

基于视频时空特征的学生课堂行为检测

刘宏¹, 张释文¹, 孙程², 聂萱², 张锦^{1,2}

(1. 湖南师范大学信息科学与工程学院, 湖南长沙 410081;
2. 长沙理工大学计算机与通信工程学院, 湖南长沙 410114)

摘要: 课堂行为能有效反映学生的学习状态, 通过深度学习技术实现课堂行为检测对于改进教学方式、提高教学质量具有重要意义。目前, 课堂行为检测较多采用基于静态图像的方式进行, 往往会忽略行为的动态特征, 对连续性行为检测效果较差。为此, 提出一种基于视频时空特征的YOLOv7-SlowFast课堂行为检测方法, 通过YOLOv7定位学生目标后采用SlowFast检测课堂行为。首先, 为提高YOLOv7在人群密集环境下的检测精度, 引入自适应空间特征融合模块解决不同尺度特征之间不一致的问题; 其次, 采用RepGhost轻量化模块改进YOLOv7网络结构, 通过结构重新参数化提高模型检测速度; 最后, 针对SlowFast时空行为检测精度低的问题, 设计基于标准化的时间注意力模块, 加强模型对时间特征的感知。实验结果表明, 改进后YOLOv7在Crowdhuman数据集上平均精度均值(mAP)达86.96%; 改进后SlowFast在自制课堂行为检测数据集上的mAP达87.28%, 能够实现课堂行为的有效检测。

关键词: 课堂行为; 时空行为检测; SlowFast; YOLOv7

DOI: 10.11907/rjdk.231427

中图分类号: TP391

文献标识码: A

开放科学(资源服务)标识码(OSID):

文章编号: 1672-7800(2024)008-0267-08



Student Classroom Behavior Detection Based on Video Spatio-Temporal Features

LIU Hong¹, ZHANG Shiwen¹, SUN Cheng², NIE Xuan², ZHANG Jin^{1,2}

(1. College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China;

2. School of Computer and Communication Engineering, Changsha University of Science & Technology, Changsha 410114, China)

Abstract: Classroom behavior can effectively reflect students' learning status, and using deep learning technology to detect classroom behavior is of great significance for improving teaching methods and enhancing teaching quality. At present, classroom behavior detection is mostly based on static images, which often ignore the dynamic characteristics of behavior and have poor performance in detecting continuous behavior. To this end, a YOLOv7 SlowFast classroom behavior detection method based on video spatiotemporal features is proposed, which locates student targets through YOLOv7 and uses SlowFast to detect classroom behavior. Firstly, in order to improve the detection accuracy of YOLOv7 in densely populated environments, an adaptive spatial feature fusion module is introduced to solve the problem of inconsistency between features of different scales. Then, the RepGhost lightweight module is used to improve the YOLOv7 network structure, and the model detection speed is improved by reparameterizing the structure. Finally, to address the issue of low spatiotemporal behavior detection accuracy in SlowFast, a standardized time attention module is designed to enhance the model's perception of temporal features. The experimental results show that the improved YOLOv7 has an average precision mean (mAP) of 86.96% on the Crowdhuman dataset. After improvement, SlowFast achieved an mAP of 87.28% on a self-made classroom behavior detection dataset, which can effectively detect classroom behavior.

Key Words: classroom behavior; spatio-temporal behavior detection; SlowFast; YOLOv7

收稿日期: 2023-04-25

扫描二维码阅读全文:



基金项目: 中央军委装备发展部全军共用信息系统装备预研项目(31511010402)

作者简介: 刘宏(1963-), 男, 硕士, 湖南师范大学信息科学与工程学院教授、博士生导师, 研究方向为人工智能、智能家居等; 张释文(1997-), 女, 湖南师范大学信息科学与工程学院硕士研究生, 研究方向为深度学习、计算机视觉; 孙程(1994-), 女, 博士, 长沙理工大学计算机与通信工程学院讲师, 研究方向为计算机视觉; 聂萱(2001-), 男, 长沙理工大学计算机与通信工程学院硕士研究生, 研究方向为高性能计算; 张锦(1979-), 男, 博士, 长沙理工大学计算机与通信工程学院、湖南师范大学信息科学与工程学院教授、博士生导师, 研究方向为人工智能、网络与信息安全等。本文通讯作者: 张锦。

0 引言

人工智能的飞速发展使各行各业都在朝着智慧化方向迈进^[1]。智慧课堂作为一种新型教学模式,正逐渐在教育领域得到广泛应用,而对学生的课堂行为进行评估与分析是智慧课堂的重要组成部分。课堂行为不仅能反映出学生的学习状态,而且能侧面验证教师的教学质量和效果,对学生的课堂行为进行观察与评估有助于及时发现教学过程中存在的不足,并有针对性地对学生的行为进行干预。因此,采用深度学习等技术自动识别和检测学生课堂行为对于协助教师全面把握学生的学习情况、发现教学问题和调整教学进度等具有重要意义^[2-3]。

