计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2022/16(07)-1479-25
doi: 10.3778/j.issn.1673-9418.2112081

面向深度学习的多模态情感识别研究进展

赵小明[1,2+],杨轶娇[1],张石清[2]
1. 浙江科技学院 理学院,杭州 310000
2. 台州学院 智能信息处理研究所,浙江 台州 318000
+ 通信作者 E-mail: tzxyzxm@163.com

摘　要:多模态情感识别是指通过与人类情感表达相关的语音、视觉、文本等不同模态信息来识别人的情感状态。该研究在人机交互、人工智能、情感计算等领域有着重要的研究意义,备受研究者关注。鉴于近年来发展起来的深度学习方法在各种任务中所取得的巨大成功,目前各种深度神经网络已被用于学习高层次的情感特征表示,用于多模态情感识别。为了系统地总结深度学习方法在多模态情感识别领域中的研究现状,拟对近年来面向深度学习的多模态情感识别研究文献进行分析与归纳。首先,给出了多模态情感识别的一般框架,并介绍了常用的多模态情感数据集。然后,简要回顾了代表性深度学习技术的原理及其进展。随后,重点详细介绍了多模态情感识别中的两个关键步骤的研究进展:与语音、视觉、文本等不同模态相关的情感特征提取方法,包括手工特征和深度特征;融合不同模态信息的多模态信息融合策略。最后,分析了该领域面临的挑战和机遇,并指出了未来的发展方向。

关键词:情感识别;多模态;深度学习;手工特征;深度特征;融合
文献标志码:A　　**中图分类号:**TP391

Survey of Deep Learning Based Multimodal Emotion Recognition

ZHAO Xiaoming[1,2+], YANG Yijiao[1], ZHANG Shiqing[2]
1. School of Science, Zhejiang University of Science and Technology, Hangzhou 310000, China
2. Institute of Intelligent Information Processing, Taizhou University, Taizhou, Zhejiang 318000, China

Abstract: Multimodal emotion recognition aims to recognize human emotional states through different modalities related to human emotion expression such as audio, vision, text, etc. This topic is of great importance in the fields of human-computer interaction, artificial intelligence, affective computing, etc., and has attracted much attention. In view of the great success of deep learning methods developed in recent years in various tasks, a variety of deep neural networks have been used to learn high-level emotional feature representations for multimodal emotion recognition. In order to systematically summarize the research advance of deep learning methods in the field of multimodal emotion recognition, this paper aims to present comprehensive analysis and summarization on recent multimodal emotion recognition literatures based on deep learning. First, the general framework of multimodal emotion recognition is given, and the commonly used multimodal emotional dataset is introduced. Then, the principle of representative deep learning techniques and its advance in recent years are briefly reviewed. Subsequently, this paper focuses on the advance of two key steps in multimodal emotion recognition: emotional feature extraction

基金项目:浙江省自然科学基金重点项目(LZ20F020002);国家自然科学基金面上项目(61976149)。
This work was supported by the Natural Science Foundation of Zhejiang Province (LZ20F020002) and the National Natural Science Foundation of China (61976149).
收稿日期:2021-12-20　　**修回日期:**2022-02-14

methods related to audio, vision, text, etc., including hand-crafted feature extraction and deep feature extraction; multimodal information fusion strategies integrating different modalities. Finally, the challenges and opportunities in this field are analyzed, and the future development direction is pointed out.

Key words: emotion recognition; multimodal; deep learning; hand-crafted feature; deep feature; fusion

情感识别是一个以人的情感状态为目标的动态过程,这意味着每个人的行为对应的情感是不同的[1-2]。日常生活中的情感识别对社会交往很重要,人类以不同的方式表达自己的感受,情感在决定人类行为中起着重要作用。为了确保有意义的交流,对这些情感的准确解读非常重要[3]。

在情感识别任务中,情感通常分为离散状态或连续状态[4]。常见的离散的情感状态有快乐、恐惧、惊讶和悲伤等;连续情感状态可以分为效价(valence)、唤醒(arousal)和支配(dominance)。唤醒表达的是激活的水平(被动或主动),并与当前情感状态的强度(积极或消极)有关;效价表示愉悦程度;支配表示情感条件施加的控制程度。由于连续情感在现实环境中的测量具有挑战性,离散情感建模更为流行[4]。

人们交流感情的方式有很多,既有口头语言,也有非口头语言,包括表达性语言、面部姿势、肢体语言等[5]。因此,来自多种模态的情绪信号可用来预测一个主体的情绪状态[6]。然而,单一的模态无法准确判断一个人的情感,单凭眼前的某个特定实体或事件无法有效判断某人的情感变化[7]。这就是情感识别应被视为多模态问题的原因之一。因此多模态情感识别考虑了多种输入模式,如语音、文本、视觉线索等,对情感信息进行建模和识别。

多模态情感识别在社交机器人、教育质量评估、安全控制、人机交互系统等方面[8-12]具有相当大的应用前景。为了推动情感识别任务的发展,近十年来出现了不同的多模态情感任务挑战赛,包括AVEC、EmotiW、MuSe、MEC等。AVEC(audio/visual emotion challenge and workshop)[13]是一项音视频挑战赛,目的是为多模态信息处理提供通用的基准测试集,并将基于听觉、视觉和视听觉情感识别任务聚集在一起。EmotiW(emotion recognition in the wild challenge)[14]是一场野外挑战竞赛,旨在为研究者提供一个平台,在代表真实世界或接近真实世界场景的数据上验证他们的方法。自2013年开始,EmotiW每年都会举办一次,挑战的子项目每年都会有所变化。MuSe 2020(multimodal sentiment)[15]是一个基于现实生活媒体(real-life media)的挑战赛,更全面地融合视听和语言模态,重点关注情感识别、情感目标参与和可信度检测三个任务,提出了一个用于竞赛的野外数据库Muse-CaR。基于MuSe 2020挑战赛的工作,MuSe 2021[16]更全面地整合了视听、语音和生理信号模态,并提供了Ulm-TSST(Ulm-trier social stress)数据集。MEC(multimodal emotion recognition)[17]挑战赛是中国模式识别大会(Chinese conference on pattern recognition,CCPR)的一部分,提供了中国自然视听情感数据库CHEAVD,定义了三个子挑战:音频、视频和多模态情感识别。

近年来,深度学习依靠强大的特征学习能力,在语音信号处理、计算机视觉、自然语言处理、情感计算等许多领域取得了巨大成功[18-20]。深度学习本质上是通过使用多个非线性变换的层次架构来获得高级的特征表示。深度信念网络(deep belief network,DBN)[21]、卷积神经网络(convolutional neural network,CNN)[22]和循环神经网络(recurrent neural network,RNN)[23]是深度学习最常用的三种方法。近年来,这些深度学习方法在多模态情感识别任务中往往用于高层次的特征学习或多模态信息的融合。为了系统地总结深度学习方法在多模态情感识别领域中的研究现状,本文拟对近年来面向深度学习的多模态情感识别研究进行系统的分析与总结。

文献[24]侧重于综述多模态信息的融合方法研究进展,没有涉及到面向深度学习的特征提取技术介绍。与上述文献不同,本文既对多模态融合方法进行详细总结与归纳,又对近年来面向深度学习的语音、视觉及文本的特征提取方法进行了分析和总结。本文主要贡献可以总结如下:(1)从多模态的角度对面向深度学习的多模态情感识别研究进行了最新的系统性文献分析与归纳,即以多模态(语音、视觉、文本等)分析主体情感为中心,对手工情感特征提取、与深度学习技术相关的深度情感特征提取以及多模态信息融合方法进行了分析与总结。(2)分析了该领域面临的挑战和机遇,并指出了未来的发

展方向。

图1给出了一般的多模态情感识别框架。由图1所示，一般的多模态情感识别系统包括三个步骤：特征提取、多模态信息融合和情感分类器的设计。特征提取是对语音、视觉、文本等不同模态信息分别提取与情感表达相关的特征参数。多模态信息融合指的是采用不同的融合策略对两种及以上的单模态信息进行融合。常见的多模态信息融合方法有特征层融合、决策层融合、模型层融合等。情感分类器的设计是采用合适的分类器来学习提取的特征表示与相关识别的情感之间的映射关系，从而获得最终的情感识别结果。根据单一模态信息的不同，一个多模态情感识别系统往往包括多个单一模态情感识别子任务，如采用语音信号的语音情感识别、采用视觉信息的视觉情感识别以及采用文本信息的文本情感识别等。

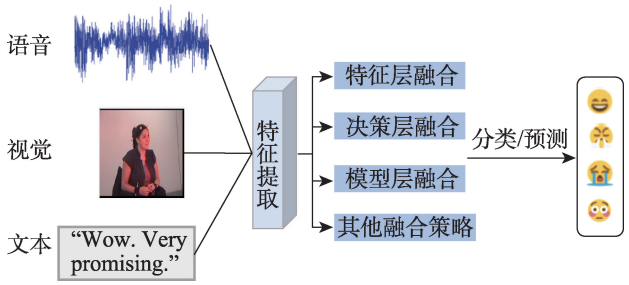

图1 多模态情感识别框架

Fig.1 Framework of multimodal emotion recognition

多模态情感识别本质上是一个模式识别问题。目前，现有的机器学习方法大都可以用于情感的分类。代表性的情感分类器主要有：贝叶斯网络（Bayesian networks，BN）、最大似然原理（maximum likelihood principle，MLP）、支持向量机（support vector machine，SVM）等[25-27]。考虑到情感分类器的设计已经比较成熟，因此本文只针对多模态情感识别系统中的前两个关键步骤——情感特征提取和多模态信息融合，系统地阐述其近年来的发展现状和未来的展望。

1 多模态情感数据集

多模态情感数据集是指包含动态情感变化和多种情感类别的数据集，数据集中包含常见的模态信息有语音、视觉、文本等。本章将重点介绍近些年来常用的多模态情感数据集，如表1所示。

eNTERFACE’05[28]：该数据集是一个试听数据

表1 多模态情感数据集

Table 1 Multimodal emotional datasets

数据集	年份	模态	简要介绍	情感标签
eNTERFACE’05[28]	2006	语音、视觉	1 277个视听样本，来自14个不同国家的42名参与者	愤怒、厌恶、恐惧、快乐、悲伤、惊讶
RML[29]	2008	语音、视觉	720个由视听情感表达的样本组成，8名参与者	愤怒、厌恶、恐惧、幸福、悲伤、惊讶
IEMOCAP[30]	2008	语音、视觉、姿势、文本	10 039段对话：平均持续时间为4.5 s，平均单词数为11.4；10名演员	中性、快乐、悲伤、愤怒、惊讶、恐惧、厌恶、沮丧、兴奋等；维度标签：效价、唤醒和支配
AFEW[31]	2012	语音、视觉	1 426个视频片段组成	愤怒、厌恶、恐惧、幸福、悲伤、惊讶、中性
BAUM-1s[32]	2016	语音、视觉	1 222个视频样本，31名土耳其受试者	快乐、愤怒、悲伤、厌恶、恐惧、惊讶
CHEAVD[17]	2016	语音、视觉	来自电影、电视剧、电视节目的140 min的自发情感片段，238名说话者	有26种非原型的情感状态，前8个主要情感为愤怒、快乐、悲伤、担心、焦虑、惊讶、厌恶、中性
CMU-MOSI[33]	2016	语音、视觉、文本	2 199个评论的话语、93段说话者视频	消极、积极
RAMAS[34]	2018	语音、视觉、姿势、生理信号	大约7 h的高质量特写视频记录，10位演员	愤怒、厌恶、快乐、悲伤、恐惧、惊讶
RAVDESS[35]	2018	语音、视觉	60段演讲，44首歌曲，24位演员	中性、平静、快乐、悲伤、愤怒、恐惧、厌恶、惊讶
CMU-MOSEI[36]	2018	语音、视觉、文本	来自1 000多名在线YouTube演讲者的3 837段视频	快乐、悲伤、愤怒、恐惧、厌恶、惊讶
MELD[37]	2019	语音、视觉、文本	包含了电视剧 *Friends* 中1 433段对话中的13 000句话	愤怒、厌恶、恐惧、喜悦、中立、悲伤、惊讶；正面、负面和中性
CH-SIMS[38]	2020	语音、视觉、文本	2 281个野外视频片段	消极、弱消极、中性、弱积极、积极
HEU-part1[39]	2021	视觉、姿势	总共19 004个视频片段，根据数据源分为两部分，共有9 951名受试者	愤怒、无聊、困惑、失望、厌恶、恐惧、快乐、中立、悲伤、惊讶
HEU-part2[39]	2021	语音、视觉、姿势		

集，由1 277个视听样本组成，由来自14个不同国家的42名参与者（8名女性）完成。每个参与者都被要求连续听6篇短篇小说，每一篇都能引起一种特定的情感。受试者必须对每一种情况做出反应，两位人类专家判断这些反应是否以明确的方式表达了预期的情感。六种特定的情感分别为：愤怒、厌恶、恐惧、快乐、悲伤和惊讶。

RML[29]：该数据库由720个包含视听情感表达的样本组成，每个视频的持续时间在3~6 s之间，包含了愤怒、厌恶、恐惧、幸福、悲伤、惊讶六种基本情绪。录音是在安静明亮的背景氛围中进行的，使用数码相机。8名受试者进行了录音，并会说六种语言，分别为英语、普通话、乌尔都语、旁遮普语、波斯语和意大利语，英语和普通话的不同口音也包括在内。采用16位单通道数字化，以22 050 Hz的频率记录样本。记录速度被设置为30 frame/s。

IEMOCAP[30]：该数据集是由南加州大学的Sail实验室收集的一个包含动作、多模态和多峰值的数据库。它包含10个说话者在分成话语的双向对话中的行为，包括视频、语音、面部动作捕捉和文本转录，所有视频中对话的媒介都是英语。总共包含10 039段对话，平均持续时间为4.5 s，平均单词数为11.4。参与者表演即兴表演或脚本场景。被众多注释者标注为中性、快乐、悲伤、愤怒、惊讶、恐惧、厌恶、挫折、兴奋等类别标签和配价、激活、支配等维度标签。

AFEW[31]：该数据集是在具有挑战性的条件下录制的动作面部表情数据集，由1 426个视频片段组成。这些视频片段被标记为6类基本情感（生气、高兴、悲伤、惊讶、厌恶、恐惧）和中性情感之一。该数据集捕捉了不同的面部表情、自然的头部姿势运动、遮挡物，来自不同种族、性别、年龄的受试者和一个场景中的多个受试者。

