计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(4):113−122 [doi: 10.15888/j.cnki.csa.009450]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于改进 YOWO 算法的学生课堂行为识别①

徐鑫磊, 张景异

(沈阳理工大学 自动化与电气工程学院, 沈阳 110159)
通信作者: 徐鑫磊, E-mail: 1981626804@qq.com

摘　要: 当前, 大部分的学生课堂行为识别工作主要基于单帧图像进行, 忽略了行为的连贯性, 因此不能充分利用视频信息来对学生的课堂行为进行准确刻画. 所以, 本文提出一种改进的 YOWO 算法模型, 有效利用视频信息对学生课堂行为进行识别. 首先, 本文采集某高校真实课堂教学中的授课录像, 制作出包含 5 类学生课堂行为的 AVA 格式视频数据集; 其次, 采用时移模块 TSM (temporal shift module), 用来增强模型获取时间上下文信息的能力; 最后, 采用非局部操作模块 non-local 来提高模型提取关键位置信息的能力. 实验结果表明, 通过对 YOWO 模型的优化, 使得网络的识别性能更佳. 在学生课堂行为数据集上, 改进后的算法的 *mAP* 值为 95.7%, 相较于原 YOWO 算法在 *mAP* 值上提高了 4.6%; 模型参数量为 81.97×10^6, 计算量为 22.6 GFLOPs, 参数量和计算量分别降低 32.3% 和 9.6%; 检测速度为 24.03 f/s, 提升了约 3 f/s.

关键词: YOWO 算法; TSM; non-local; 学生课堂行为; 行为识别; 注意力机制

引用格式: 徐鑫磊,张景异.基于改进 YOWO 算法的学生课堂行为识别.计算机系统应用,2024,33(4):113–122. http://www.c-s-a.org.cn/1003-3254/9450.html

Classroom Behavior Recognition of Students Based on Improved YOWO Algorithm

XU Xin-Lei, ZHANG Jing-Yi

(School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang 110159, China)

Abstract: At present, since the recognition of most students' classroom behavior is mainly based on a single frame image and ignores behavior coherence, video information cannot be made full use of to accurately depict students' classroom behavior. Therefore, this study proposes an improved YOWO algorithm model to effectively employ video information to identify students' classroom behavior. First, this paper collects teaching videos from real classroom teaching in a university and produces an AVA format video dataset containing five types of students' classroom behavior. Second, the temporal shift module (TSM) is adopted to enhance the ability of this model to obtain time context information. Finally, a non-local operation module is utilized to improve the ability of the model to extract key location information. The experimental results show that by optimizing the YOWO model, the recognition performance of the network is better. In the classroom behavior dataset, the *mAP* value of the improved algorithm is 95.7%, 4.6% higher than that of the original YOWO algorithm. The parameter number in the model is reduced by 32.3% at 81.97×10^6 and the calculation amount is decreased by 9.6% at 22.6 GFLOPs. The detection speed is 24.03 f/s, an increase of about 3 f/s.

Key words: YOWO (you only watch once) algorithm; temporal shift module (TSM); non-local; student classroom behavior; behavior recognition; attention mechanism

在现代教育中, 智能化越来越重要. 近年来, 随着计算机视觉和人工智能等技术的迅速发展以及教育采集设备与录制系统不断完善, 促使着越来越多的研究人员去将教学和人工智能技术结合起来, 慢慢地出现

① 收稿时间: 2023-09-09; 修改时间: 2023-10-08, 2023-11-03; 采用时间: 2023-11-15; csa 在线出版时间: 2024-01-18
CNKI 网络首发时间: 2024-01-19

了许多相关的研究工作.

周鹏霄等[1]提出, 把视频帧转换成图像, 根据图像中的人脸数量、轮廓特征和主体动作幅度 3 个特征, 利用贝叶斯模型对学生的行为进行预测, 最终实现了对课堂教学视频的 S-T 行为的智能分析. 党冬利[2]对学生在课堂上说话的行为进行了识别, 根据特定的研究情境和学生的特征, 采用贝叶斯分类器来识别学生的姿势, 其中包括了 3 种最常见的学生姿势: 举手, 听讲, 站立.

前面所说的研究者所采用的研究方法, 都是以传统的机器学习为基础的, 然而, 由于传统的机器学习方法无法对特征进行有效的抽取, 所以, 也有研究者从深度学习的角度对学生的课堂行为进行了研究. 例如, 廖鹏等[3]建立了一套以深度学习为基础的学生课堂中的异常行为检测与分析系统, 利用 VGG 模型对学生在课堂中的异常行为进行特征提取, 进而可以发现在上课时睡觉、玩手机等不正常行为; 秦道影[4]在 ResNet50 预训练模型的基础上, 提出了一种迁移学习方法, 该方法建立了 ResNet50 网络, 可以对学生的 7 种行为进行识别, 分别是: 看书、写字、举手、听讲、站立、睡觉和左顾右盼; 蒋沁沂等[5]针对卷积神经网络随网络层次的增加而出现的性能下降, 提出了一种以残差结构为基础的深层残差网络, 并成功地识别出了 6 种学生行为. 这些研究者采用的是经典的深度学习特征提取网络, 再根据提取到的特征进行学生行为的识别.

