中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2024)06-1607-21

论文引用格式:Tao J H, Fan C H, Lian Z, Lyu Z, Shen Y and Liang S. 2024. Development of multimodal sentiment recognition and understanding. Journal of Image and Graphics, 29(06):1607-1627(陶建华, 范存航, 连政, 吕钊, 沈莹, 梁山. 2024. 多模态情感识别与理解发展现状及趋势. 中国图象图形学报, 29(06):1607-1627)[DOI:10.11834/jig.240017]

多模态情感识别与理解发展现状及趋势

陶建华[1],范存航[2],连政[3],吕钊[2],沈莹[4],梁山[5*]

1. 清华大学自动化系,北京　100084; 2. 安徽大学多模态认知计算安徽省重点实验室,合肥　230601; 3. 中国科学院自动化研究所,北京　100190; 4. 同济大学软件学院,上海　457001; 5. 西安交大利物浦大学智能工程学院,苏州　215123

摘　要:情感计算是人工智能领域的一个重要分支,在交互、教育、安全和金融等众多领域应用广泛。单纯依靠语音、视频单一模态的情感识别并不符合人类对情感的感知模式,在受到干扰的情况下识别准确率会迅速下降。为了充分挖掘不同模态数据的互补性,多模态融合的情感识别研究正日益受到研究人员的广泛重视。本文分别从多模态情感识别概述、多模态情感识别与理解、抑郁症情感障碍检测及干预3个维度介绍多模态情感计算研究现状。本文认为具备可扩展性的情感特征设计、基于大模型迁移学习的识别方法将是未来的发展方向,并在解决抑郁、焦虑等情感障碍方面的作用日益凸显。

关键词:情感识别;多模态融合;人机交互;抑郁状态评估;情感障碍干预;认知行为疗法

Development of multimodal sentiment recognition and understanding

Tao Jianhua[1], Fan Cunhang[2], Lian Zheng[3], Lyu Zhao[2], Shen Ying[4], Liang Shan[5*]

1. Department of Automation, Tsinghua University, Beijing　100084, China; 2. Anhui Province Key Laboratory of Multimodal Cognitive Computation, Anhui University, Hefei　230601, China; 3. Institute of Automation, Chinese Academy of Sciences, Beijing　100190, China; 4. School of Software Engineering, Tongji University, Shanghai　457001, China; 5. School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou　215123, China

Abstract: Affective computing is an important branch in the field of artificial intelligence (AI). It aims to build a computational system that can automatically perceive, recognize, understand, and provide feedback on human emotions. It involves the intersection of multiple disciplines such as computer science, neuroscience, psychology, and social science. Deep emotional understanding and interaction can enable computers to better understand and respond to human emotional needs. It can also provide personalized interactions and feedback based on emotional states, which enhances the human-computer interaction experience. It has various applications in areas such as intelligent assistants, virtual reality, and smart healthcare. Relying solely on single-modal information, such as speech signal or video, does not align with the way humans perceive emotions. The accuracy of recognition rapidly decreases when faced with interference. Multimodal emotion understanding and interaction technologies aim to fully model multidimensional information from audio, video, and physiological signals to achieve more accurate emotion understanding. This technology is fundamental and an important prerequisite for achieving natural, human-like, and personalized human-computer interaction. It holds significant value for ushering in the era of intelligence and digitalization. Multimodal fusion for sentiment recognition receives increasing atten-

收稿日期:2024-01-06;**修回日期**:2024-02-18;**预印本日期**:2024-02-25

*通信作者:梁山　shan.liang@xjtlu.edu.cn

基金项目:国家自然科学基金项目(62201572,62201002,62101553,62306316)

Supported by:National Natural Science Foundation of China(62201572, 62201002,62101553, 62306316)

tion from researchers in fully exploiting the complementary nature of different modalities. This study introduces the current research status of multimodal sentiment computation from three dimensions: an overview of multimodal sentiment recognition, multimodal sentiment understanding, and detection and assessment of emotional disorders such as depression. The overview of emotion recognition is elaborated from the aspects of academic definition, mainstream datasets, and international competitions. In recent years, large language models (LLMs) have demonstrated excellent modeling capabilities and achieved great success in the field of natural language processing with their outstanding language understanding and reasoning abilities. LLMs have garnered widespread attention because of their ability to handle various complex tasks by understanding prompts with minimal or zero-shot learning. Through methods such as self-supervised learning or contrastive learning, LLMs can learn more expressive multimodal representations, which can capture the correlations between different modalities and emotional information. Multimodal sentiment recognition and understanding are discussed in terms of emotion feature extraction, multimodal fusion, and the representation and models involved in sentiment recognition under the background of pre-trained large models. With the rapid development of society, people are facing increasing pressure, which can lead to feelings of depression, anxiety, and other negative emotions. Those who are in a prolonged state of depression and anxiety are more likely to develop mental illnesses. Depression is a common and serious condition, with symptoms including low mood, poor sleep quality, loss of appetite, fatigue, and difficulty concentrating. Depression not only harms individuals and families but also causes significant economic losses to society. The detection of emotional disorders starts from specific applications, which selects depression as the most common emotional disorder. We analyze its latest developments and trends from the perspectives of assessment and intervention. In addition, this study provides a detailed comparison of the research status of affective computation domestically, and prospects for future development trends are offered. We believe that scalable emotion feature design and large-scale model transfer learning based methods will be the future directions of development. The main challenge in multimodal emotion recognition lies in data scarcity, which means that data available to build and explore complex models are insufficient. This insufficiency causes difficulty in creating robust models based on deep neural network methods. The above mentioned issues can be addressed by constructing large-scale multimodal emotion databases and exploring transfer learning methods based on large models. By transferring knowledge learned from unsupervised tasks or other tasks to emotion recognition tasks, the problem of limited data resources can be alleviated. The use of explicit discrete and dimensional labels to represent ambiguous emotional states has limitations due to the inherent fuzziness of emotions. Enhancing the interpretability of prediction results to improve the reliability of recognition results is also an important research direction for the future. The role of multimodal emotion computing in addressing emotional disorders such as depression and anxiety is increasingly prominent. Future research can be conducted in the following three areas. First, research and construction of multimodal emotion disorder datasets can provide a solid foundation for the automatic recognition of emotional disorders. However, this field still needs to address challenges such as data privacy and ethics. In addition, considerations such as designing targeted interview questions, ensuring patient safety during data collection, and sample augmentation through algorithms are still worth exploring. Second, more effective algorithms should be developed. Emotional disorders fall within the psychological domain, and they can also affect the physiological features of patients, such as voice and body movements. This psychological-physiological correlation is worthy of comprehensive exploration. Therefore, improving the accuracy of algorithms for multimodal emotion disorder recognition is a pressing research issue. Finally, intelligent psychological intervention systems should be designed and implemented. The following issues can be further studied: effectively simulating the counseling process of a psychologist, promptly receiving user emotional feedback, and generating empathetic conversations.

Key words: sentiment recognition; multimodel fusion; human-computer interaction; depression detection; emotion disorder intervention; cognitive behavior therapy

0 引 言

情感计算旨在构建一个可以对人类情感自动感知、识别、理解与反馈的可计算系统，是人工智能领域的重要研究方向，是一个涉及计算机科学、脑与心理科学以及社会科学等多学科交叉的研究领域。深度的情感理解与交互可以使计算机更好地理解和响应人类的情感需求，并根据情感状态进行个性化的交互和反馈，从而提升人机交互的体验。相关技术在智能助理、虚拟现实和智慧医疗等领域都有广泛的应用场景。

多模态情感理解与交互技术旨在充分建模音、视频以及生理信号多个维度信息，以实现更准确的情感理解是实现自然化、拟人化和人格化人机交互的基础性技术和重要前提，对开启智能化、数字化时代具有重大价值。随着人工智能技术应用的不断深化，人类对计算机情感理解与交互的需求愈加广泛。提高计算机情感理解与交互的深度对推动产业变革、技术演化和行业进步有重大价值。情感计算研究也成为国际学术界的研究热点，受到广泛关注。根据中国科学院科技战略咨询研究院发布的《2021年研究前沿热度指数》报告显示，“多模态情感计算”相关研究热度指数位居前列。

多模态情感理解与交互技术应用潜力广泛体现在多个领域中，包括智能助手领域、社交媒体舆情分析与个性化内容推荐以及智慧医疗等多个领域，有着重大研究价值。因此，本文通过综述多模态情感计算涉及的基本定义、特征提取和模型融合等方面的最新研究进展，并结合抑郁症这一数量最多的情感障碍的早期评估及干预，探讨情感计算的具体应用，希望本论文可以帮助初学者快速了解和熟悉多模态情感计算的基本概念、最新发展以及具体应用，启发相关科研人员做出更有价值的研究工作。

本文分别从多模态情感识别概述、多模态情感识别与理解、抑郁症情感障碍检测及评估3个维度介绍多模态情感计算研究现状。其中情感识别概述分别从学术定义、主流数据集和国际比赛3个方面展开；多模态情感识别与理解分别从情感特征提取、多模态融合以及预训练大模型背景下的情感计算3个维度对情感识别涉及的表征与模型进行介绍；情感障碍检测从具体应用出发，选取抑郁症这一最为广泛的情感障碍作为典型，分别从评估与干预两个角度分析其最新进展与发展趋势。此外，本文还详细比对了情感计算国内外的研究现状，并对未来发展趋势进行了展望。

1 国际研究现状

1.1 多模态情感识别概述

情感识别是一项通过分析情感表达时所引发的生理和行为反应来识别情感状态的技术。作为人工智能领域的一个重要分支，情感识别在交互、教育、安全和金融等众多领域应用广泛。目前，通过分析语音信号、心电信号、人脸表情以及其他生理信号等单一模态数据来识别人类情感状态的研究已经取得了一定的进展。但是，单一模态的情感识别实际上并不符合人类对情感的感知模式。当人类试图隐藏情感信息或者单一通道的信号受到其他模态干扰时，情感识别性能会明显下降。鉴于不同模态的互补性，多模态融合的情感识别研究正日益受到学术界和工业界的重视，构建多模态情感识别系统已被认为是提高情感识别性能和鲁棒性有效手段之一（Sebe等，2005；Soleymani等，2017）。本节将介绍多模态情感理解的基本定义、主流数据集以及主要国际比赛等内容，对情感计算技术进行概述。

1.1.1 学术定义

情感是人对客观事物是否满足自己的需要而产生的态度体验，是人类的一种重要本能。在人们的日常生活交流中，情感也是不可或缺的重要部分（Shott，1979）。计算机图灵奖得主马文·明斯基教授在《脑智社会（The Society of Mind）》专著中专门指出情感在交互过程中的重要性：“问题不在于智能机器能否有情感，而在于没有情感的机器能否实现智能”（Minsky，1988）。情感有助于快速传递信息和理解真实意图，是人机交互的关键组成部分。

除了喜怒哀乐等基本情感外，人类表达的情感有时非常复杂。例如，悲喜交加、百感交集等情感是多种基本情感的复合体。目前，学术界常用的情感表示模型主要归为两大类：离散情感模型和维度情感模型。

离散情感模型将情感表示为离散实体。尽管有文献对离散情感进行了归纳，但是目前学术界并没有针对离散情感状态达成共识。例如，Tomkins

(1962)认为人的情感包括恐惧、愤怒、痛苦、高兴、厌恶、惊奇、关心和羞愧。Shaver等人(1987)认为情感有6种基本类别,分别是爱、喜悦、惊奇、愤怒、悲伤和恐惧。美国心理学家Ekman(1999)将离散情感与面部表情等属性进行关联,认为离散情感包括生气、厌恶、恐惧、高兴、悲伤和惊讶,这也是目前使用最广泛的离散情感表示方式。

维度情感模型用连续的维度空间来描述情感,它认为情感是高度相关的连续体,能够用多个取值连续的维度进行刻画。因而,情感可以表示为多维空间中的坐标点,不同情感在维度空间中的相对距离能够衡量它们之间的相似性和差异性。在当今情感识别领域,认可度最高的维度模型为"愉悦—唤醒—支配"模型(Mehrabian,1996)。其中,愉悦度也称为效价度(valence),用于描述情感的正负向;唤醒度也称为激活度(activation),用于描述情感的兴奋水平;支配度表示主体对情感状态的主观控制程度,高支配度是一种有力、主宰感,而低支配度是一种退缩、软弱感。

心理学研究表明,离散情感模型和维度情感模型在应用上各有优势和劣势。离散情感简单直观,在当前情感识别研究中应用最多,但是离散情感只能表示有限种类的情感类型,难以完全反映人类复杂、细微的情感状态。维度情感虽然理论上能够反映所有情感状态,但由于维度情感中的坐标值一般不易被人理解和体验,导致标注人员难以正确感知维度情感,容易造成标注结果方差较大。

