

DOI: 10.13718/j.cnki.xdzk.2024.12.018

石敏, 许莉钧, 彭华东. 可穿戴生理传感器驱动的深度情绪识别模型在心理健康评估中的应用[J]. 西南大学学报(自然科学版), 2024, 46(12): 189-201.

可穿戴生理传感器驱动的深度情绪识别模型在心理健康评估中的应用

石敏¹, 许莉钧², 彭华东³

1. 福州外语外贸学院 教育学院, 福州 350202; 2. 南京工程学院 艺术与设计学院, 南京 211167;
3. 福州外语外贸学院 艺术与设计学院, 福州 350202

摘要: 准确的情绪识别对心理健康问题的早期诊断和干预具有重要意义. 可穿戴生理传感器在情绪识别中展现出创新性的技术应用. 然而, 整合多种生理信号检测情绪是一项复杂而充满挑战的任务. 针对这些挑战, 提出了一种深度学习模型驱动的时间多模态融合方法, 以捕捉脑电图(EEG)和血容量脉搏(BVP)信号之间以及内部的非线性情绪相关性, 并提高情绪分类性能. 采用端到端的时间多模态深度学习模型, 将来自轻量级传感器的 EEG 和 BVP 信号进行融合, 以执行情绪识别任务. 通过使用卷积神经网络(ConvNet)和长短期记忆网络(LSTM)模型, 研究整合 EEG 和 BVP 信号, 共同学习并探索跨模态高度相关的表示形式. 通过智能可穿戴传感器收集数据集验证时间多模态融合方法, 并与最新研究结果进行比较. 实验结果表明, 该方法在情绪唤醒度估计上取得了 89.16% 的准确率, 与其他先进方法达到了相似水平. 将该方法应用于焦虑治疗评估, 以验证深度学习技术在心理健康应用中的有效性. 实验结果表明, 该方法成功提取了愉悦和唤醒估计值, 有效评估了时间域内的不同情绪变化, 为心理健康研究和治疗提供了有益的参考.

关键词: 情绪识别; 深度学习; 智能穿戴; 生理信号; 心理健康

中图分类号: TP393

文献标志码: A

文章编号: 1673-9868(2024)12-0189-13

开放科学(资源服务)标识码(OSID):



Application of Wearable Physiological Sensor-Driven Deep Learning Emotion Recognition Model in Mental Health Assessment

SHI Min¹, XU Lijun², PENG Huadong³

1. School of Education, Fuzhou University of International Studies and Trade, Fuzhou 350202, China;
2. Institute of Art and Design, Nanjing Institute of Technology, Nanjing 211167, China;
3. School of Art and Design, Fuzhou University of International Studies and Trade, Fuzhou 350202, China

Abstract: Accurate emotion recognition has significant implications for the early diagnosis and intervention of mental health issues. Wearable physiological sensors have exhibited innovative technological applica-

收稿日期: 2024-02-20

基金项目: 国家自然科学基金项目(61702225); 2023 年江苏省研究生教育教学改革项目(2023YJYJG02); 福州外语外贸学院课题(FWKQJ201910).

作者简介: 石敏, 博士, 副教授, 主要从事工业设计和感性工学研究.

tions in emotion recognition. However, integrating various physiological signals to detect emotions is a complex and challenging task. In response to these challenges, this paper proposes a deep learning model-driven time multimodal fusion approach to capture the nonlinear emotion correlations both internally and between electroencephalogram (EEG) and blood volume pulse (BVP) signals, and enhance emotion classification performance. The study employs an end-to-end time multimodal deep learning model, fusing EEG and BVP signals from lightweight sensors to perform emotion recognition tasks. Specifically, by utilizing Convolutional Neural Network (ConvNet) and Long Short-Term Memory (LSTM) models, this research integrates EEG and BVP signals, jointly learning and exploring highly correlated representations across modalities. The proposed multimodal fusion method is validated using a dataset collected through intelligent wearable sensors and compared with the latest research results. Experimental results demonstrate that the proposed method achieves an accuracy of 89.16% in estimating arousal levels, reaching a comparable level to other advanced methods. Additionally, this study applies the method for anxiety therapy assessment to validate the effectiveness of deep learning technology in mental health applications. Experimental results indicate that the proposed method successfully extracts valence and arousal estimates, and effectively evaluates different emotional changes in the time domain, providing valuable insights for mental health research and treatment.

Key words: emotion recognition; deep learning; wearable technology; physiological signals; mental health

情绪是人类行为中至关重要的因素,对每一个社会行为都产生影响。情绪识别一直是研究领域中蓬勃发展的话题,因为它涉及到人际互动,从而有助于通过工程学、心理学和认知科学更好地理解人类的情绪状态^[1]。通过这些方法进行情绪识别有助于在不使用外部医疗设备的情况下准确地理解心理健康和人类行为。情绪识别具有广泛的应用价值,例如可以帮助监测危急情况下个体的情绪状态,在临床环境中可用于监测患者的心理状况,在娱乐/视频游戏行业中可用于识别用户对特定视频、电影剪辑或游戏的情感反应,在消费者服务行业中则可以通过辨识用户对产品的反馈来改善营销策略或提升用户体验。

在心理学和精神医学领域,情绪在塑造人类行为、决策以及与他人互动方面扮演着至关重要的角色。情绪识别作为心理治疗的一部分,有助于及时发现和治疗情绪和精神障碍。对于某些心理障碍,如创伤后应激障碍(PTSD)或自闭症谱系障碍(ASD)等,当个体无法有效表达其情感状态时,情绪识别技术可以帮助医生和治疗师更好地了解患者的情感需求,进而制定更具针对性的治疗计划。从经济角度来看,影响识别系统也是有益的。例如,欧盟心理健康报告统计数据显示,欧洲每年因工作引起抑郁症所产生的成本高达数千亿欧元。

以往的研究已经验证了负面情绪与心血管疾病之间的关联。愤怒、焦虑或抑郁等负面情绪是心血管疾病患者的重要心理危险因素。这些负面情绪会引发生理反应,如高心率、高血压和低心率变异性(HRV),从而导致心血管疾病发作和死亡风险增加。INTERHEART 研究招募了来自 52 个国家的 24 767 名受试者,结果显示 33% 的急性心肌梗死事件可归因于社会心理危险因素,超过了与其他风险因素(如糖尿病、吸烟、肥胖以及不良饮食习惯和久坐不动)相关联的影响。此外,该项调查还发现抑郁和焦虑患者表现出脑电图皮层多动和额叶 α 波不对称性显著增加。然而,在探索使用多种生理信号监测情绪对待有心血管系统问题患者时,并没有太多相关的文献可借鉴。从心理健康角度出发,检测人类情绪与生理反应之间的关联具有重要意义。