随着时间的推移, 一些研究者将经典的目标检测算法应用到了学生课堂行为识别当中. 例如, 周叶[6]提出了在 Faster R-CNN 检测框架的基础上进行改进, 利用特征金字塔和主要样本注意机制, 分别解决了不同尺度的学生课堂行为检测和数据类别不平衡问题, 从而实现了以图像为基础的小学生课堂行为检测. 杜俊强[7]提出, 首先利用 YOLOv3 定位目标物体, 并对图像进行裁剪, 然后利用 ResNet50 与 VGG16 模型, 利用双网络模型的融合框架, 来完成对学生上课行为的辨识. 董琪琪等[8]提出基于改进的 SSD 目标检测算法进行教室学生课堂行为识别, 实现了准确检测学生睡觉、举手、回答、写字、听讲 5 种动作. 孙绍涵等[9]在 YOLOv4 的基础上提出改进, 并通过构建学生的课堂行为状态数据集, 对教室中学生的课堂行为状态进行识别和分析.

以上研究者采用的数据类型都是单一图片数据, 为了增强学生行为识别效果, 一些研究者采用多模态数据的方法. 例如, 林灿然等[10]将人体关键点与 RGB 图像特征相结合, 采用多模态数据方法, 对学生的课堂行为进行了检测; 郭俊奇等[11]根据课堂教学情景, 构建了能提取时间上下文信息的三维卷积神经网络, 实现对教师课堂行为的识别; 还提出了改进损失函数的 YOLOv5 模型, 实现了多目标的学生课堂行为识别. 郑丹[12]提出了一种将改进的双流卷积神经网络用于学生课堂行为的识别中, 它以 VGG16 为空间特征, 利用光流学习网络来提取时间流特征, 从而提高了学生行为识别的精度.

总体而言, 当前在课堂学习行为识别方面, 有两个方面的困难: 一是难以获得课堂上的学生学习行为数据, 而这一类型的数据在公共的行为识别数据集中又非常匮乏. 因此, 研究者们就必须对课堂上的视频进行专门的收集, 并对视频进行筛选划分、分割和标注, 这是一项非常耗费时间的工作. 二是已有学者对课堂学习行为进行识别, 多以检测为主, 仅根据单一画面进行分类, 未充分考虑到学习行为的时序性, 无法对学习行为进行有效的识别和分析.

在时空动作检测任务中需要在网络结构中融入两种不同的信息: 一种是前一帧的时间信息; 另外一种是从关键帧中得到的空间信息. YOWO (you only watch once) 网络之前提出的方法通常是通过单独的网络提取这些信息, 另外再使用一种额外的融合机制来获取检测结果. 而 YOWO[13−16]是一个具有双分支的单阶段网络, 能够在推理过程中实现时间信息和空间信息的同步获取, 并且能够从视频中直接预测出目标物体的边界框与行为类别. 因为整体的网络结构是一致的, 所以能够端到端的进行优化.

因此, 本研究采用 YOWO 网络作为基线模型[13], 通过改进 YOWO 网络的不足, 有效利用行为的时序性, 实现检测视频中的学生行为. 它能使教师更好地掌握学生的学习状况, 并能适时地调整教学方式, 使教学流程得到优化.

1 YOWO 网络结构

YOWO 模型由德国慕尼黑工业大学人机通信研究所提出, YOWO 的网络结构如图 1 所示. YOWO 结构是一个单阶段网络, 具有两个分支, 能进行端到端的训练. 2D-CNN 分支提取关键帧 (即当前帧) 的空间特征, 3D-CNN 分支提取先前帧的时间特征, 为了更好地汇总两个分支的特征, 采用了通道融合注意机制来充分发挥通道之间的相关性. 最后, 将融合后的图像特征

用于帧水平的检测. 采用 YOWO 网络作为基线模型是因为它具有很强的灵活性, 它的 2D-CNN 和 3D-CNN 分支的架构可以被任意的 CNN 架构去替换, 可以根据学生课堂行为识别任务去调整网络结构达到最佳效果.

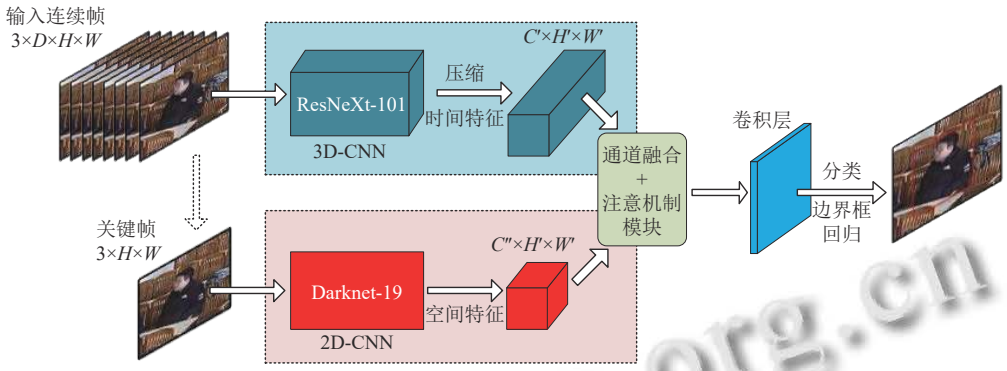


图 1 YOWO 网络结构图

在 YOWO 结构中采用 3D-CNN 分支, 是由于动作的前后帧信息对于动作的识别起着至关重要的作用. 3D-CNN 能在时间维度上进行运算, 用来获取上下文特征信息. YOWO 之所以采用 3D-ResNeXt-101 结构, 是由于该结构在 Kinetics 数据集上效果很好. 3D 网络的输入是连续帧, 形状为 $3\times D\times H\times W$, 输出为经过压缩的 $C'\times H'\times W'$, 压缩是为了能够匹配 2D-CNN 的输出特征大小, 方便进行特征融合.

