




视频分析课堂学生集中学习 监视系统

Mu-Chun Su , IEEE 高级会员, Chun-Ting Cheng, IEEE 会员,
张明庆,  IEEE 高级会员和谢益增 (Yi-Zeng Hsieh) , IEEE 会员

摘要:自动学习反馈监控和分析在现代教育中变得至关重要。我们提出了一个视频分析系统,能够监控课堂学生的学习并为教师提供反馈。如今,学生在课堂上使用笔记本电脑和手机做电子笔记或在线浏览是一种常见的做法。然而,技术的使用也会影响学生的注意力并影响学习,如果控制不当,可能会严重阻碍他们的学习进度。在这项开创性的研究中,我们提出了一种基于非侵入式深度学习的计算机视觉系统,通过提取和推断高级视觉为线索(包括他们的面部表情、手势和活动)来监控学生的注意力。我们的系统可以自动实时协助教练进态感知。我们假设仅 RGB 彩色图像作为边缘设备上的输入和可运系统,以便于部署。我们提出了两个用于学生为分析的视频分析组件:(1)面部分析组件基于 Dlib 面部检测和面部标志跟踪来定位每个学生并分析他们的面部方向、眨眼、凝视和面部表情。(2)活动检测和识别组件基于 OpenPose 和 COCO 对象检测进操作,可以识别举手、打字、接听电话、歪头、办公桌上打瞌睡等8种课堂手势和为。在新收集的真正课堂学生活动数据集 (ICSAD) 上执,我们实现了近 80% 的活动检测率。我们的系统在处理面部和姿势方向时与视图无关,平均角度误差 < 10°。这项工作的源代码位于: <https://github.com/YiZengHsieh/ICSAD>。

索引术语 视频分析、深度学习、学生注意力、为线索、面部检测、地标跟踪、面部方向、面部表情、眨眼、打哈欠、姿势分析、物体检测、活动检测。

稿件于2021年5月21日收到; 2021 年 10 月 14 日修订; 2021年11月6日接受。出版日期2021年11月10日;当前版本的日期为 2021 年 12 月 23 日。这项工作得到了台湾科学技术部的部分支持,拨款 MOST 110-2221-E-019-051、拨款 MOST 109-2622-E-019-010、授予大多数 109-2221-E-019-057、授予大多数 109-2622-E-008-018-CC2、授予大多数 109-2221-E-008-059- MY3、授予大多数 110-2634-F-019- 001、授予多数 110-2634-F-008-005、授予多数 109-2221-E-008-059-MY3、授予多数 107-2221-E-008-084- MY2、授予多数 109-2634-F- 008-007、授予 MOST 109-2221-E-019-057、授予 MOST 107-2221-E-019-039-MY2 和授予 MOST 109-2634-F-019-001,部分由 LSH-NCU 提供联合研究基金会资助 NCU-LSH-109-B-010、资助 NCU-LSH-108-A-007 和资助 NCU-LSH-108-A-009。(通讯作者:谢益增)

国立中央大学计算机科学与信息工程系,台湾 桃园 320317

Ming-Ching Chang 大学计算机科学系
纽约州立大学奥尔巴尼分校,Albany, NY 12222 USA。
Yi-Zeng Hsieh 就职于国立台湾海洋大学电气工程系和海洋工程卓越中心,台湾基隆 20224 (电子邮件:yzhsieh@mail.ntou.edu.tw)。

数字对象标识符 10.1109/TCE.2021.3126877

1558-4127 c 2021 IEEE。允许个人使用,但重新发布/重新分发需要 IEEE 许可。
有关更多信息,请参阅 <https://www.ieee.org/publications/rights/index.html>。

授权许可使用仅限于:西安理工大学。于 2023 年 10 月 12 日 06:27:26 UTC 从 IEEE Xplore 下载。存在限制。

一、简介

移动技术的发展势在必行
时间影响我们的日常生活。然而全球范围内

智能手机的日益普及被视为好坏参半,尤其是对青少年而言。最近的研究表明,在教育和学习中,学生的注意力会随着智能手机和笔记本电脑的使用而降低[2],[25]。是否允许或规范课堂技术使用的问题正在成为一个主要争论。人们关心的是如何在最好地利用技术力量的同时让学习者免受不必要的干扰。针对数字干扰的规范化和政策可能会对学习为、出勤率和学习成果产生长期影响。从教师的角度来看,技术使用造成的学生分心也很重要。如果能够及时意识到学生分心的情况,则可以立即采取行动来提高学生的注意力。在这项研究中,我们提出了一种用于学生学习和为注意力监控的非侵入式计算机视觉系统。通过检测和识别学生的面部注意力和身体姿势,我们的视频分析方法可以识别学生的为并监督课堂注意力状态。

所提出的视频分析系统由安装在典型教室环境中的摄像机和边缘计算机组成。摄像机拍摄课堂上学生的 RGB 彩色视频。我们的系统旨在识别学生的疲劳程度以及课堂为和活动,以估计整体注意力状态。该系统可以通过估计每个学生坐标系中的面部和身体姿势方向来自动处理视图变化。

我们开发了两个用于学生为和注意力监控的视频分析组件。我们的系统利用流的开源深度神经网络 (DNN)包进有效的视觉检测和跟踪,见图1:(1)面部分析组件基于 Dlib [11]面部检测面部标志跟踪进操作,可以定位每个学生并分析他们的面部方向、眼神交流、凝视和面部表情。(2)基于OpenPose [5]运的活动检测和识别组件可以分析身体姿势以进活动检测。我们开发了一个 DNN 模型来检测和识别多达 8 种与注意力和分心分析相关的学生为。

