

基于人体骨架和深度学习的教师体态语言智能测评

王永固 马家荣 王瑞琳

(浙江工业大学 教育科学与技术学院, 浙江杭州 310023)

[摘要] 深度学习技术与教师教育融合创新,可赋能教师教学体态语言行为精准测评。针对教师体态语言人工视频分析技术滞后问题,本研究采用人物目标检测、目标追踪和体态骨架图生成三种方法,创建真实教学情境下基于人体骨架图的专家型教师标准体态语言数据集,使用金字塔卷积模块优化 ResNet50 残差网络,构建金字塔残差神经网络模型,并基于数据集开展识别模型的训练、验证和测试实验。结果显示,体态语言识别模型的准确率、精确率和召回率均达到 95% 以上,识别准确率由高到低依次为工具性体态语言、巡视性体态语言、指示性体态语言、常规性体态语言、描述性体态语言。本研究基于上述训练良好的模型,开发教师体态语言智能感知测评系统,提出教师体态语言特征的人体骨架表征方法、教师体态语言数据集创建方法及智能识别神经网络模型构建方法,构建了大数据驱动的教师课堂体态语言智能测评方法体系。

[关键词] 教师体态语言; 智能识别; 深度学习; 人体骨架; 神经网络模型

[中图分类号] G420

[文献标识码] A

[文章编号] 1007-2179(2023)01-0100-12

一、引言

体态语言是教师传递教学信息和组织课堂活动的重要途径。教师利用不同的体态语言(如手势、走动等)引发学生注意,增强教师语言表达,激活学生学习动机,辅助学生理解内容,提高课堂教学效果(王继新等, 2020; Bosch et al., 2018; 李爽等, 2020)。相关研究发现,职前师范生和在职教师在正确规范、恰当运用体态语言方面存在较多问题(张倩等, 2019; 牟智佳等, 2020)。为解决这一现实问题,学界采用人工视频分析方法测评教师体态语言,使用

手工标记教师体态语言类型,统计不同类型体态语言运用频次和方式等,提出相应干预方法与改进建议(蒋立兵等, 2018; 周鹏生, 2006)。但是,该方法存在效度低、反馈慢和耗时长等缺点。

教育部(2018)发布的《关于实施卓越教师培养计划 2.0 的意见》强调,利用大数据、深度学习等技术,监测课程教学实施情况,有效诊断评价职前教师学习状况和教学质量。因此,教师体态语言测评需与深度学习技术融合,探索测评的精准化、个性化和规模化路径,助力教师专业能力科学发展。随着人工智能领域深度学习技术的发展,机器能够自

[收稿日期] 2022-09-23

[修回日期] 2022-12-14

[DOI 编码] 10.13966/j.cnki.kfjyyj.2023.01.011

[基金项目] 2021 年国家自然科学基金面上项目“多模态特征融合的自闭症教育机器人情感社交智能感知模型及应用研究”(62177043);浙江省 2020 年哲学社会科学规划重大项目“‘互联网+’残疾人社会支持智慧服务构建与创新”(20XXJC01ZD)。

[作者简介] 王永固,博士,教授,浙江工业大学教育科学与技术学院,研究方向:学习科学与技术、智能教育(wangyg@zjut.edu.cn);马家荣、王瑞琳,硕士研究生,浙江工业大学教育科学与技术学院,研究方向:人工智能教育。

[引用信息] 王永固,马家荣,王瑞琳(2023). 基于人体骨架和深度学习的教师体态语言智能测评[J]. 开放教育研究,29(1): 100-111.

动识别和理解视频中人物的体态语言动作(王萍, 2020), 这为教师体态语言智能测评提供了可行性。

鉴于此, 本研究采用人体骨架和深度学习方法, 设计教师体态语言智能识别与测评方案, 建立真实教学情境下教师标准化体态语言数据集, 生成教师体态语言智能识别深度学习模型, 开发教师体态语言智能测评系统, 探索教师体态语言的精准测评。

二、文献综述

随着人工智能技术与教师教育的深度融合, 深度学习技术正在创新教师体态语言的智能测评方法与系统, 目前已取得阶段性研究成果。

(一) 人工视频分析测评方法

教师体态语言测评研究最先采用人工视频分析技术, 包含主观测量和客观测量两种方法。1) 主观测量法。它采用学习者主观报告方式, 评估人员记录学习者对教师体态语言的主观感受(Harrigan et al., 2008; McArthur, 2022), 其评测结果受评估者和学习者知识结构和主观经验的影响。巴尔马克等(Barmaki et al., 2015)采用主观测量方法, 测评教师口头语言和体态语言之间的和谐程度, 发现数学教师的体态频次明显高于生物教师。2) 客观测量法。它通过自然情境的非参与式观察, 控制体态语言的要素变量, 分析体态语言的特定要素对教学效果的影响机制。巴尔马克(Barmaki, 2014)应用信息技术构建客观测量系统, 分析教师与虚拟学生的体态语言交互过程。研究发现, 更加开放、更长时间的体态语言与学生的学业表现正相关。以上两种方法属于人工视频分析范畴, 主观性强、反馈慢和耗时长, 难以实现教师体态语言评测的精准、即时和高效。

(二) 基于深度学习的智能测评方法

深度学习技术正被应用于教师体态语言的智能测评研究, 它借助红外成像和骨骼成像两种姿态特征数据采集技术, 智能感知和测评教师体态语言教学行为(徐欢云等, 2019)。在红外成像技术方面, 王继新等(Wang et al., 2020)提出基于卷积神经网络的红外体态估计方法, 识别真实教室中教师的指示体态, 识别率为 92%。但是, 红外成像技术需要专业的红外成像设备, 其教育场景应用成本高。与

