计算机科学与探索
Journal of Frontiers of Computer Science and Technology
ISSN 1673-9418,CN 11-5602/TP

《计算机科学与探索》网络首发论文

题目：多模态的情感分析技术综述
作者：刘继明，张培翔，刘颖，张伟东，房杰
网络首发日期：2021-03-25
引用格式：刘继明，张培翔，刘颖，张伟东，房杰．多模态的情感分析技术综述．计算机科学与探索．
https://kns.cnki.net/kcms/detail/11.5602.TP.20210324.1400.016.html

网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

网络首发时间：2021-03-25 17:10:46
网络首发地址：https://kns.cnki.net/kcms/detail/11.5602.TP.20210324.1400.016.html

计算机科学与探索
Journal of Frontiers of Computer Science and Technology doi: 10.3778/j.issn.1673-9418.2012075

多模态的情感分析技术综述

刘继明[1+]，张培翔[2]，刘　颖[2,3,4]，张伟东[2,4]，房　杰[2,3,4]

1. 西安邮电大学 通信与信息工程学院，西安 710121
2. 西安邮电大学 图像与信息处理研究所，西安 710121
3. 陕西省无线通信与信息处理技术国际合作研究中心，西安 710121
4. 西安邮电大学 电子信息现场勘验应用技术公安部重点实验室，西安 710121
+ 通信作者 E-mail:liuying_ciip@163.com

摘　要：情感分析是指利用计算机自动分析确定人们所要表达的情感，其在人机交互和刑侦破案等领域都能发挥重大作用。深度学习和传统特征提取算法的进步为利用多种模态进行情感分析提供了条件。结合多种模态进行情感分析可以弥补单模态情感分析的不稳定性以及局限性等缺点，能够有效提高准确度。近年来，研究者多用面部表情信息、文本信息以及语音信息三种模态进行情感分析。本文主要从这三种模态对多模态情感分析技术进行综述：首先对多模态情感分析的基本概念以及研究现状进行简要介绍；其次总结了常用的多模态情感分析数据集；然后分别对现有的基于面部表情信息、文本信息和语音信息的单模态情感分析技术进行简要叙述；接下来详细介绍了模态融合技术，并依据不同的模态融合方式对多模态情感分析技术的现有成果进行重点描述；最后讨论了多模态情感分析存在的问题以及未来的发展方向。

关键词：多模态；情感分析；模态融合

文献标志码：A　**中图分类号：**TP391

Summary of Multi-modal Sentiment Analysis Technology

LIU Jiming[1+], ZHANG Peixiang[2], LIU Ying[2,3,4], ZHANG Weidong[2,4], FANG Jie[2,3,4]

1.School of Communications and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China
2.Center for Image and Information Processing, Xi'an University of Posts and Telecommunications, Xi'an 710121, China
3.International Joint Research Center for Wireless Communication and Information Processing Technology of Shaanxi Province, Xi'an 710121, China
4. Key Laboratory of Electronic Information Application Technology for Crime Scene Investigation, Ministry of Public Security, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

Abstract: Sentiment analysis refers to the use of computers to automatically analyze and determine the emotions that people want to express. It can play a significant role in human-computer interaction and criminal investigation and solving cases. The advancement of deep learning and traditional feature extraction algorithms provides conditions for the use of multiple modalities for sentiment analysis. Combining multiple modalities for sentiment analysis can make

基金项目：国家自然科学基金青年项目（No.61801381）。
This work was supported by the National Natural Science Foundation of China Youth Project (No. 61801381).

up for the instability and limitations of single-modal sentiment analysis, and can effectively improve accuracy. In recent years, researchers have used three modalities of facial expression information, text information, and voice information to perform sentiment analysis. This article mainly summarizes the multi-modal sentiment analysis technology from these three modalities: firstly, it briefly introduces the basic concepts and research status of multi-modal sentiment analysis; secondly, it summarizes the commonly used multi-modal sentiment analysis data sets; A brief description of the existing single-modal emotion analysis technology based on facial expression information, text information and voice information; the following describes the modal fusion technology in detail, and the multi-modal emotion analysis technology based on different modal fusion methods The current results of the paper are mainly described; finally, the problems of multi-modal sentiment analysis and the future development direction are discussed.

Key words: Multimodal; Sentiment Analysis; Modal Fusion

情感是生物对外界价值关系产生的主观反应，也是生物智能的重要组成部分[1]。在日常生活中，我们一般都是通过面部表情来获取他人的情感状态，但是某一些情况下，我们也会根据语气、肢体动作等其他一些细微的变化来获取他人的情感状态。在服务型机器人、审讯、娱乐等方面需要通过计算机的帮助来获得人类准确的情感状态，所以情感分析体现了越来越重要的研究价值。

情感分析的理论和算法构建涉及人工智能（Artificial Intelligence, AI）、计算机视觉（Computational Vision, CV）和自然语言处理（NLP）等多个方面，是一个多学科交叉的研究领域。早在二十世纪的时候，Ekma[2]等人就将人类的情感分为愤怒、厌恶、恐惧、快乐、悲伤和惊讶六种基本情感，奠定了当今表情识别的基础。在后来的研究中，蔑视也被认为是人类的基本情感之一。

在现有的文献中，主要根据面部表情、文本、以及语音中的一种模态来对情感进行分析。在面部表情识别(facial expression recognition, FER)中，传统的方法主要有基于几何和外观的方法。基于几何的方法虽然简单易行，但是容易忽略局部细节信息。基于外观的方法主要是根据面部的纹理变化来判断情绪的变化，具有良好的光照不变性。在面部的纹理特征提取中，局部二值模式(local binary pattern, LBP)和 Gabor 小波因具有较好的性能而被广泛应用。情感极性是指积极、消极以及中性的情感状态。通过文本分析得到情感极性的方法又称为意见挖掘，传统的方法是基于情感词典，该方法通过人为构建情感词典并将其作为工具来判断情感极性。由于情感词典中情感词的不完整，该方法具有很大的局限性。语音情感分析主要是提取语音中的韵律、音质等特征来进行分析。近年来，随着深度学习的发展，面部表情、文本和语音三种模态都尝试用深度学习的方法来进行情感分析。在基于深度学习的方法中，面部表情信息主要用卷积神经网络（convolutional neural networks, CNN）、深度神经网络（deep neural networks, DNN）以及与传统方法相结合进行情感分析；文本信息主要用循环神经网络（necurrent neural network, RNsN）、长短期记忆网络（Long Short-Term Memory, LSTM）来进行情感分析；语音情感分析主要用支持向量机（support vector machine,SVM）、隐马尔科夫模型(HMM)等来进行分析。基于深度学习的方法在这三种模态的情感分析中都取得了不错的效果，但是由于数据集等原因，在训练模型时仍然存在一些不可避免的误差。

在情感分析的发展过程中，许多研究者用一种模态来进行情感分析。由于用单模态来进行情感分析时

只能在该模态获得情感信息，所以在某些情况下有很多局限性。如图 1 所示，在对人物进行情感分析时，若仅仅考虑文本信息，会得到一样的结果，只有结合面部表情后才能得到正确的情感极性。随着研究的深入，为了解决单模态的局限性，研究者开始结合两种或两种以上的模态来实现跨模态的情感分析。多模态的情感分析有效解决了单模态的局限性，并且提高了结果的准确度。图 2 显示了一个多模态情感分析的框架。该框架包含两个基本步骤：分别处理单模态的数据和将处理后的数据进行融合。这两个步骤都很重要，如果单一模态的数据处理不好，会对多种模态的情感分析结果产生负面影响，而融合方式的性能不好会破坏多模态系统的稳定性[3]。

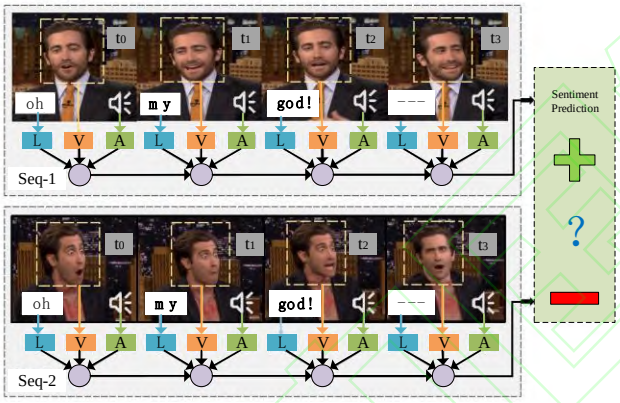


Fig.1 Limitations of a single mode

图 1 单一模态的局限性

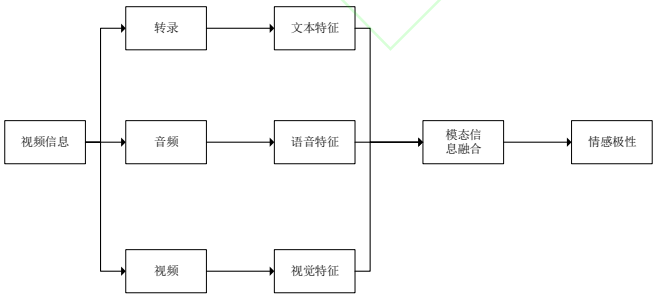


Fig.2 Framework for multimodal sentiment analysis

图 2 多模态情感分析的框架

在情感分析中，目前常用的信息有面部表情信息、文本信息和语音信息，也有一些研究者尝试用姿态、脑部信息来进行情感分析。多模态的情感分析是指由两种及两种以上的模态信息结合来进行情感分析。在特征提取阶段，多模态的情感分析与单模态的特征提取方法相同。利用多模态和单模态进行情感分析最大的区别就是在于多模态需要将单模态的信息进行融合，从而得到情感极性。结合现有文献，模态融合主要包括三种方法，分别是特征级融合、决策级融合以及混合融合。