1.1.2 主流数据集

情感识别的一个关键点在于数据库采集,数据库质量的好坏直接决定了由它训练的模型性能的好坏。随着多模态情感识别技术的发展,学术界和工业界构建了大量数据库。下面将重点介绍目前国际上主流的多模态情感数据集。表1对主流数据集的数据量、发布日期等信息进行了汇总。

表1 情感识别主流数据集

Tabel 1 Main datasets for emotion recognition

数据集	发布年份	数据量	标注方式
IEMOCAP	2008	1 039段对话	离散,维度
CMU-MOSI	2017	2 199段视频	情感强度
CMU-MOSEI	2018	23 453段视频	情感强度
MELD	2019	约13 000段视频	离散

1) IEMOCAP数据库。IEMOCAP(interactive emotional dyadic motion capture database)情感数据集(Busso等,2008)于2008年由美国南加州大学发布,总共包括1 039段对话,总视频时长约为12 h。该数据集由10名专业演员(5男5女),在有台词或即兴的场景下进行录制,参与者在预先定义的情境中表达情感。数据采集内容包含音频、视频和文本信息,以及通过附加传感器收集的姿势信息。数据采集结束后,标注人员对对话进行切分,然后从10种情感状态中选择最合适的标签,包括高兴、快乐、悲伤、愤怒、惊讶、害怕、恶心、沮丧、兴奋和其他情感。

2)CMU-MOSI数据库。CMU-MOSI(Camegie Mellon University multimodal opinion sentiment intensity)数据库(Zadeh等,2017b)由美国卡内基梅隆大学构建,从YouTube中筛选了部分视频构建数据集,视频内容涵盖各种主题。数据筛选过程中确保了每个视频都是单人面对摄像头,以便清晰捕捉面部表情,筛选过程没有对摄像机型号、距离或演讲者场景进行限制。最终,CMU-MOSI数据库包括89位英语演讲者,包括41位女性和48位男性。所有数据进一步划分为2 199个片段,每个片段都进一步标注情感强度,数值范围从强烈的负向情感(-3)到强烈的正向情感(+3)。

3) CMU-MOSEI数据库。CMU-MOSEI(CMU multimodal opinion sentiment and emotion intensity)数据集(Zadeh等,2018a)进一步将CMU-MOSI数据库扩充,总共包含3 228段YouTube视频。这些视频被切分成23 453个片段,涵盖了3种不同模态信息,包括文本、视觉和声音。该数据集包含1 000名演讲者,涵盖250个主题,为研究人员提供了多元化的视角。数据集中的所有视频均为英语,标注人员为每段视频额外提供了情感状态和情感强度的标注。情感状态包括快乐、悲伤、愤怒、害怕、恶心和惊讶,情感强度标记范围从强烈的负面情绪到强烈的正面情绪(-3到3)。

4)MELD数据库。MELD(multimodal emotionlines dataset)数据集(Poria等,2019)源于电视剧《老友记》,包含文本、音频和视频信息。MELD数据集由1 400个视频组成,并进一步切分为13 000个片段。标注过程中总共涉及7种情感类别:愤怒、厌恶、悲伤、喜悦、中性、惊讶和害怕。此外,每个片段额外提供了情绪标签,包括积极、消极和中性。

1.1.3　国际比赛

为了促进情感计算技术发展，国际上围绕多模态情感识别组织了多场比赛，最有名的包括EmotiW（emotion recognition challenges in the wild）比赛和AVEC（audio/visual emotion recognition challenge）比赛。

1）EmotiW 比赛。Dhall 和 Goecke 自 2013 年起连续举办多届 EmotiW 比赛（Dhall 等，2013，2015），比赛任务主要针对多模态离散情感识别。数据集在完全自然的条件下采集，包含背景噪声、大范围头部运动、光照和遮挡等多种干扰因素。随着研究人员的努力，EmotiW 比赛中涉及的 6 种基本情感识别准确率逐年提高。但是，情感识别系统性能仍然难以满足实际应用需求，这反映了自然条件下基本情感识别是一个极具挑战性的问题。除了离散情感识别任务外，EmotiW 比赛也关注情感计算中的其他研究方向，包括注意力分析与群体情感识别等任务（Dhall 等，2016，2017）。

2）AVEC 比赛。Bjorn Schuller、Fabien Ringeval 等人自 2011 年开始举办了音视频维度情感竞赛 AVEC（Schuller 等，2011；Valstar 等，2013）。区别于 EmotiW 离散情感识别比赛，AVEC 比赛侧重于维度情感识别，包含激活度、愉悦度等预测任务。随着 AVEC 比赛的开展，越来越多的研究人员参与到维度情感识别研究中，情感识别也逐渐从基于正负极性的情感分类过渡到了基于连续值的情感回归。每年的 AVEC 赛题都有所侧重，例如在 2013 年，组织者引入了抑郁症检测任务（Valstar 等，2013）。在 2018，组织者引入了情感障碍检测与跨文化情感识别任务（Ringeval 等，2018）。

1.2　多模态情感识别与理解

多模态情感识别与理解，是指从语音、图像和文本等模态信息中提取有效的情感特征，采用特征交互、特征融合等方法，实现对情感状态的准确解析。如图 1 所示，涉及的相关技术分别介绍如下。

1.2.1　情感特征提取

情感特征提取试图从不同的人类情感表达方式中获取到蕴含的情感信息，不同的信息模态具有不同的情感特征获取方式。本节主要从文本、图像和语音 3 个方面综述国际上的情感特征获取技术。

自然语言与情感之间存在着紧密的联系，是人类情感表达的主要媒介，通过词汇、语法和语气等方式，人们能够表达出的情感状态多种多样，包括喜悦、愤怒、悲伤和惊讶等。受到文化和社交背景的影响，语言进行情感表达变得更加多样化。自然语言中的情感探究已成为多个领域的研究焦点。从自然语言中获取情感信息的方法有很多，以下是一些常见的方法和技术，用于提取文本的情感信息，包括：情感词典法、情感特征的向量表示和深度学习方法。

1）情感词典法。早期对于文本情感信息的抽取主要采用的是构建情感词典的方式。情感词典具有优秀的准确性、适用性广泛以及较强的可解释性，同时可根据自身任务进行扩展，使其包含特定领域或任务相关的情感词汇，从而适应不同的情感分析需求。这使得它可以灵活地应对多样化的任务。Esuli 和 Sebastiani（2006）提出 SENTIWORDNET 词典描述了同义词集中包含的术语的客观、积极和消极程度。Mohammad 和 Turney（2013）提出了一个大型的“词—情感”关联词典 NRC Emotion Lexicon，使用 Plutchik（Imbir，2020）的 8 种基本情绪来注释单词。尽管情感词典在情感分析中具有许多优点，但也存在一些限制，如对文本上下文和语义的理解能力较弱，难以处理多义词，以及可能对特定领域的情感词汇不敏感。因此，根据具体情感分析任务的需求，情感词典方法可以与其他技术（如深度学习模型）结合使用，以提高情感分析的准确性和鲁棒性。

2）情感特征的向量表示。情感特征的向量表示是一种将文本中的情感信息以向量的形式表示的技术。这种表示方法的主要目的是将文本中的情感内容嵌入到向量空间中，以便进行情感分析、情感分类或其他自然语言处理任务。Mikoov 等人（2013）提出 Word2Vec。Word2Vec 是一种基于神经网络的词嵌入方法，它使用连续词袋（combining bag of words，CBOW）和跳跃语境（skip-gram）模型来学习单词向量，考虑了单词的上下文关系，使得相似的单词在向量空间中接近。Pennington 等人（2014）提出 GloVe。GloVe 旨在将单词映射到连续的向量空间中，以捕获单词之间的语义关系。GloVe 的关键特点是它使用全局语境统计信息来生成单词向量，强调了单词之间的共现频率和关联。Joulin 等人（2016）提出 FastText。FastText 基于字符级别的 n-grams，是一种快速文本分类和词嵌入方法，可以捕获单词内部的子词信息，具有出色的性能和速度。

3）深度学习方法。基于深度学习方法的情感信

息提取具有多重优势，包括高性能、上下文感知、迁移学习的支持、多模态数据的处理能力以及自动特征学习。深度学习模型在情感分析任务中表现出色，能够更好地捕捉文本中的复杂情感内容，并适用于多种数据类型。Hochreiter和Schmidhuber（1997）提出长短期记忆（long short-term memory，LSTM），广泛用于带有时序信息的特征。随着深度学习和计算机算力的提升，预训练模型大量出现，如Devlin等人（2019）提出BERT（bidirectional encoder representations from Transformers）文本预训练模型。

人的面部表情、肢体动作等都与人类的情感表达有极强的相关性。早期卷积神经网络（convolutional neural network，CNN）系列，如ResNet（He等，2016），相关的预训练模型系列，如CLIP（contrastive language-image pre-training）等也常常用于图像信息的提取（Radford等，2021）。目前，大多数特征都是通过神经网络或公共库来提取的。最常用的公共库包括CERT（computer expression recognition toolbox）（Littlewort等，2011）、FacNet（Schroff等，2015）和OpenFace（Amos等，2016）。

此外，语音蕴含丰富的情感信号，对于全面的情感理解，语音也必不可少。长短期记忆（LSTM）和双向LSTM广泛应用于人工提取的声学特征。目前，大多数多模态情感分析模型使用OpenEAR（open-source emotion and affect recognition）（Eyben等，2009）、openSMILE（open-source large-scale multimedia feature extractor）（Eyben等，2010）、COVAREP（Degottex等，2014）、LibROSA（McFee等，2015）等开源库来提取声学特征。

1.2.2　基于多模态融合的情感识别

不同模式的融合是使用多种模式的情绪分析的中心。多模态融合是一种从各种来源接收到的数据中进行过滤、提取和组合所需特征的过程。然后对这些数据进一步分析评估态度（Gandhi等，2023）。目前的融合技术包括特征级融合、决策级融合和混合融合。

1）特征级融合。特征融合，又称早期融合，是指一种进行单模态特征提取，然后对其各自的模态特征向量进行拼接、位乘或位加法的融合方法（Xiao和Luo，2022）。Zadeh等人（2018a）在每个时间步长中连接来自不同模式的输入，并将其作为单个LSTM的输入。Pham等人（2019）提出基于中期模型的多模态融合方法涉及将多模态数据输入网络，模型的中间层在不同模态之间进行特征融合。特征级融合的好处是，它允许在不同的多模态特征之间的早期相关性，从而更好地完成任务。整合不同的要素是应用这一战略的挑战之一。这种融合方法的缺点是时间同步，因为收集到的特征属于几种模式，并且在许多领域可能有很大的不同（Gandhi等，2023）。

2）决策级融合。决策融合又称为后期融合，首先基于每种模态进行情绪分析，然后提出不同的机制，将单峰情感决策纳入最终决策，包括平均、多数投票、加权和或可学习模型。Yu等人（2017）利用不同模态之间的互补性来提高模型的泛化能力和鲁棒性，采用时间选择注意机制，能够有效地处理多模态情绪识别任务，具有高精度和鲁棒性。Han等人（2021）基于互信最大化来提升不同模态之间的相似度，使用对比学习约束模型学习更优良的表征，最终将不同模态特征拼接融合。Ghorbanali等人（2022）用集成迁移学习方法，提出了一种基于加权卷积神经网络的混合MSA（multimodal sentiment analysis）模型，利用扩展的Dempster-Shafer（Yager）理论，融合文本和图像分类器的输出，在决策层确定最终极性。

3）混合融合。它是早期融合和晚期融合技术的混合体。这种融合方法结合了特征级和决策级的融合技术。研究人员使用混合融合来利用特征级和决策级融合程序的好处，同时避免了它们各自的缺点。Poria等人（2015）使用不同的核函数来表示不同的模态信息，并通过优化目标函数来选择核函数的最优组合来实现多模态信息的融合。Wu等人（2023）提出信息瓶颈模块在模型中间层进行不同模态的信息传递，经过多层的聚合达到最优的融合效果。

在多模态情绪分析中，各模态之间的交互作用往往呈现出间接性和不稳定性，因此建立一个准确、全面的多模态交互模型具有挑战性。交互问题可以从不同的角度分为两类：模态内交互问题和模态间模态交互问题（Zhang等，2023）。Zadeh等人（2017a）采用模态内和模态间的动态建模，提出张量融合网络。Ando等人（2022）使用预训练模型编码器生成的特征与常规启发式特征进行比较，并进行了模态间的交互。