红外成像技术相比, 骨骼成像技术不需要额外的成像设备, 使用深度学习的人体姿态估计器, 如 AlphaPose、OpenPose 等, 提取二维空间的教师骨骼关节点及其坐标信息, 可以忽略服装背景对人体的干扰, 保持视觉不变, 更适合真实教学场景的教师体态语言智能测评。巴尔马克(Barmaki, 2015)借助骨骼手势生成器构建闭合性手势数据库, 自动统计专家型教师与新手教师闭合性手势的频次和时长差异。刘婷婷等(Liu et al., 2019)使用 OpenPose 骨架估计技术构建教师指示体态数据库, 通过神经网络自动检测教师的指示体态, 识别准确率达 90%。陈少辉(2020)建立 CCNU-Edu-Pose 数据集, 基于教师轮廓骨架融合特征识别教师的 ok、安静、竖起拇指、指向黑板等手势, 自动量化教师的教学手势行为。以上两种技术的先行探索为教师体态语言智能测评提供了方法论参考。然而, 教师课堂体态语言智能测评研究存在两个亟待解决的问题: 1) 缺乏真实课堂教师标准化体态语言行为大数据集样本; 2) 智能测评模型的识别准确率低于 95%, 达不到真实教学应用标准。针对以上问题, 本研究提出真实教学情境的教师体态语言数据集创建方法, 即建立教师课堂教学五种类型体态语言数据集, 构建金字塔残差深度神经网络模型, 开展基于大数据集的识别模型训练与验证实验, 优化残差深度神经网络模型, 最后开发教师体态语言智能感知测评系统, 可视化分析教师体态语言的频次、间隔和时序变化等测度指标, 生成教师体态语言智能测评诊断报告。

三、研究方法

(一) 研究设计

教师体态语言智能测评研究方案由数据采集、数据集建立、识别模型构建、识别模型训练与评估、测评系统开发五个关键环节组成(见图 1)。在数据采集环节, 研究人员使用视频采集设备, 拍摄专家型教师真实教学场景的教学视频, 采集教师标准化体态语言视频素材数据; 在数据集建立环节, 研究人员使用人体目标检测技术、目标追踪技术和姿态估计技术, 将教师体态语言视频转换为人体骨架表征的图片文件, 开发并应用体态语言类别标签标注工具, 为教师姿态骨骼特征图片标注类别标签,

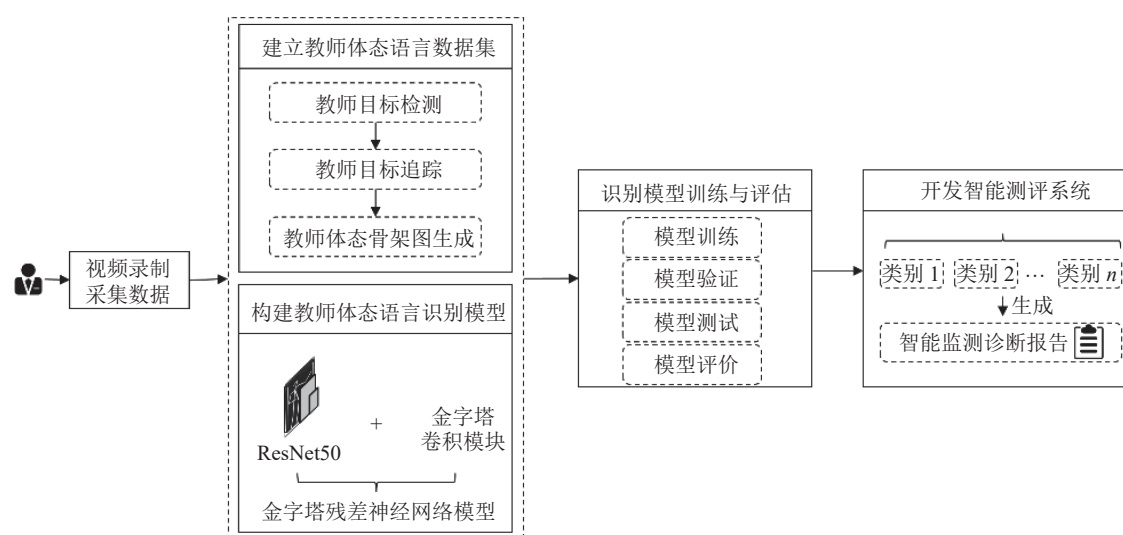


图1 教师体态语言智能测评研究设计

建立教师课堂体态语言大数据集;在识别模型构建环节,研究人员利用金字塔卷积单元优化残差网络,建立金字塔残差神经网络,提取教师体态图像的细粒度特征;在识别模型训练与评估环节,研究人员将环节二的数据集作为模型输入,训练教师体态语言识别模型,计算识别模型的准确率、损失率变化情况,评估模型的优缺点;在测评系统开发环节,研究人员应用 PyQt5 和 Pyecharts 可视化工具包,设计可交互的教师体态语言智能测评系统,将训练良好的识别模型应用到教师体态语言测评实践。经过以上五个环节,教师体态语言能够被自动识别、准确理解、智能测评,并为教师提供个性化反馈。

(二)关键技术

教师体态语言智能识别需要应用两种关键技术:人体骨架表征技术和金字塔残差神经网络识别模型。前者用于提取二维空间的教师体态关节坐标和骨架信息,即使用人体骨架信息表征教师体态特征,提高教师体态特征表征的准确度;后者利用金字塔卷积模块替换 ResNet50 特征提取网络的卷积层,增强识别模型的语义信息表达能力。

1. 教师体态语言的人体骨架表征

文献分析显示,2D 骨架具有视觉不变性、抗干扰能力强、生成骨架图像简单等优点。2D 骨架生成方法为 OpenPose 实时检测算法(Cao et al., 2017),该算法先检测人体的若干关节点,将关节点连接形成骨架图,再通过图优化方法剔除错误的连接,最

终生成人体 2D 骨架图。本研究即采用 OpenPose 人体姿态估计算法,将教师体态语言行为转变成人体骨架图的高级表征形式。人体骨架表征过程分三个步骤:首先,应用 VGG-19 卷积神经网络提取教师体态语言的原始特征;其次,将原始特征作为两个分支输入,第一个分支获取教师体态的 25 个关节点置信度图,第二个分支获取部分亲和域,预测各关节点之间的关联程度;最后,根据关节点置信度图和关联度绘制人体骨架图。通过以上三个步骤,教师体态语言特征被转换为骨架节点及其关系的骨架图。

2. 金字塔残差神经网络模型构建

教师体态语言特征提取的深度学习由 ResNet50 和金字塔卷积模块组成(见图 2)。ResNet50 网络在加深层数的同时引入四组残差单元,使神经网络在采样过程中保留更多的原始特征,防止梯度爆炸问题(He et al., 2016)。特征提取过程包括三个环节:第一,将输入特征 x 同时输送至两个分支,第一个分支负责卷积核的乘积累加运算,输入特征 x 依次经过三个卷积层(其中 3×3 卷积层被金字塔卷积模块代替),输出教师体态特征图 $F(x)$,另一个分支通过跳级连接方式横跨三个卷积层,保留输入特征 x 信息;第二,将特征图 $F(x)$ 与输入特征 x 拼接;第三,利用 relu 激活函数输出特征 $H(x)$, $H(x)=F(x)+x$ 。