1 相关研究

目前,许多学者采用目标检测^[4]、姿态识别^[5]、视频行为识别与检测^[6]等方法解决课堂行为检测问题。其中,目标检测方法是将被检测课堂图片中的学生行为看作目标对象进行检测。例如,Li等^[7]采用改进的YOLOv5s模型检测多个学生的行为,利用增强型超分辨率生成对抗网络提高图像的视觉清晰度,增加微小目标的有效特征,有效提高了检测准确率;朱先远等^[8]针对YOLOv5模型对目标密集、遮挡严重的课堂异常行为检测效果不佳的问题,提出了融合MobileNetV3的YOLOv5模型,对相互遮挡的多目标检测达到了更好的检测效果。然而目标检测方法会忽略行为的动态特征及上下文信息,无法有效检测长时间序列中学生的连续动作。

姿态识别方法通过提取人体关键点信息识别和检测学生的动作姿态。例如,闫兴亚等^[9]使用基于MobileNetV3的轻量级模型BetaPose获得检测位置后提取人体关键点进行行为分类,在拥挤场景下的姿态识别中获得了较高的准确性;苏超等^[10]采用融合注意力机制的目标检测算法提取学生位置,然后采用改进OpenPose模型提取人体骨骼关键点信息进行行为分类,实现了学生课堂动作的准确检测。基于姿态识别的课堂行为检测方法在复杂场景下的检测效果较差,且人体关键点数据的采集和处理过程较为复杂,具有一定的应用局限性。

视频行为识别与检测方法能够较好地提取学生的动态信息,有效检测连续性行为。例如,黄勇康等^[11]构建了学生课堂行为数据集,结合目标检测与跟踪算法获取学生实时图片流,利用深度残差卷积神经网络识别学生课堂行为;Albert等^[12]通过实时目标检测与跟踪得到每个学生的实时图片流,然后采用3D-CNN网络学习每个学生行为的时空特征,从而实现多个目标学生的课堂行为实时识别。视频行为识别与检测方法能够有效捕获学生行为的动态特征,提高检测的准确性和可靠性,但由于需要对行为的

长时间序列进行操作以提取时空特征,导致其计算复杂度较高,在实时性方面表现不佳。

为有效实现基于视频时空特征的课堂行为检测,同时兼顾检测准确率和实时性需求,本文提出一种基于YOLOv7-SlowFast的课堂行为检测方法,主要贡献为:①提出YOLOv7-SlowFast课堂行为检测方法,实现课堂环境下的多学生常见行为检测;②为提高YOLOv7的检测效果,引入自适应空间特征融合模块(Adaptively Spatial Feature Fusion, ASFF)和RepGhost轻量化模块改进模型结构;③为提高SlowFast模型的检测准确率,采用基于标准化的时间注意力模块对模型进行改进,提高其特征提取能力。

2 方法介绍

2.1 YOLOv7-SlowFast时空行为检测模型

在行为检测过程中,每段视频都需要通过目标检测模型获取学生位置,并且需根据学生目标框提取特征图,因此目标检测结果将直接决定提取行为时空特征的结果,从而影响后续课堂行为分类。此外,目标检测算法的检测速度也会影响整体视频行为检测的效率。SlowFast是用于视频行为识别与检测任务的3D卷积神经网络模型,通常采用Faster R-CNN作为目标检测器^[13-14]。与Faster R-CNN算法相比,YOLOv7在检测速度和准确性方面都更为优秀^[15]。因此本文选择YOLOv7作为目标检测器,结合SlowFast模型进行课堂行为检测。

如图1所示,SlowFast模型包含两条不同时间和空间分辨率的路径,分别为Slow路径和Fast路径,以3D ResNet50作为主干网络,其中res2-res5表示残差模块。两条路径以不同帧率运行,具有不同的采样间隔和卷积通道数,以获取视频中更全面的时空信息。其中,Slow路径以低时间分辨率和高空间分辨率运行,主要捕获视频中的空间语义信息;Fast路径以高时间分辨率和低空间分辨率运行,主要捕获行为快速变化的动态信息,并采用更小的卷积宽度降低通道容量,以减少计算量。Slow与Fast路径的帧速率比为 $1:\alpha$ ($\alpha > 1$),通道数比为 $1:\beta$ ($\beta < 1$)。网络通过多次横向连接融合双流分支的特征,Fast路径的特征图经过3D卷积进行尺度变换后与Slow路径特征融合。进行时空行为检测时需要首先进行目标定位,然后提取行为时空特征并进行分类。当视频输入模型时,按照30帧/s的速率剪切视频帧。SlowFast网络每次处理的视频长度为 $\tau \times T$ 帧,其中Slow路径按照步长为 τ 采样(τ 一般取16),采样 T 帧输入网络(T 一般取4);而Fast路径按照步长 τ/α 采样 αT 帧输入网络(α 一般取8)。同时,抽取第 $(\tau \times T)/2$ 帧作为该时间段内的关键帧,将其输入到YOLOv7模型中检测学生位置,并将结果输入RoI Align模块。该模块将目标框投影到3D特征图上以获取相应的特征矩阵,通过将每个2D RoI按时间维度进行扩展得到3D RoI。最后将统一尺