在多模态情感分析发展过程中，学者从不同的角度对现有的技术进行了总结。文献[4]通过基于视觉信息、语音信息、文本信息以及脑部信息的情感分析分别对现有的技术进行了总结。文献[5]对情感识别、意见挖掘和情绪分析做了详细介绍和区分，并且对情感分析所用到的文本、语音和视觉三种模态的技术做了分类总结。文献[6]对现有的单模态情感分析技术进行讨论，然后对近几年的多模态情感分析文献进行概括总结的同时指出了其模态融合的方法。文献[7]从基于深度学习的角度对现有的模态融合算法进行了归纳总结。与上述综述相比，本文在介绍单模态情感分析技术的基础上着重对多模态情感分析进行归纳总结，并且对文中提到的算法进行对比分析，最后重点介绍了多模态融合技术并对现有问题进行总结。

在本文中，第 1 节总结了现有的多模态的情感分析数据集，第 2 节讨论了单模态的情感分析技术，之后在第 3 节对多模态情感分析技术以及现有的模态融合技术进行了详细阐述，最后在第 4 节对情感分析技术存在的挑战进行简要叙述。

1 多模态情感分析数据集

目前国内外多模态情感数据库大多来源于网络视频评论或人为制作，对于科研领域仍是半公开或者不公开的状态。由于模态选择的不同以及数据集的局限性，一些研究者会根据自己的需求来建立所需要的情感数据集。用于多模态情感分析的可用数据集大多是从不同在线视频共享平台上的产品评论收集的。表 1 总结了常用的多模态情感分析数据集。

Table 1 Summary of commonly used multi-modal sentiment analysis data sets

表 1 常用的多模态情感分析数据集汇总

数据集名称	语言	所含模态	情感标签
SEED 数据集	中文	脑电	积极、中性、消极三分类
新浪微博数据集	中文	文本、图像	积极、消极、中性三分类
Yelp 数据集	英文	文本、图像	1-5 的五个情感分数
Multi-ZOL 数据集	英文	文本、图像	1-10 的十个分数
DEAP 数据集	英文	脑电、视觉	消极到积极 1-9 的九个分数
CH-SIMS 数据集	中文	文本、图像、音频	-1(负)、0(中性)、1(正)
YouTube 数据集	英文	文本、图像、音频	积极、消极、中性三分类
ICT-MMMO 数据集	英文	文本、图像、音频	积极、消极、中性三分类
MOSI 数据集	英文	文本、图像、音频	从-3 到+3 的七类情感倾向
News Rover Sentiment 数据集	英文	文本、图像、音频	积极、消极、中性三分类
IEMOCAP 数据集	英文	文本、图像、音频、姿态等	快乐、愤怒、悲伤等十个标签

SEED 数据集[8]该数据集收集了 15 名(男性 7 名，女性 8 名)受试者在观看 15 个中国电影剪辑时的脑电信号。其标签为积极、中性和消极三种。

新浪微博数据集[9]数据集收集了新浪微博中关于新闻以及娱乐八卦的评论,共包括 6171 条评论,其中有 4196 条肯定消息，1354 条否定消息和 621 条中性消息，5859 条消息具有一个伴随图像。情感标注为三分类。

Yelp 数据集[10]该数据集从 Yelp.com 评论网站收集关于餐厅和食品的评论。一共有 44305 条评论和 233569 张图片，其中每条评论有 13 个句子，23 个单词。情感标注为 1-5 的五个分数。

Multi-ZOL 数据集[11]该数据集收集了关于 5288 条多模态的关于手机的评论信息，其中每条数据至少包含一个文本内容和一个图像级。情感标注为 1-10 的十个分数。

DEAP 数据集[12]该数据集收集了 32 名(一半男一半女) 受试者在观看音乐视频时的生理信号和受试者对视频的 Valence, Arousal, Dominance, Liking 的心理量表，同时也包括前 22 名参与者的面部表情视频。标签为消极到积极 1-9 的九个分数。

CH-SIMS 数据集[13]该数据集中包含 60 个原始视频，剪辑出 2281 个视频片段，每个片段长度不小于一秒且不大于十秒。在每个视频片段中，除了说话者的面部以外不会出现其他面部，且只包含普通话。数据集的情感标注为-1(负)、0(中性)或 1(正)三种。

YouTube 数据集[14]该数据集包含从 YouTube 上收集整理的 47 个不同产品的评论视频。视频由不同年龄、不同种族背景的 20 名女性以及 27 名男性对产品的观点讲述组成，且所有视频长度都被规范为 30 秒。在进行标注时，三名人员随机观看并用积极、消极、中性三种标签对视频进行标注。该数据集共包含 13 个积极、22 个中性以及 12 个消极标签的视频序列。

ICT-MMMO 数据集[15]该数据集包含了来自 YouTube 和 ExpoTV 中的 370 个关于电影评论的视频。视频中不同的人对着摄像机表达 1 ~ 3 分钟的电影评论。此数据集中包括 228 个正面评论、23 个中立评论和 119 个负面评论。

MOSI 数据集[16]该数据集包含了 YouTube 上的 93 个关于电影评论的视频博客。视频中包括年龄为 20 ~ 30 岁以及来自不同种族背景的 41 位女性和 48 位男性的 2 ~ 5 分钟的电影评论。数据集中拥有从-3 到+3 的视频标签，代表七类情感倾向。

News Rover Sentiment 数据集[17]该数据集是新闻领域的数据集，由各种新闻节目和频道视频中的 929 个 4 到 15 秒的视频组成。该数据集的标注为三分类。

IEMOCAP 数据集[18]该数据集包含了 5 个男演员和 5 个女演员在情感互动过程中的大约 12 个小时视听数据，该数据包括对话者的音频、视频、文本、面部和姿态信息等。情感标签为愤怒、快乐、悲伤、中立等十个标签。

2 单模态的情感分析算法

情感分析主要是通过一些表达情感的方式（比如面部表情等）对人们的情感进行分析。目前，主流的单模态的情感分析主要有基于面部表情信息和基于文本信息的情感分析。

不同的人物在表达情感时的方式不同，当一个人趋向于用语言表达情感时，那么其音频特征可能包含较多的情感线索，如果一个人趋向于用面部表情来进行情感表达，那么其面部表情特征可能包含较多的情感线索。由于人们多用说话方式的改变、音调的高低或者面部表情的变化对自己的情感状态进行表达，所以本节将重点介绍基于面部表情信息、文本信息以及语音信息的情感分析技术。

2.1 基于面部表情的情感分析

在日常生活中，面部表情信息是人们相互获得情感状态的常用方式，所以面部表情信息在情感分析的过程中有很重要的意义。根据特征表示的不同，FER 系统可分为静态图像的 FER 和动态序列的 FER 两大类[9]。在动态序列的 FER 中，面部表情呈现出两个特点：空时性和显著性。动态序列的 FER 中常常忽略面部表情的显著性，为了解决这一问题，文献[19]提出一种基于空时注意力网络的面部表情识别方法，该方法在空域子网络和时域子网络中加入相应的注意力模块，来提高 CNN 和 RNN 提取特征时的性能。

面部表情识别过程包括三个阶段，分别是人脸检测，特征提取与选择以及分类。根据所采用的特征表示，可分为传统方法和基于深度学习的方法。

2.1.1 传统的 FER 方法

目前，FER 中常用特征有几何特征、外观特征、统计特征和运动特征等。基于几何特征的方法是对人脸构建几何特征矢量，且每幅图像只保存一个特征矢量；基于外观特征的方法主要对面部的纹理特征进行提取，目前常用的纹理特征主要有：LBP、基于频率域的 Gabor 小波特征等；基于整体统计特征的方法可以尽可能多的保留图像中的主要信息，目前主要有主成分分析(principal component analysis, PCA)和独立主元分析(independent component correlation algorithm, ICA)；基于运动特征的方法对动态图像序列中的运动特征进行提取，常用的是光流法。表 2 从概念和优缺点两方面对传统的 FER 特征提取方法进行了总结。

Table 2　Traditional FER feature extraction methods

表 2　传统的 FER 特征提取方法

特征提取方法	方法简单描述	优点	缺点
基于几何	提取人脸眼睛、鼻子等重要特征点的位置，按照相同的比率将人脸用集合特征矢量表示。	存储量小，对光照变化不敏感。	容易造成部分重要信息丢失。
LBP	一种用来描述图像局部纹理特征的算子。	具有旋转不变性和灰度不变性等、对光照变化不敏感。	作为低层次的特征，不易直接用于匹配和识别。
Gabor 变换	通过定义不同的带宽和方向对图像进行多分辨率分析，能有效提取图像中的纹理特征。	有明显的方向选择性和频率选择性，且对光照变化不敏感。	特征维度过大时，难以找到合适的参数变量。
PCA	最大程度保留原始人脸表情中的特征，通过对整幅人脸表情图像进行变换来获取特征进行识别。	具有较好的可重建性。	可分性较差、受光照等外来因素影响较大。
光流法	将运动图像函数 f (x,y,t)作为基本函数，根据图像强度守恒原理建立光流约束方程，通过求解约束方程，计算运动参数。	反映人脸表情变化的实际规律，对光照变化不敏感。	识别模型和算法较复杂，计算量大。

2.1.2 基于深度学习的 FER 方法

近年来，研究者尝试用深度学习的方法进行面部表情识别，令人惊喜的是，深度学习在面部表情识别中也取得了良好的效果，研究者对面部表情识别的研究也逐渐从传统的方法转向深度学习方法。

文献[20]提出一种基于 CNN 集成的面部表情识别方法，该方法在一组 CNN 网络中设计了 3 个不同的结构化子网络，分别包含 3、5、10 个卷积层，图 3 为集成 CNN 的框架。该模型包括两个阶段，第 1 阶段将面部图像作为输入，并将其提供给三个 CNN 子网络，这是该模型的核心部分；第 2 阶段则根据前一阶段的输出预测表情，将这些子网络输出结合起来，以获得最准确的最终决策。

