Statistics and Application 统计学与应用, 2023, 12(1), 128-138
Published Online February 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.121015

Hans汉斯

衡量视觉信息的多模态情感分析综述

孙　睿[1]，周艳聪[2*]

[1]天津商业大学理学院，天津
[2]天津商业大学信息工程学院，天津

收稿日期：2023年1月21日；录用日期：2023年2月11日；发布日期：2023年2月23日

摘　要

人们逐渐以多元化的形式来表达自己的情感，于是多模态情感分析应运而生。所谓视觉信息，分为在图片中提取和在视频中提取，如何有效衡量这部分情感倾向，如何将其与文本模态内容结合，合理完成多模态情感分析任务是该领域非常值得探究的问题。本文从相应的多模态数据集展开综述，给出了目前多个经典数据集的详细概况，后梳理了各模态特征抽取方法、模态融合技术、基于深度学习的前沿技术等方面的内容。最后给出了对未来研究的展望，为选择合适匹配的数据集和探索多模态情感分析体系提供思路启发。

关键词

视觉信息，多模态情感分析，特征抽取，模态融合

A Review of Multimodal Sentiment Analysis for Measuring Visual Information

Rui Sun[1], Yancong Zhou[2*]

[1]Department of Science, Tianjin University of Commerce, Tianjin
[2]Department of Information Engineering, Tianjin University of Commerce, Tianjin

Received: Jan. 21st, 2023; accepted: Feb. 11th, 2023; published: Feb. 23rd, 2023

Abstract

People gradually express their emotions in diverse forms, so multimodal sentiment analysis is born. The so-called visual information, divided into extracted in pictures and extracted in videos, how to effectively measure this part of sentiment tendency and how to combine it with text modal

*通讯作者。

文章引用: 孙睿, 周艳聪. 衡量视觉信息的多模态情感分析综述[J]. 统计学与应用, 2023, 12(1): 128-138.
DOI: 10.12677/sa.2023.121015

content to reasonably accomplish the multimodal sentiment analysis task is a very worthy question in this field. In this paper, we review the corresponding multimodal datasets, give a detailed overview of several classical datasets at present, and then sort out the content of each modal feature extraction method, modal fusion techniques, and frontier techniques based on deep learning. Finally, an outlook on future research is given to provide inspiration for selecting suitable matching datasets and exploring multimodal sentiment analysis systems.

Keywords

Visual Information, Multimodal Sentiment Analysis, Feature Extraction, Modality Fusion

1. 引言

现今人们处在一个人人可以表达观点、情感，人人可以输出价值、知识的时代，海量且多模态的数据内容很容易获得，且大众更易受其影响。《心理学大辞典》中认为：“情感是人对客观事物是否满足自己的需要而产生的态度体验。”而我们对于现实世界的感知和感受，包括我们做出的任何选择，在很大程度上受到他人对于当前世界的洞察和观点的影响。于是，情感分析“应运而生”。其可以从文本中分析出人们对实体及其属性所表达的观点、情感、评价、态度和情绪。从消费产品到健康医疗，再到旅游、医院、金融服务、社会事件乃至政治选举，情感分析日益显现出其价值。

近些年，随着研究的深入，不同于以往只通过文本传递信息，越来越多的用户倾向于使用多种媒体形式(如文本加上图像、文本加上歌曲、文本加上视频等)共同表达他们的态度与情感[1]。这种来源丰富的数据内容可以帮助研究者们更准确地提取表达者对某一对象不同方面的情感，结合两种及以上模态的数据建模分析，就是多模态情感分析，是一新兴领域。目前，衡量视觉信息(图片或视频)的情感分析研究相对较少，视觉信息与文本之间的语义匹配关系较为复杂，存在着不同层面上的语义对齐关系。综合考虑不同视觉信息与文本不同层面的语义段之间的匹配关系，充分地挖掘这各模态之间的情感交互特征，对于精确的多模态情感分析具有重要的作用和意义。

