Computer Science and Application 计算机科学与应用, 2022, 12(4), 1163-1174
Published Online April 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.124119

Hans汉斯

融合脸部外观和多行为特征的学生专注度识别网络

陆玉波，战荫伟，杨　卓，李学聪

广东工业大学，计算机学院，广东　广州

收稿日期：2022年3月20日；录用日期：2022年4月21日；发布日期：2022年4月28日

摘　要

在线学习环境中，专注度是衡量用户学习体验的重要指标。提高专注度识别的准确率可以帮助老师及时获得课程反馈，提升用户的学习体验。然而大多数现有的基于视频的专注度识别方法都只利用用户面部外观信息。除了面部外观信息之外，头部姿态和注视角度以及眨眼频率在内的细粒度行为线索也和学习专注度密切相关，但是，前人在专注度识别任务中没有很好地综合考虑以上特征。因此，本文提出一种新的专注度识别模型。该方法结合深度残差网络(ResNet)提取的脸部特征和基于OpenFace捕获的行为特征，这些特征输入到时序卷积网络(TCN)用于分析视频帧时空上的变化，以此识别出学习专注度。我们的模型在大型公开的专注度检测数据集DAiSEE上训练，在专注度四分类达到61.4%的准确率，实验结果表明，我们的方法超过DAiSEE上专注度识别的最先进方法。

关键词

专注度识别，深度残差网络，时序卷积网络，面部特征，头部姿态，注视角度，眨眼率

A Novel Engagement Recognition Network by Fusion Facial Appearance and Multi-Behavioral Features

Yubo Lu, Yinwei Zhan, Zhuo Yang, Xuecong Li

Department of Computer Science and Technology, Guangdong University of Technology, Guangzhou Guangdong

Received: Mar. 20th, 2022; accepted: Apr. 21st, 2022; published: Apr. 28th, 2022

Abstract

Engagement is an important measure of users' learning experience in an online learning envi-

文章引用: 陆玉波, 战荫伟, 杨卓, 李学聪. 融合脸部外观和多行为特征的学生专注度识别网络[J]. 计算机科学与应用, 2022, 12(4): 1163-1174. DOI: 10.12677/csa.2022.124119

ronment. Improving the accuracy of engagement recognition can help the instructors get timely feedback on the courses, and enhance users' learning experience. However, most existing video-based engagement recognition methods only use the user's facial appearance information. In addition to facial appearance, fine-grained behavioral cues such as head pose, eye gaze and blink rate are also closely related to engagement. But most researchers don't comprehensively consider these features. Therefore, in this paper, we propose a novel engagement recognition model: our proposed method combines facial features extracted by Deep Residual Network (ResNet) and behavioral features captured by OpenFace. These features are fed into temporal convolutional network (TCN) to analyze the temporal changes in video frames to detect the level of engagement. Our model trained on a large publicly available student's engagement detection dataset, DAiSEE. We achieved 61.4% in top-1 accuracy in the problem of four classifications for engagement. The results show that our method outperforms state-of-the-art methods.

Keywords

Engagement Recognition, Deep Residual Network, Temporal Convolutional Network, Facial Appearance, Head Pose, Eye Gaze, Blink Rate

1. 引言

在线学习不受时间与地域的限制，在现代教育中得到广泛使用。然而，与传统课堂相比，在线教育授课的过程中也带来其他问题。例如，学习者在接受知识的过程中，由于缺乏有效地监督，导致其学习效果无法得到保证，因此，需要对学习者的注意力状态进行有效监督，以保证在线学习者的学习质量。

对于专注度检测的研究数据包括学生的图像[1]、视频[2]、音频[3]和心电图(ECG) [4]。网络摄像头相比生物传感器，在获取学生上课数据方面，成本更低且更加便捷的。因此，最近大多数关于学生专注度检测的研究都是在网络摄像头获取到的学生上课数据基础上进行的，并使用计算机视觉技术进行专注度检测[5] [6]。

基于计算机视觉的学生专注度检测方法可以分为基于图像和基于视频的检测方法。前一种方法是从单个图像或从视频中提取的单个帧来检测专注度，这种方法的主要限制是只利用单个帧的空间信息，而专注度检测是一种时空情感行为。因此本研究主要基于视频对学生专注度检测。基于视频的专注度识别可以分为基于端到端模型检测方法和基于特征的检测方法。在基于端到端模型检测方法中，连续的原始视频帧被输入到卷积神经网络(CNNs)，而后再用递归神经网络回归出专注度级别[7]。基于特征的方法中，从视频帧中提取手工特征，并通过递归神经网络或机器学习方法进行分析，输出专注度级别[8]。现有的专注度识别研究使用各种特征，包括高层特征，如注视方向、头部姿势和面部动作单元等行为特征，以及低层特征，如 LBP-TOP [9]和 Gabor 特征[10]。我们发现专注度识别中特征提取很大程度依赖于人脸特征的提取，目前人脸特征提取方法主要有两种思路，一种是使用卷积神经网络(CNN)提取脸部空间特征，但这无法直接提取到脸部细粒度特征。另一种则基于单一手工特征或多个特征简单的组合进行特征提取，但大多数研究人员往往只关注几个特征，而没有全面考虑面部外观、头部姿态、注视角度、眨眼率等特征。在专注度识别中，如何有效地将深度网络提取的粗粒度面部特征信息和头部姿态在内的多种视觉线索的细粒度行为特征相结合，这一问题尚未得到深入的探讨。

