E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2025)06-2120-19

论文引用格式:Wang S M, Liu C G, Chen S Y and Liu Q S. 2025. A survey of multimodal emotion recognition from facial expressions, audios, and language. Journal of Image and Graphics, 30(6):2120-2138(王善敏, 刘成广, 陈胜宇, 刘青山. 2025. 面向表情、语音和语言的多模态情感识别综述. 中国图象图形学报, 30(6):2120-2138)[DOI:10.11834/jig.250168]

面向表情、语音和语言的多模态情感识别综述

王善敏[1],刘成广[2],陈胜宇[2],刘青山[1,2*]

1. 南京邮电大学,南京　210023; 2. 南京信息工程大学,南京　210044

摘　要:情感识别旨在利用计算机建立情感表达的载体与情感度量之间的关系,并赋予计算机识别、理解人类情感的能力,在人机交互的多个领域发挥重要作用,是人工智能领域的重要发展方向。心理学研究表明人们内心情感状态的外在表现通常是通过面部表情、语音和语言等多模态行为信号进行联合表达的,并伴随脑电、心电等生理指标的变化。相较于脑电、心电等生理信号,面部表情、语音和语言等行为信号凭借其非侵入式的信号采集方式,在无感式情感识别中展现出显著优势。本文回顾国内外无感式情感识别的发展历程,并从5个方面对当前研究现状进行系统梳理与归纳。首先,阐述了离散情感模型和维度情感模型对情感的表达作用;其次,对常用的情感数据库进行了统计分析;在此基础上,围绕情感的无感式表达,分别从表情、语音、语言和多模态4个维度,详细阐述了表情识别、语音情感识别、语言情感识别和多模态情感识别的相关技术,并针对其中的关键技术挑战,对相关研究工作进行了总结与归纳;进一步,介绍了无感式情感识别在医疗健康、社会服务和智能驾驶等多个智能交互领域的应用案例;最后,总结了情感识别领域目前面临的挑战,并对未来发展方向进行了展望。本文旨在对情感识别相关研究进行深入的分析与总结,为该领域的研究者提供有价值的参考。

关键词:表情识别;语音情感识别;语言情感识别;多模态情感识别;智能交互

A survey of multimodal emotion recognition from facial expressions, audios, and language

Wang Shanmin[1], Liu Chengguang[2], Chen Shengyu[2], Liu Qingshan[1,2*]

1. Nanjing University of Posts and Telecommunications, Nanjing　210023, China;

2. Nanjing University of Information Science & Technology, Nanjing　210044, China

Abstract: Emotion recognition is an essential research direction in artificial intelligence (AI), focusing on modeling the relationship between emotional expressions and measurable features. It enables computers to recognize and understand human emotions, thereby playing a crucial role in various domains of human-computer interaction. From intelligent assistants to mental health monitoring and social robotics, emotion-aware systems are becoming increasingly pervasive, making affective computing a cornerstone of future intelligent technologies. Psychological studies have long confirmed that human emotions are typically expressed through a combination of behavioral cues, most notably facial expressions, speech prosody, and language content. These multimodal signals are often intertwined, with each modality providing complementary information that reflects the inner emotional state of an individual. In addition to these observable behaviors, physi-

收稿日期:2025-04-08;**修回日期**:2025-04-10;**预印本日期**:2025-04-17

*通信作者:刘青山　qsliu@njupt.edu.cn

基金项目:国家重点研发计划资助(2024YFC3308400);江苏省前沿引领技术基础研究专项(BK20192004B)

Supported by: National Key R&D Program of China (2024YFC3308400); Natural Science Foundation of Jiangsu Province, China(BK20192004B)

ological signals such as electroencephalogram and electrocardiogram also vary with emotional changes. However, collecting physiological signals often requires contact-based or invasive sensors, limiting their practical use in daily environments. By contrast, facial expressions, voice, and language are naturally occurring, easily accessible, and can be captured unobtrusively by cameras and microphones. These advantages make non-invasive emotion recognition approaches particularly attractive for real-world applications, where user comfort and ease of deployment are critical. The field of non-invasive emotion recognition has made significant progress in recent years, supported by advancements in deep learning, multimodal fusion techniques, and the increasing availability of emotion-labeled datasets. At the theoretical level, emotions are typically represented through two main models: discrete and dimensional. The discrete model classifies affect into a set of basic categories such as happiness, sadness, anger, fear, and surprise. This model is intuitive and aligns well with how humans often perceive emotion. Alternatively, the dimensional model maps emotions onto a continuous space, commonly defined by valence (positive to negative), arousal (calm to excited), and sometimes dominance (submissive to dominant). This model allows for more nuanced and dynamic descriptions of emotional states. Both models form the basis for designing annotation protocols and developing machine learning algorithms for emotion detection. A fundamental driver of research in this field is the availability of multimodal emotional datasets. Numerous public databases support facial expression recognition, speech emotion analysis, and affective text mining. For example, facial datasets like AffectNet and real-world affective faces database(RAF-DB) contain millions of annotated images representing a broad spectrum of expressions, while audio datasets like interactive emotional dyadic motion capture database(IEMOCAP) and Ryerson audio-visual database of emotional speech and song(RAVDESS) offer rich emotional speech recordings. Language datasets, including those developed for sentiment analysis challenges, provide labeled corpora for detecting emotion in both written and spoken text. Multimodal datasets like Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSI) and Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity(CMU-MOSEI) integrate synchronized recordings of facial expressions, voice, and transcribed language, supporting more comprehensive studies on multimodal fusion and cross-modal emotion reasoning. Technological progress in this area has been fueled by innovations in machine learning and signal processing. Facial expression recognition has evolved from handcrafted feature extraction to end-to-end deep learning frameworks using convolutional neural networks (CNNs), attention mechanisms, and, more recently, vision transformers. Temporal modeling techniques have also been used to capture dynamic facial cues, including micro-expressions that may reveal concealed emotions. In speech emotion recognition, acoustic features such as pitch, intensity, Mel-frequency cepstral coefficients (MFCCs), and spectral entropy are used to infer emotion. Recurrent neural networks (RNNs), convolutional layers, and transformers have also shown strong performance, especially when trained on large emotional corpora. For language-based emotion analysis, both traditional natural language processing (NLP) techniques and modern pre-trained language models like bidirectional encoder representations from Transformers (BERT) and generative pretrained Transformer(GPT) have proven effective in extracting emotional content from text, including subtle and context-dependent cues. A major trend in the field is the integration of multiple modalities to improve recognition accuracy and robustness. Multimodal emotion recognition systems aim to leverage complementary information from facial, vocal, and linguistic cues. To achieve this, various fusion strategies have been developed: early fusion, which combines raw features; late fusion, which merges decision outputs; and hybrid fusion methods, which integrate features at intermediate stages. Cross-modal attention, contrastive learning, and reinforcement learning have been used to address issues such as modality imbalance, missing data, and asynchronous signals. Therefore, multimodal approaches consistently outperform single-modality systems, especially in real-world conditions where one modality might be unreliable. The applications of emotion recognition are also expanding into several domains. In healthcare, non-invasive emotion recognition is used for monitoring mental health, detecting depression, and supporting emotional therapy. In education, it enables intelligent tutoring systems to gauge student engagement and adapt teaching strategies. In intelligent driving, it helps assess driver fatigue and stress to enhance road safety. Customer service chatbots and entertainment platforms are also increasingly equipped with emotion-aware capabilities to personalize interactions. The ability to recognize and adapt to human emotions not only improves user satisfaction but also fosters more natural and effective communication between humans and machines. Despite these advances, the field of emotion recognition continues to face numerous challenges.

Cultural and individual differences in emotional expression can hinder the generalizability of models trained on specific datasets. Emotions are often ambiguous and context-dependent, making accurate annotation and recognition difficult. Real-time deployment requires models that are accurate, lightweight, and efficient. Moreover, ethical concerns surrounding emotion recognition—especially regarding privacy, consent, and potential misuse of affective data—are increasingly coming into focus. Thus, emotion-aware systems must be developed and deployed responsibly, with transparency and fairness. Several promising directions are emerging. Looking ahead, transfer learning and domain adaptation techniques can help models generalize better across diverse user groups and environments. Furthermore, self-supervised and unsupervised learning approaches may alleviate the dependency on large annotated datasets. Researchers are also exploring explainable AI methods to increase the interpretability and trustworthiness of emotion recognition systems. Additionally, the fusion of modalities through generative techniques and cross-modal alignment is opening new avenues for robust and flexible emotion understanding, especially in noisy or incomplete input scenarios. In conclusion, this review presents a comprehensive analysis of non-invasive emotion recognition based on facial expressions, speech, and language. By summarizing theoretical foundations, available resources, technical advancements, practical applications, and ongoing challenges, it aims to provide researchers and practitioners with a clear understanding of the current landscape and future potential of this rapidly evolving field.

Key words: facial expression recognition; audio emotion recognition; language emotion recognition; multimodal emotion recognition; intelligent interaction

0 引　言

人工智能(artificial intelligence, AI)奠基人之一、图灵奖得主、美国麻省理工学院的 Minsky(1986)在其著作《心智社会》中提出“问题不在于智能机器能否有情感,而在于缺乏情感的机器能否实现智能”,强调了情感在机器智能发展中的关键作用。美国麻省理工学院的 Picard (2000)在《情感计算》一书中进一步阐述了情感对机器智能的影响主要体现在主观感受、理性思考和行为决策等方面。由此可见,成熟的情感识别算法是机器理解、反馈与模拟人类行为,进而实现高级机器智能的重要基石。2024 年,中国科学技术协会发布的重大科学问题中,“情智兼备数字人与机器人的研究”位列首位(中国科协,2024),充分表明情感识别已成为当今时代一个具有广阔发展前景和重大研究价值的热点领域(彭晓兰 等,2024)。

心理学研究表明,人们内心情感状态的外在表现通常是通过面部表情、语音和语言等多模态行为信号进行联合表达的,并伴随脑电、心电等生理指标的变化(Izard, 1977)。相较于脑电、心电等生理信号,面部表情、语音和语言等行为信号凭借其非侵入式的信号采集方式,在无感式情感识别中展现出显著优势。基于情感表达的媒介,现有的无感式情感识别研究可划分为4个主要子方向:面部表情识别、语音情感识别、语言情感识别和多模态情感识别。表情、语音和语言3种模态既可以独立地表达情感,也可以相互增强和补充,共同构成一个完整的情感表达。具体而言,情感的产生会触发大脑神经活动,进而控制面部肌肉的运动。这些肌肉的协同运动形成了各种不同的面部表情,例如嘴角下垂表示厌恶的情感(Uono 等,2017)。语音情感识别则是基于情感的产生会影响呼吸、声带振动和声道形状,从而改变语音的声学特征。不同的声学特征组合形成了各种情感语音,例如语速过快对应于兴奋或紧张的情绪(Bhangale 和 Kothandaraman, 2023)。语言情感识别则依赖于人们在表达情感时会选择特定的词汇、句子结构和表达方式,通过分析语言文本中的情感词汇、句子结构、修辞手法和上下文语境,可以识别出语言文本所表达的情感倾向(Lin, 2024)。在面部表情、语音和语言等无感信号中,面部表情由于具备出现频次高、生物学普适性(跨文化6类基本情绪)、高辨识度和直观性等特点,通常被认为是情感推断的主导模态(Ekman 和 Friesen, 1971)。

