第 38 卷 第 5 期 广西师范大学学报(自然科学版) Vol. 38 No. 5
2020 年 9 月 Journal of Guangxi Normal University (Natural Science Edition) Sept. 2020

DOI: 10.16088/j.issn.1001-6600.2020.05.001 http://xuebao.gxnu.edu.cn

基于多路特征融合的 Faster R-CNN 与迁移学习的学生课堂行为检测

白 捷[1,2]，高海力[3]，王永众[4]，杨来邦[4]，项晓航[4]，楼雄伟[1,2,5*]

(1. 浙江农林大学 信息工程学院，浙江 杭州 311300；2. 浙江省林业智能监测与信息技术研究重点实验室，浙江 杭州 311300；
3. 浙江省林业局，浙江 杭州 311300；4. 杭州感知科技有限公司，浙江 杭州 311300；
5. 林业感知技术与智能装备国家林业和草原局重点实验室，浙江 杭州 311300)

摘 要： 课程教学质量是衡量学校教学水平的一项核心内容，其教学效果可以从学生听课状态进行直观反映。为提升学生上课状态，督促课堂纪律，本文提出一种基于多路特征融合的 Faster R-CNN 与迁移学习的学生课堂行为检测的方法。首先，利用浙江农林大学监控视频进行手工标注图片，并利用数据增强方式增加图片规模，建立常见学生课堂行为数据集；然后，以预训练的 Inception-ResNet-v2 网络用于特征提取，目标检测框架选用 Faster R-CNN，通过迁移学习方式实现对正常学习、睡觉、低头等课堂行为的检测；最后，通过多路特征融合方式，在拥有丰富语义信息的深层中融入更多细节信息的浅层特征，得到改进的学生课堂表现检测模型。实验结果显示：该模型的查准率均值可达 76.32%，在原有算法基础上提升了 12.22 个百分点，取得较好的检测效果。该模型对学生课堂行为具有较高的查准率，表明多路特征融合的 Faster R-CNN 在学生课堂行为检测具有一定的应用前景，可为提高课堂教学质量提供新的参考。

关键词： 课堂行为检测；Faster R-CNN；特征融合；迁移学习

中图分类号： TP181 **文献标志码：** A **文章编号：** 1001-6600(2020)05-0001-11

引用格式： 白捷，高海力，王永众，等. 基于多路特征融合的 Faster R-CNN 与迁移学习的学生课堂行为检测[J]. 广西师范大学学报(自然科学版)，2020，38(5)：1-11.

BAI J，GAO H L，WANG Y Z，et al. Detection of students' classroom performance based on Faster R-CNN and transfer learning with multi-channel feature fusion[J]. Journal of Guangxi Normal University (Natural Science Edition)，2020，38(5)：1-11.

Detection of Students' Classroom Performance Based on Faster R-CNN and Transfer Learning with Multi-Channel Feature Fusion

BAI Jie[1,2]，GAO Haili[3]，WANG Yongzhong[4]，YANG Laibang[4]，XIANG Xiaohang[4]，LOU Xiongwei[1,2,5*]

(1. School of Information Engineering，Zhejiang Agriculture and Forestry University，Hangzhou Zhejiang 311300，China；
2. Key Laboratory of Forestry Intelligent Monitoring and Information Technology of Zhejiang Province，Hangzhou Zhejiang 311300，China；
3. Forestry Department of Zhejiang Province，Hangzhou Zhejiang 311300，China；
4. Hangzhou Perception Technology Company Limited，Hangzhou Zhejiang 311300，China；
5. Key Laboratory of State Forestry and Grassland Administration on Forestry Sensing Technology and Intelligent Equipment，Hangzhou Zhejiang 311300，China)

Abstract: Course teaching quality is a core content to measure the teaching level of a school，and teaching effect can be directly reflected from the state of students' attendance. In order to improve students' class status and promote class discipline，this paper proposes a detection method for students' classroom behavior based on Faster

收稿日期： 2020-01-08

基金项目： 浙江省重点研发项目(2018C02013)；浙江省科技计划项目(2017C02044)

通信联系人： 楼雄伟(1979—)，男，浙江东阳人，浙江农林大学副教授，博士。E-mail：fly_pugongying@163.com

R-CNN and transfer learning with multi-channel feature fusion. Firstly, images are obtained through the monitoring video of Zhejiang Agriculture and Forestry University and manually annotated, and data augmentation method is used to increase the scale of the images to establish the dataset of common students' normal classroom behavior. Then, the Inception-ResNet-v2 network based on pre-training is applied for feature extraction, and the target detection framework adopts Faster R-CNN to realize the detection of normal learning, sleeping, lowering head and other student behaviors through transfer learning. Finally, through multi-channel feature fusion method, the shallow features of more detailed information are integrated in deep layers of rich semantic information, so as to gain the improved detection model of the students' classroom performance. Experimental results show that the mean average precision of the model can reach 76.32%, which is 12.22 percentage points higher than original algorithm, and good detection effect can be achieved. This model has a high accuracy rate for students' classroom behavior, which indicates that Faster R-CNN with multi-channel feature fusion has a good application prospect in students' classroom behavior detection, and can provide a new reference for improving classroom teaching quality.

