基于音视频信息的深度多模态抑郁症识别综述

张石清[1,2],张星楠[1,2],赵小明[2]
(1. 浙江理工大学 信息学院,浙江 杭州 310023;
2. 台州学院 智能信息处理研究所,浙江 台州 318000)

摘 要:抑郁症是一种精神疾病,严重时会导致自杀行为的发生。当前抑郁症患者人数正变得越来越多,越来越普遍化、年轻化。采用机器学习方法开展面向音频、视频等模态信息的多模态抑郁症识别研究已成为一个计算机科学、心理学、医学等多学科交叉的热点课题。近年来,新发展起来的深度学习技术也逐渐被应用于面向音频、视频等模态信息的多模态抑郁症识别中的深度特征提取任务。为了系统总结和归纳近年来深度学习技术在多模态抑郁症识别领域的研究进展,首先介绍了抑郁症的临床表现及心理学诊断方法,随后简要总结了现有的抑郁症数据集,并阐述了代表性深度学习技术的基本原理及进展情况;然后,系统分析和总结了面向音频、视频的多模态抑郁症识别涉及到的关键技术,包括手工特征提取和深度特征提取,以及多模态信息融合策略;最后,指出了该领域存在的机遇与挑战,并对下一步的研究方向进行了总结与展望。

关键词:抑郁症;深度学习;音频;视频;特征提取;多模态;融合方法
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2023)07-0001-11
doi:10.3969/j.issn.1673-629X.2023.07.001

A Survey of Deep Multimodal Depression Recognition Based on Audio-visual Cues

ZHANG Shi-qing[1,2],ZHANG Xing-nan[1,2],ZHAO Xiao-ming[2]
(1. School of Information,Zhejiang Sci-Tech University,Hangzhou 310023,China;
2. Institute of Intelligent Information Processing,Taizhou University,Taizhou 318000,China)

Abstract:Depression is a mental illness that can lead to suicidal behavior in severe cases. At present,depression is becoming larger,more common and younger. The use of machine learning methods to carry out multimodal depression recognition research oriented to the fusion of audio,video and other modal information has become a hot topic in computer science,psychology,medicine and other interdisciplinary subjects. In recent years,some newly deep learning techniques have also been gradually applied to the deep feature extraction task in multimodal depression recognition integrating audio,video and other modal information. In order to systematically summarize and conclude the research progress of deep learning technology in the field of multimodal depression recognition in recent years,we firstly introduce the clinical manifestations and psychological diagnosis methods of depression,and then briefly summarize the existing depression datasets, and analyze the basic principles and progress of representative deep learning techniques. Then,we systematically analyze and summarize the key technologies involved in multimodal depression recognition fusion audio and video,including manual feature extraction and deep feature extraction,as well as multimodal information fusion strategies. Finally,the opportunities and challenges in this field are pointed out,and the next research direction is summarized and prospected.

Key words:depression;deep learning;audio;video;feature extraction;multimodality;fusion method

0 引 言

近年来,抑郁症对社会和个人的影响越来越深,它存在于各个年龄段。抑郁症患者通常情绪低下并且寡言少语,与人沟通较少,难以专注于工作,而这种行为对于医生诊断抑郁症也造成了一定的难度[1]。抑郁症已经在世界范围内被公认为一种严重疾病,对医疗系统造成了很大的负担[2]。虽然药物治疗和精神治疗对于情绪改善具有一定的作用,但抑郁症的诊断对治疗

收稿日期:2022-07-22 **修回日期**:2022-11-24
基金项目:国家自然科学基金项目(62276180,61976149);浙江省自然科学基金项目(LZ20F020002)
作者简介:张石清(1980-),男,博士,教授,CCF 会员(42009M),通信作者,研究方向为模式识别、情感计算;张星楠(1996-),男,硕士,CCF 会员(I0508M),研究方向为图像处理。

至关重要。目前,抑郁症的诊断方法主要依赖患者自我报告的诊断和症状严重程度的临床判断等主观行为[3],受环境影响非常大。

面部非语言行为的动态激活对于测量抑郁的严重程度至关重要[4-5]。针对面部活动和表情[6-7]、头部姿势和运动[8-9],以及注视和眼睛活动[10],研究者已经提出了一些客观性的自动抑郁估计(Automatic Depression Estimation,ADE)技术,用来对抑郁症进行估计和分类。研究表明,抑郁症患者在行为、语音、面部动作等方面和正常人有所不同[11-12]。例如,Giannakakis 等[13]研究了从眼睛活动、口腔活动和头部运动中获取面部信息,用来识别和分析患者的压力和焦虑状态。现阶段抑郁症的诊断主要依靠经过长期训练的专业心理医生,成本高且效率低,而且结果往往带有主观性。因此,迫切需要一个客观的自动抑郁检测系统,作为一个辅助手段来帮助医生诊断抑郁症。目前,面向机器学习的自动抑郁检测技术逐渐兴起,备受关注。

早期面向机器学习的抑郁症自动检测技术大多采用手工设计的音频和视频特征参数和经典的机器学习方法。然而这些手工设计的特征参数可靠性不够,导致抑郁症自动识别效果不是很理想,有待进一步提高。近年来,新发展起来的深度学习方法[14-16]为解决该问题提供了线索。深度学习方法的本质是通过多层的网络结构从输入数据中自动学习高层次的特征表示。鉴于所具有的强大特征学习能力,目前深度学习方法已经在计算机视觉[17-18]、语音信号处理[19-20]、自然语言处理[21]等领域取得了巨大的成功。

近年来,深度学习方法也开始被应用于抑郁症自动识别领域,并取得了一些成果。为了总结近年来深度学习方法在抑郁症自动识别领域的研究现状和进展,拟在总结现有多模态抑郁识别数据集的基础上,对面向音视频信息的深度多模态抑郁症识别研究进展进行系统性分析和归纳,并指出该领域未来的研究机遇与挑战。

1 多模态抑郁识别数据集

目前,大多数抑郁症识别最常用的公开数据集是来源于 Audio/Visual Emotion Challenge (AVEC)挑战系列数据集。表 1 列出了一些常见的多模态抑郁识别数据集。目前,拥有完整面部元图像的数据集主要有 AVEC2013[22]和 AVEC2014[23]。包含音频的数据集主要有 AVEC2013[22]、AVEC2014[23]、DAIC-WOZ[24]、DementiaBack[25]和 FORBOW[26]。包含视频图像的数据集主要有 AVEC2013[22]、AVEC2014[23]、DAIC-WOZ[24]、DementiaBack[25]和 BlackDog[27]。包含文本的数据集主要有 AVEC2016[28]、ACEV2017[29]、AVEC2019[30]、Crisis Text Line[31]和 ReachOut Trigage Shared Task[32]。