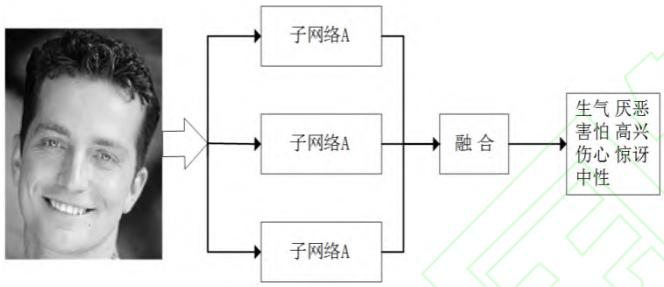


Fig.3 The framework of integrated CNN

图 3 集成 CNN 的框架

由于传统方法中的 LBP 具有旋转不变性和对光照不敏感等优点，文献[21]提出基于 VGG-NET 的特征融合 FER 方法，该方法将 LBP 特征和 CNN 卷积层提取的特征送入改进的 VGG-16 的网络连接层中进行加权融合，最后将融合后的特征送入 Softmax 分类器获取各类特征的概率，完成基本的 6 种表情分类。图 4 为该方法的基本框架。

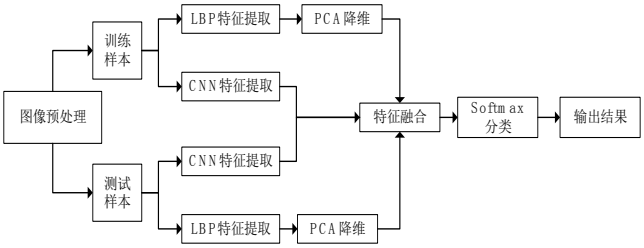


Fig.4 VGG-NET based feature fusion for FER

图 4 基于 VGG-NET 的特征融合 FER 方法

基于深度学习的方法弥补了传统方法在面部表情特征提取方面的缺点，提升了识别效果，同时也存在着一些问题。基于深度学习的方法需要大量的样本来进行模型的训练，以训练出稳定、可靠的面部表情识别模型。但是目前的面部表情数据集中的图像数量较少，在对模型训练时可能会存在过拟合的现象。为了减轻过拟合问题，研究者对扩充 FER 数据库进行了研究。文献[22]提出一种基于 cBGAN 的数据扩充方法，这种方法收敛速度快，并且可以通过添加辅助条件标签信息来控制生成数据的类别。图 5 为 cBGAN 模型，其中 G，D，Enc，Dec，Rlr 和 Rlg 分别代表生成器，鉴别器，编码器，解码器和两个重建损耗。

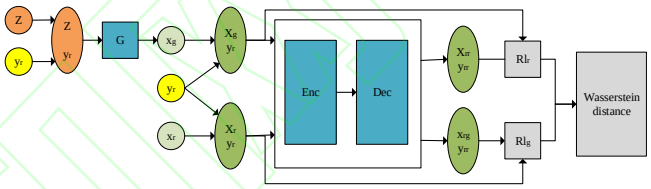


Fig.5 cBGAN model

图 5 cBGAN 模型

数据集中也存在着一些不可避免的问题：一个是在对图像标注时，依赖标注人员的主观判断，可能会出现标记错误的现象。另一个是数据中存在一些模糊的或者有遮挡的图像。用存在问题的数据集进行模型的训练时，可能会使模型在优化的初期就不合逻辑[23]。针对模糊的图像以及错误标签的问题，文献[23]提出一种自修复网络(SelfCure Network，SCN)，该网络为了防止样本的过拟合问题将数据集中的样本进行排序正则化加权。在排名最低的组中通过重标记机制改变这些样本标签来对错误标签进行修改。

由于文化背景以及采集条件的不同，数据集中的数据可能会产生明显的偏差，文献[24]深入研究了这种偏差，首次探索了数据集差异的内在原因，提出了深层情感适应网络(ECAN)，该方法可以同时匹配域间的边缘分布和条件分布，并且通过一个可学习的重加权参数来解决被广泛忽视的表达式类分布偏差。由于数据集中的数据较少，以及数据集中的问题，有些研究者提出用迁移学习的方法来弥补 FER 数据集少的缺点，但是迁移学习也会产生一些冗余信息。文献[25]基于面部肌肉运动产生面部表情变化的原理，提出了

一种新的端到端的深度网络框架以解决此问题。

2.2 基于文本的情感分析

文本情感分析是指从文本中提取可以表达观点、情感的信息。文本情感分析的应用有很多，包括获取用户满意度信息、根据用户情绪推荐产品、预测情绪等。涉及人工智能、机器学习、数据挖掘、自然语言处理等多个研究领域。文献[26]将文本情感分析分为两部分：第一部分是观点挖掘，处理意见的表达，第二部分是情感挖掘，关注情感的表达。观点挖掘更关注的是文本中表达的观点的概念，这些观点可以是积极的，消极的，也可以是中性的，而情感挖掘则是研究反映在文本中的情绪（如快乐、悲伤等）。

在文本情感分析中，情感信息抽取是最重要的部分。情感信息抽取的效果直接影响文本情感分析的效果。情感信息的抽取就是对文本中情感词的抽取，情感词汇可以分为三种类型:1.只包含情感词的词汇(单词列表); 2.由情感词和极性取向构成的词汇(只有正负注释的单词列表)；3.具有方向和强度的情感词[27]。

随着对文本情感分析研究的深入以及大量带有情感色彩的文本信息的出现，研究者从刚开始对情感词进行分析逐渐转变到句子以及篇章级别的研究。目前，基于情感词典和深度学习的方法是文本情感分析的两种主要方法。

2.2.1 基于情感词典的方法

基于情感词典的方法首先对情感词进行抽取，然后根据情感词典中包含的单词及相关词汇的情感极性来进行情感估计[28]。常用词典包括 WordNet、General Inquirer(GI) 词典等。基于情感词典的方法在识别中具有简单且识别速度快的特点，但同时也存在一些不可能避免的缺点。一个缺点就是这种方法比较依赖情感词的个数，另一个就是有一些词语一词多义，在识别时可能会造成误判。为了增加情感词典跨领域的适应性，文献[27]利用分布式语义的概念，提出了一种将语义相似度与嵌入表示相结合的情感分类模型，该方法通过计算输入词与词汇之间的语义相似度来提取文本的特征，有效地解决了情感词典中词汇覆盖率和领域适应方面的局限性。文献[29]提出了一种基于多源数据融合的方面级情感分析方法，该方法可以从不同类型的资源中积累情感知识，并且利用 BERT 来生成用于情感分析的方面特定的句子表示来使模型能够做出更准确的预测。

2.2.2 基于传统机器学习的方法

在文本情感分析领域中，传统的机器学习方法也广泛用于建立情感分析模型，这些方法首先建立一个训练集，并通过情感来标记训练数据，然后从训练数据中提取一组特征，并将其送到分类器模型中进行分析，常用的分类模型有逻辑回归、支持向量机、随机森林、最大熵分类等[30]。2002 年，文献[31]首次将朴素贝叶斯、最大熵分类和 SVM 三种机器学习方法用在文本情感分析中，取得了不错的准确度。文献[32]基于多特征组合的方式用 SVM 和条件随机场(Conditional random field，CRF)分别进行文本情感分析，通过实验表明在选用的特征中情感词对结果的影响最大，程度副词对结果的影响最小，并且还可能降低结果的准确度，同时还表明在相同的特征条件下，CRF 的效果比 SVM 好。为了提高机器学习算法在文本情感分析的准确度，文献[33]利用集成学习的方法结合多种分类器来进行情感分析。该文将常用的七个不同的传统机器学习分类模型用 bagging 和 AdaBoost-r 集成在两个不同的数据集上进行交叉验证。实验结果表明用集成学习方法比单一分类器的准确度高，并且在集成学习模型中，bagging 的表现优于 AdaBoost-r。

2.2.3 基于深度学习的方法

基于深度学习对文本进行情感分析的原理是将提取后的文本特征由计算机根据某种算法进行处理，然后对其分类。由于 CNN 在文本挖掘和 NLP 任务方面表现出了良好的适应性,研究人员用 CNN 进行了一系列实验,证明 CNN 在句子级的情感分析任务上表现出了良好的性能。受此启发，文献[34]提出了一种基于 CNN 的文本分类模型，通过使用二维 TF-IDF 特征代替预先训练的方法，得到了较好的识别准确度，图 6 为该模型的基本结构。由于在文本情感分析中文本词向量作为特征对 CNN 进行训练时无法充分利用其情感特征等问题，文献[35]提出了一种基于多通道卷积神经网络（MCCNN）的中文微博情感分析模型，该

模型可以通过多方面信息学习不同输入特征之间的联系，挖掘出更多的隐藏特征信息。该模型在多个数据集上进行实验，都取得了良好的效果。

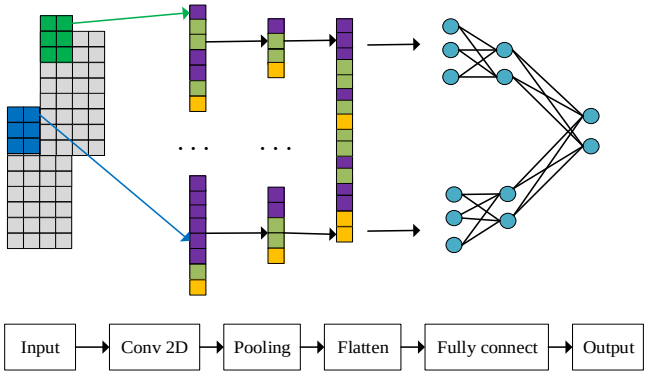


Fig.6 A text classification model basedon convolutional neural network

图 6 卷积神经网络的文本分类模型

由于现有文本情感分析算法中网络输入单一，同时缺乏考虑相似文本实例对整体分类效果的影响，文献[36]提出一种融合 CNN 和注意力的评论文本情感分析模型。在文本情感分析中，人们常常会忽略词语和上下文之间的关系，近而影响情感分析的准确度。文献[37]提出一种基于 BGRU 深度神经的中文情感分析方法，该方法通过 BGRU 对文本信息的上下文提取进行分析，通过实验表明，加入上下文信息后可以有效提高准确度。文献[38]提出了一个 CNN 和 RNN 的联合架构，该方法利用 CNN 生成的粗粒度局部特征作为 RNN 的输入来对短文本进行情感分析。神经网络模型在自然语言处理中非常强大，但该模型有两个主要缺点：训练数据集较小时，该模型可能会过拟合；当类别数较大时，它不能精确地限定类别信息。为了解决这两个缺点，文献[39]提出了一种文本生成新模型 CS-GAN，它是 RNN、生成对抗网络(generative adversarial networks, GAN)和强化学习(reinforcement learning, RL)的集合。该方法不仅可以通过 CS-GAN 扩展任何给定的数据集，还可以直接用 GAN 学习句子结构，提高该模型在不同数据集上的泛化能力。

