Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2021, 10(2), 123-132
Published Online May 2021 in Hans. http://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2021.102013

Hans汉斯

基于深度学习的学生课堂状态检测算法与应用

史　雨[1]，辛　宇[1]，袁　静[1]，欧阳群波[1]，杨德贺[2]，袁国铭[1]

[1]防灾科技学院，河北　廊坊
[2]国家自然灾害防治研究院，北京
Email: 303071589@qq.com

收稿日期：2021年4月9日；录用日期：2021年4月23日；发布日期：2021年5月26日

摘　要

课堂教学评价是教学管理的重要组成部分之一，但依赖于督导教师完成该项工作的管理形式难以全面评价并反馈学生课堂学习状态。同时，我国高校的课堂监控视频数据被大量搁置未发挥作用。基于此，本文将传统教学管理与人工智能有机结合，提出学生课堂学习状态智能检测算法，采用K-means++聚类算法对目标候选框的个数和宽高比进行聚类分析，搭建双YOLO网络模型对课堂监控视频中学生的课堂行为进行分析，实时、精准地反馈学生的课堂学习状态，并对结果进行评分分级辅助督导教师进行课堂教学评价任务以提高教学管理效率。经过测试实验，本章提出的双YOLO网络模型的准确率为86.62%，且每帧教室监控图像的计算时间0.2 s。

关键词

人工智能，学生课堂学习状态检测，双YOLO网络，K-means++

Deep Learning Based on Student Classroom Status Detection Algorithm and Application

Yu Shi[1], Yu Xin[1], Jing Yuan[1], Qunbo Ouyang[1], Dehe Yang[2], Guoming Yuan[1]

[1]Institute of Disaster Prevention, CIDP, Langfang Heibei
[2]Institute for Natural Disaster Prevention and Control, ICP, Beijing
Email: 303071589@qq.com

Received: Apr. 9th, 2021; accepted: Apr. 23rd, 2021; published: May 26th, 2021

Abstract

Classroom teaching evaluation is one of the important parts of teaching management, but it is difficult to comprehensively evaluate and feedback students' classroom learning status in the management form that relies on the supervisor to complete the work. At the same time, the classroom surveillance video data of colleges and universities in China has been largely shelved and has not

文章引用: 史雨, 辛宇, 袁静, 欧阳群波, 杨德贺, 袁国铭. 基于深度学习的学生课堂状态检测算法与应用[J]. 人工智能与机器人研究, 2021, 10(2): 123-132. DOI: 10.12677/airr.2021.102013

played a role. Based on this, this article will combine traditional teaching management and artificial intelligence, students' classroom learning intelligent detection algorithm is put forward, by the method of K-means++ clustering algorithm, the number and the aspect ratio of target candidate box for clustering analysis, build the double YOLO network model for monitoring video classroom of students classroom behavior analysis, real-time, accurate feedback on students' classroom learning state, and the results are teachers' classroom teaching evaluation rating auxiliary supervision task to improve the efficiency of teaching management. After testing and experiments, the accuracy rate of the TWO-STAGE YOLO network model proposed in this chapter is 86.62%, and the calculation time of each frame of classroom monitoring image is 0.2 s.

Keywords

Artificial Intelligence, Student Classroom Learning State Detection, TWO-STAGE YOLO Network, K-means++

1. 引言

国务院于 2012 年 10 月 1 日实施了《教育督导暂行规定》，明确将教育教学水平、教育教学管理等教育教学工作情况纳入教育督导工作范围，为高校建立和完善教学质量保障与监控体系提供了重要依据。通常，教学评价需要涉及整个教育教学工作：对学生的知识掌握程度的评估，对课堂学生学习投入度专注度参与率的考察，对教师教学水平、教学方式和教学过程的评价。目前，高校课堂教学质量的评价主要由教学督导工作组[1]完成，督导评价结果也成为了衡量和提升教学质量的核心内容。但是依赖于督导组进行课堂教学评价现阶段仍存在以下 2 个方面的问题：