表 1 抑郁检测数据集总结

数据集	评估方法	数据类型	人数	样本数
AVEC2013[22]	BDI-II	视频/音频	82	150
AVEC2014[23]	BDI-II	视频/音频	82	300
AVEC2016[28]	PHQ-8	音频/文本/视觉特征	N/A	N/A
AVEC2017[29]	PHQ-8	音频/文本/视觉特征	142	189
AVEC2019[30]	PHQ-8	音频/文本/视觉特征	N/A	275
Crisis Text Line[31]	Crisis counselor judgment	文本	843 982	4 800 万
DAIC-WOZ[24]	PHQ-8	视频/音频	142	189
DementiaBank [25]	HAM-D	视频/音频	226	N/A
BlackDog[27]	QIDS-SR	视频/音频	130	N/A
ORI[33]	LIFE	视频	8	N/A
FORBOW[26]	MADRS	音频	526	N/A

AVEC2013[22]和 AVEC2014[23]都采用了视听抑郁语言语料库的子集。AVEC2013 数据集包含了 3 个部分,分别是 train、dev 和 test。其中每个部分包含了 50 个视频,共 150 个视频。AVEC2014 子集中的录音只包括原始录音中 14 项任务中的两项任务:Northwind 和 Freeform。其中,Northwind 表示参与者大声朗读德语寓言《风与太阳》的节选,而 Freeform 表示参与者使用德语回答了一些问题,比如:“你最喜欢的菜是什么?”“你最好的礼物是什么,为什么呢?”等。AVEC2014 数据集也包含了 3 个部分:train、dev 和 test,其中每个部分又包含了 Northwind 和 Freeform 两个部分,共包含了 300 个视频。

AVEC2016[28]、AVEC2017[29]和 AVEC2019[30]包含了抑郁预测挑战,它们都采用了 DAIC-WOZ

(Distress Analysis Interview Corpus - Wizard of Oz)[24]的子集。与 AVEC2013、AVEC2014 不同的是,AVEC2016、AVEC2017 和 AVEC2019 都没有提供原始视频,只提供了原始的音频信号和提取的视频特征。DAIC-WOZ 数据集包含临床访谈。该访谈由一个叫 Ellie 的动画虚拟采访者进行,由另一个房间的采访者控制。该访谈被设计用来支持诊断诸如焦虑、抑郁、创伤后精神识别等心理疾病状态。DAIC-WOZ 包括记录了大量问答的音频和视频数据。DAIC-WOZ 标签则使用标准化的自我评估主观抑郁问卷 PHQ-8 进行诊断,每个记录都会被标记一个单独的值。

2　深度学习及抑郁检测中的应用

深度学习是一种纯粹自动从数据中学习特征的方法[34]。它以分层的层次网络结构直接从原始数据中学习出高层次的特征表示[35-39],现已在各种任务中表现出了优越的性能。在多模态抑郁识别中,深度学习模型可以提取多模态抑郁数据集中的深度特征,经过融合后预测抑郁水平。下面先介绍深度学习中经典的模型结构。

2.1　卷积神经网络(CNN)

卷积神经网络(Convolutional Neural Networks,CNN)启发于动物的视觉系统[40],最早由 Fukushima[41]提出。CNN 主要包括三部分:卷积层、池化层和全连接层。给定一个输入图片,经过多层卷积,每一层都经由一个激活函数,由卷积核提取出图片的高级特征。然后,经过全连接层,将提取的高级特征映射到一个一维向量。目前,CNN 在众多领域都取得了良好的特征学习性能,如人脸识别[42-43]、计算机视觉[44]、语音信号处理[45]、自然语言处理[46]等。

CNN 在图像领域有着得天独厚的优势,由于拥有共享卷积核,可以处理高维数据,自动提取特征。但是 CNN 也存在诸多的缺陷,比如当网络层次太深时,采用反向传播修改参数会使靠近输入层的参数变化较慢;采用梯度下降算法很容易使训练结果收敛于局部最小值而非全局最小值;池化层会丢失大量有价值信息,忽略局部与整体之间的关联性;由于特征提取的封装,为改进性能增加了不确定性。

为了克服 CNN 当前的不足和缺陷,许多新的 CNN 结构被提出来。Szegedy 等[47]提出了一种叫 GoogleNet 的 CNN 模型,提升了网络深度,同时使用了稀疏连接的卷积,使得大量参数同时避免了过拟合。Krizhevsky 等[48]提出了一种叫 AlexNet 的 CNN 模型。该网络拥有大量的参数和神经元,使用了非饱和神经元和 GPU 运算的卷积操作,并开发了一种“Dropout”的正则方法用于降低过拟合。He 等[49]提出了一个深度的残差学习框架,名为深度残差网络(Residual Net,Resnet)。除此以外,其它代表性的 CNN 模型包括 VGGNet[50]、DenseNet[51]、ShuffleNet[52]、MobileNet[53]、3D ResNet[54]、C3D[55]等。

2.2　循环神经网络(RNN)

循环神经网络(Recurrent Neural Networks,RNN)是一种具有前向传播的定向循环网络。每个输出不仅和现在的输入有关,还和之前所有的输入相关。虽然 RNN 可以有效处理时间序列数据,并应用于语音识别或者手写字识别任务[56],但是 RNN 也存在许多缺陷,比如在反向传播的过程中,存在梯度消失的问题[57]。此外,RNN 训练比较困难,所以 RNN 只能处理短时的时间序列问题。

为了解决传统 RNN 存在的问题,近年来研究者提出了更为先进的结构,以便可以处理更长的时间序列。Hochreiter 等[16]提出了一种名为长短期记忆(Long Short-Term Memory,LSTM)网络的模型。LSTM 避免了 RNN 中存在的梯度消失问题。Chao 等[58]提出了门循环单元(Gated Recurrent Unit,GRU)。Zhang 等[59]提出了双向长短期记忆(Bi-direction Long Short-Term Memory,BiLSTM)网络。近年来,LSTM 改进的模型还有 Tree-LSTM[60]、Graph LSTM[61]、SENTENCE LSTM[62]、LSTM-CNN 等。

2.3　基于深度学习的自动抑郁检测

目前,深度学习方法被大量应用于抑郁症识别领域。其中,CNN 模型常用于视频信号的抑郁检测,而 RNN/LSTM 等模型则用于音频信号的抑郁检测。Melo 等[63]提出了一种基于最大差分(Maximization-Differentiation)的深度神经网络模型,用于视频抑郁症识别。Zhou 等[64]提出了一种采用深度联合标签分布(Deep Joint Label Distribution)与度量学习(Metric Learning)的面部抑郁识别方法。李金鸣等[65]提出一种基于深度学习的音频抑郁症识别方法。赵张等[66]提出一种融合注意力机制和双向 LSTM 的音频抑郁识别方法。