这些包括写笔记、打字、打电话发短信、举手、歪头、在办公桌上打瞌睡、单手低头和打电话。活动检测准确度可以提高

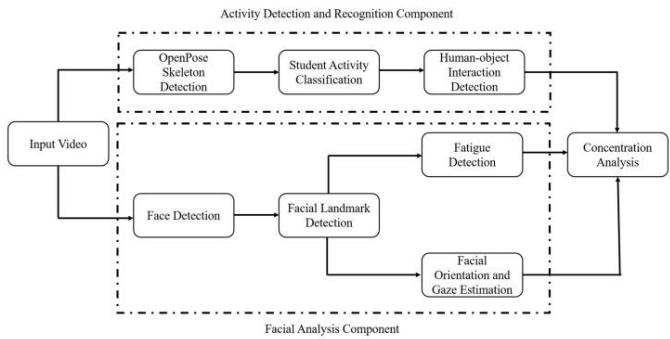


图 1. 拟议的课堂学生集中度分析流程概述。

通过使用 YOLO 视觉对象检测器 [20] 检测常见的课堂对象（例如手机、笔记本、计算器），更好地识别人与对象的交互。

本文的贡献总结如下。

- 1)我们提出了一种自动、非侵入式视频分析系统,通过大量的实验和评估来监控课堂学生的注意力。
- 2)提出的面部分析流程包括面部检测和地标跟踪,用于一系列全面的面部表情分析,包括每个学生的每帧面部方向、眨眼、凝视和面部表情分析的检测。
- 3)所提出的学生活动分析流程基于学生骨骼姿势的检测和常见课堂对象的识别。该系统可以通过观点不变分析来识别多达 8 种类型的学生活动。
- 4) 在新收集的真实课堂学生活动数据集 (ICSAD)上实验。混淆矩阵结果表明,我们的系统可以实现近 80% 的活动检测率,平均角度误差 < 10°。 TCSAD 数据集由 10 个学生科目的 400 个视频片段,3 个视角的 8 个课堂活动组成,这些数据将在论文发表时发布。对面部方向和表情（打哈欠检测）的评估也在 YawDD [1] 数据集上实验。
- 5)所提出的分析算法可以在轻量级边缘设备上使用标准 RGB 相机实时运行,无需将视觉数据存储或传输到远程云。这种在线边缘计算设计可确保轻松部署,并可在很大程度上缓解隐私问题。

本文的结构如下。第二部分回顾背景知识和文献调查。第三节描述了我们的面部分析流程,它可以提取学生的注意力和潜在的疲劳线索。第四节描述了我们的姿势和活动分析流程,可以直接监控课堂上学生的为、手势和动作,以推断注意力状态。第五节描述了实验验证、评估指标和结果

开放数据集和我们新收集的课堂学生活动数据集。第六节总结了这项工作。

二.背景

A. 浓度估算的回顾

专注通常被定义为持续关注地接收外部信息[10]。在这项工作中,我们将学习注意力视为在课堂上保持长期注意力。由于学生的学习注意力与他们的学习成绩密切相关,因此教育研究中大量的工作都在研究如何最好地评估和保持学习注意力。人类注意力的经典研究主要依赖于基于侵入式传感器的方法,例如脑电图 (EEG)传感。在[6]和[14]中,支持向量机 (SVM)模型被训练来识别脑电图脑电波,以评估学生的注意力和专注力。这些基于传感器的方法很难在现实世界的课堂环境中实施。

随着深度学习和人工智能的最新进步,视频传感方法变得越来越多样化。

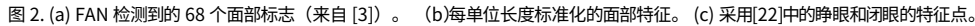
这种视频分析方法是侵入式的,现在可以可靠地识别丰富的视觉信号以进行情感和情绪分析。然而,隐私和道德问题伴随着这些基于视频的方法而出现[9]。在[19]中,三个摄像头用于捕捉教室环境中的学生视图,其中学生的轻微动作和身体动作仅被粗略估计。由于视觉信息不足,该方法无法可靠地估计学生的注意力。在[26]中,学生的注意力被估计为五个级别。通过基于 Kinect 的动作检测和视频中每一秒学生注意力水平的手动估计,并开发了用于学生注意力分类的机器学习模型。

我们的模型与[26]中提出的模型之间的主要区别如下: (1)我们的方法适用于单个 RGB 相机（易于部署）,而他们需要 Kinect 传感器才能工作。 (2)我们的系统估计了学生详细的注意力测量（包括疲劳状态、眨眼、面部方向、目光注视以及八种课堂手势和为）,而[26]仅估计了三级量表（高、中、低）的学生注意力。我们的系统为报告的学生注意力状态提供丰富的上下文信息和可解释的线索。 (3)对于检测涉及身体动作的学生活动,我们检测了八种课堂学生活动（举手、打字、接听电话、歪头、书桌打瞌睡等）,而[26]仅检测四项学生活动（观察幻灯片、写笔记、向前倾斜和向后倾斜）。

我们的监控系统（目前状态）是在课堂环境中实施的,为教师提供分析结果以供课后查看。作为未来的工作,我们计划在在线视频会议环境中实施我们的系统,以便教师可以收到学生的实时关注信息。

B. 头部/身体姿势方向估计的回顾

头部姿势或身体方向可以用欧拉角表示,即偏航（沿y轴旋转）、俯仰



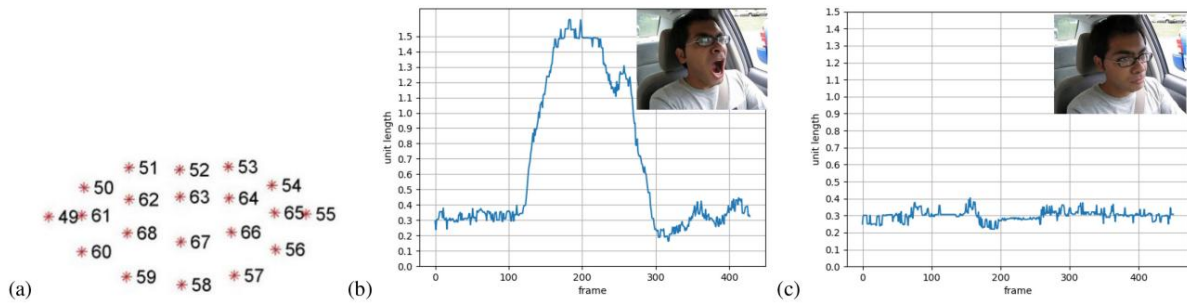


图 3.打哈欠检测: (a) 放大到[3]中图 2a 中的嘴部标志点。(b)显示了打哈欠的情况,(c)显示了典型的闭嘴情况,其中y轴描绘了方程式中的Mouth。(6) x 轴是视频帧(时间)。

$$p \sim x(t) = \frac{p^x(t)}{du(t)}, \quad (4)$$

其中 $p \sim x(t)$ 表示在时间t进平移和缩放归一化后的每个特征点x。

B. 闭眼和眨眼检测

我们采用[22]中的快速方法,根据图2c中的6个眼睛标志点(p1到p6)来检测闭眼眼睛,以计算眼睛长宽比(EAR):

$$\text{EAR} = \frac{|p2 - p6| + |p3 - p5|}{2|p1 - p4|}. \quad (5)$$

最低的EAR值用于在阈值化后确定闭眼。具体来说,如果在150ms的预设时间内 $\text{EAR} < 0.2$,则检测到闭眼。如果EAR值在该时间段内振荡,则检测到眨眼。