对于情感识别相关领域的研究已有三十多年的历史。面部表情识别、语音情感识别和语言情感识别均经历了从基于手工设计特征到基于端到端的情感识别方法的演进历程。在早期阶段,研究者们主要依赖于人工提取的特征,例如面部的局部二值特征(Ojala 等,2002)、语音的梅尔频率倒谱系数 (Mel-frequency cepstral coefficient, MFCC)(Schuller 等,

2004）和语言文本的情感词袋模型（Esuli 和 Sebastiani，2006）等，结合支持向量机（support vector machine，SVM）（Suthaharan，2016）或隐马尔可夫模型（hidden Markov model，HMM）（Eddy，1996）等传统分类器进行情感识别。然而，这种方法对人工经验依赖性强，且难以捕捉复杂的情感表达。随着深度学习的兴起，端到端的情感识别模型，如卷积神经网络（convolutional neural network，CNN）（He 等，2016）、循环神经网络（recurrent neural network，RNN）（Grossberg，2013）和 Transformer（Vaswani 等，2017）等，能够直接从原始数据中学习情感特征，极大地简化了情感识别流程，并取得了显著的性能提升。

本文对情感识别领域的相关方面进行了综合性介绍，涵盖情感表示模型、情感语料库以及经典的情感识别方法。具体而言，首先，重点阐述了离散情感模型和维度情感模型，并分析了各自的优缺点；接着对主流的情感数据库进行了简要的信息概括，详细阐述面部表情识别、语音情感识别和语言情感识别等子任务中的经典研究，重点介绍多模态情感识别方法，对相关技术进行总结与归纳，并对经典研究的结果进行展示分析；然后，介绍无感式情感识别的应用；最后，对情感识别领域当前面临的挑战和未来的发展方向进行总结与展望。

1 情感表示模型

情感的定义是理解和识别人类情绪的基础。在情感计算领域，存在两种通用的情感模型：离散情感模型和维度情感模型。离散情感模型以其简洁直观的特点得到广泛应用，而维度情感模型则能够更细粒度地描述情感的变化。本节将分别介绍这两种经典的情感模型。

离散情感模型将情绪视为有限的基本类别集合，可分为基础和复合情感。Ekman（Ekman 和 Friesen，1971）的 6 种基本情绪模型（生气、厌恶、恐惧、高兴、悲伤、惊讶）因其跨文化一致性而在情感计算领域得到广泛认可。不同学者对基本情绪的定义略有差异，例如表 1 中的 Tomkins（1962，1963）和 Izard（2013）的离散情感模型。Plutchik（2001）的情感轮模型定义了 8 种基本情绪及其对立关系，且基本情绪可组合成更复杂的情感（如喜悦 + 信任 = 爱）。离散模型易于理解，但仅能定性描述有限的情感状态，且部分类别间易混淆。

表 1　多种离散情感模型

Table 1　Various discrete emotion models

情感模型	情感类别
Ekman 情感模型	生气、厌恶、恐惧、高兴、悲伤、惊讶
Tomkins 情感模型	愤怒、兴趣、轻蔑、厌恶、痛苦、恐惧、欢乐、羞耻
Plutchik 情感轮模型	喜悦、信任、害怕、惊讶、悲伤、厌恶、生气、期待
Izard 情感模型	生气、轻视、厌恶、悲痛、恐惧、内疚、有趣、开心、羞耻、惊喜

维度情感模型将人类情感视为连续变化而非离散类别，通常采用多维空间描述情感，以更细腻地刻画情绪变化。Russell（1980）的 Valence（效价）-Arousal（唤醒度）模型是一种经典的维度情感模型，如图 1 所示，该模型包含愉悦度和唤醒度两个维度，分别表示情感的正负性和激活程度。此外，Russell 和 Mehrabian（1977）还提出了 PAD（pleasure-arousal-dominance）模型，增加了支配度维度，衡量个体对情境的控制感。与离散模型相比，维度模型既可定性又可定量描述情感，反映情感的连续变化，因此近年来应用增多。然而，细粒度的情感维度标注在数据库构建方面仍面临挑战。

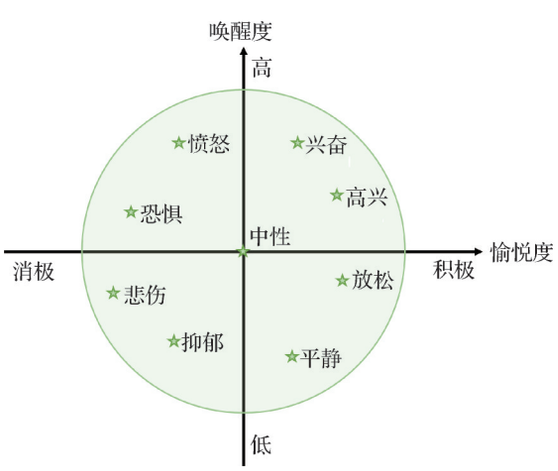


图 1　维度情感模型

Fig. 1　Dimensional emotion model

2 情感数据库

在情感识别研究中，高质量的数据库是推动技

术进步的关键因素。无论是表情识别、语音情感识别、语言文本情感识别，还是多模态情感识别，数据库的构建方式和数据质量都会直接影响模型的性能，尤其是在深度学习等数据驱动方法上表现尤为明显。因此，本文首先对情感识别领域的常用数据库进行介绍。目前各类情感数据库的构建并无统一标准，通常根据不同的任务需求进行设计。现有的情感数据库可分为表演型和自发型。表演型数据由职业演员录制，具有较强的情感表达，而自发型数据则更贴近实际应用，但识别难度更高。表情识别和语音情感识别数据库涵盖了实验室环境下表演型的标准化数据和真实世界场景中自发产生的自然数据。语言文本情感数据库则广泛来源于社交媒体、新闻评论等真实文本，多模态情感数据库综合了表情、语音及语言文本等多种模态信息，能够提供更全面的情感特征表达。表2总结了情感识别领域常用的数据库。

2.1 表情数据库

CK+（extended Cohn-Kanad）数据集是表情识别早期研究的常用数据集，包含 593 个来自 123 位受试者的面部表情视频序列。视频记录了实验室环境下受试者由中性表情到极端表情的面部变化，其中 327 个视频被标注为 Ekman 6种基本情绪、中性或蔑视表情（Lucey 等，2010）。

Oulu-CASIA（Oulu-CASIA NIR&VIS）数据集包含 80 位受试者在不同光照条件下的6种基本面部表情图像，每种光照约 18～30 幅，共 2 880 个图像序列。该数据集同时提供近红外和可见光图像（Zhao 等，2011）。

MMI 数据集包含超过 2 900 个视频和 1 500 幅来自 75 位受试者的高分辨率图像，标注了6种基本表情和中性表情。该数据集还在帧级别上进行了部分面部动作单元（action unit，AU）编码，指明了每一帧的 AU 状态（中性、起始、顶点、偏移），以便于研究者分析表情的细微变化（Pantic 等，2005）。

SFEW（static facial expression in the wild）数据集模拟真实场景，提供了一个野外环境下的静态面部表情识别挑战。它通过人脸关键点聚类算法从 AFEW（acted facial expression in the wild）数据库选择关键帧创建，SFEW 2.0 作为 EmotiW 2015 SReco 子挑战赛的基准数据，包含958个训练样本、436个验证样本和372个测试样本，每幅图像标注7种表情之一（Dhall 等，2011）。

FER2013（facial expression recognition 2013）数据集于 ICML2013（international conference on machine learning 2013）提出，是情感分析和人机交互领域常用的表情识别数据集，包含 35 887 幅灰度面部图像，标注为6类基本表情或中性。数据集分为 28 709 个训练集样本、3 589 个公共测试样本和 3 589 个私有测试样本。标签为众包标注，存在噪声或不一致性的情况（Goodfellow 等，2013）。

RAF-DB（real-world affective faces database）数据集包含29 672幅从互联网下载的面部图像，展示真实环境中的情感表达。每幅图像由约40名标注者独立标注，涵盖7种基本情绪和12类双标签复合表情，并提供人脸关键点、边界框、种族、年龄和性别等属性标注（Li 等，2017a）。

AffectNet 数据集包含超过 100 万幅互联网收集的面部图像，涵盖了6类基础情绪、中性和蔑视的离散情感标注以及效价和唤醒度的维度情感标注（Mollahosseini 等，2019）。

ExpW（expression in the wild）数据集整理了 91 793 幅从互联网收集的自然场景面部图像，覆盖广泛的年龄、性别和种族，包含多种光照、姿态和背景变化。数据集中每幅图像都经过人工标注，提供7种情绪类别标签（Zhang 等，2018）。

DFEW（dynamic facial expression in the wild）是一个专为动态面部表情识别设计的大型视频数据集，包含 16 372 段从 YouTube 收集的自然场景视频片段，聚焦于动态表情变化过程。这些片段由 12 位专家提供7种情绪类别标注（Jiang 等，2020）。

2.2 语音情感数据库

IEMOCAP（interactive emotional dyadic motion capture database）是一个常用的语音和多模态情感识别数据集，包含10名演员在剧本表演和即兴表演中录制的约12 h的音视频数据。数据集标注了9种情感类别，并提供效价、唤醒度、支配度等维度型标注（Busso 等，2008）。

CASIA（CASIA Chinese emotional speech database）是一个中文语音情感数据集，包含4名演员录制的语音，标注为6种情感类别（Zhang 和 Jia，2008）。

EmoDB（Berlin emotional speech database）是一个德语语音情感数据库，包含10名专业演员录制的语音，标注为8种情感类别（Burkhardt 等，2005）。

CHEAVD（Chinese emotion audio-visual database）是一个中文自然情感音视频数据库，包含从电影和电视节目中提取的2 322个片段，标注为8类情绪（Li 等，2017b）。

RAVDESS（Ryerson audio-visual database of emotional speech and song）包含24名演员录制的语音和视频，涵盖 8 种情感类别，包括语音和歌曲两种模式，常用于语音情感识别和多模态情感识别（Livingstone 和 Russo，2018）。

MSP-Podcast（Montreal Speech Prosody Podcast）是一个大型的、更贴近真实场景的英文语音情感数据集，从播客节目中提取。它包含超过 62 140 段语音片段，并进行了多维度标注，包括9种情感类别和效价、唤醒度、支配度 3 个维度（Lotfian 和 Busso，2019）。