BAUM-1s[32]：该数据集是一个视听自发数据集，包含来自31名土耳其受试者的1 222个视频样本。该数据集有六种基本情绪（快乐、愤怒、悲伤、厌恶、恐惧、惊讶）以及无聊和蔑视。它还包含四种精神状态，即不确定、思考、专注和烦恼。为了获得自发的视听表达，采用了观看电影的情感激发方法。

CHEAVD[17]：该数据集为中国自然情感视听数据库，提取了34部电影、2部电视剧、2部电视节目、1部即兴演讲和1部脱口秀节目中的140 min的自发情感片段，其中电影和电视剧占大部分。该数据集有238名说话者，覆盖了从儿童到老年人，其中男性比例为52.5%，女性比例为47.5%；总共有26种非原型的情感状态，包括基本的6种，由4个讲母语的人标记。前8个主要的情感为愤怒、快乐、悲伤、担心、焦虑、惊讶、厌恶和中性。

CMU-MOSI[33]：该数据集是一个富含情感表达的数据集，由2 199个评论的话语、93段说话者（含89个说话者）视频组成。这些视频涉及大量主题，如电影、书籍和产品。视频是从YouTube上抓取的，并被分割成话语。每个分割情感标签由5个注释者在+3（强阳性）到−3（强阴性）之间评分，将这5个注释的平均值作为情感极性，因此只考虑了两类（积极和消极）。训练集由数据集中的前62段视频组成，测试集包含剩余的31段视频。在训练和测试中分别包含了1 447个话语（含467个否定话语）和752个话语（含285个否定话语）。

RAMAS[34]：该数据集是第一个俄罗斯多模态情感数据库。他们认为专业戏剧演员可能会使用动作模式的刻板印象，因此选用半职业演员在情感情境中表演动作。10名半职业演员（5名男性和5名女性）参与了数据收集，年龄在18~28岁，母语为俄语。半职业演员在设定的场景中表达了一种基本的情感（愤怒、厌恶、快乐、悲伤、恐惧、惊讶）。数据库包含大约7 h的高质量特写视频记录，采集了音频、运动捕捉、特写和全景视频、生理信号等多种数据。

RAVDESS[35]：该数据集由24位专业演员录制，包括60段演讲和44首带有情绪的歌曲（包含中性、平静、快乐、悲伤、愤怒、恐惧、厌恶、惊讶）。每个演员录制的作品有三种形式：视听、视觉和语音。录音是在专业工作室录制的，镜头中只有演员和绿色屏幕可见。为了确保相机能够捕捉演员的头和肩膀，相机的高度随时调整。工作室提供全光谱照明来最小化面部阴影。

CMU-MOSEI[36]：该数据集是迄今为止最大的多模态情感分析和情感识别数据集，包含来自1 000多名在线YouTube演讲者的3 837段视频，其中包含六种情绪类别：快乐、悲伤、愤怒、恐惧、厌恶和惊讶。它在话语层面进行注释，共有23 259个样本。CMU-MOSEI中的样本包括三种模式：采样率为44.1 kHz的音频数据、文本转录和以30 Hz的频率从视频中采样的图像帧。该数据集是性别平衡的，所有的句子都是从各种主题和独白视频中随机选择的，视频被转录并标记正确的标点符号。

MELD[37]：该数据集是从EmotionLines数据集演

变而来的。EmotionLines只包含电视剧*Friends*中的对话。MELD是一个多模态的情感对话数据集，包含语音、视觉和文本信息。MELD包含了电视剧*Friends*中1 433段对话中的13 000句话，每段对话包含两个以上的说话者。由于数据仅从一部电视剧中获得，参与人数有限，84%的场次由6位主演获得。对话中的每一句话都被标记为这七种情感标签中的任何一种——愤怒、厌恶、悲伤、喜悦、中立、惊讶和恐惧。MELD还对每个话语都有情绪（正面、负面和中性）注释。

CH-SIMS[38]：该数据集是一个中文单模态和多模态情感分析的数据集，其中包含2 281个经过精炼的野外视频片段，具有多模态和独立的单模态注释。它允许研究人员研究模态之间的相互作用或使用独立的单模态注释进行单模态情感分析。该数据集只考虑普通话，对口音材料的选择持谨慎态度。剪辑长度不少于1 s，也不超过10 s。对于每个视频剪辑，除了演讲者的脸外，不会显示其他脸。每个片段包含15个单词，平均长度为3.67 s。每个剪辑都由人类注释者根据5个情感分数的平均值进行标记，五类分别为消极{−1.0，−0.8}，弱消极{−0.6，−0.4，−0.2}，中性{0}，弱积极{0.2，0.4，0.6}和积极{0.8，1.0}。

HEU Emotion[39]：该数据库包含总共19 004个视频片段，根据数据源分为两部分。第一部分包含从Tumblr、Google和Giphy下载的视频，包括10种情绪和两种模式（面部表情和身体姿势）；第二部分包括从电影、电视剧和综艺节目中手工获取的语料，包括10种情绪和3种形式（面部表情、身体姿势和情绪言语）。该数据库是迄今为止最广泛的多模态情绪数据库，共有9 951名受试者，他们是来自不同文化背景的人，如中国人、美国人、泰国人和韩国人。在大多数情况下，他们说自己的母语。因此，该数据库是一个具有多种语言的情感数据库。

2 深度学习技术回顾

深度学习被认为是机器学习中一个新兴的研究领域，近年来得到了更多的关注。与传统方法相比，用于情感识别的深度学习技术具有许多优点，比如能够检测复杂的结构与特征，而无需手动进行特征提取等[40]。在本章中，简要回顾了几种有代表性的深度学习方法及其最新的改进方法。

2.1 深度信念网络

深度信念网络（DBN）是由Hinton等[21]于2006年提出的一种生成式模型，旨在获取输入数据的高阶层次特征表示。DBN是一种多层深结构，由一系列叠加的限制玻尔兹曼机（restricted Boltzmann machine，RBM）构建而成[41]。RBM由两层神经元构成：可见层和隐藏层。每个神经元与另一层的神经元完全连接，但同一层的神经元之间没有连接。训练DBN需要两个阶段：预训练和微调。预培训是通过一种有效的逐层贪婪学习策略[42]以无监督的方式实现的。在预训练过程中，采用对比发散（contrastive divergence，CD）[43]算法对DBN中的RBM进行训练，以优化DBN模型的权重和偏差。然后，使用反向传播（back propagation，BP）算法进行微调以更新网络参数。DBN的主要优点是它具有快速学习和提供高效表示的趋势，它通过层层预训练来实现这一点[21]。同时，DBN也存在局限性。如在训练DBN模型时计算成本高；DBN不能考虑输入图像的二维结构，这可能会影响它们在计算机视觉等领域中的性能和适用性。

近年来，不少研究者提出了一些基于DBN模型的改进方法。Lee等[44]提出了一种用于全尺寸图像的分层生成式模型，即卷积深度置信网络（convolutional deep belief network，CDBN），由多个基于最大池化的卷积RBM（convolutional restricted Boltzmann machine，CRBM）相互堆叠而成。CDBN能够从未标记的对象图像和自然场景中学习有用的高级视觉特征。Wang等[45]提出了一种基于迁移学习的增长型DBN（growing DBN with transfer learning，TL-GDBN）。TL-GDBN通过迁移学习将学习到的权重参数转移到新添加的神经元和隐藏层，从而实现结构增长，直到满足预训练的停止标准。然后采用自上而下逐层偏最小二乘回归法对TL-GDBN预训练得到的权值参数进行了进一步的微调，避免了传统的基于反向传播算法的微调问题。Deng等[46]提出了一种基于改进的量子启发差分演化（quantum-inspired differential evolution，MSIQDE）算法，然后利用具有全局优化能力的MSIQDE对DBN的参数进行优化，构造了一个最优DBN模型，并进一步应用该模型提出了一种新的故障分类方法，即MSIQDE-DBN方法。MSIQDE-DBN可以消除人为因素的干扰，自适应地选择DBN的最佳参数，从而有效地提高分类精度，满足实际要求。

2.2 卷积神经网络

卷积神经网络（CNN）最初是于1998年由LeCun等[47]提出的，并被广泛使用和改进。CNN的基本结

构包括卷积层、池化层和全连接层。卷积层采用多个可学习滤波器对整个输入图像进行卷积运算，从而产生相应的激活特征映射。池化层连接于卷积层之后。池化层通过使用非线性下采样方法实现平移不变性，用于对提取到的特征进行降维，保留主要特征。常用的池化方法有最大池化和平均池化。全连接层通常位于CNN的末端，它用于激活上一层以生成最终的特征表示和分类结果。近年来，各种改进的CNN架构被提出，并应用于大量领域。代表性的CNN架构有AlexNet[48]、VGGNet[49]、GoogleNet[50]、ResNet[51]、DenseNet[52]等。

与2D-CNN相比，Tran等[53]提出的用于大规模视频数据集上训练的三维卷积神经网络（3D-CNN），是一种简单而有效的时空特征学习方法，可以同时对外观和动作进行建模。由于三维卷积比二维卷积涉及更多的参数，计算成本较高，Yang等[54]提出了一种近似于传统的三维卷积网络的模型——基于微网（MicroNets）的非对称单向三维卷积网络（asymmetric 3D convolutional neural networks）。为了提高其特征学习能力，该模型采用了一组局部三维卷积网络，引入了多尺度三维卷积分支。然后，利用微网构建非对称3D-CNN深度模型，用于动作识别任务。Kumawat等[55]提出了LP-3DCNN（local phase in 3D convolutional neural networks），使用校正局部相位体积（rectified local phase volume，ReLPV）块代替传统的3D卷积块，ReLPV块通过提取输入图中每个位置的3D局部邻域中的相位来获得特征图。Chen等[56]提出了一种频域紧致三维卷积神经网络（frequency domain compact 3D convolutional neural networks），利用一组学习到的具有很少网络参数的最优变换，将时域转换为频域来实现3D卷积操作，从而消除三维卷积滤波器的时间维冗余。

总之，卷积神经网络的权值共享网络结构网络模型的复杂度，减少了权值的数量。该优点在网络的输入使多维图像上表现得更为明显，使图像可以直接作为网络的输入，避免了传统识别算法中复杂的特征提取和数据重建过程。卷积神经网络的局限性有：无法表示高层特征与低层特征之间的位姿（平移和旋转）关系，以及底层对象之间的空间关系。因此，CNN在识别具有空间关系特征时存在不足；池化层可能会丢失有价值的信息等。

2.3 循环神经网络

循环神经网络（RNN）[23]能够从序列数据中捕获时间信息，因此通常用于序列处理。作为一个单前馈神经网络，RNN采用隐状态上的递归连接来捕获序列数据的历史信息。此外，RNN在所有时间步长上共享相同的网络参数。对于训练RNN，采用传统的时间反向传播（backpropagation through time，BPTT）[57]算法。然而当网络需要训练的参数很多时，RNN容易造成梯度消失或梯度爆炸问题。

长短期记忆网络（long short-term memory，LSTM）[58]于1997年被提出，是一种新的循环网络结构。LSTM主要用于缓解RNN训练过程中产生的梯度消失和梯度爆炸问题。LSTM单元中有三种类型的门：输入门、遗忘门和输出门。输入门用于控制有多少当前输入数据流入网络的存储单元。遗忘门作为LSTM单元的关键部件，用于控制哪些信息需要保留，哪些信息需要遗忘，并以某种方式避免梯度损失和爆炸问题。输出门控制存储单元对当前输出值的影响。基于这三个特殊门，LSTM能够对序列数据中的长期相关性进行建模。

近年来，出现了各种RNN或LSTM的改进。Chung等于2014年提出了循环门控单元（gated recurrent unit，GRU）[59]。GRU使每个循环单元自适应地建模不同时间尺度的长期依赖关系。与LSTM单元不同，GRU单元内没有单独的存储单元。Zhao等[60]于2019年提出了一种基于卷积LSTM的贝叶斯图，用于识别基于骨架的动作。Zhang等[61]于2019年提出了一种用于语音情感分类的多尺度深卷积LSTM。Xing等[62]于2020年提出了一种新的脉冲卷积递归神经网络（spiking convolutional recurrent neural network，SCRNN），借助卷积运算和递归连接，从基于事件的序列数据中建模时空关系。

3 特征提取

语音、视觉、文本是情感表达最常见的三种模态。针对语音、视觉、文本信息的情感特征提取是多模态情感识别任务的一个关键问题。根据特征类型的不同，可以分为手工特征和深度特征两大类。下面将对语音、视觉、文本三种模态信息分别阐述其手工特征提取和深度特征提取技术的进展。

3.1 语音情感特征提取

语音情感识别是通过说话人的声音来识别他人的情绪。语音情感特征提取是决定语音情感识别精度高低的一个关键因素。语音情感特征主要分为低层次的手工语音情感特征和通过深度学习技术得到

的深度语音情感特征。

3.1.1 手工语音情感特征

早期用于自动语音情感识别的语音情感特征是手工制作的低层次描述（low-level descriptors，LLD）特征，如韵律特征（基频、能量）、音质特征（共振峰、声道参数）、谱特征（线性预测倒谱系数（linear predictive cepstral coefficient，LPCC）、Mel频率倒谱系数（Mel-frequency cepstral coefficients，MFCC））等[63-66]。

Liscombe等[67]提取了一系列基于基音周期、振幅和频谱倾斜的连续语音特征，并评估了其与各种情感的关系。Yacoub等[68]提取了37个韵律学特征，包括音高（基频）、响度（能量）和音段（可听持续时间）等，分别比较了使用神经网络、支持向量机、K-近邻算法和决策树在语音情感分类中的结果。Schmitt等[69]使用由MFCC和能量低级描述符（LLD）创建的音频词袋（bag-of-audio-words，BoAW）方法作为特征向量和简单的支持向量回归（support vector regression，SVR）来预测唤醒和效价维度。孙韩玉等[70]考虑了不同特征包含的信息，使用频谱图特征和LLD特征分别输入到双通道卷积门控循环网络。

Luengo等[71]从语音信号中提取声学参数：韵律学特征、谱相关特征和语音质量特征。对单个参数和组合特征进行研究分析，在参数级（早期融合）和分类器级（后期融合）研究了不同参数类型的组合，判别这些特征在情感识别中的不同性能。