在本文中，我们提出一种新的专注度识别框架，将头部姿态在内的多种行为特征与面部外观相结合，用于专注度识别。图 1 展示了该模型的总体框架。首先利用 OpenFace2.0 [11]工具箱检测人脸区域，然后在大规模人脸识别数据集 VGGFace2.0 [12]中用 SE-ResNet-50 (SENet) [13]进行预训练并提取人脸空间特征。同时，我们利用 OpenFace 从视频帧中提取出头部姿态和注视角度这些行为特征，并和人脸空间特征连接成特征向量，输入到时域卷积网络中进行联合训练，我们在一个公开的在线教育情感状态数据集(DAiSEE) [7]上评估所提出方法的性能，实验结果表明，我们所提出的方法优于最先进算法。同时还比较了头部姿态在内的多种行为特征与面部外观特征的组合，综合考虑这些特征可以得到最好的性能，我们还展示了不同方法在不同专注度层级中的混淆矩阵，结果表明我们在加入头部姿态、注视角度、眨眼率这些行为特征之后能够有效提升模型对低专注度样本的识别能力。

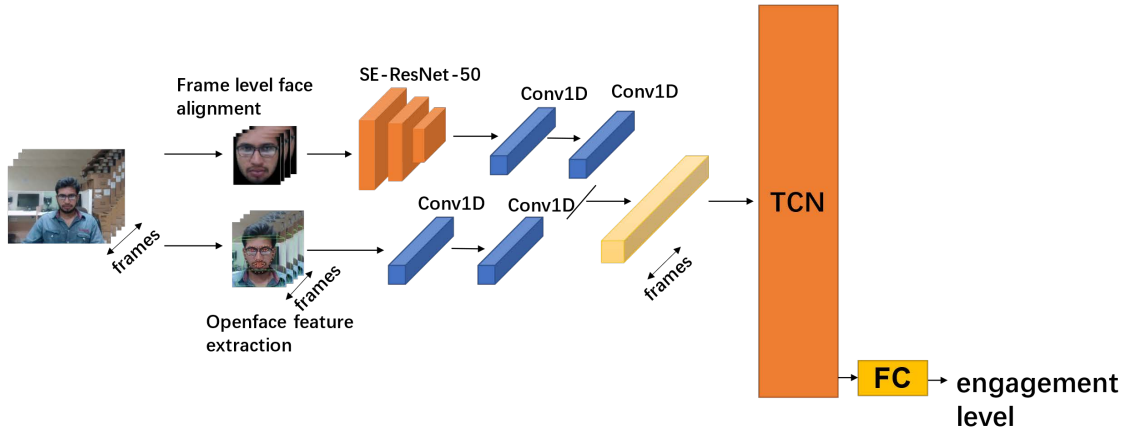

Figure 1. The framework proposed in this paper
图 1. 本文提出的框架

2. 相关工作

近年来，不同类型的数据比如视频[2]、音频[3]、ECG 数据[4]、压力传感器数据[14]、心率数据[15]都用于测量学生专注度水平，本文中，我们只探讨基于视频数据的专注度识别工作，在这些方法中，机器学习和深度学习算法被用来分析视频中人物的专注度，基于视频的专注度识别算法又可以分为基于端到端模型检测的方法和基于特征的方法。

基于特征的专注度识别方法，其步骤是先从视频或者图像中提取手工特征，然后将其输入到分类器或者回归器中，在先前的方法[9] [10]，提取传统的视觉特征(例如 Gabor 或 LBP-TOP)用于专注度检测，而随着深度学习技术的引入，视觉线索(例如注视方向、头部姿态和面部动作单元等)通过开源视觉工具箱如 OpenFace [11]，OpenPose [16]来提取的，使用这些提取技术获得多模态特征，组合成每个视频帧的特征向量，这些特征向量将会输入到序列模型中进行分析，并输出专注度预测结果。Yang 等人[17]使用 OpenFace [11]和 OpenPose [16]来跟踪面部特征和姿体动作，建立了基于多示例学习机制的 LSTM 多模型回归方法，并在 2018 年情感识别子挑战中获得最佳。Niu 等人[18]设计了一个 gaze-au-pose (GAP)特征，从每个视频帧中提取 GAP 特征，并使用具有门控递归单元(GRU)的递归神经网络进行专注度预测。黄等人[19]提出了深度专注度识别网络(Deep Engagement Recognition Network, DERN)，该网络结合了双向 LSTM 和注意力机制，并对从人脸中提取的特征进行分类，他们在 DAiSEE 数据集上的专注度识别精确度达到 60%。

基于端到端模型的专注度识别方法，通常是将视频帧输入到深度卷积网络以检测专注度。Gupta 等人

[7]建立了 DAiSEE (在线学习情感状态数据集)，并使用不同的端到端视频分类技术，包括 InceptionNet [20]、C3D [21]和 LRCN [22]，建立基准测试结果，准确率分别为 46.4%、56.1%和 57.9%。

Zhang 等人[23]对膨胀的 3D 卷积网络(I3D)进行优化，得到了对专注层级二分类的最高精度。Liao 等人[24]提出深度面部时空网络(DFSTN)，他们的专注度识别模型包含两个模块，一个预训练的 SE-ResNet-50 用于从人脸中提取空间特征，以及带有全局注意力机制的 LSTM，并在 DAiSEE 数据集上达到 58.84%的专注度识别准确率。