金字塔卷积模块通过改变卷积核的尺度和数

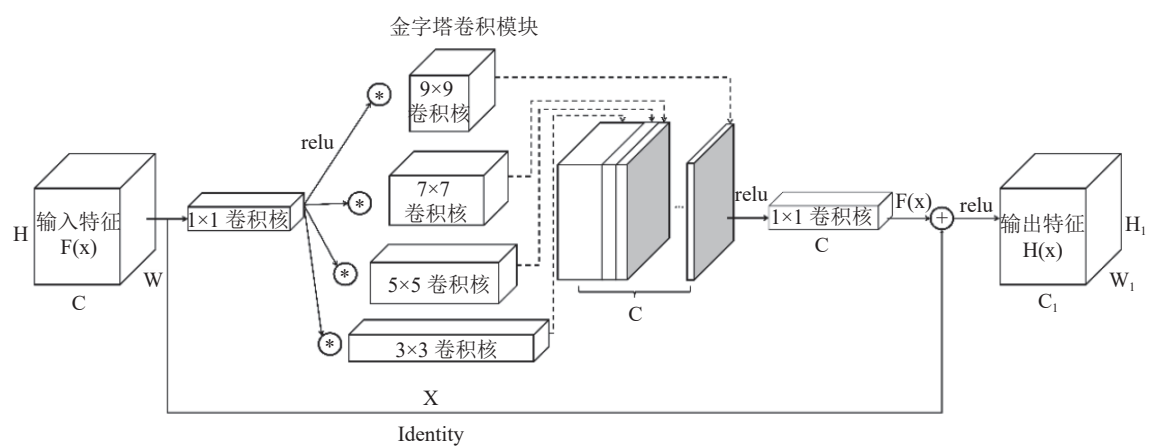


图2 金字塔残差神经网络模型

量,保留教师体态图像更多的层次特征,有效缓解采样过程中的局部信息丢失问题,增强模型的学习能力(Duta et al., 2020)。首先,金字塔卷积模块将ResNet50的3×3卷积核替换成深度、尺度(9×9、7×7、5×5、3×3)不同的金字塔卷积模块,每个模块均对整个输入数据进行卷积操作;然后,拼接9×9卷积核、7×7卷积核、5×5卷积核、3×3卷积核提取的特征,得到输出特征图。在金字塔卷积模块中,3×3小尺度卷积核专注于学习教师体态的深层细节信息,5×5、7×7等大尺度卷积核学习教师体态浅层语言信息,这能使深度神经网络识别的语义信息更丰富。

四、研究实施

本研究的实施过程分建立数据集、训练识别

模型、确定评价指标和开发测评系统四个环节。其中,建立数据集是基础,即为模型训练提供特征数据集;训练识别模型是核心,即基于数据集训练生成识别模型;确定评价指标是保障,评估模型输出结果,衡量前两个环节的优缺点;开发测评系统是关键,即为职前师范生和在职教师开展教学体态语言训练提供评价和改进建议。

(一)建立数据集

教师体态语言数据集建立过程分数据采集、数据处理、数据增强、标签定义、标签标注五个步骤(见图3)。

1. 数据采集

本研究选取浙江省某中职学校32名专家型教师(女18名,男14名),采集真实教学场景下讲授型课堂教学视频1440分钟。实验教师均自愿参加,

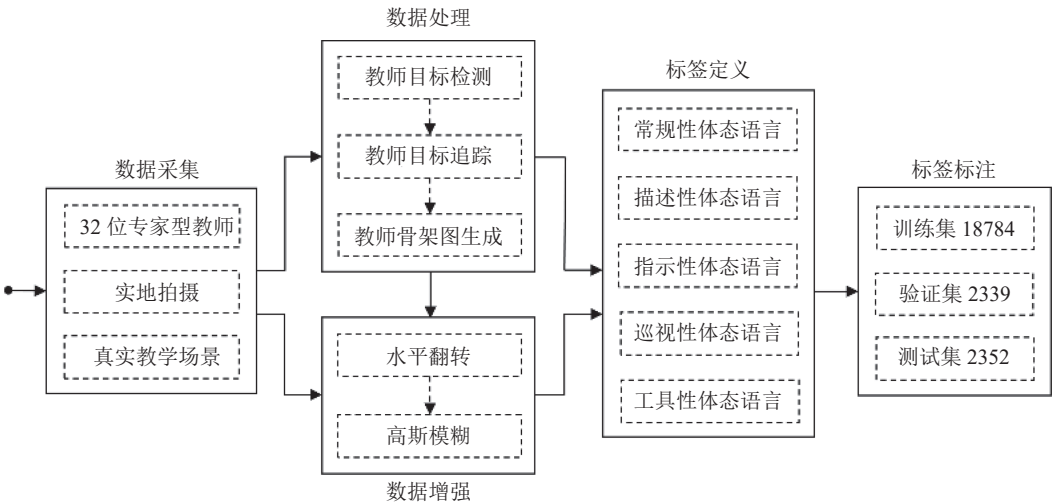


图3 教师体态语言数据集建立过程

同意并知情实验内容和被采集的体态语言类型。

2. 数据处理

教师体态特征数据处理采用人物目标检测、目标追踪和姿态骨架图生成三种技术,先检测教师目标所在区域,再追踪特定区域的教师目标,最后生成教师体态骨架图(见图4)。

第一,检测教师目标。本研究采用目标检测算法 yolov3(Redmon et al., 2018),检测课堂教学场景的教师和学生目标(见图4a)。具体过程分三步:1)调整帧图片大小,使用 OpenCV 库将课堂教学视频流的帧图片大小统一为 416×416 像素,将其输入到特征提取网络,得到 13×13 、 26×26 、 52×52 三种帧图片;2)结果解码,确定边界框坐标值;3)确定人物目标的位置,按照置信度得分排序预测边界框,利用非极大值抑制筛选概率最大边界框,得到每幅帧图片的教师和学生位置。

