融合多模态深度学习的高校课堂行为感知与教学优化策略研究

初奕萱

天津外国语大学英语学院，天津市，300204；

摘要：高校课堂教学面临行为感知手段单一、教学反馈滞后、课堂互动数据挖掘不足等问题。为此，提出一种基于多模态深度学习与改进 YOLOv5-TAM 模型的课堂行为感知与教学优化系统。系统整合课堂视频、语音、文本等多源数据，通过改进的目标检测网络实现学生面部表情、身体姿态等视觉特征的精准提取，采用 Transformer 结构完成图像、音频、文本三类特征的时序对齐与深度融合，实时识别学生注意力水平、情绪状态及互动活跃度。在此基础上，构建基于 TOPSIS 的教学质量综合评价模型，设计可视化反馈界面与教学策略推荐引擎。真实课堂环境测试表明，系统行为识别平均精度达到 84%以上，反馈延迟低于 2 秒，显著提升了教师对课堂状态的感知能力与教学调控的实时性。本研究突破传统单一模态评价的局限，为高校智慧课堂建设提供了可复用的技术方案与实践参考。

关键词：多模态融合，深度学习，课堂行为感知，YOLOv5-TAM，教学优化策略

DOI：10.69979/3029-2735.26.05.028

传统课堂教学质量评估多依赖课后问卷调查或教师主观判断，缺乏对课堂实时行为数据的有效采集与量化分析。学生注意力分散、情绪低落、互动不足等问题往往在课后才被发现，教师难以在教学过程中及时调整策略。随着深度学习与多模态融合技术的发展，基于计算机视觉的课堂行为分析逐步成为研究热点，但现有方案多局限于单一视频模态，忽略语音、文本等信息对理解学生状态的补充价值。同时，传统循环神经网络在处理长时序课堂行为数据时存在梯度消失问题，难以捕捉学生状态变化与教学环节之间的长距离依赖关系。因此，设计融合图像、音频、文本多模态数据并引入注意力机制的深度学习系统，实现对课堂行为的实时感知与教学策略的动态优化，对于提升高校教学质量具有重要的理论价值与现实意义。

1 高校课堂行为感知与教学优化的技术痛点分析

高校课堂环境具有人员密集、行为多样、干扰因素复杂等特点，现有教学辅助系统在行为感知全面性、反馈及时性与模型泛化能力方面存在明显不足。

数据模态单一导致感知维度缺失。 当前多数课堂分析系统仅依赖监控视频，通过检测学生头部姿态、面部表情判断注意力水平。然而，单纯视觉信息无法反映学生的情绪波动（如焦虑、疲惫）和参与意愿（如发言频率、语调变化）。例如，一个学生可能坐姿端正但心不在焉，仅凭图像难以准确识别其“伪听讲”状态。同时，课堂中的师生互动内容、教学 PPT 文本等语义信息被完全忽略，导致系统缺乏对教学情境的理解能力。

教学反馈滞后，缺乏实时干预手段。 传统教学评价属于“课后总结式”，教师通常在课程结束后才能通过作业或问卷了解学生掌握情况。对于学生注意力集体下降、互动沉默等课堂危机，教师无法在黄金干预窗口内获得提醒，只能凭借个人经验临时调整，优化效果因人而异。

模型泛化能力与实时性不足。 现有深度学习模型在实验室受控环境下表现良好，但迁移到真实课堂后，面临光照变化（如阴天、投影光干扰）、学生遮挡（举手、低头、前排遮挡后排）、摄像头视角局限等问题，识别精度显著下降。此外，部分复杂模型计算开销大，无法在普通教学硬件上实现实时处理，难以大规模推广。

上述痛点制约了智能课堂分析系统的实际应用效果，亟需构建一种融合多模态数据、具备长时序建模能力且满足实时性要求的解决方案。

2 多模态融合与深度学习模型的系统构建

针对上述技术痛点，从数据采集层、特征提取层、

模型感知层与反馈优化层四个维度，构建基于多模态深度学习的高校课堂行为感知与教学优化系统，实现从数据整合到教学干预的全流程闭环。

2.1 多模态数据采集与预处理

系统依托教室部署的高清摄像头、定向麦克风及教学终端，同步采集三类数据：视频数据（学生面部表情、身体姿态、视线方向）、音频数据（教师讲授、学生发言的语音及语调变化）、文本数据（PPT 内容、课堂互动记录）。针对不同模态特点设计预处理策略：视频数据执行帧提取、去噪及人体姿态关键点检测；音频数据采用 VAD（语音活动检测）与梅尔频谱提取；文本数据进行分词与语义清洗。所有预处理后的数据统一转换为时序对齐的特征向量，为后续融合奠定基础。

2.2 改进型 YOLOv5-TAM 单模态特征提取

图像模态是课堂行为识别的核心。项目在 YOLOv5 骨干网络中引入 Triplet Attention 三重注意力机制，提出改进型 YOLOv5-TAM 模型。该机制通过三个并行分支捕获通道、高度和宽度维度上的交互信息，显著提升了对小目标（如微表情、视线偏移）和遮挡目标的检测能力。同时结合 Mosaic 数据增强与自适应锚框优化，模型在课堂复杂环境下的检测平均精度达到 84.68%，实时帧率超过 60 帧/秒。音频模态采用卷积递归神经网络（CRNN）提取语音时序特征，文本模态基于中文 BERT 模型获取教学内容语义向量。