C. 打哈欠检测

打哈欠时的吸气和呼气会导致嘴巴张开一段时间。因此,可以通过检测张嘴和闭嘴来稳健地识别打哈欠,如图3所示。我们计算上唇和下唇中点之间的垂直距离dMouth(分别为图3a中的p52和p58),以确定打哈欠张嘴动作单元:

$$d\text{Mouth} = |p52 - p58|. \quad (6)$$

在我们的实验中,对于非打哈欠的情况,在地标归一化后,dMouth通常约为0.2至0.4个单位。打哈欠时,dMouth在100帧内增加到约0.5个单位。

当dMouth超过肩宽0.5的阈值时,我们计算打哈欠的开始。如果该打哈欠状态超过 $d\text{Mouth} > \text{肩宽}$ 的时间段,则检测到打哈欠动作单元。

D. 面部方向和注视估计

我们使用欧拉角估计面部和凝视方向。具体来说,如图4a所示,偏航角方差被计算为鼻尖(图2a中的p34)和两个眼角(p37和p46)之间的距离。在图4b中,与深度信息相关的俯仰角方差被计算为鼻尖(p34)到两个嘴角(p49和p55)之间的距离。

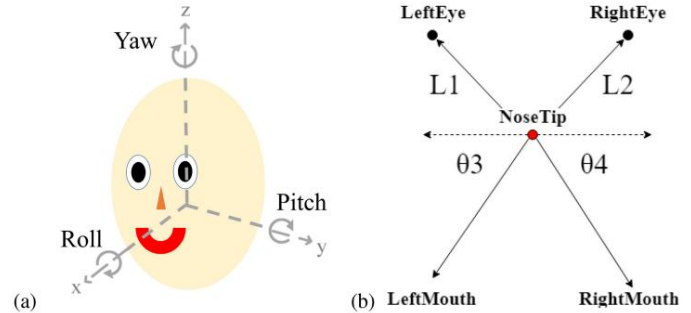


图 4. (a)头部姿势和面部方向估计的偏航、俯仰和原始(图采用自[24,图5])。(b)Yaw表示方向角。面部朝向角度估计特性示意图。

再次使用上述五个面部特征点(左眼角p46、右眼角p37、鼻尖p34、左嘴角p55、右嘴角p49)通过滚动角方差计算面部方向。接下来我们详细描述面部方向估计过程。

面部标志点的距离方差: Yaw和Pitch角度的变化会影响左右眼到鼻尖的距离,以及左右嘴角到鼻尖的距离。因此,本研究采用眼角和嘴角作为四个特征点来计算它们与鼻尖之间的距离。计算过程表示为:

$$dx = px - p\text{NoseTip}, \quad (7)$$

其中dx是特征点到鼻尖的距离;px为特征点,pNoseTip为鼻尖;L1、L2为左右眼到鼻尖的距离,如图4所示。

角度变化的点特征: Yaw和Pitch的角度变化是通过眼睛到鼻尖的距离、嘴巴的水平角度和鼻尖周围的变化来识别的,而Roll角度的变化是通过眼睛到鼻尖的距离的变化来识别的。左右眼角与鼻尖成水平角。这项研究使用与眼睛相关的角度来确定面部方向。嘴巴周围的四个特征点是测定中所表征的水平角的尖端,利用等式(8)和(9)导出:

$$u = px - p\text{NoseTip}, \quad (8)$$

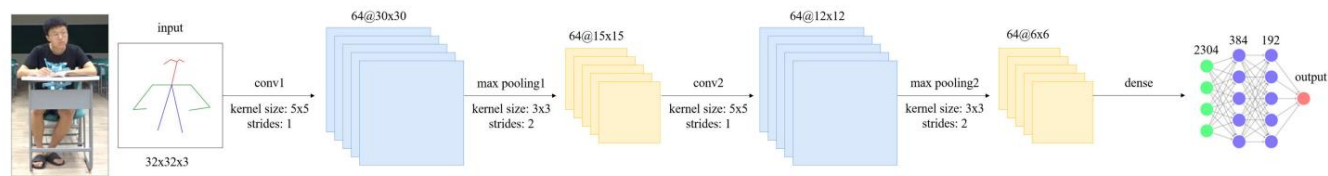


图 5. 输入图像和学生姿势骨架被输入到所提出的 CNN 中以进课堂活动识别。

表一
学生面部朝向计算

Student focus	Student pitch	x Student yaw
Left blackboard	0° ~ 30°	-45° ~ -15°
Center blackboard	0° ~ 30°	-15° ~ 15°
Right blackboard	0° ~ 30°	15° ~ 45°
Student's desktop	-30° ~ 0°	-45° ~ 45°

$$\theta x = \cos - 1 \frac{u \cdot v}{|u||v|}, \tag{9}$$

其中u是鼻子的特征向量， v是水平向量， θx 是鼻尖与特征点之间的水平角度。图4中以嘴部两侧为例,虚线为水平向量， $\theta 3$ 和 $\theta 4$ 为鼻尖与嘴部左右两侧之间的角度。一旦估计出特征向量,就可以计算出学生的面部方向。表 1 列出了使用所提出的方法计算的学生面部方向。

四.姿势和活动分析

学生的注意力集中程度可以通过他们的身体为来确定。例如,用一只手支撑头部坐着可能表明学生感到无聊。本研究采用OpenPose提取骨架特征向量,然后将其用作所提出的神经网络的输入来评估学生的为。该方法涉及多个过程,如下所述。

我们使用 Openpose 来提取骨骼信息。 Openpose是一种开源的骨架姿态估计方法[5]。提取骨骼信息后,对其进归一化,然后用作神经网络的输入,以识别身体位置和运动。采用滑动窗口来增加系统的稳定性。最后,将姿势信息与运动检测、分心为检测和面部方向相结合,以分析受试者的注意力水平。