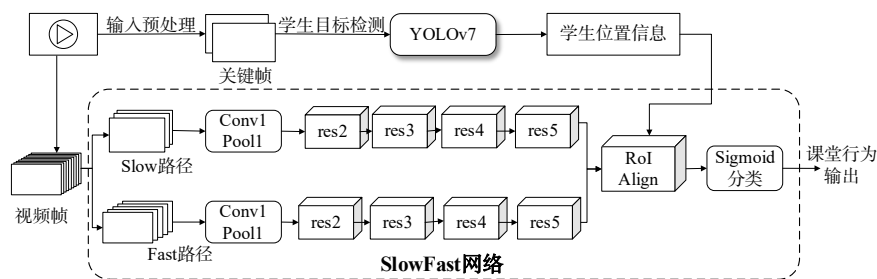


Fig. 1 YOLOv7-SlowFast model structure

图1 YOLOv7-SlowFast模型结构

寸后的特征图经池化和全连接层处理,采用 Sigmoid 分类器进行课堂行为分类。

2.2 改进 YOLOv7 网络

YOLOv7 延续了 YOLO 系列算法的设计思想,具有检测速度快、精度高和易于部署等优点。YOLOv7 的主干网络包括最大池化与卷积结合的 MPConv 和高效聚合模块 (Efficient Layer Aggregation Networks, ELAN); 颈部通过特征金字塔 (Feature Pyramid Networks, FPN) 结构和路径融合网络 (Path Aggregation Network, PANet) 融合高层和低层特征信息,提高模型中语义信息的丰富性和位置特征的提取

能力;检测部分采用重参数化卷积模块 RepConv 提高推理检测速度,同时采用辅助头训练和动态标签分配策略提升整体检测性能。课堂环境较为复杂,人群密集且遮挡严重。为提高检测精度,本文提出一种改进的 YOLOv7 目标检测算法。首先,采用 ASFF 模块提高模型对多尺度特征的利用效果;其次,为提高模型检测速度,利用 RepGhost 轻量化模块改进颈部网络结构,降低模型参数量和计算量,减少模型推理时间。改进后的 YOLOv7 网络结构如图 2 所示,其中在检测头部采用 ASFF 模块,在颈部网络中将 ELAN 模块替换为轻量化 RepGhost 模块。

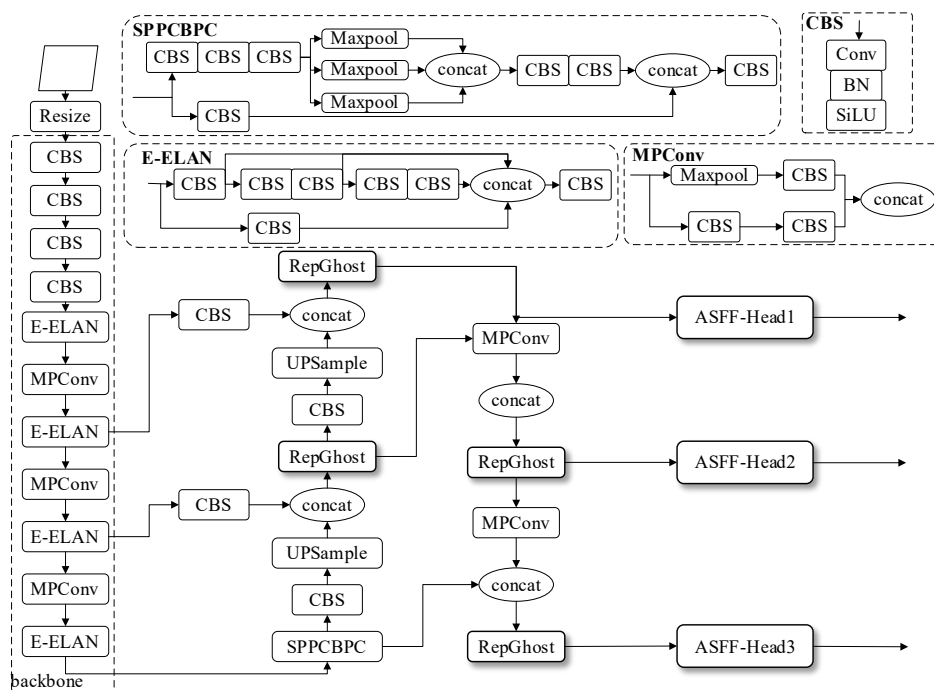


Fig. 2 Improved YOLOv7 network structure

图2 改进 YOLOv7 网络结构

2.2.1 自适应空间特征融合

FPN 和 PANet 结构虽然能有效融合多尺度特征信息,但也存在不同特征层目标尺寸不一致的问题。网络预测前无法充分利用不同分支的特征信息,限制了特征信息的交互和性能的提升。针对该问题,本文采用 ASFF 检测头,在 PANet 网络后过滤 3 个不同尺度的特征,充分利用低层细粒度特征和高层语义信息^[16]。图 3 为 ASFF 特征融合示意图,其中 X^1 、 X^2 和 X^3 分别表示 PANet 网络输出的不同大