3　音视频抑郁特征提取

3.1　手工音频特征提取

早期面向音频信号的抑郁症识别采用的手工特征主要有响度、音高、共振峰、音质特征、频谱特征(Spectral Features)[67],以及 Mel 频率倒谱系数(MFCC)等。

Otero 等[68]提出一种基于音频手工特征的抑郁预测模型。该方法提取音频的手工特征,包括 MFCC、频谱变换-感知线性预测(RASTA-PLP)、能量(Energy)和谱特征。然后,将每个特征集的段向量进行拼接,输

入到支持向量回归(SVR)获得抑郁预测结果。Cummins 等[69]采用高斯混合模型(Gaussian Mixture Model,GMM)提取语音抑郁特征,然后使用支持向量机(SVM)进行抑郁预测。Yalamanchili 等[70]利用提取的低层次(Low-level Descriptor,LLD)声学特征,如韵律特征、音质特征、谱特征等,训练一个抑郁分类模型,以便实现抑郁和非抑郁的二分类任务。Simantiraki 等[71]提取了声源(Glottal Source)相关的相位失真方差(Phase Distortion Deviation,PDD)特征用于抑郁检测。该特征通过相位成分估计声源特征,而声源特征和抑郁具有相关性。

手工音频特征提取方法比较简单,而且也取得了较好的抑郁识别性能。但是手工提取的音频特征是属于低层次的,可靠性不够,与高层次的抑郁音频特征存在"语义鸿沟"问题。

3.2 深度音频特征提取

目前,各种代表性的深度学习方法,如 DBN、CNN、RNN/LSTM 等,被应用于抑郁症识别中的音频特征提取任务,即从原始的音频信号中学习出高层次的音频特征用于后续的抑郁症识别。

Dong 等[72]提出一种基于声音和情绪线索的抑郁检测层次模型。该模型利用预训练好的深度残差网络(Resnet)模型从原始音频信号中提取说话人识别特征,并从频谱图中提取语音情感识别特征。然后,为了充分利用说话人的声音和情感差异之间的互补信息,将这两种深度语音特征结合起来,输入到一个由全连接层和模糊分类器构成的抑郁症检测层次化模型实现抑郁症严重程度的预测。He 等[73]提出一种基于 CNN 的面向音频信号的抑郁识别方法。该方法首先采用 CNN 从原始音频信号和低级描述符(Low-level Descriptors,LLD)特征中提取高层次特征;然后从音频信号频谱中提取一种鲁棒性的中位值扩展的局部二元模式特征(Median Robust Extended Local Binary Patterns,MRELBP);最后,将所有深度特征拼接后经过全连接层得到抑郁预测结果。Ma 等[74]提出一种名为 DepAudioNet 的音频抑郁分类方法。该方法将 CNN 和 LSTM 结合来编码声音通道中的抑郁特征用于抑郁识别。输入的音频信号经过 3 个一维卷积运算之后,采用 LSTM 进一步提取 128 维的深度特征,然后经过全连接层获得最后的抑郁预测结果。Zhao 等[75]提出一种层次化注意力转移网络用于音频抑郁识别。该方法由四个部分组成:(1)一个由编码解码器构成的教师(teacher)网络,用于训练语音识别以获得最初的注意力图(Attention Map);(2)一个较浅的学生(Student)网络作为模型的主体结构,用于训练抑郁识别,模拟教师网络;(3)一个层次化注意力自动编码器,用于获得丰富的特征表示,在此基础上可以进行监督训练;(4)主体学生抑郁模型加上一个层次化注意力网络,获得最终的抑郁识别结果。

综上,相比于手工音频特征提取方法(见表 2),深度音频特征提取方法可以通过搭建深度的神经网络模型来学习更高层次的抽象特征表示用于抑郁症识别,但是由于深度神经网络模型采用黑盒子(Black-box)的特征提取操作,导致它们往往无法给出其解释意义。

表 2 音视频抑郁特征提取方法的比较

模态	提出者	时间	特征表示	数据集	回归/分类	性能指标
音频	Cummins 等[69]	2013	Spectral 等	Mundt、Blackdog	分类	ACC:63.3%
音频	Otero 等[68]	2014	RPLP、SDC 等	AVEC2013	回归	MAE:8.38
音频	Ma[74]	2016	CNN、LSTM	DAIC-WOZ	分类	ACC:52%
音频	Simantiraki 等[71]	2017	声源、PDD	AVEC2014	分类	ACC:67%
音频	He 等[73]	2018	DCNN	AVEC2013、AVEC2014	回归	MAE:8.20 MAE:8.19
音频	Zhao 等[75]	2020	MFCC	AVEC2017	回归	MAE:4.20
音频	Yalamanchili 等[70]	2020	韵律、语音质量和光谱特征等	AVEC2016	分类	ACC:90%
音频	Dong 等[72]	2021	Resnet	AVEC2013、AVEC2014	回归	MAE:7.31 MAE:6.79
视频	Jan 等[76]	2014	运动历史直方图	AVEC2014	回归	MAE:8.30
视频	Kächele 等[77]	2014	LPQ	AVEC2013	回归	MAE:8.72
视频	Dhall 等[78]	2015	LBP-TOP、FisherVector	AVEC2014	回归	MAE:7.08
视频	Wen 等[79]	2015	LBP-TOP	AVEC2013	回归	MAE:8.22
视频	Zhu 等[80]	2018	双流 DCNN	AVEC2014	回归	MAE:7.58 MAE:7.47

续表 2

模态	提出者	时间	特征表示	数据集	回归/分类	性能指标
视频	Melo 等[81]	2019	3DCNN	AVEC2013、AVEC2014	回归	MAE:6.40 MAE:6.59
视频	He 等[82]	2020	CNN	AVEC2013、AVEC2014	回归	MAE:6.59 MAE:6.51
视频	Jazaery 等[83]	2021	3DCNN、RNN	AVEC2013、AVEC2014	回归	MAE:7.37 MAE:7.22
视频	周炫余等[84]	2021	BiLSTM、VGG16	JA-IPAD	分类	ACC:90.6%

注释:性能指标 ACC: Accuracy (准确率), MAE:Mean Absolute Error(平均绝对误差)

3.3　手工视频特征提取

一般的手工视频特征提取方法有特征动态历史直方图(Feature Dynamic History Histogram,FDHH)、运动历史直方图(Motion History Histogram,MHH)、三个正交平面的局部二值模式(Local Binary Pattern from Three Orthogonal Planes,LBP-TOP)[85]、局部相位量化(Local Phase Quantization,LPQ)、时空兴趣点(Space-Time Interest Points, STIP)[86]、局部二值模式(Local Binary Pattern, LBP)、局部三元模式(Local Ternary Pattern,LTP)[87]等。