2.3 基于语音的情感分析

在日常生活中，以语音进行交流是必不可少的方式之一。在语音中含有丰富的情感信息，不仅仅只是文本信息，还包括音调、韵律等可以显示情感的特征。近年来，利用多媒体计算机系统研究语音中的情感信息越来越受到研究者的重视，分析情感特征、判断和模拟说话人的喜怒哀乐成为一个意义重大的研究课题。在现有的文献中，基于语音的情感分析研究大部分集中在识别一些声学特征中，如韵律特征、音质特征和谱特征。目前主要分为基于传统机器学习的方法和基于深度学习的方法。

2.3.1 基于传统机器学习的方法

在语音情感分析中，有一些研究集中在情感语音数据库的构建、语音特征提取、语音情感识别算法等方面。现有成果中，传统的情感识别的主要方法 SVM、K 最近邻法(K-nearest neighbor, KNN)、HMM[40]、高斯混合模型(gaussian mixture model, GMM)等。如文献[41]通过基于机器学习的 PPCA 工具包来提取韵律特征进行情感分析。文献[42]通过使用预先训练的 SVM 和线性判断分析(linear discriminant analysis, LDA)分类器将语音情感特征分类输入来完成语音情感分析。

目前仍无法准确地确定各类情感的本质特征由哪些语音情感特征参数决定，理论上说，提取统计的特征参数越详细，情感类型越容易辨识，但实际上必须在大量情感信息中挑选出能准确反映情绪状况的特征参数，才能获得良好的语音情感识别性。通过对声学特征的对比分析，文献[43]结合韵律特征和质量特征导出 MFCC、LPCC 和 MEDC 三种特征来训练 SVM 进行情感分析，取得了不错的效果，并且该方法具有较好的鲁棒性。

2.3.2 基于深度学习的方法

随着深度学习的日益发展，其被更多的研究者用于识别语音中的情感分析中。文献[44]利用 CNN 从音频中提取情感特征，然后将提取到的特征送入分类器进行情感分类识别。在大规模的网络语音数据中进行情感分析一直以来是一个挑战，为解决这个问题，文献[45]提出了一个深度稀疏神经网络(DSNN)模型，该模型提取话语中三个方面的特征：声学特征（音调、能量等）、内容信息(如描述性相关和时间相关性)和地理信息（如地理-社会相关性），然后融合所有的特征来自动预测情感信息。

2.4 小结

本小节主要介绍了现有的单模态的情感分析方法。如图 7 所示，根据模态不同分别对文献进行叙述。在 FER 中，现有算法多用传统方法与深度学习相结合的方法来进行情感分析，在数据集方面用 GAN、迁移学习等进行扩充。

在文本情感分析中，由于传统方法中情感词典受情感词数量和个数的限制，大多数研究者使用深度学习中的 RNN、LSTM 等模型来进行分析，同时加入注意力机制来提高分析效果；在语音情感分析中，多用深度学习的方法来进行分析，而难以采集到大量包含情感的语音数据是限制对其深入研究的主要因素之一。

由于从单模态中获得的信息量有限，想要进一步提高情感分析的准确度变得十分困难。所以有研究者尝试从多种模态中获取更多的信息进行情感分析来提高准确度。

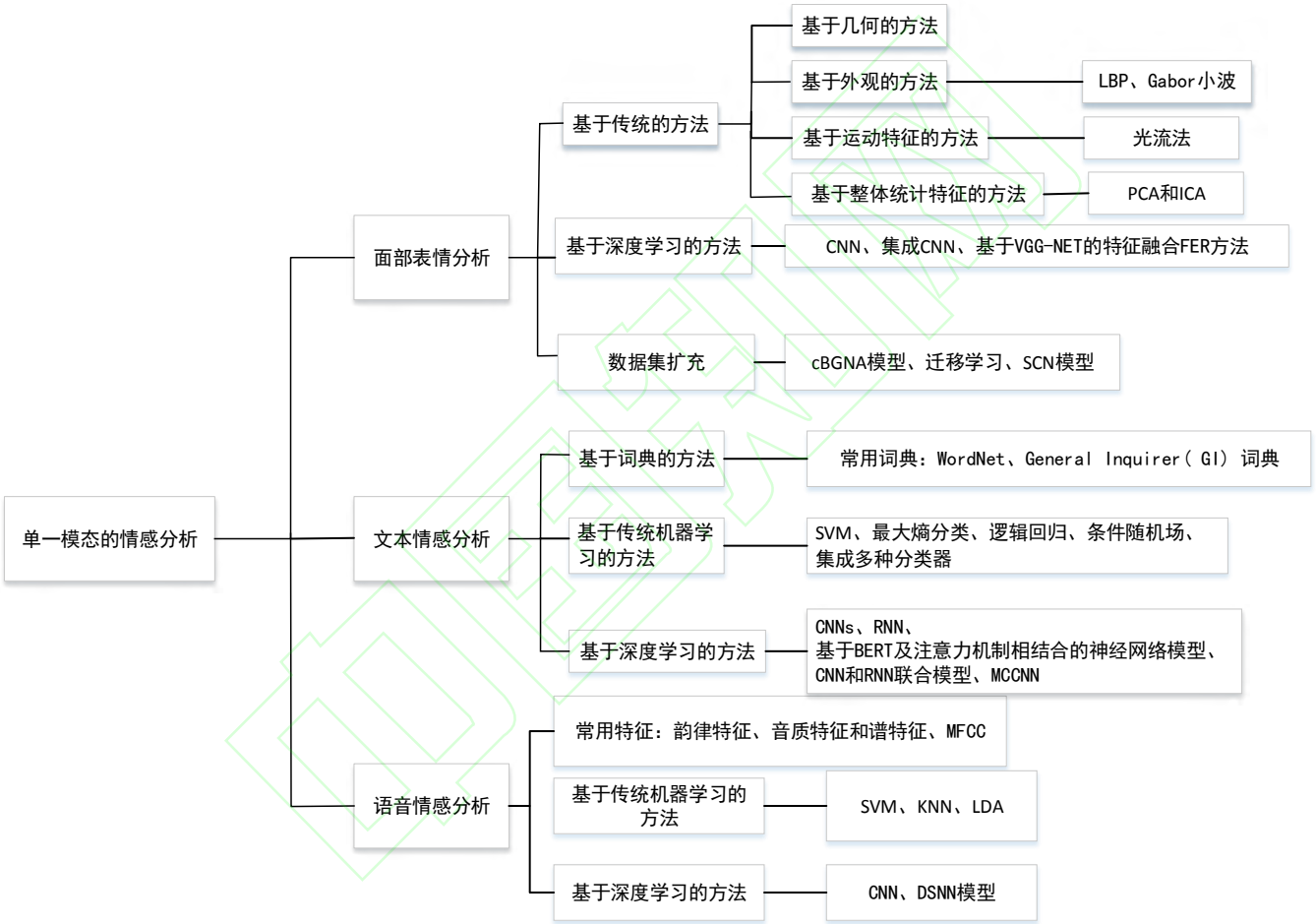


Fig.7　Current research status of monomodal sentiment analysis

图 7　单模态情感分析研究现状结构框图

3 多模态情感分析

用单模态进行情感分析有识别率低、稳定性差等局限性，在情感分析的发展过程中，研究者利用多种模态进行情感分析来提高其准确性以及稳定性。在多模态情感分析中，模态融合的效果会直接影响结果的准确性[46]。因此对单模态的信息处理完成时，还需要根据所用模态的不同以及模态中信息的不同选择适当的模态融合方法。

本小节中，先对近几年的多模态情感分析文献根据模态融合方式的不同进行归纳总结，然后讨论了现有的模态融合算法，最后对文献中出现的算法进行对比分析。

3.1 基于多模态的情感分析

在现有的文献中，基于多模态的情感分析除了单模

态的特征提取外，还需要进行模态融合。融合不同模态的信息是任何多模态任务的核心问题，它将从不同的单模态中提取到的信息集成一个多模态特征[47]。多种模态信息的融合可以为决策提供更加全面的信息，从而提高决策总体结果的准确度[48]。目前模态融合的方式主要分为特征级融合、决策级融合和混合融合三种。

3.1.1 特征级融合

特征级融合也称早期融合，在进行特征提取后立即集成，通常只是简单连接它们的表示，广泛出现在多模态学习任务中[49]。

在基于特征级融合的文献中，文献[50]建立了首个在话语层面进行注释的 MOUD 数据集并且提出了一种基于话语级的情感分析方法。该方法用 OpenEAR、CERT 提取语音和面部的情感特征，将视频中出现频率低的单词删除，剩余单词与每个话语转录内频率的值相关联得到简单的加权图特征作为文本情感特征，然后使用特征级融合的方法将三种特征进行融合送入 SVM 进行分析得到情感极性。

由于视频中的话语之间存在相互依赖和联系，一些文献在对视频中人物的情感分析过程中利用这种依赖和联系，取得了不错的情感分析效果。文献[51]提出了一种基于 LSTM 的情感分析模型，该模型在进行特征特征提取时分为两部分，第一部分用 CNN、3d-CNN 和 openSMILE 对文本信息、面部表情信息以及音频进行特征提取，第二部分用 bc-LSTM 提取语境话语层面的特征。文献[52]提出了一种多模态神经网络结构，此结构用 LSTM 整合了随时间变化的视觉信息，并将其与音频和文本信息通过特征级融合的方式进行情感分析，图 8 为该结构的基本框架。文献[53]提出了一种卷积递归多核学习（CRMKL）模型。在特征提取时，用 openSMILE 提取音频中音高和声音强度；在视频中，为了捕捉时间相关性，将时间 t 和 t+1 的每对连续图像转换成单个图像，作为 RNN 的输入，输出为“正”或“负”；在文本中，先将西班牙语转换为英语，用 word2vec 字典进行预处理形成 300 维的向量作为 CNN 的输入来提取特征。在模型中，将提取的特征用基于循环相关的特征子集(CFS)和 PCA 进行特征选择降低特征维度，然后用 MKL 将特征进行特征级融合，最后进行分析得到情感极性。通过实验表明，加入上下文之间的联系进行分析时，可以有效的提高情感分析的准确度。