YOWO 结构中的 2D-CNN 分支是用来进行提取关键帧中的空间特征, 可以为学生行为定位提供有效的信息. YOWO 中使用 Darknet-19 作为基本架构, 是因为它能很好地兼顾效率和准确率. 3D 网络输入连续帧中的最新帧作为 2D 网络关键帧的输入, 因此, 没有必要再额外地重新加载数据. Darknet-19 的输出特征图大小和 3D 网络的输出保持一致, 形状为 $C''\times H'\times W'$.

YOWO 结构中的特征融合模块, 将 3D-CNN 和 2D-CNN 两个分支得到的特征进行有效融合.首先, 将两个特征沿着通道维度进行堆叠, 然后在此基础上, 利用格拉姆矩阵对多个通道的相关性进行映射, 并将各通道的相关性与原有特征的权重加权求和, 建立多通道特征映射间的长时间语义相关性的模型, 从而实现多通道特征的融合. 该方法突出了上下文关系, 提高了特征的辨识能力.

YOWO 结构的最后使用一个卷积层来预测行为类别和边界框大小. 采用的方法与 YOLOv2 的一样, 每个网格都有 5 个相应的先验框, 是使用 K-means 算法进行计算得到. 每个先验框都会得到 *NumCls* 个行为分类分数, 4 个坐标分数, 1 个置信度分数, 然后对先验框回归得到预测框.

2 模型改进

YOWO 算法在自制的行为数据集上, 由于样本特征有限, 使得模型学习到的特征也有限, 从而导致模型的检测精度不高, 通过分析发现, YOWO 的网络结构存在两个问题: 一是 YOWO 的 2D 特征提取网络没有获取时间上下文信息的能力, 因此针对具有相似的行为容易误判; 另一个是 YOWO 的 2D 特征提取网络缺乏关键信息的获取能力, 因此图片中的细节特征不容易被关注到, 从而会损失精度, 比如看书和玩手机这两种行为, 区别就在于手正在触摸的东西是什么. 本研究基于这两点问题改进 YOWO 算法网络结构, 实现在原始的单阶段人体行为定位框架 YOWO 上的创新, 同样可以实现端对端的联合训练, 提高模型的检测精度, 并在自制的学生课堂行为数据集上实验验证, 实现效果更好的学生行为的正确识别.

2.1 改进后网络结构

图 2 显示了改进后的系统框架. 在 2D 主干网络中, 首先将 Darknet-19 替换为参数量更小的 ResNet18 网络, Darknet-19 网络的卷积层数比 ResNet18 多一层, 并且它有 3 个通道数为 1 024 维的卷积核而 ResNet18 没有, 所以导致了它的参数量大; 其次在 ResNet18 中融入 TSM 模块和 non-local 模块, 来提升 2D 网络的时间上下文信息获取能力和关键信息提取能力, 并且 TSM 模块不会增加网络的参数, 只是沿时间维度移动部分通道, non-local 模块中的卷积核大小都为 1×1, 加入到模型中不会让网络参数有相同数量级的增加. 以此来达到既减少参数量又改善网络模型性能的效果. 在自制的学生课堂行为数据集上, 让网络更快更有效地提

取特征, 进一步增强模型的检测精度.

2.2 TSM

TSM 是用于高效视频理解的时移模块[17−20], TSM 的原理示意图如图 3 所示, TSM 沿时间维度移动部分通道, 用来促进相邻帧之间的信息交换, 它能够方便地插入到二维 CNN 中, 在零计算和零参数的条件下, 实现时间建模. 因此, 可以将它和 YOWO 网络的 2D-CNN 结合, 来增强网络在时间上的特征提取能力, 当存在遮挡时, 可以通过多帧图片以进行信息互补, 从而丰富特征信息来提高准确率.

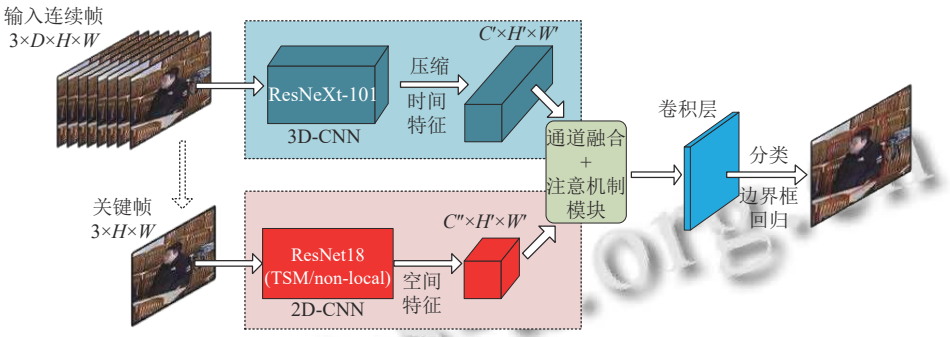


图 2 改进后 YOWO 网络结构图

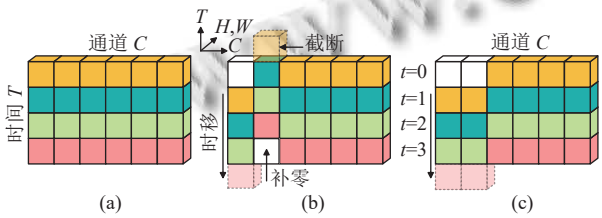


图 3 TSM 原理示意图

传统的 2D CNN 在维度 T 上独立运行, 因此, 不会影响时间建模 (图 3(a)). 相比之下, TSM 在时间维度上, 向前和向后移动部分通道, 相邻帧的信息会与当前帧混合在一起, 时间信息会更丰富 (图 3(b)). 但是实时在线视频推理, 未来的帧不能移动到现在, 因此使用单向 TSM (图 3(c)) 来执行在线视频推理.

