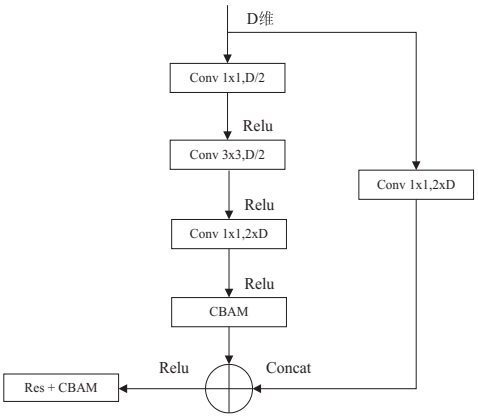

图 4 残差网络中加入 CBAM 模块

(4)改善损失函数计算,训练所用到的损失函数 L 应为人体目标置信度损失 L_t、人体目标坐标损失 L_c 和人体目标行为判别损失 L_a 之和。其中 L_t 为二分类交叉熵损失,存在人体目标时标签值为 1,不存在人体目标时标签值为 0,由于大部分网格没有人体目标,加入 λ 项来平衡损失值。L_c 为 MSE 损失,只有存在人体目标时才计算这部分损失。L_a 也在有人体目标时才计算,损失函数二分类交叉熵。具体公式如式(1)~式(4)。

$$L = L_t + L_c + L_a \tag{1}$$

$$L_t = \sum_{i=0}^{s^2}\sum_{j=0}^{B} l_{ij}^{\text{obj}} (c_i - \hat{c}_i)^2 + \lambda_n \sum_{i=0}^{s^2}\sum_{j=0}^{B} l_{ij}^{\text{noobj}} (c_i - \hat{c}_i)^2 \tag{2}$$

$$L_c = \lambda_c \sum_{i=0}^{s^2}\sum_{j=0}^{B} l_{ij}^{\text{obj}} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] + \lambda_c \sum_{i=0}^{s^2}\sum_{j=0}^{B} l_{ij}^{\text{obj}} [(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2] \tag{3}$$

$$L_a = \sum_{i=0}^{s^2}\sum_{j=0}^{B} l_{ij}^{\text{obj}} \sum_{k=0}^{L} y_{ijk} * \log(p_{ijk}) + \tag{4}$$

$$(1-y_{ijk})*\log(1-p_{ijk}) \tag{4}$$

(5)进行预训练,本地采集视频数据非常耗时,为了提高人体行为判别最后的准确率,需要利用主流数据集的预训练模型来初始化网络参数。结合考场任务的需求,最终选择了 COCO 数据集,COCO 数据集是标准的 YOLOv3 数据集,一共包含 20 万个图像,80 个类别中有超过 50 万个目标标注,平均每个图像的目标数为 7.2,是最广泛公开的目标检测数据集。COCO 数据集与其他数据集对比见表 2。

表 2　目标检测数据集对比

数据集	图片数	类别数
COCO	200K	80
ImageNet	450K	200
Pascal VOC	12K	20
Oxford-IIIT	7K	37
KITTI Vision	7K	3

预训练模型输出使用 COCO 数据集格式,YOLOv3 会利用特征金字塔进行上采样,并产生三种不同维度的输出。任务需要根据数据集选择合适的维度和 anchor box。每个维度输出 3 个 anchor box,每个 anchor box 携带 80 个类别的标签向量。在 YOLOv3 每一个残差块后添加一个 CBAM 模块,以提升整体准确率。

预训练完成后初始化网络,输出修改为任务所需的 1 个 anchor box 并携带 8 个行为判别标签向量,在本地数据集上进行再次训练,获得最终模型。

1.3　模型部署

模型运行阶段,利用部署在考场的高清摄像头采集视频数据流,并接入到视频处理设备。视频处理设备主要进行视频解帧与帧图像预处理,以符合神经网络数据输入。

将模型导入考场边缘设备中,接入视频处理设备的输出数据帧,将视频处理设备的输出视频帧作为输入数据,边缘设备同时输出所有考场人员目标检测框和 8 个行为类别判断,并根据类别判断的结果做出实时的考场行为判断,整体部署如图 5 所示。