A. 人体姿态估计和姿态特征归一化

骨架大小与相机的位置有关,并且会根据拍摄对象距相机的距离而相应变化。除了每个人的骨骼都不同之外,这意味着骨骼信息必须标准化以避免错误。标准化包括数据移位和数据缩放。

在数据移位中,以颈为原点,从颈位置减去每个点的位置来计算新的位置,如公式 (10)所示:

$$j x(t) = j x(t) - j neck(t), \tag{10}$$

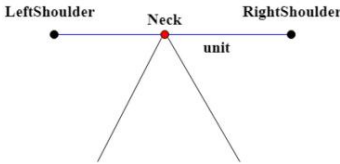


图 6.每单位长度归一化的骨架信息。

其中jneck(t)是时间 t 时的颈部坐标; j x(t)为t时刻特征点x平移得到的新坐标; jx(t)是特征点x在时刻的新坐标。

在数据缩放中,肩长是单位长度。使用等式 (11)和 (12)将原始坐标除以单位长度,产生新坐标:

$$Lunit(t) = jleftShoulder(t) - jrightShoulder, \tag{11}$$

$$j_x(t) = \frac{j x(t)}{junit(t)}, \tag{12}$$

其中Lunit(t)是时间 t 的单位长度;平移和缩放后的j坐标 $j_x(t)$ 是新的在图6中,蓝线是肩长。

B.学生活动分类使用图像作为卷积神经网络

络的输入会导致一些由背景干扰或其他物体引起的错误。文献[16]中的错误识别可能就是由于这个原因。在[18]中,数据集被分为四类,但这显然不能完全呈现课堂上的情况。

我们提出的方法采用从课堂上的八种常见手势中提取的骨架特征作为卷积神经网络的输入以提高性能。骨骼图像的种类如图11所示。

本研究的运动场景是坐着的情况,因此忽略了下半身的骨骼。标准化骨架分为三个部分。第一部分是头部,包括左耳、右耳、左眼、右眼和鼻子等特征。第二部分是上半身,包括左腕、左肘、左肩、颈部、右肩、右肘、右腕等特征。第三部分是下半身,包括左臀部和右臀部。图5显示了映射到人体的骨骼。

CNN 架构:所提出的卷积神经网络 (CNN)的架构如图 5 所示。输入图像是调整大小为 32x32 的骨架图像。 CNN 由两个卷积层、两个池化层和一个全连接层组成。内核大小为 5x5,步幅大小为 1。

池化大小为3x3,步幅大小为2。采用零填充。全连接层有2个隐藏层

1个输出层。第一隐藏层由 384 个神经元组成,第二隐藏层由 192 个神经元组成。输出层由手势数量决定。

C. 检测人与物体的交互

由于骨架图像的某些动作难以区分,因此需要额外的信息来确定这些动作。本研究中最常见的场景是学生使用书籍、笔记本、智能手机等物体。

根据做出的手势来识别骨骼;然后通过对一些对象进行分类来确认手势。

因此,本研究提出了两种对象分类方法来提高手势识别的准确性。

第一种方法是图像分类,搜索对象的区域,然后根据分类器对对象进行分类;本研究的课堂场景中识别了 8 个手势。这项研究通过识别两只手对物体执的手势,对使用笔记本和智能手机发短信进笔记书写进了分类。一只手支撑头部,另一只手拿着智能手机的手势属于单手使用物体的手势。

本研究介绍了两个对象片段区域类别如下。

首先是两只手放在桌子上的类别。两只手的中心也是物体的中心,如式(13)所示:

$$\frac{jleftWrist(t) + jrightWrist(t)}{2} = omid(t), \tag{13}$$

其中omid(t)是时间 t 时对象的中心; jLeftWrist(t)为t时刻左手腕中心坐标; jRightWrist(t)为t时刻右手腕的中心坐标。

然后,为了覆盖物体的范围,本研究利用肩长的中心来扩展范围,如式 (14)、表 II (15)和 (16)所示:

$$Lshoulder(t) = jleftShoulder(t) - jrightShoulder(t), \tag{14}$$
$$otopleft(t) = omid(t) - \frac{Lshoulder(t)}{2}, \tag{15}$$
$$obottomRight(t) = omid(t) + \frac{Lshoulder(t)}{2}, \tag{16}$$

其中Lshoulder(t)是时间 t 时的肩长; jleftShoulder(t)和jrightShoulder(t)分别是时间t时的左肩坐标和右肩坐标; otopLeft(t)和obottomRight (t)分别是t时刻的左上角和右下角坐标。

第二类是单手使用物体。
使用物体时,参考点通常是手腕。
然而,在这种情况下,由于对象的范围位于手腕的顶部,因此参考以对象的中心为中心向上移动了肩长的一半。右手或左手的区别是通过手的y坐标来区分的。

方程 (17)和 (18)用于此类:






$$或(t) = \begin{cases} j左手腕(t), & j左手腕(t) < j右手腕(t) \\ j右手腕(t), & j右手腕(t) < j左手腕(t) \end{cases}, \tag{17}$$

$$Lshoulder(t) = jleftShoulder(t) - jrightShoulder(t), \tag{18}$$

$$omid(t) = ox_{ref}(t), \quad Oref(t) = \frac{Lshoulder(t)}{2}, \tag{19}$$

表二

对象区域搜索的两类

motion	Writing note	Using notebook	Texting smartphone
object	note	notebook	smartphone
result			
motion	One hand head	Answering smartphone	
object	none	smartphone	
result			

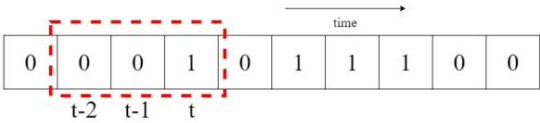


图 7. 滑动窗口。

其中oref(t)是时间 t 的参考坐标;和
jleftWrist(t)和jrightWrist(t)分别是时间 t 时左右手腕的 y 轴。由于物体尺度较小,本研究将范围从物体中心扩展到肩长的一半,如式 (20)和 (21)所示:

$$otopLeft(t) = omid(t) - \frac{LShoulder(t)}{2}, \tag{20}$$

$$obottomRight(t) = omid(t) + \frac{LShoulder(t)}{2}, \tag{21}$$

本研究采用的分类器是VGGNet[21],它在ImageNet大规模视觉识别挑战赛 (ILSVRC)中获得一等奖。本研究采用表中的架构D,输入图像尺寸为 224×224。有16个卷积层和4个池化层,与由2个隐藏层和1个输出层组成的全连接层连接。架构D也称为VGG16。