推理时, 对每一帧都保留了前 1/8 的特征, 并进行缓存. 在下一帧中, 将当前特征图的前 1/8 替换为缓存的特征. 将当前的 7/8 特征与之前的 1/8 特征相结合, 以产生下一层, 然后反复进行, 如图 4 所示.

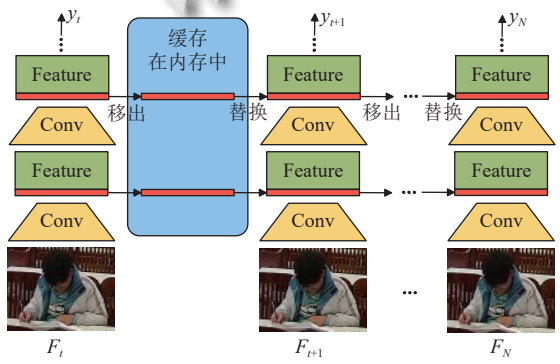


图 4 单向 TSM 推理过程

TSM 模块的插入位置为 ResNet18 网络残差块的内部, 如图 5 所示. 插入到残差块的内部是为了保证网络的空间特征学习能力, 如果将 TSM 直接插入每个卷积层或残差块之前, 它会损害主干模型的空间特征学习能力, 特别是当移动大量通道时, 当前帧存储的信息在移动通道中会丢失, 所以将 TSM 放在残差块中的残差分支内, 这种方式特征在通过时间偏移后, 原始激活的所有信息仍然可以访问.

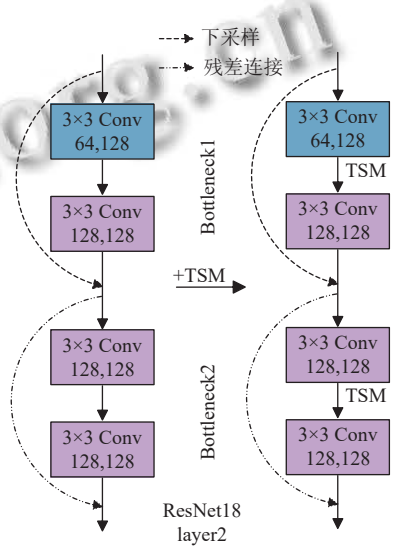


图 5 ResNet18 中 TSM 插入位置

2.3 Non-local

Non-local 操作[21]是一种高效、简单且通用, 用于捕获深度神经网络的远程依赖性的模块. 直观上理解, non-local 操作为计算某个位置的响应是输入特征图中

所有位置特征的加权和, 并且是一个灵活的构建块, 可以很容易地与卷积层一起使用, 加入到 ResNet18 网络中, 可以增强网络关键信息的获取能力. 因为学生课堂行为存在因细微的变化而学生行为不同的情况, 所以通过 non-local 操作可以加强网络对相似行为之间细节判断的能力. 同时对模型的运行速度不会产生明显影响.

Non-local 运算定义如式 (1):

$$y_i = \frac{1}{c(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \tag{1}$$

其中, i 表示待求响应在空间、时间或时空上的位置索引, j 表示该响应的可能位置索引. x 为输入信号 (图像、序列、视频); y 是与 x 大小一样的输出信号. 函数 f 计算 i 与 j 的相关程度. 函数 g 对输入信号在位置 j 上乘以一个参数矩阵进行线性映射计算, 可以理解为嵌入操作. $c(x)$ 是为了进行归一化, 让输出值保持在一定的范围内, 有助于模型收敛, 是通过 Softmax 的方式进行归一化.

式 (1) 是计算跟 i 位置可能相关的其他位置 j 的结果, 相关的位置可以是前后帧图像上的位置, 而不是像传统的卷积只是计算相邻位置上的信息, 所以能够提取到更广泛范围的信息.

式 (1) 中的 non-local 操作可以融合到一个 non-local 块中, 该块可以方便地加入到许多现有网络结构中. 将 non-local 块定义如式 (2):

$$z_i = w_z y_i + x_i \tag{2}$$

其中, y_i 在式 (1) 中给出, W_z 是用来线性映射的参数, 可以随着网络训练进行变化, "$+x_i$"代表残差连接. 残差连接可以在不改变原有结构的情况下, 将一个新的 non-local 块插入到已有的模型中. 图 6 展示了一个 non-local 块的示例, 图中的 θ、ϕ、g、W_z 代表着维度为 1×1×1 的卷积核, 输入 X 维度为 T×H×W×1024, T 代表着输入图片的帧数, H 和 W 表示输入的高和宽, 1 024 代表着通道数. 首先将输入分别经过 θ、ϕ、g 的卷积运算得到 $\theta(x_i)$、$\phi(x_j)$、$g(x_j)$, 然后将 $\theta(x_i)$ 和 $\phi(x_j)$ 点乘得到 $f(x_i, x_j)$, 再通过 Softmax 进行归一化, 再和 $g(x_j)$ 相乘就会得到 y_i, 也就是式 (1) 的结果, 最后再经过 W_z 卷积运算并加上原始输入 X_i 得到最终输出 Z_i. 可以发现, non-local 主要通过一系列的卷积运算来实现, 所以可以方便地和其他网络融合发挥它的作用.