1) 课堂教学评价数据客观性不强：一方面，针对某门课程的课堂评价通常是由督导组随机选择两次听课的时间段，通过课堂旁听的方式综合老师的授课内容以及学生的课堂状态进行现场评价，未进行其他课时的课堂评价，评价参考数据不全。

2) 课堂教学评价任务繁重：由于督导老师人数有限，学生课时数量巨大，完全依赖于督导老师对全部的课堂质量进行评价是极具挑战性的。

综上，仅依赖于督导老师的课堂教学质量评价难以全面并实时反馈课堂学生学习状态，利用人工智能技术对海量的监控视频数据进行智能评价是教学管理亟需的一项现代教育信息技术。

近年来，深度卷积网络在计算机视觉和模式识别中得到了广泛的应用，例如首先生成一系列的候选区域，然后再进行精确分类的 Fast R-CNN [2]、Faster R-CNN [3]等，然而以上方法由于存在重复卷积、计算开销大问题难以用于实时的场景检测；相反，以 SSD [4]和 YOLO [5] [6] [7]三个版本为代表的单阶段检测算法，可以从原始图像直接进行目标预测，在速度和精度上大大提升，其在人脸识别[8]、车牌号识别[9]等目标检测领域得到了广泛的应用，其中 YOLOv3 的精度和 SSD 相当，但速度要快上 3 倍，更适用于对实时性要求高的场景。

目前基于课堂监控视频的研究主要是针对课堂人数统计[10]和人脸识别考勤[11]，未涉及对学生课堂状态的检测，本文在 YOLOv3 网络的基础上，提出基于高校监控视频的学生课堂学习状态智能检测算法，其创新点主要体现在以下几个方面：

1) 首次将人工智能技术应用于海量的课堂监控视频中，实时检测学生的课堂状态并对其检测结果进行统计分析，为课堂教学质量评价提供参考；

2) 在课堂监控视频中，单个 YOLOv3 网络模型检测的后排学生课堂学习状态无法达到很好的检测效果。造成该结果主要有三个原因：一是后排的学生由于距离原因动作轮廓比较模糊，二是监控设备像素有限无法保证教室后方学生行为的清晰度，三是在教室人数较多的前提下，学生过于密集造成的部分遮挡。针对以上的图像多尺度问题，本算法设计了双 YOLOv3 神经网络算法，先训练人体定位检测模型用于检测学生的位置，在尽可能较全定位教室人体位置后，在人体定位检测模型的基础上对相应位置的学生做二次检测，进一步检测学生的行为，同时使用 K-means++聚类算法对标注的训练集数据进行聚类分析，然后预测出学生的课堂学习状态，在提高检测精度的结果上改进图片中远景角度的学生漏检问题。

2. 学生课堂学习状态智能检测算法概述

2.1. K-means++聚类算法

YOLOv3 [7]在对预测框进行预测的时候，借鉴了 Faster R-CNN 中的先验框思想。先验框是一组可以通过人工标注得到的宽高固定的初始预测框，不同的数据集聚类得到的先验框的个数和宽高各不相同。Yolov3 使用了 K-means 聚类算法对数据集中的样本框进行聚类分析，得到最有可能的样本框的形状，但是 K-means 聚类算法的结果受聚类初始化中心点选取的影响较大。因此本文采用 K-means++聚类算法对学生课堂视频数据集进行了聚类分析，从输入的数据集中随机选取一个点作为第一个中心点，以初始中心点彼此尽可能远离为原则，减少了初始化聚类中心点选取对聚类结果的影响，聚类得到的先验框更贴近数据集标注的目标框。

$$\mathrm{d}(\mathrm{box},\mathrm{centroid}) = 1 - \arg\max \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_k} IOU(\mathrm{box},\mathrm{centroid})}{n} \tag{1}$$

公式(1)中，box 表示样本，centorid 表示簇的中心，表示第 k 个聚类中心样本数，n 表示总样本数。表示簇的中心矿和聚类框的交并比，IOU 越大，距离越小。