随着人工智能的出现,利用多模态信号解决情感识别问题是情感计算的新兴趋势。多项研究利用最先进的深度学习方法,结合生理信号,如心电图(ECG)、EEG、皮肤温度以及面部表情、声音、姿势等,对情绪进行分类识别。此外,使用轻量级人体可穿戴传感器和先进的机器学习技术自动情绪识别已应用于不同的领域,如电脑游戏^[2]、电子健康^[3]和道路安全^[4]。然而,尽管情感计算作为一个科学领域已经确立了几十年,情感系统仍未成为我们日常生活的一部分。原因之一是可穿戴传感器数据与人体心理、生理状态之间的关系,并不像可穿戴传感器数据与人体物理状态之间的关系那样明确。例如,智能手机可以计算步数

并识别人类的身体活动(例如, 跑步与步行), 但无法高精度地识别情绪和相关情感状态(例如, 认知负荷)。人们能够使用智能健身设备监测运动和运动强度, 但其在捕捉个体的复杂情感状态, 如心情变化和焦虑水平方面仍然存在困难。

研究表明, 多模态数据能够显著提升情绪分类识别性能。不同来源的数据之间存在相互关联, 可为个体情绪提供互补信息。为了捕捉这种信息, 使用一组紧凑的潜在变量来捕获模态之间的相关性至关重要。然而, 在各种异质生理数据中学习潜在情绪信息是一个具有挑战性的问题, 这是因为不同生理信号由异构时间序列数据组成, 并且随着时间推移, 模态内部和模态之间存在某些情绪结构。

为了充分发挥多模态生理信号的优势, 构建一个更加准确和具备鲁棒性的情绪识别系统, 需要建立一种多模态融合模型。该模型能够捕捉每个模态内部以及模态之间的内在变化, 并且具备处理不同类型数据的能力。因此, 在可穿戴生理传感器和深度学习基础上, 本文提出一种时间多模态融合模型, 旨在提高可穿戴传感器的多模态生理信号情绪分类性能。

- 本文的创新之处/新颖点在于:
- 1) 引入一种时间多模态融合方法, 旨在捕捉脑电图(EEG)和血容量脉搏(BVP)信号之间以及内部的非线性情绪相关性, 从而提高情绪分类性能。
 - 2) 通过使用卷积神经网络(ConvNet)长短期记忆(LSTM)模型, 本文方法能够在时间上融合不同的生理信号, 从而更全面地理解情感的时间结构。
 - 3) 通过端到端的方式进行情感识别, 使用深度学习模型一次性学习每个模态的特征, 然后共同学习和探索跨模态的情感相关表示。这种端到端的方法有助于简化模型设计和训练流程, 并提高情感识别的性能。
 - 4) 本文不仅仅停留在情感计算的理论层面, 还将提出的方法应用于焦虑疗法评估, 拓展了情感计算技术在心理健康领域的实际应用范围。

1 理论分析与研究现状

1.1 情绪定义和生理信号研究

在许多与情感相关的研究中, 通过使用基于情感圆环模型的二维情感分类模型, 将人类情感分为唤醒和愉悦两个维度, 并将支配性应用于二维和三维扩展的情感模型。在当前研究中, 推断情绪状态的常用方法包括多种途径, 如面部表情、言语、身体手势和生理信号等。尽管面部表情因其直观性越来越受欢迎, 但与其他方式相比, 生理信号具备独特的优势。① 在可用性方面不断增强, 并且目前可穿戴设备的兴起也为其提供了更广泛的应用场景。② 对外界视觉噪声(如照明)有较高的抵抗能力, 因此非常稳定且具有广泛的适应性。③ 保真度高, 很难复制或掩盖这些信号来模拟特定情绪。④ 由于尺寸结构相对较小, 可以提高加工效率。

生理信号可以通过非侵入式传感器和可穿戴设备轻松测量, 包括 ECG、肌电图(EMG)、EEG、皮肤电反应(GSR)、皮肤温度(SKT)和呼吸频率(RSP)等多种与人类情感密切相关的信号。不同的研究采用不同数量的信号; 一些研究者使用 7 个生理信号来进行情感分类, 而有些研究者则仅利用 EEG 信号进行情感分类。许多情感分类研究已通过结合各种类型的信号、获取两个或更多信号(即多模态信号)得到实施。多模态信号是由具有多种感觉模式的信号组件构成, 有许多方法可以应用于多模态信号。雪雯等^[5]运用面部表情和 EEG 信号进行了情感识别和分析, 范习健等^[6]则利用面部表情和声音进行了情感识别和分析。需要注意的是, 虽然在相同条件下重复实验可能会导致结果差异, 但这并非与使用的特征数量直接相关。

此外, Cha 等^[7]对 EMG 活动进行了线性判别分析, 最早使用生理数据进行自动情绪估计, 指出个体生物信号与情绪活动之间存在相关性。Pinto 等^[8]提出一种多模态方法, 结合了 4 种不同的模态: 心率、皮肤电导(EDA)、呼吸和 EMG。通过这些更广泛的数据, 取得了相对更好的结果。

Dzedzickis 等^[9]设计了一个实验, 在其中引入包含特定类型情绪剪辑的影片供参与者观看。他们采用 k 最近邻算法、判别函数分析和 Marquardt 反向传播算法, 并利用了来自多个模态的特征(如 GSR、温度和心率)。Prabhakaran 等^[10]则采用支持向量机(SVM)分类器, 运用基于小波特征的 EDA 信号执行一种自动化

情感分类方法. 另一方面, Skaramagkas 等^[11]引入了 eSEE-d 数据库, 利用深度多层感知器(DMLP)网络对眼动和凝视特征进行分类, 并实现了对情感激发视频中的情感唤起和愉悦度水平的准确识别.

由于缺乏大量数据进行比较, Yin 等^[12]创建了 DEAP 数据集, 该数据集提供了大量的反应特征, 如 ECG、GSR 或 EDA、眼电图(EOG)和 EMG, 以扩大分析此类反应的可能性. Arpaia 等^[13]利用该数据集, 进一步引入双模深度去噪自编码器, 以双模 LSTM 为瓶颈提取生物信号和视觉信息的高级表示, 分析附加时间信息的影响. Hasnul 等^[14]的研究进一步探讨了从生物信号中构建潜在特征的可行性, 并有效压缩了类似 ECG 的数据. 然而, 迄今为止这些方法还没有完全分析和量化提取出来的生物潜在特征对最终模型估计的重要性.

1.2 多模态情绪识别

尽管大多数用于情绪识别的机器学习和深度学习框架都使用来自一种模式的数据, 即视频、音频或 EEG, 但最近研究者对上述模式的数据产生了相当大的兴趣. 传感器数据融合在提高情感检测的可靠性和准确性方面具有极大的优势, 且多模态系统已显示出优于单模态系统的性能^[15]. 多模态融合将来自许多不同类型传感器的数据结合在一起, 这样的融合主要在两个不同的水平上执行, 即特征级融合和决策级融合^[16].

1.2.1 特征级融合

在特征级融合方法(早期融合)中, 从不同模态派生的特征被组合成一个特征向量, 然后可以在该向量上训练分类器. 人类在面对面的互动中使用和整合多种感官线索来检测情感状态, 这是特征级融合背后的基本思想. 特征级融合的主要优点是, 在早期阶段多模态特征之间的关联可以带来更好的性能, 只需要在特征向量上进行一个学习阶段. 然而, 特征级融合也存在一些挑战. 由于从不同模态获得的特征可能具有不同的时间尺度, 因此将特征同步到相同格式可能会面临困难且计算昂贵. 由于使用特征级融合获得了大量的特征集, 如果训练数据集有限, 分类准确性可能会受到严重影响. 此外, 学习异构特征之间的交叉相关性也存在较大的困难.