1.2.3　基于大模型的多模态情感识别

大语言模型（large language model，LLM）已经展

示出优秀的建模能力以及出色的语言理解和推理能力，在自然语言处理领域取得了巨大成功（Chowdhery 等，2022；Hoffmann 等，2022）。LLM 可以通过以少量或零样本的方式理解提示来处理各种复杂的任务，因此获得了广泛关注。LLM 可以通过自监督学习或对比学习等方法，学习到更丰富、更有表达力的多模态特征表示。这些特征表示能够捕捉到不同模态之间的关联性和情感信息。Su 等人（2020）提出了一种基于 Transformer 架构的多模态预训练模型 VL-BERT，该模型可以同时处理文本和图像信息，并在多个任务上取得了优异的性能，包括视觉问答和情感分类等。Bao 等人（2020）提出了一种基于伪掩码机制的多模态预训练模型 UniLMv2，该模型在多个任务上取得了优越的性能，包括情感分类和图像描述生成等。

大模型可以同时处理多个与情感相关的任务，例如情感分类、情感生成等。通过多任务学习，模型可以共享底层表示并相互促进，提高情感计算的准确性和鲁棒性。Liu 等人（2016）提出了一种基于共享内存的深度多任务学习方法。该方法通过在网络中引入共享内存模块，实现了不同任务之间的信息共享和交互，从而提高了多任务学习的性能。He 等人（2018）研究了跨领域情感分类的多任务学习方法。该方法通过共享底层特征表示和多个上层任务网络，同时处理不同领域的情感分类任务，实现了更好的泛化性能。

2023 年，LLaVA（Liu 等，2023）和 MiniGPT-4（Zhu 等，2023a）使用“图像—指令”对应数据集开发指令跟随图像—LLM。PandaGPT（Su 等，2023）利用 ImageBind 的多模态编码器（仅在“图像—指令”对上进行训练），使大型模型能够理解 6 种模态。多模态大模型有着更好的音视频理解能力，通过注意力机制、融合网络等方法，将不同模态的信息进行有效地融合，以提高情感计算的性能，受到了学术界广泛关注。

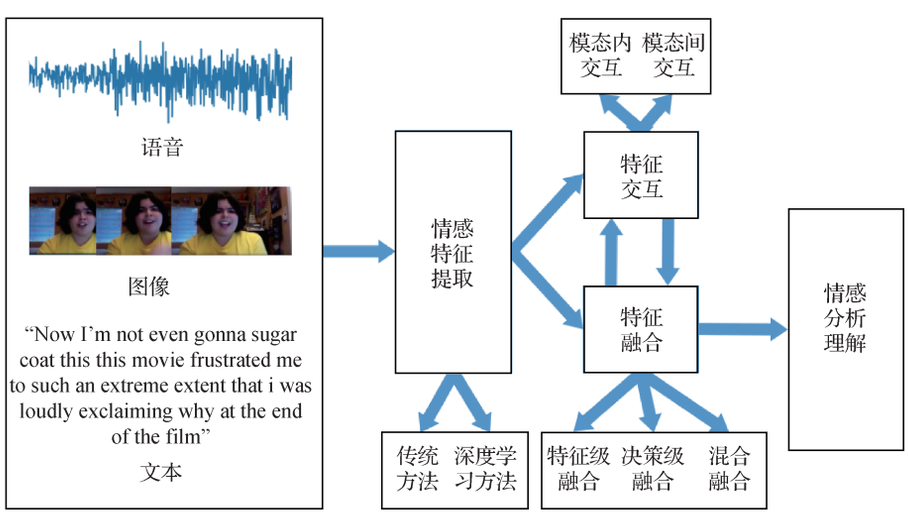

图 1　多模态情感分析与理解流程图

Fig. 1　The diagram of multimodal emotion analysis and understanding

1.3　抑郁症情感障碍检测及干预

随着社会的高速发展，人们面对的压力逐渐增大，内心容易产生抑郁、焦虑等情绪。长期处于抑郁和焦虑状态的人容易患上精神疾病。抑郁症是其中比较常见且危害较大的一种，典型症状包括情绪低落，睡眠质量下降和食欲减退，身体疲倦，注意力不易集中等（World Health Organization，2020a）。在中国有超过 5 400 万人患有抑郁症，占总人口数的 4.2%（World Health Organization，2020b）。抑郁症不仅对个人和家庭造成严重伤害，也给社会带来巨大的经济损失。

研究人员尝试利用算法提取患者的语言表达、面部表情和声音特征并对患者的抑郁状态进行自动评估。他们基于患者的声音、语言和表情提取了多种特征并提出多种基于传统分类器和深度神经网络（deep neural network，DNN）的评估模型。在心理干预领域，聊天机器人技术已被用于治疗焦虑、抑郁和物质使用障碍等心理疾病并且被证实具有一定的治疗效果（Ahmed 等，2021）。结合了专业心理治疗方法（如认知行为疗法（cognitive behavior therapy，

CBT))的人工智能(artificial intelligence, AI)聊天机器人可以对患者进行心理引导和提供专业建议,进而实现对患者的心理治疗。尽管目前AI聊天机器人还不能够取代专业心理咨询师,但它可以作为辅助心理干预的方式,随时随地陪伴用户,及时缓解用户的焦虑和抑郁情绪。近年来,研究人员不断发展互联网认知行为疗法(internet-delivered cognitive behavior therapy, ICBT),提供高质量的线上心理健康服务(Bakker等,2016)。ICBT系统的发展与快速增长的患者需求相符合,约70%的患者对使用移动应用程序进行自我监控和心理健康管理表现出强烈的兴趣(Torous等,2014)。ICBT系统通常采用聊天机器人的方式与用户进行交互。实验表明,ICBT系统对焦虑症和抑郁症的干预结果与专业心理治疗师使用CBT方法的干预结果比较接近(Spek等,2007;Barak等,2008;Andersson和Cuijpers,2009)。

1.3.1 抑郁状态自动评估

抑郁状态自动评估涉及的主要流程如图2所示。Cohn等人(2009)利用面部动作编码系统(facial action coding system, FACS)、主动外观建模(active appearance model, AAM)技术来提取患者表情特征。结合表情特征和声调等声音特征,他们利用一些传统分类器(如支持向量机和逻辑回归分类器)对患者的抑郁程度进行评估,并达到了较高的准确性。此后,基于机器学习的抑郁状态评估问题吸引了越来越多研究人员的兴趣(Cummins等,2015;Joshi等,2013;Scherer等,2014;Morales等,2017)。

该领域的研究人员早期在特征选择和提取方面投入了大量的精力。他们致力于人工分析每个用于临床心理检查的问题,保留那些与患者抑郁情绪高度相关的问题,并将其用于智能代理的心理检查提问。这些问题使得被访谈者的反馈内容更有区分性,使得抑郁状态自动评估算法能够从反馈内容中提取出更有效的特征。Arroll等人(2005)证明在心理检查中使用的一些关键问题能够有效提升抑郁症诊断的准确性。Yang等人(2016)对抑郁症患者心理检查的回答文本进行了手动分析,人工选择了一些与抑郁症紧密相关的问题。他们根据所选择的问题构建了一棵决策树来预测患者的抑郁状态。类似地,Sun等人(2017)分析了抑郁患者心理检查的文字记录,并从某些主题(如睡眠质量、最近的情绪状态、PTSD(post-traumatic stress disorder)等)下的回答中提取文本特征。基于提取的文本特征,他们利用随机森林来评估患者的抑郁程度。Gong和Poellabauer(2017)对心理检查的提问问题按主题进行建模。他们将患者的回答内容剪切成片段并按照不同主题进行归类,从同一类片段中提取音频、视频和语义特征,然后采用特征选择算法,筛选出最具区分性的特征。此外,他们还比较了传统机器学习算法(如随机森林和支持向量机)在抑郁状态评估问题中的性能。Williamson等人(2016)构建了一系列的背景指标,包括患者是否经过抑郁症诊断、是否进行过医疗/心理治疗等,然后使用Gaussian Staircase模型对患者的抑郁状态进行评估并取得了较好的结果。

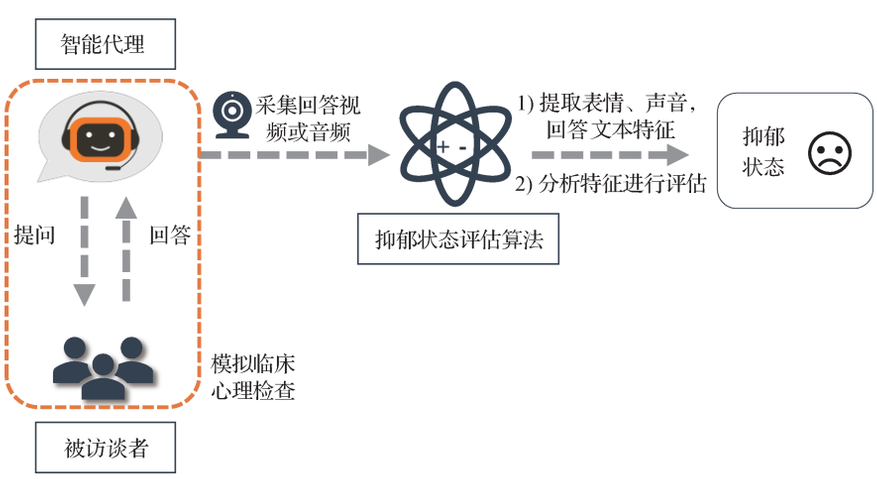

图2 抑郁状态自动评估流程示例

Fig. 2 An example of automated depression state assessment

受到深度学习技术的启发，Mendels 等人（2017）利用声学特征和文本特征联合训练深度模型，并将训练好的模型用于抑郁状态评估。他们发现，使用深度网络并融合来自多个模态（如声音、文本）的特征可以达到更高的评估准确度。Yang 等人（2017）提出了一个基于深度卷积神经网络（deep convolutional neural network，DCNN）的抑郁症检测模型。为了训练该模型，他们还额外设计了一组新的视频和音频特征用于模型输入。Alhanai 等人（2018）使用长短期记忆（LSTM）网络来评估患者的抑郁程度。他们通过计算皮尔逊相关系数来选择与抑郁程度密切相关的音频特征和文本特征。针对患者谈话样本，Lam 等人（2019）提出了一种基于主题建模的多模态数据增强方法。他们首先对每段心理检查视频和音频标记主题标签。针对每个患者，他们提取同主题标签下的部分视频/音频数据，并合并成一个新的训练样本；然后使用 Transformer 对文本内容进行特征建模，利用深度一维卷积神经网络（1D CNN）进行音频特征建模。实验结果表明，两种模型的结合使得抑郁状态评估的准确率得到了进一步提升。Ma 等人（2016）使用 LSTM 和卷积神经网络（CNN）提取抑郁症患者的声音特征并用于抑郁状态评估。另外，为了解决抑郁数据集中普遍存在的正负样本不平衡问题，他们对患者心理检查的回答音频进行了随机过采样以增加抑郁类别的样本数量。Haque 等人（2018）提出了一个因果卷积网络（Causal CNN）模型。该模型将音频特征、视频特征和语言特征编码为多模态嵌入，并用该嵌入预测患者的抑郁状态。Dinkel 等人（2019）提出了一种基于文本数据的多任务双向长短时记忆网络（bidirectional LSTM，BiLSTM）评估模型。该模型将预训练的词嵌入作为输入特征。另外该模型使用了一个新提出的损失函数，能够同时判断对象是否患有抑郁症以及评估患者抑郁程度。

1.3.2 面向抑郁人群的智能心理干预系统

一般的心理干预系统如图 3 所示，主要通过持续进行的对话和用户进行交互，帮助用户缓解焦虑或者抑郁情绪。随着计算机技术的发展，利用软件系统对用户进行心理干预的方式越来越常见。在精神治疗领域，SimCoach（Rizzo 等，2011）利用虚拟代理和用户进行互动，旨在提升那些不愿意与传统咨询师交谈的用户的参与程度，让用户增强对自身精神症状和治疗方法的认识。Pasikowska 等人（2013）提出另一种解决方案，即利用网络摄像头向用户提问，进而评估用户的压力和焦虑程度，帮助他们应对压力并减少焦虑和愤怒情绪。