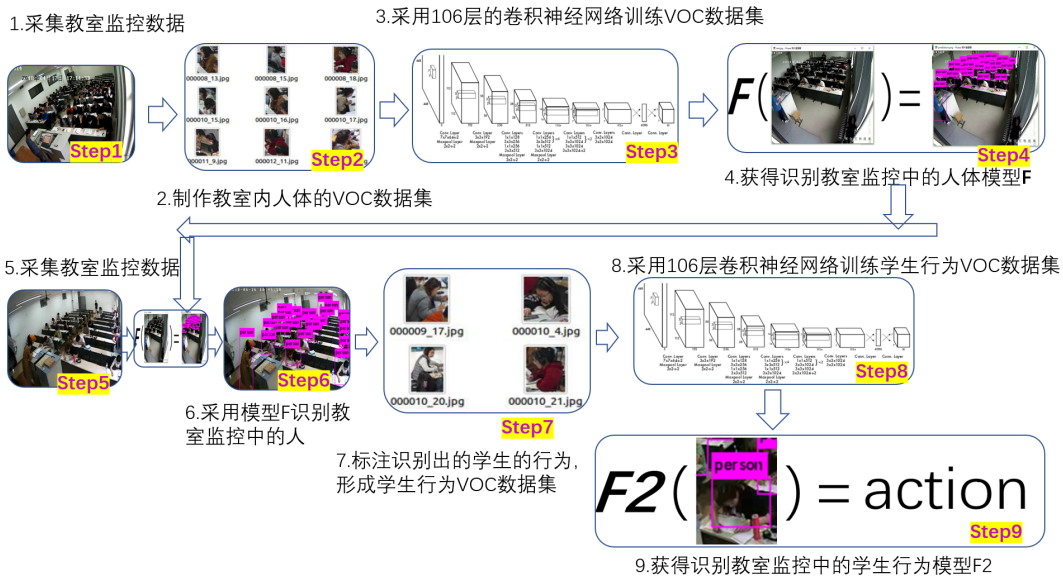


Figure 1. Training flow chart of dual YOLOV3 network model
图 1. 双 YOLOv3 网络模型训练流程图

Figure 2. Dual YOLOV3 network model test flow chart
图 2. 双 YOLOv3 网络模型测试流程图

按照上述方法，在学生课堂状态数据集中使用 K-means++算法重新对标注信息进行聚类，得到的 9 组 anchor 值为：(10,13)、(16,30)、(33,23)、(30,61)、(62,45)、(59,119)、(116,90)、(156,198)、(344,319)。

2.2. 双 Yolov3 网络模型

YOLO 是一个基于卷积神经网络(CNN)的实时物体检测算法。神经网络把图片分成不同的区域，然后给出每个区域的边框预测和概率，并依据概率大小对所有边框分配权重。作为目前最先进的目标检测算法，YOLO v3 主要经过了三次迭代，检测效果更加准确，功能也更强大，更实用。

针对运用 yolo 算法进行学生课堂行为识别以达到检测教学质量这一目的，本文将学生的课堂学习状态分为五类：正常听课(Normal)，阅读和记笔记(Read)，玩手机(Phone)，睡觉(Sleep)，左顾右盼等其他未正常听课行为(Abnormal)。算法包括课堂学生人体定位模型和学生课堂学习状态识别模型，其整体流程分为训练和测试两个阶段，主要流程如图 1、图 2 所示：在训练阶段，首先通过图像采集系统采集课堂监控视频数据，分别对人体定位和学生课堂学习状态识别两个特征进行数据标注，然后搭建双 YOLOv3 网络并设置网络参数，将标注的数据集输入到双 YOLOv3 网络，最后训练人体定位模型和学生课堂学习状态识别模型。在测试阶段，首先使用人体定位模型检测出学生的位置，然后使用学生课堂学习状态识别模型在其检测出的学生位置预测框中进一步识别学生的学习状态，最后统计分析检测结果并存入数据库，有效改进原始 YOLOV3 网络模型漏检率高的问题。