总体而言现有的研究方法要么只考虑单一行为特征或特征之间简单的组合，要么则使用端到端模型对面部区域进行专注度预测。而在专注度识别中，如何有效地将深度学习模型捕获的面部空间特征与多行为特征相结合，尚未得到深入的探讨。在下一节中，我们将提出一种新的专注度识别方法，我们融合端到端模型提取的面部特征以及 OpenFace 提取的多行为特征(眨眼频率、注视角度、头部姿势)，并将上述融合后的特征输入到时序卷积网络(TCN)进行专注度识别。

3. 本文方法

我们在图 1 展示了本文提出的方法框架。首先，从给定的视频序列中按照指定的时间间隔提取视频帧，这些视频帧经过人脸检测并对齐后输入到 SE-ResNet-50 [13]中生成脸部空间特征。同时，我们使用 OpenFace 来提取人脸的行为特征，如头部姿态、注视角度、眨眼率这些特征将连接并组成特征向量，然后分别将面部特征以及 OpenFace 提取的特征输入到 1D CNN 层，这个层负责学习高级时序特征并进行回归，该层输出的两种不同模态的特征图连接作为人脸的粗粒度信息和细粒度信息的联合表征，每个视频帧提取到的特征将作为 TCN 的一个时间步长的输入，TCN 的最后一个时间步长紧接着一个全连接层输出预测的专注度层级。

3.1. 脸部特征提取

脸部是反映一个人情绪状态最具表现力的区域，老师通常通过观察学生的脸部来判断他们的参与度。直观上理解，脸部包含人类大量的情感信息，并且直接与一个人的专注程度相关联[6]。在本文方法中，我们利用 CNN 来提取面部空间特征，这些特征对低光照度和遮挡等恶劣条件更具鲁棒性。

我们采用 SE-ResNet-50 [13]作为骨干网络进行特征提取。SE 注意力模块通过对不同特征通道的依赖关系建模，自适应地对特征通道进行加权，增强有效特征的表达，通过这些模块与主流的 CNN 架构(如 ResNet)集成，可以显著提高网络表征能力。在训练之前，我们首先使用 OpenFace2.0 工具箱在每一帧中裁剪面部区域。我们之所以采用 OpenFace2.0 作为人脸检测器，是因为与其他人脸检测算法[25] [26] [27]相比，OpenFace 2.0 在速度和准确度上有更多的优势。每个人脸图像的大小都被调整为 224 × 224，以匹配我们模型的输入大小。虽然深度学习可以提取出更鲁棒性的特征，但它需要大量的数据来训练模型，对于我们的任务，视频的数量比较少，视频帧之间的差异很小。因此，采用长期递归卷积网络[22]等方法很难获得鲁棒的人脸空间特征。受面部表情识别相关工作的启发[27]，我们利用在人脸识别数据库预训练的模型，并将其应用于我们的任务。在实验中，我们采用 SE-ResNet-50 [13]在 VGGFace2 [12]上进行预训练，该数据集包含大量不同姿态不同光照的人脸识别数据，能够帮助我们提取到更多判别性的脸部特征。

对于每一帧的人脸图像，我们使用 SE-ResNet-50 的最后一个池化层的 2048 维特征向量来进行表示。同样的，受之前工作[28]所启发，1D CNN 能够学习到短时间粒度的高阶时序表征，正如图 1 所示，我们把静态面部特征送入 1D CNN 层去获取时序上语义表征。

3.2. 行为特征提取

一个人的专注状态在行为上的表现往往体现在眼部行为和头部姿态活动上。

Ranti 等人[29]证明，眨眼率是衡量人对视觉内容专注程度的一个重要指标。他们推断，眨眼会在特定的时间内停止，以最大限度地减少眨眼过程中视觉信息的损失。从频率来看，一个人越专注当前视觉内容时，停止眨眼的频率越大。所以我们认为眨眼率是比较重要的专注行为的特征。面部动作单元 AU45 的强度表示眼镜的闭合程度[11]。因此 AU45 的强度，也将作为其中一种专注行为特征。

研究表明，注意力高度集中的情况下，人的视线方向和头部姿态往往更倾向于静止，反之亦然[8] [10]。此外，在高专注度的情况下，人的眼睛注视方向是视觉内容。因此，视频中的眼睛位置、头部姿态和注视方向可以认为专注状态的特征。

对于以上所提及的专注行为特征，我们使用 OpenFace 工具箱分别提取注视方向、头部姿态以及代表眨眼率的脸部动作单元 AU45 强度。这些特征的输出格式细节描述在[30]中。

上述描述的这些特征从每一帧中提取，并组合作为每帧的专注行为特征向量，每一帧中提取的特征维度如下：

1) 注视角度：包含左右眼 2 个 3 维的注视角度向量和一个 2 维的左右眼平均注视角度(弧度制)。注视角度特征总的维数为 8。

2) 头部姿态：头部的任意姿态可转化为 6 个参数(yaw, roll, pitch, x, y, z)，前三个为旋转参数，后三个为位置参数。头部姿态特征的维数为 6。

3) 眨眼：脸部动作单元(Facial Action Unit)中代表“眨眼”程度是 AU45 的强度(一个连续值从 0 到 5 表示该面部单元的运动强度)，该特征维数为 1。

然而根据 Wu 等人[31]的研究，OpenFace 提取的特征在一个小的时间尺度上涉及许多噪声，为了有效地利用这些特征，它们需要将经历一维卷积处理和时序池化，以减少小的时间粒度中的噪声。因此，与提取面部静态特征之后的操作类似，我们同样使用两层 1D CNN 对多行为特征进行一维卷积处理。