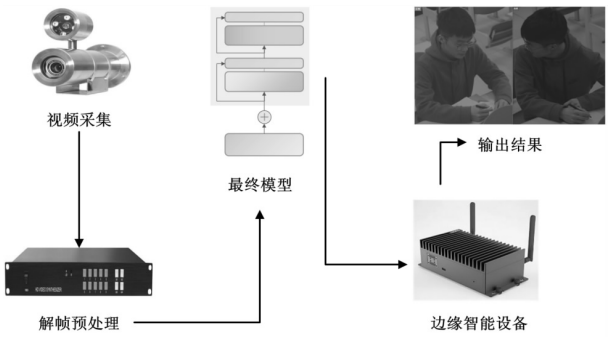

图 5　整体部署

2　实　验

实验运行部分结果如图 6 所示,可以发现,只有当 8 种行为类别均未检出才会判定为正常行为;如果检测出 1 条或多条行为类别,将判定为违规行为。

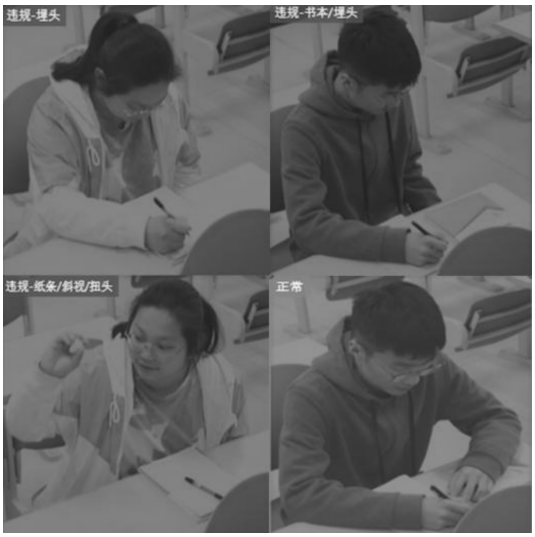

图 6　模型运行结果

为了更好地分析模型在实际监控中的效率与准确性,利用 COCO 数据集、本地采集小数据集、COCO+本地数据集三种数据组合分别测试了算法 Faster RCNN、SSD、RetinaNet 和本实验改进的 YOLOv3 模型(简称 Model)表现。

4 种方案都需要在 COCO 数据集上进行预训练并作本地数据集的迁移学习,才能在实际的考场视频上进行行为的识别。四种方案训练迭代过程对比如图 7 所示,其中横坐标表示迭代次数,纵坐标表示损失值。

可以发现 Model 相比 SSD、Faster RCNN 和 RetinaNet,具有优良的训练曲线,曲线在中前段就很快收敛到低值,中后段也更加平稳,反映出模型泛化推理阶段具有更好的鲁棒性。

将 4 种方案部署于考场,并对考场行为视频长时运行汇总后进行分析评价,评价指标采用目标检测常用的 mAP(mean Average Precision),并比较了四种方案在 COCO+本地数据集上的推理速度,实验分析结果见表 3。

表 3　行为识别对比实验分析 mAP

方案	COCO/%	Local/%	COCO+Local/%	推理速度/ms
Faster RCNN	41.3	66.5	82.4	52
SSD	39.2	63.3	83	48
RetinaNet	40.5	64.3	79.5	70
Model	41.5	67.2	88.3	23

可以看到模型的 mAP 在三组数据下均超过了其他方法,且具有最快的推理速度,具体原因可以总结

如下：

(1)YOLOv3 是一种集成了 SSD(多尺度预测)、FCN(全卷积)、FPN(特征金字塔)、DenseNet(特征通道)等多种特性的快速目标检测算法；

(2)改进的网络中加入了 CBAM 注意力模块，提取了丰富的人体局部特征，能有效地提升准确率和效率；

(3)具有单一评价指标的端到端训练很好地弥补了 YOLOv3 先天准确率上低于其他多阶段模型的缺点；

(4)本地小数据集采集自室内环境、场景单一、没有明显遮挡、只有 1 个 anchor box、输出类别数量只有 8 个等因素，使得 COCO+Local 的迁移学习方案获得了良好的效果。