第二种对象分类方法包括对象检测和定位对象的位置,因此采用 You Only Look Once (YOLO) [20] 方法。其优点是处理速度更快,同时仍保持良好的精度。YOLO的主要特点是采用CNN模型来训练整个图像。在预测整个图像后,图像被分割成许多区域,每个区域的中心可以确定许多边界框,每个边界框都有置信度分数和类别。

每个边界框都有 5 个值,包括中心坐标、高度、宽度和置信度得分。

滑动窗口后处理:这项研究识别了相机输入的每一帧中的每个动作。为了保证系统稳定性,并避免噪声干扰,采用滑动窗口和“胜者通吃”的方法来实现顺序运动检测。使用滑动窗口,可以识别 3 帧中的每个动作。帧t - 2、t - 1 和t分别被识别为 0、0 和 1。然后使用“赢家通吃”方法将操作识别为 0。

表三
疲劳情况

Situational description	Fatigue behavior	Facial orientation	Motion recognition
Non-class distraction for a long time		Desktop facing	Laptop usage, phone texting
Insufficient motivation, bored by class content	Closed eyes, blinking, yawning		Call answering
			Sleepy, one-handed head

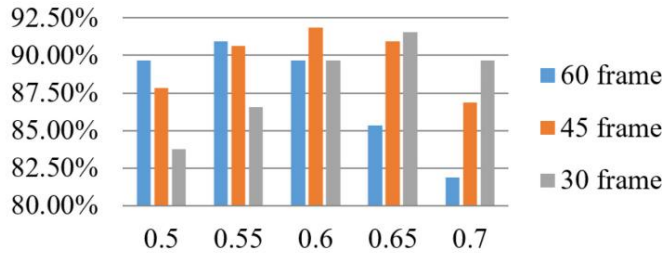


图 8. 打哈欠检测结果。

D. 浓度分析

这项研究将注意力分为两个级别:正常和疲劳。首先,假设学生的初始注意力是正常的。在对面部信息进行分类并确定运动类别后,如果系统检测到主体处于疲劳状态,则将其状态从正常浓度水平改变为疲劳浓度水平。表三描述了疲劳情况。

五、实验验证及结果

A. 哈欠检测评估

该方法可以通过打哈欠特征提取来检测打哈欠为。实验表明,该方法能够准确检测打哈欠为,并使用开放数据在现实情况下验证该方法。

YawDD开放数据集[1]包括不同光照环境下的两种视频。主要场景是一个人开车。第一个数据集由放置在后视镜下方的摄像机捕获的镜头组成;每个测试人员都会通过三种不同的摄像机视图进行测试:正常、说话和打哈欠。共有320部电影。第二个数据集由放置在仪表盘上的摄像机拍摄的镜头组成,共有 29 部影片。

打哈欠检测结果:打哈欠检测实验确定镜头中是否出现打哈欠为,并在不同时间段进。通过分析距离和时间阈值来测试打哈欠影片。最佳距离为40像素;不同时间阈值有 45 帧。45 帧每帧长约 1.5 秒。失败案例发生在口鳞尺寸太小,或者打哈欠为被手遮住时,如图8所示。

B. 面部朝向评估

我们使用[7]中的Pointing 04数据集进面部方向训练和测试的实验

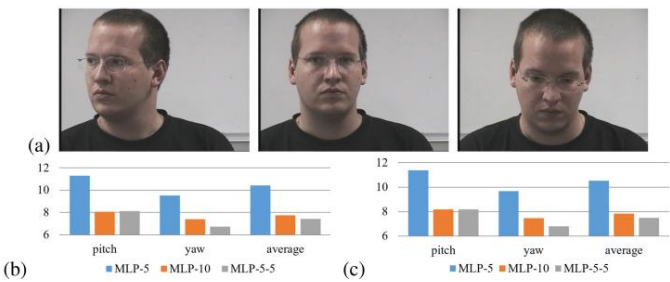


图 9.(a) 面部方向。(b) 每种角度类型的训练误差。(c) 每种角度类型的测试误差。

表四
课堂景观

Shot angle	View	Shot angle	View
Front-Rear		Left 45°	
Front-Far		Right 45°	
Overlook			

估计器网络。该数据集包括 15 名测试人员;每个测试者被录制两个不同的面部朝向视频,其中偏航角在-90度到90度之间,俯仰角在-90度到90度之间。总共有2940个视频。

本研究采用5个面部特征(左眼角、右眼角、鼻尖、左嘴角、右嘴角),调整角度范围作为实验数据,使得偏航角在-45度到45度,俯仰角在-30度到30度之间。

角度以15度间隔变化,共有1050个视频。采用5折交叉验证,通过5次结果的平均值来训练模型。多层神经网络有两个隐藏层架构:5,10 和 5,5。使用 5,5 架构实现了最佳性能,但俯仰误差大于偏航误差。仰视和俯视所涉及的特征并不明显;然而,如图2和3所示,误差小于5-10度。 9b.c.

C. 活动检测和识别评估

本研究构建了 5 个不同的场景,涉及两个距离(远和近)、两个水平角度(-45 度和 45 度)和两个垂直角度(向上和向下),如表 IV 所示。

课堂学生活动数据集 (ICSAD): 共有 10 名受试者参与实验,其中 6 名男性和 4 名女性,如表 V 所示。该实验总结了课堂环境中的 8 种常见为:写笔记、使用笔记本电脑,用手机发短信,拿着东西

表五
课堂科目

#	Gender	Height	#	Gender	Height
1	M	170	6	F	170
2	M	172	7	F	162
3	M	175	8	M	172
4	F	162	9	M	170
5	F	157	10	M	170

Writing notes:



Using laptop:



Phone texting:



Raising hand:



Crooked head:



One-handed head:



Desk napping:



Phone answering:



图 10. 课堂活动。

一手侧头,一手支撑头,睡觉,接电话。该方法录制了5次,动作如图10所示。实验研究共录制了400部电影。

使用 Openpose 方法提取三个摄像机观察 (0、-45 和 45 度)的骨骼图像,并对骨骼图像进行归一化以输入 CNN,以便对每个运动进行分类。骨骼图像的种类如图11所示。