3. 实验与分析

3.1. 训练数据集介绍

考虑到不同教室的摄像头角度、教室中学生座位疏密及学生每节课的座位不同等多种因素可能会导致学生课堂学习状态识别模型训练测试结果造成偏差，本研究设置以下数据采集方案：

1) 采集同样规模教室的两个月时间间隔内学生上课时间段的监控视频，尽量选择摄像头视角较为清晰，教室上课人数不算过于密集且学生人员位置不固定的教室。

2) 人工筛选出了 1000 张具有代表性的样本分布相对均匀的学生课堂学习状态数据集。

3) 使用标注工具 labelImg 对图片进行标注，对象类别信息和位置信息在标注过程中会被保存到对应的 xml 文件下。

图 3 是数据集标注样本框的中心点的分布散点图，数据集标注样本框的中心点主要分布在图片长宽

比例的 0.2~0.8 之间。

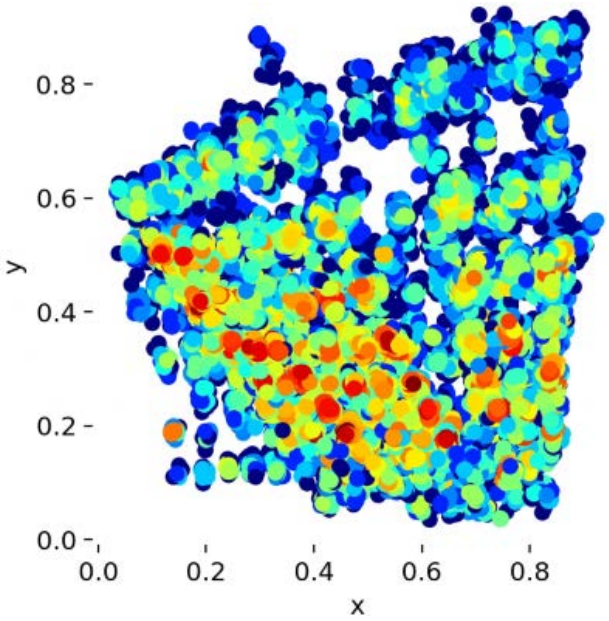


Figure 3. Scatter diagram of center points of the labeled sample box
图 3. 标注样本框的中心点的分布散点图

3.2. 训练环境和训练参数

本文数据集的训练基于 Windows10 64 位系统，使用 Nvidia GeForce RTX2080Ti 显卡、CUDA10.0 和 cuDNN 调用 GPU 加速训练。图像训练前被降采样到 416 像素 × 416 像素，训练中每轮迭代(Iteration)的样本数量(batch)为 64；为了节省内存，每一轮样本又被平均分成 16 组(subdivision)输入网络参与训练前向传播。batch 越大，每次训练提取的样本特征越多，训练效果越好，subdivision 越大，内存占用越小。为了防止模型训练出现过拟合，本研究将模型初始学习率(Learning rate)设置为 0.001，并在迭代 2000 次、4000 次时衰减 10 倍以减少 loss 波动。通过公式(2)计算卷积核个数(filters)到人体定位模型的卷积核大小为 18，学生课堂学习状态识别模型的卷积核大小为 30。