Keywords: classroom behavior detection; Faster R-CNN; feature fusion; migration learning

近年来,普通高校学生上课的主动接受能力呈下降趋势,学生的学习积极性低,在课堂中聊天、睡觉、玩手机等情况普遍存在。如何科学地提升学生上课状态,对学校全面提高教学质量具有重要的意义[1]。

传统的课堂行为分析主要以人工观察监控视频为主[2],主流方法包括 S-T 教学分析[3]、弗兰德斯互动分析系统[4]等。由于监控视频数量庞大,通过人工方式进行处理,易出现疲劳、效率较低等问题,同时耗费大量的人力成本。随着人工智能的发展,利用机器学习和图像视频处理等对监控视频智能化识别与分析,能够较好地减少传统课堂行为分析对人力的依赖性[5]。

在机器学习领域中,学生课堂行为检测的研究主要利用人体骨架向量、流光特征、全局运动方向特征等方法进行特征提取,并结合朴素贝叶斯或支持向量机等分类器进行人体行为识别[6]。例如:张鸿宇等[7]通过选用人体骨骼向量,采用 SVM 分类器对姿态向量特征进行分类,有效地识别出多个学习者的举手、正坐和低头等课堂行为;党冬利[8]通过提取运动历史图的 Zernike 矩特征[9]、流光特征及全局运动方向等特征,利用朴素贝叶斯分类器对在背景复杂的教室环境下的学生举手、站立和坐下 3 种动作进行了有效识别。传统的目标检测常用手工提取局部特征方式表示,局部特征[10]相比全局特征[11]对遮挡、扭曲、噪音等不敏感方面具有优势。如:尺度不变特征变换(scale-invariant feature transform,SIFT)算子[12]利用构建高斯差分金字塔,找出关键点定位,构建关键点描述等过程获取局部特征,提取出的特征具有一定的平移缩放稳定性和抗干扰性等优点;方向梯度直方图[13]能够较好地提取图像边缘信息。然而这些传统的机器学习方法需要依赖大量的人工提取特征并且准确率较低。

相比于传统方法,深度学习能够从大量数据中自动学习特征数据,克服了人工提取特征的局限性,以 VGGNet[14]、GoogLeNet[15]、ResNet[16]等为代表的卷积网络模型在分类方面都取得较好的识别效果。另外,在目标检测领域还将特征提取过程和分类器统一在一个框架中,能够快速适应不同的分析任务[17]。近年来,深度学习也被应用于行为识别,如廖鹏等[18]利用背景差分法提取目标区域,通过 VGG 网络模型提取特征,实现了对正常上课、睡觉、玩手机 3 种课堂行为的检测。可见深度学习应用于课堂行为检测具有一定的理论可行性。

目标检测是图像处理研究的核心之一,利用基于深度学习的目标检测算法来实现对学生课堂表现的行为检测,对提高教学质量具有重要的意义。针对目标检测的研究,基于深度学习的方法正在不断的普及。如今,基于深度学习的目标检测框架主要分为两大类[19]:一类是以 YOLO(you only look once)[20]和 SSD(single shot multibox detector)[21]为代表的一阶段方法的目标识别算法,这些检测系统同时计算类别概率和位置信息;另一类是以 Faster R-CNN 系列[22-26]为代表的两阶段方法,通过区域建议网络 RPN 生成候选区域再精细计算类别概率,检测的准确率也更高。因此,本文选择 Faster R-CNN 进行课堂行为检测研究。

考虑到训练深度网络需要大量样本集的支持,而本文建立的学生课堂行为数据集较小,易导致模型准确率低、泛化能力差等问题,因此,本文基于迁移学习的思想,选用在 ImageNet 训练好的 Inception-ResNet-

v2-Faster R-CNN 作为预训练模型，将模型参数迁移[27]到课堂检测模型中，调整最后的输出通道数，以期实现对学生常见课堂行为即正常学习、睡觉、低头(低头、玩手机等)的检测。

为了进一步提升检测的准确率，以接近实际应用，考虑到主干网络 Inception-ResNet-v2[28]模型的深层网络的特征图，经多次的特征提取后，在细节信息上有所丢失，而低层网络特征图的视觉信息明显，因此，本文在主干网络 Inception-ResNet-v2 模型基础上，增加了连接浅层网络到深层网络的通路，以增强特征信息，以期达到提高目标检测效果的目的。

1 Faster R-CNN 与迁移学习模型构建

1.1 Faster R-CNN 模型

Faster R-CNN 算法总体结构如图 1 所示，其总体主要结构大致分为 3 个部分：用于提取特征数据的卷积网络部分、生成候选框的区域建议网络(RPN)部分和检测子网络部分。