1.2.2 决策级融合

在决策级融合方法(后期融合)中, 首先获取基于每种模态派生特征所做出的决策, 然后利用局部决策得到融合决策向量, 且该向量可用于最终的决策或分类. 与特征级融合相比, 决策级融合具有明显的优势, 因为所有的决策都具有统一格式, 从而可以轻松整合以避免同步问题. 此外, 在采用决策级融合时还能够根据不同模态应用最佳分类器或方法, 提供更大灵活性. 已经有研究注意到, 在情感研究中常使用决策水平融合, 并且研究人员更倾向于选择这种类型的融合方式而非特征级融合.

2 方法

本文提出的基于深度学习模型的时间多模态融合方法(EEG 和 BVP 信号), 用以捕捉模态内和模态间的时间情感结构. 深度学习模型基于端到端的卷积神经网络(ConvNet LSTM 网络)和两种不同的融合方法: 早期融合和后期融合. 在早期融合模型中, 原始脑电图和脉搏波形数据被输入 ConvNet 网络以提取特征, 然后所有生成的特征被串联形成联合特征向量. 创建的联合特征向量被送入 LSTM 网络, 接着是一个密集层和 Softmax 层用于情感分类. 在后期融合模型中, 原始脑电图和脉搏波形数据被输入 ConvNet, 然后是 LSTM 网络和密集层. 每个网络生成的特征被合并, 然后输入密集层和 Softmax 层进行情感分类.

这些模型在使用可穿戴生理传感器(Empatica E4 和 Emotiv Insight)收集的数据集上进行了评估. Empatica E4 和 Emotiv Insight 分别捕获脉搏波形和脑电图信号.

2.1 数据准备

每位参与者观看了 9 个视频片段, 因此每位参与者进行了 9 次试验. 每个试验都被标记为不同的情感类别. 每个试验记录 6 个信号通道: 5 个 EEG 通道和 1 个 BVP 通道.

为了为时间多模态学习准备数据, 本文在每个试验的每个通道上使用了滑动窗口策略. 我们应用一个滑动窗口, 创建了一组连续的固定大小的窗口, 具有固定的重叠度. 将这 6 个通道的输入表示为长度为 N 的序列, 即:

$$EEG_ch_1 = (ch_1^1, \dots, ch_1^{n-1}, ch_1^n, \dots, ch_1^N) \quad (1)$$

$$EEG_ch_2 = (ch_2^2, \dots, ch_2^{n-1}, ch_2^n, \dots, ch_2^N) \quad (2)$$

$$EEG_ch_3 = (ch_3^3, \dots, ch_3^{n-1}, ch_3^n, \dots, ch_3^N) \quad (3)$$

$$EEG_ch_4 = (ch_4^4, \dots, ch_4^{n-1}, ch_4^n, \dots, ch_4^N) \quad (4)$$

$$EEG_ch_5 = (ch_5^1, \dots, ch_5^{n-1}, ch_5^n, \dots, ch_5^N) \quad (5)$$

$$BVP = (BVP^1, \dots, BVP^{n-1}, BVP^n, \dots, BVP^N) \quad (6)$$

其中 $ch_1^n, \dots, ch_5^n$ 和 BVP^n 分别表示 $EEG_ch_1, \dots, EEG_ch_5$ 和 BVP 在时间 n 处的窗口。

所有生成的窗口都被视为新的训练数据示例, 其标签与原始试验相同。然后, 将每个通道分割成具有不同窗口大小(2 s、3 s、5 s 和 10 s)和 50% 重叠的连续窗口。将带通滤波(6 阶巴特沃斯滤波)、陷波滤波和独立成分分析(ICA)等预处理技术应用于脑电信号。为了消除 BVP 信号中的噪声和伪影, 应用了 3 Hz 低通巴特沃斯滤波器。此外, 本文使用零均值和单位方差对数据进行归一化。

2.2 基于早期融合的时间多模态深度学习

本节介绍了采用早期融合方法的时间多模态深度学习模型。在本文所提出的模型中, 时间生理信号在早期阶段(ConvNet 之后)被融合到联合表示序列中。图 1 描述了本文所提出的模型架构。