3.3. 时序卷积网络

时间卷积网络(Time Convolutional Network, TCN [32])，它是一种能够捕获上下文信息的时间序列模型，越来越多用于动作分割[33]、动作定位[34]和其他序列建模任务，同时 TCN 也在连续维度情感识别里获得最佳的性能[35]。因此，我们利用 TCN 作为时序预测模型，TCN 网络中有三个关键设计：因果卷积、扩张卷积和残差连接[36]，这赋予了 TCN 强大的时序分析能力，下面我们分别介绍。

3.3.1. 因果卷积(Casual Convolution)

TCN 使用了因果卷积，在这种卷积中，如图 2 所示，t 时刻的输出仅与来自 t 时刻和更早之前的时刻的信息相关，由于使用的是卷积结构，因此 TCN 更适合于并行计算，相比于处理时间序列数据的递归神经网络 RNN 的串行计算的方式，计算效率更高。

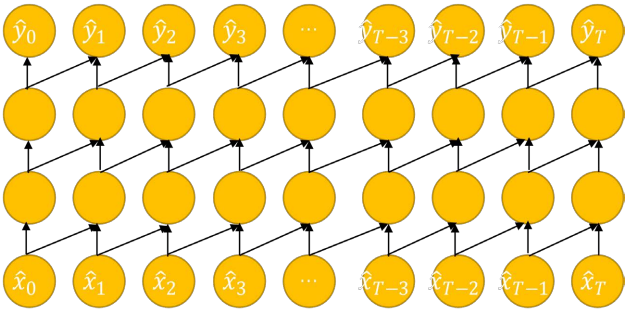

Figure 2. Causal convolution structure
图 2. 因果卷积结构

这种卷积计算方式具备如下特性：1) 不存在信息从未来泄露到过去，当前时刻的输出只与观测到的数据有关；2) 堆叠的因果卷积层越深，网络越能追溯到越久远的历史。因果卷积的条件概率公式可以这样表示：

$$p(x)=\prod_{t=1}^{T} p(x_t \mid x_1,\cdots,x_{t-1}) \tag{1}$$

其中，x_t 表示 t 时刻的信息。

从图 2 中可以看到，当一个节点需要考虑较久远的信息，那么需要更深的卷积层数，但是此时网络的参数也会增加，因此为了缓解这一问题，便需要引入扩张卷积模块：

3.3.2. 扩张卷积(Dilated Convolution)

因为因果卷积的感受野和卷积层数是呈线性相关的，为了提高计算效率，TCN 应用了扩张卷积(图 3)，即在普通卷积中注入一定长度的空洞，使得在同等卷积核大小的情况下，相比普通卷积的感受野要更大。综合上述因果卷积的特点可得出，扩张卷积不仅可以捕获过去时间点上的时序信息，而且具备长期记忆性。对于一维序列输入 $X\in\mathbb{R}^n$ 和一维卷积核 $f:\{0,\cdots,k-1\}\to\mathbb{R}$，对序列元素 s 的扩张卷积运算 F 定义为：

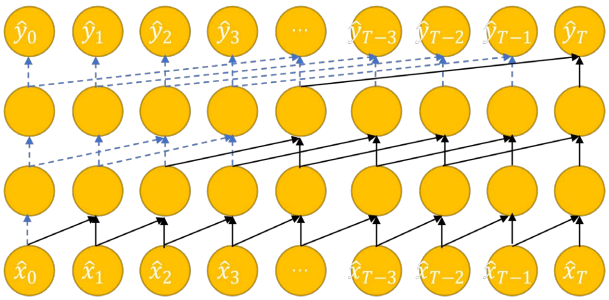

Figure 3. Dilated convolution structure
图 3. 扩张卷积结构

$$F(s)=(X *_d f)(s)=\sum_{i=0}^{k-1} f(i)\cdot X_{s-d\cdot i} \tag{2}$$

式子中 k 为卷积核尺寸，d 为扩张因子，用于控制扩张卷积中注入的空洞个数，当设置 $d=1$ 时，扩张卷积退化为普通的因果卷积。d 数量上随着网络深度的增加而指数增长，一般的，网络中第 i 层的扩张因子为 2^d。这使得有较大扩张因子的网络顶层可以捕获到更长的时间依赖关系(long-term dependencies)。

3.3.3. 残差连接(Residual Connections)

深层卷积神经网络可以更好地提取到数据更深层次的特征，因此在许多任务上，性能表现优越。然而，随着网络层数地加深，网络末端的梯度信息很难通过反向传播进行平稳传递，进而使得深层网络的训练和学习变得困难。

在本工作中，TCN 利用残差块(图 4)代替了 TCN 层与层之间的连接，由于输入和输出张量可以由不同的尺寸大小，因此 TCN 应用额外的 1 × 1 卷积层以保持输入和输出张量形状相同，同时使用残差块替换卷积层，如图所示，每个残差块包含两组扩张卷积和非线性层 ReLU，并添加了权重归一化层(weight normalization layer)和 dropout 层进行正则化。

我们将行为特征和脸部外观特征拼接在一起构成每一个视频帧的特征向量，从每个视频帧提取的特征向量将作为 TCN 的一个时间步长的输入，TCN 的最后一个时间步长，接着一个全连接层，输出预测的视频中人物专注度类别。