3.1.2 深度语音情感特征

近年来，深度学习技术广泛应用于语音情感识别任务，用于深度语音情感特征提取。常见的用于语音情感识别的深度学习方法有CNN、DBN、RNN等。

Dutta等[72]提出了一种基于线性预测编码（linear predictive coding，LPC）和MFCC的语音识别模型。LPC和MFCC特征由两种不同的RNN网络进行提取，用于识别阿萨姆语。

Mao等[73]提出了将CNN应用于语音情感识别的特征提取。CNN有两个学习阶段：在第一阶段，利用未标记样本通过一种稀疏自动编码器来学习局部不变特征；在第二阶段，局部不变特征被用作特征提取器的输入，即显著判别特征分析（salient discriminative feature analysis，SDFA），以学习显著判别特征。

陈婧等[74]提出了一种新的多粒度特征提取方法。该方法基于不同的时间单位，包括短时帧粒度、中时段粒度以及长时窗粒度。为了融合这些多粒度特征，提出了一种基于认知机理的回馈神经网络（cognition-inspired recurrent neural network，CIRNN）。CIRNN组合不同的时间级特征来模拟人类对音频信号的逐步处理，通过同时突出情感的时间序列和内容信息的作用，实现多级信息融合。

俞佳佳等[75]提出了一种针对语音原始信号的特征提取方法，利用SincNet滤波器从原始语音波形中提取一些重要的窄带情感特征，再利用Transformer模型的编码器提取包含全局上下文信息的深度特征。

Zhang等[76]利用DBN对提取的低阶声学特征进行无监督特征学习，根据DBN隐含层的学习结果，对多层感知器（multi-layer perceptron，MLP）进行初始化，并用于汉语语音情感分类。

Ottl等[77]以两种不同的方式从视频中提取特征，其一使用深度频谱（deep spectrum）工具包从音频频谱图中学习深度表示，再采用各种流行的卷积神经网络结构进行图像识别预训练；此外，使用OpenSMILE工具[78]提取了6 373维的手工特征表示，包括语音质量特征，如抖动和微光，以及频谱、MFCC和与发声相关的低级描述符（LLD）等。最后，对深度特征和手工特征进行了早期和晚期融合。

从上述已有的手工语音情感特征和深度语音情感特征文献来看：（1）采用OpenSMILE工具提取较高维度的LLD特征，已成为手工语音情感特征的主流方法。（2）采用CNN从原始语音信号直接提取高层次的语音情感特征，已成为深度语音情感特征的主流方法。（3）手工语音情感特征和深度语音情感特征各有优缺点。近年来将这两种特征进行融合用于语音情感识别，是一个有意义的研究方向。

3.2 视觉情感特征提取

视觉情感识别通过提取面部表情图像的外观和几何特征并感知其变化来识别静态图像或视频序列中的情感[19,79-82]。根据视觉输入数据的类型，基于视觉的情感识别可分为两种：基于静态面部图像的表情识别和基于动态视频序列的表情识别。下面将针对静态面部图像和动态视频序列分别阐述各自的手工特征提取和深度特征提取的进展。

3.2.1 手工视觉情感特征

（1）静态面部图像

静态图像是指不包含音频和时间信息的静止图像，先对其进行一系列的预处理，如旋转、人脸定位、对齐、归一化等，再提取图像信息中的几何图形和外貌特征来获得人脸表情特征。用于传统面部情感识

别的典型特征主要是手工制作的特征，对于静态面部图像主要的手工特征提取方法有：局部二值模式（local binary pattern, LBP）[83-84]、尺度不变特征变换（scale invariant feature transform，SIFT）[85]、方向梯度直方图（histograms of oriented gradients，HOG）[86]、Gabor小波法[87]等。

刘军等[88]提出了一种新的基于主导近邻像素的人脸图像表示——局部Gabor空间直方图（local Gabor spatial histogram based on dominant neighboring pixel，LGSH-DNP）。首先对人脸图像进行Gabor滤波器组滤波，过滤后的图像中的每个像素都由具有最高值的两个相邻像素的位置标记，从而得到多个描述子图像。其次从这些描述子图像中提取空间直方图。最后采用加权交集直方图相似度测度实现人脸分类。在进行了大量实验之后验证了所提出的LGSH-DNP方法的有效性。

Bah等[89]提出了一种新的方法，利用LBP算法，结合对比度调整、双边滤波、直方图均衡化和图像融合等先进的图像处理技术，解决了影响人脸识别精度的一些问题，从而提高LBP编码的识别率，提高了整个人脸识别系统的准确率。

Deeba等[90]开发了一个基于局部二值模式直方图（local binary pattern histogram，LBPH）方法的人脸识别系统，LBPH算法是LBP和HOG算法的组合，用于处理低层和高层图像中的实时人脸识别。使用LBPH，可以用一个简单的特征向量表示人脸图像。

Zhang等[91]从每幅人脸图像中提取对应于每幅人脸图像的一组标志点的尺度不变特征变换（SIFT）特征。然后，将由提取的SIFT特征向量组成的特征矩阵作为输入数据，输送到设计良好的深度神经网络模型，用于学习分类的最佳鉴别特征。

（2）动态视频表情序列

面部表情涉及一个动态过程，并且动态信息（例如面部标志的移动和面部形状的变化）包含可以更有效地表示面部表情的有用信息。因此，捕获这样的动态信息以便识别整个视频序列中的面部表情非常重要。基于动态视频表情序列的算法主要包括光流法和模型法。

在视频表情序列中，光流分析已被应用于检测面部部件的运动，通过测量两个连续帧之间面部特征点的几何位移来确定这些部件的运动[92]。Fan等[93]使用两种类型的动态信息来增强识别：一种基于梯度金字塔直方图（pyramid histogram of gradients，PHOG）[94]的新型时空描述符来表示面部形状的变化，以及密集光流来估计面部标志的移动。将图像序列视为时空体，并使用时间信息来表示与面部表情相关联的面部地标的动态运动。在此背景下，将表示空间局部形状的PHOG描述符扩展到时空域，以捕获时间维度中面部子区域局部形状的变化，从而给出额头、嘴、眉毛和鼻子的三维面部组件子区域。他们将这个描述符称为PHOG-TOP（PHOG-three orthogonal planes）。通过结合PHOG-TOP和面部区域的密集光流，利用鉴别特征的融合进行分类，从而识别面部表情。

刘涛等[95]提出了一种新的面部情感识别的方法，通过对人脸表情图像与中性表情图像之间的光流特征的提取来体现人脸表情变化的差异，采用高斯线性判断分析（linear discriminant analysis，LDA）方法对光流特征进行映射，得到人脸表情图像的特征向量，采用多类支持向量机分类器实现人脸情感分类。

Happy等[96]探讨了与面部微运动相关的时间特征，并提出了用于微表情识别的光流方向模糊直方图（fuzzy histogram of optical flow orientation，FHOFO）特征。FHOFO使用直方图模糊化从光流矢量方向构造合适的角度直方图，对时间模式进行编码，以对微观表达式进行分类。

邵洁等[97]针对RGB-D图像序列，提出了一种自发的人脸表情识别算法。在对图像对齐和归一化进行预处理后，提取四维时空纹理数据作为动态特征。然后采用慢速特征分析方法检测表情的顶点，建立顶点图像的三维人脸几何模型作为静态特征。将这两种特征结合起来，通过主成分分析进行降维，最后利用条件随机场对特征进行训练和分类。

Yi等[98]利用特征点的运动趋势和特征块的纹理变化，提出了一种截取视频序列的面部情感识别框架。首先，采用主动外观模型（active appearance model，AAM）对特征点进行标记，选择其中最具代表性的24个特征点。其次，通过确定情感强度最小和最大的两个关键帧，从人脸视频中截取人脸表情序列。然后，拟合代表任意两个特征点之间欧氏距离变化的趋势曲线，并计算趋势曲线上特定点的斜率。最后，将计算得到的斜率集合与所提出的特征块纹理差（feature block texture difference，FBTD）相结合，形成最终的表情特征，并输入一维卷积神经网络进行情感识别。

3.2.2 深度视觉情感特征

尽管传统的人脸识别方法通过提取手工特征取得了显著的成功，但近年来由于深度学习方法高度的自动识别能力逐渐应用于情感识别，用于提取高级特征。

（1）静态面部图像

对于静态面部图像的深度特征提取，主要采用的是基于卷积神经网络的一些模型框架。Yolcu等[99]提出了检测面部重要部位的方法，使用三个结构相同的CNN，每一个都能检测到脸部的一部分，如眉毛、眼睛和嘴巴。在将图像引入CNN之前，要进行裁剪和面部关键点的检测，结合原始图像获得的标志性人脸被引入第二类CNN以检测面部表情。研究人员表明，这种方法比单独使用原始图像或图像化人脸更准确。

Sun等[100]用光流表示静态图像中的时间特征，提出了一种多通道深度时空特征融合神经网络（multi-channel deep spatial-temporal feature fusion neural network，MDSTFN），用于静态图像的深度时空特征提取与融合。该方法的每个通道都是从预先训练好的深卷积神经网络进行微调。结果表明，该方法可以有效地提高静态图像的人脸表情识别性能。

张鹏等[101]提出了一种基于多尺度特征注意机制的人脸表情识别方法，采用两层卷积层提取浅层特征信息。其次，在Inception结构上并行加入空洞卷积（dilated convolution），用于提取多尺度特征，再引入通道注意力机制，加强了模型对有用的特征信息的利用。

Sepas-Moghaddam等[102]首先利用VGG16卷积神经网络提取空间特征。然后，利用Bi-LSTM从视点特征序列中学习空间角度特征，探索前向和后向角度关系。此外，通过注意力机制选择性地关注最重要的空间-角度特征。最后，采用融合方法获得情感识别分类结果。

崔子越等[103]提出了一种改进的Focal Loss和VGGNet相结合的人脸表情识别算法，利用新设计的输出模块来改进VGGNet模型，提高了模型的特征提取能力。通过设置概率阈值来避免错误标记样本对模型性能的负面影响，Focal Loss得到了改进。

郑剑等[104]提出了一种深度卷积神经网络FLF-TAWL（deep convolutional neural network fusing local feature and two-stage attention weight learning）用于融合局部特征和两阶段注意力权重学习。该网络能够自适应地捕获人脸的重要区域，提高人脸表情识别的有效性。

（2）动态视频表情序列

对于动态视频表情序列的深度特征提取，常用的方法有CNN、RNN、LSTM等。Jung等[105]提出了一种联合微调方法来整合两个独立的深层网络，分别使用图像序列和面部标志点进行训练，以学习时间外观特征和时间几何特征。Jaiswal等[106]提出了一种通过使用CNN和BiLSTM的组合来获取时间信息的方法。Fan等[107]提出了一种混合网络，该网络使用3DCNN体系结构进行特征提取，并进一步选择RNN来捕获面部信息的时间相关性。

Kim等[108]研究了情绪状态下面部表情的变化，他们提出了一种结合CNN和LSTM的框架。面部表情的特征编码为两部分：第一部分，CNN学习情绪状态所有帧中面部表情的空间特征；第二部分，通过LSTM来学习时间特征。

Yu等[109]提出了一种称为时空卷积嵌套LSTM（spatio-temporal convolutional features with nested LSTM，STC-NLSTM）的新体系结构，该体系结构基于三个深度学习子网络：用于提取时空特征的3DCNN，用于保持时间动态的时间T-LSTM，对多级特征进行建模的卷积C-LSTM。3DCNN用于从表示面部表情的图像序列中提取时空卷积特征，T-LSTM用于对每个卷积层中的时空特征的时间动态进行建模，并采用C-LSTM将所有T-LSTM的输出集成在一起，从而对网络中间层编码的多级特征进行编码。

Liang等[110]提出了一种用于面部情感识别的深度卷积双向长短时记忆（Bi-LSTM）融合网络，它可以利用空间和时间特征。该框架主要由三部分组成：用于区分性空间表示提取的深空间网络（deep spatial network，DSN）、用于学习时间动力学的深时间网络（deep temporal network，DTN）和用于长期时空特征积累的循环网络。给定一个表示情感类的图像序列，具有更深和更大架构的DSN从序列中的每一帧中学习细微特征，而DTN则通过将两个相邻帧作为输入来关注短期表达变化。为了更好地识别时空信息，Bi-LSTM网络被进一步用于发现数据之间的相关性。此外，该框架是端到端可学习的，因此可以调整时间信息以补充空间特征。

司马懿等[111]使用预先训练好的Inception ResNet v1网络提取每一帧的特征向量，然后计算特征向量之间的欧氏距离来定位表情强度最大的完整帧，从

而得到标准化的人脸表情序列。为了进一步验证定位模型的准确性,分别采用VGG16网络和ResNet50网络对定位后的完整帧进行面部表情识别。

Meng等[112]提出了帧注意网络(frame attention networks,FAN),将具有可变数量人脸图像的视频作为其输入,并生成固定尺寸的表示。整个网络由两个模块组成。特征嵌入模块是一个深度卷积神经网络,它将人脸图像嵌入到特征向量中。帧注意模块学习多个注意权重,这些权重用于自适应地聚合特征向量以形成单个判别视频表示。

Pan等[113]提出了一种基于深度时空网络的视频面部表情识别方法。首先采用空间卷积神经网络和时间卷积神经网络,分别提取视频序列中的高级时空特征。然后组合提取的空间和时间特征输入到融合网络中,进行基于视频的面部表情分类任务。

从上述已有的手工视觉情感特征和深度视觉情感特征文献来看:(1)基于视觉的情感识别可分为基于静态面部图像的表情识别和基于动态视频序列的表情识别。(2)对于静态面部图像的手工特征提取,主要是通过提取图像信息中的几何图形和外貌特征来获得人脸表情特征,常用的方法有LBP、HOG、SIFT等及其改进的方法;对于静态面部图像的深度特征,主要采用基于CNN的网络模型进行面部图像的深度特征提取;对于动态视频表情序列的手工特征提取,捕获视频序列的动态信息才能更有效地表示面部表情的有用信息,常用的方法主要包括光流法和模型法;对于动态视频表情序列的深度特征提取,考虑视频序列的时空性,通常采用基于CNN和RNN的模型来分别提取空间深度特征和时间深度特征。(3)根据现有的文献表明,将视觉手工特征和深度特征相结合,是一个值得深入研究的方向。