Non-local 的插入位置如图 7 所示. 在 ResNet18 网络的第 2, 3 个 layer 中, 每个 Bottleneck 块的最后一层后面插入 non-local 块, 用来对前面提取到的学生课堂行为特征进行全局信息建模, 从而加强特征细节上的有效性.

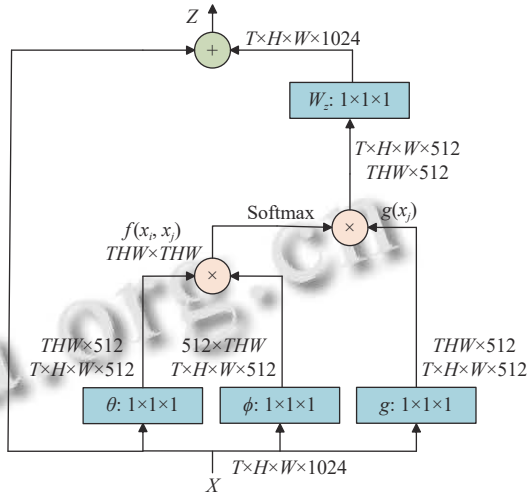


图 6 Non-local 块

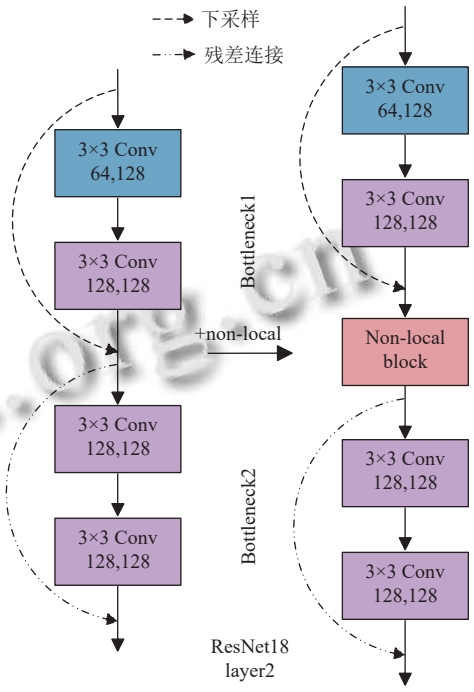


图 7 ResNet18 中 non-local 插入位置

2.4 学生课堂行为识别流程

将改进后的 YOWO 模型, 对学生在课堂上行为进行识别的流程如图 8 所示.

在训练过程中: (1) 首先要对数据预处理. 利用教室中的摄像头, 收集学生在课堂上的真实学习行为录像; 通过数据筛选, 以收集到的学生课堂学习行为录像

为基础, 选取出其中所要研究的部分学生行为; 将视频分割成连续多帧图像, 对每个帧图像中的学生行为进行标注, 并划分为训练集、验证集和测试集. (2) 然后将数据集输入到检测网络中, 得到目标的分类、坐标回归和置信度的值. (3) 最后, 使用优化算法计算梯度并将其用于反向传播, 更新模型的参数.

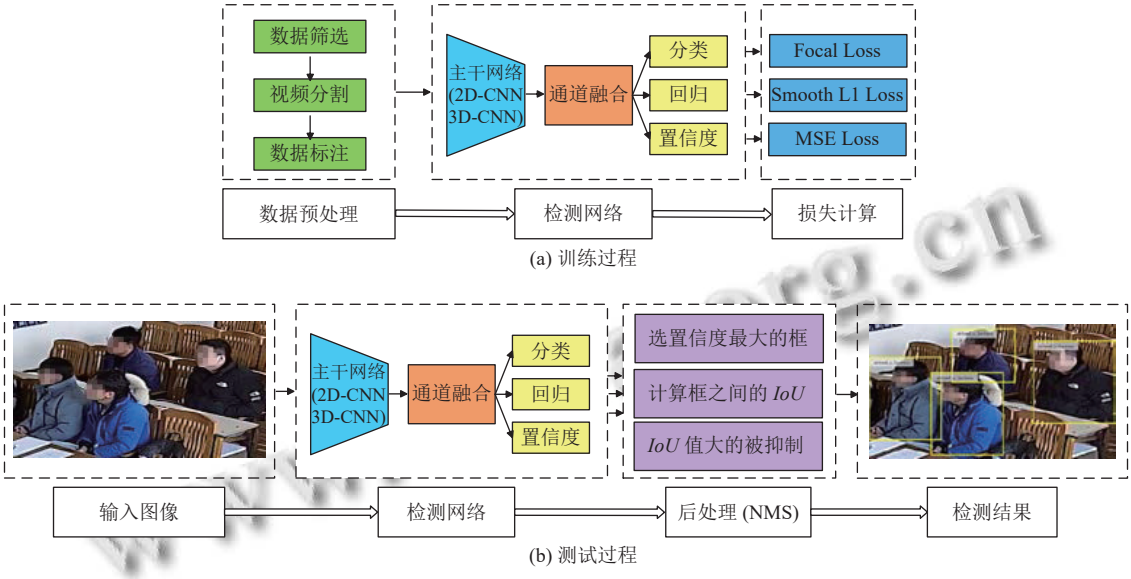


图 8 学生课堂行为识别流程图

在测试过程中: 首先将图像输入到训练好的检测网络中, 经过网络推理得到大量的预测框, 然后通过后处理操作进行去重, 最后输出检测结果.