Note writing:



Using laptop:



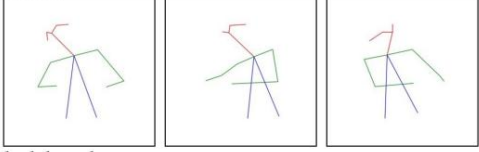
Phone texting:



Raising hand:



Crooked head:



One-handed head:



Desk napping:



Phone answering:



图11.课堂活动-OpenPose人体骨骼检测结果:第一列: 0°,第二列:右45°,第三列:左45°。

由于在实际应用中不可能预先记录每个学生的动作数据然后使用神经网络进行训练,因此预训练的数据就显得尤为重要,以保证预训练的神经网络模型能够根据不同的人的实际情况来应用。

因此,这个实验测试了论文在不同现实场景中呈现的特征。该实验还增加了在不同人身上提升原始图像和骨骼特征的能力。

表六
CNN的混淆矩阵

		Prediction classification							
Ground truth	Note writing	47.35	15.64	14.26	0.23	5.74	12.19	3.57	1.02
	Typing	17.08	59.59	12.89	0.86	6.45	0.00	0.44	2.69
	Phone texting	21.66	12.51	51.98	0.00	4.32	6.51	2.25	0.77
	Raising hand	0.03	1.21	0.00	96.00	1.20	0.04	0.28	1.24
	Crooked head	7.14	7.39	12.00	0.81	56.71	4.04	11.22	0.69
	Desk napping	2.16	0.62	0.50	0.79	1.95	93.01	0.53	0.44
	One-handed head	5.14	0.14	2.28	0.09	9.24	1.70	65.30	16.11
	Phone talking	1.49	4.63	2.21	0.06	1.01	0.00	15.68	74.92

在这个实验中,不同人的数据并不是特别混合,因此训练和测试时使用的受试者是不同的人,5折交叉验证分为两个人的5等份。

特征比较使用原始图像和骨架图像。使用原始图像的框架是AlexNet,而使用骨架图像的框架是所提出的CNN。两种图像在训练时的识别率几乎都是100%,但在测试过程中,原始图像的识别率不足60%,而骨骼图像的识别率提高了10%左右超过原始图像。这一结果证明骨架特征更受不同人群的欢迎,并且比原始图像更好,但识别率低于70%,这意味着系统稳定性较差。

需要从混淆矩阵中的测试结果中识别出动作类别的问题。

测试混淆矩阵是所提出的 CNN 的结果,如表 VI 所示。在识别笔记、使用笔记本和在智能手机上发短信等为中都会出现错误,因为这些动作中的骨架图像都非常相似:都涉及使用桌子上的物体。将头放在一只手上和接听电话之间也会发生混淆,因为两个骨架也非常相似。因此,有必要进对象检测来防止这种混淆并提高识别精度。

D.人与物体交互评估上述实验结果表明需要进物

体识别以提高识别结果,因此所提出的物体识别方法有助于区分相似的骨架动作。首先,子骨架被定义为同一类别中的相似动作,然后使用对象识别来区分它们。本研究采用YOLO目标检测方法,并使用COCO数据集[13]作为训练数据。定义了三个对象:书籍、笔记本电脑和手机。

运动检测结合物体检测方法提高了识别率,如表七所示。

准确率提升10%,测试准确率从68%提升至77%。

写笔记、使用笔记本和单手支撑头部动作的表现有所改善,但在智能手机上打字和单手接听电话的表现仍然较差

表七
不同个体的准确率

Motion recognition class	Testing accuracy
Skeleton feature	68.10%
Skeleton + object detection	77.89%

表八
骨架图像与物体结合的混淆矩阵
检测

		Prediction classification							
Ground truth	Note writing	76.67	9.85	0.00	0.06	4.73	5.82	2.74	0.12
	Typing	2.00	90.25	0.00	0.28	3.75	0.76	1.72	1.26
	Phone texting	20.85	4.98	64.11	0.12	2.11	4.34	2.10	1.40
	Raising hand	0.56	0.08	0.17	90.84	7.06	0.90	0.32	0.07
	Crooked head	27.62	0.83	4.59	1.91	54.03	6.33	4.12	0.57
	Desk napping	2.48	1.90	0.00	0.59	0.51	93.13	0.45	0.94
	One-handed head	3.78	0.36	0.52	0.32	9.45	3.21	82.36	0.00
	Phone talking	8.44	0.22	0.00	1.91	2.14	2.75	12.76	71.78

表九
角度改善结果

Camera angle	Training accuracy	Testing accuracy
Front rear	94.43%	85.64%
Right 45°	94.04%	69.65%
Left 45°	94.49%	78.69%
Front-far	95.59%	81.77%
Front-overlook	93.91%	79.38%
Average	94.49%	79.03%

由于智能手机的面积较小,导致握持智能手机的手被遮挡。物体识别后,由于每个人的头部倾斜方向不同,倾斜头部运动的识别精度没有提高,如表八所示。

角度提升测试:因教室环境不同,摄像头设置可能无法捕捉到学生正面的动作,且不保证学生面对摄像头。因此,本实验测试了不同拍摄角度下是否能够成功识别。收集数据记录时,每个参与者都从5个不同的角度进记录:分别是正面(近、远、俯视)和侧面(左45度和右45度)。本实验和数据混合在一起,然后分成5等份,基于5个角度进5倍交叉验证。