2.3 语言文本情感数据库

SST-5（Stanford sentiment treebank-5 classes）是一个细粒度的情感分析数据集，包含约 11 855 段电影评论，标注为5种情感极性（Socher 等，2013）。

IMDB（internet movie database）是一个常用的二分类情感分析数据集，包含 50 000 条电影评论，标注为正面或负面情感（Maas 等，2011）。

Twitter US Airline Sentiment 是一个情感分析数据集，包含 14 640 条关于美国航空公司的 Twitter 推文，标注为正面、负面或中性情感（Rane 和 Kumar，2018）。

SemEval-2013（international workshop on semantic evaluation 2013）包含从 Twitter 收集的 15 196 条推文，标注为正面、负面或中性情感（Martínez-Cámara 等，2014）。

ChnSentiCorp（Chinese sentiment corpus）是一个中文情感分析数据集，包含约 10 800 条评论，涵盖酒店、笔记本电脑和书籍3个领域，标注为正面或负面情感（Tan 和 Zhang，2008）。

NLPCC-2014（National Language Processing and Chinese Computing Conference 2014）是由中国中文信息学会（NLPCC）从新浪微博采集的 14 000 条微博文本构建的中文情感分析数据集，每条微博文本都被标注为7类情绪中的一种（Xu 等，2016）。

Weibo Emotion Corpus 是一个大规模中文微博情感数据集，包含约4万条微博，标注 7 个情感类别。每条微博可能被标注为多个情感标签，适用于多标签情感分类任务（Cheng 等，2017）。

2.4 多模态情感数据库

CMU-MOSI（Carnegie Mellon University multimodal opinion sentiment and emotion intensity）（Zadeh 等，2016）和 CMU-MOSEI（Carnegie Mellon University multimodal opinion sentiment and emotion intensity）（Zadeh 等，2018c）是一个多模态情感分析数据集，分别包含 2 199 和 22 856 段来源于 YouTube 电影评论视频的片段，标注为 −3（强负面）到 +3（强正面）的连续情感强度值，以及7种情绪类别。

MELD（multimodal emotionlines dataset）是一个多模态对话情感数据集，包含从美剧《老友记（Friends）》中提取的 1 433 段对话，拆分为 13 708 个句子，每个句子使用7类情感标注（Poria 等，2019）。

MERR（multimodal emotion recognition in response）数据集包含 28 618 段视频，提供了粗粒度的9类情感标签，以及 4 487 段视频，提供细粒度的情感描述和推理过程。该数据集涵盖了多种情感类别，帮助模型从不同场景中学习并能够推广到实际应用中（Cheng 等，2024）。

EMER-Coarse（emotion recognition in the wild-coarse）数据集由中国科学院自动化研究所于2024年构建，包含 115 595 段以人为中心的视频样本，标注了5种情感类别，并为每个视频提供了详细的情感解释（Lian 等，2024b）。

3 基于表情、语音和语言的多模态情感识别方法

3.1 面部表情识别

面部表情识别是指利用计算机技术自动分析和识别面部表情所传达的情感状态。如图 2 所示，在人脸检测、人脸对齐等预处理步骤之后，表情识别流程主要包括表情特征提取和情感分类两个关键环节，其中表情特征提取是核心。依据特征提取方法的不同，表情识别方法可分为基于手工特征和基于端到端学习两大类。在早期阶段，传统方法主要依赖局部二值模式（Ojala 等，2002）、基于面部关键点的形状特征（Ghimire 和 Lee，2013）以及小波变换（Gu 等，2012）等手工设计的特征来描述面部表情变化。深度学习发展后，表情特征提取逐渐向数据驱动方式转变，卷积神经网络等深度模型能够自动学

表2　无感式情感识别数据库

Table 2　Contactless emotion recognition datasets

	数据库	样本数量	离散标签	维度标签	模态	类型
表情数据库	CK+	593个序列	6类基本表情、中性、蔑视		视觉	表演
	Oulu-CASIA	2 880个序列	6类基本表情		视觉	表演
	MMI	2 900段视频	7类基本表情		视觉	表演
	SFEW	1 766段视频	6类基本表情、中性		视觉	自发
	FER2013	35 887幅图像	6类基本表情、中性、蔑视		视觉	自发
	RAF-DB	29 672幅图像	6类基本表情、中性		视觉	自发
	AffectNet	45万幅图像	6类基本表情、中性、蔑视	效价、唤醒度	视觉	自发
	ExpW	91 793幅图像	6类基本表情、中性		视觉	自发
	DFEW	16 372段视频	6类基本表情、中性		视觉	自发
语音情感数据库	IEMOCAP	10 039个句子	6类基本表情、中性、沮丧	效价、唤醒度、支配度	语音、多模态	表演
	CASIA	9 600个句子	生气、高兴、恐惧、悲伤、惊讶、中性		语音	表演
	EmoDB	800个句子	愤怒、无聊、厌恶、焦虑、恐惧、快乐、悲伤、中性		语音	表演
	CHEAVD	2 322个句子	6类基本表情、中性、焦虑		语音	表演
	RAVDESS	7 356个句子	6类基本表情、平静		语音、多模态	表演
	MSP-Podcast	62 140个句子	6类基本表情、中立、蔑视、其他	效价度、唤醒度、支配度	语音	自发
语言问题情感数据库	SST-5	11 855个句子	非常正面、正面、中性、负面、非常负面		语言	电影评论
	IMDB	50 000个句子	正面、负面		语言	电影评论
	Twitter US Airline	14 640条推文	正面、负面、中性		语言	Twitter推文
	SemEval-2013	15 196条推文	正面、负面、中性		语言	Twitter推文
	ChnSentiCorp	10 800条评论	正面、负面		语言	网页评论
	NLPCC-2014	14 000 条评论	6类基本表情、喜爱		语言	新浪微博
	Weibo Emotion Corpus	4万条微博	6类基本表情、喜爱		语言	新浪微博
多模态情感数据库	CMU-MOSI	2 199段视频	极度消极、消极、轻微消极、中性、轻微积极、积极、极度积极	−3到+3的情绪值	多模态	视频片段
	CMU-MOSEI	22 856段视频	极度消极、消极、轻微消极、中性、轻微积极、积极、极度积极	−3到+3的情绪值	多模态	视频片段
	MELD	13 708段视频	6类基本表情、中性		多模态	电影片段
	MERR	28 618段视频	6类基本表情、中性、怀疑、蔑视		多模态	网页搜集
	EMER-Coarse	115 595段视频	愤怒、快乐、悲伤、担忧、惊讶		多模态	网页搜集

习深层次的表情特征(Wang等,2022,2023a),提高了识别的准确性和鲁棒性。

随着深度学习的深入应用,表情识别在多个关键问题上取得进展,包括缓解模型参数与有限样本量的不匹配、提升对个体属性差异的适应性、增强遮挡与姿态变化下的特征完整性以及提高模型的泛化能力。深度学习早期,模型参数量远超表情样本量,易导致过拟合。常见方法是先在大规模人脸数据库上预训练,再用表情数据微调,如FaceNet2-ExpNet模型采用两阶段微调策略提升特征学习效果(Ding

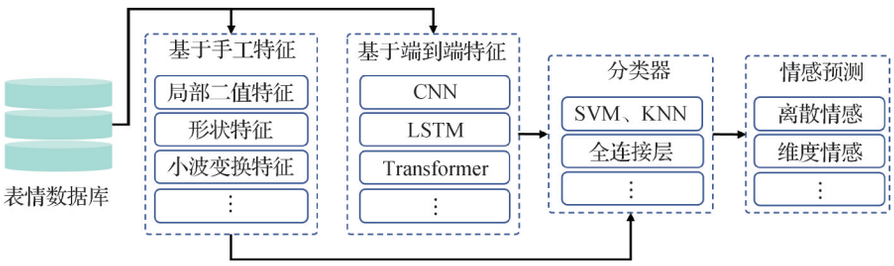


图2　表情识别框架

Fig. 2　Framework of facial expression recognition

等，2017）。随着大规模表情数据库的出现，该问题有所缓解，但个体属性（如身份、年龄、性别）对表情特征提取的干扰仍然突出（Wang 等，2024c）。对此，相关研究主要从表征约束和属性解耦入手，如 Cai 等人（2018）改进中心化损失增强类间区分度，Yang 等人（2018）提出 DeRL 模型通过残差学习提取身份无关的表情特征。

随着表情识别技术的日趋成熟，研究对象逐渐从实验室环境转向真实世界人脸，随之而来的面部遮挡和头部姿态偏转等自遮挡问题日益突出（王宇键 等，2025；卢莉丹 等，2024；崔鑫宇 等，2024）。针对这一问题，研究者主要采用缺失区域补全和未遮挡区域特征筛选两类方法。例如，Lee 等人（2023）结合自编码器和遮挡检测器修复遮挡区域纹理，Li 等人（2019）提出注意力 CNN 通过动态权重调整，强化未遮挡区域的特征提取。同时，真实场景的表情识别还面临模型域泛化能力弱的问题，由此衍生了跨域表情识别研究，其核心在于源域与目标域样本的分布对齐，以保证模型的跨域适应性（Chen 等，2022；Li 等，2021；Gao 等，2024b；Ji 等，2023）。然而，由于表情数据类内差异大、类间相似性强，简单的边缘分布对齐难以满足需求，因此部分方法进一步提出对齐不同域的条件分布，增强样本与标签间的映射一致性（Li 和 Deng，2022；Wang 等，2025b）。

近年来，大语言模型（large language model，LLM）推动了视觉情感识别向跨模态与可解释方向演进。以 CLIPER 为代表的模型利用 LLM 强大的语义表征能力，通过扩展情感标签并引入标签文本与表情特征的对比学习，有效提升了识别准确性，在 RAF-DB 等多个数据集上取得了显著进展（Li 等，2024）。然而，基于单一表情的情感识别在实际应用中仍然存在情感歧义、易受光照等因素干扰，缺乏上下文情境，难以准确反映个体真实情绪。结合语音、语言等多模态信息，有助于弥补面部表情的局限，提升情感识别的准确性与鲁棒性。

3.2　语音情感识别

语音情感识别是指通过计算模型解析语音信号中隐含的情感状态，其核心在于情感相关的声学特征提取。如图 3 所示，随着研究的深入，语音情感识别已从传统的声学特征工程逐渐转向深度语义特征的提取与学习。早期的研究侧重于手工设计的声学特征集，包括音高、语速及音量等韵律特征，梅尔频率倒谱系数、频谱包络、共振峰等谱特征以及抖动、闪烁、声带振动等音质特征。例如，Schuller 等人（2004）通过将音高、频率和音色等基本声学特征与支持向量机相结合，进行情感分类。Huang 等人（2020）则分析了不同情感类别的语音信号特征差异，并对各个特征的贡献度进行了排序，研究发现基