Dhall 等[78]提出一种用于抑郁分析的时间分段 Fisher 向量方法。该方法使用 LBP-TOP 方法提取视频时空特征,然后计算出 Fisher 向量,输入到支持向量回归(SVR)获得抑郁识别结果。该方法对统计聚合技术进行了分析和比较,以便选取具有判别性的视频特征表示。Jan 等[76]从相应的视频和音频信号中提取表示抑郁状态下的面部和声音特征。然后,基于运动历史直方图提出了动态特征生成方法,用于提取视频中的动态特征。最后,利用偏最小二乘法(Partial Least Square,PLS)和回归法进行抑郁预测,并采用决策融合获得最终的抑郁检测结果。Wen 等[79]采用 LBP-TOP 方法提取面部区域子集中的时间信息及动态特征描述符,然后利用稀疏编码方法实现抑郁症的预测。Kachele 等[77]采用局部相位量化(Local Phase Quantization,LPQ)提取和抑郁相关的面部表达特征,然后结合支持向量机和多层感知器实现最终的抑郁症预测。

综上所述,手工视频特征提取方法,可以提取低层次的视频特征信息用于抑郁症识别,操作比较简单。但是和手工音频特征提取方法类似,该方法可靠性不够,提取的视频特征参数同样与高层次的抑郁视频特征存在“语义鸿沟”问题。

3.4　深度视频特征提取

目前,一些典型的深度学习方法,如 CNN、C3D、LSTM 等,被广泛用于提取视频图像的深度特征,用于抑郁识别。

Zhu 等[80]提出了一种基于双流(two-stream)CNN 的视频抑郁预测方法。该方法使用一个带有两个全连接层的双流 CNN 架构来联合学习视频中面部外观和动态的抑郁特征,并设置了一个集成外观和动态信息的联合调优层。He 等[82]提出了一种深度局部全局注意力卷积神经网络(Deep Local Global Attention Convolutional Neural Network,DLGA-CNN)的视频抑郁识别方法。该方法采用基于局部注意力的 CNN(Local Attention Based CNN,LA-CNN)关注局部面部抑郁特征,而使用基于全局注意力的 CNN(Global Attention Based CNN,GA-CNN)从整个面部区域学习全局抑郁模式。Jazaery 等[83]提出基于视频的深度时空特征编码的抑郁水平分析方法。该方法使用三维卷积神经网络(3D-CNN)学习两个不同尺度的时空特征,然后利用递归神经网络(RNN)进一步学习视频的时空特征。Melo 等[81]提出一种结合全局和局部的面部三维卷积抑郁检测方法。该方法将三维全局平均池化集成到 3DCNN 中,分别处理全脸区域和眼睛区域的视频片段,用于关注与分析抑郁高度相关的局部面部区域。周炫余等[84]提出了一种基于多模态数据融合计算的大学生心理健康自动评估方法。该方法采用的多模态数据包含文本数据、图像数据和学生特定时间段的网络数据。其中,使用 VGG16 提取图像数据的特征。

该方法在自构建的多模态心理评估数据集(JA-IPAD)上的测试表明,该模型能够精准评估大学生的心理健康状态。

综上所述,相比于手工视频特征提取方法(见表 2),深度视频特征提取方法不依赖于专业知识和繁琐的步骤,具有自动学习高层次的视频特征能力,受到外界影响(如光照、姿态等等)小。利用深度学习网络,可以在具有一定时间长度的视频中,提取静态和动态特征,或者提取全局特征和局部特征,也可以利用循环神经网络学习不同尺度的时空特征,往往取得比手工视频特征更好的抑郁识别性能。

4 多模态信息融合策略

多模态信息融合方法一般可以分为三种:特征层融合(Feature-level Fusion)、决策层融合(Decision-level Fusion)和模型层融合(Model-level Fusion)。这些融合方法各有优势和各自的应用场景。

4.1 特征层融合

特征层融合也叫早期融合(Early Fusion,EF)。在特征层融合中,将输入的多个特征数据直接级联得到一个总的特征向量,用于后续的分类或回归任务。在特征层融合中,用来融合的特征包括视觉特征、文本特征、音频特征和运动特征等。但是,特征层融合容易导致级联后的特征向量维度过高。

He 等[88]提出一种视听多模态抑郁识别方法。对于音频数据,提取说话速率以及低水平描述符(LLD)特征;对于视频数据,提取 LGBP-TOP、头部姿势、STIP 以及 Divergence-Curl-Shear (DCS)描述符特征。在特征层融合中,对于每个视频序列,通过主成分分析(PCA)处理后的音频和视觉特征被连接到一个高维特征向量中,输入到 SVR 进行抑郁预测。Joshi 等[89]提出一种包括视听融合的多模态抑郁症诊断方法。该方法使用 BoA(Bag Of Audio)框架获得音频特征;视频特征则使用 BoV(Bag Of Video)框架,计算 LBP-TOP 和 STIP。在特征融合方法中,为了避免拼接特征导致的数据过大,对组合特征进行主成分分析(PCA),然后使用 SVM 进行分类。Cummins 等[90]提出一种融合听视觉的多模态抑郁识别方法。该方法将 GMM-UBM 范式和包含一阶二阶的 MFCC 结合起来,用于提取音频特征;采用时空趣点(Space-Temporal Interesting Point,STIP)和定向梯度的金字塔直方图(Pyramid Histogram of Oriented Gradients,PHOG)来提取视频特征。该方法分别测试了单音频、单视频和音视频融合的抑郁评估结果。在多模态信息融合中,采用特征融合方法实现。考虑各特征间时间维度的不相关性,对上述特征进行长度方向的拼接融合,最后使用 SVR 进行抑郁评估。

4.2 决策层融合

决策层也叫晚期融合(Late Fusion,LF)。在每个模态获得各自的决策结果之后,再将这些决策结果按照某种代数运算规则,比如最小值、最大值、平均值等,进行组合,得到最终的结果。但是,这种基于规则的决策层融合方法将不同模态独立开来,因而可能无法揭示不同模态之间的关系。