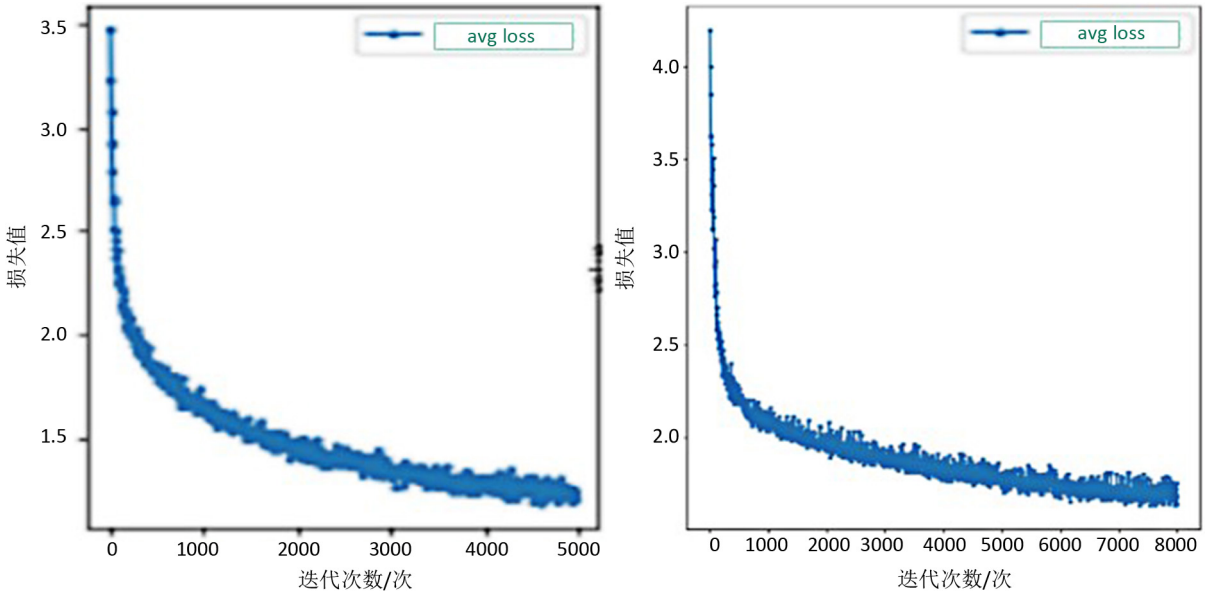
$$\text{filters} = \text{filters} = 3\times\left(5+\text{num}\left(\text{classes}\right)\right) \tag{2}$$

由图 4 所示，图 4(a)和图 4(b)分别为人体定位模型和学生课堂学习状态识别模型训练的平均损失函数图，随着训练迭代次数的增加，损失值逐渐减少，趋势逐渐平稳，当平均损失值在 1.0 上下浮动，达到理想的训练效果。因此本研究选取迭代 5000 次的模型作为最终的人体定位模型。选取迭代 8000 次的模型作为最终的学生课堂学习状态识别模型。

3.3. 评价指标

1) 混淆矩阵实验及结果

混淆矩阵是用来总结一个分类器结果的矩阵。用于观察模型在各个类别上的表现，可以计算模型对应各个类别的准确率、漏检率、误报率。本研究混淆矩阵评价的符号定义如表 1 所示，$Y=0$ 表示样本的实际类别为负例，$Y=1$ 表示样本的实际类别为正例；$\hat{Y}=0$ 表示模型预测的结果为负例，$\hat{Y}=1$ 表示模型预测的结果为正例。



(a) 人体定位模型平均损失函数图 (b) 学生课堂学习状态识别模型平均损失函数图

Figure 4. Model loss function image
图 4. 模型损失函数图像

Table 1. Confusion matrix
表 1. 混淆矩阵

		模型预测的结果	
		$\hat{Y}=1$	$\hat{Y}=0$
样本的实际类别	$Y=1$	真正例(True Positive, TP)	假负例(False Negative, FN)
	$Y=0$	假正例(False Positive, FP)	真负例(True Negative, TN)

Table 2. Test results of original YOLO network model and double YOLO network model
表 2. 原始 YOLO 网络模型和双 YOLO 网络模型的测试结果

	原始 YOLO 网络模型			双 YOLO 网络模型		
	预测结果	样本总数	所占比例	预测结果	样本总数	所占比例
准确识别的数量	14,521	22,534	64.44%	19,518	22,534	86.62%
误报的数量	2995	22,534	13.29%	507	22,534	2.25%
漏检的数量	5018	22,534	22.27%	2509	22,534	11.13%

其中，NTP 表示样本的真实类别是正例，其模型的预测结果也是正例的真正例样本数量；NTN 表示样本的真实类别是负例，其模型将其预测成正例的真负例样本数量；NFP 表示样本的真实类别是负例，但模型将其预测成为正例的假正例样本数量；NFN 表示样本的真实类别是正例，模型将其预测成为负例的假负例样本数量。三种评价指标的定义如下：