小特征图。首先对特征图进行尺度变换,通过上下采样的卷积操作调整特征空间分辨率及通道数。 $X^{2 \rightarrow 1}$ 表示 X^2 调整至 X^1 相同尺寸后的特征图,特征调整后进行特征层的自适应融合。 Y^1 、 Y^2 和 Y^3 表示融合后的特征。 Y_{ij}^l 表示输出特征图 Y^l 在 (i, j) 位置的特征向量。计算公式为:

$$Y_{ij}^l = \alpha_{ij}^l \times X_{ij}^{1 \rightarrow l} + \beta_{ij}^l \times X_{ij}^{2 \rightarrow l} + \gamma_{ij}^l \times X_{ij}^{3 \rightarrow l} \quad (1)$$

式中: $X_{ij}^{n \rightarrow l}$ 表示从 n 层特征调整至 l 层特征在 (i, j) 位置的特征向量; α_{ij}^l 、 β_{ij}^l 和 γ_{ij}^l 分别表示在 (i, j) 位置特征融合

时的空间滤波权重参数,在模型训练时通过反向传播更新学习。式(2)为 α_{ij}^l 的计算公式, β_{ij}^l 和 γ_{ij}^l 的计算方法与之相同。

$$\alpha_{ij}^l = \frac{e^{\lambda_{\alpha_{ij}}^l}}{e^{\lambda_{\alpha_{ij}}^l} + e^{\lambda_{\beta_{ij}}^l} + e^{\lambda_{\gamma_{ij}}^l}} \quad (2)$$

式中: $\lambda_{\alpha_{ij}}^l$ 、 $\lambda_{\beta_{ij}}^l$ 和 $\lambda_{\gamma_{ij}}^l$ 分别表示 α_{ij}^l 、 β_{ij}^l 和 γ_{ij}^l 的调控参数,通过卷积计算得到,且需满足 $\alpha_{ij}^l + \beta_{ij}^l + \gamma_{ij}^l = 1$ 和 $\alpha_{ij}^l, \beta_{ij}^l, \gamma_{ij}^l \in [0, 1]$ 。

2.2.2 RepGhost轻量模块

为提高模型检测速度,实现模型的轻量化,减少网络推理时间,采用RepGhost轻量级卷积模块改进模型网络结构^[17]。RepGhost基于GhostNet模块^[18]进行改进,采用Add操作替换Ghost模块中低效的Concat运算,以隐性实现特征重用。由于内存复制的复杂性,Concat运算在硬件上的运算效率较低,而Add操作可以在权值空间中完成特征融合过程,且不会引入额外的推理时间。

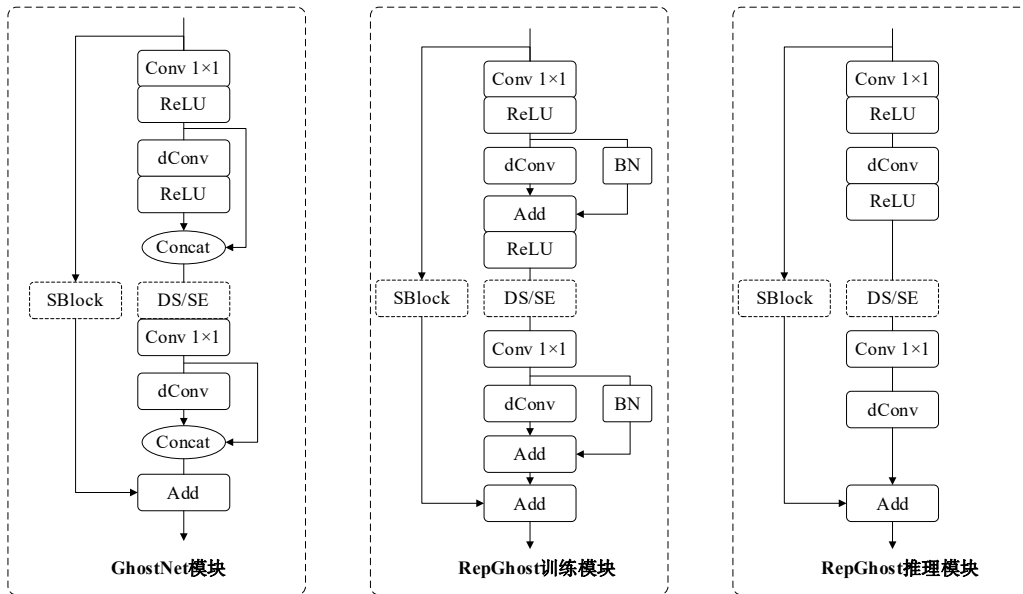


Fig. 4 RepGhost module structure

图4 RepGhost模块结构

2.3 改进SlowFast网络

在视频行为检测任务中,由于被检测对象的行为是动态的,目标的外观和位置不断变化,加之视频中复杂背景的干扰,较易出现漏检和错检等现象,而引入注意力机制能有效提高模型的检测效果。例如,文献[19]提出一种时空注意力模块提高SlowFast模型对视频中时空行为的检测效果。受此启发,为了更好地捕获行为的动态信息,同时考虑到时间特征为视频处理中区别不同动作的重要信息,本文借鉴标准化注意力机制(Normalization-based Attention Module, NAM)^[20],设计用于时间维度的注意力模块NTAM(Normalization-based Temporal Attention Module),利用权重的贡献改善注意力机制。结构如图5所示。