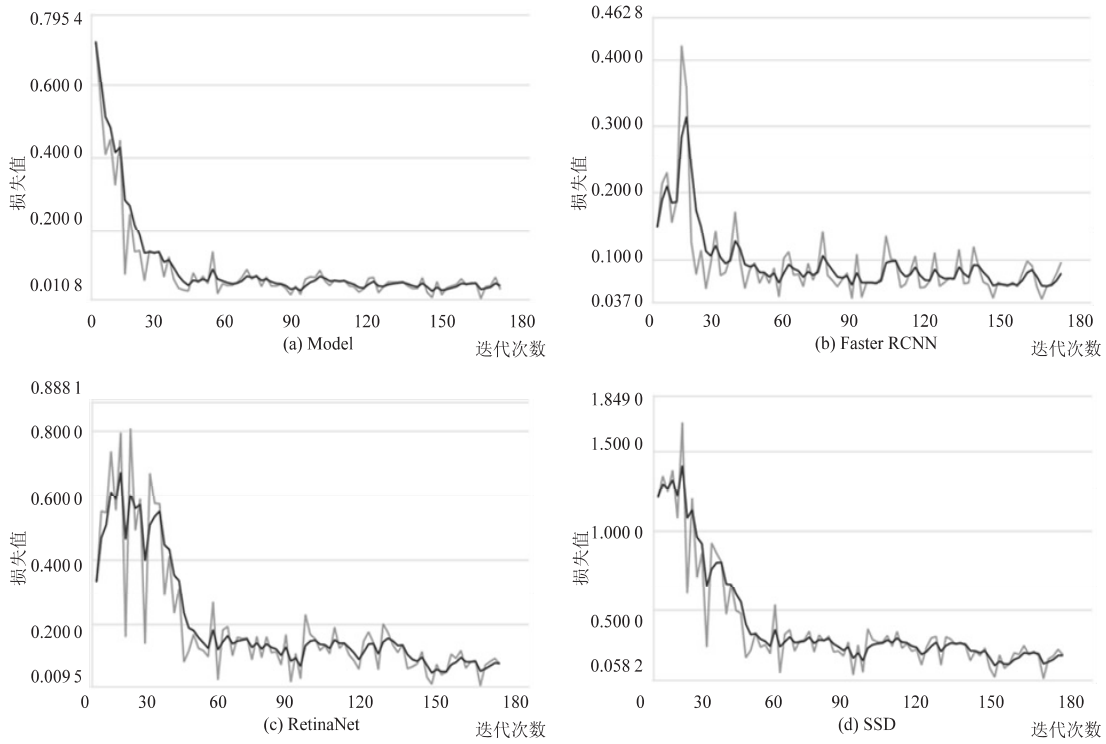

图 7 模型训练过程对比

对各种行为判别准确率的进一步分析可以发现，由于加入了通道注意力和空间注意力机制，在 COCO 数据集上可以提取到非常良好的人体上半身局部特征，因此最终在扭头、说话、埋头、斜视、探头等与人体局部特征相关的行为判别上表现明显好于其他非人体局部特征的行为，如图 8 所示。

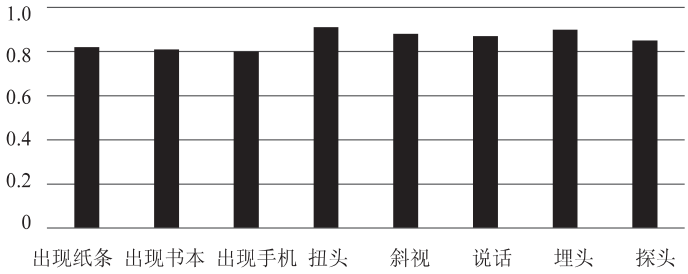

图 8 各行为判别对比

3 结束语

虽然行为识别领域由于问题场景的复杂性、拍摄角度不同、相同目标类内差异巨大、光线变化、类内和类间遮挡等问题，使得行为识别在自然采集的数据集上仍然存在很大的挑战。但是在实际的考场监控场景中，由于摄像头角度固定、室内无明显光学变化、检测的目标单一、类内无遮挡等天然的优势，使得利用 YOLOv3 改进后的模型能较好地同时执行多目标人体检测和行为判别任务。

端到端的多目标行为识别算法也具有一些不可避免的缺点：

(1)由于问题的特殊性，需要对本地采集数据进行精心的标注；

(2)由于数据集数量有限，导致无法胜任实际应用，需要借助其他数据集的预训练参数；

(3)算法属于静态帧的行为判别方法，没有加入时序信息，在一些情况下会造成错误的判定，比如捡东

西、举手等。鉴于此应加入人工复核机制,或增加采集到的视频时长,这些措施无疑都提高了算法的复杂性与操作的难度。

参考文献:

[1] 张恒鑫,叶颖诗,蔡贤资,等. 基于人体关节点的高效动作识别算法[J]. 计算机工程与设计,2020,41(11):3168-3174.