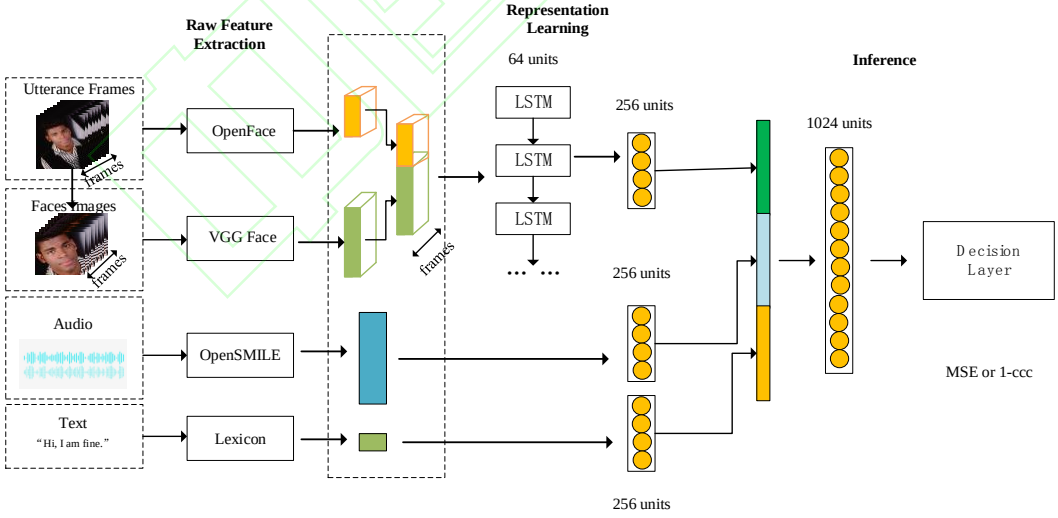


Fig.8 Multimodal neural network framework for emotion recognition

图 8 多模态神经网络情感识别框架

由于在用语音特征区分愤怒和开心时准确率过低，文献[54]结合文本和语音来区分愤怒和开心两种情绪。该方法用 openSMILE 提取声学特征，用基于词典的方法提取文本特征，然后进行特征级融合，将融合后的结果分别送入 SVM 和 CNN 中进行对比分析。通过实验证明文本和语音中包含的情感信息进行互补，提高了愤怒和开心的区分准确率。

由于注意力机制和门控循环单元在一些领域取得

了不错的效果，所以在多模态情感分析的研究中，研究者尝试将注意力机制和门控循环单元引入其中进行分析，如文献[55]结合音频和文本进行情感分析，提出了一种多特征融合和多模态融合的新策略(DFF-TMF)。在特征提取时，用 Librosa 工具包在音频中提取声学特征，用 BERT 模型在文本中提取文本特征，然后将其分别输入到改进的 Bi-LSTM 和 CNN 串行神经网络中，结合注意力机制对情感特征进行改善，分别得到其情感向量，随后用多模态注意力机制和 Bi-LSTM 编码器来选择性学习这些输入进行特征级融合，最后用 softmax 进行情感分析。此方法在进行模态融合时用多模态注意力机制重点融合来自音频和文本互补的情感信息，减少了特征融合的数量。文献[56]用视频信息和文本信息提出了一种改进的多模态情感分析方法。该方法使用自注意力机制获得视频上下文的相关性，使用交叉注意力机制学习不同模态之间的相互作用，使用交叉相互的门控机制来克服单个模态中存在的噪声，选择性学习融合特征向量，随后使用 Bi-GRU 来学习每个模态的深度特征向量，最后将每个模态的深度多模态特征向量连接用 softmax 进行情感分析。

文献[57]利用图像的深度语义信息提出了一种深度语义以及多主体网络，从图像中提取包括对象和场景在内的深度语义特征作为情感分析的附加信息。在视觉信息中，分别选用 VGG 模型和 Scene-VGG 模型在 ImageNet 以及 dataset-Place365 数据集上进行预训练，然后采用迁移学习来克服数据集之间的类别差异，将学习到的参数转移到情感分析任务中，来获得视觉特征以及场景特征。在文本信息中，引入注意力机制和 LSTM 模型提取文本特征。

由于大多数现有的任务方法在进行情感分析时主要依赖文本内容，而没有考虑其他重要的模态信息，基于此问题，文献[58]提出了一种用于实体级多模态情感分类的实体敏感注意和融合网络。在文本特征中，将文本分为左上下文、有上下文和目标实体三部分，用三个 LSTM 获得其上下文信息以及情感特征；在视觉特征中，用残差网络(ResNet)来提取视觉特征并用注意力机制来获得其每部分的权重信息，然后加入 GRU 来滤除图像噪声，最后通过特征级别融合的方式将两种模态的特征融合后送入 softmax 中进行情感分析。虽然此方式在几个评论数据集上都取得了较好的效果，但是其网络结构较为复杂，运行时间较长。

3.1.2 决策级融合

决策级融合也称后期融合。在这个融合过程中，每个模态的特征被独立地分析，将分析结果融合为决策向量以获得最终的决策结果。决策级融合的优点是当任何一个模态缺失时，可以通过使用其他模态来做出决策，这时需要一个智能系统来检测缺失的模态。由于在分析任务中使用了不同的分类器，因此在决策级融合阶段，所有这些分类器的学习过程都变得繁琐而耗时[3]。

在基于决策级融合的方式中，部分文献仅用单模态提取的特征进行情感分析。文献[59]提出了一种基于深度 CNN 的微博视觉和文本的情感分析方法，在该方法中，用 CNN 和 DNN 分别对文本信息和视觉信息进行情感分析，最后用平均策略和权重对两种模态的分析结果进行融合。由于中文微博数据集较小，在构建 DNN 模型时加入 DropConnect 防止过拟合。文献[60]使用文本、视频和音频三种模态提出了一种擅长于异构数据的基于深层 CNN 的特征提取方法。该方法在文本特征提取时，用 CNN 对其情感特征进行提取；在面部特征提取时，将视频逐帧剪辑获取静态图像，然后从静态图像中提取面部特征点；在音频特征中，用 openSMILE 软件来提取与音调、声音强度相关的音频特征，最后将所提取的特征送入单独地分类器中进行分析，将结果在决策级进行可并行化的融合。该文和文献[53]都用基于循环相关和主成分分析来减少特征分析时的数量。特征选择虽然加快了情感分析的速度，但同时可能丢失较为重要的细节情感特征信息，对结果产生负面影响。

由于多模态情感分析数据集较少，且注释的数据集中的示例较少，在情感分析模型训练时，得到的结果可能会与人物的身份特征相关联。为了解决此类问题，文献[61]提出了一个选择加性(SAL)学习程序来改善神经网络在多模态情感分析中的泛化能力。SAL 程序一共分为选择阶段和添加阶段两部分。在选择阶段，SAL 从神经网络学习的潜在表征中识别混杂因素。在加法阶段，SAL 通过在这些表示中添加高斯噪声，迫

使原始模型丢弃混杂元素。将文献[44]中的情感分析方法用 SAL 增加其泛化能力和预测情绪后得到 SAL-CNN,通过实验证明，SAL-CNN 在有限数据集上得到了不错的效果，并且该方法在不同的数据集上进行测试时，也获得了良好的预测精度。

文献[62]介绍了一种新的损失函数的回归模型称为 SDL 并且提出了一个时间选择性注意的模型(TASM)，该模型由注意力模块、编码模块和说话人分布损失函数三部分组成。注意力机制通过明确分配注意权重来帮助模型选择显著的时间步长，在注意力模块用 LSTM 对序列进行预处理，编码阶段用 Bi-LSTM 对序列观测值进行编码并加权组合作为该模块的输出，最后送到 SDL 中进行情感分析。在模态的特征提取中，用 openFace 提取面部外观特征，用 COVAREP 提取声学特征,文本用 Glove 得到词向量。通过实验表明，加入注意力机制之后的模型能够关注以人为中心的视频序列的显著部分，并且取得了不错的效果。

3.1.3 混合融合

混合融合是特征级融合和决策级融合方法的结合。这种融合方法结合了特征级融合和决策级融合的优点，同时模型复杂度和实现难度也随之增加。

由于注意力机制和 GRU 在情感分析中表现出较好的性能，文献[63]提出了一种带有时间注意门控的多模态嵌入 LSTM 模型,该模型在单词级上进行融合，并且可以关注到最重要的时间帧，解决了“在每一时刻要寻找什么样的情况”和“在交流中什么时候说话最重要”这两个关键问题。在本文中，首次提出了一个注意层和一个强化学习训练的输入门控制器来解决模态中的噪声问题。文献[64]提出了一种端到端的 RNN 模型用来对情感进行分析。此模型可以捕捉所有模态对话上下文、听者和说话者情绪状态之间的依赖性以及可用模态之间的相关性。在结构上，使用两种门控循环单元(gaterecurrent unit, GRU): sGRU 和 cGRU 来为对话者的状态和情感建模。除此之外，使用一个互连的上下文网络来学习上下文表示，并且使用成对的注意力机制来对每种模态的有用信息进行简单的表示。此文通过实验表明成对的注意力在多模态数据上具有最先进的性能。

文献[65]和文献[66]引入了几种不常用的模态进行情感分析。文献[65]基于面部表情、皮肤电反应和脑电图提出了一种基于混合融合的多模态情感分析系统，图 9 为其结构框图。该系统用 8898 张图片训练得到 CNNF 模型，输出为七个离散情感类别的概率向量。CNNV 和 CNNA 由脑电图和 GSR 模态的网络进行训练得到。加权单元分别计算 CNNV 和 CNNA 输出的化合价和唤醒的加权和，然后将其送到距离计算器计算情感距离，最后将情感距离送到决策树，与 CNNF 得到的结果进行融合得到情感状态。文献[66]结合视频中的面部表情和姿态提出一种基于视觉的多模态情感分析框架，从视频序列中自动识别面部表情和上身手势特征进行特征级融合，随后将分析结果用乘积和加权的方法进行决策级融合得出结果。由于多模态情感分析的数据集较少，该篇文章所用的数据集是自建的一个面部表情和姿态的视频数据库。