本文详细总结了带有视觉信息的多模态情感分析数据集、模态特征提取算法、不同模态数据融合方法、深度学习前沿技术，旨在为未来研究提供更多可能的思路。

2. 多模态情感分析数据集

本节总结了多模态情感分析中较流行的几个数据集概况，形成表 1。其中，数据集的名称显示在第一列，数据集发布的年份显示在第二列，每个数据集中包含的视频数量显示在第三列，话语的数量显示在第四列，在某一些数据集中，由于作者没有提及关于话语数量的指标，故该列为空值。第五列表示在整个数据集中，不同发言者的数量，第六列和第七列分别是数据集使用的语言和数据来源。第八列为数据集的下载地址，第九列说明了数据所使用的标签，最后一列显示了该数据集所属的领域，比如产品评论、电影评论、辩论等等。

2.1. YouTube 数据集

YouTube 数据集由 Morency 等人[2]于 2011 年开发。其来源于 YouTube 网站，且不局限于单一特定

的主题，其内容侧重以下关键词：意见、评论、产品评论、牙膏、战争、工作、商业、化妆品评论、相机评论、婴儿产品评论，“我讨厌”，“我喜欢”等等。该数据集内容丰富，由 47 个视频组成，每个视频包含 3~11 个话语。讲述者的年龄从 14 岁到 60 岁不等，其中 40 个视频由女性表达，其余的视频由男性表达。虽然所有的讲述者都来自不同的文化，但他们统一用英语表达了自己的观点。数据集中的每个视频都以积极、消极或中性这三个标签标记，最后形成 13 个积极的视频，12 个消极的视频和 22 个中性的视频。

2.2. MOUD 数据集

Multimodal Opinion Utterances Dataset (MOUD)是由 Perez-Rosas 等人[3]在 2013 年创建的，其从 YouTube 网站收集了 80 个视频，主要包括产品评论或推荐方面的内容。讲述者的年龄从 20 岁到 60 岁不等，其中 15 个视频由女性者表达，其余的视频由男性表达。所有的视频均用西班牙语录制。数据集中的每个视频都以积极、消极或中性这三个标签标记，最后形成了 182 个积极，231 个消极和 85 个中性标签的共 498 个视频的多模态数据集，且平均持续时间为 5 秒。

2.3. ICT-MMMO 数据集

The Institute for Creative Technologies Multi-Modal Movie Opinion (ICT-MMMO)数据集由 Wollmer 等人[4]于 2013 年创建。该数据集包括从 YouTube 和 ExpoTV 两个网站上收集的 370 个在线英文电影评论视频。数据集的标签为以下几种：强积极、弱积极、中性、强消极和弱消极。

2.4. POM 数据集

Persuasive Opinion Multimedia (POM)数据集是 S. Park 等人[5]从 ExpoTV 上收集的 1000 篇影评，所有的影评均为英文表达。每篇电影影评都是一段讲述者谈论某一特定电影的视频，以及该讲述者对电影的评分，从 1 星(最负面)到 5 星(最正面)。数据集中的每个视频均是某一个人谈论一部特定电影的正面拍摄，视频的平均长度约为 93 秒。该数据集有两个方面的广泛应用：其一，它被用来研究在线社交多媒体背景下的说服力。每个视频都从 1 (非常没有说服力)到 7 (非常有说服力)。其二，它被用来识别说话者的特征。每个视频都有可能带有讲述者的以下特征：自信、娱乐、信任、激情、放松、说服、支配、紧张、可信、娱乐、保留、信任、放松、紧张、幽默和有说服力。共 903 个视频，其中 600 个用于训练，100 个用于验证，203 个用于测试。