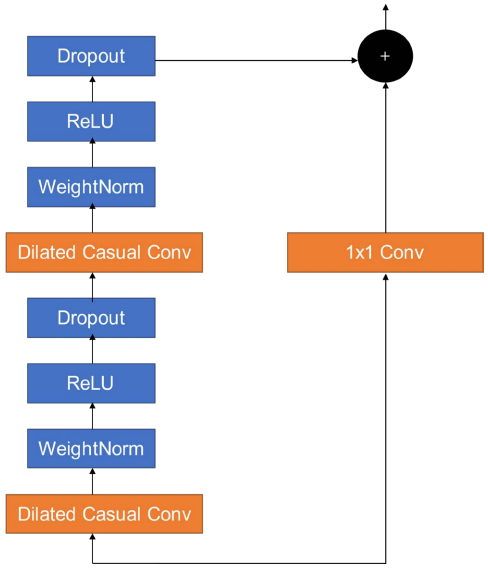

Figure 4. Residual block structure
图 4. 残差块网络模型结构

4. 实验

4.1. 数据集

我们在一个公开的在线教育情感状态数据集(DAiSEE) [7]进行实验，DAiSEE 由 9068 个视频组成，每个视频时长 10 秒，总共记录了 113 名印度学生的在线学习情况。每段视频的情绪状态可以分为无聊、迷茫、专注和郁闷四种状态类型，每个状态都在四个级别(序数类)中的一个，级别 0 (非常低)、1 (低)、2 (高)和 3 (非常高)。在本文中，重点只放在研究学生在网络学习环境中的专注程度情况。这些视频的帧率和分辨率分别为 30 帧/秒(Fps)和 640 × 480 像素。数据集分为训练集、验证集和测试集，比例为 6:2:2。

4.2. 评价指标

不同的研究人员采用不同的衡量标准来衡量参与度。Gupta 等人[7]将此问题作为分类任务进行处理，并选取 Top-1 准确率来评估算法在 DAiSEE 上的性能。在我们的实验中，我们使用相同的度量方式来验证我们方法的有效性。Top-1 精确率是指排名第一的类别与实际结果相符的准确率，Top-1 准确率如下：

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3}$$

其中 TP 代表真正例，FP 代表假正例。

4.3. 实验配置

4.3.1. 网络参数

我们采用一个带有挤压激励(SE)模块的 50 层 ResNet 进行面部特征提取，并将最后一个平均池化层的激活输出作为该视频帧的脸部空间特征。为了避免过拟合，我们对每个卷积块都采用 Spatial Dropout 策略，丢弃率设置为 0.2。神经网络中所有神经元权重都使用高斯分布进行初始化。在面部特征提取网络之后堆叠的两个一维 CNN 层的输出通道分别为 1，1。并且他们的卷积核大小被设置为 5、3。池化层的步长分别被设置为 2、4。另外，在提取多行为特征之后叠加的两个一维 CNN 层的输出通道分别为 1、1。卷积核大小分别设置为 5、3。TCN 中总共有 8 层隐层，每层神经元个数为 128，卷积核大小为 7，丢弃

率为 0.25。在 TCN 最后一个时间步长，一个带有 4 个输出神经元的全连接层用于四分类(对应 DAiSEE 数据集专注度级别数)。

4.3.2. 网络参数

我们在神经网络训练过程中，批次大小设置为 128，训练周期设置为 500。采用修正线性单元(RELU)作为激活函数，优化器采用 RMSprop 优化器，学习率为 0.0001。优化器采用 KERAS 框架中的默认参数，损失函数采用分类交叉熵函数。

4.3.3. 实验步骤

本文提出的融合脸部外观和多行为特征的学生专注度识别网络，其详细如图 1 所示。我们的实验步骤如下：

步骤 1：对专注度数据集的视频序列解码提取成帧并保存，并利用 OpenFace 对原始视频帧进行人脸检测并单独裁剪出人脸数据集，同时，利用 OpenFace 提取视频帧包含头部特征、视线方向等多行为特征信息并保存。

步骤 2：使用在大型人脸数据集 VGGFace2 上预训练的 SE-ResNet 模型，对步骤 1 得到的人脸数据集提取其人脸外观特征信息并保存。

步骤 3：将步骤 2 得到的人脸外观特征信息和步骤 1 得到的多行为特征信息进行特征融合，以此得到融合脸部外观和多行为信息的特征。

步骤 4：构建深度学习网络模型，其网络结构见图 1，详细网络参数和训练过程见 4.3.1 和 4.3.2 节，其中在训练时，步骤 3 得到的融合后的特征将作为模型的输入。

步骤 5：对训练后的模型，使用 4.2 节中的评价指标，对模型性能进行评估。对不同特征组合的性能评估，只需在步骤 3 中对特征进行组合，并调整深度学习网络模型中相对应的特征通道，其余步骤不变。

4.4. 实验结果与分析

我们的方法在 DAiSEE 数据集上实验，表 1 展示和其他工作比较的结果，我们的方法分别与基于端到端模型的方法：C3D [21]，LRCN [7]，DFSTN [24]进行比较，以及当前最优算法 DERN [19] (基于特征的方法)进行比较，可以看到，我们提出的方法性能要优于以往无论是基于端到端学习的方法还是基于特征的方法，我们的方法比最先进的 DERN 算法(60%)提高 1.4%，比起最好的端到端模型 DFSTN 方法(58.8%)提高 2.6%。

Table 1. Accuracy of different methods on DAiSEE [7] dataset
表 1. 不同方法在 DAiSEE [7]数据集上的精确度