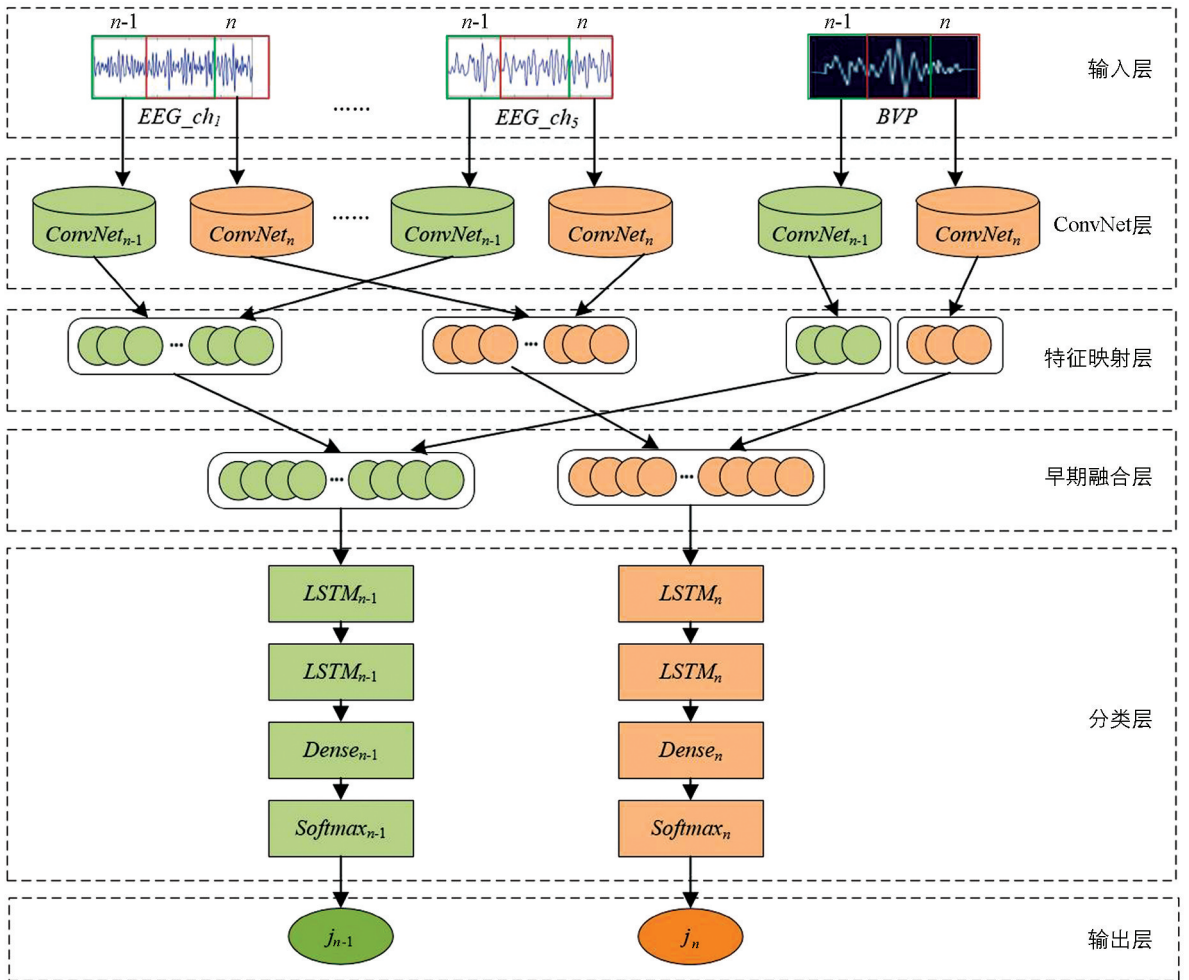


图 1 ConvNet LSTM 早期融合的时间多模态深度学习流程图

该系统的输入是 EEG(5 个 EEG 通道)和 BVP(1 个通道)信号。输入被分割成具有一定程度重叠(50% 重叠)的连续固定大小窗口。在时间点 n 将 6 个通道中每个通道创建的窗口馈送到两块 ConvNet 中提取特征图, 并将每个通道在时间点 n 的输出特征图连接起来构建联合表示。随后在时间 n 处创建的联合表示被馈送到 LSTM 的两个层中, 然后是密集层和 Softmax 层进行情感分类。

该模型由输入层、ConvNet 层、特征映射层、早期融合层、分类层和输出层组成。

输入层: 时间多模态深度学习模型很大程度上依赖于输入。为了应用时间多模态学习, 对每个 EEG 和 BVP 通道都采用了滑动窗口策略。来自每个通道时间 n 处的窗口被视为输入, 将被馈送到卷积神经网络进行训练。

ConvNet 层: 每个通道在时间点 n 的输入(即切片窗口)将被传递至两个特征提取器。在卷积神经网络的特征提取阶段, 考虑到不同通道之间的时间关联性, 每个通道在时间点 n 的输入都经过两个特征提取器处理, 包括卷积、激活、归一化和最大池化层。通过这些层次化的处理, 旨在学习每个通道在时间上的分层特征。鉴于所处理的是生理信号, 采用了 1D(One-Dimensional)卷积层。该设计的目标是在时间维度上捕获信号的动态特征, 以便更有效地体现时间关联性, 从而最终影响情感分类结果。基于该架构, 每个 EEG 和 BVP 信号在时间点 n 的窗口被引入 ConvNet 结构, 使得每个通道在时间点 n 的卷积神经网络输出对应于相应的特征映射。

特征映射层: 如果 $ConvNet_{ch1}^n$ 表示 EEG_{ch1} 的卷积神经网络, 并且 FM_{ch1}^n 表示时间点 n 处的相应特征映射, 则:

$$FW_{ch1}^n = ConvNet_{ch1}^n(ch_1^n) \quad (7)$$

为了实现每个通道的时间卷积神经网络学习, 需要考虑当前输入及其历史记录。同时, 为了获得特征图表示, 时间点 n 最近的每个模态历史($ConvNet^{t-1}$)被附加到当前窗口。

将每个脑电通道在时间点 n 处准备的特征图连接起来, 形成在时间点 n 处的 EEG 联合表示特征图。

早期融合层: 在这一层中, 每个时间步(n)的 EEG 联合表示特征映射和 BVP 特征映射被连接在一起, 构建一个单一的特征映射向量。

分类层: 在该层中, 使用两层 LSTM 网络后跟一个密集层和 Softmax 层来模拟时间 n 处多模态特征表示的整体时间动态。需要注意的是, LSTM 网络可以帮助学习时间情感结构, 因为 LSTM 网络包含了隐藏状态或记忆, 有助于存储先前的信息(隐藏层)和学习时间情感结构。

因此, LSTM 在时间点 n 的输出取决于先前的隐藏状态($n-1$)以及当前状态, 这些状态也可以捕获先前联合表示的时间模式。值得强调的是, 本文所提出的模型能够分别学习每个通道的时间模式, 同时使用联合表示跨模态学习时间模式。

2.3 基于后期融合的时间多模态深度学习

在本文所提架构中, EEG 通道和 BVP 信号基于后期融合方法进行时间融合。该模型的架构由输入层、ConvNet 层、特征映射层、LSTM 网络层、后期融合层和输出层组成(图 2)。此架构中的输入层、ConvNet 层与基于早期融合的时间多模态学习模型相同。

在输入层中, 首先是将 6 个通道(5 个 EEG 通道和 1 个 BVP 信号)在时间点 n 的窗口送入 ConvNet 层。

在 ConvNet 层中, 每个通道在时间点 n 时刻的输入(即切片窗口)会被传入两个特征提取器。此构架中 ConvNet 层与早期融合模型相同。

在特征映射层中, 每个通道的特征映射由 ConvNet 生成。将每个 EEG 通道的特征映射串联起来, 形成 EEG 模态的联合表示特征映射。每种模态的特征映射被首先送入两层 LSTM 网络, 然后被送入一个密集层。

在后期融合层中, 将两层 LSTM 生成的高级特征映射和每个模态在时间点 n 生成的密集层相结合, 构建联合表示层。

分类: 为了对不同的情绪进行分类, 在时间 n 的联合表示层被馈送到一个密集层和一个 Softmax 层。

该系统接收 EEG 信号(5 个通道)和 BVP 信号(1 个通道), 将其分成连续的固定大小窗口, 重叠程度为 50%。模型输出 4 种情绪(HA-P、HA-N、LA-P 和 LA-N)之一。6 个通道在时间点 n 生成的窗口被传递到单独的两块卷积神经网络中提取特征映射, 并结合 EEG 通道特征映射输出构建联合表示。每个模态(EEG 和 BVP)在时间点 n 处创建的联合表示被馈送到两个 LSTM 层和一个密集层中。两个模态在时间点 n 时密集层的输出被组合成一个联合表示, 然后被送入一个密集层, 再后送入一个 Softmax 层用于情感分类。