3 学生课堂行为数据集构建

由于目前国内没有公开的学生课堂行为数据集, 因此以 2023 年辽宁省沈阳市某大学真实课堂教学环节中的授课视频作为原始数据.

在收集到的原始资料基础上, 对学生的课堂行为特征进行分析, 总结出了具有代表性的学生的课堂行为, 包括听课、看书、玩手机、书写和不学习 (包含吃东西、左顾右盼、睡觉等行为) 5 类, 以此来构建相应的学生课堂行为视频数据集, 每种行为具体的动作状态如表 1 所示. 听课、看书、书写代表着学生在课堂上积极的学习行为, 而玩手机、不学习代表着学生在课堂上消极的行为. 因此, 可以较为全面地将学生在课堂上的学习状况表现出来.

制作数据集的原始数据是约 1 500 min 的课堂视频, 为了确保了数据集的多样性, 所收集的实际课堂情景为 15 个, 每个都包含 2 节课, 总时长在 90–100 min 之间. 为了对学生课堂行为进行识别, 因此在选择数据时, 应该尽可能多地反映出学生上课时的正面信息. 因此, 将教室前方的摄像头用作数据集的原始数据, 但是会存在学生课堂行为部分遮挡的特点. 数据集示例如图 9 所示.

表 1 学生课堂行为的动作状态

学生课堂行为	动作状态
听课 (attend_a_lecture)	眼注视前方+端坐
看书 (looking_book)	手触碰书本+低头
玩手机 (playing_phone)	手触碰手机+低头
书写 (writing)	手握笔+低头
不学习 (no_studying)	吃东西、睡觉等

4 实验分析

为检验改进后的 YOWO 模型在学生课堂行为识别中的效果, 本研究基于视频的学生课堂行为识别数据集, 针对改进点设计了消融实验.

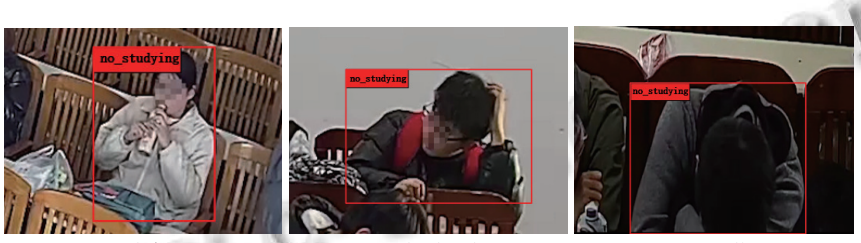
4.1 实验环境配置

实验采用的硬件配置为 Intel(R) Core(TM) i7-7700 CPU @ 3.60 GHz, 内存为 16 GB, GPU 型号为 NVIDIA GeForce GTX 1070 Ti, 深度学习框架为 PyTorch 1.9.0 版本, CUDA 版本为 10.2, Python 版本为 3.9.16.

模型在训练时, 输入图片大小均为 224×224, batch size 均为 16, 采用 Adam 优化器进行优化, 初始学习率

为 0.000 1, 动量为 0.9. 采用的分类损失是 Focal Loss、边界框损失为 Smooth L1 Loss、置信度损失采用的 MSE Loss, 训练轮次为 50 轮, 同时会将验证效果最好的模型权重进行保存.

(a) 听课 (b) 看书 (c) 书写 (d) 玩手机



(e) 吃东西 (f) 左顾右盼 (g) 睡觉

图 9 学生课堂行为数据集示例

4.2 评价指标

本文通过精确率 (Precision, P)、召回率 (Recall, R)、平均精度 (AP) 和平均精度均值 (mAP) 来评价模型的性能. AP 值通过 IoU 设置为 0.5 时的平均精度来测量, 即 AP@0.5. 计算公式如式 (3)–式 (6) 所示:

$$P = \frac{TP}{TP+FP} \tag{3}$$

$$R = \frac{TP}{TP+FN} \tag{4}$$

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{5}$$

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{6}$$

其中, TP 表示标注框与预测框的 IoU 大于 0.5 的检测框数量; FP 表示 $IoU \leqslant 0.5$ 的检测框数量; FN 表示未检测到的标注框数量; $P(R)$ 代表 P-R 曲线, 其中精确率 P 为纵坐标, 召回率 R 为横坐标; AP 代表通过插值计算的 P-R 曲线下的面积; C 代表共 C 个类; mAP 为所有类别的 AP 值的平均值.

本文还选择 FPS 作为模型性能的评价指标之一, 它代表着每秒可处理图像帧的数量. 帧数越大, 表示检测速度越快, 计算公式如式 (7) 所示:

$$FPS = \frac{Num}{Time} \tag{7}$$

其中, Num 固定时间内检测总帧数, $Time$ 是固定的时间间隔长度.

4.3 实验与结果分析

在相同的实验环境和超参的情况下, 本研究针对改进 YOWO 的 2D-CNN 网络结构进行消融实验, 实验中所采用的 3D-CNN 网络都为 ResNeXt-101, 使用的数据集都为自制的学生课堂行为数据集, 实验结果如表 2 所示. 模型 1 是未做改进的 baseline 网络; 模型 2 是将 2D 主干网络替换 ResNet18; 模型 3 是在模型 2 的基础上加入 TSM 模块; 模型 4 是在模型 2 的基础上加入 non-local 模块; 模型 5 为同时加入 TSM 和 non-local 模块.