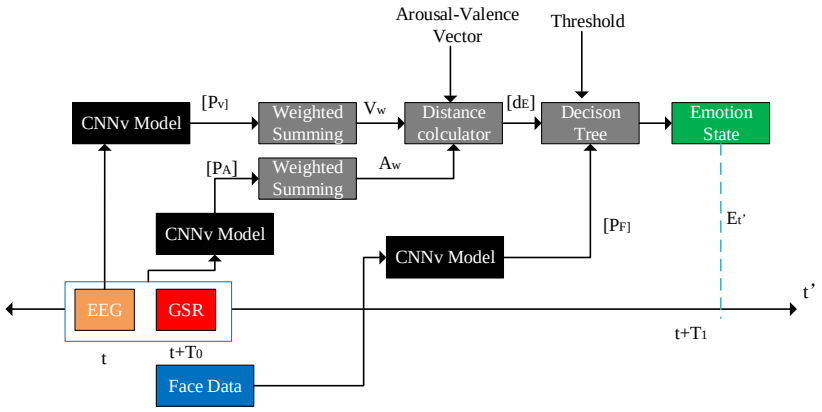


Fig.9 Cross-subject multi-modal emotion recognition based on hybrid fusion

图 9 基于混合融合的跨主体多模态情感识别

3.2 模态融合相关算法

由于多模态技术的发展，模态融合技术也受到了研究者的广泛关注。在模态融合技术发展的初期，大多数研究者都是使用基于机器学习的方法训练模态的分类器，如 SVM、Logistic 回归、K-近邻分类器等，还有一些研究者使用基于规则的方法，如线性加权和多数表决等[67]。随着深度学习的发展，研究者将深度学习的方法引入到模态融合中。近年来提出了一些经典的模态融合方法如下：

一些研究者将注意力机制和 GRU 引入模态融合中，通过注意力机制来获得模态的特征向量的权重。文献[68]提出了一种融合抽取(FEN)模型。该模型用一种细粒度的注意力机制来交互学习视觉和文本信息的跨模态融合表示向量，可以融合两种单一模态中对情感最有用的信息。在该方法中引入门控机制调节多模态特征融合时的权重以进行情感分析。文献[69]提出了一种基于双向门控递归单元模型获取话语间的上下文关系的信息增强融合算法框架，通过 CT-BiLSTM 获取文本、音频和视频的上下文相关的单模态特征，然后用 AT-BiGRU 模型来放大与目标话语高度相关的上下文信息。此算法可以优先选择对情感分析有较大影响的模态，能够增强对目标话语正确分类结果影响较大的情感信息。

也有一些研究者将张量引入模态融合中计算模态间的交互作用，如文献[70]提出了一种张量融合网络(TEN)新模型，可以端到端地学习模态内和模态间的动态特性，它明确地聚合了单模态、双模态和三种模态的相互作用。通过三种模态嵌入子网络分别对语言、视觉和声学模态进行模态内动力学建模。文献[71]将张量的方法和一些神经网络结合起来提出了一种深度高阶序列融合网络(Deep-HOSeq)，通过从多模态时间序列中提取两种对比信息进行多模态融合。第一种是模态间信息和模态内信息的融合，第二种是多模态交互的时间粒度信息。该网络用 LSTM 从每个单模态中获得模态内信息，然后将每个模态内信息合并为多模态张量，取其外积。另一方面利用前馈层从每个单模态中获得潜在特征，然后在每个时间步骤中获得模态内的作用，用卷积层和全连接层进行特征提取，最后通过池化操作统一来自所有时态步骤的信息。在获得这两种信息后与一个融合层相结合来进行情感分析。

在处理模态融合时，保持单模态神经网络的性能是至关重要的，基于此观点文献[72]提出了一种多层的多模态融合方法，该方法引入了一个特定的神经网络称为中央网络，该中央网络不仅可以将不同的特征联合起来，而且通过使用多任务学习来规范各个模态的网络。此融合方法可以通过将相应的单模态网络层和其前一层的加权和作为每个层的输入。中央网络的优点为中央网络的损失函数不仅允许学习如何组合不同的模态，而且还增加了对特定模态的网络的限制，从而增强了模态间的互补性。文献[73]用分层的方式对模态信息进行融合，在该方法中引入 RNN 和 GRU 分别用来获取周围话语信息以提高特征向量的质量和对上下文信息进行建模。

3.3 不同算法对比

为了得到影响情感分析准确率的因素，此小节将前文中对视频信息进行情感分析所提到的算法进行对比研究，对比结果如表 3～6 所示。以下表中的评价指标都为 Accuracy(%),并且表中的模态信息 A、V、T 分别代表 Audio，Video，Text。

Table 3　Comparison of single-modal sentiment analysis

表 3　单模态的情感分析 Acc 比较

算法	T+V	T+A	V+A
LSTM-based model[51]	80.22	79.33	62.17
Utterance-level[50]	72.39	72.88	68.86
C-MKL[60]	85.46	84.12	83.69
CRMKL[53]	96.21	84.12	95.68
Multilogue-Net[64]	80.06	80.18	75.16
SAL-CNN[61]	73.00	72.50	62.10

Table 4 Comparison of different algorithms on the MOSI data set

表 4 MOSI 数据集上不同算法 Acc 比较

算法	T	V	A	A+V+T
LSTM-based model[51]	78.12	55.80	60.31	80.30
GME-LEST[63]	\	\	\	76.50
DFF-TMF[55]	\	\	\	80.98
Gated mechanism for attention[56]	\	\	\	83.91
TSAM[62]	74.50	61.80	60.90	75.10
Multilogue-Net[64]	\	\	\	81.19

Table 5 Comparison of different algorithms on the MOUD dataset

表 5 MOUD 数据集上不同算法 Acc 比较

算法	T	V	A	A+V+T
LSTM-based model[51]	52.17	48.58	59.99	68.11
Utterance-level[50]	70.94	67.31	64.85	74.09
CRMKL[53]	79.77	94.50	74.22	96.55
SAL-CNN[61]	73.20	63.60	61.80	73.00

Table 6 Accuracy comparison of the same algorithm on different datasets

表 6 相同算法在不同数据集上 Acc 比较

算法	数据集	A+V+T
LSTM-based model[51]	MOSI	78.12
	MOUD	52.17
MultiSentiNet-Att[57]	MVSA-Single	69.84
	MVSA-Multi	68.86
Multilogue-Net[64]	MOSI	81.19
	MOSEI	82.10
DFF-TMF[55]	MOSI	80.98
	MOSEI	77.15
	IEMOCAP	81.37
Gated mechanism for attention[56]	MOSI	83.91
	MOSEI	81.14

通过表 3 可以看出，大多数算法用 T+V 和 T+A 进行情感分析时的准确率都要高于 V+A，说明在基于多模态的情感分析中，文本信息仍然是重要的情感线索。

通过表 4 可得，在 MOSI 数据集上用三种模态进行情感分析时，Gated mechanism for attention 算法的准确率最高，说明门控单元在模态选择时的重要性，此外，对每种模态的特征进行除噪也可以提高准确率。其次，Multilogue-Net 和 DFF-TMF 这两种算法的准确率也较高，可以看出注意力机制以及模态间的相关性在提高准确率方面也有重要的价值。

通过表 5 可得，在 MOUD 数据集上用三种模态进行情感分析时，CRMKL 算法的准确率最高，说明视频中上下文的信息、文本信息的预处理以及模态融合的选择对提高准确率很有帮助。

通过表 6 可得，大多数算法在不同数据集上的鲁棒性较好。LSTM-based model 方法在不同数据集上的准确度相差较大，产生这种效果的原因是模型用 MOSI 训练，MOSI 是英语，而 MOUD 是西班牙语，语言不同，因而情感表达方式不同，分析方式也不同。

通过表 3 ~ 5 中相同的算法进行对比可以看出，用

三种模态进行情感分析的准确率高于用两种和一种模态方法的准确率，说明结合多种模态信息进行情感分析的必要性。

3.4 小结

本小节主要对现有的多模态情感分析技术以及模态融合技术进行了总结。在多模态情感分析技术中，部分文献仅在单模态的特征提取上进行改进提高准确度，而忽略了视频序列中的上下文信息，导致对不同模态的特征挖掘不充分。随着研究的深入，研究者引入RNN、LSTM、GRU 等网络提取上下文信息进而提高情感分析准确度，但处理长时间的序列容易出现信息丢失问题。现阶段可以考虑用多层级GRU 编码上下文信息解决长时间依赖问题，从而获得更为全面的情感信息。

在模态融合技术方面，研究者利用多层融合的方法来进行模态融合，这种方法可以提高单模态特征向量的质量，在数据较大的情况下可以获得较好的效果，但在小样本中可能导致过拟合问题。由于注意力机制在模态融合中寻找最优权值时具有重要的作用，张量可以将所有模态的特征投影到同一空间获得一个联合表征空间，易于计算模态间的交互作用，所以近年来主流的模态融合方法是基于注意力机制的方法和基于张量的方法。

4 总结和展望

随着深度学习和一些融合算法的兴起，多模态情感分析技术得到了快速的发展，本文通过对多模态情感分析研究现状的认识，总结出其面临的挑战与发展趋势如下：

（1）多模态情感分析数据集。在多模态情感分析中，数据采集时的花费以及如何在人们自然表达问题的情况下进行数据的采集是目前存在的主要问题之一。数据集较少且多是由视觉、文本和语音三种模态组成，缺少姿态、脑电波等模态数据。所以需要高质量且规模较大的数据集来提高情感分析的准确度。