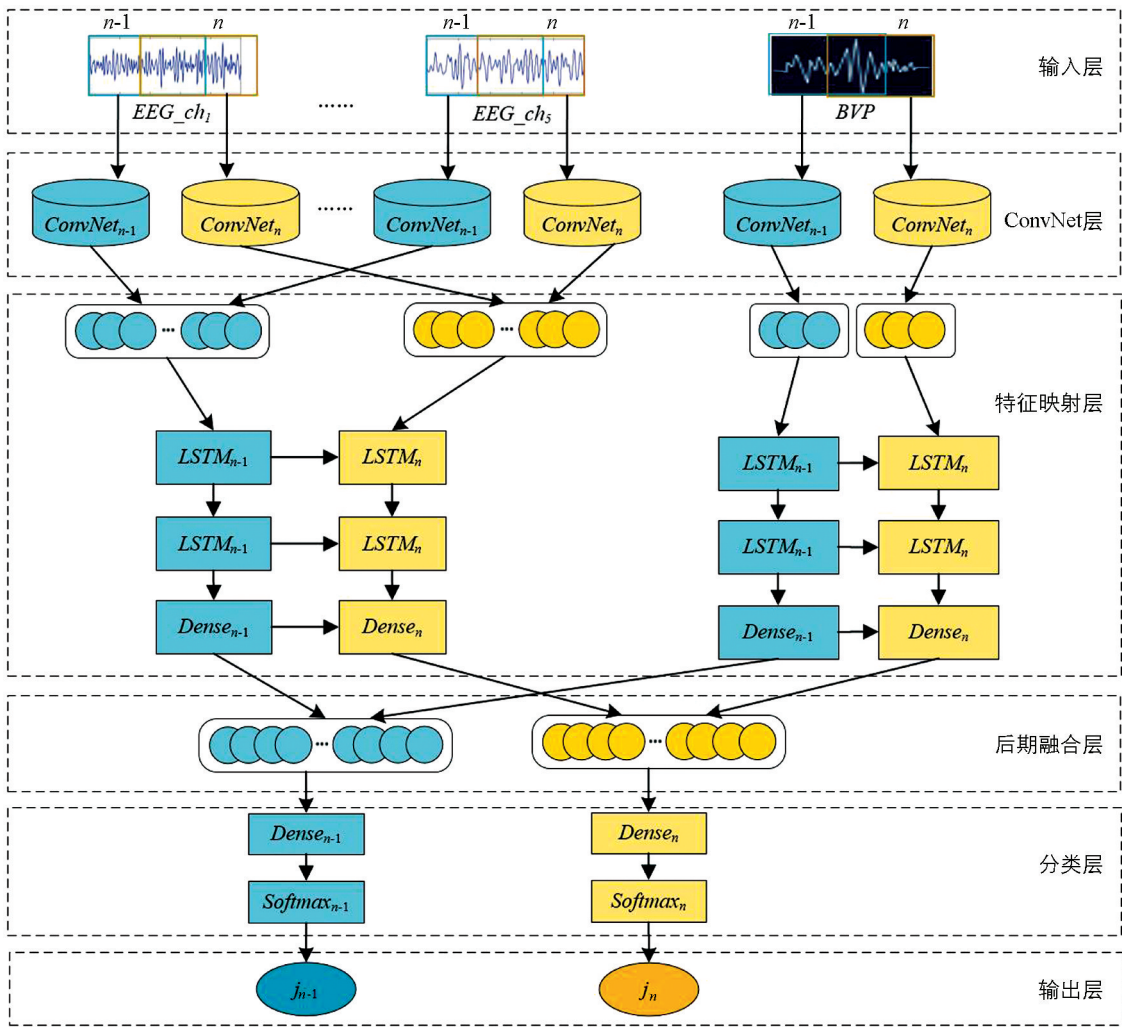


图 2 ConvNet LSTM 后期融合的时间多模态深度学习流程图

2.4 原始生理信号的卷积神经网络架构

使用 ConvNet, 首先学习局部非线性特征, 然后从较低层次的特征生成高级特征. ConvNet 由卷积层和多层处理层组成, 卷积层可以使用一组可学习的滤波器生成较低级别的特征, 多层处理层可以表示较高级别的特征. 许多 ConvNet 网络使用池化层来控制过拟合, 池化层可减少网络中的参数数量和表示的空间大小, 有助于避免过度拟合.

利用卷积神经网络, 首先学习局部非线性特征, 然后由低级特征生成高级特征. 卷积网络由卷积层和多层处理组成. 卷积层可以使用一组可学习的滤波器生成较低级别的特征, 多层处理可以表示较高级别的特征. 池化层常被用于卷积网络中控制过拟合. 池化层的主要作用是通过减少网络的参数数量和表示空间大小来降低模型的复杂度, 从而避免过拟合问题发生. 在卷积神经网络中, 池化层通常位于卷积层之后, 可以是最大池化层、平均池化层、自适应池化层等不同类型的池化层. 最大池化层会选取输入特征图中的最大值, 平均池化层会选取输入特征图中的平均值, 自适应池化层则会根据输入特征图的形状自适应地计算池化核的大小. 这些池化操作可以有效地降低特征图的维度, 减少模型的参数数量, 从而控制过拟合的发生.

本文提出了一个由两块卷积最大池化层组成的 ConvNet 网络(图 3). 卷积层、指数线性单元(ELU)、批标准化层和最大池化层构成了 ConvNet 的每个模块. 在卷积层中, 将时刻 n 的当前输入/窗口或前一层带有一组滤波器(Z)的输出进行卷积. 这一层能够使用可训练的过滤器捕获时间信息, 每个滤波器的输出根据以下公式计算得到:

$$j = frame^n * Z + h$$

(8)

其中, j 为输出结果, $frame$ 表示输入数据在时间点 n 处的值, Z 为卷积核, h 为偏置项, $*$ 为卷积算子. 本文所提出的方法使用的激活函数为指数线性单元(ELU), 通过以下函数映射前一层的输出:

$$ELU(i) = \alpha * (\exp(i) - 1), i < 0 \qquad ELU(i) = i, i \geq 0$$

(9)

其中: $ELU(i)$ 表示输入 i 经激活函数处理后的结果, i 为输入值, α 为 ELU 函数中的一个超参数, $\exp(i)$ 为指数函数.

批标准化层, 用于规范前一特征映射的输出. 本文采用了一个最大池化层, 以减少参数数量. 实际上, 最大池化层通过在定义的窗口(局部邻域)上取最大特征映射来对输入进行降采样.

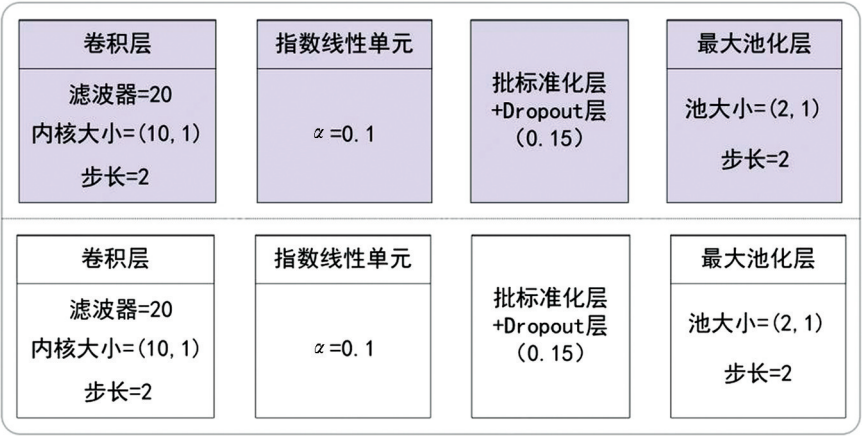


图 3 ConvNet 架构的两个模块

3 结果与分析

3.1 数据集

本文对 30 名年龄在 20 至 50 岁范围内受试者收集的数据集上的性能进行研究. 每个参与者观看了 MAHNOB 数据集中使用的 9 个视频片段, 以诱发不同的情绪. 这些视频剪辑包含从《纽约黑帮》《教父》《钢琴家》等热门电影中选出的电影场景, 并由心理学专家进行了注释.