2.5. Yelp 数据集

该数据集来自 Yelp.com，从食品和餐馆类别中抓取的在线评论，涵盖了美国 5 个不同的主要城市：波士顿(BO)、芝加哥(CH)、洛杉矶(LA)、纽约(NY)和旧金山(SF)。其中洛杉矶的数据量是最庞大的，拥有最多的文件和图像。波士顿最小。然而，文档的长度，就句子数和单词数而言，在这五个城市中是非常相似的。该数据集总共有超过 44,000 条评论，其中包括 24.4 万张图片，且每篇评论至少有 3 张图片，并包括一个 1 到 5 分的评分。

2.6. CMU-MOSI 数据集

CMU-MOSI 数据集由 Amir Zadeh 等人[6]于 2016 年开发。该数据集由 YouTube 网站上收集到的 93 个 vlog 组成。这种类型的视频通常只有一个讲述者，讲述者的年龄从 20 岁到 30 岁不等，其中 41 个视频由女性表达，其余的视频由男性表达，讲述者均用英语表达观点。

该数据集的一个优势是能处理多样性的内容，包含噪音。另外需要注意，所有视频都是在不同的设

置下录制的，会存在差异，一些用户使用高科技的麦克风和摄像头，而另一些用户则使用不太专业的录音设备。其次用户与相机的距离不同，背景和照明条件不同，也会存在一些差异。这些视频保持了原来的分辨率，没有任何质量方面的提高。数据集的标签分为：强烈阳性(+3)、阳性(+2)、弱阳性(+1)、中性(0)、弱阴性(−1)、阴性(−2)、强烈阴性(−3)。

2.7. CMU-MOSEI 数据集

CMU-MOSEI 数据集由 Zadeh 等人[7]于 2018 年开发。CMU-MOSEI 是一个更大规模的数据集，由 3228 个视频组成，包含了 22777 个来自 1000 多个 YouTube 用户(57%男性，43%女性)的话语。这些视频讨论了 250 个不同的主题，其中最常见的 3 个主题是：评论(16.2%)、辩论(2.9%)和咨询(1.8%)。和 CMU-MOSI 数据集一样，CMU-MOSEI 也可以处理多样性并包含噪声。数据集中的每一个话语都被标记为以下八种情绪中的一种：强烈积极(+3)、阳性(+2)、弱阳性(+1)、中性(0)、弱阴性(−1)、阴性(−2)、强烈阴性(−3)。

最近越来越多的研究使用了 CMU-MOSI 和 CMU-MOSEI 这两个数据集来评估他们的模型在多模态情感分析中的性能。

2.8. CH-SIMS 数据集

CH-SIMS 数据集是由 Yu 等人[8]于 2020 年开发的。该数据集由 60 个视频组成，包括从电影、电视剧和综艺节目中收集的 2281 个话语。话语的平均长度为 3.67 秒，在每个视频中，除了说话者的脸外，没有其他面孔出现。对于每个话语，每个视频给出一个多模态标签和三个单模态标签。这可以帮助研究者使用 SIMS 来进行单模态和多模态情感分析任务。该数据集中的标签为：正、弱正、中性、弱负、负。

2.9. 其他常用数据集总结

表 2 列举了其他常用数据集的基本概况，供有兴趣的学者自行按需下载使用。

Table 1. Summary of typical multimodal dataset profiles
表 1. 典型的多模态数据集概况总结

Dataset	创建年份	视频/图片总数	话语总数	讲述者数量	语言	来源	情感标签	讨论主题	下载地址
YouTube	2011	47	280	47	英语	YouTube	积极、消极、中性	产品评论	发送邮件到 stratou@ict.usc.edu
MOUD	2013	80	498	80	西班牙语	YouTube	积极、消极、中性	无特定主题	http://web.eecs.umich.edu/~mihalcea/downloads.html
ICT-MMMO	2013	308	/	370	英语	YouTube、ExpoTV	[−2,+2]	电影评论	发送邮件到 stratou@ict.usc.edu
POM	2014	1000	/	352	英语	ExpoTV	[1,7]	电影评论	/
Yelp	2014	244,000	44,000	/	英语	Yelp	[1,5]	餐厅评论	https://www.yelp.com/dataset
CMU-MOSI	2016	93	2199	89	英语	YouTube	[−3,+3]	无特定主题	https://www.amir-zadeh.com/datasets