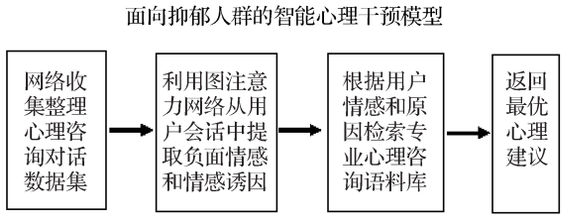

图 3 智能心理干预模型示例

Fig. 3 An example of intelligent psychological intervention model

聊天机器人是精神治疗领域最常用的干预技术。1964 年 ELIZA 由美国麻省理工学院人工智能实验室开发，用于模拟参与初次心理访谈的心理治疗师的行为（Weizenbaum，1966）。尽管 ELIZA 是一个简单的基于规则式的聊天机器人，但是它很好地完成了心理治疗任务，有效缓解了用户的压力。Bickmore 等人（2010）设计了一个带有虚拟形象的共情聊天机器人用于与抑郁患者的交流。该聊天机器人可以通过文字及声音向病人介绍治疗方案及注意事项。实验结果表明患者对该系统非常满意。Shim（Ly 等，2017）是一个基于文本的聊天机器人，通过结合 CBT 方法有效缓解了用户焦虑的情绪。Gardiner 等人（2017）利用聊天机器人向参与人员提供生活方式相关的建议。在实验过程中，参与者有效地降低了酗酒量，增加水果摄入量，减少了压力情绪。另外，Bhakta 等人（2014）发现用户认为向聊天机器人披露敏感信息是“安全的”。

用于抑郁症治疗的聊天机器人交互示例如图 4 所示，最有代表性的应用是 Woebot（Fitzpatrick 等，2017）和 Wysa（Inkster 等，2018）。Woebot 通过简短的日常对话和情绪跟踪的形式为用户提供认知行为治疗，改善用户抑郁和焦虑情绪。Wysa 则综合了认知行为治疗、行为强化和正念等多种心理疗法来帮助抑郁症患者。

1.3.3 多模态抑郁数据集

对于抑郁状态自动评估领域来说，大规模高质量的多模态抑郁数据集是必不可少的。通常来说，多模态抑郁数据集包含了抑郁患者和作为对比的健

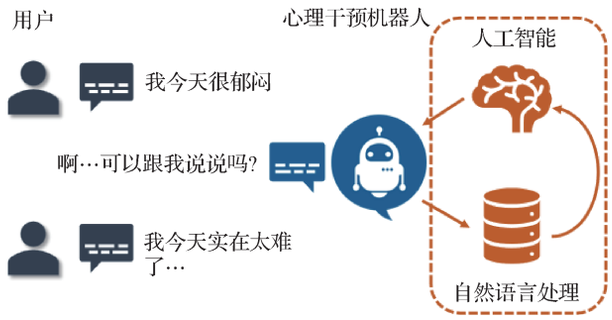

图4　用户和心理干预机器人交互示例

Fig. 4　An example of user interaction with a psychological intervention robot

康人在心理检查中的音频/视频和相应的文字记录。DAIC-WoZ(Gratch 等,2014)和 AViD-Corpus(Valstar 等,2013)是多模态抑郁状态评估领域最常用的两个数据集。其他非公开抑郁情绪数据集包括黑狗数据集(black dog dataset)(Alghowinem 等,2013)、虚拟人类压力评估访谈语料库(virtual human distress assessment interview corpus)(Scherer 等,2013)和基于汉密尔顿评分量表的抑郁症数据集(Hamilton rating scale for depression corpus)(Yang 等,2013)。

DAIC-WoZ 数据集的创建目的是为了支持心理障碍(包括焦虑、抑郁和创伤后应激障碍)自动诊断技术。心理访谈由一个名为 Ellie 的虚拟访谈者主持,两名场外工作人员控制 Ellie 的非语言行为(如 Ellie 的点头和面部表情)和语言行为。在心理访谈中,志愿者独自坐在一个房间里与 Ellie 进行互动。Ellie 会询问一些提前设计好的问题,例如“谁是对你的生活有积极影响的人?”和“你在生活中最自豪的是什么?”。如果志愿者的回答太短,Ellie 会用一些预设的句子(例如“这有多难?”)来引导他说更多的话。研究人员共招募了 142 名志愿者,每个人的访谈时长在 8～30 min 之间。每次访谈收集的数据包括录音、录像、文字记录以及患者心理健康问卷(patient health questionnaire for depression,PHQ-8)的得分和回答。PHQ-8 是一种在大型临床研究中常用的抑郁程度测量问卷(Gilbody 等,2007)。PHQ-8 问卷的得分反映了被访谈者的抑郁严重程度。

AViD-Corpus 数据集包含了 84 名抑郁症和非抑郁症志愿者的音频和视频数据。志愿者需要完成两个交互任务,整个过程由摄像头和麦克风记录。第 1 个任务是朗诵任务,志愿者需要朗读德国寓言《北风和太阳》的节选片段。第 2 个任务是问答任务,志愿者需要回答电脑屏幕上显示的问题(例如“您最喜欢的菜是什么”,“您最好的礼物是什么,为什么”,“讨论一个悲伤的童年记忆”等)。朗诵或回答的音频由连接到笔记本电脑内置声卡的耳机录制,视频录制的帧率和分辨率分别为 30 帧/s 和 640 × 480 像素。每个志愿者还需要填写 BDI-Ⅱ(beck depression inventory-Ⅱ)问卷(Rush 等,2006)。与 PHQ-8 问卷类似,BDI-Ⅱ问卷用于度量志愿者的抑郁严重程度,共有 21 个描述项目,得分在 0～63 之间。若志愿者的 BDI-Ⅱ得分在 20～28 分之间,则表明该志愿者可能具有中度抑郁症状。如果志愿者的 BDI-Ⅱ得分超过 29 分,则表明其可能患有重度抑郁症。

黑狗数据集是由黑狗研究所收集的视听语料库。黑狗数据集的数据收集过程包含两部分,分别为句子阅读任务和心理访谈任务。在心理访谈任务中,医生发起提问,参与者对该问题进行回答与描述。阅读任务需要参与者阅读 20 个具有消极或积极意义的句子。黑狗研究所在获得参与人员的知情同意书后,收集了 40 多名抑郁志愿者和 40 名健康志愿者参与心理访谈和阅读任务的音频和视频。这些音频和视频记录了志愿者的言语、面部表情和肢体语言。

虚拟人类压力评估访谈语料库的数据收集方式与 DAIC-WoZ 类似,也是利用一个虚拟代理与参与人员进行互动。参与者与虚拟代理的互动流程如下:首先,虚拟代理解释该访谈的目的,并提出一系列问题,与参与者建立融洽的关系。然后,虚拟代理开始进行访谈。整个访谈分为 3 个阶段,即积极阶段、消极阶段和第 2 个积极阶段。在积极阶段,访谈者会询问一些引发积极情绪的问题,如“你认为你最优秀的品质是什么?”。在消极阶段,访谈者会询问一些包含消极情绪的问题,如“你有不安的想法吗?”。每个参与者的访谈音频都会被记录并被收集起来。该数据集共包含 43 名参与者的访谈记录。另外,参与者需要完成若干问卷调查,包括 PTSD 检查表民用版(PTSD CheckList —— civilian version,PCL-C)(Ruggiero 等,2003)和患者健康问卷(PHQ-9)(Kroenke 等,2001)的抑郁模块。PCL-C 和 PHQ-9 量表都被用于评估参与者的精神状况。

抑郁症汉密尔顿评分量表语料库是美国匹兹堡大学研究人员在治疗抑郁症患者期间收集的一个经过临床验证的抑郁数据集。数据集中共有 57 名符

合 DSM-IV（diagnostic and statistical manual）重度抑郁症标准（Bell，1994）的参与者。在抑郁症治疗过程中，11 名专业的临床心理医生会在第 1、7、13 和 21 周对每位患者进行访谈，然后评估患者的抑郁程度。医生采用汉密尔顿抑郁症评分量表（Hamilton rating scale for depression，HRSD）评估患者的抑郁程度。HRSD 得分为 15 分及以上表示患者有中度或重度抑郁症，得分为 7 分及以下表示精神处于正常的状态（Fournier 等，2010）。该语料库的数据由 4 台摄像机和 2 个麦克风采集得到，共包含了 49 名参与者的 130 个疗程的访谈录音和录像。

2 国内研究进展

2.1 多模态情感识别概述

2.1.1 学术定义

人类在认识客观实际时，会表现出喜、怒、哀、乐等各种主观体验，这些主观体验复杂繁多，很难明确地定义其客观规范。如何界定情感状态是一个颇具争议的问题。目前，学术界关于情感的表示并没有完全统一的认识，也没有一个定性和定量的评价标准，主要采用离散模型或维度模型表示情感。考虑到情感本身具有的模糊性（即不同标注人员针对相同数据可能会感受到不同的情感），明确的离散或维度标签难以准确刻画模糊的情感状态。针对上述问题，Lian 等人（2023c）提出了一种新的情感表示方法，他们采用带有推理属性的情感描述去刻画情感状态，推理过程的合理性将作为模型的评价指标。他们试图利用这种情感表示方法，减弱情感标注过程中的模糊性和歧义性，促进情感计算技术的应用。

2.1.2 中文数据集

为了促进中文情感识别技术的发展，国内学者提出了多个围绕中文场景的多模态情感数据集，比较有代表性的是 CH-SIMS（Chinese multimodal sentiment analysis dataset with fine-grained annotations of modality）数据集、M3ED（multi-modal multi-scene multi-label emotional dialogue database）数据集以及 MER（multimodal emotion recognition ）数据集。表 2 对中文数据集的数据量、发布日期等信息进行了汇总。相比于现有数据集，MER2023 额外提供了大量未标注样本用于评测半监督学习算法的性能。

1）CH-SIMS 数据集。CH-SIMS 数据集（Yu 等，2020）总共包含 2 281 个视频片段，每个片段只包含一个角色。该数据集专门针对中文语言，涵盖了多种场景，并且每个模态都有各自的情感标注，便于研究人员分析模态互斥或模态互补问题。CH-SIMS 数据集利用情感强度作为标签，强度范围从负向情绪到正向情绪（−1～1）。此外，该数据集进一步提供了年龄和性别等属性的标注，便于研究人员分析人物属性和情感状态之间的关联性。

2）M3ED 数据集。M3ED 数据集（Zhao 等，2022）是一个多模态、多场景和多标签的情感对话数据集。在对话中，角色的情感状态受到诸多因素影响，包括对话场景、对话主题和角色交互模式等。针对多模态对话情感分析任务，M3ED 数据集提供了 56 部电视剧中的 9 082 轮对话。所有对话切分为 24 449 个片段，每个片段提供了 3 种模态信息（声音、视觉和文本）。此外，标注人员从 7 种情感类别（快乐、惊讶、悲伤、厌恶、愤怒、害怕和中性）中选择了一个或多个标签，它对于跨文化情感分析和识别具有重要价值。

3）MER 数据集。MER 数据集（Lian 等，2023b）面向中文语言，侧重于评价多模态情感识别系统的鲁棒性。MER 数据集包含 3 部分数据，MER-MULTI、MER-NOISE 和 MER-SEMI。其中，MER-MULTI 数据集提供了离散标签和维度标签，便于研究人员分析多标签相关性对于情感识别性能的影响；MER-NOISE 数据集提供了语音噪声和视频模糊处理后的数据，便于研究人员评测情感识别模型在干扰环境下的鲁棒性；MER-SEMI 提供了大规模无标记视频样本，便于研究人员评测无监督学习或者弱监督学习对于多模态情感识别任务的作用。

表 2 主要中文情感数据集

Table 2 The Chinese emotion datasets

数据集	发布年份	数据量	标注方式
CH-SIMS	2020	2 281 个视频	情感强度
M3ED	2022	9 082 轮对话	离散
MER2023	2023	5 030 标注+73 148 未标注	离散，维度

2.1.3 组织的国际比赛

2016 年，中国科学院自动化研究所举办了首届多模态情感识别比赛（Li 等，2016），包括音频情感识

别、表情识别和多模态情感识别3个子任务。自2018年起，中国科学院心理研究所联合芬兰奥卢大学等国内外机构连续多年举办微表情识别挑战赛，促进了微表情技术的发展(Yap等，2018)。2023年，中国科学院自动化研究所联合清华大学、英国帝国理工学院、芬兰奥卢大学和新加坡南洋理工学院，在国际会议ACM Multimedia上组织了首届MER 2023挑战赛(Lian等，2023b)，主要针对多模态情感识别中的系统鲁棒性问题，并组织了3个子赛道，用于评价干扰环境下的系统鲁棒性以及多标签相关性和无标注样本对于系统鲁棒性的影响。

2.2 多模态情感识别与理解

2.2.1 情感特征提取

国内关于情感特征提取方法蓬勃发展。对于文本，徐琳宏等人(2008)提出中文情感词汇本体库，从不同角度描述一个中文词汇或者短语，包括词语词性种类、情感类别、情感强度及极性等信息。Ku和Chen(2007)提出中文情感极性词典(National Taiwan University Semantic Dictionary，NTUSD)。该词典为简体的情感极性词典，共包含2 812个正向情感词和8 278个负向情感词，可以用于二元情感分类任务当中。视觉特征方面，可从视频中提取说话人情绪相关的面部表情特征和身体姿势(Zhu等，2023b)。