Meng 等[91]提出一种融合音频和视频的抑郁识别方法。对于视频数据,该方法先采用运动历史直方图提取动态特征,然后提取 LBP 和边缘定向直方图(Edge Orientation Histogram,EOH)特征,并将 LBP 和 EOH 特征在特征层上直接拼接输入到偏最小二乘(Partial Least Square,PLS)进行抑郁预测,得到视频抑郁检测结果;对于音频数据,先提取低水平描述符(LLD),然后使用 MHH 提取音频动态特征,并采用 PLS 得到音频预测结果。最后,使用线性联合先验(Linear Opinion Pool,LOP)方法对结果做决策融合,并得到最终的抑郁检测结果。Yang 等[92]提出一种集成深度和浅层模型混合架构的多模态抑郁分析方法。对于音视频数据,该方法首先采用 CNN 模型分别对音频和视频进行训练,之后冻结 CNN 的权重值并丢弃其最后一个全连接层参数,同时接入一个新的六层 DNN,获得音视频的抑郁预测结果。对于文本数据,提取 5 个段落矢量(Paragraph Vector,PV)描述符输入到 SVM 获得文本的抑郁识别结果。最后,建立一个抑郁分类的随机森林(Random Forest,RF)模型,对上述获得的音视频结果和文本结果做决策融合获得最终的抑郁检测结果。Yang 等[93]提出一种基于决策树的融合音视频和语言信息的抑郁分类方法。该决策树将语言信息与低层次音视频特征取得的结果进行决策融合。音频特征采用了共振峰、韵律和音质特征;视频特征使用了 HOG、眼睛注视特征和头部姿态特征的直方图。该方法针对男女性别分开训练,使用 SVR 及 LLR 进行测试,获得 PHQ 值。

4.3 模型层融合

模型层融合方法是对每个模态分别进行建模,并同时考虑模态之间的相互关联性。目前,采用神经网络的模型层融合方法被广泛应用于抑郁症识别。

Lin 等[95]提出一种基于 BiLSTM 和 CNN 的自动抑郁检测方法。该方法由三个部分组成:第一部分为带注意力层的 BiLSTM 用来学习访谈序列的语言特征;第二部分为一维 CNN 学习语音信号 Mel 频谱特征;第三部分由一个全连接层将前两个模型的输出进行融合,获得最终的抑郁检测结果。Ray 等[94]提出了一种基于多层次注意力的融合文本、音频和视频的多模态抑郁症预测方法。该方法对输入特征采用一种多层次注意力操作,以便让更有影响力的特征获得更大的权重。对于视频数据,该方法提取姿势、凝视和面部动作单元(Facial Action Unit,FAU)特征,输入到 BiLSTM 进行时间动态建模获得视频特征;对于音频数据,提取包含 MFCC 的 LLD 特征输入到 BiLSTM 进行时间动态建模获得音频特征;对于文本数据,采用预训练好的通用句子编码器(Universal Sentence Encoder)[97]提取文本特征,输入到 BiLSTM 进行上下文建模获得文本特征。最后,将上述得到的视频、音频和文本特征通过多层注意力网络进行融合获得最终的抑郁预测结果。Zhang 等[96]提出一种基于多模态深

度去噪自编码器(Multimodal Deep Denoising Autoencoder,MultiDDAE)的抑郁症识别方法。该方法采用多模态深度去噪自动编码器提取视听特征,然后使用 Fisher 向量编码产生会话级(Session-level)特征表示。对于文本数据,使用段落矢量(Paragraph Vector,PV)方法提取文本特征。最后,将视听特征与文本特征进行串联,然后输入到一个多任务的深度神经网络上进行融合,输出最终的抑郁症识别结果。

表 3　多模态抑郁症识别中的融合方法比较

方法	时间	融合方法	数据集	特征表示	性能指标
Joshi 等[89]	2013	特征层融合	Blackdog	音频:基频、响度、强度和 MFCC,视频:STIP、LBP-TOP	ACC:91.7%
Cummin 等[90]	2013	特征层融合	AVEC2013	音频:MFCC,视频:STIP、PHOG	N/A
He 等[88]	2015	特征层融合	AVEC2014	音频:MHH,视频:MHH、Bow、VLAD	MAE:6.16
Meng 等[91]	2013	决策层融合	AVEC2013	音频:LLD、MHH,视频:MHH、LBP、EOH	MAE:8.72
Yang 等[93]	2016	决策层融合	AVEC2016	音频:共振峰、韵律、语言特征,视频:HOG、眼睛注视、头部姿势	MAE:6.70
Yang 等[92]	2021	决策层融合	AVEC2016	音频:共振峰、韵律、语音质量,视频:位移范围的直方图 HDR	MAE:5.38
Ray 等[94]	2019	模型层融合	E-DAIC、DAIC-WOZ	音频:MFCC、EGE、BoAW,视频:pose、gaze、AUs	MAE:4.02
Lin 等[95]	2020	模型层融合	DAIC-WoZ、AViD-Corpus	音频:1DCNN,文本:BiLSTM、Attention	MAE:9.30
Zhang 等[96]	2020	模型层融合	AVEC2019	音频:MFCC,视频:Lankmarks、Gaze、AUs	ACC:89.3%

注释:性能指标 ACC: Accuracy (准确率), MAE:Mean Absolute Error(平均绝对误差)

综上所述,现有的多模态信息融合方法主要包括特征层融合、决策层融合和模型层融合(见表 3)。其中,特征层融合方法最简单,但是容易导致级联后的特征向量维度过高而出现“维度灾难”问题。决策层融合方法采用某种代数运算规则对不同模态取得的结果进行组合,从而获得最终的结果。然而,这种基于代数运算规则的决策层融合方法是将不同模态相互独立出来,没有考虑不同模态之间的相互关联性。模型层融合方法是一种同时考虑模态之间的相互关联性的方法,通常能够获取比特征层融合方法、决策层融合方法更好的性能。目前,采用注意力机制[94,97]在模型层上实现多模态信息的交互融合,已经成为当前一种主流的模型层融合方法。但是,模型层融合方法的计算复杂度一般比较高。因此,如何设计计算复杂度低而性能又好的模型层融合方法,是未来一个重要研究方向。

5　挑战和机遇

5.1　数据集问题

第一,医院需要保护患者诊断数据的隐私,使得不同的医疗机构无法收集和共享数据,这极大地影响了模型的准确性[98]。由于单一医疗机构无法收集到足够的高质量数据,该模型的预测能力无法达到临床辅助的作用。其次,虽然有许多隐私保护机器学习算法,但很难获得良好的训练效果。

第二,由于各医疗机构之间的巨大差距,它们所拥有的患者数据差异很大。为了处理各种情况,算法和模型需要具有较高的泛化能力,而在没有数据交换情况下,模型很难获得足够的准确性和特异性。

第三,缺乏足够的标签数据已经严重阻碍深度学习等技术在自动抑郁识别中的应用。虽然,深度学习在算法和模型上有了很大的进步,但在自动抑郁识别上还是受限于标签数据的缺乏。如果拥有更大的标签数据集,并且数据分布平衡,会对基于深度学习的自动抑郁识别领域产生积极的影响[12]。

第四,现有数据集的模态选择还不够丰富,缺乏诸如和抑郁症相关的脑电图[99-100]、人格特性和情绪的数据信息,这些特征可能和抑郁症有直接的关联,并且可以作为多模态抑郁检测模型的输入,提高模型的准确性。