Continued

CMU-MOSEI	2018	3228	22,777	1000	英语	YouTube	[−3,+3]	无特定主题，占比前三的主题是：评论(16.2%)、辩论(2.9%)和咨询(1.8%)	http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/
CH-SIMS	2020	60	2281	474	中文	/	[−2,+2]	电影、电视剧、综艺节目	https://github.com/thuiar/MMSA

Table 2. Other multimodal datasets
表 2. 其他多模态数据集

Dataset 名称	下载地址
IEMOCAP	https://sail.usc.edu/iemocap/
EmoReact—Children Emotion Dataset	http://multicomp.cs.cmu.edu/resources/emoreact-dataset/
MVSA—multiview social dataset	http://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data/
RECOLA	https://diuf.unifr.ch/main/diva/recola/
VAM	https://sail.usc.edu/VAM/vam_info.html
RAVDEES	https://zenodo.org/record/1188976#.YHL_yegzbIU

3. 特征抽取

各模态特征抽取是多模态情感分析的第一步，本节对这部分技术进行对应梳理。

3.1. 文本特征抽取

文本特征主要采用单词袋(BOW)、术语频率和逆文档频率(TF-IDF)、N-grams 和单词嵌入等技术进行提取。Dash，Rout 等人[9]研究了不同文本特征的有效性，识别 Twitter 数据集中存在的情感。他们提取了文本特征，并通过三种机器学习技术进行情感分类，利用 SVM 分类器得到了最佳的分类精度。Vinodhini 和 Chandrasekaran [10]提出了一种有效的在线评论情感分类的方法，使用了单语、双语和三语文本特征的组合。情感分类采用不同的技术，通过概率神经网络获得了最高的性能。也有学者 Kaibi 和 Nfaoui [11]使用 word2vec (一种基于神经网络架构的流行单词嵌入模型)、全局向量(GloVe)模型和 FastText 模型进行文本特征提取，在 Twitter 数据集上进行实验。其中，FastText 模型在使用 SVM 分类器的情况下情感分类是最有效的。Ahuja，Chug，Kohli 等人[12]研究了两种文本特征提取方法，即 TF-IDF 和 N-Grams 对情感分析的影响。结果显示，情感分析的 TF-IDF 词级性能比基于 N-gram 的特征性能高出约 3%~4%。Mohey [13]提出了一个增强的 BOW 模型，并对 CiteULike 网站上的文本评论进行了情感分析。该增强算法的分类准确率为 83.5%，优于标准的 BOW 算法(62%)。Poria，Cambria 等人[14]的研究证明了基于 CNN 的方法在提取文本特征方面是有效的。它构造了包含每个文本的重要信息的特征向量，并且该向量具有代表整个文本的特征的集合。而 CNN 的输出可以被提供给一个更简单的有助于训练网络的分类器，CNN 作为一种有监督的方法，能很好地适应数据集中的异常现象。

3.2. 视觉信息抽取

视觉信息提取的部分，分为两种，从图片中提取和从视频中提取。若为视频，其涉及到将每个视频

剪辑分割成若干帧，并从每一帧中提取出各种特征。特别利用身体手势特征识别情绪方面，Piana，Staglianò等人[15]通过提取手势中的某些特征来自动识别与运动学相关的情感，他们使用了 SVM 分类器来自动执行情感预测的任务。Noroozi 等人[16]利用面部特征与手势特征的结合，进行了有效的情感分析。Yakaew，Dailey 等人[17]对 RAVDEES 数据集中出现的视频进行了情感分析。该数据集主要是从专业演员那里收集的视频，并有着不同的情感类别，如中性、快乐、冷静、悲伤、恐惧、愤怒、惊讶和厌恶。他们将视频分成若干帧后，提取出了反映演员情感的唇部和面部区域。对于人脸检测，他们使用了带有 SVM 的梯度直方图。