准确率 = NTP/(NTP + NFP)：被正确分类的学生学生课堂学习状态样本比例；

误报率 = NFP/(NFP + NTP)；被错误分类的学生学生课堂学习状态样本中假正例的比例；

漏检率 = NFN/(NFP + NTP)；被错误分类的学生学生课堂学习状态样本中假负例的比例。

测试数据中含有 500 张教室监控图片，内含 22,534 条学生课堂行为记录，人体定位模型的检测效果

如图 5 所示，原始 YOLO 模型和双 YOLO 模型的测试对比结果如表 2 所示，教室学生课堂学习状态检测结果图对比如图 6 所示。

分析表 2 得，使用双 YOLO 网络的学生课堂学习状态智能检测算法的准确率为 86.62%，漏检率为 11.13%，误报率为 2.25%，相较于原始 YOLO 网络算法的准确率为 64.44%，漏检率为 22.27%，误报率为 13.29%，本文算法的准确率更高，漏检率更低。

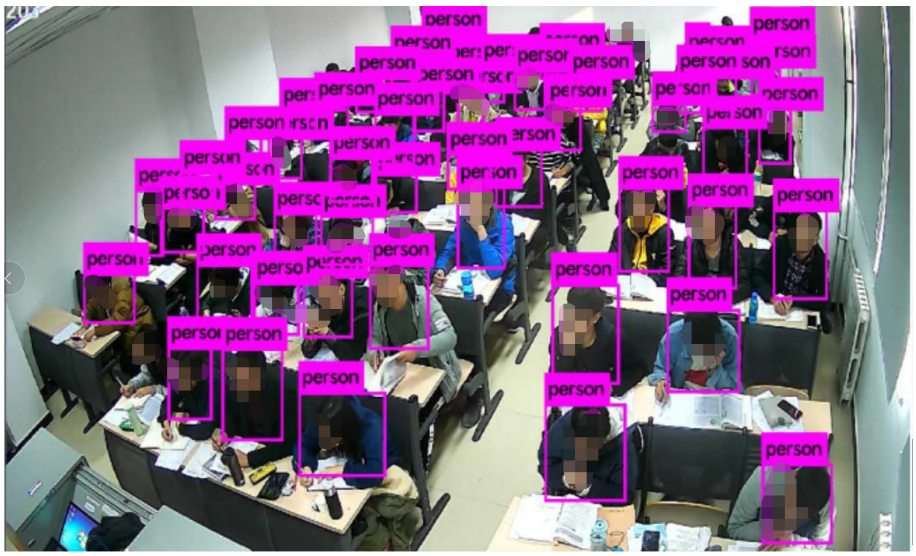


Figure 5. Detection effect of human body positioning model
图 5. 人体定位模型的检测效果

(a) 原始 YOLOV3 算法　　(b) 本文算法

Figure 6. Test results of classroom learning status of students
图 6. 教室学生课堂学习状态检测结果图

4. 学生课堂状态智能检测系统设计

本文在基于双 YOLO 网络的学生课堂学习状态智能检测算法上，设计了学生课堂状态智能检测系统，通过训练好的 YOLOv3 模型检测出教室的学生人数和学生的学习状态，面向教学管理人员，软件操作流程精简，减轻教学管理人员的负担，对学生的出勤率和课堂关注度等情况作出评价。

学生课堂状态智能检测系统主要包括 3 大功能：预览课堂监控功能，学生人体定位检测功能和学生课堂学习状态检测功能，在加载模型后，可以对课堂监控视频进行实时的学生人数统计和学生状态识别，图 7 为本系统的流程图：加载学生课堂状态智能检测模型后，首先进行课堂监控视频的选择与预览，接