2.2.2 基于多模态融合的情感识别

多模态数据从不同的角度描述对象，比单模态数据提供更多的信息。来自不同模式的数据信息可以相互补充(Lai等，2023)。国内对于多模态融合的探索非常丰富。中国科学院自动化研究所Lian等人(2021)提出了一套基于Transformer结构的细粒度多模态信息融合方法。区别于传统意义上将多模态特征压缩到句子级别再进行融合的策略，他们直接利用细粒度特性信息，在模型内部实现了跨模态异构特征动态对齐与融合，避免了特征压缩造成的信息丢失问题。哈尔滨工业大学Hu等人(2022)提出了一个多模态情感知识共享框架(unified multimodal sentiment analysis，UniMSE)，该框架将来自特征、标签和MSA(multimodal sentiment analysis)及ERC(emotion regulation checklist)任务统一起来。在句法和语义层面进行情态融合，并引入情态和样本之间的对比学习，以更好地捕捉情感与情绪之间的差异和一致性。清华大学计算机科学与技术系智能技术与系统国家重点实验室Yu等人(2021)提出一个基于自监督学习策略的标签生成模块来获取独立的单峰监督。然后对多模态任务和单模态任务进行联合训练，分别学习一致性和差异性。哈尔滨工业大学Wu等人(2022)提出情感词感知多模态改进模型(sentiment word aware multimodal，SWRM)，该模型可以利用多模态情感线索动态地改进错误的情感词。宁波大学刘婷婷等人(2021)从表情、语音、姿态、生理信号和文本信息等多通道信息分析用户的情绪状态，归纳了情绪识别中的一些机器学习方法，展望了在智能体应用中的多模态情感计算的研究方向。

国内对于多模态交互有更丰富的探索。中山大学计算机科学与工程学院Yang等人(2022)提出了一个多模态框架——两阶段多任务情感分析(two-phase multi-task sentiment analysis，TPMSA)。它采用两阶段训练策略来充分利用预训练模型，并采用了一种新的多任务学习策略来研究每个表征的分类能力，同时进行模态内和模态间的交互。河北大学数学与信息科学学院Zhang等人(2022)提出模拟情感连贯性的方法，提出的DEAN(deep emotional arousal network)模型由3个组成部分组成，即跨模态Transformer模拟人类感知分析系统；多模态开发系统模拟认知比较器；引入多模态门控块模拟人类情绪唤醒模型的激活机制，同时进行模态间和模态内的交互。中国科学院自动化研究所Lian等人(2023a)将多模态交互与半监督学习相结合。该方法利用了无情感标注的多模态数据，采用模态之间的翻译任务作为预训练任务，学习模态交互模式，然后在下游情感识别任务上微调，获取更鲁棒的情感特征表示。西安电子科技大学计算机科学与技术学院Wang等人(2023)提出了一种新的文本增强融合网络(text enhanced Transformer fusion network，TETFN)方法，该方法学习面向文本的成对跨模态映射，以获得有效的统一多模态表示，在中间进行了模态间的交互。

2.2.3 基于大模型的多模态情感识别

中国科学院自动化研究所Lian等人(2023b)提出了基于多模态线索推理的情感数据集，借助多模态线索将模糊的情感感知过程明确化，提升了情感标签的可靠度，并利用多模态大模型结合该任务进行微调，进一步提升了情感预测结果的可解释性与可靠性。华中科技大学Lei等人(2023)提出了InstructERC，率先将传统的判别式框架的ERC模型

转变为与 LLM 结合的生成式的模型框架，在多个数据集上取得 SOTA（state of the art）效果。安徽大学 Yi 等人（2023）采用大模型进行不同模态情感特征提取，并通过选择最优融合参数实现最佳识别结果，在 Muse 2023 取得赛道 1 冠军。新疆大学 Hu 等人（2023）提出了一个单独频谱模型和一个结合了大模型的情感识别联合网络，通过设计不同的交互注意力模块将两个中间特征进行融合，并设计多分支训练策略对联合网络进行优化，从大模型和基于频谱的模型得到共性特征和特性特征，取得了良好的效果。

2.3 抑郁症情感障碍检测及干预

如前所述，抑郁症是最常见的情感障碍，其早期筛查和干预是亟需解决的问题。2020 年国家卫生健康委办公厅发布的《探索抑郁症防治特色服务工作方案》将开展抑郁症早期筛查和心理干预列为重点任务（《国家卫生健康委办公厅关于探索开展抑郁症、老年痴呆防治特色服务工作的通知》）。随着人工智能领域的发展，国内在结合人工智能技术解析抑郁状态评估和心理干预方面也开展了广泛的研究。

2.3.1 抑郁状态自动评估

在国内，研究者和机构已经积极投入这一领域，不断推动相关研究的发展。早在 2016 年，北京航空航天大学的智能识别与图像处理实验室（Ma 等，2016）就参加了音频/视频情感识别挑战赛（AVEC'16）的抑郁分类（depression classification sub-challenge，DCC）子赛，提出了 DepAudioNet 算法用于解决该任务。DepAudioNet 利用 CNN 和 LSTM 深度网络提取代表性的音频特征用于抑郁状态的分类。2019 年中国科技大学（Yin 等，2019）参加了 AVEC'19 挑战赛的抑郁检测子赛。团队利用深度神经网络从音频样本中提取了音频、视频和文本 3 种特征，将 3 种特征拼接后再利用 BiLSTM 网络预测抑郁状态的严重程度。随后齐鲁工业大学 Ye 等人（2021）设计了一组实验，让参与人员朗读带有不同情感的文本，记录参与者的朗读音频。通过分析参与者的音频特征变化来对抑郁程度进行预测。上海交通大学 X—跨媒体语言智能实验室 Zhang 等人（2021）利用自监督学习技术学习域内域外的音频特征嵌入，然后利用 BiLSTM 网络预测音频样本的抑郁程度。同济大学 Lin 等人（2020）利用 1D CNN 和 BiLSTM 分别提取声音和文本嵌入，利用融合的嵌入特征和全连接网络对样本抑郁状态进行分类。为了解决抑郁训练样本不足的问题，合肥工业大学 Guo 等人（2024）利用提示词按照不同主题向参与者提问，通过总结参与者不同答复的情感倾向对参与者进行抑郁状态分类。东北师范大学 Fang 等人（2023）利用多层次注意力机制融合音频、视频和文本特征，使得预测准确性进一步提升。除了利用常规的音频、识别和文本特征，兰州大学尝试利用运动设备采集的信号来预测参与者的抑郁状态，通过分析参与者走路的动能和势能信息来预测参与者的抑郁状态（Sun 等，2020）。

此外，产业界也开始对抑郁状态的自动评估表现出兴趣。一些公司已经推出了基于智能手机和可穿戴设备的应用程序（如闻心在身边 App），通过监测用户的身体状况来预测用户的情感状态。

2.3.2 面向抑郁人群的智能心理干预系统

黄智生等人（2019）建立了一个基于语义数据处理平台的网络机器人，该系统每天从微博中抓取数据并提取出自杀信息，然后自动发送给心理辅导团队以提供相应的救援服务。北京理工大学研发的“便携式三导脑电抑郁状态采集与分析系统”能够持续性检测抑郁障碍人群，为抑郁状态的快速、精准识别提供了科学依据。深圳镜象科技和华东师范大学合作，上线了一款名为“EMO-GPT”的情感陪伴聊天机器人，可以随时随地为用户提供倾诉陪伴服务，帮助他们处理焦虑，应对压力，实现心理健康陪伴。从 2021 年开始，各种针对心理健康问题的筛查、预防的创业公司纷纷涌现，包括清华大学的聆心智能、中国科学院自动化研究所的中科智极等。但是，目前成熟的可直接用于心理障碍评估和干预的中文系统还不多见。

2.3.3 多模态抑郁数据集

国内的研究者已经开始积极构建多模态抑郁数据集，以促进抑郁状态的跨领域研究。兰州大学团队（Guo 等，2021）采集了 104 名抑郁患者的视频样本来研究多模态特征（即音频和视频特征）对于抑郁症检测的影响。作为对比，团队还采集了 104 名健康人的视频样本，从而构造了一个中文多模态抑郁数据集。出于保护隐私的目的，该数据集并未公开。兰州大学团队（Cai 等，2022）又发布了一个包含音频和 EEG（electroencephalogram）信号的中文多模态数据集。该数据集的样本来自 52 名志愿者的访谈、阅读和图像描述记录，其中 23 人确诊患有抑郁症。同济大学发布了首个公开的中文多模态抑郁数据集

EATD-Corpus(emotional audio-textual depression corpus)(Shen等,2022)。该数据集包含了162位志愿者的心理访谈问答音频及文本,其中共有74名男性和88名女性。162名志愿者中有30人达到抑郁症标准,其余132人处于健康状态。音频时长共计2小时15分钟,每个问题的回答音频平均时长为15 s。北京科技大学(Zou等,2023)提出了一个公开的中文多模态抑郁数据集CMDC(Chinese multimodal depression corpus)。该数据集包含78名志愿者参与诊断过程的视频样本,其中26人确诊为抑郁患者,52人为健康的对照样本。基于视频样本,团队提取了文本、音频、图像3种类型的特征。

3 国内外研究进展比较

国外在多模态情感理解领域较早开展了一系列基础性研究工作,包括离散和维度情感表示方法、多模态情感数据集构建,并较早组织了情感识别比赛。近年来,国内面向中文语言,开展了大量针对多模态情感理解方面的工作,并且构造了中文情感数据集。因此,从整体上看,国内外研究进展趋同。此外,国内针对中文语言的工作,有助于在国际上推进跨文化情感识别研究。国内提出了基于多模态线索描述的情感表示方法,相比于国际上常用的维度情感表示方法和离散情感表示方法,有希望解决情感固有的模糊性对于现有技术的制约。

3.1 多模态情感识别与理解

国际研究在情感特征提取方面广泛采用深度学习方法。深度神经网络(DNN)和卷积神经网络(CNN)等技术已经在情感识别和情感分析中取得显著的进展。国内的情感特征提取研究还关注了文化因素对情感表达的影响。研究人员经常关注中文情感表达和文化差异,以更好地理解和解释情感。情感特征提取研究通常涉及跨学科合作。研究者将计算机科学、心理学、语言学和社会科学等领域的知识融合在一起,以更全面地研究情感的特征提取和分析。中国和国际研究在多模态融合领域都取得了显著进展,国际研究强调深度学习、多模态数据集和虚拟现实,而中国研究侧重多样的应用领域、文化多样性和跨学科合作。两者的合作和知识共享将有助于推动多模态融合技术的不断发展,以实现更多领域的创新应用。

国际研究着重于多模态表示学习,旨在将不同模态的数据映射到共享的表示空间,以更好地理解它们之间的关联。国内应用领域多样化:中国的研究者在多模态融合中广泛应用,包括虚拟现实、医疗保健、智能交互和情感分析等领域。多模态融合技术在中国的应用领域多元化。文化和语言多样化:中国的研究通常考虑文化和语言多样性,尤其是在处理中文多模态数据时,这包括了解不同地区和文化之间的差异。两者的合作和知识共享将有助于推动多模态融合技术的不断发展,以实现更多领域的创新应用。

3.2 抑郁症情感障碍检测及干预

在抑郁状态自动评估领域,国内研究人员早在2016年就参加了AVEC挑战赛并且取得了不错的成绩。基于AVEC'16挑战赛给出的DAIC抑郁检测数据集,国内此后相继提出各种基于音频、视频、文本和EEG信号特征的抑郁检测模型。总体来说,抑郁状态自动评估问题无论在国内还是国外都处于探索阶段,研究人员都在努力尝试提取更有效的抑郁相关的特征,设计出更有针对性的检测模型。在本领域国内外的技术较为接近。在多模态抑郁数据集领域,国内近年来发展迅速,相继提出了若干内容为中文的公开多模态抑郁数据集,如EATD-Corpus、CMDC等。对比国外的成果,目前能获取到的仅有一个英文多模态抑郁数据集,即DAIC数据集。虽然之前发表的国外工作中提出了很多英文抑郁数据集,但出于隐私保护,这些数据集都未能公开,仅仅在论文中给出关于数据集的部分统计结果。然而缺乏公开数据集将导致下游分析研究难以开展。因此,国内涌现的中文多模态抑郁数据集将极大地促进针对国内患者的自动抑郁检测的研究。在智能心理干预领域,国外起步较早。如已经上线应用商店的Woebot和Wysa两个App,它们分别于2017年和2018年被设计实现,并用于实际心理干预中。但是目前国内并无类似的成熟产品。尽管国内已有许多具有应用潜力的研究,但这些研究成果尚未转化为商业产品和服务。国内应用情感计算技术的行业发展相对较慢,与国际领先水平相比,市场应用有限。