为避免SENet等注意力机制使用的全连接层和卷积

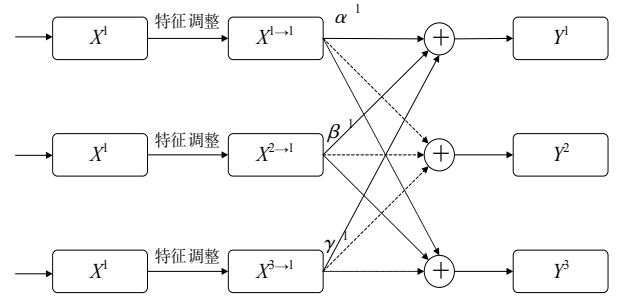


Fig. 3 Schematic diagram of ASFF feature fusion

图3 ASFF特征融合示意图

RepGhost模块结构如图4所示。与GhostNet模块相比,RepGhost在重参数化时采用Add操作,在分支中添加批归一化(Batch Normalization, BN),在训练过程中带来非线性。将ReLU在深度卷积层向后移动,满足结构重新参数化的规则。在训练阶段存在BN分支,在推理阶段则只有卷积层和ReLU激活层,以使模块具有较高的硬件效率。

层^[21],采用标准差代表权重的重要性,使用BN的比例因子表示时间的变化强度,并将注意力模块引入3D残差块中,使模型能够捕获更多运动信息,从而提高其对时间特征的感知能力。通过比例因子计算时间维度方差,并将其作为衡量权重重要程度的依据。计算公式为:

$$B_i = BN(B_{input}) = \alpha \frac{B_{input} - \mu_\beta}{\sqrt{\sigma_\beta^2 + \varepsilon}} + \beta \quad (3)$$

式中: μ_β 为小批量 B 的均值; σ_β 为标准差; α 和 β 分别为尺度和位移可训练的放射变换参数; ε 用于防止方差为0。每个时间维度权重的计算公式为:

$$\omega_\alpha = \frac{\alpha_i}{\sum_{i=0} \alpha_i} \quad (4)$$

式中: α_i 为每个时间维度的比例因子。时间注意力计算后输出特征的计算公式为:

$$M_t = \text{sigmoid}(\omega_\alpha(BN(F_{input}))) \quad (5)$$

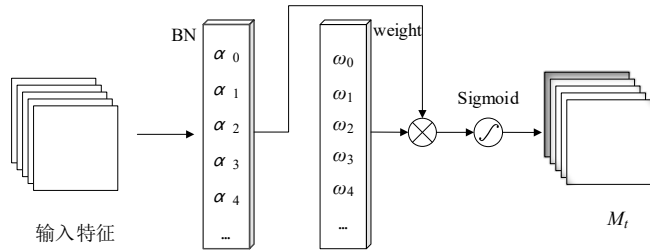


Fig. 5 NTAM module structure

图 5 NTAM 模块结构

3 实验方法与结果分析

3.1 实验环境

YOLOv7 模型训练时设置迭代次数 epoch 为 300, 权重衰减系数为 0.000 5, 初始学习率为 0.01, batchsize 大小为 32。SlowFast 模型训练时目标检测模块未参与联合训练, 优化器采用随机梯度下降 (Stochastic Gradient Descent, SGD), 权重衰减系数为 0.000 01, 初始学习率为 0.012 5, batchsize 为 4, 迭代次数 epoch 为 30。

3.2 实验数据集

选择公开的密集人群检测数据集 Crowdhuman^[22], 该数据集提供了头部、完整人体、可见区域等多种标注信息, 包含多种人群密集的场景。本文实验只对完整人体进行训练与测试, 训练集、验证集和测试集的图像数量分别为 15 000、4 370 和 5 000。

同时自制面向时空的课堂行为检测数据集, 参考公开的时空行为检测数据集 AVA 的标注方式进行构建^[23]。数据集制作步骤为: ①将经过筛选、截取处理后的视频按照 30 帧/s 切分视频帧; ②按照步幅为 1s 提取视频中的关键帧, 将这些关键帧输入 Faster R-CNN 目标检测器自动标注图像中的学生位置信息; ③采用标注工具对检测出的学生目标框进行微调, 再进行学生的多标签课堂行为标注, 生成 AVA 格式数据集标注文件^[24]。