（2）单模态情感分析。在 FER 中，一方面是不同的数据之间存在一定的差异性，由于不同的采集条件和注释的主观性，数据偏差和注释不一致在不同的数据集中非常常见。另一方面是对一些表情识别不准确，在对高兴、伤心等表情识别时很容易，但在捕获令人反感、愤怒和其他较不常见的表情信息时非常具有挑战性；在基于文本信息的情感分析中，由于不同领域的情感表达差别较大，导致情感词典的构建较难。在含有许多隐喻、反话等复杂的语言形式中进行情感分析得到的效果并不理想，所以如何提取对情感分析具有更大价值的特征依然是一个有待完善的课题；除此之外，情感分析技术在语音、姿态等一些模态中的不成熟制约了多模态情感分析技术的发展。

（3）模态间相关性。从不同模态中提取的特征之间存在一定的相关性，在现有的模态融合算法中，常常会忽略不同特征间的相关性，所以如何有效利用模态间的相关性来提高情感分析的准确度是未来的研究方向之一。

（4）算法复杂度。在进行多模态情感分析时，模态过多会提高融合算法的复杂度，模态过少会影响结果的准确性，所以如何选择最佳的模态进行融合也是一个急需解决的问题。

（5）模态融合时模态的权值问题。在模态融合时，不同环境中不同模态的最优权值分配是影响情感分析结果的重要因素之一。在完成不同分析任务时，不同的模态对分析结果的影响不同，所以如何将对分析结果影响最大的模态赋予较大的权值是接下来模态融合的重点方向之一。

5 结束语

本文对多模态的情感分析领域的现有研究成果进行了总结，介绍了常用的多模态情感分析数据集；然后将近几年中单模态的情感分析技术的文献根据面部表情信息、文本信息以及语音信息进行分类叙述；随后对多模态的情感分析技术的文献进行总结，并且对现有的模态融合技术进行了详细的描述；最后对情感分析中存在的问题进行了讨论。

参考文献：

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

[2] Ekman P, Friesen W V. Constants across cultures in the face and emotion[J]. Journal of personality and social psychology, 1971, 17(2): 124.

[3] Poria S, Cambria E, Bajpai R, et al. A review of affective computing: From unimodal analysis to multimodal fusion[J]. Information Fusion, 2017, 37: 98-125.

[4] Peng X J. Multi-modal Affective Computing: A Comprehensive Survey[J]. Journal of Hengyang Normal University, 2018, 039(3):31-36.

[5] Soleymani M, Garcia D, Jou B, et al. A survey of multimodal sentiment analysis[J]. Image and Vision Computing, 2017, 65: 3-14.

[6] Huddar M G, Sannakki S S, Rajpurohit V S. A survey of computational approaches and challenges in multi-

modal sentiment analysis[J]. Int J Comput Sci Eng, 2019, 7(1): 876-883.

[7] Gao J, Li P, Chen Z, et al. A survey on deep learning for multimodal data fusion[J]. Neural Computation, 2020, 32(5): 829-864.

[8] Zheng W L, Lu B L. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks[J]. IEEE Transactions on Autonomous Mental Development, 2015, 7(3): 162-175.

[9] Li S, Deng W. Deep facial expression recognition: A survey[J]. IEEE Transactions on Affective Computing, 2020.

[10] Zhang Y, Lai G, Zhang M, et al. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis[C]//Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, Queensland, Australia, 6 Jul, 2014.New York: ACM, 2014: 83-92.

[11] Xu N, Mao W, Chen G. Multi-interactive memory network for aspect based multimodal sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence, Hawaii, USA, 17 Jul, 2019. Menlo Park: AAAI, 2019, 33: 371-378.

[12] Koelstra S, Muhl C, Soleymani M, et al. Deap: A database foremotion analysis; using physiological signals[J]. IEEE transactions on affective computing, 2011, 3(1): 18-31.

[13] 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, Online, July ,2020. Stroudsburg: ACL, 2020: 3718-3727.

[14] Morency L P, Mihalcea R, Doshi P. Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web[C]//Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011, Alicante, Spain, November 14-18, 2011. New York: ACM, 2011: 169-176.

[15] Wöllmer M, Weninger F, Knaup T, et al. Youtube movie reviews: Sentiment analysis in an audio-visual context[J]. IEEE Intelligent Systems, 2013, 28(3): 46-53.

[16] Zadeh A, Zellers R, Pincus E, et al. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos[J]. arXiv preprint arXiv:1606. 62 59,2016.

[17] Ellis J G, Jou B, Chang S F. Why we watch the news: a dataset for exploring sentiment in broadcast video news[C]//Proceedings of the 16th international conference on multimodal interaction, Istanbul Turkey, Nov 12-16,2014. New York: ACM, 2014: 104-111.

[18] 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.

[19] Feng X Y, Huang D, Cui S X, et al. Spatial-temporal attention network for facial expression recognition[J]. Journal of Northwest University (Natural Science Edition), 2020,50(3):319-327.

[20] Lu J H, Zhang S M, Zhao J L. Facial Expression Recognition Based on CNN Ensemble[J]. Journal of Qingdao University (Engineering & Technology Edition), 2020,35(2):24-29+42.

[21] Li X L, Niu H T. Facial expression recognition using feature fusion based on VGG-NET[J]. Computer Engineering & Science,2020,42(3):500-509.

[22] Luo Y, Zhu L Z, Lu B L. A GAN-Based Data Augmentation Method for Multi-modal Emotion Recognition[C]//International Symposium on Neural Networks, Cairo, Egypt, Oct 1, 2020. Berlin: Springer, Cham, 2019: 141-150.

[23] Wang K, Peng X, Yang J, et al. Suppressing uncertainties for large-scale facial expression recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, June 14-19, 2020. Piscataway: IEEE, 2020: 6897-6906.

[24] Li S, Deng W. A deeper look at facial expression dataset bias[J]. IEEE Transactions on Affective Computing, 2020.

[25] Dai R. Facial Recognition Method Based on Facial Physiological Features and Deep Learning[J]. Journal of Chongqing University of Technology(Natural Science), 2020,34(6):146-153.

[26] Yadollahi A, Shahraki A G, Zaiane O R. Current state of text sentiment analysis from opinion to emotion mining[J]. ACM Computing Surveys (CSUR), 2017, 50(2): 1-33.

[27] Araque O, Zhu G, Iglesias C A. A semantic similarity-based perspective of affect lexicons for sentiment analysis[J]. Knowledge-Based Systems, 2019, 165: 346-359.

[28] Zhao Y Y, Qin B, Liu T. Sentiment Analysis[J]. Journal of Software,2010,21(8):18 34-1848.

[29] Chen F, Yuan Z, Huang Y. Multi-source data fusion for aspect-level sentiment classification[J]. Knowledge-Based Systems, 2020, 187: 104831.

[30] Li Z, Fan Y, Jiang B, et al. A survey on sentiment analysis and opinion mining for social multimedia[J]. Multimedia Tools and Applications, 2019, 78(6): 6939-6967.

[31] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment Classification using Machine Learning Techniques[J]. Proceedings of Emnlp, 2002.

[32] LI T T , JI D H. Sentiment analysis of micro-blog based on SVM and CRF using various combinations of features[J]. Application Research of Computers, 2015, 32(4): 978-981.

[33] Prusa J, Khoshgoftaar T M, Dittman D J. Using ensemble learners to improve classifier performance on tweet sentiment data[C]//2015 IEEE International Conference on Information Reuse and Integration, San Francisco, CA, USA, 26 Oct 2015. Piscataway: IEEE, 2015: 252-257.

[34] Chen J, Yan S, Wong K C. Verbal aggression detection on Twitter comments: Convolutional neural network for short-text sentiment analysis[J]. Neural Computing and Applications, 2018: 1-10.

[35] CHEN K, LIANG B, KE W D, et al. Chinese Micro-Blog Sentiment Analysis Based on Multi-Channels Convolutional Neural Networks[J]. Journal of Computer Research and Development, 2018, 55(5): 945-957.

[36] Zhu Y, Chen S P. Commentary Text Sentiment Analysis Combining Convolution Neural Network and Attention[J]. Journal of Chinese Computer Systems,2020,41 (3):551-557.

[37] CAO Y, LI T R, JIA Z, et al. BGRU: New Method of Chinese Text Sentiment Analysis[J]. Journal of Frontiers of Computer Science and Technology, 2019, 13(6): 973-981.

[38] Wang X, Jiang W, Luo Z. Combination of convolutional and recurrent neural network for sentiment analysis of short texts[C]//Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical pa-pers, Osaka, Japan, Dec 11-16, 2016. Stroudsburg: ACL, 2016: 2428-2437.

[39] Li Y, Pan Q, Wang S, et al. A generative model for category text generation[J]. Information Sciences, 2018, 450: 301-315.

[40] Pao T L, Chen Y T, Yeh J H, et al. Detecting emotions in Mandarin speech[C]//International Journal of Computational Linguistics & Chinese Language Processing, ROCLING/IJCLCLP, March 2005. Stroudsburg: ACL, 2005: 347-362.

[41] Li Y, Ishi C T, Ward N, et al. Emotion recognition by combining prosody and sentiment analysis for expressing reactive emotion by humanoid robot[C]//2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Malaysia, Dec 12-15, 2017. Piscataway: IEEE, 2017, 1356-1359.

[42] Semwal N, Kumar A, Narayanan S. Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models[C]//2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA), New Delhi, India, 15 Jun 2017. Piscataway: IEEE, 2017.7947681.

[43] Samantaray A K, Mahapatra K, Kabi B, et al. A novel approach of speech emotion recognition with prosody, quality and derived features using SVM classifier for a class of North-Eastern Languages[C]//2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), Kolkata, India, 3 Sep 2015. Piscataway: IEEE, 2015: 372-377.

[44] Huang Z, Dong M, Mao Q, et al. Speech emotion recognition using CNN[C]//Proceedings of the 22nd ACM international conference on Multimedia, Orlando, Florida, USA, Nov 3-7,2014. New York: ACM, 2014: 801-804.

[45] Ren Z, Jia J, Guo Q, et al. Acoustics, content and geo-information based sentiment prediction from large-scale networked voice data[J]. Research & Exploration in Laboratory, 2014:1-4.