若从图片中提取特征，Song [18]认为图片情感分析是对图像所传达信息的一种高级语义理解，它可以弥合低层次视觉特征和高层次情感之间的巨大情感鸿沟。图片情感分析方法主要包括基于传统的机器学习的方法和基于深度学习的方法两大类。传统图片情感分析方法，首先需要从数字化的图片中提取出不同维度的视觉特征，通过情感空间模型的映射，使用机器学习算法和模型训练数字化图片中隐含的情感[19]。近些年来，深度学习技术在图像处理、机器翻译和语音识别等领域取得了很多成果，通过对底层特征进行一系列的非线性变换和组合，抽象概括出具有高阶语义信息的高层特征[20]。You 等人[21]设计了 CNN 架构来处理情感分析任务，对于大规模弱标记的数据集，在首次训练结束后，去除该训练数据集中情感分类置信度较低的样本，然后再使用过滤后的数据集进一步调整模型的参数，通过这种渐进式训练策略降低训练集中噪音样本对模型学习的影响。此外，作者通过迁移学习机制，使用 Twitter 图片数据集来微调该 CNN 网络参数，实验结果显示该 CNN 网络具有较好的泛化性能。Mittal 等人[22]介绍了在图像情感分析中的 DNN、CNN、基于区域的 CNN (Regional CNN, R-CNN)和 Fast R-CNN，并研究了它们的适应性与局限性。

近年来 transformer 方法也逐渐受到关注。Transformer 首先应用于自然语言处理领域，是一种主要基于自我注意机制的深度神经网络，与 2017 年被引入[23]。由于其强大的表现能力，在视觉模态提取部分提供了帮助。Felicia 等人[24]提出了混合 LSTM-Transformer 分类器，在 RAVDESS 和 Emo-DB 两个数据集上准确率分别达到了 75.62%和 85.55%。Heusser 等人[25]提出了一种基于预训练的 Transformer 语言模型，在 IEMOCAP 数据集上进行了评估，识别率达到了 73.5%。在文献[26]中，提出了一种改进的 Transformer 模型，并在多头注意中使用了不同的位置编码和泰勒线性注意(TLA)方法。该模型在 Emo-DB 数据集上测试时达到了 74.9%的准确率，在 URDU 数据集上测试时达到了 80%。然而，Transformer 的缺点是，它丢失了其位置的顺序信息，需要在每个时间步长中重新计算上下文窗口中的整个历史记录[27]。

4. 模态融合技术

在执行最后的分类任务之前要将集成的各模态特征(文本、音频和视频)进行融合。多模态数据的融合过程可以提供额外的情感信息，从而提高总体结果的精度。目前的多模态融合技术有以下几种：早期或特征级融合、后期或决策级融合、混合融合、模型级融合和规则级融合。

4.1. 特征级融合

该方法通过将各模式中提取的特征组合为一个单一的特征向量，其主要优点是可以用来在早期阶段识别多模态数据特征之间存在的相关性，从而提供准确的结果。从不同的模式(文本、音频和视频)中提取的特征在融合过程开始时被转换为相同的格式，采用特征级融合得到了单模态融合的最佳结果，与现有方法相比，处理时间也有所提高。Monkaresi 等人[28]结合了头部运动、生理学和面部颜色进行情感分类。与单模态方法相比，它取得了统计上更高的精度。

4.2. 决策级融合

该融合方法接受每个模态独立特征抽取后的独立分类结果，将局部结果融合为决策向量给出最终的决策结果。这种方法的优点是每个模态可以使用其最适合的分类器来学习其特征，但也因此而耗时过大。Cai 等人[29]和 Dobriek 等人[30]已经证明，对于情感预测，后期融合比早期融合获得了更好的结果。