表 2 消融实验结果 1 (%)

模型	2D主干网络	精确率	召回率	mAP@0.5
1	Darknet-19	86.7	87.8	91.1
2	ResNet18	87.2	85.0	90.3
3	ResNet18+TSM	94.3	89.9	95.4
4	ResNet18+non-local	92.5	90.5	95.1
5	ResNet18+TSM+non-local	93.8	92.1	95.7

由表 2 的实验结果可知, 将 YOWO 网络的 2D-CNN 网络换为 ResNet18 时, mAP 值下降了 0.8%, 是因为它的结构与 Darknet-19 相比网络层数和参数量更小, 所以提取到的学生行为特征不够, 从而学习能力有所下降, 导致对学生行为的判断准确率下降.

模型 3 与模型 2 相比, 它的精确率、召回率与

mAP@0.5 分别提升了 7.1、4.9 和 5.1 个百分点. 对于学生课堂行为的识别, 行为的时序性是至关重要的, 模型 3 通过加入 TSM 模块, 解决了 2D 主干网络不能获取上下文信息的问题, 同时为了避免 2D 主干网络的空间特征提取能力受到明显影响, 只对网络中的部分层进行 TSM 操作, 来达到空间特征和时间特征上的平衡. 实验结果表明, TSM 结构有助于检测出更多的目标且准确率更高.

将模型 4 和模型 2 对比, 可以发现它的精确率、召回率与 *mAP*@0.5 分别提升了 5.3、5.5 和 4.8 个百分点. 由于学生课堂行为的动作幅度比较小且微小的变化就会导致行为发生改变, 因此需要着重关注带来行为转变的区域. 模型 4 在 2D 主干网络中引入 non-local 模块, 用来计算影响学生行为转变相关性最大区域, 并在训练时给予更大的权重, 来提高关键信息的利用能力. 实验结果表明, non-local 模块的确有助于检测到更多目标, 提升检测精度, 同时比模型 3 有更高的召回率.

模型 5 的实验结果表明, 将 TSM 和 non-local 同时加入到网络中, 二者可以起到一定的互补作用, 进一步提高网络的效果, *mAP*@0.5 的值达到了 95.7%.

为了更好地验证改进后模型的训练效果, 将模型的训练损失进行对比实验, 结果如图 10. 从图 10 中可以看出模型 3 和模型 4 的收敛速度比模型 2 要快, 并且损失值也要小, 并在 50 轮时达到了收敛, 说明 non-local 和 TSM 发挥了作用.

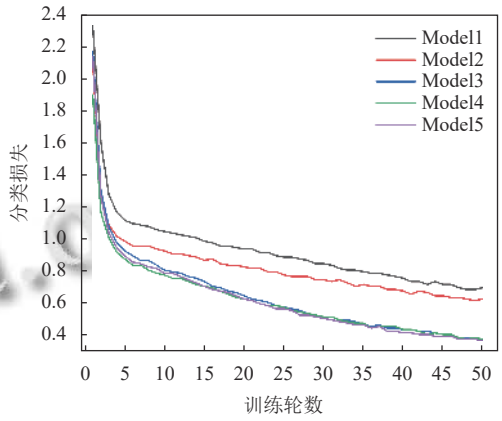


图 10 损失变化曲线

表 3 主要是针对 5 个模型的 Params (参数量)、FLOPs (浮点运算次数) 与 *FPS* (每秒处理图像帧数) 进行实验对比.

表 3 消融实验结果 2

模型	2D主干网络	*mAP*@0.5 (%)	Params ($\times 10^6$)	FLOPs ($\times 10^6$)	*FPS* (f/s)
1	Darknet-19	91.1	120.99	24 989.24	21.04
2	ResNet18	90.3	81.81	22 526.95	26.99
3	ResNet18+TSM	95.4	81.81	22 526.95	26.28
4	ResNet18+non-local	95.1	81.97	22 578.93	24.42
5	ResNet18+TSM+non-local	95.7	81.97	22 578.93	24.03

通过表 3 对比发现, YOWO 网络的改进使得模型对学生课堂行为识别的准确率上升了, 并且模型参数量减少为原来的 68%, 计算量减少为原来的 90%, *FPS* 提升了 5 帧左右, 让模型的整体效果有了一定的提升.

为了更加清晰地分析网络对于每一类学生课堂行为的检测效果, 采用 *AP* 作为算法性能评估指标, 针对上面 5 种模型进行对比.

为了更直观地展现各类别的检测效果, 将各类别 *AP* 值绘制为图 11. 从图 11 中可以看出 attend_a_lecture (听课) 类别的准确率由模型 5 达到了最高的 97.2%, 比最初的模型 1 提高了 2.8 个百分点. Looking_book (看书) 类别模型 1 的准确率为 83.1% 比较低, 是因为看书与书写这两种行为有比较强的相似性, 区别在于手部的细节动作, 比较难以区分, 当加入 non-local 和 TSM 模块后准确率达到了 93.5%, 提升了 10.4 个百分点. 对于 no_studying (不学习)、playing_phone (玩手机) 与 writing (书写) 这 3 个类别分别提升了 4.7、1.7 与 3.7 个百分点.