这些视频基于不同情绪类别中标签数量最多的标准进行选择. 例如, 选择具有最多幸福标签的视频片段诱发幸福情绪. 参与者在观看视频片段时, 他们的脑活动(EEG)和心脏活动(BVP)使用轻便传感器进行捕捉. 本文使用 Emotiv Insight 和 Empatica E4 分别捕捉 EEG 和 BVP 信号(图 4). Emotiv Insight 仅包含 5 个通道(AF3、AF4、T7、T8 和 Pz)和 2 个参考通道(CMS、DRL). 通道的定位依据国际 10~20 系统(图 5).



图 4 Emotiv Insight 耳机和 Empatica E4 腕带

为了获取原始 BVP 和 EEG 信号, 本文使用了 Empatica Connect 和 TestBench 软件. 在观看每个视频

片段后, 要求每个参与者使用键盘表达他们的情绪状态(愉悦、唤醒、快乐、兴奋、中性、愤怒、沮丧、抑郁、悲伤、镇定和平静)。

本文所呈现的情绪状态被映射到情绪维度的四个象限。在第一步中, 要求参与者闭上眼睛放松约 1 min, 同时记录他们的基线 EEG 和 BVP 信号, 以减少眼部噪声。每个视频片段之间允许 1 min 的静音, 帮助防止当前情绪与之前情绪混淆。每个参与者观看 9 个视频片段, 并被要求在每个片段后报告他们的情绪(通过自我评估), 如图 6 所示。

经过人工分析, 本文对收集到的信号进行了筛选, 以确保数据质量, 并忽略了由轻量级 Emotiv 传感器产生的噪声和低质量的 EEG 信号, 这些噪声可能是由于电极移动或接触松动导致的。在去除噪声数据后, 本文使用了 30 名参与者中 25 名(12 名女性和 13 名男性)的 EEG 信号。这些传感器具有轻便且无线化等优势, 使其成为自然环境下进行自由生活研究最合适的选项。

对于使用可穿戴生理传感器(Empatica E4 和 Emotiv Insight)收集的数据集, 本文首先提供分析, 查看模型使用的每个模态对其最终估计结果的影响, 包括使用内部提取的潜在特征。其次, 为了从提出的方法中了解影响估计的质量, 将本文结果与该数据集相关文献报告的结果进行比较。然后, 使用性能最好的模型来验证本文收集的数据集上每个人的情绪状态。为了评估影响估计的质量, 本文遵循该数据集的原始人独立性协议。最后, 使用精确度分数(以百分比为单位)来判断每个估计影响标签的质量。

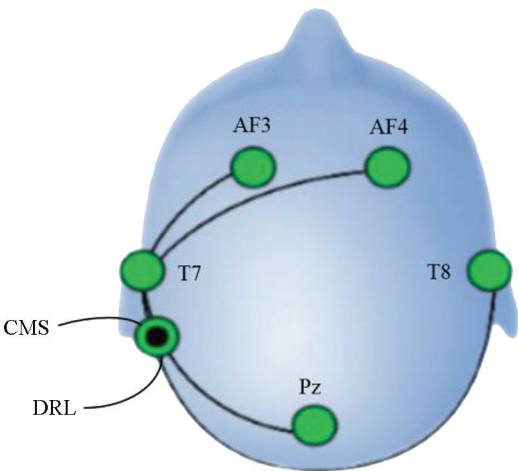


图 5 Emotiv 传感器 5 个通道位置示意图

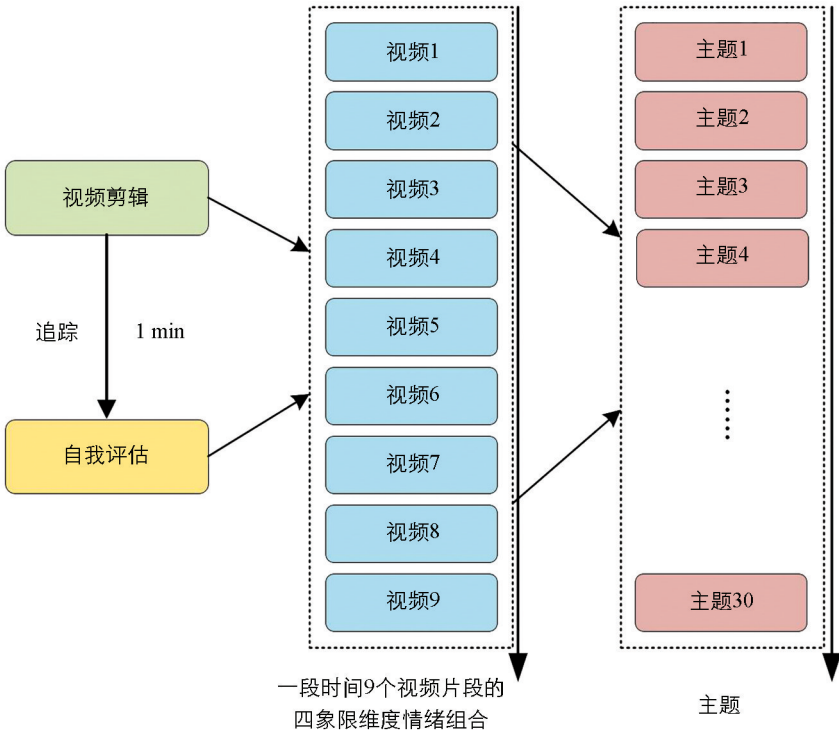


图 6 情绪诱发的实验方案图

3.2 评估指标

本文采用愉悦度(*Valence*)和唤醒度(*Arousal*)作为情绪识别性能的评价指标。这两个维度是情感状态的关键组成部分。

愉悦度表示情绪体验的积极或消极方向,是一个连续的维度,通常在情感空间的水平轴上表示.积极的情绪体验对应较高的愉悦度值,而消极的情绪体验对应较低的愉悦度值.这一维度通常涵盖了从极度不愉快到极度愉快的范围.

愉悦度(*Valence*)的计算公式为:

$$Valence = \sum_{i=1}^N P_i \cdot V_i$$

(10)

其中, P_i 为模型输出中第 i 个样本的真实标签(例如,愉悦或不愉悦), V_i 是对应样本的情绪预测得分.唤醒度表示情绪体验的活跃程度或强度,是情感空间垂直轴上的维度.高唤醒度对应情绪体验的强烈、兴奋或激动,而低唤醒度表示较为冷静和平静的情绪状态.唤醒度反映了情绪的生理激活水平.