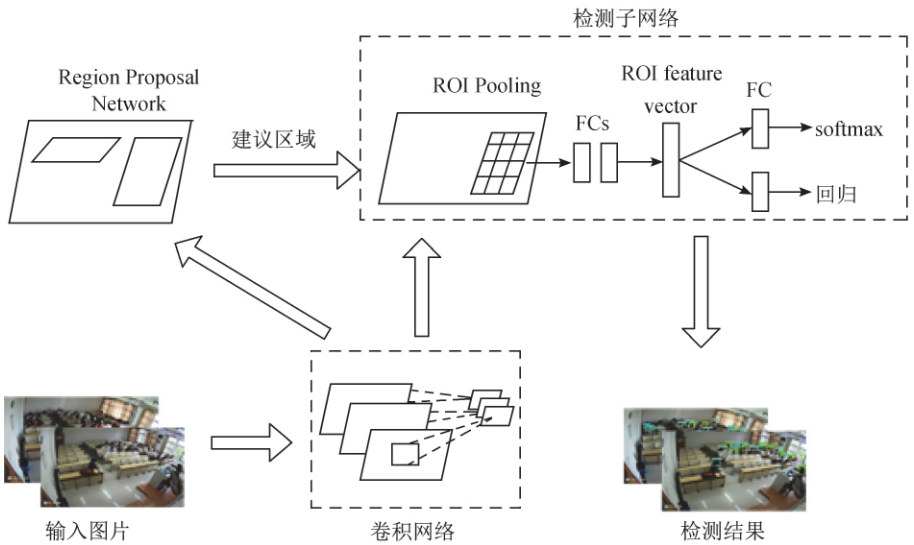


图 1 Faster R-CNN 总体结构

Fig. 1 Overall structure of Faster R-CNN

待检测课堂图片经由卷积网络层提取卷积特征图，在卷积网络的最后一个卷积层获取特征图，分别用于输入到后面的 RPN 网络和检测子网络，以适应不同输入尺寸。本文的基础主干网络选择 Inception-ResNet-v2 模型，该模型引入了 Inception 模型结构，可使同一层特征图能够使用多个尺寸不同的卷积核，以获得不同尺度的特征，使得网络的宽度能够提高。同时引入了正则化和 ResNet 模型的残差网络结构，有利于减缓网络性能退化和梯度消散问题，使得网络的层数可以加深。

提取候选区域的任务通过 RPN 实现，核心思想在于使用卷积神经网络 CNN 产生区域建议框，通过使用 9 个不同尺度比例的 anchor boxes 在卷积层上滑动，而这 9 个 anchor boxes 和边框回归可以得到多尺度比例的候选区域，然后将候选框输入 Faster R-CNN 中做更精细的分类和位置修正。

为了训练 RPN，对每个锚点 anchor 进行二分类，分为对象类和非对象类。其损失函数由分类误差和回归模型产生的误差组成，其中分类误差和回归模型比重为 1∶2。其定义为

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{\text{cls}}}\sum_i L_{\text{cls}}(p_i,p_i^*) + \lambda\frac{1}{N_{\text{reg}}}\sum_i p_i^* L_{\text{reg}}(t_i,t_i^*), \tag{1}$$

式中：N_{cls} 表示分类锚点数量；指示函数 p_i^* 表示当锚点 i 为对象(即为 positive)时，其值为 1，否则为 0；p_i 为锚点 i 是对象的概率值；L_{cls} 表示分类的 log 损失函数，仅当锚点 i 为对象时，计算回归损失；λ 值为 2；$L_{\text{reg}}(t_i,t_i^*)$ 表示 $\text{smooth}_{L1}(t_i - t_i^*)$ 函数，t_i 表示预测边框坐标向量，t_i^* 表示与正向锚相关联的真实坐标框(ground-truth box)向量。具体公式为：

$$t_x=(x-x_a)/w_a,t_y=(y-y_a)/h_a,t_w=\log(w/w_a),t_h=\log(h/h_a),$$
$$t_i^*=(x^*-x_a)/w_a,t_y^*=(y^*-y_a)/h_a,t_w^*=\log(w^*/w_a),t_h^*=\log(h^*/h_a)。\quad(2)$$

式中:变量 x、y、w、h 表示预测框的中心坐标及其宽度和高度;下标为 a 的 x_a、y_a、w_a、h_a 表示锚框的中心坐标及其宽度和高度;上标为* 的 x^*、y^*、w^*、h^* 表示真实坐标框的中心坐标及其宽度和高度。smooth_{L1} 函数为:

$$\text{smooth}_{L1}(x)=\begin{cases}0.5x^2, & \text{if}\ |x|<1,\\ |x|-0.5, & \text{else}。\end{cases}\quad(3)$$

区域建议网络 RPN 生成 300 个区域建议框,每个建议框带有该框是对象的概率信息,经卷积网络得到的最末卷积层的特征图,使用 3×3 的卷积核卷积特征,得到 256 维特征向量,将 256 维特征向量分别经由 1×1 的卷积核进行降维后送入 Softmax 和回归器,生成 k 组建议框的偏移量和 k 组表示物体的概率。本文 anchor 个数 $k=9$,其尺度大小分别为 128^2、256^2、512^2,取 3 个尺度长宽比为 1∶1、1∶2、2∶1。每个像素点共 9 个尺度不同的候选框。

检测子网络部分包括兴趣区域池化(region of interest pooling,ROI pooling)层和全连接层。待检测的课堂图片经卷积网络提取特征图后,再经 RPN 网络生成候选框,按比例在该特征图中找到对应区域,利用 ROI 最大池化将不同尺寸的特征映射到固定长度的向量。将 ROI 池化后的输出结果,分别输入到用于全连接网络组成的回归层和分类层。回归层对 RPN 区域建议框中的目标位置进行回归和分类层通过 Softmax 分类,最后由非极大值抑制输出检测结果。