自制课堂行为检测数据集中单条视频包含学生 7~20 人。各行类别及标签数量分别为看黑板 (26 131)、低头 (10 780)、扭头/转身 (5 419)、讲话 (2 909)、站立 (1 585)、举手 (2 551)、趴在桌上 (2 010), 总标注量为 51 387。按照 4:1:1 的比例划分数据集、验证集和测试集。图 6 为数据集中部分行为示意图。

3.3 模型评价指标

采用精确率 (Precision)、召回率 (Recall)、各类别的平均精度 (Average Precision) 和平均精度均值 (mean Average Precision, mAP) 评价模型检测性能。计算公式分别为:



Fig. 6 Examples of classroom behavior in dataset

图 6 数据集中课堂行为示例

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

$$AP = \int_0^1 P(R) dR \quad (8)$$

$$mAP = \frac{1}{n} \sum_{i=1}^n AP \quad (9)$$

式中: TP 表示预测为正样本且实际标签为正的样本数量; FP 表示预测为正样本但实际标签为负的样本数量; FN 表示预测为负样本但实际标签为正的样本数量; AP 为以召回率为横坐标、精确率为纵坐标绘制的 P-R (Precision-Recall) 曲线下面积。

同时采用每秒帧数 (Frames Per Second, FPS) 评估模型的检测速度。采用参数量、模型大小评估模型复杂度。

3.4 实验结果与分析

3.4.1 目标检测实验结果

为验证改进模块对模型性能的贡献, 在 Crowdhuman 数据集上进行消融实验。结果如表 1 所示。可以看出, 引入 ASFF 模块后, 模型 mAP 提高了 1.57%, 但增加了模型参数量和计算量, FPS 减少了 4.45。引入 RepGhost 轻量化模块后, 模型 mAP 提高了 1.05%, FPS 增加了 5.71, 能够提高检测精度并达到更快的推理速度。同时引入两种改进模块后, mAP 与原模型相比提高了 1.31%, FPS 有所降低, 但与单独引入 ASFF 模块相比 FPS 增加了 2.03, 在实现高精度检测的同时兼顾了检测速度, 能够更好地适应人群密集环境。

Table 1 YOLOv7 ablation experiment result

表 1 YOLOv7 消融实验结果

模型	参数量/ 10 ⁶	FPS	精确 率/%	召回 率/%	mAP/%
YOLOv7	36.48	40.16	85.19	76.95	85.65
YOLOv7+ASFF	58.14	35.71	85.99	78.67	87.22
YOLOv7+RepGhost	30.13	45.87	84.86	78.95	86.70
YOLOv7+ASFF+RepGhost	51.79	37.74	84.92	78.79	86.96

将改进后的 YOLOv7 模型与 Faster R-CNN、YOLOv3 和 YOLOv5 等模型进行比较, 结果如表 2 所示。可以看出,

改进后的 YOLOv7 模型在检测准确率和检测速度方面均表现较好。

Table 2 Comparison of detection performance of different models

表 2 不同模型检测性能比较

模型	参数量/ 10 ⁶	FPS	精确率/%	召回率/%	mAP/%
Faster R-CNN	186.00	3.44	71.66	72.06	72.50
YOLOv3	61.53	14.49	85.25	85.06	81.50
YOLOv5s	7.02	97.08	84.94	68.51	79.46
YOLOv5m	20.88	68.03	86.65	87.51	82.10
YOLOv5l	46.13	37.04	86.86	73.47	83.49
YOLOv7-tiny	6.00	47.85	82.71	72.29	81.24
YOLOv7	36.48	40.16	85.19	76.95	85.65
改进后 YOLOv7	51.79	37.74	84.92	78.79	86.96

改进后的 YOLOv7 模型在 Crowdhuman 数据集和课堂环境下的检测效果如图 7 所示。可以看出,该模型在人群密集环境下有较好的检测效果,能够检测出图像中存在部分遮挡的学生。

3.4.2 行为检测实验结果

为验证 NTAM 模块对模型的改进效果,在自制课堂行为检测数据集上进行实验,其中 SlowFast 采用 3D ResNet50 作为主干网络,两条通道帧速率比 α 取值为 8,通道数之比 β 取值为 1/8。结果如表 3 所示。可以看出,引入 NTAM 模块后,SlowFast 模型的 mAP 提高了 1.81%,对“讲话”这种对时序特征要求较高的行为检测效果提升明显,AP 值提高了 11.01%。NTAM 模块使模型更加关注时间维度的重要

信息,提取更多有效的行为动态特征,实现更好的课堂行为检测效果。

Table 3 Detection results of various types of behaviors before and after model improvement

表 3 模型改进前后各类行为检测结果 %

模型	各类行为 AP							mAP
	抬头	低头	扭头/ 转身	讲话	站立	举手	趴在 桌上	
SlowFast	95.56	89.09	81.13	62.04	91.07	85.40	93.98	85.47
SlowFast+NTAM	95.68	91.34	79.58	73.05	91.15	85.76	94.37	87.28