5.2　集成更多模态问题

第一,目前,多模态抑郁症识别主流的方法是采用音视频信息为主,而忽视了将音视频信息与其它模态信息,如网络社交信息文本、脑电信号等相融合,以便进一步改善多模态抑郁症识别性能。采用网络社交的文本信息进行抑郁分析,近年来获得越来越多的关注。该数据可以从网络大量获得,并且从社交文本中可提取和抑郁症相关的显著特征。脑电信息则是和抑郁症高度相关的特征,抑郁症的脑电信息是一种非常有效

的用于抑郁检测的生理特征。因此,在采用的音视频信息基础上,集成网络社交信息文本、脑电信号等其它模态信息用于多模态抑郁症识别将是一个非常有意义的研究课题。

第二,现有文献大多只关注音频和视频的副语言信息,如说话率、面部动作单位(AUs),而不是关注口语内容中的语言信息[92]。然而,后者可以直接反映个人的睡眠状态、情绪状态、感觉和其他精神分析症状。

5.3 深度学习技术自身缺陷问题

第一,传统的RNN结构,包括LSTM、BiLSTM和GRU,可以有效处理短期时间序列。然而,它们并不能有效地处理长期序列。随着序列长度的增加,由于这些RNN结构模型自身存在的遗忘问题,导致它们的性能会迅速下降。由于抑郁症的复杂性,抑郁症的检测过程往往需要进行更长时间的诊断才有效。因此,对于抑郁症的检测,如何更加有效地处理长序列的视听数据,将是一个极具挑战的问题。

第二,抑郁症的症状与情感、个性等其他心理学因素密切相关。因此,采用多任务学习(Multi-task Learning)方法,结合情感、个性等因素开展自动抑郁检测是一个非常有意义的研究课题。此外,当前的数据集数据量很少,多任务学习也可以在数据稀疏时防止过拟合。由于有不同的任务参与训练,可以增强模型的泛化性。

第三,目前,大部分抑郁症识别采用的是监督学习方法,而对于自监督方法的报道甚少。自监督学习可以通过数据本身创建一个类似于标签的向量,使得大量的无标签数据可以利用。对于拥有较少标签数据量的自动抑郁识别来说,自监督学习的探索是一个很好的方向。

6 结束语

该文系统性总结了深度学习在面向音视频信息的多模态抑郁症识别中的应用现状及研究进展。首先,回顾了深度学习技术的发展历史,并介绍了基本的深度学习模型原理。在音视频特征提取的部分,总结和归纳了手工音频特征、深度音频特征、手工视频特征和深度视频特征的提取方法,并对各种特征提取技术进行了比较。对于面向音视频信息的多模态信息融合方法,重点分析了特征层融合、决策层融合和模型层融合等方法的应用。最后,指出了当前的自动抑郁检测中存在的问题和未来的发展方向。

参考文献:

[1] KATON W,SULLIVAN M D. Depression and chronic medical illness[J]. J Clin Psychiatry,1990,51(Suppl 6):3-11.

[2] DEWA C S,CHAU N,DERMER S. Examining the comparative incidence and costs of physical and mental health-related disabilities in an employed population[J]. Journal of Occupational and Environmental Medicine,2010,52(7):758-762.

[3] MUNDT J C,SNYDER P J,CANNIZZARO M S,et al. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology[J]. Journal of Neurolinguistics,2007,20(1):50-64.

[4] ELLGRING H. Non-verbal communication in depression[M]. Cambridge:Cambridge University Press,2007:76-87.

[5] PHILIPPOT P,FELDMAN R S,COATS E J. Nonverbal behavior in clinical settings[M]. New York:Oxford University Press,2003:171-209.

[6] COHN J F,KRUEZ T S,MATTHEWS I,et al. Detecting depression from facial actions and vocal prosody[C]//2009 3rd international conference on affective computing and intelligent interaction and workshops. Amsterdam:IEEE,2009:1-7.

[7] STRATOU G,SCHERER S,GRATCH J,et al. Automatic nonverbal behavior indicators of depression and PTSD:exploring gender differences[C]//2013 Humaine association conference on affective computing and intelligent interaction. Geneva:IEEE,2013:147-152.

[8] JONES I H,PANSA M. Some nonverbal aspects of depression and schizophrenia occurring during the interview[J]. Journal of Nervous and Mental Disease,1979,167(7):402-409.

[9] MURPHY-CHUTORIAN E,TRIVEDI M M. Head pose estimation in computer vision:a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2009,31(4):607-626.

[10] ALGHOWINEM S,GOECKE R,WAGNER M,et al. Eye movement analysis for depression detection[C]//2013 IEEE international conference on image processing. Melbourne:IEEE,2013:4220-4224.

[11] LOW L S A,MADDAGE N C,LECH M,et al. Detection of clinical depression in adolescents' speech during family interactions[J]. IEEE Transactions on Biomedical Engineering,2011,58(3):574-586.

[12] PAMPOUCHIDOU A,SIMOS P G,MARIAS K,et al. Automatic assessment of depression based on visual cues:a systematic review[J]. IEEE Transactions on Affective Computing,2019,10(4):445-470.

[13] GIANNAKAKIS G,PEDIADITIS M,MANOUSOS D,et al. Stress and anxiety detection using facial cues from videos[J]. Biomedical Signal Processing and Control,2017,31:89-101.

[14] XU S,WANG J,SHOU W,et al. Computer vision techniques in construction:a critical review[J]. Archives of Computational Methods in Engineering,2021,28(5):3383-3397.

[15] YADAV U, SHARMA A K. Review on automated depression detection from audio visual clue using sentiment analysis [C]//2021 second international conference on electronics and sustainable communication systems (ICE- SC). Coimbatore:IEEE,2021:1462-1467.

[16] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[17] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. San Juan:IEEE, 2016:770-778.

[18] 姜建勇,吴　云,龙慧云,等. 基于 CenterNet 的实时行人检测模型[J]. 计算机工程,2021,47(10):276-282.

[19] AFOURAS T,CHUNG J S,SENIOR A,et al. Deep audio-visual speech recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 44 (12): 8717-8727.

[20] PURWINS H,LI B,VIRTANEN T,et al. Deep learning for audio signal processing[J]. IEEE Journal of Selected Topics in Signal Processing,2019,13(2):206-219.

[21] GAO Z,FENG A,SONG X,et al. Target-dependent sentiment classification with BERT[J]. IEEE Access, 2019, 7: 154290-154299.

[22] VALSTAR M, SCHULLER B, SMITH K, et al. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge[C]//Proceedings of the 3rd ACM international workshop on audio/visual emotion challenge. New York:Association for Computing Machinery,2013:3-10.

[23] VALSTAR M, SCHULLER B, SMITH K, et al. AVEC 2014:3D dimensional affect and depression recognition challenge[C]//Proceedings of the 4th international workshop on audio/visual emotion challenge - AVEC '14. Orlando:ACM Press,2014:3-10.