1.2 迁移学习模型

迁移学习可定义为:给定一源域 D_S 和学习任务 T_S,一个目标域 D_T 和学习任务 T_T,则目的是使用 D_S 和 T_S 中的知识帮助提高 D_T 中的目标预测函数 $f_T(x)$ 的学习,其中 $D_S\neq D_T$ 或 $T_S\neq T_T$[29]。针对样本数据不足的情况,利用经预训练的成熟神经网络模型进行迁移,通过共享卷积池化层的权值参数,对仅采集了少量数据的问题进行求解。这不仅有利于减少对数据样本数量的要求,而且缩短了训练所需的时间。本文采用的迁移学习模型结构如图 2 所示。

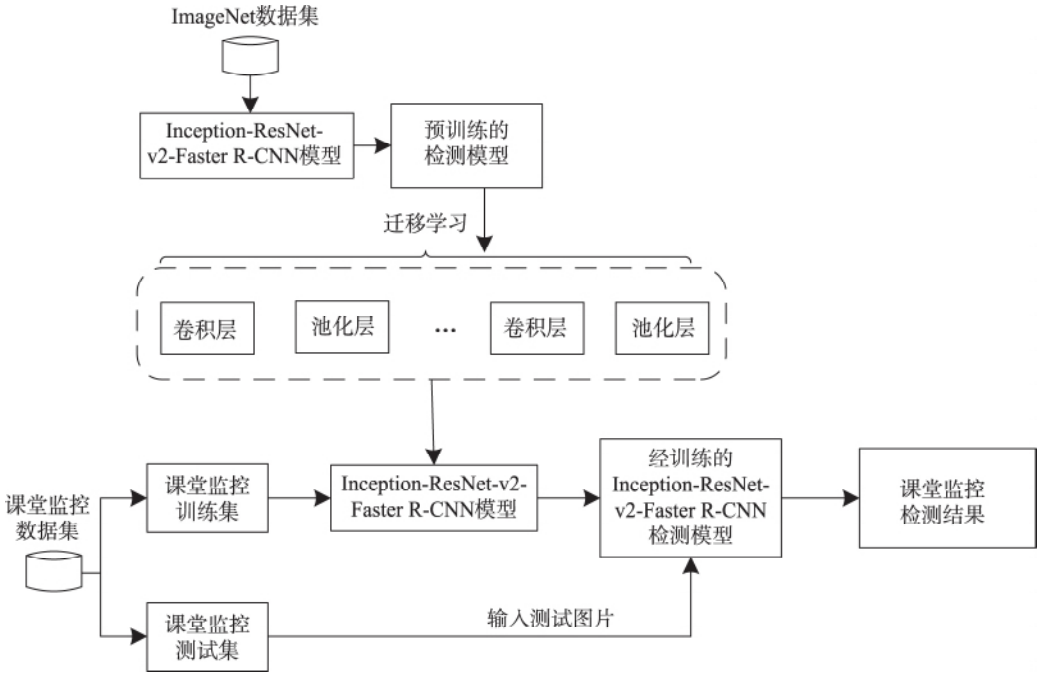


图 2 Inception-ResNet-v2-Faster R-CNN 迁移学习模型结构

Fig. 2 Inception-ResNet-v2-Faster R-CNN transfer learning model structure

本文学生课堂行为检测模型的构建思路为:首先,选用在 ImageNet 预训练的 Inception-ResNet-v2-Faster R-CNN 模型,将参数迁移到本文的课堂行为检测模型中;然后,调整最后一层全连接层的输出通道

数为本文数据集的类别数，即类别数为 3; 最后，在本文建立的数据集上进行训练学习，并利用测试集测试训练后的结果。

2 特征融合的卷积网络

在目标检测算法 Faster R-CNN 中，卷积网络一般使用 VGG16 或 ZF 网络进行特征提取。为了更好地提取特征数据，本文采用网络层数更多的 Inception-ResNet-v2 模型作为骨干卷积网络模型。在深度卷积网络提取深层特征图过程中，距离网络输入部分越近的网络层，其提取到的特征图分辨率越大，目标位置准确; 距离网络输出部分越近的网络层，其特征图的分辨率越小，提取到的特征语义信息越丰富，但细节信息有所丢失，而丢失的细节信息可以在浅层的特征图中进行获取[30]。基于该思想，本文将采用深层特征通过融合浅层特征的方式来提高对学生课堂表现的检测效果。

本文的骨干卷积网络 Inception-ResNet-v2 经前人的大量研究和实验，具有较强的借鉴性。同时，本文通过融合浅层特征网络思想在该模型基础上进行了适当改进。在网络设计思路方面，本文借鉴文献[31]提出的多路融合网络的思想来实现深层特征融合浅层特征，该模型融合了浅层和深层特征后，通过上采样将特征还原到一定大小，而本文并不需要通过上采样操作，仅采用了该模型的多路特征融合的思想。