唤醒度(*Arousal*)的计算公式为:

$$Arousal = \sum_{i=1}^N Q_i \cdot A_i$$

(11)

其中, Q_i 为第 i 个样本的真实标签(例如,高唤醒或低唤醒), A_i 为对应样本的情绪预测得分.在情感计算领域,使用这两个维度的组合可以更全面地描述和区分各种情绪状态.例如,高愉悦度和高唤醒度可能对应于兴奋或愉快的情绪,而低愉悦度和低唤醒度可能对应于沮丧或平静的情绪.这种细粒度情感描述对于情绪识别系统的性能评估至关重要,因为它能够更准确地捕捉个体的情绪状态.

3.3 消融研究

本文在消融研究中提供两个主要分析:模态和潜在特征影响分析.模态分析举例说明了各个模态作为模型输入的影响,并建立了基线准确性.潜在特征影响分析证实了纳入隐藏特征的好处,可以更准确地估计联合特征向量对模型的影响.

3.3.1 模态分析

为了查看每种模态对 ConvNet LSTM 网络最终估计值的单独贡献,本文将一种模态移到另一种模态来分别训练每种模态的 ConvNet LSTM 网络.这样产生两个经过训练的 ConvNet LSTM 网络,其中一个仅根据 EEG 信号进行训练,另一个仅使用 BVP 信号.然后,将它们的结果与使用多模态信号的本文模型进行比较,结果如表 1 所示.

表 1 消融实验对比结果

%

标签	模式		
	EEG 信号	BVP 信号	多模态
愉悦	57.3	54.1	67.2
唤醒	75.2	76.1	84.0
快乐	65.1	66.2	73.9
兴奋	72.5	69.2	81.3
中性	47.8	44.5	56.6
愤怒	36.4	33.1	45.2
沮丧	52.9	49.6	61.7
抑郁	55.7	52.4	64.5
悲伤	42.2	38.9	51.0
镇定	70.4	67.1	79.2
平静	71.6	68.3	80.4
平均值	58.8	56.3	67.7

从表 1 的结果中可知,仅利用 EEG 信号相比仅使用 BVP 信号的结果更好.主要原因是,EEG 信号在影响标签上产生了更稳定的准确性,平均准确率为 58.8%.虽然使用 BVP 信号会产生相似但不稳定的估计值,但在某些影响标签(例如唤醒和快乐)中具有更高的准确性,而在其他标签中则较低,平均准

确率为 56.4%。这种不稳定性可能是本文空间网络中包含高参数的结果。本文使用两种模态的平均准确率为 67.7%，且在所有情绪类别中都具有较高的准确率，表明多模态融合有助于对这一特定情绪维度进行分类。

3.3.2 潜在特征分析的影响

表 2 显示了本文模型与其他两个基线模型(文献[17]和文献[18])的对比结果。根据影响标签的总准确度，相较于其他两个基线模型，本文模型表现出更高的平均准确率。与文献[17]中的模型相比较，差异幅度较大。按准确率排序，在愤怒、中性和沮丧等情绪维度上，联合特征向量的引入显著提升了本文模型的性能。

表 2 本文模型与基线模型在不同情绪维度上的对比 %

标签	模型		
	文献[17]模型	文献[18]模型	本文模型
愉悦	66.8	69.9	70.7
唤醒	77.2	78.3	83.4
快乐	78.7	79.8	82.6
兴奋	53.8	54.9	57.2
中性	88.7	86.8	91.1
愤怒	88.6	87.1	94.7
沮丧	78.4	79.5	86.3
抑郁	66.5	65.6	67.9
悲伤	67.8	68.9	73.2
镇定	69.5	71.2	72.9
平静	72.1	74.7	75.5
平均值	73.5	74.2	77.8

进一步比较本文模型和文献[18]模型的对比结果，发现愤怒和沮丧情绪的准确率提高了约 8%，悲伤和中性情绪的准确性提高了约 5%，表明特征向量可能有助于估计这些情绪。需要注意的是，这些情绪状态对唤醒水平要求较高，这或许可以解释唤醒估计中约 6%的准确性增益(从文献[17]模型的 77.2%增加到本文模型的 83.4%)。基于这些发现本文得出结论：提取的潜在特征对模型估计有益，因为其在适当的集成情况下可以作为所有模态原始信号的补充信息。

3.4 与其他研究的比较

表 3 显示了最新研究方法与本文模型在收集的数据集上结果的比较。我们根据愉悦和唤醒领域来评估它们的结果，并考虑到它们在共识中的广泛使用，以及它们在所有比较研究中的可用性。

表 3 与其他最新研究的准确性比较

方法	模式	愉悦度/%	唤醒度/%
雪雯等 ^[5]	面部表情和 EEG 信号	47.65	58.76
范习健等 ^[6]	面部表情和声音	43.24	60.38
Cha 等 ^[7]	EMG 信号	40.57	55.29
Pinto 等 ^[8]	多模态	68.72	66.35
Dzedzickis 等 ^[9]	多模态	75.44	75.77
Skaramagkas 等 ^[11]	多模态	77.36	90.43
Arpaia 等 ^[13]	多模态	85.28	85.71
本文方法	多模态	84.81	89.16

由表 3 可以发现: ① 与其他方法相比, 本文模型产生了相对较好的结果, 最高唤醒估计准确率为 89.16%。尽管与 Skaramagkas 等^[11] 和 Arpaia 等^[13] 的研究相比, 在价精度上稍有下降, 但这两种模型使用了更复杂的脑电图特征, 并结合了其他 EOG 和 EMG 等生理信号及技术, 这要求更成熟、更先进的仪器设备。而本文所提出的模型不同于其他模型, 只需有效地利用 EEG 和 BVP 信号进行推理即可获得结果, 因此相对容易实施。另外, 一些不涉及面部和生物信号融合模态的方法与其他多模态方法(包括本文提出的方法)相比表现较差, 这一发现符合本文在模态分析中得出的结论: 多种输入方式对于结果具有重要性。

3.5 焦虑症治疗评估

使用基线模型中性能最佳的文献[18]模型, 并利用本文收集的数据集对焦虑疗法进行评估。我们在数据的前 15 min 子集上运行本文的模型, 以获取每位患者治疗前情绪状态的平均值。然后, 收集最后 15 min 的另一个估计值来代表治疗后患者的状态。本文累积了所有患者的这些结果, 并将其显示在相应象限位置上(图 7), 以观察变化情况。

从图 7 中可以看出, 本文模型能够产生相当合理的结果, 它显示了治疗前后情绪变化的理想趋势(从第二象限到第四象限)。此外, 我们还注意到, 本文模型估计位于每个象限的中心附近(即中性情绪), 这可能是由于潜在特征造成的, 这些特征取决于训练数据集所具有的特定属性。大多数样本都靠近中性位置, 因此仍然需要更多极端示例来完善训练。尽管如此, 在不太极端情况下, 本文方法仍然能够突出治疗的预期趋势, 并表明其在真实世界应用方面具有一定价值。