[2] 何嘉宇,雷　军,李国辉. 特征金字塔结构的时序行为识别网络[J]. 中国图象图形学报,2021,26(7):1637-1647.

[3] 董琪琪,刘剑飞,郝禄国,等. 基于改进 SSD 算法的学生课堂行为状态识别[J]. 计算机工程与设计,2021,42(10):2924-2930.

[4] 王昊飞,李俊峰. 基于注意力机制的改进残差网络的人体行为识别方法[J]. 软件工程,2021,24(11):51-54.

[5] 齐　琦,钱慧芳. 基于融合 3DCNN 神经网络的行为识别[J]. 电子测量技术,2019,42(22):140-144.

[6] 周　鹏,袁国良,张　颖,等. 基于改进的 DCNN 人体行为识别[J]. 传感器与微系统,2021,40(10):125-128.

[7] 窦雪婷. 基于 ResNet-LSTM 的行人过街行为识别方法[J]. 计算机与数字工程,2021,49(9):1872-1877.

[8] 陈京荣. 基于姿态估计与 Bi-LSTM 网络的跌倒行为识别[J]. 现代计算机,2021(20):80-85.

[9] 张　俊,李　昌. 基于 LSTM 多传感器数据融合人体行为识别方法[J]. 芜湖职业技术学院学报,2021,23(2):32-35.

[10] 罗毛欣,王天赋,白晓晨,等. 基于多分支 LSTM 网络的行为识别[J]. 科学技术创新,2021(13):13-14.

[11] 叶黎伟,宋宗珀. Yolo-v5 改进模型的课堂行为实时检测方法[J]. 长江信息通信,2021,34(7):41-45.

[12] 窦　刚,刘荣华,范　诚. 基于卷积神经网络的考场不当行为识别[J]. 中国考试,2021(2):56-62.

[13] 马钰锡,谭　励,董　旭,等. 面向智能监控的行为识别[J]. 中国图象图形学报,2019,24(2):282-290.

[14] 陈亚晨,韩　伟,白雪剑,等. 基于改进 YOLO-v3 的眼机交互模型研究及实现[J]. 科学技术与工程,2021,21(3):1084-1090.

[15] 孙宝聪. 基于图像检测的机场人员异常行为分析技术研究[J]. 数字通信世界,2020(1):26.

[16] GHARAHDAGHI A,RAZZAZI F,AMINI A. A non-linear mapping representing human action recognition under missing modality problem in video data[J]. Measurement,2021,186:110123.

[17] GUTOSKI M,LAZZARETTI A. Incremental human action recognition with dual memory[J]. Image and Vision Computing,2021,116:104313.

[18] SONG L W,XIANG C C,GUO H F,et al. Activity recognition using multiplex limited penetrable visibility graph[J]. Journal of the Mechanical Behavior of Biomedical Materials,2021,124:104880.

[19] BRIGHI M,FRANCO A,MAIO D. ActivityExplorer:a semi-supervised approach to discover unknown activity classes in HAR systems[J]. Pattern Recognition Letters,2021,151:340-347.

(上接第 166 页)

Hook:Curran Associates Inc.,2013:3111-3119.

[14] STRAKOVA J,STRAKA M,HAJIC J,et al. Neural architectures for nested NER through linearization[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. Florence:Association for Computational Linguistics,2019:5326-5331.

[15] XU M,JIANG H,WATCHARA W S,et al. A local detection approach for named entity recognition and mention detection[C]//Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1:long papers). Vancouver:Association for Computational Linguistics,2017:1237-1247.

[16] ZHANG S L,JIANG H,XU M B,et al. The fixed-size ordinally-forgetting encoding method for neural network language models[C]//Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2:short papers). Beijing:Association for Computational Linguistics,2015:495-500.

[17] LI X,YAN H,QIU X,et al. FLAT:Chinese NER using flat-lattice transformer[C]//Proceedings of the 58th annual meeting of the association for computational linguistics.[s. l.]:Association for Computational Linguistics,2020:6836-6842.