4.3. 混合融合

顾名思义，该方法结合了特征级和决策级融合技术，利用这两者的优点来完成情感分析任务。Wöllmer 等人[31]采用了混合融合，运用 Bi-LSTM 组合音频和视觉特征，其在由在线评论视频(370 个)组成的 ICT-MMMO 数据集上得到验证，通过融合这些模态，获得了 65.7%的 F1 值，高于单模态性能。Siddiquie 等人[32]通过利用语音的声谱图特征和视频的情感本体融合了多模态信息，因每个模态包含对其他模态的补充信息，最终提高了准确率。为了从面部表情和语音中识别人类情感，Mansoorizadeh 和 Charkari [33]通过创建混合特征空间的新方法建立了模型，与基于单模态的方法相比，所提出的融合模型实现了更高的准确率。

4.4. 模型级融合

该方面的技术基于 Lin 等人[34]在文章中指出了在不同模式下提取的数据之间的关系。Zeng 和 Hu 等人[35]引入了一个多流融合的隐马尔可夫模型，用于视听内容的影响识别。该方法在 11 种情感状态的识别中进行了实验，结果发现即使在有音频噪声干扰的情况下，它也表现良好。为了从各模态中识别情感，Sebe，Cohen 等人[36]通过基于概率的方法融合模态，使用贝叶斯网络模型，该算法在不同情感状态下采集的视听数据上进行了实验，取得了较好的情感识别效果。Song 等人[37]提出了一种三重隐马尔可夫模型，该模型基于视听输入对相关属性进行建模，该模型可以自动识别情绪，其结果显示了其优于单模态方法。

4.5. 规则级融合

规则级融合采用加权融合和投票表决等技术来实现模态融合。在加权融合的情况下，来自多个模态的特征使用乘积或和这样的操作符来组合。根据 Al-Azani 和 El-Alfy [38]，这种加权方法成本较低，并且单个模态被赋予归一化权重。但是这种方法的问题是，为了更好地执行，权重必须被适当地规范化。在基于投票表决的融合中，多数分类器做出的决策起着至关重要的作用。Corradini，Mehta 等人[39]融合了语音和 2D 手势，其主要是与玩游戏的计算机系统的交互过程中获得。他们开发了一种人机交互系统，可以识别用户的手势并采取相应的行动。Iyengar 等人[40]使用标准化语音模型和面部分数的线性加权来检测视频中的独白，他们的结论是加权总和法可能优于乘积加权法。

5. 基于深度学习的前沿技术

回归到情感分析的技术发展，从方法的演进上大致经过以下几个阶段：构建情感词典、基于机器学习的方法和基于深度学习的方法[41]。考虑到现在研究的发展倾向，本文对基于深度学习的方法做一些阐述与总结。深度学习的概念是机器学习的一个高级领域，正如 Silver 等人[42]指出的那样，它使用多层网络来克服 CNN 的障碍。深度学习的主要优势是它不需要任何预定义的功能，可以自己学习输入数据的特征，能通过有效捕捉用户和项目之间的非线性关系来更好的完成情感分析任务。

5.1. 使用自动编码器的情感分析

自动编码器是一种有效压缩和编码数据的前馈神经网络。其基于无监督的学习概念，将编码输出重

构为更接近原始输入的值。有编码方法、解码和损失函数几个要素。自动编码器被有效地用于执行分类任务，Vincent 等人[43]提出了一种新颖的上下文自动编码器并被成功地用于情感分析，在来自 facebook 和 Twitterd 网站的数据集上实验时，优于传统的词袋方法和其他方法。