3.3 文本情感特征提取

文本情感通常是利用文本信息来传达人的情感。提取文本情感特征是文本情感识别任务的一个关键。首先对文本字符进行转化,转化为可被计算机识别的数值,得到初步的文本特征表示。在此基础上,为了减少信息冗余、降低数据维度,对其进行有效特征提取,用于输入到下游的神经网络中训练,实现最终的情感识别。因此文本情感特征提取是实现文本情感分类的关键,主要分为手工文本情感特征提取和深度文本情感特征提取。

3.3.1 手工文本情感特征

常用的手工文本情感特征提取方法是词袋模型(bag-of-words model,BoW)[114-115]。该模型将文档映射成向量,如 $\boldsymbol{v}=[x_1, x_2, \cdots, x_n]$,其中 x_i 表示基本术语中第 i 个词的出现。这些基本术语是从数据集中收集的,通常是出现频度最高的前 n 个词。出现特征的值可以是二进制、术语频率或TF-IDF(term frequency-inverse document frequency)。二进制值表示在文本中是否出现第 i 个单词,不考虑单词的权重。术语频率表示每个单词出现的次数。一般来说,文本中的高频词汇可以体现文本的某些代表性思想,但某些词在所有文本中可能高频出现。TF-IDF平衡始终具有高频率的单词的权重。它假设一个词的重要性与它在文档中的频率成比例增加,但被它在整个语料库中的频率所抵消[116-117]。尽管BoW模型简单且常用,但它存在高维稀疏性和词间关系缺失的问题,是一种低层次的文本特征表示方法。

为了改进BoW模型,Deerwester等[118]提出了潜在语义分析(latent semantic analysis,LSA),LSA利用奇异值分解(singular value decomposition,SVD)将原始的BoW特征表示转换为具有较低维度的向量。如果原始向量基于频率,则转换的向量也与术语频率近似线性相关。Hofmann[119]提出概率潜在语义分析(probability latent semantic analysis,PLSA),PLSA基于LSA引入了统计概率模型,与标准LSA相比,它的概率变体具有良好的统计基础,并定义了一个合适的数据生成模型,解决了一义多词和一词多义的问题。Blei等[120]提出了潜在狄利克雷分布模型(latent Dirichlet allocation,LDA),是一种离散数据集合(如文本语料库)的生成概率模型。LDA是一个三层贝叶斯概率模型,包含词、主题、文档三层结构,通过将词映射到主题空间,计算出每个词的权重,从而选择文本特征。

3.3.2 深度文本情感特征

词嵌入(word embedding)是一种基于分布式语义建模的技术,一些预训练好的面向深度学习的词嵌入模型被广泛应用于文本情感提取任务。根据编码信息的不同重点,词嵌入可分为两类:典型词嵌入和情感词嵌入。前者侧重于通过建模一般语义和上下文信息来学习连续单词嵌入,而后者侧重于将情感信息编码到单词嵌入中[121]。

早期的词嵌入模型通常基于句法上下文进行训练。他们认为出现频率较高的词在某些语义标准上往往是相似的,例如word2vec[122]和Glove[123]。它们在大量未标记数据上进行训练,目的是捕获细粒度的语法和语义规则。预训练的词嵌入模型比随机初始

化的单词向量具有更好的性能，并且在NLP任务中取得了巨大的成功。然而，早期的词嵌入模型假设“一个词由唯一向量表示”，并忽略了不同上下文信息的影响。它们将每个单词嵌入一个唯一的向量，无论是单义还是多义。这种局限性阻碍了早期单词嵌入模型的有效性。

近年来，受迁移学习的启发，预训练语言模型的出现开启了NLP领域的突破。ELMo（embeddings from language models）[124]是一种新型的深层语境化（deep contextualized）单词表示方法。ELMo是一个深层的双向语言模型，通过捕获词义随上下文的变化动态生成单词嵌入。它可以模拟词语的复杂特征（如同义词和语义）及在不同的语言语境中的语义变化（即多义词）。ELMo可以很容易地转移到现有模型中，并显著改进了六个具有挑战性的NLP问题的最新技术，包括问题回答、情感分析等领域。

近年来，OpenAI提出了基于Transformer[125]的语言模型GPT（generative pre-training）[126]。与ELMo不同，GPT利用上文预测下一个单词。GPT采用两阶段过程，首先在未标记的数据上使用语言建模目标来学习神经网络模型的初始参数，随后使用相应的监督目标使这些参数适应目标任务。GPT在GLUE基准测试的许多句子级任务上取得了先前的最新成果。

BERT（bidirectional encoder representations from transformers）[127]是一种基于Transformer[125]的双向编码表示模型，在所有层中对上下文进行联合调节，通过无监督学习预测上下文中隐藏的单词，从未标记文本中预训练深层双向表示。BERT打破了11项NLP任务的最佳记录。随后越来越多的预训练模型及改进不断出现，如GPT-2[128]、GPT-3[129]、Transformer-XL[130]、XLNet[131]等，推动着NLP领域的不断进步与成熟。

在传统词嵌入的推动下，情感词嵌入在不同的情感任务中取得较大的贡献，如情感分类和情感强度预测。为了将情感信息纳入词语表示，Tang等[132]提出了情感特定词嵌入（sentiment-specific word embeddings，SSWE），它在向量空间中编码情感（积极或消极）和句法上下文信息。与其他词嵌入相比，这项工作证明了将情感标签纳入与情感相关任务的词级信息的有效性。

Felbo等[133]通过训练一个名为DeepMoji的两层Bi-LSTM模型，使用1.2亿条推特数据预测输入文档的情感，在情感任务方面取得了良好的效果。

Xu等[134]提出了Emo2Vec，将情感语义编码为固定大小的实值向量的词级表示，采用多任务学习的方法对Emo2Vec进行了6个不同的情绪相关任务的训练。对Emo2Vec的评估显示，它优于现有的与情感相关的表示方法，并且在训练数据更小的十多个数据集上取得了更好的效果。当与GloVe[123]级联时，Emo2Vec使用简单的逻辑回归分类器在几个任务上取得了与最新结果相当的性能。

Shi等[135]提出了一种新的学习领域敏感和情感感知嵌入的方法，该方法同时捕获单个词的情感语义信息和领域敏感信息，可以自动确定并产生域通用嵌入和域特定嵌入。域公共词和域特定词的区分，使得多个域的通用语义数据增强的优势得以实现，同时捕获不同域的特定词的不同语义。结果表明，该模型提供了一种有效的方法来学习领域敏感和情感感知的单词嵌入，这有利于句子和词汇层面的情感分类。

从上述已有的手工文本情感特征和深度文本情感特征文献来看：（1）常用的手工文本情感特征提取采用的是词袋模型BoW，但它存在高维稀疏性和词间关系缺失的问题，是一种低层次的文本特征表示方法。为了改进BoW模型，继而出现了一系列改进的模型，如LSA、PLSA、LDA等。（2）深度文本情感特征主要以词嵌入的形式表示，一些预训练好的面向深度学习的词嵌入模型被广泛使用于文本情感提取任务，主要分为典型词嵌入和情感词嵌入。常用的词嵌入为word2vec、Glove、BERT等。

4 多模态信息融合方法

情感是以非言语方式发生的动态心理生理过程，这使得情感识别变得复杂。近年来，尽管单模态情感识别任务取得了一些研究成果，但研究表明，多模态的情感识别任务效果优于单一模态[136-137]。研究尝试结合不同模式的信号，如语音、视觉、文本等信息，从而提高情感识别任务的效率和精确度。这部分将重点介绍多模态情感识别中的多模态信息融合方法。常见的融合方法有[24,137]：特征层（feature-level）融合、决策层（decision-level）融合、模型层（model-level）融合等。

特征层融合也被称为早期融合（early fusion，EF），是一种复杂度较低、相对简单的融合方法，考虑了模式之间的相关性。对于多模态，特征层融合直接将单模态提取到的特征级联成一个特征向量，并对其训练分类器，用于情感识别。然而，集成多模态

模式中不同度量级别的特征将显著增加级联特征向量的维数,容易导致维度过高以至于训练模型困难。

决策层融合也被称为后期融合(late fusion,LF),采用某种决策融合规则,将不同的模态视为相互独立的,组合多种单模态的识别结果,得到最终的融合结果。常用的决策融合规则包括“多数投票”“最大”“总和”“最小”“平均”“乘积”等[138]。虽然基于规则的融合方法易于使用,但基于规则的融合面临的困难是如何设计好规则。如果规则过于简单,它们可能无法揭示不同模式之间的关系。决策级融合的优点是来自不同分类器的决策易于比较,并且每个模态可以使用其最适合任务的分类器。

模型层融合近年来广泛应用于情感识别任务,其旨在对每个模态分别建模,同时考虑模态之间的相关性。因此,它可以考虑不同模式之间的相互关联,并且降低了这些模态时间同步的需求。此外,混合融合方法是特征级和决策级策略的组合,因此结合了早期融合和晚期融合两者的优点。

根据采用的单一模态信息的数量,常见的多模态情感识别可以分为双模态情感识别和三模态情感识别。本章从基于双模态和三模态的情感识别这两方面对这些融合方法进行分析,结果如表2所示。

4.1 双模态情感识别

常见的双模态情感识别可以分为:融合语音和视觉信息的音视频情感识别以及融合语音和文本的双模态情感识别。

4.1.1 融合语音和视觉信息的音视频情感识别

Huang等[139]提出利用Transformer模型在模型层面上融合视听模式。利用OpenSMILE提取声学参数集(eGeMAPS)作为音频特征,视觉特征由几何特征构成,包括面部地标位置、面部动作单位、头部姿态特征和眼睛注视特征。多头注意力在编码音视频后,从公共语义特征空间产生多模态情感中间表征,再将Transformer模型与LSTM相结合,通过全连接层得到回归结果,进一步提高了性能。在AVEC 2017[140]数据库上的实验表明,模型级融合优于其他融合策略。

刘菁菁等[141]提出一种基于LSTM网络的多模态情感识别模型。对语音提取了43维手工特征向量,包括MFCC特征、Fbank特征等;对面部图像选取26个人脸特征点间的距离长度作为表情特征。采用双路LSTM分别识别语音和面部表情的情感信息,通过Softmax进行分类,进行决策层加权特征融合。在eNTERFACE’05数据集上,传统情感六分类的准确率达到74.40%;另外,模型层特征融合方法采用双层LSTM的结构,将情感分类特征映射到激活度-效价空间(arousal-valence space),在两个维度上的准确率分别达到84.10%和86.60%。

Liu等[142]提出了一种新的表示融合方法,称为胶囊图卷积网络(capsule graph convolutional network,CapsGCN)。首先,从语音信号中提取声谱图,通过2D-CNN进行特征提取;对图像进行人脸检测,通过VGG16进行视觉特征提取。将提取出的音视频特征输入到胶囊网络,分别封装成多模态胶囊,通过动态路由算法有效地减少数据冗余。其次,将具有相互关系和内部关系的多模态胶囊视为图形结构。利用图卷积网络(GCN)学习图的结构,得到隐藏表示。最后,将CapsGCN学习到的多模态胶囊和隐藏关系表示反馈给多头自注意力,再通过全连接层进行分类。实验表明提出的融合方法在eNTERFACE’05[28]

表2 多模态信息融合方法

Table 2 Multimodal information fusion methods

时间	作者	模态	特征提取	融合方式	分类/回归	数据集	识别结果
2020	Huang等[139]	语音、视觉	语音:eGeMAPS 视觉:几何特征	模型层融合 (Transformer+LSTM)	全连接层	AVEC 2017	CCC(Arousal维度):0.654 CCC(Valence维度):0.708
2020	刘菁菁等[141]	语音、视觉	语音:MFCC、Fbank等 视觉:人脸特征点间的距离长度	特征层融合 决策层融合 模型层融合 (双层LSTM)	Softmax	eNTERFACE’05	Acc(6-class):74.40%
2021	Liu等[142]	语音、视觉	语音:声谱图+2D-CNN 视觉:VGG16	模型层融合 (GapsGCN)	全连接层	eNTERFACE’05	Acc(6-class):80.83% F1-score:80.23%
2021	王传昱等[143]	语音、视觉	语音:DBM+LSTM 视觉:LBPH+SAE+CNN	决策层融合	Softmax	CHEAVD	Acc(6-class):74.90%
2018	Hazarika等[144]	语音、文本	语音:MFCC等 文本:FastText+CNN	特征层融合 (Self-Attention)	Softmax	IEMOCAP	Acc(4-class):71.40% F1-score:71.30%

表2 (续)

时间	作者	模态	特征提取	融合方式	分类/回归	数据集	识别结果
2020	Priyasad等[146]	语音、文本	语音：SincNet层+DCNN 文本：DCNN+Bi-LSTM	特征层融合 决策层融合 模型层融合 （Cross-Attention）	Softmax	IEMOCAP	Acc（4-class）：80.51%
2020	Krishna等[147]	语音、文本	语音：CNN+Bi-LSTM 文本：Glove+CNN	模型层融合 （Cross-Modal Attention）	Softmax	IEMOCAP	Acc（4-class）：72.82%
2021	Lian等[148]	语音、文本	语音：eGeMAPS 文本：Common Crawl and Wikipedia词嵌入	模型层融合 （CTNet：Transformer+GRU）	Softmax	1. IEMOCAP 2. MELD	1. Acc（4-class）：83.60% Acc（6-class）：68.00% F1-score：67.50% 2. Acc（7-class）：62.00% F1-score:60.50%
2021	王兰馨等[149]	语音、文本	语音：IS10 paraling 文本：word2vec+Bi-LSTM+CNN	特征层融合 决策层融合	Softmax	IEMOCAP	Acc（4-class）：69.51%
2017	Poria等[151]	语音、视觉、文本	语音：手工特征，如声音强度、音调及其统计数据 视觉：3D-CNN 文本：word2vec	模型层融合 （AT-Fusion+CAT-LSTM）	Softmax	CMU-MOSI	Acc（2-class）：78.30%
2020	Pan等[152]	语音、视觉、文本	语音：IS13-ComParE 视觉：3D-CNN 文本：word2vec	混和融合（MMAN）	Softmax	IEMOCAP	Acc（4-class）：73.94%
2020	Mittal等[153]	语音、视觉、文本	语音：声学特征，如音高等 视觉：面部识别模型、面部动作单元和面部地标中获得的特征组合 文本：Glove词嵌入	模型层融合（M3ER）	全连接层	1. IEMOCAP 2. CMU-MOSEI	1. Acc（4-class）：82.70% 2. Acc（6-class）：89.00%
2020	Siriwardhana等[154]	语音、视觉、文本	语音：Wav2Vec提取SSL特征 视觉:Fabnet提取SSL特征 文本：RoBERTa提取SSL特征	模型层融合 （Transformer）	Softmax	1. IEMOCAP 2. CMU-MOSEI 3. CMU-MOSI 4. MELD	1. Acc（4-class）：84.65% 2. Acc（7-class）：55.70% 3. Acc（7-class）：55.50% 4. Acc（7-class）：64.30%
2020	Mai等[157]	语音、视觉、文本	语音：双向GRU 视觉：双向GRU 文本：双向GRU	混和融合（MFRM）	全连接层	1. IEMOCAP 2. CMU-MOSEI 3. CMU-MOSI 4. MELD	1. Acc（4-class）：83.45% 2. Acc（7-class）：51.00% 3. Acc（7-class）：39.40% 4. Acc（2-class）：88.19%
2020	Wang等[159]	语音、视觉、文本	语音：OpenSMILE提取声学特征 视觉：3D-CNN 文本：CNN	特征层融合	全连接层	1. IEMOCAP 2. CMU-MOSI 3. MELD	1. Acc（6-class）：60.81% 2. Acc（2-class）：82.71% 3.Acc（7-class）：61.95%
2021	Dai等[160]	语音、视觉、文本	语音：光谱图+CNN 视觉：CNN 文本：Transformer	模型层融合（MESM）	前馈网络	1. IEMOCAP 2. CMU-MOSEI	1. Acc（6-class）：84.40% 2. Acc（6-class）：66.80%
2021	Ren等[161]	语音、视觉、文本	语音：IS13 ComParE 视觉：3D-CNN 文本：Glove	模型层融合（IMAN）	全连接层	IEMOCAP	Acc（6-class）：65.00% F1-score：64.50%
2021	Khare等[162]	语音、视觉、文本	语音：声学特征 视觉：VGG-16 文本：Glove	模型层融合 （Transformer）	全连接层	CMU-MOSEI	Acc（6-class）：66.60% F1-score：78.50%