着进行学生人体定位检测或学生课堂学习状态检测两种功能选择，最终根据课堂监控视频，实时显示可视化结果并进行相关统计。

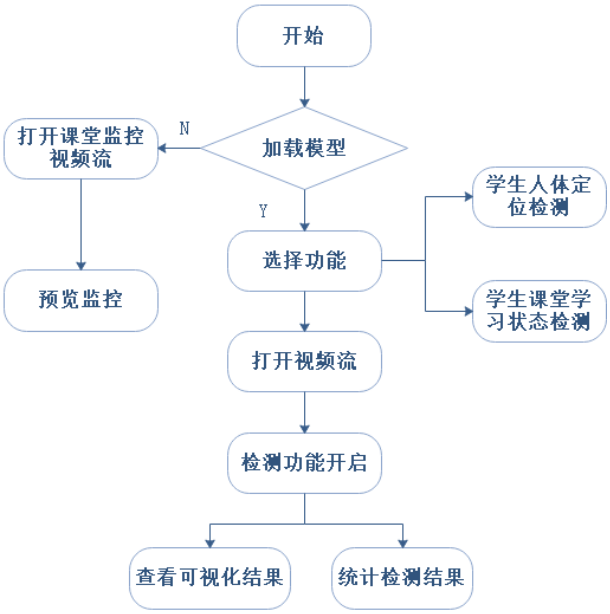
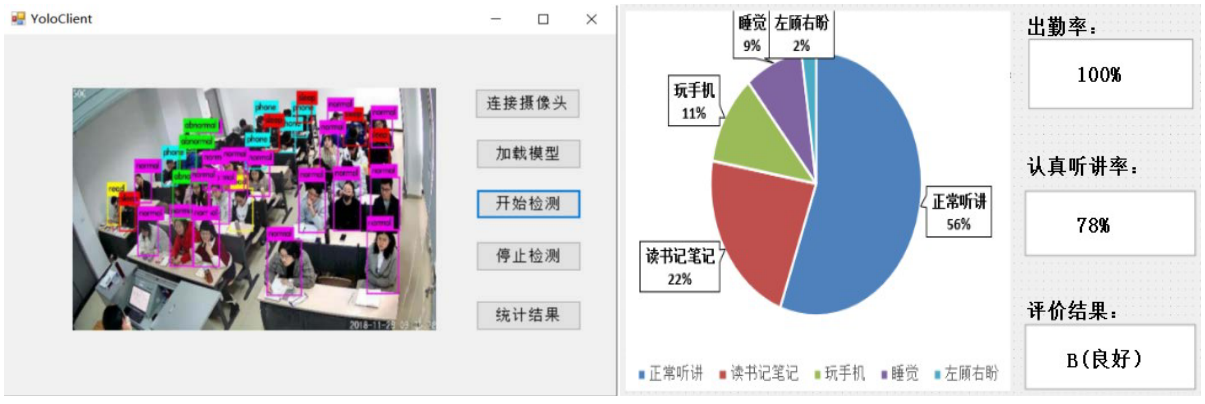


Figure 7. Flowchart of students' classroom status intelligent detection system
图 7. 学生课堂状态智能检测系统流程图

Table 3. Classroom evaluation score evaluation criteria and situation analysis
表 3. 课堂评价得分评判标准及情况分析

课堂评价等级	课堂评价得分	情况分析
A (优秀)	≥85	班级 85%以上学生都在认真听讲或记笔记，学生精力集中，课堂专注度高
B (良好)	≥70	班级 70%以上学生在认真听讲或记笔记，学生课堂专注度良好
C (一般)	≥50	班级仅有超过 50%的学生在认真听讲或记笔记，学生课堂专注度一般
D (不佳)	<50	班级学生认真听讲或记笔记人数不到总人数的 50%，学生课堂专注度不佳



(a) 用户界面　　(b) 课堂评价结果

Figure 8. Display of application user interface and classroom evaluation results
图 8. 应用程序用户界面展示及课堂评价结果展示图