第二,追踪教师目标。目标追踪算法监测追踪帧图片的教师目标,过滤学生目标(见图4b)。具体过程分四步:1)锁定第 $t-1$ 帧教师目标的坐标。2)查看第 t 帧图像是否存在。3)若第 t 帧图像存在,锁定第 t 帧教师位置坐标,将 $t-1$ 帧和 t 帧的坐标进行 IOU 匹配。若 IOU 匹配成功,将第 $t-1$ 帧教师目标位置坐标更新为第 t 帧坐标,继续执行第一步直至第 t 帧图像不存在;若 IOU 匹配失败,借助卡尔曼滤波算法,预测第 t 帧教师目标位置,并将第 $t-1$ 帧教师目标位置的坐标更新为预测位置坐标,继续执行第一步直至第 t 帧图像不存在。4)结束教师目标追踪,得到只包含教师目标的图片流。

第三,生成教师体态骨架图。在课堂教学中,

学生的视觉注意不会聚焦教师的脚部。因此,本研究改进 OpenPose 算法,得到能够检测 19 个教师关节的姿态估计器(见图4c)。

3. 数据增强

在保证图像特征自然表达的情况下,研究人员采用水平翻转、高斯模糊操作方法,增加数据样本,减少模型过拟合,增强模型泛化能力。

4. 标签定义

本研究参考埃克曼(Ekman)体态语言分类框架(孔令智,1987),结合团队前期研究成果,界定课堂教学场景的教师体态语言的动作特征及其教学价值(徐佳吟,2021),将教师体态语言分为常规性、描述性、指示性、巡视性和工具性五类,界定每种体态语言的动作特征,描述教师体态语言教学效能(见表一)。

5. 标签标注

本研究由两名标注人员和一名教育专家按表一的五种体态语言类别,使用 PyQt5 工具包开发标注程序进行标签标注。数据集按 8:1:1 比例划分,得到 18784 张训练集、2339 张验证集、2352 张测试集的教师体态语言数据集(见表二)。

(二)训练识别模型

本研究实验环境分硬件环境、软件环境和实验参数。1)硬件环境:采用深度学习工作站,内存容量为 32 GB,显存容量为 10 GB,以保障模型训练的高效。2)软件环境:采用 PyTorch 深度学习框架和 Python 编程语言,以保证模型训练的灵活性。3)设置实验参数,批尺寸设置为 16,学习率设置为 $1e-3$,优化器选择 Adam,增强模型收敛的稳定性。以上三个步骤为模型训练提供了高效、灵活、稳定

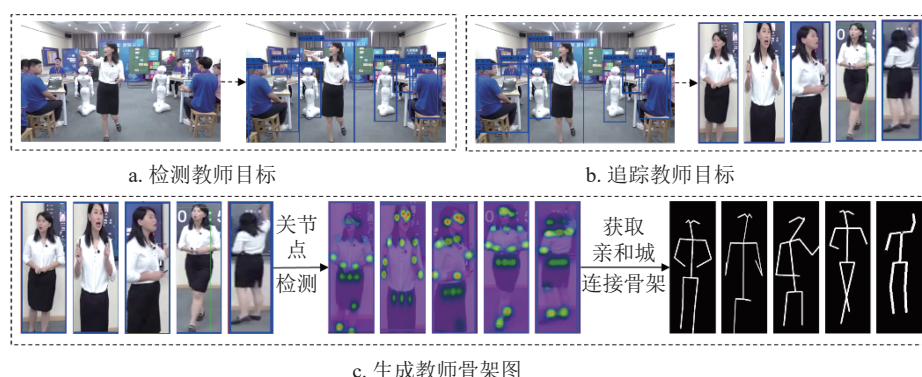


图4 教师体态语言数据处理过程

表一 教师体态语言类别、特征界定和教学效能

类别标签	体态动作特征界定	教学效能
常规性体态语言	教师面向正前方或身体轻微转向,男教师一般手臂自然下垂,女教师通常手臂自由弯曲、双手交叉置于腹部	呈现教师身份,规范教师角色地位
描述性体态语言	教师一个手臂或两个手臂弯曲手肘,使用手部形态向学生表达特定含义,传递教学信息,如使用单手或双手表达特定语音符号	传递教学信息,解释教学内容,降低学生认知负荷
指示性体态语言	教师手臂和手指指向前方学生,或者指向媒体显示区域,手臂伸直或弯曲,如指向 PPT 内容,引导学生视觉注意方向,或指向前方某一学生,示意学生发言	调节学生注意指向,增强课堂学习氛围
巡视性体态语言	教师走动,双腿骨架呈交叉状或接近交叉状	营造良好学习氛围,激励学生投入精力
工具性体态语言	教师转身板书,或操作电子白板	辅助教师语言表达,执行教学行为动作

表二 教师课堂体态语言数据集构成及数量分布

体态语言类别	训练集数量(张)	验证集数量(张)	测试集数量(张)	总计(张)
常规性	3731	506	472	4709
描述性	3985	503	499	4987
指示性	3767	459	470	4696
巡视性	3903	476	488	4867
工具性	3398	395	423	4216
总计(张)	18784	2339	2352	23475

的实验环境。后续模型训练实验按以下三个步骤依次展开。

1)归一化处理。研究人员将教师骨架图像保持长宽比例不变缩放至 224×224 像素,对缩放后的图像进行归一化处理,然后将训练数据集导入优化模型进行训练。2)模型训练与验证,选择最优模型。研究人员在模型训练过程中,以训练数据的损失函数为线索寻找最优权重和偏置参数,损失函数越小,神经网络的模型效果越好;利用验证集数据评估模型,调整模型超参数,经过不断的迭代训练和验证,得到能够识别教师体态语言的最优模型。3)测试模型性能。研究人员将测试集数据导入最优模型,根据评价标准评估模型的泛化能力和准确性能。

(三)确定评价指标

针对教师体态语言识别的多分类场景特征,研究人员将准确率、精确率和召回率作为识别模型的评价指标,计算公式如(1)、(2)、(3)所示:

$$\text{准确率} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{精确率} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{召回率} = \frac{TP}{TP + FN} \quad (3)$$