[24] GRATCH J,ARTSTEIN R,LUCAS G,et al. The distress analysis interview corpus of human and computer interviews [C]//Proceedings of the ninth international conference on language resources and evaluation (LREC'14). Reykjavik: European Language Resources Association (ELRA),2014: 3123-3128.

[25] BECKER J T, BOILER F, LOPEZ O L, et al. The natural history of Alzheimer´s disease: description of study cohort and accuracy of diagnosis[J]. Archives of Neurology,1994, 51(6):585-594.

[26] DUMPALA S H,RODRIGUEZ S,REMPEL S,et al. Significance of Speaker embeddings and temporal context for depression detection[J]. arXiv:2107.13969,2021.

[27] MCINTYRE G,GÖCKE R,HYETT M,et al. An approach for automatically measuring facial activity in depressed subjects[C]//2009 3rd international conference on affective computing and intelligent interaction and workshops. Amsterdam:IEEE,2009:1-8.

[28] VALSTAR M, GRATCH J, SCHULLER B, et al. AVEC 2016:depression, mood, and emotion recognition workshop and challenge[C]//Proceedings of the 6th international workshop on audio/visual emotion challenge. Amsterdam: ACM,2016:3-10.

[29] RINGEVAL F,SCHULLER B,VALSTAR M,et al. AVEC 2017: real-life depression, and affect recognition workshop and challenge[C]//Proceedings of the 7th annual workshop on audio/visual emotion challenge. Mountain View California:ACM,2017:3-9.

[30] RINGEVAL F,SCHULLER B,VALSTAR M,et al. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition[C]// Proceedings of the 9th international on audio/ visual emotion challenge and workshop. New York:ACM,2019:3-12.

[31] THOMPSON L K,SUGG M M,RUNKLE J R. Adolescents in crisis:a geographic exploration of help-seeking behavior using data from crisis text line[J]. Social Science & Medicine,2018,215:69-79.

[32] MILNE D N, PINK G, HACHEY B, et al. CLPsych 2016 shared task: triaging content in online peer-support forums [C]//Proceedings of the third workshop on computational linguistics and clinical psychology. San Diego: Association for Computational Linguistics,2016:118-127.

[33] HOPS H, BIGLAN A, TOLMAN A, et al. Living in family environments (LIFE) coding system: reference manual for coders[M]. Eugene: Oregon Research Institute, 1995: 20-37.

[34] LECUN Y,BENGIO Y,HINTON G. Deep learning[J]. Nature,2015,521(7553):436-444.

[35] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[C]//Proceedings of the thirteenth international conference on artificial intelligence and statistics. Sardinia: PMLR,2010:249-256.

[36] ERHAN D, COURVILLE A, BENGIO Y, et al. Why does unsupervised pre-training help deep learning? [C]//Proceedings of the thirteenth international conference on artificial intelligence and statistics. Sardinia: JMLR, 2010: 201-208.

[37] LE ROUX N,BENGIO Y. Representational power of restricted Boltzmann machines and deep belief networks[J]. Neural Computation,2008,20(6):1631-1649.

[38] BENGIO Y. Learning deep architectures for AI[J]. Foundations and Trends® in Machine Learning, 2009, 2(1): 1-127.

[39] 王志鹏,王　涛. 基于 Faster RCNN 的穿越围栏违规行为检测[J]. 计算机系统应用,2022,31(4):346-351.

[40] HUBEL D H,WIESEL T N. Receptive felds,binocular interaction,and functional architecture in the cat's visual cortex [J]. Journal of Physiology,1962,160:106-154.

[41] FUKUSHIMA K. Neocognitron:a self-organizing neural network model for a mechanism of pattern recognition unafected by shif in position[J]. Biological Cybernetics,1980,36(4):193-202.

[42] SHARMA S,SHANMUGASUNDARAM K,RAMA-SAMY S K. FAREC-CNN based efficient face recognition technique using Dlib[C]//2016 international conference on advanced communication control and computing technologies (ICACCCT). Kovilpatti:IEEE,2016:192-195.

[43] BEN FREDJ H,BOUGUEZZI S,SOUANI C. Face recognition in unconstrained environment with CNN[J]. The Visual Computer,2021,37(2):217-226.

[44] YANG R,SINGH S K,TAVAKKOLI M,et al. CNN-LSTM deep learning architecture for computer vision-based modal frequency detection[J]. Mechanical Systems and Signal Processing,2020,144:106885.

[45] MUSTAQEEM,KWON S. A CNN-assisted enhanced audio signal processing for speech emotion recognition[J]. Sensors,2020,20(1):183.

[46] YIN W,KANN K,YU M,et al. Comparative study of CNN and RNN for natural language processing[J]. arXiv:1702.01923,2017.

[47] SZEGEDY C,ZAREMBA W,SUTSKEVER I,et al. Intriguing properties of neural networks[J]. arXiv:1312.6199,2014.

[48] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM,2017,60(6):84-90.

[49] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. New York:IEEE,2016:770-778.

[50] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv:1409.1556,2015.

[51] HUANG G,LIU Z,VAN DER MAATEN L,et al. Densely connected convolutional networks[J]. arXiv:1608.06993,2018.

[52] ZHANG X,ZHOU X,LIN M,et al. ShuffleNet:an extremely efficient convolutional neural network for mobile devices[C]//2018 IEEE/CVF conference on computer vision and pattern recognition. Salt Lake City:IEEE,2018:6848-6856.

[53] HOWARD A G,ZHU M,CHEN B,et al. MobileNets:efficient convolutional neural networks for mobile vision applications[J]. arXiv:1704.04861,2017.

[54] HARA K,KATAOKA H,SATOH Y. Learning spatio-temporal features with 3D residual networks for action recognition[C]//2017 IEEE international conference on computer vision workshops (ICCVW). Venice:IEEE,2017:3154-3160.

[55] TRAN D,BOURDEV L,FERGUS R,et al. Learning spatio-temporal features with 3D convolutional networks[C]//2015 IEEE international conference on computer vision (ICCV). Santiago:IEEE,2015:4489-4497.

[56] GRAVES A,LIWICKI M,FERNÁNDEZ S,et al. A novel connectionist system for unconstrained hand-writing recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2009,31(5):855-868.

[57] JOZEFOWICZ R,ZAREMBA W,SUTSKEVER I. An empirical exploration of recurrent network architectures[C]//Proceedings of the 32nd international conference on machine learning. Berlin:PMLR,2015:2342-2350.

[58] CHO K,VAN MERRIENBOER B,GULCEHRE C,et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv:1406.1078,2014.

[59] ZHANG S,ZHENG D,HU X,et al. Bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 29th Pacific Asia conference on language,information and computation. Shanghai:ACL,2015:73-78.

[60] TAI K S,SOCHER R,MANNING C D. Improved semantic representations from tree-structured long short- term memory networks[J]. arXiv:1503.00075,2015.