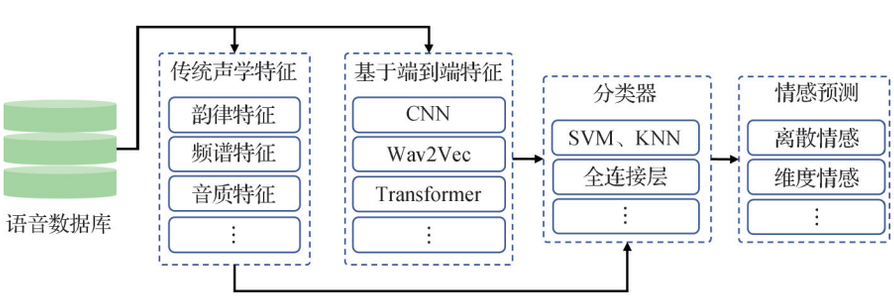


图3　语音情感识别框架

Fig. 3　Framework of audio emotion recognition

频(F0)、音强和时长对情感类别的分类贡献较大。此外,Eyben等人(2010)开创性地提出了openSMILE工具集,集成了MFCC、音高和能量等声学特征的提取方法。

随着深度学习技术的崛起,语音情感识别研究逐步转向声学特征与语义信息的联合表征,利用端到端的深度模型提高情感识别的准确性。例如,Zou等人(2022)充分利用了Transformer模型在时序特征提取方面的优势,结合卷积神经网络(CNN)提取梅尔谱特征、双向LSTM (long short-term memory)模型提取MFCC特征、Transformer提取Wav2Vec2特征,使用共同注意力机制融合多种特征进行情感表达。Ye等人(2023)则指出,传统的时序情感识别模型通常只能感知固定时间尺度的信息,缺乏对长范围上下文的建模能力,提出了时序的双向多尺度网络TIM-Net,通过膨胀因果卷积设计的双向结构来融合过去和未来的互补信息,并通过动态融合模块处理不同尺度的语音信号。Chen等人(2023)进一步提出了处理语音副语言信息的框架SpeechFormer++,该框架通过多层级Transformer建模句子、单词、音素与帧4个粒度的语音关联性,探索了副语言信息在情感表达中的增强作用。总体来看,语音情感识别正朝着多尺度感知、上下文建模与副语言增强等方向不断深化,为更全面、准确的情感理解提供基础。然而,语音情感识别虽然能捕捉语调、节奏等情绪线索,但易受口音、语速和噪声等因素干扰,且难以准确传达复杂情感和语境信息。结合表情、语言等模态能更全面地理解情绪状态,增强识别系统的稳定性与泛化能力。

3.3 语言文本情感识别

语言文本情感识别是指通过计算模型解析语言文本中隐含的情感语义,核心在于构建语言符号系统到情感空间的映射关系,如图4所示,其技术的发展可分为情感词典、传统机器学习方法和深度学习方法3个阶段。在早期,研究主要依赖人工构建的语法规则和情感词典。例如,Esuli和Sebastiani(2006)创建了情感词典SentiWordNet,为每个同义词集合分配了正面、负面和中性的情感分数,这一情感词典广泛应用于基于词汇的情感分析任务,并为后续研究提供了标准化资源。然而,基于情感词典的方法对多义词和隐性情感的处理能力有限,难以准确识别没有明显情感词的句子或存在讽刺、反语等现象的文本。随着传统机器学习方法的应用,情感识别技术开始采用文本特征工程(例如,N-gram、词袋模型、TF-IDF(term frequency-inverse document frequency)等)将文本转换为数值特征向量,再通过机器学习算法进行情感分类。例如,Pang等人(2002)首次将朴素贝叶斯和支持向量机等机器学习算法应用于电影评论情感分类。然而,传统机器学习方法依赖人工设计特征,特征设计过程需要大量领域知识和实验验证,且对长距离依赖和上下文细节的捕捉能力有限,难以有效处理复杂语义和噪声干扰。

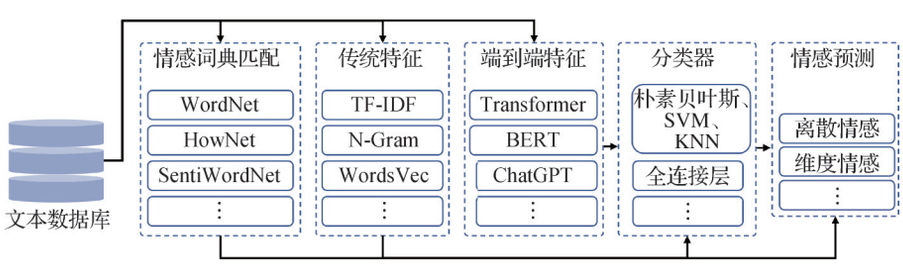


图4 语言文本情感识别框架

Fig. 4 Framework of language emotion recognition

随着深度学习的兴起,情感识别方法开始向更自动化、端到端的特征学习转变。基于深度学习的方法,如卷积神经网络、时序网络以及长短期记忆等,能够自动从原始文本中学习多层次的特征表示,并结合注意力机制等技术进一步捕捉上下文信息和复杂语义。例如,Chen(2015)提出了加权记忆和多重注意力机制,建模目标词与文本之间的远距离交互关系,在多个数据集上取得了当时最优结果。Transformer(Vaswani等,2017)结构的提出极大地推动了自然语言处理(natural language processing, NLP)领域的创新,并为后续的预训练语言模型奠定了基础。基于Transformer的编码器结构,BERT(bidirectional encoder representations from Transformers)模型通过两个预训练任务捕捉深层次的语义信

息(Koroteev，2021)。BERT模型预训练完成后，可以通过少量参数微调应用于各种下游任务(如文本分类、问答、命名实体识别等)，并在多个NLP任务上刷新了性能记录。此后，许多基于BERT及其变体的研究不断优化情感分析模型，如XL-NET(generalized autoregressive pretraining for language understanding)、GPT(generative pretrained Transformer)、LLaMA(large language model meta AI)等，推动情感识别技术向更高精度、更广泛应用的方向发展。然而，语言文本具有高度的抽象性和歧义性，特别是在缺乏语调、表情等非语言线索的情况下，难以准确传达语者的真实情感。讽刺、反语等情绪表达在纯文本中常难以辨识。此外，语言内容可能与实际情绪状态不一致，导致识别偏差。因此，融合语音和表情等多模态信息，有助于弥补单一语言模态的局限，提升情感识别的准确性和鲁棒性。

3.4 基于表情、语音和语言的多模态情感识别

基于表情、语音和语言的多模态情感识别结合多个感官通道的信息，以提高情感识别的准确性和鲁棒性(王顺杰 等，2023；陶建华 等，2024)。如图5所示，现有多模态情感识别研究可分为模态融合和跨模态特征解耦两部分。

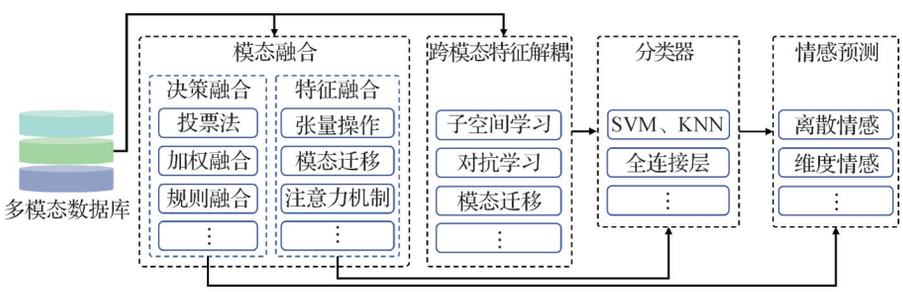


图5 多模态情感识别框架

Fig. 5 Framework of multimodal emotion recognition

3.4.1 多模态融合

在多模态情感识别的早期研究中，模态融合是核心研究方向之一。根据模态融合的阶段，现有方法主要分为特征级融合、决策级融合和混合融合。特征级融合通过整合多个模态的特征，构建统一的情感表示并进行情感预测(Zadeh 等，2017)。决策级融合独立地对每个模态进行情感识别，并在最终决策阶段融合多个模态的预测结果，以推断情绪状态(Gkoumas 等，2021)。混合融合结合了特征级和决策级的优势，在特征表示和决策阶段均进行模态融合，同时也集成了两者的局限性(Sun 等，2021)。具体地，决策级融作用于多个预测向量，对融合策略要求相对较低，但容易导致模型对某一模态过度依赖，限制模态间有效交互。相比之下，特征级融合受到更加广泛的关注，衍生出基于张量、基于图、基于模态迁移以及基于注意力机制的多种融合方法。例如，TFN(tensor fusion network)模型通过扩展单个模态的维度，并计算模态间张量的笛卡尔积，以同步建模模态内和模态间的关联性(Zadeh 等，2017)。然而，该方法的高维度扩展增加了计算复杂度。LMF(low-rank multimodal fusion)模型(Liu 等，2018)和MRRF(multiresolution recurrence fusion network)模型(Barezi 等，2023)采用低秩分解降低计算成本，有效缓解了TFN的计算开销问题。MFN(multimodal fusion network)模型(Zadeh 等，2018a)和MARN(multi-attention recurrent network)模型(Zadeh 等，2018b)由美国卡内基梅隆大学的Paul Pu Liang博士提出，分别通过门控单元和注意力机制，在时间维度上对多模态特征进行整合，以增强时序建模能力。HHMPN(heterogeneous hierarchical message passing network)模型通过构建图节点及其关联性，建立模态特征、特征与标签以及标签间的关系，并利用图传播机制实现模态融合(Zhang 等，2021)。ARGF(adversarial representation graph fusion)模型采用对抗性编解码模型对齐不同模态的分布，以减少模态间差异，同时构建图融合网络整合对齐后的多模态表示(Mai 等，2020)。具体而言，该方法以模态交互表示作为图节点，交互特征的重要性作为边权重，并通过3层图结构完成多模态融合。近年来，Transformer模型的核心机制——自注意力被拓展至跨模态领域，并广泛应用于模态融合。其中，MulT(multimodal Transformer)是较为经典的代表方法，通过

多组 Transformer 结构实现模态的两两交互(Tsai 等,2019)。得益于 Transformer 在建模跨模态长时依赖关系方面的优势,即使模态特征未经过严格的时序对齐,该方法仍能取得良好的融合效果。

上述特征级模态融合通常伴随单模态特征选择,以构建统一的多模态表征。然而,由于不同模态的数据在构成元素、元素分布、噪声程度、情感信息量以及时序等方面存在严重差异,即异构性和异步性,直接进行模态融合往往难以充分建模跨模态交互,导致融合效果受限。