注：Acc代表平均准确率，F1-score代表F1得分，CCC代表一致性相关系数。

上取得了80.83%的准确率和80.23%的F1得分。

王传昱等[143]提出了一种基于音视频的决策融合方法。对视频图像，利用局部二进制模式直方图（local binary patterns histograms，LBPH）、稀疏自动编码器（sparse auto-encoder，SAE）和改进的CNN来实现；对于语音模态，基于改进深度受限波尔兹曼机和LSTM来实现。在单模态识别后，根据权重准则将两种模态的识别结果进行融合，通过Softmax进行分类。在CHEAVD数据集上的实验结果表明，识别率达到了74.90%。

4.1.2 融合语音和文本的双模态情感识别

Hazarika等[144]提出了一种基于自注意力（self-attention）的特征级融合方法。对语音提取高维手工特征，如响度、音高、声音质量、梅尔光谱、MFCC等；对文本采用FastText[145]嵌入字典进行编码，再用CNN进行特征提取；该注意力机制为这些模态分配适当的分数，然后将这些分数用作加权组合的权重，最后通过Softmax进行分类。在IEMOCAP（前四个会话作为训练集，第五个会话作为测试集）数据集上的实验表明，该融合方法在四分类的情感识别率达到了71.40%。

Priyasad等[146]提出了一种基于深度学习的方法来融合文本和声音数据进行情感分类。利用SincNet层和深度卷积神经网络（DCNN）从原始音频中提取声学特征，级联两个并行分支（其一为DCNN，其二为Bi-RNN与DCNN串联）进行文本特征提取，再引入交叉注意力（cross-attention）来推断从Bi-RNN收到的隐藏表示上的N-gram级相关性，最后通过Softmax进行分类。该方法在IEMOCAP（10折交叉验证）数据集上进行了评估，实验结果表明，该系统性能优于现有方法，加权精度提高0.052。

Krishna等[147]提出了一种利用跨模态注意力（cross-modal attention）和基于原始波形的一维卷积神经网络进行语音-文本情感识别的新方法。他们使用音频编码器（CNN+Bi-LSTM）从原始音频波形中提取高级特征，并使用文本编码器（词嵌入Glove[123]+CNN）从文本中提取高级语义信息；使用跨模态注意力，其中音频编码器的特征关注文本编码器的特征，反之亦然，再通过Softmax进行分类。实验表明，该方法在IEMOCAP（四个会话作为训练集，一个会话作为测试集，做交叉验证）数据集上获得了最新的结果。与之前最先进的方法相比，得到0.019的精度绝对提升。

Lian等[148]提出了一个用于会话情感识别的多模态学习框架，称为CTNet（conversational transformer network），使用基于Transformer来建模多模态特征之间的模态内和模态间的交互。利用OpenSMILE提取88维的声学特征（eGeMAPS），在Common Crawl and Wikipedia数据集上训练的300维词向量作为文本特征。为了建模上下文敏感和说话人敏感的依赖关系，使用了基于多头注意力的双向GRU网络和说话人嵌入，通过Softmax进行分类。在IEMOCAP（前四个会话用作训练集和验证集，第五个会话用作测试集）和MELD（十折交叉验证）数据集上的实验结果表明了该方法的有效性，与其他方法相比在加权平均F1得分上表现出0.021~0.062的性能提升。

王兰馨等[149]提出了基于Bi-LSTM-CNN的语音-文本情感识别算法。提取word2vec[150]词嵌入作为文本特征，再经过Bi-LSTM和CNN模型进行文本特征提取，对语音利用OpenSMILE进行手工声学特征提取（IS10_paraling），将两者特征融合的结果作为联合CNN模型的输入，通过Softmax进行分类，进行情感识别。基于IEMOCAP（四个会话作为训练集，一个会话作为测试集）的结果表明，情感识别准确率达到了69.51%。

4.2 基于三模态的情感识别

Poria等[151]提出了一个能够捕捉话语间上下文信息的循环模型。他们使用CNN进行文本特征提取，将话语表示为word2vec向量的矩阵；使用OpenSMILE提取音频特征，提取的特征由几个底层描述符组成，如声音强度、音调及其统计数据；用3D-CNN对视频中图像序列进行特征提取。他们提出了一个基于上下文注意力的LSTM（contextual attention-based LSTM，CAT-LSTM）模型来模拟话语之间的上下文关系，之后引入了一种基于注意力的融合机制（attention-based fusion，AT-Fusion），它在多模态分类融合过程中放大了更高质量和信息量的模式，最后通过Softmax进行分类。结果显示，该模型在CMU-MOSI（训练集（含1 447个话语）、测试集（含752个话语）划分与说话人无关）数据集上比最先进的技术提高了0.06~0.08。

Pan等[152]提出了一种称为多模态注意力网络（multi-modal attention network，MMAN）的混合融合方法。利用OpenSMILE提取语音手工特征（IS13-ComParE），通过3D-CNN提取视觉特征，提取word2vec词嵌入作为文本特征。他们提出了一种新

的多模态注意力机制（cLSTM-MMA），通过三种模式促进注意力，并选择性地融合信息，最后通过Softmax进行分类。MMAN在IEMOCAP（训练集、测试集随机划分）情感识别数据库上实现了最先进的性能。

Mittal等[153]提出了一个使用乘法融合层的多模态情感识别模型，称为M3ER。该方法学习更可靠的模态，并在样本基础上抑制较弱的模态。提取Glove词嵌入作为文本特征，对语音模态提取声学特征，如音高等，从最先进的面部识别模型、面部动作单元和面部地标中获得的特征组合作为视觉特征。通过引入典型相关分析来区分无效模态和有效模态，再生成代理功能来代替无效的模态，最后通过全连接层进行分类。实验结果表明，在IEMOCAP上的平均准确率为82.70%，在CMU-MOSEI（随机划分为训练集（70%）、验证集（10%）和测试集（20%））上的平均准确率为89.00%，总体来说比以往研究提高了约0.05的准确率。

Siriwardhana等[154]首次使用从独立预训练的自监督学习（self supervised learning，SSL）模型中提取的SSL特征来表示文本（采用RoBERTa[155]）、语音（采用Wav2Vec）和视觉（采用Fabnet）的三种输入模态。鉴于SSL特征的高维特性，引入了一种新的Transformer和基于注意力的融合机制，最后通过Softmax获得最终分类结果。该机制可以结合多模态SSL特征并实现多模态情感识别任务的最新结果。对该方法进行了基准测试和评估，在四个数据集IEMOCAP（前四个会话作为训练集，第五个会话作为测试集）、CMU-MOSEI（使用了CMU-SDK[156]中提供的标签和数据集拆分）、CMU-MOSI（使用CMU-SDK[156]中提供的标签和数据集拆分）、MELD上的结果表明该方法优于最先进的模型。

Mai等[157]提出了多融合残差记忆网络（multi-fusion residual memory network，MFRM）来识别话语级情感。对语音、视觉及文本模态采用双向GRU模型来获得每个模态的特征表示。在MFRM中，提出了情感强度注意，使MFRM能够关注发生强烈情感或重大情感变化的时间步长，并引入时间步长级融合来建模时间受限的模式间交互。此外，还提出了残差记忆网络（residual memory network，RMN）来处理融合特征。最后，通过全连接层得到分类结果。大量实验表明，MFRM在CMU-MOSI（1 284个话语作为训练集，686个话语作为测试集）、CMU-MOSEI（16 265个话语作为训练集，4 643个话语作为测试集）、IEMOCAP（前四个会话作为训练集，第五个会话作为测试集）、IMDB[158]数据集上实现了最先进的结果。

Wang等[159]受Transformer最近在机器翻译领域取得成功的启发，提出了一种新的融合方法Trans-Modality来解决多模态情感分析的任务。文本、视觉和声学特征分别通过CNN、3D-CNN和OpenSMILE进行提取。通过Transformer，学习的特征体现了源模态和目标模态的信息，再通过全连接层进行分类。在多个多模态数据集CMU-MOSI（训练集、验证集包含1 447个话语，测试集包含752个话语）、MELD（训练集、验证集包含11 098个话语，测试集包含2 610个话语）、IEMOCAP（训练集、验证集包含5 810个话语，测试集包含1 623个话语）上验证了该模型。实验表明，提出的方法达到了最先进的性能。

Dai等[160]提出了一个完全端到端的模型（multi-modal end-to-end sparse model，MESM）将特征提取和多模态建模这两个阶段连接起来，并对它们进行联合优化。对于语音和视觉模态中的每个光谱图块和图像帧，采用CNN进行特征提取；对文本采用Transformer进行编码。为了减少端到端模型带来的计算开销，引入了稀疏跨模态注意力（cross-modal attention）进行特征提取，最后通过前馈网络得到分类结果。在IEMOCAP（将70%、10%和20%的数据分别随机分配到训练集、验证集和测试集）和CMU-MOSEI（随机划分）上的实验结果表明，完全端到端模型明显优于基于两阶段的现有模型。此外，通过添加稀疏的跨模态注意力，该模型可以在特征提取部分以大约一半的计算量保持相当的性能。

Ren等[161]提出了一种新的交互式多模态注意网络（interactive multimodal attention network，IMAN）用于对话中的情绪识别。利用OpenSMILE对语音信息提取声学特征（IS13 ComParE），利用3D-CNN提取视觉特征，提取Glove词嵌入作为文本特征。IMAN引入了一个跨模态注意融合模块来捕获多模态信息的跨模态交互，并采用了一个会话建模模块来探索整个对话的上下文信息和说话者依赖性，最后通过全连接层得到分类结果。在IEMOCAP（前四个会话作为训练集，最后一个为测试集）数据集上的实验结果表明，IMAN在加权平均精度和F1-得分方面分别达到了0.004和0.002的提升。

Khare等[162]将自监督训练扩展到多模态情感识

别中,对一个基于Transformer训练的掩码语言模型进行预训练,使用音频(声学特征)、视觉(VGG16提取的深度特征)和文本(Glove词嵌入)特征作为输入,最后通过全连接层进行分类。该模型对情感识别的下游任务进行了微调。在CMU-MOSEI数据集上的研究结果表明,与基线水平相比,自监督训练模型可以提高高达0.03的情感识别性能。

来自不同模态的信息对最终情感识别性能的贡献是不同的,模型应该更加关注融合过程中提供更多信息的模态。传统的特征融合和决策融合方法无法考虑模态之间的交互影响,因此近年来逐渐从传统融合方法走向模型层融合。随着注意力机制的不断改进,考虑到注意力机制能够学习不同模态对识别性能的影响,注意力机制在多模态融合中扮演着越来越重要的作用。

5 挑战与机遇

5.1 深度学习技术的自身缺陷

到目前为止,各种深度学习方法已经成功地应用于学习高级特征表示以进行情感特征识别。此外,这些深度学习方法通常优于基于手工特征的其他方法。然而,这些使用的深度学习技术具有大量的网络参数,导致其计算复杂度高。为了缓解这个问题,越来越多的学者着手对深度网络的压缩和加速的研究。剪枝(pruning)[163-164]是减少深度神经网络(DNN)参数数量的一种强有力的技术。在DNN中,许多参数是冗余的,在训练过程中对降低误差没有很大的贡献。因此,在训练之后,这些参数可以从网络中移除,移除这些参数对网络精度的影响最小。剪枝的主要目的是减少模型的存储需求并使其便于存储。如He等[163]引入了一种新的通道剪枝方法来加速深度卷积神经网络。给定一个训练好的CNN模型,提出了一个迭代的两步算法,通过基于LASSO(least absolute shrinkage and selection operator)回归的通道选择和最小二乘重建来有效地修剪每一层。进一步将该算法推广到多层和多分支的情况。修剪后的VGG16以5倍的加速达到了最先进的结果,同时对ResNet、Xception等网络实现了2倍的加速。

尽管就各种特征学习任务的性能衡量而言,深度学习已经成为一种最先进的技术,但黑盒问题仍然存在。深层模型的多个隐藏层究竟学习到了什么样的内部表示尚未可知。由于其多层非线性结构,深度学习技术通常被认为不透明,其预测结果往往无法被人追踪。为了缓解这个问题,直接可视化学习到的特征已经成为理解深度模型的广泛使用的方式[165]。然而,这种可视化的方式并没有真正提出相关的理论来解释这个算法到底在做什么。因此,从多模态情感识别的理论角度探讨深度学习技术的可解释性[166]是一个重要的研究方向。