表九显示了五个角度和平均个体测试识别率。分别将每个角度作为测试数据,其余四个角度作为训练数据。

结果表明,右45度角和左45度角的识别率较差,因为每个骨骼图像的侧视图和正视图不同,因此难以正确分类。

滑动窗口测试:本研究提出使用滑动窗口方法来提高识别稳定性。

该实验网络架构使用CNN,测试数据集记录了三个模拟场景,由八个动作组成,每个动作持续约5秒。

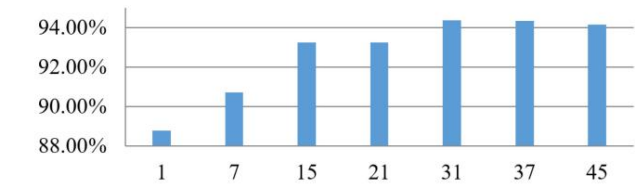


图 12. 使用不同滑动窗口大小的识别精度。



图 13.来自我们的 ICSAD 数据集的真实课堂测试示例的两个视图。

表十
真实世界测试场景的混淆矩阵（单位：百分比）

		Prediction classification							
		Note writing	Using laptop	phone texting	Raising hand	Crooked head	Desk napping	One-handed head	Phone talking
Ground truth	Note writing	89.33	2.22	0.00	0.00	0.00	8.44	0.00	0.00
	Typing	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone texting	13.33	0.00	78.57	0.00	8.10	0.00	0.00	0.00
	Raising hand	0.00	0.00	0.00	97.12	0.00	2.88	0.00	0.00
	Crooked head	3.32	0.00	0.00	0.00	69.10	25.57	0.00	0.00
	Desk napping	1.00	0.00	0.00	0.00	0.00	99.00	0.00	0.00
	One-handed head	0.00	0.00	0.00	0.00	0.00	0.00	83.27	16.73
	Phone talking	0.00	0.00	0.00	0.00	0.00	0.00	1.44	98.56

由于这8个动作都是静态运动,短时间内不会发生变化,因此本次实验窗口测试尺寸较大,以便比较7、15、21、31、37、45和1帧组成的窗口的性能（不使用滑动窗口），如图12所示。当不使用滑动窗口时,识别率最低,而当窗口为31帧时,识别率最高。因此,31帧滑动窗口最准确地识别了本研究中定义的动作；31帧相当于0.5秒。

真实场景测试:以上实验基于有限的数据进行训练和测试。然而,为了测试所提出的方法在分类结果的实际上下文中的识别性能,本研究记录了

真实教室中的一些视频,并识别本文定义的动作,如图13所示,其中红色框表示学生使用笔记本,蓝色框表示学生用一只手支撑头部。在图13b中,红色框表示学生正在写笔记。本实验的网络架构是所提出的CNN,滑动窗口大小为31。

表X中的混淆矩阵显示平均识别率为89.37%。倾斜头部的动作与头枕在桌子上睡觉的动作相混淆,因为倾斜头部的动作使骨骼显得更小。由于智能手机对象太小,在智能手机操作上打字会导致错误。

锯。结论

本文提出了一种确定学生学习焦点深度的深度学习方法。研究通过分析学生的面部信息和动作为,重点关注参考程度,以便教师了解学生当前的学习状况。这些信息可用于调整教学安排和方法,以提高学生在课堂上的注意力。

利用面部信息,本文设计了眨眼和打哈欠两种疲劳为的检测,使得识别学生何时处于精神不适状态成为可能。此外,所提出的方法还估计了面部方向角度,从而识别学生何时分心。本研究还定义了学生常见的八种动作,其中四种是使用物品的动作。

在使用电子设备方面,所提出的方法能够使用面部方向数据来识别学生何时在课堂上查询相关材料,或何时因分心而盯着屏幕。一般为可以用来判断学生是否积极参与课堂。

设计了三个实验来测试面部和骨骼特征的可靠性。片中打哈欠为识别准确率超过90%。面部方向估计的误差在10度或更小。不同角度、不同人体识别率的识别操作,准确率接近80%,最后结合滑动窗口机制,实景测试识别率接近90%。上述实验结果证明,该方法能够为教师提供准确的学生学习状况信息。

本研究以面部信息和动作为作为注意力分析的基础。面部方向估计可用于训练CNN将2D面部特征点转换为相应的3D模型。更准确的估计面朝角落和动作为,虽然8种动作的记录已经包含了大部分学生动作;不可避免地还会有除这些之外的动作,那么我们就需要确定未定义的动作。

参考

[1] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi and B. Hariri, “YawDD: 打哈欠检测数据集”, Proc. ACM多媒体系统。会议。（MMSys），2014年,第24-28页。[在线的],可用: <http://dx.doi.org/10.21227/e1qm-hb90>

[2] NA Attia,L. Baig,YI Marzouk 和 A. Khan,“技术和干扰对本科生注意力的潜在影响”,巴基斯坦医学杂志.科学,卷。 33,没有。第 4 页.修订版 860–865,

[3] A. Bulat 和 G. Tzimiropoulos,“利用有限资源进入体态姿势估计和人脸对齐的二值化卷积地标定位器”,Proc.国际.会议.计算.维斯, 2017 年,第 3726–3734 页。

[4] A. Bulat 和 G. Tzimiropoulos,“我们距离解决 2D 和 3D 人脸对齐问题还有多远? (以及 230,000 个 3D 面部标志的数据集),” Proc.国际.会议.计算.维斯, 2017 年,第 1021–1030 页。

[5] Z. Cao,GH Martinez,T. Simon.S. Wei 和 YA Sheikh,“OpenPose:使用部分亲和力和场进实时多人 2D 姿势估计” IEEE 传输.模式肛门.马赫.英特尔,卷。 43,没有。 1,页。 172–186,一月。 2021年

[6] C.-M.陈,J.-Y.王,和C.-M. Yu,“通过使用基于脑电波信号的新型注意力感知系统来评估学生的注意力水平”, Brit. J.教育.技术,卷。 48,没有。 2,第 348–369 页,2015 年。

[7] N. Gourier,D. Hall 和 JL Crowley,“通过对显着面部特征的稳健检测来估计面部方向”, Proc.国际.会议.模式识别。(ICPR) 国际.车间视觉观察.指示手势, 2004 年,第 1–9 页。

[8] K. Hara 和 R. Chellappa,“通过分类生长回归森林:在物体姿态估计中的应用”, Proc.欧元.会议.计算.维斯, 2014 年,第 552–567 页。

[9] LP Hartman,“技术与道德:工作场所的隐私”, Bus. 社会出版社,卷。 106,没有。 1,页。 2004 年 1 月 27 日。

[10] W.詹姆斯,《心理学原理》。美国纽约州纽约市:亨利 霍尔特,1890。

[11] DE King,“Dlib-ml 机器学习工具包”, J. Mach.学习.资源, 卷.第 10 页。 1755–1758,2009 年 12 月。

[12] C. Li,A. Pourtaherian,L. van Onzenoort,WETA Ten 和 PHN de With.,“用于实时和远程不适检测的婴儿监控系统”, IEEE Trans.消耗.电子,卷。 66,没有。 4,第 336–345 页,2020 年 11 月。