Method	Top1-acc
C3D (Fine Tunning) [21]	57.6%
LRCN [7]	57.9%
DFSTN [24]	58.8%
DERN [19]	60.0%
Ours	**61.4%**

同时我们为了进一步研究各特征在专注度识别中的重要性，我们比较面部外观特征和行为特征之间不同组合的实验，实验结果如表 2，在只有面部外观而不加其他行为特征的时候，我们的模型也有 58%的准确度，稍微优于表 1 中基准模型 LRCN。这表明我们 SE-RESNET-50 + TCN 的骨干模型，比起普通

CNN + LSTM 的网络的性能要更好，这也得益于其在大型人脸数据集上充分的预训练，以及带有注意力模块的残差网络强大的特征提取能力，同时，可以看到，在面部外观特征的基础上，单独加入头部姿态、凝视、以及 AU45 这些特征，都比起单独的面部特征的性能要更好：1) 在这些行为特征中，提升性能最好的是头部姿态，相比于仅使用面部外观的模型提升了 1%，它与[37]中的发现是一致的，[37]认为头部的运动，例如摇头、点头也被认为是区分专注程度的标志。2) 其次是 AU45 (眨眼率)，相比于面部外观的模型提升了 0.5%，这与 ranti 等人[29]的研究相一致，研究表明眨眼率为个人行为和认知参与视觉内容提供一个可靠的衡量标准。3) 最后是凝视，相比于面部外观的模型提升了 0.3%，许多研究[38] [39]表明，人的视线角度和注意力焦点有关。4) 而当我们综合考虑这些特征，我们的模型具有更好的性能。

Table 2. Performance of different feature combinations
表 2. 不同特征组合的性能表现

Feature	Top1-acc
Facial Appearance	58%
Head Pose + Facial Appearance	59%
Eye Gaze + Facial Appearance	58.3%
AU45 + Facial Appearance	58.5%
Head Pose + Eye Gaze + Facial Appearance	60.3%
Head Pose + Eye Gaze + AU45 + Facial Appearance	**61.4%**

在 DAiSEE 中，专注度四种标签的数量之比为 1:8:73:67，由于高度不平衡的数据分布，过去的方法没有一个能够正确对少数类别(低专注度类别)的样本进行正确分类(专注度 0 和 1)。图 5 展示之前端到端方法、以及本文提出方法的混淆矩阵，纵轴代表真实类别，横轴代表预测类别。混淆矩阵中的数字代表预测为该类别的样本数量占其真实样本数量的比例，可以看到没有一种端到端方法能够正确对少数类别(专注度 0 和 1)的样本进行分类。然而，我们提出的脸部外观+行为特征的方法(图 5(c))，能够对类 1 中的部分样本进行正确分类，这表明加入行为特征可以提高对低专注度类别的样本分类。

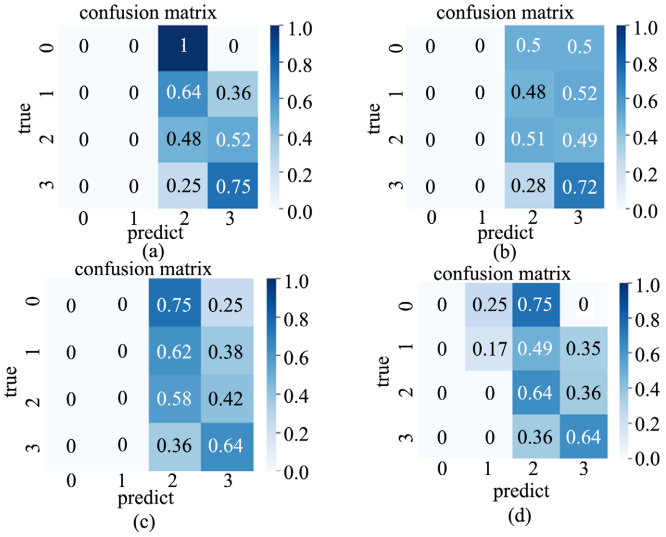

Figure 5. Engagement-level confusion matrices of different methods on the DAiSEE dataset [6], (a) C3D fine tuning, (b) LRCN, (c) facial appearance only (proposed), (d) Face appearance + behavioral features(proposed)
图 5. 不同方法在 DAiSEE [6]数据集上的专注度层级的混淆矩阵，(a) C3Dfine tuning，(b) LRCN，(c) 仅使用脸部外观(本文方法)，(d) 使用脸部外观 + 行为特征(本文方法)

与高专注度样本相比，低专注度样本的身体姿势和面部表情变化更大，见图 6。由于我们的面部特征提取网络在大型人脸识别数据集上(包括大量样本)进行预训练，并且在特征融合期间加入从视频帧中提取头部姿态、注视方向、以及眨眼率这些与专注行为有关的特征，使得该模型在识别少数的低专注度类别样本优于先前方法。

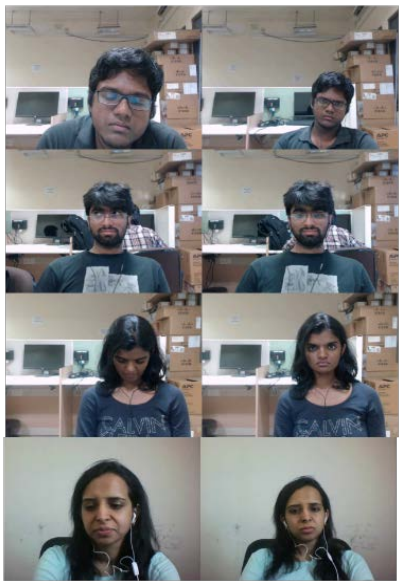

Figure 6. Representative disengaged (left) and engaged (right) samples
图 6. 代表性的低专注度(左)和高专注度(右)样本