本文根据 Inception-ResNet-v2 的实际网络结构，在 Inception-ResNet-v2 模型基础上增加了 2 条通路，用于融合浅层特征图和深层特征图。Inception-ResNet-v2 网络的 PreAuxLogits 层所提取出的特征图用于特征输出前，先将精心选取 Inception-ResNet-v2 的浅层特征图与 PreAuxLogits 层提取的深层特征图进行特征融合处理后，再作为网络的特征输出。该模型结构的框架见图 3，以输入图片大小为 299×299 的彩色图片为例。

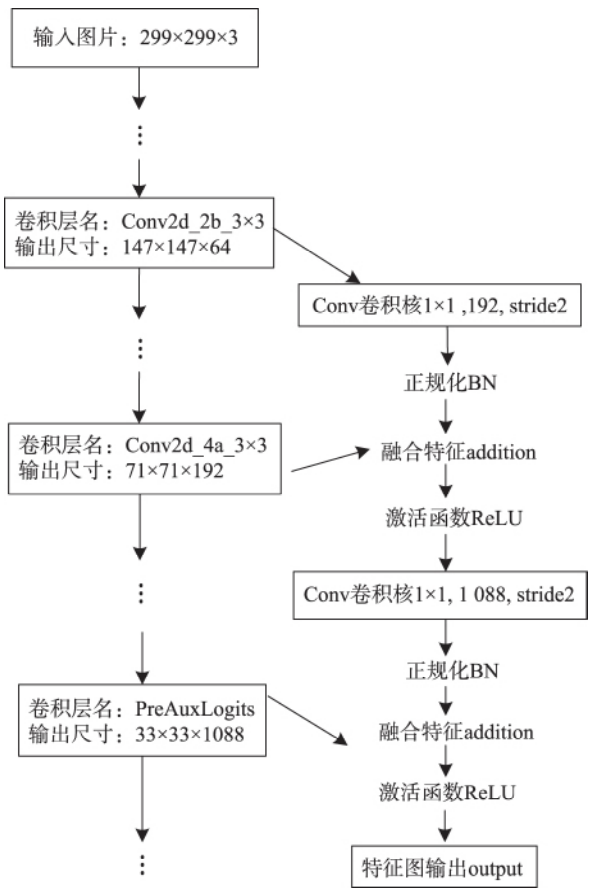


图 3 改进的 Inception-ResNet-v2 模型结构

Fig. 3 Improved Inception-ResNet-v2 model structure

图3左侧自上而下的网络路线,即从输入图片到卷积层PreAuxLogits的网络路线,是文献[28]提出的Inception-ResNet-v2网络结构。而图中右侧支线网络对应着本文改进的融合网络,Conv代表卷积操作;BN代表BatchNorm正规化;addition代表融合特征;ReLU是特征激活函数。

该卷积网络模型对输入的图片按自上而下的网络路线进行特征提取。在Conv2d_2b_3×3网络层提取的特征图经1×1大小、stride为2的卷积后,再经过BatchNorm正规化,而后与Conv2d_4a_3×3提取的特征图特征融合,经由ReLU激活,实现了浅层Conv2d_2b、Conv2d_4a层的特征融合。然后,将该特征融合的输出值经1×1大小、stride为2的卷积核处理后,经BatchNorm正规化,再与PreAuxLogits网络层提取的特征图融合,并经由ReLU激活输出,实现了融合PreAuxLogits层的深层特征。最后将该层的输出取代原先的PreAuxLogits层输出。

3 实验及分析

3.1 数据集设置

目前现有的目标检测公开数据集如PASCAL VOC、MS coco等,是专用于检测类别的,并不适合本研究,因此,本文将利用真实的浙江农林大学监控数据自行制作。选取数据集图片共300张,每张图片对象个数为20~40,总共超8 000个目标对象。其中200张图片以上课睡觉、趴着或玩手机的人为主,100张图片以正常上课的人为主。这200张和100张图片分别采样于2个不同教室的视频数据,每秒采样1帧图片。随机采取60%的数据用于训练集,剩下的40%数据用于测试集。生成的数据集详细信息如表1。数据集的图片中包含多名学生对象,姿态也是多样的,本文将抬头注视讲台、身体坐姿端正或看书等状态的学生标注为正常上课状态;将趴在桌上、靠着头等状态的学生标注为上课睡觉状态;将低头、玩手机等状态的学生标注为低头状态。数据集样本如图4所示。

表1 标注图片信息

Tab. 1 Label picture information

数据集	图片数	正常人数	睡觉人数	低头人数	总人数
训练集	240	1 909	646	2 343	4 898
测试集	60	1 256	465	1 585	3 306
样本数	300	3 165	1 111	3 928	8 204

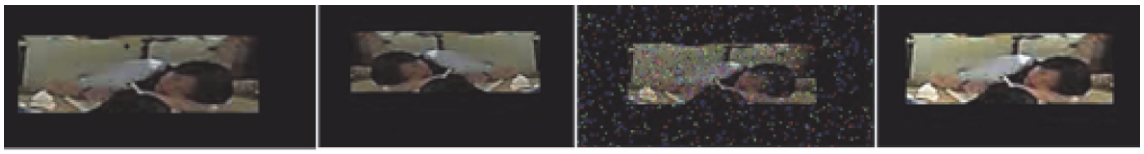
图4 数据集样本

Fig. 4 Dataset sample

3.2 数据增强

由于制作的数据集中,睡觉的学生样本数与低头和正常上课的学生样本数相差过大,容易导致模型出现过拟合问题,然而在实际情况下,睡觉样本的数据集是不易采集的。因此本文采用数据增强技术来增加睡觉类别的图片数量,以避免过拟合问题,从而提高网络性能。本文选取部分图片,将图中的睡觉学生进行裁剪选出并进行增强:对裁剪图片进行左右对称镜像变换;向裁剪出的图片加入高斯噪音;对裁剪出的