表 3 中所提到的课堂评价得分结合了权重为 80%的认真听课率与权重为 20%的课堂出勤率：其中认真听讲率为双 YOLO 网络模型检测出的学生课堂中正常听讲和阅读记笔记两种学习状态占总体学习状态的比例；课堂出勤率为通过人体定位模型检测出的学生人数占班级实际人数的比例。通过划分课堂评价得分标准，本研究将学生课堂状态从班级大多学生注意力集中、专注度高的课堂状态到班级大部分学生注意力不集中、专注度的课堂状态低分为 A (优秀)，B (良好)，C (一般)，D (不佳)四个等级。自习课的学生学习状态评判同理。其中在 C (一般)，D (不佳)两个等级中，根据实际上课情况和老师教学任务的安排，如出现 phone (玩手机)、abnormal (左顾右盼)等学生课堂状态大于 40%的情况可能为教师在课堂上发布了雨课堂签到答题、学生小组学术讨论活动等教学任务，此时应以实际情况确定最终标准。

如图 8(a)所示为在一段实时监测的教室人数为 45 人的监控视频中截取的应用程序用户界面展示，通过课堂评价结果展示图 8(b)可看出该堂课的出勤率为 100%，认真听讲率为 78%，学生课堂状态为 B(良好)等级，正常听课及阅读约超过课堂总人数的 70%，学生在课堂上精力较为集中，课堂专注度良好。

5. 结论

本文利用人工智能领域中的计算机视觉技术，创新搭建双 YOLO 网络模型并结合 K-means++聚类算法对课堂监控视频中学生的课堂状态进行检测和分析，有效提高了模型检测结果的准确率。但是本研究提出的项目还存在一些待改进之处，如教室内部人数过多造成的拥挤现象导致学生行为(尤其是手部动作特征)被遮挡住，导致神经网络模型的识别准确率较低。本文算法可以充分利用课堂视频监控资源进行学生人体定位和学生课堂学习状态的实时检测，并由学生课堂状态智能检测系统通过课堂评分向高校老师及领导准确实时反馈课堂及自习课的学生学习状态，打破高校监控视频被机械记录并搁置的现状，推动了人工智能技术在高校教学服务中的应用，提高了教学管理的效率各个高校的教学，实现了共同构建融合信息技术的高校人才培养体系的目的。

基金项目

中央直属高校基本科研业务经费 ZY20180122 资助。

参考文献

[1] 王莉芬, 杨勇, 朱惠延. 地方高校教学督导助推青年教师教学能力提升的对策思考[J]. 高教论坛, 2019(11): 33-36.

[2] Girshick, R. (2015) Fast R-CNN. *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 1440-1448. https://doi.org/10.1109/ICCV.2015.169

[3] Ren, S., He, K., Girshick, R., *et al*. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[4] Liu, W., Anguelov, D., Erhan, D., *et al*. (2016) SSD: Single Shot Multibox Detector. *International Conference on Computer Vision* (*ICCV*), arXiv:1512.02325, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[5] Redmon, J., Divvala, S., Girshick, R., *et al*. (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91

[6] Redmon, J. and Farhadi, A. (2017) YOLO9000: Better, Faster, Stronger. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6517-6525. https://doi.org/10.1109/CVPR.2017.690

[7] Edmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement. *Computer Vision and Pattern Recognition*, arXiv preprint arXiv: 1804.02767.

[8] 吴松伟. 基于人脸识别技术的智能视频监控系统设计与实现[J]. 计算机时代, 2019(11): 62-66.

[9] 刘飞鹏, 沈希忠. 图像处理技术在车牌号识别领域中的算法改进[J]. 电视技术, 2016, 40(12): 28-33.

[10] 闫敬文, 樊秋月. 基于视频图像处理的人数统计方法[J]. 汕头大学学报(自然科学版), 2008, 23(2): 69-73.

[11] Lukas, S., Mitra, A.R., Desanti, R.I. and Krisnadi, D. (2016) Student Attendance System in Classroom Using Face Recognition Technique. 2016 *International Conference on Information and Communication Technology Convergence* (*ICTC*), Jeju, 19-21 October 2016, 1032-1035. https://doi.org/10.1109/ICTC.2016.7763360