为了验证本文模型的课堂行为检测效果,在自制数据集上与 SlowOnly、MViT^[25] 和 LSTC^[26] 和改进前 SlowFast 模型进行比较。结果如表 4 所示。可以看出,改进后的 SlowFast 模型取得了最高的检测准确率。

Table 4 Classroom behavior detection results of different models

表 4 不同模型课堂行为检测结果

模型	模型大小/MB	参数量/10 ⁶	mAP/%
SlowOnly (3D ResNet 50)	121.50	31.80	75.38
SlowOnly (3D ResNet 101)	194.15	50.80	82.84
MViT	138.51	36.30	74.48
LSTC	274.23	71.80	86.78
SlowFast	128.64	33.66	85.47
改进后 SlowFast	128.64	33.66	87.28

3.4.3 Faster R-CNN/YOLOv7-SlowFast 课堂行为检测结果比较

SlowFast 原始模型在检测时空行为时采用 Faster R-



(a) Crowdhuman dataset detection performance

(a) Crowdhuman 数据集检测效果



(b) Detection effect in classroom environment

(b) 课堂环境下检测效果

Fig. 7 Improved YOLOv7 detection effect

图 7 改进后 YOLOv7 检测效果

CNN 作为目标检测器。表 5 为 Faster R-CNN-SlowFast 与本文提出的 YOLOv7-SlowFast 的时空行为检测结果,其中目标检测部分置信度阈值为 0.8,行为检测置信度阈值为 0.5。可以看出,与 Faster R-CNN-SlowFast 模型相比,YOLOv7-SlowFast 模型可有效减少目标检测时间,且能够检测到视频中更多的学生目标。

图 8 为 Faster R-CNN-SlowFast 和 YOLOv7-SlowFast 模型的课堂行为检测结果。可以看出,YOLOv7-SlowFast 模型能够检测出更多学生目标,对学生的定位更加准确,对后排被遮挡的学生也有较好的检测效果。而在图 8(e)中,Faster R-CNN-SlowFast 模型将两名学生标注进同一检测

Table 5 Faster R-CNN/YOLOv7-SlowFast Spatio-temporal behavior detection results

表 5 Faster R-CNN/YOLOv7-SlowFast 时空行为检测结果					
视频	目标检测模型	关键帧数量	目标检测时间/s	时空行为检测时间/s	检测人数
视频①	Faster R-CNN	105	9.52	23.17	9
	YOLOv7		8.14	21.97	15
视频②	Faster R-CNN	86	6.45	17.65	9
	YOLOv7		4.50	15.67	10
视频③	Faster R-CNN	87	6.46	17.62	10
	YOLOv7		4.77	15.96	13

框中,出现了明显的错检现象。



Fig. 8 Classroom behavior detection results of Faster R-CNN-SlowFast and YOLOv7-SlowFast

图 8 Faster R-CNN-SlowFast 和 YOLOv7-SlowFast 模型课堂行为检测结果

4 结语

本文针对学生课堂行为检测问题提出基于时空行为特征的 YOLOv7-SlowFast 模型,通过由 ASFF 模块和 RepG-

host 模块改进的 YOLOv7 检测学生位置,再通过由 NTAM 模块改进的 SlowFast 提取行为的时空特征,取得了较高的检测精度和较快的检测速度。与基于静态图像的课堂行为检测方法相比,YOLOv7-SlowFast 模型充分考虑行为的空间和运动特征,对连续性行为具有较好的检测效果。该模型可应用于学生课堂专注度与参与度评估中,通过检测整节课课堂表现分析学生行为的积极性和消极性,通过专注

度计算方法评估学生在课堂中的学习情况,为教师提供及时有效的反馈。

然而,由于需要使用目标检测模型检测视频关键帧中的学生位置,而 SlowFast 网络采用 3D CNN 的骨干网络,模型整体参数量及计算量仍较大。如何减少计算量及推理时间是未来需要研究的方向。同时可以训练模型检测学生在课堂中的异常行为,如玩手机、打架等,有助于纠正学生的不良习惯,维护课堂秩序。

参考文献:

- [1] CHANG J, ZHANG G W, CHEN W J, et al. Gas station unsafe behavior detection based on YOLO-V3 algorithm[J]. China Safety Science Journal, 2023, 33(2): 31-37.
常捷,张国维,陈文江,等. 基于 YOLO-V3 算法的加油站不安全行为检测[J]. 中国安全科学学报, 2023, 33(2): 31-37.
- [2] ZHAO C, SHU H, GU X Q, et al. The measurement and analysis of students' classroom learning behavior engagement based on computer vision technology[J]. Modern Educational Technology, 2021, 31(6): 96-103.
赵春,舒杭,顾小清. 基于计算机视觉技术的学生课堂学习行为投入度测量与分析[J]. 现代教育技术, 2021, 31(6): 96-103.
- [3] ZHENG Z J, LIANG G J, LUO H B, et al. Attention assessment based on multi-view classroom behaviour recognition[DB/OL]. <https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.12146>.
- [4] GUO Q. System analysis of the learning behavior recognition system for students in a law classroom: based on the improved SSD behavior recognition algorithm[J]. Scientific Programming, 2022, 2022: 1-11.
- [5] ZHOU J, RAN F, LI G, et al. Classroom learning status assessment based on deep learning[J]. Mathematical Problems in Engineering, 2022(3): 1-9.
- [6] XIE Y, ZHANG S, LIU Y. Abnormal behavior recognition in classroom pose estimation of college students based on spatiotemporal representation learning[J]. Traitement du Signal, 2021, 38(1): 89-95.
- [7] LI L, LIU M, SUN L, et al. ET-YOLOv5s: toward deep identification of students' in-class behaviors[J]. IEEE Access, 2022, 10: 44200-44211.
- [8] ZHU X Y, WANG S L, ZHOU Y F, et al. An improved classroom abnormal behavior detection algorithm based on YOLOv5[J]. Journal of Qiqihar University (Natural Science Edition), 2023, 39(1): 46-52.
朱先远,王松林,周叶凡,等. 一种改进 YOLOv5 的课堂异常行为感知检测算法[J]. 齐齐哈尔大学学报(自然科学版), 2023, 39(1): 46-52.
- [9] YAN X Y, KUANG Y X, BAI G R, et al. Student classroom behavior recognition method based on deep learning [J/OL]. Computer Engineering: 1-8 [2023-02-22]. <https://doi.org/10.19678/j.issn.1000-3428.0065369>.
闫兴亚,匡娅茜,白光霖,等. 基于深度学习的学生课堂行为识别方法[J/OL]. 计算机工程: 1-8 [2023-02-22]. <https://doi.org/10.19678/j.issn.1000-3428.0065369>.
- [10] SU C, WANG G Z. Research on student behavior recognition based on improved OpenPose[J]. Application Research of Computers, 2021, 38(10): 3183-3188.
苏超,王国中. 基于改进 OpenPose 的学生行为识别研究[J]. 计算机应用研究, 2021, 38(10): 3183-3188.
- [11] HUANG Y K, LIANG M Y, WANG X X, et al. Multi-person classroom action recognition in classroom teaching videos based on deep spatiotemporal residual convolution neural network[J]. Journal of Computer Applications, 2022, 42(3): 736-742.
黄勇康,梁美玉,王笑笑,等. 基于深度时空残差卷积神经网络的课堂教学视频中多人课堂行为识别[J]. 计算机应用, 2022, 42(3): 736-742.
- [12] ALBERT C C Y, SUN Y, LI G, et al. Identifying and monitoring students' classroom learning behavior based on multisource information[DB/OL]. <https://onlinelibrary.wiley.com/doi/full/10.1155/2022/9903342>.
- [13] CHRISTOPH F, FAN H, MALIK J, et al. Slowfast networks for video recognition [C]//The IEEE 2019 International Conference on Computer Vision, 2019: 6201-6210.
- [14] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149.
- [15] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors [DB/OL]. <https://arxiv.org/pdf/2207.02696.pdf>.
- [16] LIU S, HUANG D, WANG Y. Learning spatial fusion for single-shot object detection[DB/OL]. <https://arxiv.org/pdf/1911.09516.pdf>.
- [17] CHEN C, GUO Z, ZENG H, et al. RepGhost: a hardware-efficient ghost module via re-parameterization [DB/OL]. <https://arxiv.org/pdf/2211.06088.pdf>.
- [18] HAN K, WANG Y, TIAN Q, et al. Ghostnet: more features from cheap operations [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1580-1589.
- [19] YANG F. Research on students' classroom behavior based on WS-DAN and spatiotemporal attention[D]. Chengdu: Sichuan Normal University, 2022.
杨帆. 基于 WS-DAN 与时空注意力的学生课堂行为检测研究[D]. 成都: 四川师范大学, 2022.
- [20] LIU Y, SHAO Z, TENG Y, et al. NAM: normalization-based attention module[DB/OL]. <https://arxiv.org/pdf/2111.12419.pdf>.
- [21] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.
- [22] SHAO S, ZHAO Z, LI B, et al. Crowdhuman: a benchmark for detecting human in a crowd [DB/OL]. <https://arxiv.org/pdf/1805.00123.pdf>.
- [23] GU C, SUN C, ROSS D A, et al. Ava: a video dataset of spatio-temporally localized atomic visual actions [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6047-6056.
- [24] YANG F. A multi-person video dataset annotation method of spatio-temporally actions[DB/OL]. <https://arxiv.org/pdf/2204.10160.pdf>.
- [25] FAN H, XIONG B, MANGALAM K, et al. Multiscale vision transformers [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 6824-6835.
- [26] LI Y, ZHANG B, LI J, et al. LSTC: Boosting atomic action detection with long-short-term context [C]//Proceedings of the 29th ACM International Conference on Multimedia, 2021: 2158-2166.

(责任编辑:尹晨茹)