3.4.2 跨模态特征解耦

为缓解上述问题,近期研究者提出了"先交互、再融合"的范式,即在融合前先促进跨模态交互,并解耦出模态间的交互属性,以提高融合的有效性。MISA(modality-invariant and -specific representation)模型首次明确了模态间的交互包含共性和差异性两部分,并设计了双分支结构,从各模态表征中分别提取这两种特征(Hazarika 等,2020)。其中,共性特征是跨模态共享的,具有一定的语义一致性,而差异性特征则刻画了各模态独有的信息。此外,为避免信息损失,MISA 通过组合共性与差异性特征来重构原始模态表示。基于 MISA 的理论框架,近年来涌现出大量跨模态属性解耦的研究,主要可分为基于子空间学习、基于对抗学习和基于迁移学习的方法。TMBL(Transformer-based multimodal binding learning)模型将多模态特征映射到同一子空间,并通过相似性损失、差异性损失和重构损失,引导两个分支分别学习共性和差异性特征,确保解耦后的表征具有良好的判别能力(Huang 等,2024)。DTN(disentanglement Transformer network)模型采用编解码结构解耦出各模态的共性和差异性特征,并进一步引入对抗学习,以减少不同模态间共性特征的分布差异,从而增强跨模态的可对齐性(Zeng 等,2024)。CRNet(co-space representation interaction network)(Shi 等,2024),TCSP(text-centered shared-private network)(Wu 等,2021)和 CJTF(coordinated-joint translation fusion)(Lu 等,2024)分别通过编码器、序列到序列(Seq2Seq)以及 Transformer 等模型实现模态间的迁移,并在迁移过程中评估与目标模态的差异性,实现跨模态属性的解耦。然而,各模态的情感表达是随时间变化的,上述解耦过程固着于静态特征空间,未能构建时变解耦算子以刻画模态随时间演变的重要性。

3.4.3 多模态情感识别结果

在过去的十多年里,多模态情感识别得到了广泛关注,研究者们探索了多种视觉、语音和语言文本特征提取技术。由于涉及隐私保护等问题,尤其是以人为主的视频数据,多数公开数据集仅提供了预先提取的视觉和语言特征,而语言文本信息因不涉及隐私问题,通常能完整公开。早期,语言文本主要依赖于 GloVe(global vectors for word representation)特征表示。然而,随着语言大模型应用的深入,近年来多模态情感识别模型开始嵌入预训练的通用语言模型,并通过微调进行优化,使得文本特征能够端到端地表示,从而显著提升了多模态情感识别的性能(Rahman 等,2020)。

表 3 总结了多模态情感识别领域部分代表性工作在常用的情感识别数据库 CMU-MOSI 和 CMU-MOSEI 上的结果。为公平对比各个方法,在 CMU-MOSI、CMU-MOSEI 上的对比方法均采用微调的 BERT 模型提取语言文本特征。早期的特征级融合方法侧重于整合多模态特征。例如,TFN 通过计算模态间的笛卡尔积建模交互,在 CMU-MOSI 上取得了 82.6% 的 F1 分数。为了降低计算复杂度,LMF 采用低秩近似,提升了性能并减少了计算开销。MFN 通过门控机制关注时序信息,在 CMU-MOSI 上达到了 83.5% 的 F1 分数。近年来,基于 Transformer 的模型因其强大的长时依赖建模能力受到关注。MulT 通过跨模态注意力机制在 CMU-MOSI 和 CMU-MOSEI 上分别达到 83.7% 和 84.6% 的 F1 分数,证明了 Transformer 在跨模态融合任务中的有效性。随后,GFN(graph fusion network)、MAG(multimodal adaptation gate)、TFR-Net(Transformer-based feature reconstruction network)和 MCL(multimodal contrastive learning)等模型通过引入图网络和复杂的融合策略,进一步提升情感识别性能。其中,MCL 模型性能较为优越,在 CMU-MOSI 上达到了 86.1% 的 F1 分数,在 CMU-MOSEI 上为 86.2%。

考虑到直接融合异构模态的挑战,"先交互、再融合"的范式取得了显著成果。MISA 通过区分模态共有和特有信息,在 CMU-MOSI 和 CMU-MOSEI 上分别获得 83.6% 和 85.3% 的 F1 分数。基于 MISA 的思路,TMBL 和 DTN 等模型通过解耦策略进一步提升了性能。TMBL 在 CMU-MOSI 和 CMU-MOSEI 上分

表3 多模态情感识别结果

Table 3 Multimodal emotion recognition results

	CMU-MOSI					CMU-MOSEI				
	Acc7 (↑)/%	Acc2 (↑)/%	F1 (↑)/%	MAE (↓)	Corr (↑)	Acc7 (↑)/%	Acc2 (↑)/%	F1 (↑)/%	MAE (↓)	Corr (↑)
TFN(Zadeh 等,2017)	44.7	82.6	82.6	0.761	0.789	51.8	84.5	84.5	0.622	0.781
LMF(Liu 等,2018)	45.1	84.0	84.0	0.742	0.785	51.2	84.2	84.3	0.612	0.779
MFN(Zadeh 等,2018a)	44.1	83.5	83.5	0.759	0.786	52.6	84.8	84.8	0.607	0.771
MulT(Tsai 等,2019)	41.5	83.7	83.7	0.767	0.799	50.7	84.7	84.6	0.625	0.775
GFN(Mai 等,2020)	47.0	84.3	84.3	0.736	0.790	51.8	85.0	85.0	0.611	0.774
ICCN(Sun 等,2020)	39.0	83.0	83.0	0.860	0.710	51.6	84.2	84.2	0.565	0.713
MAG(Rahman 等,2020)	42.9	83.5	83.5	0.790	0.769	51.9	85.0	85.0	0.602	0.778
MISA(Hazarika 等,2020)	42.3	83.4	83.6	0.783	0.761	52.2	85.5	85.3	0.555	0.756
TFR-Net(Yuan 等,2021)	42.6	84.0	83.9	0.787	0.788	51.7	85.2	85.1	0.606	0.781
HyCon(Mai 等,2023b)	46.6	85.2	85.1	0.713	0.790	52.8	85.4	85.6	0.601	0.776
MCL(Mai 等,2023a)	49.2	86.1	86.1	0.713	0.793	53.3	86.2	86.2	0.581	0.791
MHE(Gao 等,2024a)	41.5	83.6	83.5	0.801	0.722	52.5	84.2	84.0	0.577	0.712
CMHFM(Wang 等,2024a)	37.0	81.0	81.3	0.912	0.677	52.6	84.0	83.6	0.558	0.731
CRNet(Shi 等,2024)	47.4	86.4	86.4	0.712	0.797	53.8	86.2	86.1	0.541	0.771
TMBL(Huang 等,2024)	36.3	83.8	84.3	0.867	0.762	52.4	85.8	85.9	0.545	0.766
DTN(Zeng 等,2024)	48.1	86.2	86.2	0.714	0.807	52.5	86.3	86.3	0.579	0.788
MMCL(a)(Wang 等,2025a)	50.4	87.3	87.3	0.672	0.817	54.7	86.5	86.5	0.528	0.799
MMCL(c)(Wang 等,2025a)	49.9	86.7	86.7	0.681	0.811	54.5	86.6	86.5	0.528	0.794
Co-SA(a)(Wang 等,2024b)	49.8	87.2	87.0	0.688	0.812	54.3	86.8	86.7	0.535	0.791
Co-SA(c)(Wang 等,2024b)	49.5	86.6	86.5	0.688	0.814	54.0	86.4	86.4	0.535	0.790

注:Acc7为7类情感识别准确率,Acc2为积极情绪和消极情绪识别准确率,F1为积极情绪和消极情绪的F1分数,MAE为真实情感与预测情感的平均绝对误差,Corr为真实情感和预测情感的相关性;↑表示值越高越好,↓表示值越低越好。

别取得84.3%和85.9%的F1分数,而DTN则通过对抗学习增强特征一致性,CMU-MOSI和CMU-MOSEI上的F1分数分别为86.2%和86.3%。CRNet基于模态迁移思想,在CMU-MOSI和CMU-MOSEI上也取得了优异成绩,F1分数分别为86.4%和86.1%。这些结果验证了跨模态特征解耦策略的有效性,普遍优于早期的直接融合方法。

表3数据显示,MMCL(multi-modality collaborative learning)(Wang 等,2025a)和Co-SA(cooperative sentiment agent)(Wang 等,2024b)在CMU-MOSI和CMU-MOSEI数据集上取得了最优的情感识别性能。MMCL基于特征解耦,通过模态内注意力机制和强化学习挖掘跨模态增强互补的情感特征,显著提升了判别能力。在CMU-MOSI数据集上,以简单加权的特征组合MMCL(a)实现了87.3%的二分类准确率和F1分数,以简单拼接的特征组合MMCL(c)在CMU-MOSI上取得86.7%的二分类准确率和F1分数。在CMU-MOSEI数据集上,MMCL(a)达到86.5%的二分类准确率和F1分数,MMCL(c)则为86.6%的二分类准确率和86.5%的F1分数。Co-SA引入协同情感智能体框架,精细建模情感时序动态并通过智能体协同增强模态间互补性,在CMU-MOSI数据集上,Co-SA(a)获得87.2%的正负情感识别准确率和87.0%的F1分数,在CMU-MOSEI数据集上分别为54.3%的Acc7和86.7%的F1分数。这些方法的成功表明,结合模态内信息增强、模态间

互补学习和创新交互范式可以有效提升多模态情感识别性能。

3.4.4 情感识别大模型

多模态大模型凭借其强大的语言理解和生成能力，在情感识别与情感推理等领域展现出独特优势。当前的主流研究主要集中在3个方面：1）通用大模型在情感识别中的表现评估；2）针对特定情绪分析任务的大模型优化；3）情感智能推理与可解释性分析。

尽管通用大模型（如GPT-4、LLaMA等）在语言生成与知识推理等任务中表现优异，但多项研究揭示了这些大模型在情感分析中的不足。例如，Lian等人（2024c）对GPT-4V在21个基准数据集上的表现进行了评估，结果显示其在视觉情感理解中表现强劲，但在微表情识别等需要专业知识的任务中却存在不足。类似的研究还发现，尽管GPT-4的情感智商超过了大部分人类参与者，但其情感理解机制与人类认知存在差异，可能通过非人类方式实现类似人类的情感智能（Wang等，2023b）。此外，Chen等人（2024b）的研究也表明，通用大模型在复杂情感场景中的表现仍有限，提示了其在实际应用中的改进空间。

为弥补通用大模型在情感分析任务中的局限性，研究者们通过构建高质量的数据集（如EMER-Coarse和MERR）以及针对性的模型优化策略，构建情感分析专用大模型。具体而言，面向情感识别任务的大模型优化主要集中在3个方面：1）构建具备情感特性的模型结构；2）开发情感导向的训练策略；3）设计专用的提示技术。例如，Cheng等人（2024）提出的Emotion-LLaMA模型引入了情感特定编码器，融合音频、视觉与文本信息，并在此基础上对大模型进行微调；Chen等人（2024a）设计了情感引导的掩码机制与分层自监督学习方法，以增强模型对情感特征的感知能力；Yang等人（2024）提出的EmoLLM模型则通过多视角视觉投影和情感提示机制，在多模态情感理解任务中取得了显著性能提升。这些优化措施基于对通用大模型在情绪理解能力的系统评估，从而构建起“评估—优化”的良性闭环，持续推动多模态情感识别的发展。