图片进行亮度变换,提高图片亮度。增强效果如图 5 所示。



(a) 裁剪 (b) 左右对称镜像 (c) 高斯噪音 (d) 提高亮度

图 5 数据增强效果

Fig. 5 Data enhancement effect

3.3 实验环境

本文实验过程中使用操作系统为 Ubuntu 16.04 的 PC 机,处理器为 Intel Core(TM) i7-6700 @ 3.40 GHz,16GiB RAM,显卡为 NVIDIA GeForce GTX 745。其中训练过程基于开源的 Python 语言和 TensorFlow 实现,预训练模型 Inception-ResNet-v2 源于 Google 公开的 TensorFlow Object Detection API 下的模型。

3.4 训练参数设置

本文采用 mini-batch 方式下的 momentum 梯度下降法,可以在一定程度上加快收敛,减少震荡,设置 0.9的动量参数,考虑到本实验的运行环境,本文设置 BatchSize 大小为 1。本文的网络设置调整图片的最大维度为 512,最小维度为 300。设置 6 000 为最大迭代次数,其中,前 1 800 次设置学习率为 0.003,之后的迭代设置学习率为 0.000 3。在 RPN 的卷积网络层则以 0 为均值、0.01 为标准差的截断的高斯分布随机初始化。对共享的卷积层通过 Xavier 方法进行初始化。

4 结果与分析

为了评估目标检测的结果性能,本文采用查准率均值(mean average precision, mAP)作为标准。mAP 是用于衡量多个类的检测结果,为平均精确度(average precision,AP)的平均值。而 AP 用于衡量单个类别的检测结果,为精确率-召回率曲线与坐标轴所围成的面积。其中精确率(Precision)、召回率(Recall)定义为:

$$\eta_{\mathrm{P}} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}}, \tag{4}$$

$$\eta_{\mathrm{R}} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}}。 \tag{5}$$

式中:N_{TP}表示正确检测的框的个数;N_{FP}表示错误检测的框的个数;N_{FN}表示漏检的框的个数。

4.1 多路特征融合对查准率均值的影响

通过 Faster R-CNN 和多路特征融合改进的 Faster R-CNN 在测试集上进行实验,检测结果如表 2 所示。

表 2 各组实验在测试集上检测的 AP 结果表

Tab. 2 AP results tested on test set for each group of experiments %

算法	AP			mAP
	正常上课学习	睡觉	低头	
Faster R-CNN	74.05	55.80	62.47	64.10
多路特征融合改进的 Faster R-CNN	80.42	71.16	77.38	76.32

从表 2 结果可以看出,多路特征融合改进的 Faster R-CNN 实验结果 AP 值均高于 Faster R-CNN 模型,同时在 mAP 值上提高了 12.22 个百分点。通过融合浅层特征的方式,将更多细节信息的浅层特征融入了

具有高度语义信息的深层特征图中,使得在处理学生课堂监控视频图片的数据集上,具有更好的检测结果。其对应的检测结果样例图如图6所示。图6中,多路特征融合的Faster R-CNN在检测正常上课学习、睡觉和低头的学生行为上,其检测框更准确并且漏检数量更少。通过多路特征融合方式,增强了细节特征信息,提高了检测的结果。

(a)~(c)是Faster R-CNN检测出的课堂行为结果,对应的行为分别是正常上课、睡觉、低头;
(d)~(f)是多路特征融合改进的Faster R-CNN检测出的课堂行为结果,对应的行为分别是正常上课、睡觉、低头。

图6 有无融合多路特征的Faster R-CNN部分样例检测结果对比

Fig. 6 Comparison of partial sample detection results of Faster R-CNN with or without fused multi-path features

将模型数据导出后,随机检测100张图片,计算出检测每张图片所需要的平均时间,检测结果如表3所示。

表3 各组检测1张图片的平均时间

Tab. 3 Average time for each group to test one picture s

算法模型	检测时间
Faster R-CNN	2.103
多路特征融合改进的Faster R-CNN	2.178

从表3可以看出,在本文的硬件条件下,本文算法模型的检测速度略慢于Faster R-CNN,平均每张图片的检测时间需要多花0.075 s。这表明融合浅层特征方式需要消耗更多的运算时间从而提升精度。

4.2 与已有检测算法的对比

为了验证本文算法的适应性,将本文算法在VOC2007数据集上进行训练和测试,与其他算法在该数据集上进行结果对比,结果如表4所示。

表4 不同算法在VOC2007数据集上mAP结果对比

Tab. 4 Comparison of mAP results of different algorithms on VOC2007 dataset %

算法模型	mAP
YOLO[20]	63.4
SSD300[21]	68.0
R-CNN[22]	53.7
SPP[23]	59.2
Fast R-CNN[24]	66.9
Faster R-CNN[25]	69.9
本文算法	71.52