5. 结论

在本文中，我们提出了一种新的专注度识别框架，相比于现有的端到端识别网络，我们的模型综合考虑脸部外观以及多种细粒度的行为特征进行专注度识别，一方面，我们的脸部外观网络在大型人脸数据集 Vggface2 上进行预训练，以获得更具有判别性的空间人脸特征。另一方面，我们通过 OpenFace 提取到学生的头部姿态、注视角度、眨眼率这些细粒度的专注行为特征。实验证明，我们所提出的方法，在 DAiSEE 数据集上的专注度分类精确度达到 61.4%，比现有的方法要好。同时，通过不同特征组合的实验，也证明我们所提取的细粒度行为特征和粗粒度的面部外观特征之间具有互补的作用。最后，我们还比较了不同方法在 DAiSEE 数据集上的专注度识别混淆矩阵，结果表明，我们的方法相比对比方法在低专注度类别识别效果要好。

6. 未来展望

在当前专注度识别领域，大多数专注度识别数据集仍然是基于视频信息，但音频是反映学习者专注度的一个重要信息，未来的研究者在构建专注度识别数据集时，还应该考虑加入音频信息，基于多模态信息的专注度识别可能是一个有潜力的研究方向，此外，对于专注度类别不平衡问题，可以考虑借鉴异常检测的思路，提高模型对低专注度类别的识别准确度。

基金项目

广东省重点领域研发计划项目(编号 2020B0101130019，2019B010150002)；国家自然科学基金(批准号 61907009)。

参考文献

[1] Nezami, O.M., Dras, M., Hamey, L., Richards, D., Wan, S. and Paris, C. (2018) Automatic Recognition of Student Engagement Using Deep Learning and Facial Expression. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Würzburg, 16-20 September 2019, 273-289. https://doi.org/10.1007/978-3-030-46133-1_17

[2] Dhall, A., *et al*. (2020) Emotiw 2020: Driver Gaze, Group Emotion, Student Engagement and Physiological Signal Based Challenges. *Proceedings of the* 2020 *International Conference on Multimodal Interaction*, Nicolaïkerk, 25-29 October 2020, 784-789. https://doi.org/10.1145/3382507.3417973

[3] Guhan, P., *et al*. (2020) ABC-Net: Semi-Supervised Multimodal GAN-Based Engagement Detection Using an Affective, Behavioral and Cognitive Model.

[4] Belle, A., Hargraves, R.H. and Najarian, K. (2012) An Automated Optimal Engagement and Attention Detection System Using Electrocardiogram. *Computational and Mathematical Methods in Medicine*, **2012**, Article ID: 528781. https://doi.org/10.1155/2012/528781

[5] Doherty, K. and Doherty, G. (2018) Engagement in HCI: Conception, Theory and Measurement. *ACM Computing Surveys*, **51**, 1-39. https://doi.org/10.1145/3234149

[6] Dewan, M.A.A., Murshed, M. and Lin, F. (2019) Engagement Detection in Online Learning: A Review. *Smart Learning Environments*, **6**, 1-20. https://doi.org/10.1186/s40561-018-0080-z

[7] Gupta, A., *et al*. (2016) DAiSEE: Towards User Engagement Recognition in the Wild. *Journal of Latex Class Files*, **14**, 1-12.

[8] Chen, X., *et al*. (2019) FaceEngage: Robust Estimation of Gameplay Engagement from User-Contributed (YouTube) Videos. *IEEE Transactions on Affective Computing*, 1. https://doi.org/10.1109/TAFFC.2019.2945014

[9] Zhao, G.Y. and Pietikainen, M. (2007) Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**, 915-928. https://doi.org/10.1109/TPAMI.2007.1110

[10] Whitehill, J., *et al*. (2014) The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions. *IEEE Transactions on Affective Computing*, **5**, 86-98. https://doi.org/10.1109/TAFFC.2014.2316163

[11] Baltrusaitis, T., *et al*. (2018) Openface 2.0: Facial Behavior Analysis Toolkit. 2018 13*th IEEE International Conference on Automatic Face & Gesture Recognition* (*FG* 2018), Xi'an, 15-19 May 2018, 59-66. https://doi.org/10.1109/FG.2018.00019

[12] Cao, Q., *et al*. (2018) Vggface2: A Dataset for Recognising Faces across Pose and Age. 2018 13*th IEEE International Conference on Automatic Face & Gesture Recognition* (*FG* 2018), Xi'an, 15-19 May 2018, 67-74. https://doi.org/10.1109/FG.2018.00020

[13] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[14] Rivas, J.J., *et al*. (2021) Multi-Label and Multimodal Classifier for Affective States Recognition in Virtual Rehabilitation. *IEEE Transactions on Affective Computing*, 1. https://doi.org/10.1109/TAFFC.2021.3055790

[15] Monkaresi, H., *et al*. (2016) Automated Detection of Engagement Using Video-Based Estimation of Facial Expressions and Heart Rate. *IEEE Transactions on Affective Computing*, **8**, 15-28. https://doi.org/10.1109/TAFFC.2016.2515084

[16] Cao, Z., *et al*. (2019) OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 172-186. https://doi.org/10.1109/TPAMI.2019.2929257