[61] PENG N,POON H,QUIRK C,et al. Cross-sentence nary relation extraction with graph LSTMs[J]. Transactions of the Association for Computational Linguistics,2017,5:101-115.

[62] ZHANG Y,LIU Q,SONG L. Sentence-state LSTM for text representation[C]//Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1:long papers). Melbourne:Association for Computational Linguistics,2018:317-327.

[63] DE MELO W C,GRANGER E,LOPEZ M B. MDN:a deep maximization-differentiation network for spatio-temporal depression detection[J]. IEEE Transactions on Affective Computing,2021:1-1,DOI:10.1109/TAFFC.2021.3072579.

[64] ZHOU X,WEI Z,XU M,et al. Facial depression recognition by deep joint label distribution and metric learning[J]. IEEE Transactions on Affective Computing,2022,13(3):1605-1618.

[65] 李金鸣,付小雁. 基于深度学习的音频抑郁症识别[J]. 计算机应用与软件,2019,36(9):161-167.

[66] 赵 张,汪静莹,耿馨佚,等. 融合注意力机制与双向长短时记忆网络的基于语音分析的抑郁识别方法[J]. 复旦学报:自然科学版,2021,60(6):733-739.

[67] ZHAO Z,ZHAO Y,BAO Z,et al. Deep spectrum feature representations for speech emotion recognition[C]//Proceedings of the joint workshop of the 4th workshop on affective social multimedia computing and first multi-modal affective computing of large-scale multi-media data. New York:Association for Computing Machinery,2018:27-33.

[68] LOPEZ-OTERO P, DACIA-FERNANDEZ L, GARCIA-MATEO C. A study of acoustic features for depression detection[C]//2nd international workshop on biometrics and forensics. Valletta:IEEE,2014:1-6.

[69] CUMMINS N, EPPS J, SETHU V, et al. Modeling spectral variability for the classification of depressed speech[C]//Interspeech 2013. Lyon:ISCA,2013:857-861.

[70] YALAMANCHILI B, KOTA N S, ABBARAJU M S, et al. Real-time acoustic based depression detection using machine learning techniques[C]//2020 international conference on emerging trends in information technology and engineering (ic-ETITE). Vellore:IEEE,2020:1-6.

[71] SIMANTIRAKI O, CHARONYKTAKIS P, PAMPOUCHIDOU A, et al. Glottal source features for automatic speech-based depression assessment[C]//Interspeech 2017. Stockholm:ISCA,2017:2700-2704.

[72] DONG Y, YANG X. A hierarchical depression detection model based on vocal and emotional cues[J]. Neurocomputing,2021,441:279-290.

[73] HE L, CAO C. Automated depression analysis using convolutional neural networks from speech[J]. Journal of Biomedical Informatics,2018,83:103-111.

[74] MA X, YANG H, CHEN Q, et al. DepAudioNet:an efficient deep model for audio based depression classification[C]//Proceedings of the 6th international workshop on audio/visual emotion challenge. Amsterdam:ACM,2016:35-42.

[75] ZHAO Z, BAO Z, ZHANG Z, et al. Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders[J]. IEEE Journal of Selected Topics in Signal Processing,2020,14(2):423-434.

[76] JAN A, MENG H, GAUS Y F A, et al. Automatic depression scale prediction using facial expression dynamics and regression[C]//Proceedings of the 4th international workshop on audio/visual emotion challenge-AVEC '14. Orlando:ACM,2014:73-80.

[77] KÄCHELE M, GLODEK M, ZHARKOV D, et al. Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression [C]//Proceedings of the 3rd international conference on pattern recognition applications and methods. Loire Valley:SCITEPRESS-Science and and Technology Publications,2014:671-678.

[78] DHALL A, GOECKE R. A temporally piece-wise fisher vector approach for depression analysis[C]//2015 international conference on affective computing and intelligent interaction (ACII). Xi'an:IEEE,2015:255-259.

[79] WEN L, LI X, GUO G, et al. Automated depression diagnosis based on facial dynamic analysis and sparse coding[J]. IEEE Transactions on Information Forensics and Security,2015,10(7):1432-1441.

[80] ZHU Y, SHANG Y, SHAO Z, et al. Automated depression diagnosis based on deep networks to encode facial appearance and dynamics[J]. IEEE Transactions on Affective Computing,2018,9(4):578-584.

[81] DE MELO W C, GRANGER E, HADID A. Combining global and local convolutional 3D networks for detecting depression from facial expressions[C]//2019 14th IEEE international conference on automatic face gesture recognition (FG 2019). Lille:IEEE,2019:1-8.

[82] HE L, CHAN J C W, WANG Z. Automatic depression recognition using CNN with attention mechanism from videos[J]. Neurocomputing,2021,422:165-175.

[83] AL JAZAERY M, GUO G. Video-based depression level analysis by encoding deep spatiotemporal features[J]. IEEE Transactions on Affective Computing, 2021, 12(1):262-268.

[84] 周炫余,刘　林,陈圆圆,等. 基于多模态数据融合的大学生心理健康自动评估模型设计与应用研究[J]. 电化教育研究,2021,42(8):72-78.

[85] ALMAEV T R, VALSTAR M F. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition[C]//2013 Humaine association conference on affective computing and intelligent interaction. Geneva:IEEE,2013:356-361.

[86] LAPTEV I, MARSZALEK M, SCHMID C, et al. Learning realistic human actions from movies[C]//2008 IEEE conference on computer vision and pattern recognition. Anchorage:IEEE,2008:1-8.

[87] LIAO W H. Region description using extended local ternary patterns[C]//2010 20th international conference on pattern recognition. DC:IEEE,2010:1003-1006.

[88] HE L, JIANG D, SAHLI H. Multimodal depression recognition with dynamic visual and audio cues[C]//2015 international conference on affective computing and intelligent interaction (ACII). Xi'an:IEEE,2015:260-266.

[89] JOSHI J, GOECKE R, ALGHOWINEM S, et al. Multimodal assistive technologies for depression diagnosis and monitoring [J]. Journal on Multimodal User Interfaces,2013,7(3):217-228.

[90] CUMMINS N, JOSHI J, DHALL A, et al. Diagnosis of depression by behavioural signals:a multimodal approach [C]//Proceedings of the 3rd ACM international workshop on audio/visual emotion challenge. Barcelona:ACM,2013:11-20.

[91] MENG H, HUANG D, WANG H, et al. Depression recognition based on dynamic facial and vocal expression features using partial least square regression[C]//Proceedings of the 3rd ACM international workshop on audio/visual emotion challenge. Barcelona:ACM,2013:21-30.

[92] YANG L, JIANG D, SAHLI H. Integrating deep and shallow models for multi-modal depression analysis—hybrid archi-

(下转第 40 页)