其中,T、F表示正确与错误;P、N表示模型预测类别,P代表预测为正样本,N代表预测为负样本;TP表示将正样本预测为正样本的数量,TN表示将负样本预测为负样本的数量,FP表示将负样本预测为正样本的数量,FN表示将正样本预测为负样本的数量。准确率表示预测正确的样本数量占总样本数的比例,精确率表示实际为正样本数量占预测为正样本数量的比例,召回率表示预测为正样本的数量占实际正样本数的比例。通常状况下,准确率、精确率和召回率的比值越大,说明模型性能越好,1为最佳状态。

(四)开发测评系统

本研究应用 PyQt5 和 Pyecharts 可视化工具包,设计交互式教师体态语言智能测评系统。该系统包含测评视频导入、教师检测追踪、教师骨架表征和生成测评报告四个功能模块。1)测评视频导入模块利用 PyQt5 的 QFileDialog 文件对话框组件,生成文件对话框,支持测评人员从本地系统导入待测视频文件。2)教师检测追踪模块使用 yolov3 算法和目标追踪算法,获得待测视频中仅包含教师目标的图像流。3)教师骨架表征模块应用 OpenPose 姿态估计器,生成教师体态骨架图。4)生成测评报告模块应用已训练的金字塔残差神经网络模型,统

计分析视频中教师五种体态语言类型的频次和方式,再使用 Pyecharts 工具包绘制可视化测评结果,生成教师教学体态语言测评报告。

五、研究结果

本实验在表二的数据集上进行,训练时长共 6 小时 20 分钟。为验证自建数据集对识别模型训练过程的有效性,本实验从模型训练结果、模型验证结果、模型测试结果三个方面展开:1)依据训练集准确率和损失率的收敛情况判断模型是否得到有效训练;2)依据验证集的准确率和损失率的收敛情况判断模型训练参数是否得到有效调整及模型是否过拟合;3)依据测试集的准确率、精确率和召回率值评估模型识别性能的优劣。最终实验结果见图 5。

(一)模型训练结果

预处理后的教师体态语言数据集被输入金字塔残差神经网络,经过 60 轮迭代训练,训练集的准确率和损失率收敛趋于平稳(见图 5a)。结果显示,随着迭代次数的增加,模型准确率迅速提高,最终稳定在 95% 以上;损失率不存在较大幅度波动,整体呈下降趋势,最终收敛在平稳的数值区间。以上数据说明,识别模型得到有效训练,自建的教师体态语言训练集的数量和质量满足模型学习需要。

(二)模型验证结果

相较于训练集的准确率和损失率变化情况,识别模型在验证集上的波动略大,准确率在前 10 次

迭代中提高迅速,在第 40 次迭代后准确率趋于稳定,达到 95% 以上(见图 5b);损失率在前 10 次迭代中迅速下降,在第 40 次迭代后损失率趋于收敛。因此,识别模型的准确率在训练集和验证集上均呈现先上升后收敛,说明识别模型泛化能力较强,未出现过拟合问题。

模型在验证集上对教师不同类别体态语言的识别达到高准确率。验证集的整体预测准确率高达 99.91%。其中,工具性体态语言的准确率最高,为 99.97%;描述性体态语言次之,为 99.96%;指示性体态语言和巡视性体态语言的准确率为 99.91%;常规性体态语言的准确率低于其他四种类型,为 99.79%。

(三)模型测试结果

混淆矩阵显示了识别模型在教师体态语言测试集的表现情况(见图 5c)。横坐标表示模型预测类别,纵坐标表示真实类别。其中,对角线上的数值表示预测正确的数量,非对角线上的数值表示预测错误的数量。混淆矩阵图颜色深度及数值说明,描述性体态语言预测正确的数量最高,为 479;工具性体态语言预测正确的数量最低,为 422;其他三类的数量居中。

基于测试集混淆矩阵的数值,教师五类体态语言的准确率、精确率和召回率见表三。三个评价指标值特征如下。

1)准确率。工具性体态语言的准确率最高,描述性体态语言的准确率偏低,但两者差别很小。

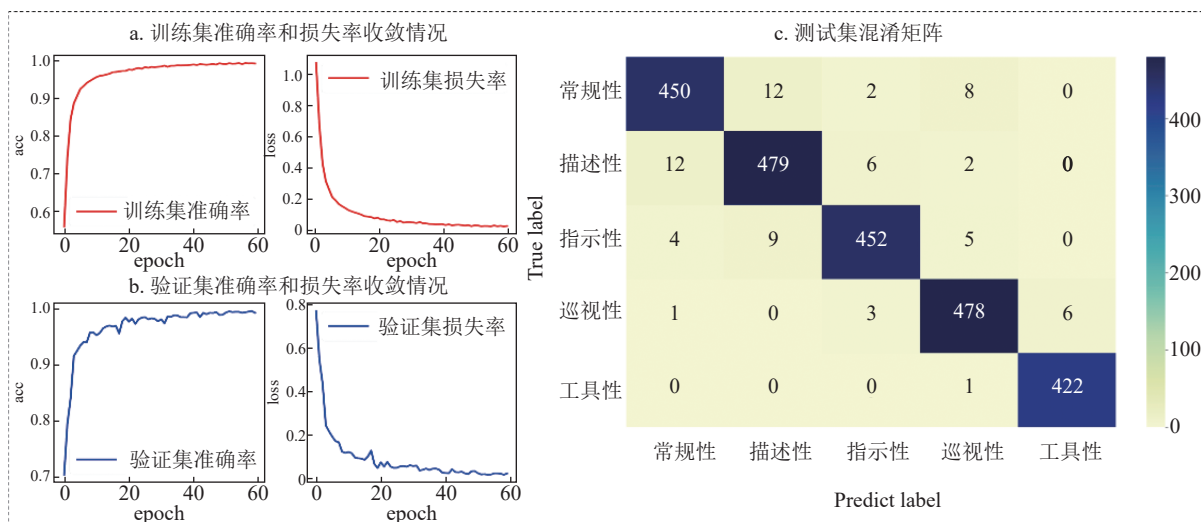


图5 实验结果

2)精确率。工具性体态语言的精确率最高(98.60%),这说明极少出现将其他类别体态语言预测为工具性体态语言;描述性体态语言的精确率最低,这说明将其他类别体态语言预测为描述性体态语言的概率最大。3)召回率。工具性体态语言的召回率最高(99.76%),这说明极少出现将工具性体态语言预测为其他类别体态语言;常规性体态语言的召回率最低,这说明将常规性体态语言预测为其他类别体态语言的概率最大,预测效果不佳。