随着大模型在情感分析中的广泛应用，情感推理能力和决策透明度逐渐成为研究热点。Lan等人（2025）提出的ExpLLM模型，通过链式思维机制，从面部动作单元分析情绪成因，并生成详细的情绪解释，在微表情识别任务中表现突出。Lian等人（2024b）提出的AffectGPT框架，通过粗粒度和细粒度数据集的两阶段训练，增强了多模态情绪识别的可解释性，生成情绪预测并提供相应解释，增强了模型的可信度。此外，Yang等人（2023）在EmoSet数据集中加入可解释的视觉特征，使得模型决策依据更加透明。一些研究还探索了情绪推理与认知科学的结合，如将Ekman的情绪理论编码进模型架构中，提升了情绪场景的解释性（Li等，2023；Lian等，2024a）。

综上，随着大模型在情感识别中的应用不断深入，研究涵盖了情感识别任务的评估、模型优化和情感智能推理等多个方向。这些研究不仅推动了情感计算领域的进展，也为构建具备情感智能的人工系统提供了理论与技术支持。未来，随着大模型技术的进一步成熟，情感识别有望在心理健康支持、人机交互等领域发挥更加重要的作用。

4 情感识别应用

1）安全驾驶。无感式情感识别在安全驾驶领域具有广阔的应用前景。通过分析驾驶员的面部表情、语音语调和肢体动作，系统能够实时监测驾驶员的情绪状态，如疲劳、焦虑或愤怒。一旦检测到驾驶员处于不适宜驾驶的状态，系统可以及时发出警告，甚至主动干预，例如调整车速、切换驾驶模式或提供休息建议，从而有效降低交通事故的发生率，提升驾驶安全性。

2）智慧医疗。在智慧医疗领域，无感式情感识别技术为患者提供个性化、主动式的关怀。通过监测患者的面部表情、语音和语言信号，系统能够评估患者的情绪状态和心理健康状况，例如抑郁、焦虑或疼痛程度。这些信息可以帮助医生更全面地了解患者的需求，制定更精准的治疗方案。此外，该技术还可以应用于远程心理咨询、精神疾病辅助诊断和术后康复监测等方面，提升医疗服务的效率和质量。

3）智能客服。无感式情感识别技术能够显著提升智能客服的智能化和人性化水平。通过分析用户在语音或文本交流中的情感倾向，系统可以更好地理解用户的需求和情感诉求，从而提供更贴心的服务。例如，当检测到用户表达不满或愤怒情绪时，系

统可以自动将服务转接给人工客服，或者主动提供解决方案，有效提升用户满意度，降低客户流失率。

4）公共安全。在公共安全领域，无感式情感识别技术可用于监测公共场所人群的情绪状态，及时发现潜在的安全隐患。通过分析视频监控中的面部表情和肢体动作，以及语音数据中的语调和内容，系统可以识别出异常行为和高危人群，例如正在发生争执或处于高度紧张状态的人群。这有助于安保人员及时采取干预措施，预防突发事件的发生，维护社会稳定。

5 挑战与展望

伴随信息技术的进步，人们对计算机的期望已不再局限于功能实现，情感层面的需求日益凸显。深度学习的突破性进展，为满足这一需求提供了强大的技术支撑。在过去十年间，情感识别已成为人机交互、计算机视觉、语音信号处理和自然语言处理等领域的核心研究方向之一。情感识别是一项交叉学科研究，尤其与应用心理学和计算机科学紧密相关。由于计算机在情感理解方面与人类存在显著差异，该领域的研究充满挑战与机遇。尽管近年来语音情感识别技术取得了长足进步，但在情感语料库构建、情感识别模型设计等方面仍存在诸多值得深入探索的问题。

1）情感语料库构建。情感语料库是情感识别研究的基础，然而，现有的情感数据库仍然面临着诸多挑战。首先，情感表达本身具有复杂性，单一的基础类别难以精确捕捉人类丰富的情感状态，因此需要借助维度情感模型进行更细粒度的描述。其次，情绪认知具有主观性，情感语料库的标注过程不可避免地受到标注者个人偏好的影响，导致数据库本身存在偏差。为解决这一问题，需要结合心理学领域的知识，设计更加客观的情感评价指标，以减少主观因素的干扰。此外，神经心理学研究表明，环境上下文对情感认知具有重要作用，而目前大多数情感语料库为了简化问题，往往去除了环境上下文信息，这在一定程度上限制了情感识别模型的性能。因此，如何构建能够反映真实情感表达，同时兼顾客观性和上下文信息的情感语料库，是当前研究面临的重要课题。

2）情感识别模型设计。情感识别模型的设计是无感式情感识别的核心环节，但现有的情感识别模型仍然面临着多重挑战。首先，由于情感表达具有个性化特征，不同个体的情绪表达习惯存在差异，难以使用统一的标准模型来准确识别所有人的情感。因此，如何结合个体的情绪表达习惯，例如特定的面部肌肉运动模式、语音语调特征以及词汇使用偏好等，建立个性化的情感评估模型，是一个亟待解决的问题。其次，真实场景中存在各种遮挡和光照变化等干扰因素，给情感识别带来了很大的困难。如何设计具有鲁棒性的情感识别模型，使其能够有效应对这些干扰因素，是模型走向实用化的关键。此外，如何将环境上下文信息融入情感推断模型，也是一个重要的研究方向。神经心理学研究表明，环境上下文对情感认知具有重要作用，将上下文信息纳入考虑有助于提高情感识别的准确性。最后，人们可能会出于某种目的伪装行为信号，这给情感识别带来了很大的挑战。如何从伪装信号中正确识别情感，例如通过微表情识别等手段，是提高情感识别可靠性的重要途径。

6 结 语

在人工智能领域，如何让机器具备感知和理解人类情感的能力，已成为重要的研究课题。与需要用户主动参与的情感识别方法不同，无感式情感识别强调在用户无感知的情况下进行情感识别，具有更广泛的应用前景。近年来，无感式情感识别技术取得了快速发展，特征提取方法从传统特征向深度特征转变，识别模型从传统机器学习算法向深度神经网络演进，情感表达方式也从离散类别向连续维度发展。

本文对基于表情、语音和语言的无感式情感识别领域研究进展进行了系统性综述。首先探讨了离散和维度情感模型等关键理论基础，随后对常用的无感式情感数据库进行归纳整理，并重点介绍基于面部表情、语音、语言和多个模态的无感式情感识别技术。此外，本文还介绍了无感式情感识别技术的应用场景，并对该领域面临的挑战与机遇进行了深入分析。

参考文献(References)

Barezi E J, Momeni P and Fung P. 2023. Modality-based factorization for multimodal fusion [EB/OL]. [2025-04-08]. https://arxiv.org/pdf/1811.12624.pdf

Bhangale K and Kothandaraman M. 2023. Speech emotion recognition based on multiple acoustic features and deep convolutional neural network. Electronics, 12(4): #839 [DOI: 10.3390/electronics12040839]

Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W F and Weiss B. 2005. A database of German emotional speech//Proceedings of Interspeech. Lisbon, Portugal: ISCA: 1517-1520 [DOI: 10.21437/Interspeech.2005-446]

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

Cai J, Meng Z B, Khan A S, Li Z Y, O'Reilly J and Tong Y. 2018. Island loss for learning discriminative features in facial expression recognition//Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018). Xi'an, China: IEEE: 302-309 [DOI: 10.1109/FG.2018.00051]

Chen T S, Pu T, Wu H F, Xie Y, Liu L B and Lin L. 2022. Cross-domain facial expression recognition: a unified evaluation benchmark and adversarial graph learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12): 9887-9903 [DOI: 10.1109/TPAMI.2021.3131222]

Chen W D, Xing X F, Chen P H and Xu X M. 2024a. Vesper: a compact and effective pretrained model for speech emotion recognition. IEEE Transactions on Affective Computing, 15(3): 1711-1724 [DOI: 10.1109/TAFFC.2024.3369726]

Chen W D, Xing X F, Xu X M, Pang J X and Du L. 2023. SpeechFormer++: a hierarchical efficient framework for paralinguistic speech processing. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31: 775-788 [DOI: 10.1109/TASLP.2023.3235194]

Chen Y H. 2015. Convolutional Neural Network for Sentence Classification. Waterloo, Canada: University of Waterloo

Chen Y Y, Wang H, Yan S Z, Liu S J, Li Y Z, Zhao Y and Xiao Y H. 2024b. EmotionQueen: a benchmark for evaluating empathy of large language models [EB/OL]. [2025-04-08]. https://arxiv.org/pdf/2409.13359.pdf

Cheng X Y, Chen Y, Cheng B X, Li S S and Zhou G D. 2017. An emotion cause corpus for Chinese microblogs with multiple-user structures. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 17(1): #6 [DOI: 10.1145/3132684]

Cheng Z B, Cheng Z Q, He J Y, Sun J D, Wang K, Lin Y X, Lian Z, Peng X J and Hauptmann A. 2024. Emotion-LLaMA: multimodal emotion recognition and reasoning with instruction tuning [EB/OL]. [2025-04-08]. https://arxiv.org/pdf/2406.11161.pdf

China Association for Science and Technology. 2024. CAST Releases 2024 Major Scientific Issues, Engineering Challenges, and Industrial Technology Problems (中国科协. 2024. 重大科学问题、工程技术难题和产业技术问题)

Cui X Y, He C, Zhao H K and Wang M L. 2024. Combining ViT with contrastive learning for facial expression recognition. Journal of Image and Graphics, 29(1): 123-133 (崔鑫宇, 何翀, 赵宏珂, 王美丽. 2024. 融合ViT与对比学习的面部表情识别. 中国图象图形学报, 29(1): 123-133) [DOI: 10.11834/jig.230043]

Dhall A, Goecke R, Lucey S and Gedeon T. 2011. Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark//Proceedings of 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). Barcelona, Spain: IEEE: 2106-2112 [DOI: 10.1109/ICCVW.2011.6130508]

Ding H, Zhou S K and Chellappa R. 2017. FaceNet2ExpNet: regularizing a deep face recognition net for expression recognition//Proceedings of the 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017). Washington, USA: IEEE: 118-126 [DOI: 10.1109/FG.2017.23]

Eddy S R. 1996. Hidden Markov models. Current Opinion in Structural Biology, 6(3): 361-365 [DOI: 10.1016/S0959-440X(96)80056-X]