5.2 跨库的多模态情感识别

多模态情感识别技术虽然有了巨大的发展,但在跨语言的环境中,仍然是一个具有挑战性但至关重要的问题。由于数据采集和注释环境的不同,不同数据集之间往往存在数据偏差和注释不一致。现在的多模态情感识别往往在同一个数据集进行训练和测试,大多数研究人员通常在一个特定的数据集中验证他们提出的方法的性能,且当下的跨库情感识别也大多为单模态的情感识别任务。由于需要联合处理多个数据源,这比单模态情感识别系统具有更大的复杂性。因此如何进行跨库的多模态情感识别也是未来的一个挑战。近年来,新发展起来的对抗学习方法是一种可行的跨库多模态情感识别策略。常见的对抗学习网络有生成性对抗网络(generative adversarial networks,GAN)[167]、对抗式自动编码器(adversarial autoencoder)[168]等。

学习各种模式的联合嵌入空间对于多模态融合至关重要。主流模态融合方法未能实现这一目标,留下了影响跨模态融合的模态缺口。Mai等[169]提出了一种新的对抗编码器-解码器-分类器框架来学习一个模态不变的嵌入空间。由于各种模态的分布在本质上是不同的,为了减少模态差异,使用对抗训练通过各自的编码器将源模态的分布转换为目标模态的分布。进一步通过引入重构损失和分类损失对嵌入空间施加额外约束。然后使用层次图神经网络融合编码表示,明确了多阶段的单峰、双峰和三峰相互作用。该方法在多个数据集上取得了最新的性能。因此在后续的多模态情感任务中,将对抗学习方法应用于多模态融合是一个值得深入研究的方向。

5.3 集合更多的模态信息

以前的研究主要集中在依靠面部表情、语音和文本来评估人类的情绪状态。然而,这些类型的输入数据是相对主观的,并且缺乏足够的客观特征来准确标记一个人的情绪。因此增加更多的模态信息进行研究是一个值得探索的问题。最近,人们开始使用基于情感识别方法的生理信号[170],这种方法更加客观,适合于对情绪状态进行连续实时监测。常

用于检测情绪的生理信号包括脑电图（electroencephalogram，EEG）、心电图（electrocardiogram，ECG）、皮肤电反应（galvanic skin response，GSR）、皮肤温度（skin temperature，ST）和光容积图（photoplethysmogram，PPG）等[171]。在情感识别系统中使用两个或两个以上的信号可以极大地提高整体准确性。

此外，虽然面部表情自动情绪识别取得了显著的进展，但身体手势的情感识别尚未得到深入的探索。人们经常使用各种各样的身体语言来表达情感，很难列举所有的情绪身体手势，并为每个类别收集足够的样本。因此，识别新的情绪性身体手势对于更好地理解人类情绪至关重要。然而，现有的方法并不能准确地确定一个新的身体姿势属于哪种情绪状态。身体语言作为传递情感信息的重要因素，在情感识别中尚未得到深入的研究[172]。人们经常使用各种各样的身体语言来表达情感，但很难列举所有的情感身体姿势种类，并为每个类别收集足够的样本。目前主流的算法主要将现有的身体检测和特征提取技术应用到情感分类任务中，但并不能准确地确定一个新的身体姿势属于哪种情感状态。因此，识别新的情感身体姿势对于情感识别至关重要。

此外，目前的一些语义融合策略，如多视图融合、迁移学习融合和概率依赖融合，在多模态数据的语义融合方面取得了一些进展[173]。因此，将深度学习和语义融合策略结合起来，可能对多模态情感识别带来一个新的研究方法。

5.4 小样本学习

在数据收集困难、缺乏数据的情况下，对情感识别任务而言是一个巨大的挑战。例如大多数基于身体手势的情感识别数据集只包含几百个样本，且大部分收集的是实验者在实验室环境中执行的行为。这种收集方法大多由实验设计者预先指定，且姿势种类较少。然而，人们表达情感的方式是不同的，随之产生不同的身体姿势。当在模型测试过程中出现一个新的身体手势时，算法很容易识别错误。解决小样本问题的一种方法是扩展训练数据集，以包括尽可能多的情感身体手势。然而，收集所有类别的标记数据都是巨大的工作量。

对于数据小样本问题，零次学习（zero-shot learning，ZSL）是一种较好的解决方法[172]。ZSL可以通过属性和语义向量的等边信息建立可见类别和不可见类别之间的关联。例如它为身体姿势这个问题提供了一个解决方法，即使用它们的语义描述来识别新的身体姿势类别，然后从身体姿势标签中推断出情感类别。因此，在情感识别任务中，对小样本学习方法的深入研究及不断改进是未来值得探索的一个方向。

6 总结

本文对近年来面向深度学习的融合语音、视觉、文本等模态信息的多模态情感识别技术进行了系统性分析与总结。详细阐述了几种具有代表性的深度学习技术，如DBN、CNN、LSTM及其改进方法；介绍了近年来国内外的多模态情感数据库，重点介绍了近年来深度学习技术在多模态情感识别领域中的研究进展，如基于深度学习的单一模态情感特征提取方法及多模态信息融合策略。此外，给出了未来进一步提高多模态情感识别性能的几个具有挑战性的研究方向。

参考文献：

[1] DINO H I, ABDULRAZZAQ M B. Facial expression classification based on SVM, KNN and MLP classifiers[C]//Proceedings of the 2019 International Conference on Advanced Science and Engineering, Duhok, Apr 2-4, 2019. Piscataway: IEEE, 2019: 70-75.

[2] PERVEEN N, ROY D, CHALAVADI K M. Facial expression recognition in videos using dynamic kernels[J]. IEEE Transactions on Image Processing, 2020, 29: 8316-8325.

[3] SHRIVASTAVA V, RICHHARIYA V, RICHHARIYA V. Puzzling out emotions: a deep-learning approach to multimodal sentiment analysis[C]//Proceedings of the 2018 International Conference on Advanced Computation and Telecommunication, Bhopal, Dec 28-29, 2018. Piscataway: IEEE, 2018: 1-6.

[4] SCHERER K R. Psychological models of emotion[J]. The Neuropsychology of Emotion, 2000, 137(3): 137-162.

[5] AMMEN S, ALFARRAS M, HADI W. OFDM system performance enhancement using discrete wavelet transform and DSSS system over mobile channel[R]. Advances in Computer Science and Engineering, 2010: 142-147.

[6] LIANG J J, CHEN S Z, JIN Q. Semi-supervised multimodal emotion recognition with improved Wasserstein GANs[C]//Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Lanzhou, Nov 18-21, 2019. Piscataway: IEEE, 2019: 695-703.

[7] AL-SULTAN M R, AMEEN S Y, ABDUALLAH W M. Real time implementation of stegofirewall system[J]. International Journal of Computing Digital Systems, 2019, 8(5): 498-504.

[8] ZHANG Y Y, WANG Z R, DU J. Deep fusion: an attention

guided factorized bilinear pooling for audio-video emotion recognition[C]//Proceedings of the 2019 International Joint Conference on Neural Networks, Budapest, Jul 14-19, 2019. Piscataway: IEEE, 2019: 1-8.

[9] CHEN J, LV Y, XU R, et al. Automatic social signal analysis: facial expression recognition using difference convolution neural network[J]. Journal of Parallel Distributed Computing, 2019, 131: 97-102.

[10] GHALEB E, POPA M, ASTERIADIS S. Multimodal and temporal perception of audio-visual cues for emotion recognition[C]//Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction, Cambridge, Sep 3-6, 2019. Piscataway: IEEE, 2019: 552-558.

[11] ABDULRAZZAQ M B, KHALAF K I. Handwritten numerals' recognition in Kurdish language using double feature selection[C]//Proceedings of the 2019 2nd International Conference on Engineering Technology and Its Applications, Al-Najef, Aug 27-28, 2019. Piscataway: IEEE, 2019: 167-172.

[12] CAIHUA C. Research on multi-modal Mandarin speech emotion recognition based on SVM[C]//Proceedings of the 2019 IEEE International Conference on Power, Intelligent Computing and Systems, Shenyang, Jul 12-14, 2019. Piscataway: IEEE, 2019: 173-176.

[13] SCHULLER B W, VALSTER M F, EYBEN F, et al. AVEC 2012: the continuous audio/visual emotion challenge[C]// Proceedings of the 2012 International Conference on Multimodal Interaction, Santa Monica, Oct 22-26, 2012. New York: ACM, 2012: 449-456.

[14] DHALL A, GOECKE R, JOSHI J, et al. Emotion recognition in the wild challenge 2013[C]//Proceedings of the 2013 International Conference on Multimodal Interaction, Sydney, Dec 9-13, 2013. New York: ACM, 2013: 509-516.

[15] STAPPEN L, BAIRD A, RIZOS G, et al. MuSe 2020 Challenge and Workshop: multimodal sentiment analysis, emotion target engagement and trustworthiness detection in real-life media: emotional car reviews in-the-wild[C]//Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop, Seattle, Oct 16, 2020. New York: ACM, 2020: 35-44.

[16] STAPPEN L, MEßNER E M, CAMBRIA E, et al. MuSe 2021 challenge: multimodal emotion, sentiment, physiological-emotion, and stress detection[C]//Proceedings of the 2021 ACM International Conference on Multimedia, Oct 20-24, 2021. New York: ACM, 2021: 5706-5707.

[17] LI Y, TAO J H, SCHULLER B W, et al. MEC 2016: the multimodal emotion recognition challenge of CCPR 2016 [C]//Proceedings of the 7th Chinese Conference on Pattern Recognition, Chengdu, Nov 5-7, 2016. Cham: Springer, 2016: 667-678.

[18] OBAID K B, ZEEBAREE S, AHMED O M. Deep learning models based on image classification: a review[J]. International Journal of Science Business, 2020, 4(11): 75-81.

[19] ZHAO X, SHI X, ZHANG S. Facial expression recognition via deep learning[J]. IETE Technical Review, 2015, 32(5): 347-355.

[20] SCHMIDHUBER J. Deep learning in neural networks: an over-view[J]. Neural Networks, 2015, 61: 85-117.

[21] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[22] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[23] ELMAN J L. Finding structure in time[J]. Cognitive Science, 1990, 14(2): 179-211.

[24] D'MELLO S K, KORY J. A review and meta-analysis of multi-modal affect detection systems[J]. ACM Computing Surveys, 2015, 47(3): 1-36.

[25] RISH I. An empirical study of the naive Bayes classifier [C]//Proceedings of the 2001 Workshop on Empirical Methods in Artificial Intelligence, Seattle, 2001: 41-46.

[26] KEERTHI S S, SHEVADE S K, BHATTACHARYYA C, et al. Improvements to Platt's SMO algorithm for SVM classifier design[J]. Neural Computation, 2001, 13(3): 637-649.

[27] WINDEATT T. Accuracy/diversity and ensemble MLP classifier design[J]. IEEE Transactions on Neural Networks, 2006, 17(5): 1194-1211.

[28] MARTIN O, KOTSIA I, MACQ B, et al. The eNTERFACE' 05 audio-visual emotion database[C]//Proceedings of the 22nd International Conference on Data Engineering Workshops, Atlanta, Apr 3-7, 2006. Washington: IEEE Computer Society, 2006: 8.

[29] WANG Y, GUAN L. Recognizing human emotional state from audiovisual signals[J]. IEEE Transactions on Multimedia, 2008,10(5): 936-946.

[30] BUSSO C, BULUT M, LEE C C, et al. IEMOCAP: interactive emotional dyadic motion capture database[J]. Language Resources and Evaluation, 2008, 42(4): 335-359.

[31] DHALL A, GOECKE R, LUCEY S, et al. Collecting large, richly annotated facial-expression databases from movies[J]. IEEE Multimedia, 2012, 19(3): 34-41.

[32] ZHALEHPOUR S, ONDER O, AKHTAR Z, et al. BAUM-1: a spontaneous audio-visual face database of affective and mental states[J]. IEEE Transactions on Affective Computing, 2016, 8(3): 300-313.

[33] ZADEH A, ZELLERS R, PINCUS E, et al. Multimodal senti-

ment intensity analysis in videos: facial gestures and verbal messages[J]. IEEE Intelligent Systems, 2016, 31(6): 82-88.

[34] PEREPELKINA O, KAZIMIROVA E, KONSTANTINOVA M. RAMAS: Russian multimodal corpus of dyadic interaction for affective computing[C]//LNCS 11096: Proceedings of the 20th International Conference on Speech and Computer, Leipzig, Sep 18-22, 2018. Cham: Springer, 2018: 501-510.

[35] LIVINGSTONE S R, RUSSO F A. The Ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English[J]. PLoS One, 2018, 13(5): e0196391.

[36] ZADEH A B, LIANG P P, PORIA S, et al. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Jul 15-20, 2018. Stroudsburg: ACL, 2018: 2236-2246.

[37] PORIA S, HAZARIKA D, MAJUMDER N, et al. MELD: a multimodal multi-party dataset for emotion recognition in conversations[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics, Florence, Jul 28-Aug 2, 2019. Stroudsburg: ACL, 2019: 527-536.

[38] YU W, XU H, MENG F, et al. CH-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Jul 5-10, 2020. Stroudsburg: ACL, 2020: 3718-3727.

[39] CHEN J, WANG C H, WANG K J, et al. HEU emotion: a large-scale database for multimodal emotion recognition in the wild[J]. Neural Computing and Applications, 2021, 33 (14): 8669-8685.

[40] DENG L, YU D. Deep learning: methods and applications [J]. Foundations and Trends in Signal Processing, 2014, 7(3/4): 197-387.

[41] FREUND Y, HAUSSLER D. Unsupervised learning of distributions of binary vectors using 2-layer networks[C]//Advances in Neural Information Processing Systems 4, Denver, Dec 2-5, 1991. San Mateo: Morgan Kaufmann, 1991: 912-919.

[42] BENGIO Y, LAMBLIN P, POPOVICI D, et al. Greedy layer-wise training of deep networks[C]//Proceedings of the 20th Annual Conference on Neural Information Processing Systems, Vancouver, Dec 4-7, 2006. Cambridge: MIT Press, 2007: 153-160.

[43] HINTON G E. Training products of experts by minimizing contrastive divergence[J]. Neural Computation, 2002, 14(8): 1771-1800.

[44] LEE H, GROSSE R B, RANGANATH R, et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations[C]//Proceedings of the 26th International Conference on Machine Learning, Montreal, Jun 14-18, 2009. New York: ACM, 2009: 609-616.