[17] Yang, J.F., *et al*. (2018) Deep Recurrent Multi-Instance Learning with Spatio-Temporal Features for Engagement Intensity Prediction. *Proceedings of the* 20*th ACM International Conference on Multimodal Interaction*, Boulder, 16-20 October 2018, 594-598. https://doi.org/10.1145/3242969.3264981

[18] Niu, X.S., *et al*. (2018) Automatic Engagement Prediction with GAP Feature. *Proceedings of the* 20*th ACM International Conference on Multimodal Interaction*, Boulder, 16-20 October 2018, 599-603. https://doi.org/10.1145/3242969.3264982

[19] Huang, T., *et al*. (2019) Fine-Grained Engagement Recognition in Online Learning Environment. 2019 *IEEE* 9*th International Conference on Electronics Information and Emergency Communication* (*ICEIEC*), Beijing, 12-14 July 2019, 338-341. https://doi.org/10.1109/ICEIEC.2019.8784559

[20] Szegedy, C., *et al*. (2015) Going Deeper with Convolutions. *Proceedings of the IEEE Conference on Computer Vision*

and Pattern Recognition, Boston, 7-12 June 2015, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[21] Tran, D., *et al*. (2015) Learning Spatiotemporal Features with 3d Convolutional Networks. *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, 7-13 December 2015, 4489-4497. https://doi.org/10.1109/ICCV.2015.510

[22] Donahue, J., *et al*. (2015) Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 677-691. https://doi.org/10.1109/CVPR.2015.7298878

[23] Zhang, H., *et al*. (2019) A Novel End-to-End Network for Automatic Student Engagement Recognition. 2019 *IEEE* 9*th International Conference on Electronics Information and Emergency Communication* (*ICEIEC*), Beijing, 12-14 July 2019, 342-345. https://doi.org/10.1109/ICEIEC.2019.8784507

[24] Liao, J.C., Liang, Y. and Pan, J.H. (2021) Deep Facial Spatiotemporal Network for Engagement Prediction in Online Learning. *Applied Intelligence*, **51**, 1-13. https://doi.org/10.1007/s10489-020-02139-8

[25] Zhu, X.X. and Ramanan, D. (2012) Face Detection, Pose Estimation, and Landmark Localization in the Wild. 2012 *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, 16-21 June 2012, 2879-2886.

[26] Viola, P. and Jones, M. (2001) Rapid Object Detection Using a Boosted Cascade of Simple Features. *Proceedings of the* 2001 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, 1.

[27] Xie, S.Y., Hu, H.F. and Wu, Y.B. (2019) Deep Multi-Path Convolutional Neural Network Joint with Salient Region Attention for Facial Expression Recognition. *Pattern Recognition*, **92**, 177-191. https://doi.org/10.1016/j.patcog.2019.03.019

[28] Trigeorgis, G., *et al*. (2016) Adieu Features? End-to-End Speech Emotion Recognition Using a Deep Convolutional Recurrent Network. 2016 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Shanghai, 20-25 March 2016, 5200-5204. https://doi.org/10.1109/ICASSP.2016.7472669

[29] Ranti, C., *et al*. (2020) Blink Rate Patterns Provide a Reliable Measure of Individual Engagement with Scene Content. *Scientific Reports*, **10**, Article No. 8267. https://doi.org/10.1038/s41598-020-64999-x

[30] Openface Output Format. https://github.com/TadasBaltrusaitis/OpenFace/wiki/Output-Format

[31] Wu, S.W., *et al*. (2019) Continuous Emotion Recognition in Videos by Fusing Facial Expression, Head Pose and Eye Gaze. 2019 *International Conference on Multimodal Interaction*, Suzhou, 14-18 October 2019, 40-48. https://doi.org/10.1145/3340555.3353739

[32] Bai, S.J., Kolter, J.Z. and Koltun, V. (2018) An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling.

[33] Lea, C., *et al*. (2017) Temporal Convolutional Networks for Action Segmentation and Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 1003-1012. https://doi.org/10.1109/CVPR.2017.113

[34] Chao, Y.-W., *et al*. (2018) Rethinking the Faster R-CNN Architecture for Temporal Action Localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City,18-23 June 2018, 1130-1139. https://doi.org/10.1109/CVPR.2018.00124

[35] Khorram, S., *et al*. (2017) Capturing Long-Term Temporal Dependencies with Convolutional Networks for Continuous Emotion Recognition. *INTERSPEECH* 2017: *Conference of the International Speech Communication Association*, Stockholm, 20-24 August 2017, 1253-1257. https://doi.org/10.21437/Interspeech.2017-548

[36] He, K.M., *et al*. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*, 27-30 June 2016, Las Vegas, 770-778. https://doi.org/10.1109/CVPR.2016.90

[37] Kapoor, A., Burleson, W. and Picard, R.W. (2007) Automatic Prediction of Frustration. *International Journal of Human-Computer Studies*, **65**, 724-736. https://doi.org/10.1016/j.ijhcs.2007.02.003

[38] Langton, S., Watt, R.J. and Bruce, V. (2000) Do the Eyes Have It? Cues to the Direction of Social Attention. *Trends in Cognitive Sciences*, **4**, 50-59. https://doi.org/10.1016/S1364-6613(99)01436-9

[39] Dong, L.G., *et al*. (2009) Visual Focus of Attention Recognition in the Ambient Kitchen. In: *Asian Conference on Computer Vision*, Springer, Berlin, 548-559.