从表 4 可以看出，与 YOLO 算法相比，本文算法的 mAP 值较高，这是由于 YOLO 采用了单个网络结构进行训练和检测，即通过一个卷积网络实现边框回归和分类的任务；而本文算法是基于 Faster R-CNN，经 RPN 网络生成候选区域的基础上再进一步进行检测。与 SSD 算法相比，本文算法的 mAP 值与其相近，且略高于 SSD 算法。因为 SSD 算法采用了多尺度特征融合的方式，通过抽取网络的不同层不同尺度的特征做预测，故 mAP 值与本文算法相近；同时，SSD 算法是基于回归的检测方式，因此查准率均值相比较低。与 Faster R-CNN 相比，本文算法在 mAP 值上提高了近 1.62 个百分点，说明本文算法具有一定的适应性。

4.3 与已有课堂姿态检测方法的对比

在已有的课堂姿态检测方法中，文献[7]、文献[8]、文献[18]和文献[32]分别进行了 SVM 分类器、朴素贝叶斯分类、VGG 网络模型与迁移学习和回归森林法的课堂姿态检测研究，本文提出的基于多路特征融合的 Faster R-CNN 与迁移学习的检测方法与已有课堂学生姿态检测方法比较结果如表 5 所示。

表 5 本文方法与已有课堂姿态检测方法比较

Tab. 5 Our method compared with existing methods for behavioral detection of students' classroom performance

检测方法来源	图像数据来源	特征提取	实验方法	检测类别
文献[7]	华中科技大学模拟课上监控视频	人工特征	Kinect 骨骼向量+SVM 分类器	举手、正坐、趴桌和起立
文献[8]	西安科技大学模拟教室课上拍摄视频	人工特征	Zernike 矩特征、流光特征及全局运动方向等+朴素贝叶斯分类	举手、站立和坐下
文献[18]	燕山大学模拟学生教室课上视频图像	自动特征	背景差分法+VGG 网络模型及迁移学习	正常上课、睡觉、玩手机
文献[32]	华中科技大学教室监控视频	人工特征	纹理特征、几何特征、Gabor 特征等及回归森林法	学生人脸朝向
本文	浙江农林大学课上监控视频图像	自动特征	多路特征融合的 Faster R-CNN 及迁移学习	正常上课、睡觉、低头

从表 5 可以看出，本文与文献[7]、文献[8]、文献[18]和文献[32]的课堂学生姿态检测方法相比的改进在于：

① 本文构建的图像数据集来源于真实情况下的学生课堂监控视频，相比于模拟学生上课的课堂视频，本文数据更符合应用实际中的学生课堂表现常见情况，可在本文建立的真实课堂数据集中增加更多高校的课堂图像，有利于进一步开展课堂学生行为检测研究。

② 文献[7]、文献[8]和文献[32]采用的方法都需要人工手动的方式选取特征，如纹理特征、几何特征等，整体检测结果依赖于手工提取的特征数据的优劣。而本文利用深度网络自动提取课堂姿态特征，充分发挥了卷积网络对数据本质的刻画能力的优势。

③ 本文将数据增强和迁移学习技术应用到基于 Faster R-CNN 的学生课堂行为检测模型中，改善了过拟合问题，提高了模型性能。在 Inception-ResNet-v2 主干网络的基础上，设计融合浅层特征的结构，得到多路特征融合的 Faster R-CNN 检测模型，提高了该模型的精确度。

5 结语

为提高课堂教学质量，保证课堂纪律，本文提出了基于迁移学习和多路融合改进的 Faster R-CNN 的学生课堂行为检测方法。利用浙江农林大学的真实课堂监控视频建立新的数据集，通过手工标注的方式生成超 8 000 个目标对象的实验数据；通过 Inception-ResNet-v2 网络自动特征提取，减少对手工特征提取的依赖；通过 Faster R-CNN 检测学生上课的 3 种状态；通过数据增强，扩充睡觉类别的数据，减轻过拟合；通过迁移学习加快训练速度，提高模型性能；进一步，将网络的深层特征多路融合浅层特征，增强特征信

息,提高检测效果。实验结果表明:本文提出的多路特征融合的 Faster R-CNN 在课堂行为测试集上可提升 mAP 值 12.22 个百分点,且具有一定的适应性,对提高课堂教学质量具有重要意义。

本文未来的工作主要为:一是继续扩充数据集,增加各个学校的课堂图片来源及识别的类型;二是考虑到本文采用图片的形式进行检测,而在实际应用中需要处理视频流,因此后续将对相应的预处理技术进行研究,为更好地应用于实际做准备;三是研究如何进一步优化模型的结构和参数,提升模型检测速度。

参 考 文 献

[1] 胡小玲. 高校课堂管理现状及对策分析[J]. 扬州大学学报(高教研究版),2018,22(3):114-120. DOI:10.19411/j.cnki.1007-8606.2018.03.018.

[2] 秦道影. 基于深度学习的学生课堂行为识别[D]. 武汉:华中师范大学,2019.

[3] 薛新国. S-T 分析法及其在教学中的应用[J]. 江苏教育研究,2019(10B):4-8. DOI:10.13696/j.cnki.jer1673-9094.2019.29.002.