2.3 基于 Transformer 的跨模态融合机制

为实现图像、音频、文本三类特征的深度整合，系统设计了基于注意力机制与 Transformer 结构的多模态融合模型。首先通过线性投影将不同模态的特征映射到统一的嵌入空间；然后利用 Transformer 编码器中的自注意力层，计算不同时刻、不同模态特征之间的关联权重，自动学习各模态在识别特定行为（如注意力集中、情绪低落、主动互动）时的贡献度。例如，识别“主动回答”时，系统提升音频模态（发言内容）和文本模态（互动记录）的权重；识别“疲劳分神”时，则强化图像模态（眨眼频率、头部下垂）与生物体征替代指标（通过音频分析语速变化）的权重。融合后的特征向量输入分类器，输出学生个体及班级整体的实时状态（注意力高/中/低、情绪积极/中性/消极、互动活跃度等）。

2.4 智能教学反馈与优化策略

基于多模态感知结果，系统开发了教师端可视化界面，以动态仪表盘、热力图、趋势曲线等形式实时展示班级注意力分布、情绪状态占比、互动频率等关键指标，并设置预警功能——当整体注意力低于阈值或情绪倾向消极时，界面自动闪烁提醒。同时，系统内置策略推荐引擎，根据当前行为状态与历史数据，为教师生成个性化建议（如“建议增加提问互动”“当前讲解速度偏快，可适当停顿”）。此外，构建基于 TOPSIS 多指标决策模型的综合教学质量评价体系，融合客观行为数据（注意力变化、参与度）与主观反馈（学生满意度问卷），生成阶段性教学分析报告，辅助教师进行课后反思与课程改进。

3 系统的技术验证与效能优化

系统开发完成后，在某合作高校的 2 个班级真实课堂环境中进行了为期 4 周的应用测试，覆盖英语翻译、综合英语等课程，累计采集 80 余课时的多模态数据。所有测试数据均来自公开课堂采集或经授权的教学观摩活动，不涉及特定学校的隐私信息。

功能验证。 测试期间，系统实时识别学生行为的平均精度达到 82.7%（接近实验室环境的 84.68%），其中注意力状态识别准确率为 85.3%，情绪识别准确率为 79.6%。系统从数据采集到反馈界面刷新的平均延迟为 1.8 秒，满足实时性要求。在光照变化、部分遮挡等干扰条件下，YOLOv5-TAM 模型相比原始 YOLOv5 精度提升约 6 个百分点，证明了注意力机制的有效性。教师反馈显示，可视化界面能够帮助其快速定位“注意力洼地”区域，并根据预警提示及时调整教学节奏。

性能优化。 针对测试中发现的复杂光照（如投影关闭瞬间）、学生大幅度动作导致目标丢失等问题，采取了以下优化措施：在数据预处理阶段引入自适应直方图均衡化与目标跟踪算法（DeepSORT），提升遮挡情况下的 ID 保持能力；对 Transformer 模型实施剪枝与量化，参数量减少约 30%，推理速度提升至实时要求以上；针对不同课程类型（理论课、互动课），设计模型参数动态调整机制，在互动课中提升音频与文本模态的融合权重。

模型迭代与部署。 建立增量学习机制，将每周采集的新标注数据用于模型微调，持续提升泛化能力。最终系统以轻量化形式部署在教室边缘服务器上，实现数据本地化处理，无需上传云端，保障了教学数据隐私。

所有数据采集均获参与者知情同意，并进行匿名化处理，符合科研伦理规范。

4 结语

融合多模态深度学习的高校课堂行为感知与教学优化系统，有效突破了传统单一模态评价的局限，实现了对学生注意力、情绪、互动行为的多维实时感知与教学策略的动态反馈。系统创新性地提出改进型 YOLOv5-TAM 模型，提升了复杂课堂环境下的检测精度；引入 Transformer 结构完成跨模态特征融合，挖掘了行为状态与教学环节之间的时序关联。真实课堂验证表明，系统具备良好的实时性、抗干扰能力与教师接受度，为智慧课堂建设提供了可复用的技术路径。未来可进一步拓展数据源，引入可穿戴设备（如心率手环）实现生理信号与行为的关联分析；探索基于联邦学习的跨校模型协同训练，在保护数据隐私的前提下提升模型泛化能力；开发学生端个性化学习状态提醒功能，推动课堂教学从“教师中心”向“学习者中心”的深度转型。

参考文献

[1]杨帆,等.基于 YOLOv7 与多模型融合的学生课堂行为检测系统[J].计算机工程与应用,2023,59(15):123-130.

[2]郑周杰,等.多视角课堂行为识别与注意力评估方法研究[J].现代教育技术,2023,33(4):56-63.

[3]张勇和,等.基于多模态融合与自然感知的学生兴趣建模[J].电化教育研究,2022,43(8):88-95.

[4]Liu T, et al. Student behavior detection in classroom video based on YOLOv3 with DropBlock[J]. Journal of Intelligent & Fuzzy Systems, 2020, 39(4): 5211-5221.

[5]Zheng W, et al. Automated multi-modal teaching behavior analysis framework for large-scale classroom monitoring[J]. IEEE Transactions on Learning Technologies, 2024, 17: 456-469.

[6]Zhao H, et al. BiTNet: A lightweight transformer-based network for real-time classroom behavior feedback[J]. Pattern Recognition, 2023, 138: 109382.

注：本文基于天津市大学生创新训练计划项目“融合多模态深度学习的高校课堂行为感知与教学优化策略研究”成果撰写，由天津外国语大学创新创业训练计划项目资助。

“天津外国语大学创新创业训练计划项目资助”