4 发展趋势与展望

多模态情感识别的主要挑战在于数据稀缺性,

即没有足够数量的数据来建立和探索复杂的深度学习模型，使得深度神经网络方法难以创建泛化能力较强的多模态情感识别模型。针对上述问题，一方面需要构建大规模多模态情感数据库；另一方面需要探索基于大模型的迁移学习方法，将无监督任务或者其他任务学习到的知识迁移至情感识别任务中，缓解数据资源匮乏的问题。情感本身具有模糊性，采用明确的离散和维度标签来表示模糊的情感状态存在着局限性。未来，需要增强预测结果的可解释性，提升识别结果的可靠度。

4.1 多模态情感识别与理解

未来情感特征提取领域展现出广阔的前景和潜力。首先，多模态情感分析将成为一个重要的研究方向，将结合文本、图像、音频和其他感知模态的信息，以更全面地理解情感表达。深度学习方法将继续引领该领域的发展，新的架构和模型将不断涌现，以更准确地捕捉和表达情感信息。自动特征工程技术将减轻研究人员的负担，提高情感特征提取的效率和可扩展性。跨语言情感分析将成为一个重要的领域，满足全球化的需求。情感识别将越来越个性化，考虑到个体差异，以更好地满足个体的情感需求和反应。情感识别也将在人机交互领域扮演更重要的角色，使计算机系统更好地理解和响应用户的情感需求，实现从虚拟助手到情感驱动的产品设计。情感交互技术将持续引领科技创新，实现更自然、沉浸式的用户体验。这将包括利用深度学习和人工智能技术，以更好地理解用户的意图和情感，从而提供高度个性化的交互反馈。情感交互技术还将跨越各种平台和设备，让用户能够在智能手机、平板电脑、智能家居设备、虚拟现实头盔和增强现实眼镜等不同设备上实现无缝的交互。此外，多模态数据的融合将成为一项重要趋势，它将整合文本、图像、音频和其他感知模态的数据，以提供更全面的信息和服务。伦理和隐私问题将成为重要关注点，需要仔细考虑数据使用和隐私保护的问题。综上所述，多模态情感识别与理解的未来发展将在多个领域取得重大突破，有望为社会带来深刻的变革和创新。

4.2 情感障碍检测及干预

多模态情感计算在解决抑郁、焦虑等情感障碍方面的作用日益凸显。未来的研究可以从以下 3 方面开展。首先是多模态情感障碍数据集的研究和构建。多模态情感障碍数据集为情感障碍的自动识别提供了坚实的基础。然而，这一领域还需要面对数据隐私和伦理等挑战。除此之外，如何设计有针对性的访谈问题、如何在数据采集时不伤害患者、如何通过算法进行样本扩增都是值得考虑的问题。其次是基于多模态情感障碍识别算法。情感障碍属于心理范畴，但是情感障碍会影响患者的生理特征，如声音、躯体动作等。这种心理—生理的关联性值得深入探索。在此基础上，如何提升多模态情感障碍识别算法的准确率是亟需研究的问题。最后是智能心理干预系统的设计与实现。如何有效模拟心理咨询师的咨询过程、如何及时接收用户情感反馈、如何生成共情会话等都是需要进一步探索的问题。

致　谢　本文由中国图象图形学学会人机交互专业委员会组织撰写，该专委会链接为 https://www.csig.org.cn/16/201708/49327.html。

参考文献（References）

Ahmed A, Ali N, Aziz S, Abd-Alrazaq A A, Hassan A, Khalifa M, Elhusein B, Ahmed M, Ahmed M A S and Househ M. 2021. A review of mobile chatbot apps for anxiety and depression and their self-care features. Computer Methods and Programs in Biomedicine Update, 1: #3100012 [DOI: 10.1016/j.cmpbup.2021.100012]

Alghowinem S, Goecke R, Wagner M, Epps J, Gedeon T, Breakspear M and Parker G. 2013. A comparative study of different classifiers for detecting depression from spontaneous speech//Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver: Canada: IEEE: 8022-8026 [DOI: 10.1109/ICASSP.2013.6639227]

Alhanai T, Ghassemi M and Glass J. 2018. Detecting depression with audio/text sequence modeling of interviews//Interspeech 2018. Hyderabad, India: [s. n.]: 1716-1720 [DOI: 10.21437/Interspeech.2018-2522]

Amos B, Ludwiczuk B and Satyanarayanan M. 2016. OpenFace: a general-purpose face recognition library with mobile applications. CMU School of Computer Science, 6(2): #20

Andersson G and Cuijpers P. 2009. Internet-based and other computerized psychological treatments for adult depression: a meta-analysis. Cognitive Behaviour Therapy, 38(4): 196-205 [DOI: 10.1080/16506070903318960]

Ando A, Masumura R, Takashima A, Suzuki S, Makishima N, Suzuki K, Moriya K, Ashihara T and Sato H. 2022. On the use of modality-specific large-scale pre-trained encoders for multimodal sentiment analysis//Proceedings of 2022 IEEE Spoken Language Technology Workshop (SLT). Doha, Qatar: IEEE: 739-746 [DOI: 10.1109/SLT54892.2023.10022548]

Arroll B, Smith F G, Kerse N, Fishman T and Gunn J. 2005. Effect of the addition of a 'help' question to two screening questions on specificity for diagnosis of depression in general practice: diagnostic validity study. BMJ, 331(7521): #884 [DOI: 10.1136/bmj.38607.464537.7C]

Bakker D, Kazantzis N, Rickwood D and Rickard N. 2016. Mental health smartphone apps: review and evidence-based recommendations for future developments. JMIR Mental Health, 3(1): #4984 [DOI: 10.2196/mental.4984]

Bao H B, Dong L, Wei F R, Wang W H, Yang N, Liu X D, Wang Y, Piao S H, Gao J F, Zhou M and Hon H W. 2020. UniLMv2: pseudo-masked language models for unified language model pre-training//Proceedings of the 37th International Conference on Machine Learning. [s.l.]: JMLR.org: 642-652

Barak A, Hen L, Boniel-Nissim M and Shapira N. 2008. A comprehensive review and a meta-analysis of the effectiveness of internet-based psychotherapeutic interventions. Journal of Technology in Human Services, 26(2/4): 109-160 [DOI: 10.1080/15228830802094429]

Bell C C. 1994. DSM-IV: diagnostic and statistical manual of mental disorders. JAMA, 272(10): 828-829 [DOI: 10.1001/jama.1994.03520100096046]

Bhakta R, Savin-Baden M and Tombs G. 2014. Sharing secrets with robots?//Proceedings of 2014 World Conference on Educational Multimedia, Hypermedia and Telecommunications. Chesapeake, VA, USA: Association for the Advancement of Computing in Education (AACE): 2295-2301

Bickmore T W, Mitchell S E, Jack B W, Paasche-Orlow M K, Pfeifer L M and Odonnell J. 2010. Response to a relational agent by hospital patients with depressive symptoms. Interacting with Computers, 22(4): 289-298 [DOI: 10.1016/j.intcom.2009.12.001]

Busso C, Bulut M, Lee C C, Kazemzadeh A, Mower E, Kim S, Chang J N, Lee S and Narayanan S N. 2008. IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4): 335-359 [DOI: 10.1007/s10579-008-9076-6]

Cai H S, Yuan Z Q, Gao Y W, Sun S T, Li N, Tian F Z, Xiao H, Li J X, Yang Z W, Li X W, Zhao Q L, Liu Z Y, Yao Z J, Yang M Q, Peng H, Zhu J, Zhang X W, Gao G P, Zheng F, Li R, Guo Z H, Ma R, Yang J, Zhang L, Hu X P, Li Y M and Hu B. 2022. A multi-modal open dataset for mental-disorder analysis. Scientific Data, 9(1): #178 [DOI: 10.1038/s41597-022-01211-x]

Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Barham P, Chung H W, Sutton C, Gehrmann S, Schuh P, Shi K S, Tsvyashchenko S, Maynez J, Rao A, Barnes P, Tay Y, Shazeer N, Prabhakaran V, Reif E, Du N, Hutchinson B, Pope R, Bradbury J, Austin J, Isard M, Gur-Ari G, Yin P C, Duke T, Levskaya A, Ghemawat S, Dev S, Michalewski H, Garcia X, Misra V, Robinson K, Fedus L, Zhou D, Ippolito D, Luan D, Lim H, Zoph B, Spiridonov A, Sepassi R, Dohan D, Agrawal S, Omernick M, Dai A M, Pillai T S, Pellat M, Lewkowycz A, Moreira E, Child R, Polozov O, Lee K, Zhou Z W, Wang X Z, Saeta B, Diaz M, Firat O, Catasta M, Wei J, Meier-Hellstern K, Eck D, Dean J, Petrov S and Fiedel N. 2022. PaLM: scaling language modeling with pathways [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/2204.02311.pdf

Cohn J F, Kruez T S, Matthews I, Yang Y, Nguyen M H, Padilla M T, Zhou F and De la Torre F. 2009. Detecting depression from facial actions and vocal prosody//Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. Amsterdam, the Netherlands: IEEE: 1-7 [DOI: 10.1109/ACII.2009.5349358]

Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J and Quatieri T F. 2015. A review of depression and suicide risk assessment using speech analysis. Speech Communication, 71: 10-49 [DOI: 10.1016/j.specom.2015.03.004]

Degottex G, Kane J, Drugman T, Raitio T and Scherer S. 2014. COVAREP — A collaborative voice analysis repository for speech technologies//Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE: 960-964 [DOI: 10.1109/ICASSP.2014.6853739]

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional Transformers for language understanding [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/1810.04805.pdf

Dhall A, Goecke R, Ghosh S, Joshi J, Hoey J and Gedeon T. 2017. From individual to group-level emotion recognition: EmotiW 5.0// Proceedings of the 19th ACM International Conference on Multimodal Interaction. Glasgow UK: ACM: 524-528 [DOI: 10.1145/3136755.3143004]

Dhall A, Goecke R, Joshi J, Hoey J and Gedeon T. 2016. EmotiW 2016: video and group-level emotion recognition challenges//Proceedings of the 18th ACM International Conference on Multimodal Interaction. Tokyo, Japan: ACM: 427-432 [DOI: 10.1145/2993148.2997638]

Dhall A, Goecke R, Joshi J, Wagner M and Gedeon T. 2013. Emotion recognition in the wild challenge 2013//Proceedings of the 15th ACM on International Conference on Multimodal Interaction. Sydney, Australia: ACM: 509-516 [DOI: 10.1145/2522848.2531739]

Dhall A, Murthy O V R, Goecke R, Joshi J and Gedeon T. 2015. Video and image based emotion recognition challenges in the wild: EmotiW 2015//Proceedings of 2015 ACM on International Conference on Multimodal Interaction. Seattle, USA: ACM: 423-426 [DOI: 10.1145/2818346.2829994]

Dinkel H, Wu M Y and Yu K. 2019. Text-based depression detection: what triggers an alert [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/1904.05154.pdf

Ekman P. 1999. Basic emotions//Dalgleish T and Power M J, eds. Handbook of Cognition and Emotion. New York, USA: John Wiley and

Sons: 45-60 [DOI: 10.1002/0470013494.ch3]

Esuli A and Sebastiani F. 2006. SENTIWORDNET: a publicly available lexical resource for opinion mining//Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy: European Language Resources Association (ELRA): 417-422

Eyben F, Wöllmer M and Schuller B. 2009. OpenEAR — introducing the Munich open-source emotion and affect recognition toolkit//Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. Amsterdam, the Netherlands: IEEE: 1-6 [DOI: 10.1109/ACII.2009.5349350]

Eyben F, Wöllmer M and Schuller B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor//Proceedings of the 18th ACM International Conference on Multimedia. Firenze, Italy: ACM: 1459-1462 [DOI: 10.1145/1873951.1874246]

Fang M, Peng S Y, Liang Y J, Hung C C and Liu S H. 2023. A multimodal fusion model with multi-level attention mechanism for depression detection. Biomedical Signal Processing and Control, 82: #104561 [DOI: 10.1016/j.bspc.2022.104561]

Fitzpatrick K K, Darcy A and Vierhile M. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Mental Health, 4 (2): #19 [DOI: 10.2196/mental.7785]

Fournier J C, DeRubeis R J, Hollon S D, Dimidjian S, Amsterdam J D, Shelton R C and Fawcett J. 2010. Antidepressant drug effects and depression severity: a patient-level meta-analysis. JAMA, 303(1): 47-53 [DOI: 10.1001/jama.2009.1943]