(四)系统测评结果

本研究采用功能测试法测试教师体态语言智能测评系统。研究人员选用一段时长为3分39秒的授课视频,分四个步骤评测:1)单击菜单栏的

“测评视频导入”选项导入教师授课视频;2)点击菜单栏的“教师检测追踪”选项,等待教师目标检测和追踪完毕;3)点击菜单栏的“教师骨架表征”选项,等待教师骨架图生成完毕;4)点击菜单栏的“生成测评报告”选项,系统显示五种体态语言行为的频次、间隔和时序变化,生成测评分析报告(见图6)。

测评分析报告雷达图显示,常规性体态语言数量最多(2972帧),描述性体态语言837帧,指示性体态语言1966帧,巡视性体态语言790帧,工具性体态语言0帧。报告热力图还显示,视频前半部分以指示性和巡视性体态语言为主,后半部分以常规性体态语言为主。这说明教师前半部分通过体态

表三 识别模型在测试集的准确率、精确率和召回率

评价指标	类别				
	常规性体态语言	描述性体态语言	指示性体态语言	巡视性体态语言	工具性体态语言
准确率(%)	98.34	98.26	98.77	98.89	99.70
精确率(%)	96.36	95.80	97.62	96.76	98.60
召回率(%)	95.34	95.99	96.17	97.95	99.76

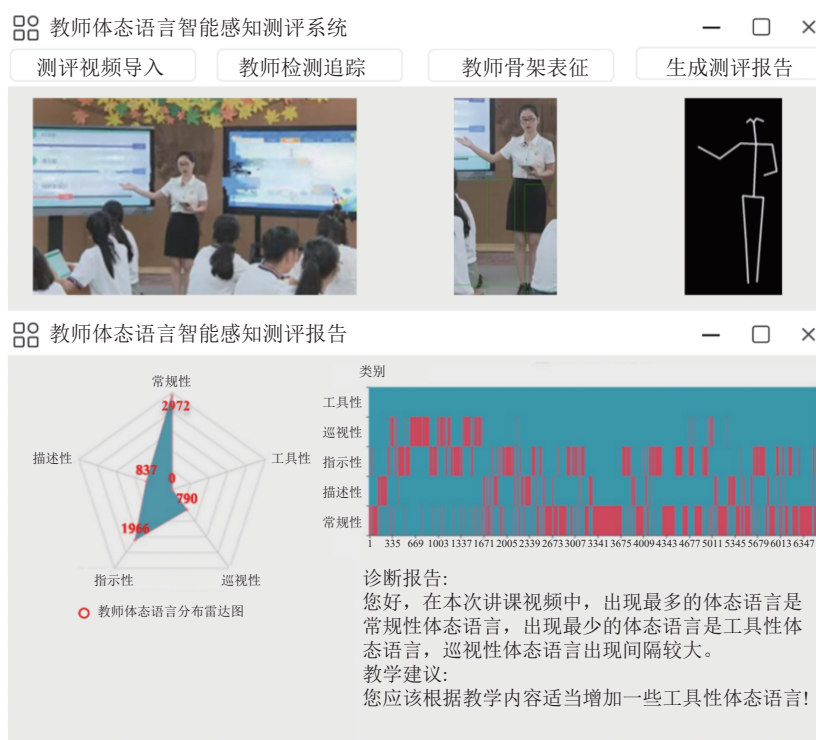


图6 教师体态语言智能感知测评系统功能测试

吸引学生注意力,使学生专注于教学媒体内容,课堂氛围较活跃;后半部分较少出现巡视走动。

六、讨论与分析

本部分依据人体骨架表征、数据集创建和神经网络模型构建等技术原理,关联和对比已有相关研究结果,重点讨论教师体态语言的人体骨架表征、数据集创建及智能识别神经网络模型构建的方法,以构建大数据驱动的教师课堂体态语言智能测评方法体系。

(一)教师体态语言特征的人体骨架表征方法

实验结果显示,教师体态语言识别的准确率、精确率和召回率均达到实用水平,其主要原因在于教师体态语言特征的人体骨架表征可显性表示教师体态特征的差异。已有研究发现,与 RGB 图像的行为识别相比,基于人体骨架的体态语言识别准确率为 97.92%,提高了 4%(何秀玲等, 2020),与本研究教师体态语言识别准确率相近。可见,与传统人类体态特征的 RGB 图像表示法相比,人体骨架表征方法聚焦人体关键关节的轮廓特征,在身体姿势特征差异表征上优势明显,且可省略 RGB 图像纹理特征,减少后期识别模型训练的冗余计算,提高识别模型的计算效率。结合人体骨架建模原理,教师体态语言特征的人体骨架表征方法包含三个关键步骤。

第一,确定教师体态识别应用场景。人体骨架模型相关研究显示,基于人体骨架的应用场景聚焦行为识别、步态识别、异常行为监测、健身康复和舞蹈训练五个领域(赫磊等, 2020)。本研究发现,教师体态语言识别属行为识别。人体骨架表征方法能最大化显现教师体态躯干角度、肢体位置等特征,是提取教师体态语言特征差异的最优方法。

第二,选择最优人体姿态估计器。人物目标检测和目标追踪方法从课堂教学场景提取教师图像,简化教师教学背景,使其不受学生、课桌椅和多媒体设备等背景的影响。在此基础上,选择自下而上的人体姿态估计器 OpenPose 加快了人体骨架生成效率,达到实时估计、快速运行的要求。

第三,建立精简的人体姿态估计模型。针对学生视觉注意极少关注教师脚部的场景特点,本研究选取包含 25 个关节点的人体骨架模型,删除对体

态语言识别贡献较少的脚跟、脚趾等特征点,保留教师体态骨架的 19 个关节点,建立了精简人体姿态估计模型。

综上所述,本研究通过以上三个步骤完成教师体态语言骨架图的特征表征,将教师姿态特征差异最大化显性表示。这个步骤是教师体态语言特征数据集前期样本数据处理的关键环节,能提高教师体态语言分类学习的准确率。