[13] TY Lin等人,“Microsoft COCO:上下文中的常见对象”, Proc.欧元.会议.计算.维斯。(ECCV), 2014 年,第 740–755 页。

[14] N.-H.刘,C.-Y.蒋和 H.-C. Chu,“利用移动传感器的脑电图信号识别人类注意力的程度”, 传感器,卷。 13,没有。 8,第 10273–10286 页,2013 年。

[15] RO Mbouna,SG Kong 和 M.-G. Chun,“用于驾驶员警觉性监控的眼睛状态和头部姿势的视觉分析”, IEEE Trans.英特尔. 运输.系统,卷。 14,没有。 3,第 1462–1469 页,2013 年 9 月。

[16] L. Mothwa,J. Tapamo 和 T. Mapati,“使用计算机视觉的智能考勤监控系统的概念模型”, Proc.国际.会议. 信号图像处理.基于互联网的系统。(SITIS), 2018 年,第 229–234 页。

[17] E. Murphy-Chutorian 和 MM Trivedi,“计算机视觉中的头部姿势估计:一项调查”, IEEE Trans.模式肛门.马赫.智力,卷。 31,没有。 4,第 607–626 页,2009 年 4 月。

[18] BN Anh等人,“基于计算机视觉的课堂学生为监控应用程序”, Appl.科学,卷。 9, 不。 22,p。 4729,2019。

[19] M. Raca 和 P. Dillenbourg,“课堂注意力评估系统”, Proc.国际.会议.学习.肛门.知道。(LAK), 2013 年,第 265–269 页。

[20] J. Redmon,S. Divvala,R. Girshick 和 A. Farhadi,“你只看一次:统一的实时对象检测”,载于Proc. IEEE 会议计算。 你要.模式识别, 2016年,第14页。 779–788。

[21] K. Simonyan 和 A. Zisserman,“用于大规模图像识别的非常深的卷积网络”, Proc.国际.会议.学习.代表, 2014 年,第 1–14 页。

[22] T. Soukupová 和 J. Cech,“使用面部标志进实时眨眼检测”, Proc. 21 日计算.维斯。 冬季研讨会, 2016 年,第 1–8 页。

[23] S. Tulyakov 和 N. Sebe,“从单张图像回归 3D 面部形状”, Proc. IEEE 国际.会议.计算.维斯。(ICCV), 2015 年,第 3748–3755 页。

[24] Y. Wang等人,“使用 RGB-D 摄像头进连续驾驶员注视区域估计”,传感器,卷。 19,没有。 6,p。 1287,2019。

[25] HH Wilmer,LE Sherman 和 JM Chein,“智能手机和认知:探索移动技术习惯与认知功能之间联系的研究综述”, Front.心理学,卷。 8,p。 605,2017 年 4 月。

[26] J. Zaletelj 和 A. Kosir,“根据 kinect 面部和身体特征预测学生在课堂上的注意力”, EURASIP J. 图像视频处理,卷。 2017 年,第 17 页。 2017 年 12 月 80 日。

[27] X.Zhen.Z.Wang,M.Yu 和 S.Li,“多输出回归的监督描述符学习”,Proc. IEEE CVPR, 2015 年,第 1211–1218 页。



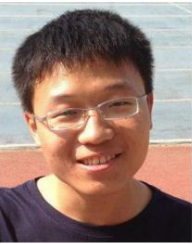
苏木春(IEEE 高级会员) 1986 年获得台湾国立交通大学电子工程学士学位,1999 年获得台湾国立交通大学电子工程硕士学位和博士学位.分别于 1990 年和 1993 年在美国马里兰大学帕克分校获得电气工程学位。



他目前是台湾国立中央大学计算机科学与信息工程教授.他撰写了 100 多篇期刊和参考会议论文.他目前的研究兴趣包括计算智能.神经网络.模糊系统.群体智能.情感计算.人机交互.机器人.模式识别.生物医学信号处理.图像处理和康复技术.他是 1991 年 IEEE Franklin V. Taylor 奖获得者.他担任许多期刊的副主编。他是 IEEE 计算智能学会和系统.人类与控制论学会的高级会员.他也是 IET 院士。



Chun-Ting Cheng (IEEE会员)于2016年获得台湾中央大学计算机科学与信息工程硕士学位.他的研究兴趣包括机器学习和深度学习。



Ming-Ching Chang (IEEE高级会员)分别于1996年和1998年获得国立台湾大学土木工程学士学位和计算机科学与信息工程硕士学位,并于1998年获得台湾大学计算机科学与信息工程博士学位. 2008年获布朗大学工程学院工程人机系统实验室博士学位.现任纽约州立大学奥尔巴尼大学工程与应用科学学院计算机科学助理教授。 2016年至2018年,他在电气与计算机工程系工作。



2012年至2016年,他担任计算机科学系的兼职教授.2008年至2016年,他担任GE全球研究中心的计算机科学家.他曾任机械研究所助理研究员

1996 年至 1998 年在台湾工业技术研究院工业研究室工作.他的研究项目由 GE 全球研究部.IARPA.DARPA.NIJ.VA 和 UAlbany 资助.他撰写了超过 85 篇同评审期刊和会议出版物.七项美国专利和 15 项披露信息.他的专业知识包括视频分析.计算机视觉.图像处理和人工智能.他曾获得 IEEE 高级视频和基于信号的监控 (AVSS) 2011 年最佳论文奖 (亚军) /IEEE 计算机视觉应用研讨会 2012 年最佳学生论文奖.GE 信念 (Stay Lean and Go Fast) 2015 年管理奖, 以及 2017 年 IEEE 智能世界 NVIDIA AI 城市挑战赛荣誉奖.他担任年度 AI 城市挑战赛 CVPR 2018–2021 研讨会的联合主席.IEEE 低功耗计算机视觉年度竞赛和研讨会 2019–2021 的联合主席.IEEE AVSS 2019 的程序主席、 2017–2019 年 IWT4S 主席.IEEE ICIP (2017 年和 2019–2021 年)和 ICME (2021 年)领域主席。



Yi-Zeng Hsieh (IEEE 会员)获得学士.硕士和博士学位.分别于 2004年.2006年和2012年获得台湾桃园国立中央大学计算机科学与信息工程博士学位.现任国立台湾海洋大学电机工程系副教授.他目前的研究兴趣包括神经网络.模式识别.图像处理.机器学习 and 深度学习。