Gandhi A, Adhvaryu K, Poria S, Cambria E and Hussain A. 2023. Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion, 91: 424-444 [DOI: 10.1016/j.inffus.2022.09.025]

Gardiner P M, McCue K D, Negash L M, Cheng T, White L F, Yinusa-Nyahkoon L, Jack B W and Bickmore T W. 2017. Engaging women with an embodied conversational agent to deliver mindfulness and lifestyle recommendations: a feasibility randomized control trial. Patient Education and Counseling, 100(9): 1720-1729 [DOI: 10.1016/j.pec.2017.04.015]

Ghorbanali A, Sohrabi M K and Yaghmaee F. 2022. Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks. Information Processing and Management, 59(3): #102929 [DOI: 10.1016/j.ipm.2022.102929]

Gilbody S, Richards D, Brealey S and Hewitt C. 2007. Screening for depression in medical settings with the patient health questionnaire (PHQ): a diagnostic meta-analysis. Journal of General Internal Medicine, 22 (11): 1596-1602 [DOI: 10.1007/s11606-007-0333-y]

Gong Y and Poellabauer C. 2017. Topic modeling based multi-modal depression detection//Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. Mountain View, USA: ACM: 69-76 [DOI: 10.1145/3133944.3133945]

Gratch J, Artstein R, Lucas G, Stratou G, Scherer S, Nazarian A, Wood R, Boberg J, DeVault D, Marsella S, Traum D, Rizzo S and Morency L P. 2014. The distress analysis interview corpus of human and computer interviews//Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland: European Language Resources Association (ELRA): 3123-3128

Guo W T, Yang H W, Liu Z Y, Xu Y P and Hu B. 2021. Deep neural networks for depression recognition based on 2D and 3D facial expressions under emotional stimulus tasks. Frontiers in Neuroscience, 15: #609760 [DOI: 10.3389/fnins.2021.609760]

Guo Y R, Liu J L, Wang L, Qin W, Hao S J and Hong R C. 2024. A prompt-based topic-modeling method for depression detection on low-resource data. IEEE Transactions on Computational Social Systems, 11(1): 1430-1439 [DOI: 10.1109/TCSS.2023.3260080]

Han W, Chen H and Poria S. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis//Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics: 9180-9192 [DOI: 10.18653/v1/2021.emnlp-main.723]

Haque A, Guo M, Miner A S and Li F F. 2018. Measuring depression symptom severity from spoken language and 3D facial expressions [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/1811.08592.pdf

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He R D, Lee W S, Ng H T and Dahlmeier D. 2018. Adaptive semi-supervised learning for cross-domain sentiment classification//Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics: 3467-3476 [DOI: 10.18653/v1/D18-1383]

Hochreiter S and Schmidhuber J. 1997. Long short-term memory. Neural Computation, 9 (8): 1735-1780 [DOI: 10.1162/neco.1997.9.8.1735]

Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, de Las Casas D, Hendricks L A, Welbl J, Clark A, Hennigan T, Noland E, Millican K, van den Driessche G, Damoc B, Guy A, Osindero S, Simonyan K, Elsen E, Rae J W, Vinyals O and Sifre L. 2022. Training compute-optimal large language models [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/2203.15556.pdf

Hu G M, Lin T E, Zhao Y, Lu G M, Wu Y C and Li Y B. 2022. UniMSE: towards unified multimodal sentiment analysis and emotion recognition [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/2211.11256.pdf

Hu Y, Hou S J, Yang H M, Huang H and He L. 2023. A joint network

based on interactive attention for speech emotion recognition//Proceedings of 2023 IEEE International Conference on Multimedia and Expo (ICME). Brisbane, Australia: IEEE: 1715-1720 [DOI: 10.1109/ICME55011.2023.00295]

Huang Z S, Hu Q, Gu J G, Yang J, Feng Y and Wang G. 2019. Web-based intelligent agents for suicide monitoring and early warning. China Digital Medicine, 14(3): 2-6 (黄智生, 胡青, 顾进广, 杨洁, 冯媛, 王刚. 2019. 网络智能机器人与自杀监控预警. 中国数字医学, 14(3): 2-6) [DOI: 10.3969/j.issn.1673-7571.2019.03.001]

Imbir K K. 2020. Psychoevolutionary theory of emotion (Plutchik)// Zeigler-Hill V and Shackelford T K, eds. Encyclopedia of Personality and Individual Differences. Cham: Springer: 4137-4144 [DOI: 10.1007/978-3-319-24612-3_547]

Inkster B, Sarda S and Subramanian V. 2018. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR mHealth and uHealth, 6(11): #12106 [DOI: 10.2196/12106]

Joshi J, Goecke R, Alghowinem S, Dhall A, Wagner M, Epps J, Parker G and Breakspear M. 2013. Multimodal assistive technologies for depression diagnosis and monitoring. Journal on Multimodal User Interfaces, 7(3): 217-228 [DOI: 10.1007/s12193-013-0123-2]

Joulin A, Grave E, Bojanowski P and Mikolov T. 2016. Bag of tricks for efficient text classification [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/1607.01759.pdf

Kroenke K, Spitzer R L and Williams J B. 2001. The PHQ-9: validity of a brief depression severity measure. Journal of General Internal Medicine, 16(9): 606-613 [DOI: 10.1046/j.1525-1497.2001.016009606.x]

Ku L W and Chen H H. 2007. Mining opinions from the web: beyond relevance retrieval. Journal of the American Society for Information Science and Technology, 58(12): 1838-1850 [DOI: 10.1002/asi.20630]

Lai S N, Hu X F, Xu H X, Ren Z X and Liu Z. 2023. Multimodal sentiment analysis: a survey [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/2305.07611.pdf

Lam G, Huang D Y and Lin W S. 2019. Context-aware deep learning for multi-modal depression detection//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE: 3946-3950 [DOI: 10.1109/ICASSP.2019.8683027]

Lei S L, Dong G T, Wang X P, Wang K H and Wang S R. 2023. InstructERC: reforming emotion recognition in conversation with a retrieval multi-task LLMs framework [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/2309.11911.pdf

Li Y, Tao J H, Schuller B, Shan S G, Jiang D M and Jia J. 2016. MEC 2016: the multimodal emotion recognition challenge of CCPR 2016//Proceedings of the 7th Chinese Conference on Pattern Recognition. Chengdu, China: Springer: 667-678 [DOI: 10.1007/978-981-10-3005-5_55]

Lian Z, Liu B and Tao J H. 2021. CTNet: conversational Transformer network for emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29: 985-1000 [DOI: 10.1109/TASLP.2021.3049898]

Lian Z, Liu B and Tao J H. 2023a. SMIN: semi-supervised multi-modal interaction network for conversational emotion recognition. IEEE Transactions on Affective Computing, 14(3): 2415-2429 [DOI: 10.1109/TAFFC.2022.3141237]

Lian Z, Sun H Y, Sun L C, Chen K, Xu M Y, Wang K X, Xu K, He Y, Li Y, Zhao J M, Liu Y, Liu B, Yi J Y, Wang M, Cambria E, Zhao G Y, Schuller B W and Tao J H. 2023b. MER 2023: multi-label learning, modality robustness, and semi-supervised learning [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/2304.08981.pdf

Lian Z, Sun L C, Xu M Y, Sun H Y, Xu K, Wen Z F, Chen S, Liu B and Tao J H. 2023c. Explainable multimodal emotion reasoning [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/2306.15401.pdf

Lin L, Chen X R, Shen Y and Zhang L. 2020. Towards automatic depression detection: a BiLSTM/1D CNN-based model. Applied Sciences, 10(23): #8701 [DOI: 10.3390/app10238701]

Littlewort G, Whitehill J, Wu T F, Fasel I, Frank M, Movellan J and Bartlett M. 2011. The computer expression recognition toolbox (CERT)//Proceedings of 2011 IEEE International Conference on Automatic Face and Gesture Recognition (FG). Santa Barbara, USA: IEEE: 298-305 [DOI: 10.1109/FG.2011.5771414]

Liu H T, Li C Y, Wu Q Y and Lee Y J. 2023. Visual instruction tuning [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/2304.08485.pdf

Liu P F, Qiu X P and Huang X J. 2016. Deep multi-task learning with shared memory [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/1609.07222.pdf

Liu T T, Liu Z, Chai Y J, Wang J and Wang Y Y. 2021. Agent affective computing in human-computer interaction. Journal of Image and Graphics, 26(12): 2767-2777 (刘婷婷, 刘箴, 柴艳杰, 王瑾, 王媛怡. 2021. 人机交互中的智能体情感计算研究. 中国图象图形学报, 26(12): 2767-2777) [DOI: 10.11834/jig.200498]

Ly K H, Ly A M and Andersson G. 2017. A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. Internet Interventions, 10: 39-46 [DOI: 10.1016/j.invent.2017.10.002]

Ma X C, Yang H Y, Chen Q, Huang D and Wang Y H. 2016. DepAudioNet: an efficient deep model for audio based depression classification//Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. Amsterdam, the Netherlands: ACM: 35-42 [DOI: 10.1145/2988257.2988267]

McFee B, Raffel C, Liang D, Ellis D, McVicar M, Battenberg E and Nieto O. 2015. Librosa: audio and music signal analysis in python// Proceedings of the 14th Python in Science Conference. 18-25 [DOI: 10.25080/majora-7b98e3ed-003]

Mehrabian A. 1996. Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. Current Psychology, 14(4): 261-292 [DOI: 10.1007/BF02686918]

Mendels G, Levitan S, Lee K Z and Hirschberg J. 2017. Hybrid acoustic-lexical deep learning approach for deception detection// Interspeech 2017. Stockholm, Sweden: ISCA: 1472-1476 [DOI: 10.21437/Interspeech.2017-1723]

Mikolov T, Chen K, Corrado G and Dean J. 2013. Efficient estimation of word representations in vector space [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/1301.3781.pdf

Minsky M. 1988. The Society of Mind. New York, USA: Simon and Schuster

Mohammad S M and Turney P D. 2013. NRC Emotion Lexicon. National Research Council of Canada [DOI: 10.4224/21270984]

Morales M R, Scherer S and Levitan R. 2017. OpenMM: an open-source multimodal feature extraction tool//Interspeech 2017. Stockholm, Sweden: ISCA: 3354-3358 [DOI: 10.21437/Interspeech. 2017-1382]

Pasikowska A, Zaraki A and Lazzeri N. 2013. A dialogue with a virtual imaginary interlocutor as a form of a psychological support for well-being//Proceedings of the International Conference on Multimedia, Interaction, Design and Innovation. Warsaw Poland: ACM: 1-15 [DOI: 10.1145/2500342.2500359]

Pennington J, Socher R and Manning C. 2014. GloVe: global vectors for word representation//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics: 1532-1543 [DOI: 10.3115/v1/D14-1162]

Pham H, Liang P P, Manzini T, Morency L P and Póczos B. 2019. Found in translation: learning robust joint representations by cyclic translations between modalities//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI: 6892-6899 [DOI: 10.1609/aaai.v33i01.33016892]

Poria S, Cambria E and Gelbukh A. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis//Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics: 2539-2544 [DOI: 10.18653/v1/D15-1303]

Poria S, Hazarika D, Majumder N, Naik G, Cambria E and Mihalcea R. 2019. MELD: a multimodal multi-party dataset for emotion recognition in conversations//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics: 527-536 [DOI: 10.18653/v1/P19-1050]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. PMLR: 139: 8748-8763

Ringeval F, Schuller B, Valstar M, Cowie R, Kaya H, Schmitt M, Amiriparian S, Cummins N, Lalanne D, Michaud A, Ciftçi E, Güleç H, Salah A A and Pantic M. 2018. AVEC 2018 workshop and challenge: bipolar disorder and cross-cultural affect recognition//Proceedings of 2018 on Audio/Visual Emotion Challenge and Workshop. Seoul, Korea (South): ACM: 3-13 [DOI: 10.1145/3266302.3266316]

Rizzo A A, Lange B, Buckwalter J G, Forbell E, Kim J, Sagae K, Williams J, Rothbaum B O, Difede J, Reger G, Parsons T and Kenny P. 2011. An intelligent virtual human system for providing healthcare information and support. Studies in Health Technology and Informatics, 163: 503-509

Ruggiero K J, Ben K D, Scotti J R and Rabalais A E. 2003. Psychometric properties of the PTSD checklist—civilian version. Journal of Traumatic Stress, 16(5): 495-502 [DOI: 10.1023/A:1025714729117]