(二)教师体态语言数据集创建方法

目前,全球尚未建立教师课堂体态语言的数据集。针对这一迫切需求,本研究基于数据驱动的研究范式,建立首个来源于真实教学场景的教师体态语言数据集。实验证明,基于该数据集训练的模型预测准确率高 97.02%,达到教师体态语言智能测评应用场景的要求。教师体态语言数据集创建包括数据采集、数据处理、数据增强、标签定义、标签标注五个环节。其中,数据处理、标签定义和标签标注是关键,决定教师体态语言数据集的质量。针对以上三个关键环节,教师体态语言数据集的数据处理、定义及标注的优化有以下三种方法。

第一,采集场景影像,提取特征数据。研究者可在真实教学场景中采集教师课堂教学视频,保证教师体态语言行为发生的情境性和真实性;然后采用人物目标检测、目标追踪和骨架图生成三个关键步骤,从真实课堂场景视频中提取教师体态语言图像数据,生成表征其体态语言特征的骨架图。

第二,聚类行为特征,定义标签类别。研究者可参考埃克曼对体态语言的分类,分析课堂场景中教师体态语言的教学价值,聚类教师体态动作行为特征,确定教师体态语言的类属:常规性、描述性、指示性、巡视性和工具性。标签定义的过程遵循科学性、一致性、可操作性等原则,形成真实教学场景中教师体态语言行为类属的特征界定。

第三,开发标注工具,开展数据标注。数据标注方法主要有开源标注工具、开发标注工具和众包三种方式(蔡莉等, 2020)。本研究采用开发标注工具方式,使用 PyQt5 工具包开发教师体态图像标签标注工具,实现图像显示、标签选择和标签保存三个数据标注功能。本研究聘请两名标注人员和一名教育专家开展多轮多人标注,由标注员使用标注工具为数据集中教师体态骨架图像添加类属标

签,教育专家进行审核和二次标注任务,最终完成数据集中教师体态骨架图数据标签标注任务。

实验证明,以上数据处理、定义及标注的方法,能提取真实教学场景中教师体态语言的显著特征,按照课堂教学效用价值科学确定教师体态类别,高效率标注数据图像标签,创建了全球首个汉语文化背景下教师体态语言数据集,为训练智能识别模型提供了高质量的数据基座。

(三)教师体态语言智能识别神经网络模型构建方法

实验结果证实,ResNet50 和金字塔卷积模块组成深度神经网络模型,其在测试数据集的平均准确率达到 97.02%,达到实用要求的识别效果。相关研究发现,ResNet50 网络模型适合人类体态行为识别与理解的应用研发(陈莹等, 2021)。但是,人体骨架表征的骨架图关系位置复杂、细节特征多,比如小臂与大臂的夹角、大腿与小腿的夹角等。ResNet50 网络模型虽然层数多,但是宽度不足,难以提取教师体态语言骨架图的多角度的细节特征。为解决这一难题,本研究在不增加计算成本和参数量的前提下,使用金字塔卷积模块替换 ResNet50 网络的单一尺度(3×3)卷积核,堆叠不同大小的卷积核,拓展网络宽度,扩大模型感受野,增加教师体态语言骨架图细粒度特征,增强网络特征提取能力,提高教师体态语言智能识别的准确率。基于以上研究分析,智能识别神经网络模型构建一般采用以下三种方法。

第一,明确神经网络模型的适用范围。不同类型神经网络模型适用于不同学习场景。例如,卷积神经网络适合图像识别、图像分类和对象检测等任务,多层感知器适用表格数据集的分类预测和回归预测等任务,循环神经网络适用语音分析、文本分析等前后存在时序依赖关系的任务。

第二,选用适合场景的神经网络模型。鉴于教师体态语言识别属图像识别领域,本研究选用卷积神经网络提取教师体态骨架图特征,并针对教师体态骨架图的细节特征,弃用浅层卷积神经网络模型,选用网络层数较深的 ResNet50 网络,提取教师体态语言更多的抽象特征,解决网络梯度消失和梯度爆炸等问题,保证模型识别效果。

第三,优化深度神经网络模型结构。神经网络

模型的优化方法有两种:调整超参数和调整网络结构。前者改变学习率、优化器和批尺寸等超参数;后者通过增加卷积层、引入注意力机制层等调整网络深度(Li et al., 2014; 王永固等, 2021),或者通过嵌入金字塔卷积、级联卷积等增加网络宽度(Zhang et al., 2016),提取更高级的语义特征。本研究采用金字塔卷积模块替换单一尺度卷积核的优化方法,增加 ResNet50 神经网络宽度。因此,模型结构优化的一般原则为依据数据集的样本细节特征,结合模型训练结果,合理增加神经网络的深度或宽度,保持适中的神经网络深度和宽度。

基于以上三种方法构建的最优的教师体态语言智能识别神经网络模型,其网络深度和宽度适合识别教师体态语言骨架图的细节特征,可满足教师体态语言识别的高准确率要求,能用于开发教师体态语言智能测评场景应用。

七、研究结论

教师体态语言自动识别是测评和干预教师课堂教学能力的关键技术之一,它能有效增强教学内容讲解和创设活跃课堂氛围,提升课堂教学的质量和效率。本研究以真实教学情境中专家型教师体态语言为研究对象,设计教师体态语言智能识别研究方案,使用人物目标检测、目标追踪和体态骨架图生成三种技术,建立了全球首个基于人体骨架的教师体态语言数据集。在此基础上,本研究针对教师体态语言的骨架图特征,构建金字塔残差卷积神经网络模型,实现教师体态语言识别准确率高于 97.02% 的研究目标,并开发教师体态语言智能感知测评系统,破解人工视频分析教师体态语言的局限性。但是,本研究的教师体态语言仅包含教师的身体、头部和上肢动作特征,尚未考虑教师的面部表情、手势语义和语音等模态特征。后续研究将采集多种模态来源数据,建立更大规模的教师体态语言数据集,探索多模态融合的教师体态语言智能感知方法,开发教师体态语言智能测评应用场景。

[参考文献]

- [1] Barmaki, R., & Hughes, C. E. (2015). A case study to track teacher gestures and performance in a virtual learning environment[C]. The Fifth International Conference on Learning Analytics and Knowledge. Poughkeepsie, NY, USA: 420-421.