Ekman P and Friesen W V. 1971. Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2): 124-129 [DOI: 10.1037/h0030377]

Esuli A and Sebastiani F. 2006. SentiWordNet: a publicly available lexical resource for opinion mining//LREC. France: ELRA: 417-422

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]

Gao Y, Fu Y Q, Sun M and Gao F. 2024a. Multi-modal hierarchical empathetic framework for social robots with affective body control. IEEE Transactions on Affective Computing, 15(3): 1621-1633 [DOI: 10.1109/TAFFC.2024.3356511]

Gao Y F, Xie Y H, Hu Z Z, Chen T S and Lin L. 2024b. Adaptive global-local representation learning and selection for cross-domain facial expression recognition. IEEE Transactions on Multimedia, 26: 6676-6688 [DOI: 10.1109/TMM.2024.3355637]

Ghimire D and Lee J. 2013. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. Sensors, 13(6): 7714-7734 [DOI: 10.3390/s130607714]

Gkoumas D, Li Q C, Dehdashti S, Melucci M, Yu Y J and Song D W. 2021. Quantum cognitively motivated decision fusion for video sen-

timent analysis//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtually: AAAI: 827-835 [DOI: 10.1609/aaai.v35i1.16165]

Goodfellow I J, Erhan D, Carrier P L, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y C, Thaler D, Lee D H, Zhou Y B, Ramaiah C, Feng F X, Li R F, Wang X J, Athanasakis D, Shawe-Taylor J, Milakov M, Park J, Ionescu R, Popescu M, Grozea C, Bergstra J, Xie J J, Romaszko L, Xu B, Chuang Z and Bengio Y. 2013. Challenges in representation learning: a report on three machine learning contests//Proceedings of the 20th International Conference on Neural Information Processing. Daegu, Korea (South): Springer: 117-124 [DOI: 10.1007/978-3-642-42051-1_16]

Grossberg S. 2013. Recurrent neural networks. Scholarpedia, 8(2): #1888 [DOI: 10.4249/scholarpedia.1888]

Gu W F, Xiang C, Venkatesh Y V, Huang D and Lin H. 2012. Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. Pattern Recognition, 45(1): 80-91 [DOI: 10.1016/j.patcog.2011.05.006]

Hazarika D, Zimmermann R and Poria S. 2020. MISA: modality-invariant and -specific representations for multimodal sentiment analysis//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 1122-1131 [DOI: 10.1145/3394171.3413678]

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

Huang J, Tao J H, Liu B and Lian Z. 2020. Learning utterance-level representations with label smoothing for speech emotion recognition//Proceedings of Interspeech. Shanghai, China: ISCA: 4079-4083 [DOI: 10.21437/Interspeech.2020-1391]

Huang J H, Zhou J, Tang Z C, Lin J Y and Chen C Y C. 2024. TMBL: transformer-based multimodal binding learning model for multimodal sentiment analysis. Knowledge-Based Systems, 285: #111346 [DOI: 10.1016/j.knosys.2023.111346]

Izard C E. 1977. Differential emotions theory//Izard C E, ed. Human Emotions. New York, USA: Springer: 43-66 [DOI: 10.1007/978-1-4899-2209-0_3]

Izard C E. 2013. Human Emotions. New York, USA: Springer

Ji Y L, Hu Y H, Yang Y and Shen H T. 2023. Region attention enhanced unsupervised cross-domain facial emotion recognition. IEEE Transactions on Knowledge and Data Engineering, 35(4): 4190-4201 [DOI: 10.1109/TKDE.2021.3136606]

Jiang X X, Zong Y, Zheng W M, Tang C G, Xia W C, Lu C and Liu J T. 2020. DFEW: a large-scale database for recognizing dynamic facial expressions in the wild//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 2881-2889 [DOI: 10.1145/3394171.3413620]

Koroteev M V. 2021. BERT: a review of applications in natural language processing and understanding [EB/OL]. [2025-04-08]. https://arxiv.org/pdf/2103.11943.pdf

Lan X, Xue J, Qi J, Jiang D M, Lu K and Chua T S. 2025. ExpLLM: towards chain of thought for facial expression recognition. IEEE Transactions on Multimedia [DOI: 10.1109/TMM.2025.3557704]

Lee I, Lee E and Yoo S B. 2023. Latent-OFER: detect, mask, and reconstruct with latent vectors for occluded facial expression recognition//Proceedings IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 1536-1546 [DOI: 10.1109/ICCV51070.2023.00148]

Li C, Wang J D, Zhang Y X, Zhu K J, Hou W X, Lian J X, Lou F, Yang Q and Xie X. 2023. Large language models understand and can be enhanced by emotional stimuli [EB/OL]. [2025-04-08]. https://arxiv.org/pdf/2307.11760.pdf

Li H T, Niu H J, Zhu Z Q and Zhao F. 2024. CLIPER: a unified vision-language framework for in-the-wild facial expression recognition//Proceedings of 2024 IEEE International Conference on Multimedia and Expo (ICME). Niagara Falls, Canada: IEEE: 1-6 [DOI: 10.1109/ICME57554.2024.10687508]

Li S and Deng W H. 2022. A deeper look at facial expression dataset bias. IEEE Transactions on Affective Computing, 13(2): 881-893 [DOI: 10.1109/TAFFC.2020.2973158]

Li S, Deng W H and Du J P. 2017a. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild//Proceedings IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2584-2593 [DOI: 10.1109/CVPR.2017.277]

Li Y, Tao J H, Chao L L, Bao W and Liu Y Z. 2017b. CHEAVD: a Chinese natural emotional audio-visual database. Journal of Ambient Intelligence and Humanized Computing, 8(6): 913-924 [DOI: 10.1007/s12652-016-0406-z]

Li Y, Zeng J B, Shan S G and Chen X L. 2019. Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Transactions on Image Processing, 28(5): 2439-2450 [DOI: 10.1109/TIP.2018.2886767]

Li Y J, Gao Y N, Chen B Z, Zhang Z, Zhu L and Lu G M. 2021. JDMAN: joint discriminative and mutual adaptation networks for cross-domain facial expression recognition//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event: ACM: 3312-3320 [DOI: 10.1145/3474085.3475484]

Lian Z, Sun H Y, Sun L C, Gu H, Wen Z F, Zhang S Y, Chen S, Xu M Y, Xu K, Chen K, Chen L, Liang S, Li Y, Yi J Y, Liu B and Tao J H. 2024a. Explainable multimodal emotion recognition [EB/OL]. [2025-04-08]. https://arxiv.org/pdf/2306.15401.pdf

Lian Z, Sun H Y, Sun L C, Yi J Y, Liu B and Tao J H. 2024b. AffectGPT: dataset and framework for explainable multimodal emotion recognition [EB/OL]. [2025-04-08]. https://arxiv.org/pdf/2407.07653.pdf

Lian Z, Sun L C, Sun H Y, Chen K, Wen Z F, Gu H, Liu B and Tao J H. 2024c. GPT-4V with emotion: a zero-shot benchmark for generalized emotion recognition. Information Fusion, 108: #102367 [DOI: 10.1016/j.inffus.2024.102367]

Lin S Y. 2024. Text emotional analysis in natural language processing. Applied and Computational Engineering, 36: 163-172 [DOI: 10.54254/2755-2721/36/20230440]

Liu Z, Shen Y, Lakshminarasimhan V B, Liang P P, Zadeh A and Morency L P. 2018. Efficient low-rank multimodal fusion with modality-specific factors [EB/OL]. [2025-04-08]. https://arxiv.org/pdf/1806.00064.pdf

Livingstone S R and Russo F A. 2018. The Ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. PLoS One, 13(5): #e0196391 [DOI: 10.1371/journal.pone.0196391]

Lotfian R and Busso C. 2019. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. IEEE Transactions on Affective Computing, 10(4): 471-483 [DOI: 10.1109/TAFFC.2017.2736999]

Lu L D, Xia H Y, Tan Y M and Song S X. 2024. Attention-guided local feature joint learning for facial expression recognition. Journal of Image and Graphics, 29(8): 2377-2387 (卢莉丹, 夏海英, 谭玉枚, 宋树祥. 2024. 注意力引导局部特征联合学习的人脸表情识别. 中国图象图形学报, 29(8): 2377-2387) [DOI: 10.11834/jig.230410]

Lu Q, Sun X, Gao Z Z Z, Long Y F, Feng J and Zhang H. 2024. Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis. Information Processing and Management, 61(1): #103538 [DOI: 10.1016/j.ipm.2023.103538]

Lucey P, Cohn J F, Kanade T, Saragih J, Ambadar Z and Matthews I. 2010. The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. San Francisco, USA: IEEE: 94-101 [DOI: 10.1109/CVPRW.2010.5543262]

Maas A L, Daly R E, Pham P T, Huang D, Ng A Y and Potts C. 2011. Learning word vectors for sentiment analysis//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, USA: ACL: 142-150

Mai S, Sun Y, Zeng Y and Hu H F. 2023a. Excavating multimodal correlation for representation learning. Information Fusion, 91: 542-555 [DOI: 10.1016/j.inffus.2022.11.003]

Mai S, Zeng Y, Zheng S J and Hu H F. 2023b. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. IEEE Transactions on Affective Computing, 14(3): 2276-2289 [DOI: 10.1109/TAFFC.2022.3172360]

Mai S J, Hu H F and Xing S L. 2020. Modality to modality translation: an adversarial representation learning and graph fusion network for multimodal fusion//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 164-172 [DOI: 10.1609/aaai.v34i01.5347]

Martínez-Cámara E, Martín-Valdivia M T, Ureña-López L A and Montejo-Ráez A R. 2014. Sentiment analysis in twitter. Natural Language Engineering, 20(1): 1-28 [DOI: 10.1017/S1351324912000332]

Minsky M. 1988. Society of mind. Simon and Schuster

Mollahosseini A, Hasani B and Mahoor M H. 2019. Affectnet: a database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1): 18-31 [DOI: 10.1109/TAFFC.2017.2740923]

Ojala T, Pietikainen M and Maenpaa T. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7): 971-987 [DOI: 10.1109/TPAMI.2002.1017623]

Pang B, Lee L and Vaithyanathan S. 2002. Thumbs up? Sentiment classification using machine learning techniques//Proceedings Conference on Empirical Methods in Natural Language Processing. Philadelphia, USA: ACL: 79-86 [DOI: 10.3115/1118693.1118704]

Pantic M, Valstar M, Rademaker R and Maat L. 2005. Web-based database for facial expression analysis//Proceedings of 2005 IEEE International Conference on Multimedia and Expo. Amsterdam, the Netherlands: IEEE: #5 [DOI: 10.1109/ICME.2005.1521424]

Peng X L, Huang J and Tian F. 2024. Emotional challenge and its prospective applications in metaverse. Journal of Image and Graphics, 29(10): 2926-2936 (彭晓兰, 黄进, 田丰. 2024. 情感挑战研究进展及其在元宇宙中的应用前景. 中国图象图形学报, 29(10): 2926-2936) [DOI: 10.11834/jig.230734]