[4] 武天宏. 弗兰德斯互动分析系统在教学中的应用[J]. 现代教育科学,2018(2):101-108,135. DOI:10.13980/j.cnki.xdjykx.2018.02.019.

[5] 王政山. 基于视频分析的学生课堂听课状态的系统研发[D]. 乌鲁木齐:新疆大学,2019.

[6] 魏艳涛,秦道影,胡佳敏,等. 基于深度学习的学生课堂行为识别[J]. 现代教育技术,2019,29(7):87-91.

[7] 张鸿宇. 课堂学习行为测量系统的设计与实现[D]. 武汉:华中科技大学,2016.

[8] 党冬利. 人体行为识别及在教育录播系统中的应用[D]. 西安:西安科技大学,2017.

[9] 郭文诚,崔昊杨,马宏伟,等. 基于 Zernike 矩特征的电力设备红外图像目标识别[J]. 激光与红外,2019,49(4):503-506. DOI:10.3969/j.issn.1001-5078.2019.04.020.

[10] 张顺,龚怡宏,王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用[J]. 计算机学报,2019,42(3):453-482. DOI:10.11897/SP.J.1016.2019.00453.

[11] 李瀚超,蔡毅,王岭雪. 全局特征提取的全卷积网络图像语义分割算法[J]. 红外技术,2019,41(7):595-599,615.

[12] 刘立,詹菡菡,罗扬,等. 尺度不变特征变换算子综述[J]. 中国图象图形学报,2013,18(8):885-892. DOI:10.11834/jig.20130801.

[13] 杨利平,辜小花. 用于人脸识别的相对梯度直方图特征描述[J]. 光学精密工程,2014,22(1):152-159. DOI:10.3788/OPE.20142201.0152.

[14] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2014-09-04) [2020-01-08]. https://arxiv.org/abs/1409.1556.

[15] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions [C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition. Los Alamtios, CA: IEEE Computer Society, 2014: 1-9. DOI: 10.1109/CVPR.2015.7298594.

[16] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. Los Alamtios, CA: IEEE Computer Society, 2016: 770-778. DOI: 10.1109/CVPR.2016.90.

[17] 严严,陈日伟,王菡子. 基于深度学习的人脸分析研究进展[J]. 厦门大学学报(自然科学版),2017,56(1):13-24. DOI:10.6043/j.issn.0438-0479.201609024.

[18] 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[J]. 电子世界,2018(8):97-98. DOI:10.19353/j.cnki.dzsj.2018.08.054.

[19] 周俊宇,赵艳明. 卷积神经网络在图像分类和目标检测应用综述[J]. 计算机工程与应用,2017,53(13):34-41. DOI:10.3778/j.issn.1002-8331.1703-0362.

[20] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. Los Alamtios, CA: IEEE Computer Society, 2016: 779-788. DOI: 10.1109/CVPR.2016.91.

[21] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector [EB/OL]. (2015-12-08) [2020-01-08]. https://arxiv.org/abs/1512.02325.

[22] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition. Los Alamtios, CA: IEEE Computer Society, 2014: 580-587. DOI: 10.1109/CVPR.2014.81.

[23] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916. DOI: 10.1109/TPAMI.2015.2389824.

[24] GIRSHICK R. Fast R-CNN [C]// 2015 IEEE International Conference on Computer Vision. Los Alamtios, CA: IEEE Computer Society, 2015: 1440-1448. DOI: 10.1109/ICCV.2015.169.

[25] REN S Q, HE K M, 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, 2017, 39 (6): 1137-1149. DOI: 10.1109/TPAMI.2016.2577031.

[26] 刘英璇，伍锡如，雪刚刚. 基于深度学习的道路交通标志多目标实时检测 [J]. 广西师范大学学报(自然科学版)，2020，38(2)：96-106. DOI：10.16088/j.jssn.1001-6600.2020.02.011.

[27] 宋光慧. 基于迁移学习与深度卷积特征的图像标注方法研究 [D]. 杭州：浙江大学，2017.

[28] SZEGEDYC, IOFFE S, VANHOUCKE V, et al. Inception-v4, Inception-ResNet and the impact of residual connections on learning [EB/OL]. (2016-02-23) [2020-01-08]. https://arxiv.org/abs/1602.07261.

[29] 秦毅，吴蔚. 基于 CNN 的计算机生成图像识别方法 [J]. 西南师范大学学报(自然科学版)，2019，44(5)：109-114. DOI：10.13718/j.cnki.xsxb.2019.05.018.

[30] 周忠义，吴谨，朱磊. 基于多路特征融合和深度学习的露霜图像分类 [J]. 计算机应用与软件，2018，35(10)：205-210，231. DOI：10.3969/j.issn.1000-386x.2018.10.037.

[31] LING S, MILAN A, SHEN C H, et al. RefineNet: multi-path refinement networks for high-resolution semantic segmentation [EB/OL]. (2016-11-20) [2020-01-08]. https://arxiv.org/abs/1611.06612.

[32] 陈靓影，刘乐元，张坤，等. 学生课堂注意力检测方法及系统：CN201410836650.X [P]. 2015-04-15.

(责任编辑　黄　勇)