Rush A J, Carmody T J, Ibrahim H M, Trivedi M H, Biggs M M, Shores-Wilson K, Crismon M L, Toprac M G and Kashner T M. 2006. Comparison of self-report and clinician ratings on two inventories of depressive symptomatology. Psychiatric Services, 57(6): 829-837 [DOI: 10.1176/ps.2006.57.6.829]

Scherer S, Stratou G, Gratch J and Morency L P. 2013. Investigating voice quality as a speaker-independent indicator of depression and PTSD//Interspeech 2013. Lyon, France: [s.n.]: 847-851 [DOI: 10.21437/Interspeech.2013-240]

Scherer S, Stratou G, Lucas G, Mahmoud M, Boberg J, Gratch J, Rizzo A and Morency L P. 2014. Automatic audiovisual behavior descriptors for psychological disorder analysis. Image and Vision Computing, 32(10): 648-658 [DOI: 10.1016/j. imavis. 2014. 06.001]

Schroff F, Kalenichenko D and Philbin J. 2015. FaceNet: a unified embedding for face recognition and clustering//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 815-823 [DOI: 10.1109/CVPR. 2015. 7298682]

Schuller B, Valstar M, Eyben F, McKeown G, Cowie R and Pantic M. 2011. Avec 2011 – the first international audio/visual emotion challenge//Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction. Memphis, USA: Springer: 415-424 [DOI: 10.1007/978-3-642-24571-8_53]

Sebe N, Cohen I, Gevers T and Huang T S. 2005. Multimodal approaches for emotion recognition: a survey//Proceedings Volume 5670, Internet Imaging VI. San Jose, USA: SPIE: 56-67 [DOI: 10.1117/12.600746]

Shaver P, Schwartz J, Kirson D and O'Connor C. 1987. Emotion knowledge: further exploration of a prototype approach. Journal of Personality and Social Psychology, 52(6): 1061-1086 [DOI: 10.1037//0022-3514.52.6.1061]

Shen Y, Yang H Y and Lin L. 2022. Automatic depression detection: an emotional audio-textual corpus and a GRU/BiLSTM-based model//

Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 6247-6251 [DOI: 10.1109/ICASSP43922.2022.9746569]

Shott S. 1979. Emotion and social life: a symbolic interactionist analysis. American Journal of Sociology, 84(6): 1317-1334 [DOI: 10.1086/226936]

Soleymani M, Garcia D, Jou B, Schuller B, Chang S F and Pantic M. 2017. A survey of multimodal sentiment analysis. Image and Vision Computing, 65: 3-14 [DOI: 10.1016/j.imavis.2017.08.003]

Spek V, Cuijpers P, Nyklícek I, Riper H, Keyzer J and Pop V. 2007. Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: a meta-analysis. Psychological Medicine, 37(3): 319-328 [DOI: 10.1017/S0033291706008944]

Su W J, Zhu X Z, Cao Y, Li B, Lu L W, Wei F R and Dai J F. 2020. VL-BERT: pre-training of generic visual-linguistic representations [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/1908.08530.pdf

Su Y X, Lan T, Li H Y, Xu J L, Wang Y and Cai D. 2023. PandaGPT: one model to instruction-follow them all [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/2305.16355.pdf

Sun B, Zhang Y H, He J, Yu L J, Xu Q H, Li D L and Wang Z Y. 2017. A random forest regression method with selected-text feature for depression assessment//Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. Mountain View, USA: ACM: 61-68 [DOI: 10.1145/3133944.3133951]

Sun S T, Chen H Y, Shao X X, Liu L L, Li X W and Hu B. 2020. EEG based depression recognition by combining functional brain network and traditional biomarkers//Proceedings of 2020 IEEE International Conference on Bioinformatics and Biomedicine. Seoul, Korea (South): IEEE: 2074-2081 [DOI: 10.1109/BIBM49941.2020.9313270]

Tomkins S S. 1962. Affect Imagery Consciousness: Volume I: The Positive Affects. New York, USA: Springer

Torous J, Chan S R, Tan S Y M, Behrens J, Mathew I, Conrad E J, Hinton L, Yellowlees P and Keshavan M. 2014. Patient smartphone ownership and interest in mobile apps to monitor symptoms of mental health conditions: a survey in four geographically distinct psychiatric clinics. JMIR Mental Health, 1(1): #5 [DOI: 10.2196/mental.4004]

Valstar M, Schuller B, Smith K, Eyben F, Jiang B H, Bilakhia S, Schnieder S, Cowie R and Pantic M. 2013. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge//Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge. Barcelona, Spain: ACM: 3-10 [DOI: 10.1145/2512530.2512533]

Wang D, Guo X T, Tian Y M, Liu J H, He L H and Luo X M. 2023. TETFN: a text enhanced Transformer fusion network for multimodal sentiment analysis. Pattern Recognition, 136: #109259 [DOI: 10.1016/j.patcog.2022.109259]

Weizenbaum J. 1966. ELIZA — a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1): 36-45 [DOI: 10.1145/365153.365168]

Williamson J R, Godoy E, Cha M, Schwarzentruber A, Khorrami P, Gwon Y, Kung H T, Dagli C and Quatieri T F. 2016. Detecting depression using vocal, facial and semantic communication cues//Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. Amsterdam, the Netherlands: ACM: 11-18 [DOI: 10.1145/2988257.2988263]

World Health Organization. 2020a. Depression 2020a [EB/OL]. [2023-12-23]. https://www.who.int/health-topics/depression

World Health Organization. 2020b. Mental health in China 2020b [EB/OL]. [2023-12-23]. https://www.who.int/china/health-topics/mental-health

Wu S X, Dai D M, Qin Z W, Liu T Y, Lin B H, Cao Y B and Sui Z F. 2023. Denoising bottleneck with mutual information maximization for video multimodal fusion [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/2305.14652.pdf

Wu Y, Zhao Y Y, Yang H, Chen S, Qin B, Cao X H and Zhao W T. 2022. Sentiment word aware multimodal refinement for multimodal sentiment analysis with ASR errors [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/2203.00257.pdf

Xiao J Q and Luo X X. 2022. A survey of sentiment analysis based on multi-modal information//Proceedings of 2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC). Dalian, China: IEEE: 712-715 [DOI: 10.1109/IPEC54454.2022.9777333]

Xu L H, Lin H F, Pan Y, Ren H and Chen J M. 2008. Constructing the affective lexicon ontology. Journal of the China Society for Scientific and Technical Information, 27(2): 180-185 (徐琳宏, 林鸿飞, 潘宇, 任惠, 陈建美. 2008. 情感词汇本体的构造. 情报学报, 27(2): 180-185) [DOI: 10.3969/j.issn.1000-0135.2008.02.004]

Yang B, Wu L J, Zhu J H, Shao B, Lin X L and Liu T Y. 2022. Multimodal sentiment analysis with two-phase multi-task learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30: 2015-2024 [DOI: 10.1109/TASLP.2022.3178204]

Yang L, Jiang D M, He L, Pei E C, Oveneke M C and Sahli H. 2016. Decision tree based depression classification from audio video and language information//Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. Amsterdam, the Netherlands: ACM: 89-96 [DOI: 10.1145/2988257.2988269]

Yang L, Jiang D M, Xia X H, Pei E C, Oveneke M C and Sahli H. 2017. Multimodal measurement of depression using deep learning models//Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. Mountain View, USA: ACM: 53-59 [DOI: 10.1145/3133944.3133948]

Yang Y, Fairbairn C and Cohn J F. 2013. Detecting depression severity from vocal prosody. IEEE Transactions on Affective Computing, 4(2): 142-150 [DOI: 10.1109/T-AFFC.2012.38]

Yap M H, See J, Hong X P and Wang S J. 2018. Facial micro-expressions grand challenge 2018 summary//Proceedings of the

13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018). Xi'an, China: IEEE: 675-678 [DOI: 10.1109/FG.2018.00106]

Ye J Y, Yu Y H, Wang Q X, Li W T, Liang H, Zheng Y S and Fu G. 2021. Multi-modal depression detection based on emotional audio and evaluation text. Journal of Affective Disorders, 295: 904-913 [DOI: 10.1016/j.jad.2021.08.090]

Yi G F, Yang Y G, Pan Y, Cao Y H, Yao J X, Lv X, Fan C H, Lv Z, Tao J H, Liang S and Lu H. 2023. Exploring the power of cross-contextual large language model in mimic emotion prediction//Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation. Ottawa, Canada: Association for Computing Machinery: 19-26 [DOI: 10.1145/3606039.3613109]

Yin S, Liang C, Ding H Y and Wang S F. 2019. A multi-modal hierarchical recurrent neural network for depression detection//Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop. Nice, France: ACM: 65-71 [DOI: 10.1145/3347320.3357696]

Yu H L, Gui L K, Madaio M, Ogan A, Cassell J and Morency L P. 2017. Temporally selective attention model for social and affective state recognition in multimedia content//Proceedings of the 25th ACM international conference on Multimedia. Mountain View, USA: ACM: 1743-1751 [DOI: 10.1145/3123266.3123413]

Yu W M, Xu H, Meng F Y, Zhu Y L, Ma Y X, Wu J L, Zou J Y and Yang K C. 2020. CH-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics: 3718-3727 [DOI: 10.18653/v1/2020.acl-main.343]

Yu W M, Xu H, Yuan Z Q and Wu J L. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis//Proceedings of the 35th AAAI Conference on Artificial Intelligence. [s. l.]: AAAI: 10790-10797 [DOI: 10.1609/aaai.v35i12.17289]

Zadeh A, Chen M H, Poria S, Cambria E and Morency L P. 2017a. Tensor fusion network for multimodal sentiment analysis [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/1707.07250.pdf

Zadeh A, Chen M H, Poria S, Cambria E and Morency L P. 2017b. Tensor fusion network for multimodal sentiment analysis//Proceedings of 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics: 1103-1114 [DOI: 10.18653/v1/D17-1115]

Zadeh A A B, Liang P P, Poria S, Cambria E and Morency L P. 2018a. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics: 2236-2246 [DOI: 10.18653/v1/P18-1208]

Zhang F, Li X C, Lim C P, Hua Q, Dong C R and Zhai J H. 2022. Deep emotional arousal network for multimodal sentiment analysis and emotion recognition. Information Fusion, 88: 296-304 [DOI: 10.1016/j.inffus.2022.07.006]

Zhang J, Xue S Y, Wang X Y and Liu J. 2023. Survey of multimodal sentiment analysis based on deep learning//Proceedings of the 9th IEEE International Conference on Cloud Computing and Intelligent Systems (CCIS). Dali, China: IEEE: 446-450 [DOI: 10.1109/CCIS59572.2023.10263012]

Zhang P Y, Wu M Y, Dinkel H and Yu K. 2021. DEPA: self-supervised audio embedding for depression detection//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China: ACM: 135-143 [DOI: 10.1145/3474085.3479236]

Zhao J M, Zhang T G, Hu J W, Liu Y C, Jin Q, Wang X C and Li H Z. 2022. M3ED: multi-modal multi-scene multi-label emotional dialogue database//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics: 5699-5710 [DOI: 10.18653/v1/2022.acl-long.391]

Zhu D Y, Chen J, Shen X Q, Li X and Elhoseiny M. 2023a. MiniGPT-4: enhancing vision-language understanding with advanced large language models [EB/OL]. [2023-12-23]. https://arxiv.org/pdf/2304.10592.pdf

Zhu L N, Zhu Z C, Zhang C W, Xu Y F and Kong X J. 2023b. Multimodal sentiment analysis based on fusion methods: a survey. Information Fusion, 95: 306-325 [DOI: 10.1016/j.inffus.2023.02.028]

Zou B C, Han J L, Wang Y X, Liu R, Zhao S H, Feng L, Lyu X W and Ma H M. 2023. Semi-structural interview-based Chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. IEEE Transactions on Affective Computing, 14(4): 2823-2838 [DOI: 10.1109/TAFFC.2022.3181210]

作者简介

陶建华，男，教授，主要研究方向为语音处理、认知推理、数据内容分析和智能交互。E-mail: jhtao@tsinghua.edu.cn

梁山，男，通信作者，副教授，主要研究方向为语音信息处理、语音伪造检测、传感器阵列语音信息处理。

E-mail: Shan.Liang@xjtlu.edu.cn

范存航，男，副教授，主要研究方向为语音增强与分离、脑机接口技术。E-mail: cunhang.fan@ahu.edu.cn

连政，男，助理研究员，主要研究方向为情感计算和人机交互。E-mail: lianzheng2016@ia.ac.cn

吕钊，男，教授，主要研究方向为脑机接口技术、情感计算和脑机协同智能。E-mail: kjlz@ahu.edu.cn

沈莹，女，副教授，博士生导师，主要研究方向为语音信号处理、自然语言理解和情感计算。

E-mail: yingshen@tongji.edu.cn