[45] WANG G M, QIAO J F, BI J, et al. TL-GDBN: growing deep belief network with transfer learning[J]. IEEE Transactions on Automation Science and Engineering, 2018, 16(2): 874-885.

[46] DENG W, LIU H L, XU J J, et al. An improved quantum-inspired differential evolution algorithm for deep belief network[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 69(10): 7319-7327.

[47] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[48] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [C]//Proceedings of the 26th Annual Conference on Neural Information Processing Systems 2012, Lake Tahoe, Dec 3-6, 2012. Red Hook: Curran Associates, 2012: 1106-1114.

[49] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv:1409.1556, 2014.

[50] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, Jun 7-12, 2015. Washington: IEEE Computer Society, 2015: 1-9.

[51] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27-30, 2016. Washington: IEEE Computer Society, 2016: 770-778.

[52] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 4700-4708.

[53] TRAN D, BOURDEV L D, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks [C]//Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Dec 7-13, 2015. Washington: IEEE Computer Society, 2015: 4489-4497.

[54] YANG H, YUAN C F, LI B, et al. Asymmetric 3D convolutional neural networks for action recognition[J]. Pattern Recognition, 2019, 85: 1-12.

[55] KUMAWAT S, RAMAN S. LP-3DCNN: unveiling local phase in 3D convolutional neural networks[C]//Proceedings

of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-20, 2019. Piscataway: IEEE, 2019: 4903-4912.

[56] CHEN H, WANG Y, SHU H, et al. Frequency domain compact 3D convolutional neural networks[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 13-19, 2020. Piscataway: IEEE, 2020: 1638-1647.

[57] WERBOS P J. Backpropagation through time: what it does and how to do it[J]. Proceedings of the IEEE, 1990, 78(10): 1550-1560.

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

[59] CHUNG J, GÜLÇEHRE Ç, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv:1412.3555, 2014.

[60] ZHAO R, WANG K, SU H, et al. Bayesian graph convolution LSTM for skeleton based action recognition[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 6881-6891.

[61] ZHANG S, ZHAO X, TIAN Q. Spontaneous speech emotion recognition using multiscale deep convolutional LSTM [J]. IEEE Transactions on Affective Computing, 2019. DOI: 10.1109/TAFFC.2019.2947464.

[62] XING Y, DI CATERINA G, SORAGHAN J. A new spiking convolutional recurrent neural network (SCRNN) with applications to event-based hand gesture recognition[J]. Frontiers in Neuroscience, 2020, 14: 1143.

[63] 高庆吉, 赵志华, 徐达, 等. 语音情感识别研究综述[J]. 智能系统学报, 2020, 15(1): 1-13.
GAO Q J, ZHAO Z H, XU D, et al. Review on speech emotion recognition research[J]. CAAI Transactions on Intelligent Systems, 2020, 15(1): 1-13.

[64] 刘振焘, 徐建平, 吴敏, 等. 语音情感特征提取及其降维方法综述[J]. 计算机学报, 2018, 41(12): 2833-2851.
LIU Z T, XU J P, WU M, et al. Review of emotional feature extraction and dimension reduction method for speech emotion recognition[J]. Chinese Journal of Computers, 2018, 41 (12): 2833-2851.

[65] 韩文静, 李海峰, 阮华斌, 等. 语音情感识别研究进展综述[J]. 软件学报, 2014, 25(1): 37-50.
HAN W J, LI H F, RUAN H B, et al. Review on speech emotion recognition[J]. Journal of Software, 2014, 25(1): 37-50.

[66] 郑纯军, 王春立, 贾宁. 语音任务下声学特征提取综述[J]. 计算机科学, 2020, 47(5): 110-119.
ZHENG C J, WANG C L, JIA N. Survey of acoustic feature extraction in speech tasks[J]. Computing Science, 2020, 47(5): 110-119.

[67] LISCOMBE J, VENDITTI J, HIRSCHBERG J B. Classifying subject ratings of emotional speech using acoustic features[C]//Proceedings of the 8th European Conference on Speech Communication and Technology, Geneva, Sep 1-4, 2003.

[68] YACOUB S M, SIMSKE S J, LIN X F, et al. Recognition of emotions in interactive voice response systems[C]//Proceedings of the 8th European Conference on Speech Communication and Technology, Geneva, Sep 1-4, 2003.

[69] SCHMITT M, RINGEVAL F, SCHULLER B W. At the border of acoustics and linguistics: bag-of-audio-words for the recognition of emotions in speech[C]//Proceedings of the 17th Annual Conference of the International Speech Communication Association, San Francisco, Sep 8-12, 2016: 495-499.

[70] 孙韩玉, 黄丽霞, 张雪英, 等. 基于双通道卷积门控循环网络的语音情感识别[J/OL]. 计算机工程与应用(2021-10-18)[2022-02-28]. https://kns.cnki.net/kcms/detail/11.2127.TP.20211015.2021.002.html.
SUN H Y, HUANG L X, ZHANG X Y, et al. Speech emotion recognition based on dual-channel convolutional gated recurrent network[J/OL]. Computer Engineering and Applications (2021-10-18)[2022-02-28]. https://kns.cnki.net/kcms/detail/11.2127.TP.20211015.2021.002.htm.

[71] LUENGO I, NAVAS E, HERNÁEZ I. Feature analysis and evaluation for automatic emotion identification in speech[J]. IEEE Transactions on Multimedia, 2010, 12(6): 490-501.

[72] DUTTA K, SARMA K K. Multiple feature extraction for RNN-based assamese speech recognition for speech to text conversion application[C]//Proceedings of the 2012 International Conference on Communications, Devices and Intelligent Systems, Kolkata, Dec 28-29, 2012. Piscataway: IEEE, 2013: 1-6.

[73] MAO Q, DONG M, HUANG Z, et al. Learning salient features for speech emotion recognition using convolutional neural networks[J]. IEEE Transactions on Multimedia, 2014, 16(8): 2203-2213.

[74] 陈婧, 李海峰, 马琳, 等. 多粒度特征融合的维度语音情感识别方法[J]. 信号处理, 2017, 33(3): 374-382.
CHEN J, LI H F, MA L, et al. Multi-granularity feature fusion for dimensional speech emotion recognition[J]. Journal of Signal Processing, 2017, 33(3): 374-382.

[75] 俞佳佳, 金赟, 马勇, 等. 基于Sinc-Transformer模型的原始语音情感识别[J]. 信号处理, 2021, 37(10): 1880-1888.
YU J J, JIN Y, MA Y, et al. Emotion recognition from raw speech based on Sinc-Transformer model[J]. Journal of Signal Processing, 2021, 37(10): 1880-1888.

[76] ZHANG S Q, ZHAO X M, CHUANG Y L, et al. Feature

learning via deep belief network for Chinese speech emotion recognition[C]//Proceedings of the 7th Chinese Conference on Pattern Recognition, Chengdu, Nov 5-7, 2016. Cham: Springer, 2016: 645-651.

[77] OTTL S, AMIRIPARIAN S, GERCZUK M, et al. Group-level speech emotion recognition utilising deep spectrum features[C]//Proceedings of the 2020 International Conference on Multimodal Interaction, Oct 25-29, 2020. New York: ACM, 2020: 821-826.

[78] EYBEN F, WÖLLMER M, SCHULLER B W. OpenSMILE: the munich versatile and fast open-source audio feature extractor[C]//Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Oct 25-29, 2010. New York: ACM, 2010: 1459-1462.

[79] 蒋斌, 钟瑞, 张秋闻, 等. 采用深度学习方法的非正面表情识别综述[J]. 计算机工程与应用, 2021, 57(8): 48-61.
JIANG B, ZHONG R, ZHANG Q W, et al. Survey of non-frontal facial expression recognition by using deep learning methods[J]. Computer Engineering and Applications, 2021, 57(8): 48-61.

[80] 李珊, 邓伟洪. 深度人脸表情识别研究进展[J]. 中国图象图形学报, 2020, 25(11): 2306-2320.
LI S, DENG W H. Deep facial expression recognition: a survey[J]. Journal of Image and Graphics, 2020, 25(11): 2306-2320.

[81] MELLOUK W, HANDOUZI W. Facial emotion recognition using deep learning: review and insights[J]. Procedia Computer Science, 2020, 175: 689-694.

[82] ZHAO X, ZHANG S. A review on facial expression recognition: feature extraction and classification[J]. IETE Technical Review, 2016, 33(5): 505-517.

[83] CHEN J, LIU X, TU P, et al. Learning person-specific models for facial expression and action unit recognition[J]. Pattern Recognition Letters, 2013, 34(15): 1964-1970.

[84] ZHANG S, ZHAO X, LEI B. Facial expression recognition based on local binary patterns and local Fisher discriminant analysis[J]. WSEAS Transactions on Signal Processing, 2012, 8(1): 21-31.

[85] CHU W S, DE LA TORRE F, COHN J F. Selective transfer machine for personalized facial expression analysis[J]. IEEE Transactions on Pattern Analysis Machine Intelligence, 2016, 39(3): 529-545.

[86] BALTRUŠAITIS T, MAHMOUD M, ROBINSON P. Cross-dataset learning and person-specific normalisation for automatic action unit detection[C]//Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, Ljubljana, May 4- 8, 2015. Washington: IEEE Computer Society, 2015: 1-6.

[87] AHSAN T, JABID T, CHONG U P. Facial expression recognition using local transitional pattern on Gabor filtered facial images[J]. IETE Technical Review, 2013, 30(1): 47-52.

[88] 刘军, 景晓军, 孙松林, 等. 一种用于人脸识别的基于主导近邻像素的局部Gabor空间直方图特征[J]. 北京邮电大学学报, 2015, 38(1): 51-54.
LIU J, JING X J, SUN S L, et al. Feature of local gabor spatial histogram based on dominant neighboring pixel for face recognition[J]. Journal of Beijing University of Posts and Telecommunications, 2015, 38(1): 51-54.

[89] BAH S M, MING F. An improved face recognition algorithm and its application in attendance management system [J]. Array, 2020, 5: 100014.

[90] DEEBA F, AHMED A, MEMON H, et al. LBPH-based enhanced real-time face recognition[J]. International Journal of Advanced Computer Science and Applications, 2019, 10 (5): 274-280.

[91] ZHANG T, ZHENG W, CUI Z, et al. A deep neural network-driven feature learning method for multiview facial expression recognition[J]. IEEE Transactions on Multimedia, 2016, 18(12): 2528-2536.

[92] YEASIN M, BULLOT B, SHARMA R. From facial expression to level of interest: a spatio-temporal approach[C]//Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, Jun 27-Jul 2, 2004. Washington: IEEE Computer Society, 2004: 922-927.

[93] FAN X, TJAHJADI T. A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences[J]. Pattern Recognition, 2015, 48(11): 3407-3416.

[94] BOSCH A, ZISSERMAN A, MUÑOZ X. Representing shape with a spatial pyramid kernel[C]//Proceedings of the 6th ACM International Conference on Image and Video Retrieval, Amsterdam, Jul 9-11, 2007. New York: ACM, 2007: 401-408.

[95] 刘涛, 周先春, 严锡君. 基于光流特征与高斯LDA的面部表情识别算法[J]. 计算机科学, 2018, 45(10): 286-290.
LIU T, ZHOU X C, YAN X J. LDA facial expression recognition algorithm combining optical flow characteristics with Gaussian[J]. Computing Science, 2018, 45(10): 286-290.

[96] HAPPY S, ROUTRAY A. Fuzzy histogram of optical flow orientations for micro-expression recognition[J]. IEEE Transactions on Affective Computing, 2017, 10(3): 394-406.

[97] 邵洁, 董楠. RGB-D动态序列的人脸自然表情识别[J]. 计算机辅助设计与图形学学报, 2015, 27(5): 847-854.

SHAO J, DONG N. Spontaneous facial expression recognition based on RGB-D dynamic sequences[J]. Journal of Computer-Aided Design & Computer Graphics, 2015, 27(5): 847-854.

[98] YI J, CHEN A, CAI Z, et al. Facial expression recognition of intercepted video sequences based on feature point movement trend and feature block texture variation[J]. Applied Soft Computing, 2019, 82: 105540.

[99] YOLCU G, OZTEL I, KAZAN S, et al. Facial expression recognition for monitoring neurological disorders based on convolutional neural network[J]. Multimedia Tools and Applications, 2019, 78(22): 31581-31603.

[100] SUN N, LI Q, HUAN R, et al. Deep spatial-temporal feature fusion for facial expression recognition in static images[J]. Pattern Recognition Letters, 2019, 119: 49-61.

[101] 张鹏, 孔韦韦, 滕金保. 基于多尺度特征注意力机制的人脸表情识别[J]. 计算机工程与应用, 2022, 58(1): 182-189.
ZHANG P, KONG W W, TENG J B. Facial expression recognition based on multi-scale feature attention mechanism[J]. Computer Engineering and Applications, 2022, 58(1): 182-189.

[102] SEPAS-MOGHADDAM A, ETEMAD S A, PEREIRA F, et al. Facial emotion recognition using light field images with deep attention-based bidirectional LSTM[C]//Proceedings of the IEEE 2020 International Conference on Acoustics, Speech and Signal Processing, Barcelona, May 4-8, 2020. Piscataway: IEEE, 2020: 3367-3371.

[103] 崔子越, 皮家甜, 陈勇, 等. 结合改进VGGNet和Focal Loss的人脸表情识别[J]. 计算机工程与应用, 2021, 57(19): 171-178.
CUI Z Y, PI J T, CHEN Y, et al. Facial expression recognition combined with improved VGGNet and Focal Loss[J]. Computer Engineering and Applications, 2021, 57(19): 171-178.

[104] 郑剑, 郑炽, 刘豪, 等. 融合局部特征与两阶段注意力权重学习的面部表情识别[J]. 计算机应用研究, 2022, 39(3): 889-894.
ZHENG J, ZHENG C, LIU H, et al. Deep convolutional neural network fusing local feature and two-stage attention weight learning for facial expression recognition[J]. Application Research of Computers, 2022, 39(3): 889-894.

[105] JUNG H, LEE S, YIM J, et al. Joint fine-tuning in deep neural networks for facial expression recognition[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Dec 7-13, 2015. Washington: IEEE Computer Society, 2015: 2983-2991.

[106] JAISWAL S, VALSTAR M F. Deep learning the dynamic appearance and shape of facial action units[C]//Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision, Lake Placid, Mar 7-10, 2016. Washington: IEEE Computer Society, 2016: 1-8.