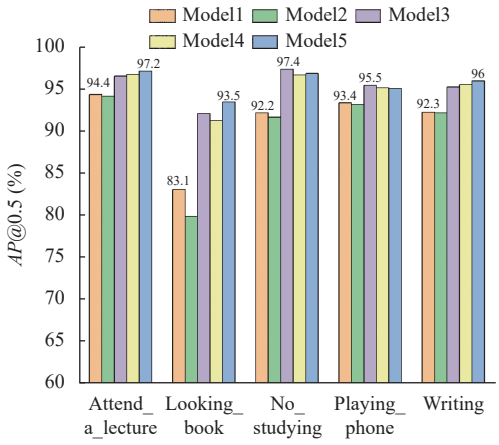


图 11 各类别 *AP* 值

为了进一步直观地展示改进方法的有效性, 图 12 展示了原始 YOWO 网络和改进后网络在学生课堂行为数据集上的检测结果. 其中第 1 行代表在两种不同的场景下 YOWO 网络上的检测结果、第 2 行代表在 YOWO 中加入 TSM 模块的检测结果、第 3 行代表在 YOWO 中加入 TSM 和 non-local 模块的检测结果. 通过对比可以发现, YOWO 所对应的检测结果准确率相对较低; YOWO+TSM 相对于 YOWO, 预测准确率明显提升; YOWO+TSM+non-local 中 non-local 的加入使得部分学生行为的准确率有了小幅度提升.

(a) Model1: YOWO

(b) Model3: YOWO+TSM

(c) Model4: YOWO+TSM+non-local

图 12 检测效果对比

5 结论与展望

本文提出了一种新的改进 YOWO 的学生课堂行为识别算法, 通过采用 TSM 和 non-local 模块, 让算法能有效利用视频的时间序列特征, 并准确识别出 5 类典型的学生课堂行为. 运用消融实验对改进后 YOWO 算法进行了验证, 结果表明, 该算法对教室环境下的学生行为进行了识别, 平均识别正确率为 95.7%. 未来的研究可以进一步扩大数据的多样性, 通过增加学生行为种类以及每种行为的数量, 以提高模型的鲁棒性和通用性. 此外, 为了提高网络的实时性, 还可以对网络进行轻量化处理, 从而达到对学生上课行为进行实时识别的目的. 从而更好地促进对课堂教学信息的实时分析, 并对课堂教学策略进行优化.

参考文献

1 周鹏霄, 邓伟, 郭培育, 等. 课堂教学视频中的 S-T 行为智

能识别研究. 现代教育技术, 2018, 28(6): 54–59. [doi: 10.3969/j.issn.1009-8097.2018.06.008]
2 党冬利. 人体行为识别及在教育录播系统中的应用 [硕士学位论文]. 西安: 西安科技大学, 2017.
3 廖鹏, 刘宸铭, 苏航, 等. 基于深度学习的学生课堂异常行为检测与分析系统. 电子世界, 2018(8): 97–98. [doi: 10.19353/j.cnki.dzsj.2018.08.054]
4 秦道影. 基于深度学习的学生课堂行为识别 [硕士学位论文]. 武汉: 华中师范大学, 2019.
5 蒋沁沂, 张译文, 谭思琪, 等. 基于残差网络的学生课堂行为识别. 现代计算机, 2019(20): 23–27. [doi: 10.3969/j.issn.1007-1423.2019.20.005]
6 周叶. 基于 Faster R-CNN 的小学生课堂行为检测研究 [硕士学位论文]. 成都: 四川师范大学, 2021.
7 杜俊强. 基于深度学习的学生课堂行为识别方法研究 [硕士学位论文]. 烟台: 山东工商学院, 2022.
8 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. 计算机工程与设计, 2021, 42(10): 2924–2930. [doi: 10.16208/j.issn1000-7024.2021.10.030]
9 孙绍涵, 张运楚, 王超, 等. 基于深度学习的学生课堂注意力评价. 计算机系统应用, 2022, 31(6): 307–314. [doi: 10.15888/j.cnki.csa.008553]
10 林灿然, 许伟亮, 李逸. 基于多模态数据的课堂学生行为识别技术的探究. 现代计算机, 2020, (6): 69–75. [doi: 10.3969/j.issn.1007-1423.2020.06.015]
11 郭俊奇, 吕嘉昊, 王汝涵, 等. 深度学习模型驱动的师生课堂行为识别. 北京师范大学学报 (自然科学版), 2021, 57(6): 905–912. [doi: 10.12202/j.0476-0301.2021207]
12 郑丹. 基于双流卷积神经网络的学生课堂行为识别研究 [硕士学位论文]. 沈阳: 沈阳师范大学, 2022.
13 Köpüklü O, Wei XY, Rigoll G. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. arXiv:1911.06644, 2019.
14 Wu T, Cao MQ, Gao ZT, *et al*. STMixer: A one-stage sparse action detector. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 14720–14729. [doi: 10.1109/CVPR52729.2023.01414]
15 Chen SF, Sun PZ, Xie EZ, *et al*. Watch only once: An end-to-end video action detection framework. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 8158–8167.
16 Chi HG, Lee K, Agarwal N, *et al*. AdamsFormer for spatial action localization in the future. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 17885–17895.
17 Lin J, Gan C, Han S. TSM: Temporal shift module for efficient video understanding. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 7082–7092.
18 Yang CY, Xu YH, Shi JP, *et al*. Temporal pyramid network for action recognition. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 588–597.
19 Feichtenhofer C. X3D: Expanding architectures for efficient video recognition. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 200–210. [doi: 10.1109/CVPR42600.2020.00028]
20 Wang LM, Tong Z, Ji B, *et al*. TDN: Temporal difference networks for efficient action recognition. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 1895–1904.
21 Wang XL, Girshick R, Gupta A, *et al*. Non-local neural networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7794–7803.

(校对责编: 孙君艳)