- [2] Barmaki, R. (2014). Nonverbal communication and teaching performance[C]. The 7th International Conference on Educational Data Mining (EDM 2014). London, UK: 441-443.
- [3] Barmaki, R. (2015). Multimodal assessment of teaching behavior in immersive rehearsal environment-teachlive[C]. The 2015 ACM on International Conference on Multimodal Interaction. Seattle, Washington, USA: 651-655.
- [4] Bosch, N., Mills, C., Wammes, J. D., & Smilek, D. (2018). Quantifying classroom instructor dynamics with computer vision[C]. International Conference on Artificial Intelligence in Education(AIED). London, UK: 30-42.
- [5] 蔡莉,王淑婷,刘俊晖,朱扬勇(2020). 数据标注研究综述[J]. 软件学报, (2): 302-320.
- [6] Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: 7291-7299.
- [7] 陈少辉(2020). 智慧教室环境下教师非言语行为量化计算研究及应用[D]. 武汉: 华中师范大学: 70-73.
- [8] 陈莹,龚苏明(2021). 改进通道注意力机制下的人体行为识别网络[J]. 电子与信息学报, 43 (12): 3538-3545.
- [9] Duta, I. C., Liu, L., Zhu, F., & Shao, L. (2020). Pyramidal convolution: Rethinking convolutional neural networks for visual recognition[J]. arXiv preprint arXiv: 2006.11538.
- [10] 赫磊,邵展鹏,张剑华,周小龙(2020). 基于深度学习的行为识别算法综述[J]. 计算机科学, 47 (S1): 139-147.
- [11] Harrigan, J., Rosenthal, R., & Scherer, K. (2008). New handbook of methods in nonverbal behavior research[M]. New York: Oxford University Press: 121-124.
- [12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: 770-778.
- [13] 何秀玲,杨凡,陈增照,方静,李洋洋(2020). 基于人体骨架和深度学习的学生课堂行为识别[J]. 现代教育技术, 30 (11): 105-112.
- [14] 蒋立兵,毛齐明,万真,沈欢(2018). 智慧教室促进高校课堂教学变革的绩效研究——基于课堂教学行为的分析[J]. 中国电化教育, (6): 52-58.
- [15] 教育部(2018). 教育部关于实施卓越教师培养计划 2.0 的意见[EB/OL]. [2018-09-30]. http://www.moe.gov.cn/srcsite/A10/s7011/201810/t20181010_350998.html.
- [16] 孔令智(1987). 社会心理学新编[M]. 沈阳: 辽宁人民出版社: 45-47.
- [17] 李爽,钟瑶(2020). 在线教师教学投入对学生学习绩效的影响——基于教师和学生的视角[J]. 开放教育研究, 26 (3): 99-110.
- [18] Li, Y., Tong, S., & Li, T. (2014). Composite adaptive fuzzy output feedback control design for uncertain nonlinear strict-feedback systems with input saturation[J]. IEEE Transactions on Cybernetics, 45(10): 2299-2308.
- [19] Liu, T., Chen, Z., & Wang, X. (2019). Automatic instructional pointing gesture recognition by machine learning in the intelligent learning environment[C]. The 2019 4th International Conference on Distance Education and Learning. New York: 153-157.
- [20] McArthur, J. A. (2022). From classroom to Zoom room: Exploring instructor modifications of visual nonverbal behaviors in synchronous online classrooms[J]. Communication Teacher, 36(3): 204-215.
- [21] 牟智佳,苏秀玲,严大虎(2020). 课堂环境下基于教学行为的教师教学投入度评测建模研究[J]. 现代远程教育, (3): 61-69.
- [22] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement[J]. arXiv preprint arXiv: 1804.02767.
- [23] Wang, J., Liu, T., & Wang, X. (2020). Human hand gesture recognition with convolutional neural networks for K-12 double-teachers instruction mode classroom[J]. Infrared Physics & Technology, 111: 103464.
- [24] 王继新,田俊,王萱,韦怡彤(2020). 基于教学行为数据分析的“互联网+在地化课堂”优化对策研究[J]. 电化教育研究, 41 (4): 93-101.
- [25] 王萍(2020). 人工智能在教育视频中的应用分析与设计[J]. 电化教育研究, 41 (3): 93-100+121.
- [26] 王永固,张晨焘,许家奇,李晓娟(2021). 基于深度学习的孤独症儿童学习情绪智能感知[J]. 现代教育技术, 31 (7): 105-111.
- [27] 徐欢云,胡小勇(2019). 借鉴、融合与创新: 教育人工智能发展的多维路向——基于 AIED(2011-2018)的启示[J]. 开放教育研究, 25 (6): 31-45.
- [28] 徐佳吟(2021). 中职教师课堂口头与体态语言能力特征及提升策略研究——基于 Praat 和 Elan 的实证分析[D]. 杭州: 浙江工业大学: 27-29.
- [29] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 23(10): 1499-1503.
- [30] 张倩,刘清堂,张文霄,吴林静,张妮(2019). 课堂师生互动视域下教师行为特征分析与策略研究——基于 Leary 模型[J]. 现代远程教育, (3): 30-37.
- [31] 周鹏生(2006). 国内外教师体态语研究之比较[J]. 外国中小学教育, (9): 26-31.

(编辑:赵晓丽)

Research on Intelligent Assessment of Teachers' Body Language Based on Human Skeleton and Deep Learning

WANG Yonggu, MA Jiarong & WANG Ruilin

(College of Educational Science and Technology, Zhejiang University of
Technology, Hangzhou 310023, China)

Abstract: *The integration and innovation of deep learning technology and teacher education enable teachers to accurately evaluate body language behavior in teaching. To address the problem of the lag of artificial video analysis technology of teachers' body language, this study used three methods of human-target detection, human-target tracking, and body skeleton map generation to construct the expert teacher's standard body language dataset based on the human skeleton map in the real teaching situation. The study also built a pyramid residual network model, which used the pyramid convolution module to optimize the ResNet50 residual network. Then, the study trained, validated, and tested the recognition model based on the dataset. The results showed that the accuracy, precision, and recall rates of the body language recognition model were more than 95%, and the recognition accuracy rates from high to low are instrumental body language, patrolling body language, indicative body language, routine body language, and descriptive body language. Based on the above-mentioned well-trained model, this study developed an intelligent perception assessment system for teachers' body language, and proposed the human skeleton representation method for teachers' body language features, a dataset creation method for teachers' body language, an intelligent recognition neural network model construction method, and the constructs of a big data-driven intelligent assessment system for teachers' body language in the classroom.*

Key words: *teachers' body language; intelligent assessment; deep learning; human skeleton; the neural network model*