[107] FAN Y, LU X J, LI D, et al. Video-based emotion recognition using CNN-RNN and C3D hybrid networks[C]//Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Nov 12-16, 2016. New York: ACM, 2016: 445-450.

[108] KIM D H, BADDAR W J, JANG J, et al. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition[J]. IEEE Transactions on Affective Computing, 2017, 10(2): 223-236.

[109] YU Z, LIU G, LIU Q, et al. Spatio-temporal convolutional features with nested LSTM for facial expression recognition[J]. Neurocomputing, 2018, 317: 50-57.

[110] LIANG D, LIANG H, YU Z, et al. Deep convolutional BiLSTM fusion network for facial expression recognition [J]. The Visual Computer, 2020, 36(3): 499-508.

[111] 司马懿, 易积政, 陈爱斌, 等. 动态人脸图像序列中表情完全帧的定位与识别[J]. 应用科学学报, 2021, 39(3): 357-366.
SIMA Y, YI J Z, CHEN A B, et al. Fully expression frame localization and recognition based on dynamic face image sequences[J]. Journal of Applied Sciences, 2021, 39(3): 357-366.

[112] MENG D B, PENG X J, WANG K, et al. Frame attention networks for facial expression recognition in videos[C]//Proceedings of the 2019 IEEE International Conference on Image Processing, Taipei, China, Sep 22-25, 2019. Piscataway: IEEE, 2019: 3866-3870.

[113] PAN X, ZHANG S, GUO W, et al. Video-based facial expression recognition using deep temporal-spatial networks [J]. IETE Technical Review, 2020, 37(4): 402-409.

[114] SOUMYA G K, JOSEPH S. Text classification by augmenting bag of words (BOW) representation with co-occurrence feature[J]. IOSR Journal of Computer Engineering, 2014, 16(1): 34-38.

[115] ZHAO R, MAO K. Fuzzy bag-of-words model for document representation[J]. IEEE Transactions on Fuzzy Systems, 2017, 26(2): 794-804.

[116] TRSTENJAK B, MIKAC S, DONKO D. KNN with TF-IDF based framework for text categorization[J]. Procedia Engineering, 2014, 69: 1356-1364.

[117] KIM D, SEO D, CHO S, et al. Multi-co-training for document classification using various document representations: TFIDF, LDA, and Doc2Vec[J]. Information Sciences, 2019, 477: 15-29.

[118] DEERWESTER S C, DUMAIS S T, LANDAUER T K, et al. Indexing by latent semantic analysis[J]. Journal of the Association for Information Science & Technology, 1990, 41(6): 391-407.

[119] HOFMANN T. Probabilistic latent semantic analysis[C]// Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, Stockholm, Jul 30-Aug 1, 1999. San Mateo: Morgan Kaufmann, 1999: 289-296.

[120] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[121] DENG J W, REN F J. A survey of textual emotion recognition and its challenges[J]. IEEE Transactions on Affective Computing, 2021: 1.

[122] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in Neural Information Processing Systems 26, Lake Tahoe, Dec 5-8, 2013. Red Hook: Curran Associates, 2013: 3111-3119.

[123] PENNINGTON J, SOCHER R, MANNING C D. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Oct 25-29, 2014. Stroudsburg: ACL, 2014: 1532-1543.

[124] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Jun 1-6, 2018. Stroudsburg: ACL, 2018: 2227-2237.

[125] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30, Dec 4-9, 2017. Red Hook: Curran Associates, 2017: 5998-6008.

[126] CHUNG Y A, GLASS J R. Generative pre-training for speech with autoregressive predictive coding[C]//Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, May 4-8, 2020. Piscataway: IEEE, 2020: 3497-3501.

[127] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Jun 2-7, 2019. Stroudsburg: ACL, 2019: 4171-4186.

[128] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8): 9.

[129] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems 33, Dec 6-12, 2020: 1877-1901.

[130] DAI Z H, YANG Z L, YANG Y M, et al. Transformer-XL: attentive language models beyond a fixed-length context [C]//Proceedings of the 57th Conference of the Association for Computational Linguistics, Florence, Jul 28-Aug 2, 2019. Stroudsburg: ACL, 2019: 2978-2988.

[131] YANG Z L, DAI Z H, YANG Y M, et al. XLNet: generalized autoregressive pretraining for language understanding[C]//Advances in Neural Information Processing Systems 32, Vancouver, Dec 8-14, 2019: 5754-5764.

[132] TANG D, WEI F, YANG N, et al. Learning sentiment-specific word embedding for twitter sentiment classification[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2014: 1555-1565.

[133] FELBO B, MISLOVE A, SØGAARD A, et al. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Sep 9-11, 2017. Stroudsburg: ACL, 2017: 1615-1625.

[134] XU P, MADOTTO A, WU C S, et al. Emo2Vec: learning generalized emotion representation by multi-task training [C]//Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Brussels, Oct 31, 2018. Stroudsburg: ACL, 2018: 292-298.

[135] SHI B, FU Z, BING L, et al. Learning domain-sensitive and sentiment-aware word embeddings[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Jul 15-20, 2018. Stroudsburg: ACL, 2018: 2494-2504.

[136] ABDULLAH S M S A, AMEEN S Y A, SADEEQ M A, et al. Multimodal emotion recognition using deep learning [J]. Journal of Applied Science and Technology Trends, 2021, 2(2): 52-58.

[137] SHOUMY N J, ANG L M, SENG K P, et al. Multimodal big data affective analytics: a comprehensive survey using text, audio, visual and physiological signals[J]. Journal of Network and Computer Applications, 2020, 149: 102447.

[138] SUN Z, SONG Q, ZHU X, et al. A novel ensemble method for classifying imbalanced data[J]. Pattern Recognition, 2015, 48(5): 1623-1637.

[139] HUANG J, TAO J H, LIU B, et al. Multimodal transformer fusion for continuous emotion recognition[C]//Procee-

dings of the IEEE 2020 International Conference on Acoustics, Speech and Signal Processing, Barcelona, May 4-8, 2020. Piscataway: IEEE, 2020: 3507-3511.

[140] RINGEVAL F, SCHULLER B W, VALSTAR M F, et al. AVEC 2017: reallife depression, and affect recognition workshop and challenge[C]//Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, Oct 23-27, 2017. New York: ACM, 2017: 3-9.

[141] 刘菁菁, 吴晓峰. 基于长短时记忆网络的多模态情感识别和空间标注[J]. 复旦学报(自然科学版), 2020, 59(5): 565-574.
LIU J J, WU X F. Real-time multimodal emotion recognition and emotion space labeling using LSTM networks[J]. Journal of Fudan University (Natural Science), 2020, 59 (5): 565-574.

[142] LIU J X, CHEN S, WANG L B, et al. Multimodal emotion recognition with capsule graph convolutional based representation fusion[C]//Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, Jun 6-11, 2021. Piscataway: IEEE, 2021: 6339-6343.

[143] 王传昱, 李为相, 陈震环. 基于语音和视频图像的多模态情感识别研究[J]. 计算机工程与应用, 2021, 57(23): 163-170.
WANG C Y, LI W X, CHEN Z H. Reserch of multi-modal emotion recognition based on voice and video images[J]. Computer Engineering and Applications, 2021, 57(23): 163-170.

[144] HAZARIKA D, GORANTLA S, PORIA S, et al. Self-attentive feature-level fusion for multimodal emotion detection[C]//Proceedings of the IEEE 1st Conference on Multimedia Information Processing and Retrieval, Miami, Apr 10-12, 2018. Piscataway: IEEE, 2018: 196-201.

[145] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146.

[146] PRIYASAD D, FERNANDO T, DENMAN S, et al. Attention driven fusion for multi-modal emotion recognition[C]//Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, May 4-8, 2020. Piscataway: IEEE, 2020: 3227-3231.

[147] KRISHNA D N, PATIL A. Multimodal emotion recognition using cross-modal attention and 1D convolutional neural networks[C]//Proceedings of the 21st Annual Conference of the International Speech Communication Association, Shanghai, Oct 25-29, 2020: 4243-4247.

[148] LIAN Z, LIU B, TAO J H. CTNet: conversational transformer network for emotion recognition[J]. IEEE/ACM Transactions on Audio, Speech, Language Processing, 2021, 29: 985-1000.

[149] 王兰馨, 王卫亚, 程鑫. 结合Bi-LSTM-CNN的语音文本双模态情感识别模型[J]. 计算机工程与应用, 2022, 58(4): 192-197.
WANG L X, WANG W Y, CHENG X. Bimodal emotion recognition model for speech-text based on Bi-LSTM-CNN[J]. Computer Engineering and Applications, 2022, 58(4): 192-197.

[150] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv: 13013781, 2013.

[151] PORIA S, CAMBRIA E, HAZARIKA D, et al. Multi-level multiple attentions for contextual multimodal sentiment analysis[C]//Proceedings of the 2017 IEEE International Conference on Data Mining, New Orleans, Nov 18-21, 2017. Washington: IEEE Computer Society, 2017: 1033-1038.

[152] PAN Z X, LUO Z J, YANG J C, et al. Multi-modal attention for speech emotion recognition[C]//Proceedings of the 21st Annual Conference of the International Speech Communication Association, Oct 25-29, 2020: 364-368.

[153] MITTAL T, BHATTACHARYA U, CHANDRA R, et al. M3ER: multiplicative multimodal emotion recognition using facial, textual, and speech cues[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, the 32nd Innovative Applications of Artificial Intelligence Conference, the 10th AAAI Symposium on Educational Advances in Artificial Intelligence, New York, Feb 7-12, 2020. Menlo Park: AAAI, 2020: 1359-1367.

[154] SIRIWARDHANA S, KALUARACHCHI T, BILLINGHURST M, et al. Multimodal emotion recognition with transformer-based self supervised feature fusion[J]. IEEE Access, 2020, 8: 176274-176285.

[155] LIU Y, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[J]. arXiv:190711692, 2019.

[156] ZADEH A, LIANG P P, PORIA S, et al. Multi-attention recurrent network for human communication comprehension[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, Feb 2-7, 2018. Menlo Park: AAAI, 2018: 5642-5649.

[157] MAI S J, HU H F, XU J, et al. Multi-fusion residual memory network for multimodal human sentiment compre-

hension[J]. IEEE Transactions on Affective Computing, 2022, 13(1): 320-334.

[158] MAAS A, DALY R E, PHAM P T, et al. Learning word vectors for sentiment analysis[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Jun 19-24, 2011. Stroudsburg: ACL: 142-150.

[159] WANG Z L, WAN Z H, WAN X J. TransModality: an End2End fusion method with transformer for multimodal sentiment analysis[C]//Proceedings of the Web Conference 2020, Taipei, China, Apr 20-24, 2020. New York: ACM, 2020: 2514-2520.

[160] DAI W L, CAHYAWIJAYA S, LIU Z H, et al. Multimodal end-to-end sparse model for emotion recognition[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Jun 6-11, 2021. Stroudsburg: ACL, 2021: 5305-5316.

[161] REN M, HUANG X, SHI X, et al. Interactive multimodal attention network for emotion recognition in conversation [J]. IEEE Signal Processing Letters, 2021, 28: 1046-1050.

[162] KHARE A, PARTHASARATHY S, SUNDARAM S. Self-supervised learning with cross-modal transformers for emotion recognition[C]//Proceedings of the 2021 IEEE Spoken Language Technology Workshop, Shenzhen, Jan 19-22, 2021. Piscataway: IEEE, 2021: 381-388.

[163] HE Y H, ZHANG X Y, SUN J. Channel pruning for accelerating very deep neural networks[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017. Washington: IEEE Computer Society, 2017: 1389-1397.

[164] LI H, KADAV A, DURDANOVIC I, et al. Pruning filters for efficient ConvNets[J]. arXiv:1608.08710, 2016.

[165] ESCALANTE H J, KAYA H, SALAH A A, et al. Modeling, recog nizing, and explaining apparent personality from videos[J]. IEEE Transactions on Affective Computing, 2020: 1.

[166] ANGELOV P, SOARES E. Towards explainable deep neural networks (xDNN)[J]. Neural Networks, 2020, 130: 185-194.

[167] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems 27, Montreal, Dec 8-13, 2014. Red Hook: Curran Associates, 2014: 2672-2680.

[168] MAKHZANI A, SHLENS J, JAITLY N, et al. Adversarial auto encoders[J]. arXiv:1511.05644, 2015.

[169] MAI S J, HU H F, XING S L. Modality to modality translation: an adversarial representation learning and graph fusion network for multimodal fusion[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, the 32nd Innovative Applications of Artificial Intelligence Conference, the 10th AAAI Symposium on Educational Advances in Artificial Intelligence, New York, Feb 7-12, 2020. Menlo Park: AAAI, 2020: 164-172.

[170] 王忠民, 赵玉鹏, 郑镕林, 等. 脑电信号情绪识别研究综述[J]. 计算机科学与探索, 2022, 16(4): 760-774.
WANG Z M, ZHAO Y P, ZHENG R L, et al. A survey of research on EGG signal emotion recognition[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(4): 760-774.

[171] YANG C J, FAHIER N, LI W C, et al. A convolution neural network based emotion recognition system using multimodal physiological signals[C]//Proceedings of the 2020 IEEE International Conference on Consumer Electronics, Taoyuan, China, Sep 28-30, 2020. Piscataway: IEEE, 2020: 1-2.

[172] WU J, ZHANG Y, ZHAO X, et al. A generalized zero-shot framework for emotion recognition from body gestures[J]. arXiv:2010.06362, 2020.

[173] GAO J, LI P, CHEN Z K, et al. A survey on deep learning for multimodal data fusion[J]. Neural Computation, 2020, 32(5): 829-864.

赵小明(1964—),男,浙江临海人,硕士,教授,主要研究方向为音频和图像处理、机器学习、模式识别等。

ZHAO Xiaoming, born in 1964, M.S., professor. His research interests include audio and image processing, machine learning, pattern recognition, etc.

杨轶娇(1997—),女,江苏南通人,硕士研究生,主要研究方向为情感计算、模式识别等。

YANG Yijiao, born in 1997, M.S. candidate. Her research interests include emotional computing, pattern recognition, etc.

张石清(1980—),男,湖南衡阳人,博士,教授,主要研究方向为情感计算、模式识别等。

ZHANG Shiqing, born in 1980, Ph.D., professor. His research interests include emotional computing, pattern recognition, etc.