[46] Wu L, Oviatt S L, Cohen P R. Multimodal integration-a statistical view[J]. IEEE Transactions on Multimedia, 1999, 1(4): 334-341.

[47] Zhang C, Yang Z, He X, et al. Multimodal intelligence: Representation learning, information fusion, and applications[J]. IEEE Journal of Selected Topics in Signal Processing, 2020.

[48] Krizhevsky A, Sutskever I, Hinton G E. Imagenet clas-sification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[49] SUN Y Y, JIA Z T, ZHU H Y. Survey of Multimodal Deep Learning[J]. Computer Engineering and Applications, 2020, 56(21): 1-10.

[50] Pérez-Rosas V, Mihalcea R, Morency L P. Utterance-level multimodal sentiment analysis[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Sofia, Bulgaria, Aug 4-9, 2013. Stroudsburg: ACL, 2013: 973-982.

[51] Poria S, Cambria E, Hazarika D, et al. Context-dependent sentiment analysis in user-generated videos[C]//Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers), Vancouver, Canada, Jul 30-Aug 4, 2017. Stroudsburg: ACL, 2017: 873-883.

[52] Deng D, Zhou Y, Pi J, et al. Multimodal utterance-level affect analysis using visual, audio and text features[J]. arXiv preprint arXiv:1805.00625, 2018.

[53] Poria S, Chaturvedi I, Cambria E, et al. Convolutional MKL based multimodal emotion recognition and sentiment analysis[C]//2016 IEEE 16th international conference on data mining (ICDM), Barcelona, Dec 12-15, 2016. New York: IEEE, 2016: 439-448.

[54] Hu T T, Shen L J, Feng Y Q, et al. Research on Anger and Happy Misclassification in Speech and Text Emotion Recognition[J]. Computer Technology and Development, 2018, 28(11):124-127+134.

[55] Chen F, Luo Z, Xu Y, et al. Complementary Fusion of Multi-Features and Multi-Modalities in Sentiment Analysis[J]. arXiv preprint arXiv:1904.08138, 2019.

[56] Kumar A, Vepa J. Gated Mechanism for Attention Based Multi Modal Sentiment Analysis[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, Oct 21, 2019. Piscataway: IEEE, 2020: 4477-4481.

[57] Xu N, Mao W. Multisentinet: A deep semantic network for multimodal sentiment analysis[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, Nov 6-10,2017. New York: ACM, 2017: 2399-2402.

[58] Yu J, Jiang J, Xia R. Entity-sensitive attention and fusion network for entity-level multimodal sentiment

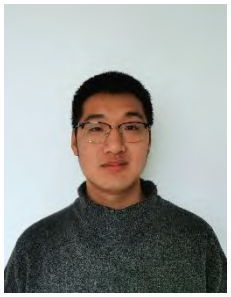
classification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 28: 429-439.
[59] Yu Y, Lin H, Meng J, et al. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks[J]. Algorithms, 2016, 9(2): 41.
[60] Poria S, Cambria E, Gelbukh A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis[C]//Proceedings of the 2015 conference on empirical methods in natural language processing, Lisbon, Portugal, Sep 17-21,2015. Stroudsburg: ACL, 2015: 2539-2544.
[61] Wang H, Meghawat A, Morency L P, et al. Select-additive learning: Improving generalization in multimodal sentiment analysis[C]//2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, Jul 10-14,2017. Piscataway: IEEE, 2017: 949-954.
[62] Yu H, Gui L, Madaio M, et al. Temporally selective attention model for social and affective state recognition in multimedia content[C]//Proceedings of the 25th ACM international conference on Multimedia, California, US, Oct, 2017. New York: ACM, 2017: 1743-1751.
[63] Chen M, Wang S, Liang P P, et al. Multi-modal sentiment analysis with word-level fusion and reinforcement learning[C]//Proceedings of the 19th ACM International Conference on Multimodal Interaction, Scottsdale Arizona, USA, November,2011. New York: ACM, 2017: 163-171.
[64] Shenoy A, Sardana A. Multilogue-Net: A Context Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation[J]. arXiv preprint arXiv:2002.08267, 2020.
[65] Cimtay Y, Ekmekcioglu E, Caglar-Ozhan S. Cross-subject multimodal emotion recognition based on hybrid fusion[J]. IEEE Access, 2020, 8: 168865-168878.
[66] Gunes H, Piccardi M. Bi-modal emotion recognition from expressive face and body gestures[J]. Journal of Network and Computer Applications, 2007, 30(4): 1334-1345.
[67] Fiérrez-Aguilar J, Ortega-Garcia J, Gon-zalez-Rodriguez J. Fusion strategies in multimodal biometric verification[C]//2003 International Conference on Multi-media and Expo, Baltimore, MD, USA, July 6-9,2003. Piscataway: IEEE, 2003, 3: III-5.
[68] Jiang T, Wang J, Liu Z, et al. Fusion-Extraction Net-work for Multimodal Sentiment Analysis[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining, Singapore, May 11-14, 2020. Berlin: Springer, Cham, 2020: 785-797.
[69] Jiang D, Zou D, Deng Z, et al. Contextual multimodal sentiment analysis with information enhancement[J]. Journal of Physics: Conference Series, 2020, 1453(1): 012159 (5pp).
[70] Zadeh A, Chen M, Poria S, et al. Tensor fusion network for multimodal sentiment analysis[J]. arXiv preprint arXiv: 1707.07 250, 2017.
[71] Verma S, Wang J, Ge Z, et al. Deep-HOSeq: Deep Higher Order Sequence Fusion for Multimodal Sentiment Analysis[J]. arXiv preprint arXiv: 2010.08218, 2020.
[72] Vielzeuf V, Lechervy A, Pateux S, et al. Centralnet: a multilayer approach for multimodal fusion[C]// Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, September 8-14, 2018. Berlin: Springer, 2018: 0-0.
[73] Majumder N, Hazarika D, Gelbukh A, et al. Multimodal sentiment analysis using hierarchical fusion with context modeling[J]. Knowledge-based systems, 2018, 161: 124-133.

附中文参考文献：

[4]彭小江. 基于多模态信息的情感计算综述[J]. 衡阳师范学院学报, 2018, 039(3):31-36.
[19]冯晓毅,黄东,崔少星,等.基于空时注意力网络的面部表情识别[J].西北大学学报(自然科学版),2020,50(3):319-327.
[20]陆嘉慧,张树美,赵俊莉.基于 CNN 集成的面部表情识别[J].青岛大学学报(工程技术版),2020,35(2):24-29+42.
[21]李校林,钮海涛.基于 VGG-NET 的特征融合面部表情识别[J].计算机工程与科学,2020,42(3):500-509.
[25] 戴蓉.基于面部生理特征和深度学习的表情识别方法[J]. 庆理工大学学报(自然科学),2020,34(6):146-153.
[28]赵妍妍,秦兵,刘挺.文本情感分析[J].软件学报,2010,21(8):1834-1848.
[32]李婷婷,姬东鸿.基于 SVM 和 CRF 多特征组合的微博情感分析[J].计算机应用研究, 2015, 32(4):978-981.
[35]陈珂,梁斌,柯文德,许波,等.基于多通道卷积神经网络的中文微博情感分析[J].计算机研究与发展,2018,55(5):945-957.
[36]朱烨,陈世平.融合卷积神经网络和注意力的评论文本情感分析[J].小型微型计算机系统,2020,41(3):551-557.
[37]曹宇,李天瑞,贾真,殷成凤.BGRU:中文文本情感分析的新方法[J].计算机科学与探索,2019,13(6):973-981.
[49]孙影影，贾振堂，朱昊宇. 多模态深度学习综述[J]. 计算机工程与应用, 2020, 56(21):1-10.
[54] 胡婷婷, 沈凌洁, 冯亚琴, 等. 语音与文本情感识别中愤怒与开心误判分析[J]. 计算机技术与发展, 2018, 28(11):124-127+134.

LIU Jiming was born in 1964. He received the Ph.D. degree from George Washington University in US. Now he is distinguished professor at Xi'an University of Posts and Telecommunications. His research interests include artificial intelligence technology and its industrialization.

刘继明(1964-)，男，福建龙岩人，在美国乔治华盛顿大学获得工业工程专业博士，现为网经科技（苏州）有限公司董事长，西安邮电大学特聘教授，主要研究领域为人工智能技术及其产业化。



ZHANG Peixiang was born in 1996. He is an M.S. candidate at Xi'an University of Posts and Telecommunications. His research interest is multimodal sentiment analysis.

张培翔(1996-)，男，山西运城人，西安邮电大学通信与信息工程学院硕士研究生，主要研究领域为多模态情感分析。

LIU Ying was born in 1972. She received the Ph.D. degree from Monash University in Australia in 2007. Now sheis a professor and M.S. supervisor at Xi'an University of Posts and Telecommunications. Her research interests include image retrieval, image enhancement, etc.

刘颖(1972-)，女，陕西户县人，2007 年于澳大利亚莫纳什大学获得博士学位，现为西安邮电大学教授、硕士生导师，图像与信息处理研究所负责人及电子信息现场勘验应用技术公安部重点实验室总工程师，主要研究领域为图像检索，图像清晰化等。主持公安部科技强警、国家自然科学基金、陕西省国际科技合作计划、陕西省教育厅专项科学研究计划等项目。

ZHANG Weidong was born in 1990. He received the Ph.D. degree from Shandong University in China. Now he is an associate professor at Xi'an University of Posts and Telecommunications. His research interest is indoor scene understanding.

张伟东(1990-)，男，陕西宝鸡人，在山东大学获得博士学位，现为西安邮电大学副教授，主要研究方向为室内场景理解。2016 年获得 CVPR 国际大规模场景理解比赛冠军。

FANG Jie was born in 1993. He received the Ph.D. degree from University of Chinese Academy of Sciences in China. Now he is an associate professor at Xi'an University of Posts and Telecommunications. His research interest is semantic understanding of visual image and its application.

房杰(1993-)，男，陕西咸阳人，2020 年于中国科学院大学获得博士学位，现为西安邮电大学副教授，主要研究方向为视觉影像的语义理解及其应用。