Picard R W. 2000. Affective Computing. Cambridge, USA: MIT Press

Plutchik R. 2001. The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American Scientist, 89(4): 344-350

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 [EB/OL]. [2025-04-08]. https://arxiv.org/pdf/1810.02508.pdf

Rahman W, Hasan M K, Lee S Zadeh A B, Mao C F, Morency L P and Hoque E. 2020. Integrating multimodal information in large pretrained transformers//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: ACL: 2359-2369 [DOI: 10.18653/v1/2020.acl-main.214]

Rane A and Kumar A. 2018. Sentiment classification system of twitter data for US airline service analysis//Proceedings of the 42nd IEEE Annual Computer Software and Applications Conference (COMPSAC). Tokyo, Japan: IEEE: 769-773 [DOI: 10.1109/

COMPSAC.2018.00114］

Russell J A and Mehrabian A. 1977. Evidence for a three-factor theory of emotions. Journal of Research in Personality, 11(3): 273-294［DOI: 10.1016/0092-6566(77)90037-X］

Russell J A. 1980. A circumplex model of affect. Journal of Personality and Social Psychology, 39(6): 1161-1178［DOI: 10.1037/h0077714］

Schuller B, Rigoll G and Lang M. 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture//Proceedings of 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing. Montreal, Canada: IEEE: I-577-I-580［DOI: 10.1109/ICASSP.2004.1326051］

Shi H, Pu Y Y, Zhao Z P, Huang J, Zhou D M, Xu D and Cao J D. 2024. Co-space representation interaction network for multimodal sentiment analysis. Knowledge-Based Systems, 283: #111149［DOI: 10.1016/j.knosys.2023.111149］

Socher R, Perelygin A, Wu J, Chuang J, Manning C D, Ng A Y and Potts C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank//Proceedings Conference on Empirical Methods in Natural Language Processing. Seattle, USA: ACL: 1631-1642

Sun J G, Yin H Q, Tian Y, Wu J P, Shen L S and Chen L. 2021. Two-level multimodal fusion for sentiment analysis in public security. Security and Communication Networks, 2021: #6662337［DOI: 10.1155/2021/6662337］

Sun Z K, Sarma P, Sethares W and Liang Y Y. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 8992-8999［DOI: 10.1609/aaai.v34i05.6431］

Suthaharan S. 2016. Support vector machine//Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. New York, USA: Springer: 207-235［DOI: 10.1007/978-1-4899-7641-3_9］

Tan S and Zhang J. 2008. An empirical study of sentiment analysis for Chinese documents. Expert Systems with Applications, 34(4): 2622-2629［DOI: 10.1016/j.eswa.2007.05.028］

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(6): 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(6): 1607-1627）［DOI: 10.11834/jig.240017］

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

Tomkins S S. 1963. Affect Imagery Consciousness: Volume II: The Negative Affects. New York, USA: Springer

Tsai Y H H, Bai S J, Liang P P, Kolter J Z, Morency L P and Salakhutdinov R. 2019. Multimodal transformer for unaligned multimodal language sequences//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL: 6558-6569［DOI: 10.18653/v1/p19-1656］

Uono S, Sato W, Kochiyama T, Sawada R, Kubota Y, Yoshimura S and Toichi M. 2017. Neural substrates of the ability to recognize facial expressions: a voxel-based morphometry study. Social Cognitive and Affective Neuroscience, 12(3): 487-495［DOI: 10.1093/scan/nsw142］

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang L, Peng J J, Zheng C Z, Zhao T and Zhu L. 2024a. A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. Information Processing and Management, 61(3): #103675［DOI: 10.1016/j.ipm.2024.103675］

Wang S J, Cai G Y, Lyu G R and Tang W B. 2023. Aspect-level multimodal co-attention graph convolutional sentiment analysis model. Journal of Image and Graphics, 28(12): 3838-3854（王顺杰，蔡国永，吕光瑞，唐炜博. 2023. 方面级多模态协同注意图卷积情感分析模型. 中国图象图形学报，28(12): 3838-3854）［DOI: 10.11834/jig.221015］

Wang S M, Liu C G and Liu Q S. 2025a. Multi-modality collaborative learning for sentiment analysis［EB/OL］.［2025-04-08］. https://arxiv.org/pdf/2501.12424.pdf

Wang S M and Liu Q S. 2025b. Unsupervised cross-domain facial expression recognition via class adaptive self-training. IEEE Transactions on Affective Computing［DOI: 10.1109/TAFFC.2025.3529947］

Wang S M, Shuai H, Liu C G and Liu Q S. 2023a. Bias-based soft label learning for facial expression recognition. IEEE Transactions on Affective Computing, 14(4): 3257-3268［DOI: 10.1109/TAFFC.2022.3220291］

Wang S M, Shuai H and Liu Q S. 2022. Phase space reconstruction driven spatio-temporal feature learning for dynamic facial expression recognition. IEEE Transactions on Affective Computing, 13(3): 1466-1476［DOI: 10.1109/TAFFC.2020.3007531］

Wang S M, Shuai H, Liu Q S and Wang F. 2024b. Cooperative sentiment agents for multimodal sentiment analysis［EB/OL］.［2025-04-08］. https://arxiv.org/pdf/2404.12642.pdf

Wang S M, Shuai H, Zhu L and Liu Q S. 2024c. Expression complementary disentanglement network for facial expression recognition. Chinese Journal of Electronics, 33(3): 742-752［DOI: 10.23919/cje.2022.00.351］

Wang X N, Li X T, Yin Z, Wu Y and Liu J. 2023b. Emotional intelligence of large language models. Journal of Pacific Rim Psychology, 17: 1-12［DOI: 10.1177/18344909231213958］

Wang Y J, He J, Zhang J X, Sun R H and Liu X L. 2025. Semi-

supervised facial expression recognition robust to head pose empowered by dual consistency constraints. Journal of Image and Graphics, 30(2): 435-450 (王宇键, 何军, 张建勋, 孙仁浩, 刘学亮. 2025. 头姿鲁棒的双一致性约束半监督表情识别. 中国图象图形学报, 30(2): 435-450) [DOI: 10.11834/jig.240205]

Wu Y, Lin Z J, Zhao Y Y, Qin B and Zhu L N. 2021. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis//Findings of the Association for Computational Linguistics: ACL-IJCNLP. Online: ACL: 4730-4738 [DOI: 10.18653/v1/2021.findings-acl.417]

Xu R F, Zhou Y, Wu D Y, Gui L, Du J C and Xue Y. 2016. Overview of NLPCC shared task 4: stance detection in Chinese Microblogs// Proceedings of the Natural Language Understanding and Intelligent Applications: the 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and the 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016. Kunming, China: Springer: 907-916 [DOI: 10.1007/978-3-319-50496-4_85]

Yang H Y, Ciftci U and Yin L J. 2018. Facial expression recognition by de-expression residue learning//Proceedings IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2168-2177 [DOI: 10.1109/CVPR.2018.00231]

Yang J Y, Huang Q R, Ding T T, Lischinski D, Cohen-Or D and Huang H. 2023. EmoSet: a large-scale visual emotion dataset with rich attributes//Proceedings IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 20326-20337 [DOI: 10.1109/ICCV51070.2023.01864]

Yang Q, Ye M and Du B. 2024. EmoLLM: multimodal emotional understanding meets large language models [EB/OL]. [2025-04-08]. https://arxiv.org/pdf/2406.16442.pdf

Ye J X, Wen X C, Wei Y J, Xu Y, Liu K H and Shan H M. 2023. Temporal modeling matters: a novel temporal emotional modeling approach for speech emotion recognition//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE: 1-5 [DOI: 10.1109/ICASSP49357.2023.10096370]

Yuan Z Q, Li W, Xu H and Yu W M. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis// Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event: ACM: 4400-4407 [DOI: 10.1145/3474085.3475585]

Zadeh A, Chen M H, Poria S, Cambria E and Morency L P. 2017. Tensor fusion network for multimodal sentiment analysis [EB/OL]. [2025-04-08]. https://arxiv.org/pdf/1707.07250.pdf

Zadeh A, Liang P P, Mazumder N, Poria S, Cambria E and Morency L P. 2018a. Memory fusion network for multi-view sequential learning//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI: 5634-5641 [DOI: 10.1609/aaai.v32i1.12021]

Zadeh A, Liang P P, Poria S, Vij P, Cambria E and Morency L P. 2018b. Multi-attention recurrent network for human communication comprehension//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI: 5642-5649 [DOI: 10.1609/aaai.v32i1.12024]

Zadeh A, Zellers R, Pincus E and Morency L P. 2016. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos [EB/OL]. [2025-04-08]. https://arxiv.org/pdf/1606.06259.pdf

Zadeh A A B, Liang P P, Poria S, Cambria E and Morency L P. 2018c. 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: ACL: 2236-2246 [DOI: 10.18653/v1/P18-1208]

Zeng Y, Yan W J, Mai S and Hu H F. 2024. Disentanglement translation network for multimodal sentiment analysis. Information Fusion, 102: #102031 [DOI: 10.1016/j.inffus.2023.102031]

Zhang D, Ju X C, Zhang W, Li J H, Li S S, Zhu Q M and Zhou G D. 2021. Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtually: AAAI: 14338-14346 [DOI: 10.1609/aaai.v35i16.17686]

Zhang J and Jia H. 2008. Design of speech corpus for mandarin text to speech//The Blizzard Challenge 2008 Workshop. Edinburgh: University of Edinburgh

Zhang Z P, Luo P, Loy C C and Tang X O. 2018. From facial expression recognition to interpersonal relation prediction. International Journal of Computer Vision, 126(5): 550-569 [DOI: 10.1007/s11263-017-1055-1]

Zhao G Y, Huang X H, Taini M, Li S Z and Pietikäinen M. 2011. Facial expression recognition from near-infrared videos. Image and Vision Computing, 29(9): 607-619 [DOI: 10.1016/j.imavis.2011.07.002]

Zou H Q, Si Y K, Chen C, Rajan D and Chng E S. 2022. Speech emotion recognition with co-attention based multi-level acoustic information//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 7367-7371 [DOI: 10.1109/ICASSP43922.2022.9747095]

作者简介

王善敏,女,讲师,主要研究方向为人脸表情识别和多模态情感识别。E-mail:smwang1994@163.com

刘青山,通信作者,男,教授,博士研究生导师,主要研究方向为计算机视觉和机器学习。E-mail:qsliu@njupt.edu.cn

刘成广,男,博士研究生,主要研究方向为表情识别和多模态抑郁评估。E-mail:liuchengguang@nuist.edu.cn

陈胜宇,男,硕士研究生,主要研究方向为多模态情感识别。E-mail:shengyuchen@nuist.edu.cn