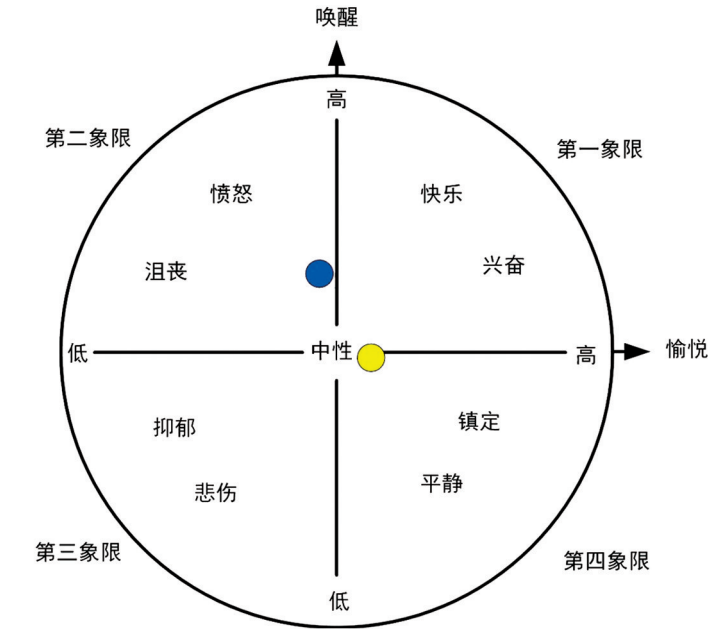


图 7 所有患者治疗前(蓝色)和治疗后(黄色)的愉悦和唤醒情绪变化

图 7 所有患者治疗前(蓝色)和治疗后(黄色)的愉悦和唤醒情绪变化

4 结论

基于生理信号的人类情绪识别研究在学术界和应用领域正受到日益广泛的关注。本文通过可穿戴生理传感器和深度学习技术, 构建了一个时间多模态融合模型, 旨在实现对情绪的准确识别。通过记录患者的 EEG 和血容量 BVP 信号, 构建了一个全面的多模态数据库, 覆盖了整个治疗周期。本文所提出的模型采用了 ConvNet LSTM 网络, 能够在时间上捕捉 EEG 和 BVP 生理信号之间的跨模态情绪相关性。深度学习的应用使得该模型能够学习非线性情绪结构, 并提高情绪分类性能。与其他相关研究相比, 实验结果表明了该模型的优越性。此外, 通过焦虑疗法评估实验, 本文模型成功地展示了患者情绪状态的理想演变趋势, 并突显了在实际应用中的潜力。这项研究不仅为构建情绪识别系统提供了先进的技术思路, 同时也为心理健康领域的实际应用提供了有益的参考和指导。

虽然本文模型在分类精度方面表现不错, 但与最新的研究方法相比, 在跨学科实验精度方面仍有很大的提升空间。未来, 我们将继续探索不同模态之间更深层次特征之间的关系, 并研究更优化的融合方法, 进一步提高多模态情绪识别性能。此外, 我们还将从模型结构和特征上进行优化改进, 提取更加优质的手动特征, 改善模型结构并降低复杂度, 提高分类精度。

参考文献:

[1] 杨集梅, 柴洁余, 邱天龙, 等. 共情与中国民族音乐情绪识别的关系: 来自 ERP 的证据 [J]. 心理学报, 2022, 54(10): 1181-1192.

[2] 邬宇涵, 苏茜, 负蓓, 等. 心理测评系统在人才测评中应用进展及启示 [J]. 护理学报, 2022, 29(21): 27-31.

[3] 刘欢欢, 王朝晖, 叶勤文, 等. 可穿戴技术在情绪识别中的应用进展及发展趋势 [J]. 纺织学报, 2022, 43(8): 197-205.

[4] 程洪, 黄瑞, 邱静, 等. 人机智能技术及系统研究进展综述 [J]. 智能系统学报, 2020, 15(2): 386-398.

[5] 雪雯, 陈景霞, 胡凯蕾, 等. 基于 EEG 和面部视频的多模态连续情感识别 [J]. 陕西科技大学学报, 2024, 42(1): 169-176.

[6] 范习健, 杨绪兵, 张礼, 等. 一种融合视觉和听觉信息的双模态情感识别算法 [J]. 南京大学学报(自然科学), 2021, 57(2): 309-317.

[7] CHA H S, IM C H. Performance Enhancement of Facial Electromyogram-Based Facial-Expression Recognition for Social Virtual Reality Applications Using Linear Discriminant Analysis Adaptation [J]. Virtual Reality, 2022, 26(1): 385-398.

[8] PINTO G, CARVALHO J M, BARROS F, et al. Multimodal Emotion Evaluation: A Physiological Model for Cost-Effective Emotion Classification [J]. Sensors, 2020, 20(12): 3510.

[9] DZEDZICKIS A, KAKLAUSKAS A, BUCINSKAS V. Human Emotion Recognition: Review of Sensors and Methods [J]. Sensors, 2020, 20(3): 592.

[10] PRABHAKARAN S, AYYAMPERUMAL N K. Automatic Detection of Artifacts and Improved Classification Models for Emotional Activity Detection from Multimodal Physiological Data [J]. Journal of Intelligent & Fuzzy Systems, 2023, 45(5): 8915-8929.

[11] SKARAMAGKAS V, KTISTAKIS E, MANOUSOS D, et al. ESEE-D: Emotional State Estimation Based on Eye-Tracking Dataset [J]. Brain Sciences, 2023, 13(4): 589.

[12] YIN G H, SUN S Q, YU D, et al. A Multimodal Framework for Large-Scale Emotion Recognition by Fusing Music and Electrodermal Activity Signals [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2022, 18(3): 1-23.

[13] ARPAIA P, CRAUSO F, DE BENEDETTO E, et al. Soft Transducer for Patient's Vitals Telemonitoring with Deep Learning-Based Personalized Anomaly Detection [J]. Sensors, 2022, 22(2): 536.

[14] HASNUL M A, AZIZ N A A, ALELYANI S, et al. Electrocardiogram-Based Emotion Recognition Systems and Their Applications in Healthcare-A Review [J]. Sensors, 2021, 21(15): 5015.

[15] TAN C, CEBALLOS G, KASABOV N, et al. Fusion Sense: Emotion Classification Using Feature Fusion of Multimodal Data and Deep Learning in a Brain-Inspired Spiking Neural Network [J]. Sensors, 2020, 20(18): 5328.

[16] SHARMA A, SHARMA K, KUMAR A. Real-Time Emotional Health Detection Using Fine-Tuned Transfer Networks with Multimodal Fusion [J]. Neural Computing and Applications, 2023, 35(31): 22935-22948.

[17] YAN X C, LIN Z H, LIN Z Y, et al. A Novel Exploitative and Explorative GWO-SVM Algorithm for Smart Emotion Recognition [J]. IEEE Internet of Things Journal, 2023, 10(11): 9999-10011.

[18] ASADZADEH S, REZAI T Y, BEHESHTI S, et al. Accurate Emotion Recognition Utilizing Extracted EEG Sources as Graph Neural Network Nodes [J]. Cognitive Computation, 2023, 15(1): 176-189.

责任编辑 夏娟