5.2. 基于递归神经网络(RNN)的情感分析

递归神经网络(RNN)可以克服神经网络遇到的缺点，在神经网络中，其当前决策是基于其从过去学习到的知识，它们就像带有记忆的顺序网络一样工作，每个隐藏层项目的值也依赖于先前的状态。RNN 被有效地用于语音和手写识别应用[44]。RNN 的增强版本是门控循环单元和长短期记忆单位(LSTM)。为了对评论进行情感分析，Pal 等人[45]应用了不同的 LSTM 架构，他们验证了 RNN-LSTM 比 CNN 更有效，双 LSTM 模型的性能优于其他 LSTM 的架构。文献[46]在 LSTM 的基础上提出了 TD-LSTM 和 TC-LSTM，这是第一次考虑方面并将方面与背景联系起来以提高分类准确性的文章。与 RNN 相比，CNN 更擅长在情感分析中捕捉本地特征。文献[47]提出了一个基于注意力的双向 CNN-RNN 深度模型，该模型通过双向 LSTM 和 GRU 层提取过去和未来的上下文，效果较优。

5.3. 基于自我注意机制的情感分析

2017 年，谷歌发表了一篇名为《注意力是你需要的全部》的文章，提出了自我注意机制[23]。后来围绕自我关注和多头自我关注开展了一些研究工作。Letarte 提出了 SAnet 模型，通过大量实验证明了自我注意机制对于情感分析的重要性[48]。有学者[49]提出了结合自我注意机制和多通道特征的情感分类 SAMFBiLSTM 模型，该模型还指出词性在情感分类中起着重要作用。文献[50]使用全局注意和局部注意来捕捉上下文和方面之间的交互。与基于 LSTM 的模型相比，自我注意可以获得单词之间的依赖关系，并且可以并行计算。

6. 现有模型比较

本文在此比较了现有的准确率较高的几个模型，标注了出处，并给出了每个模型各模态特征提取所用的方法，这些模型均在 MOSI 数据集上进行了性能评估，效果及详情见表 3。

Table 3. Comparison of the effects of existing models on the MOSI dataset
表 3. MOSI 数据集上现有模型效果比较

出处	模型	准确率	F1 值	文本模态	视觉模态	语音模态
[51]	MHSAN	78.70%	/	word2vec	3D-CNN	openSMILE
[52]	Multilogue-Net	81.19%	80.10%	CNN	3D-CNN	openSMILE
[53]	HFFN	80.19%	80.34%	Glove	FACET	COVAREP
[54]	MMMU-BA	82.31%	/	word2vec	3D-CNN	openSMILE
[55]	MARNN	84.31%	/	word2vec	3D-CNN	openSMILE
[56]	MulT	83%	82.80%	Glove	FACET	COVAREP

7. 结论

本文阐述了衡量视觉信息的多模态情感分析技术，详尽梳理了有关数据集、特征抽取、模态融合、先进技术等方面的内容，旨在帮助更多学者在多模态情感分析领域进行更广泛的研究。同时讨论了各种方法的优缺点，使研究者能够选择合适的应用方法和匹配的数据集。

对于未来可能的研究方向，总结为以下三点：1) 在多模态情感分析数据集的选择上，一是突破单一模态下的局限，尝试将更多模态的数据结合到文本信息当中是当前研究热点，目前多模态数据集还缺少身体姿态、手势等形式的数据集，在多样性上还有增益空间。2) 特征抽取和模态融合部分，要有连贯、合理的可解释性。目前一个多模态情感分析任务的有效完成需考虑好两个问题：一是要提取相邻话语之间的上下文关系，二是在融合各模态数据时，需对起重要作用的模态优先排序。3) 采用注意力机制等前沿技术能有效提高精度，且目前研究来看双头注意模块在跨视图动态建模中比自我注意模块更鲁棒。将注意模块加在多模态情感分析体系中的哪一环节，仍是接下来研究者需重点探究的问题。

参考文献

[1] 张亚洲, 戎璐, 宋大为, 张鹏. 多模态情感分析研究综述[J]. 模式识别与人工智能, 2020, 33(5): 426-438.

[2] Morency, L.P., *et al*. (2011) Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web. *Proceedings of the* 13*th International Conference on Multimodal Interfaces*, Alicante, 14-18 November 2011, 169-176.

[3] Perez-Rosas, V.M. and Morency, L.-P. (2013) Utterance-Level Multimodal Sentiment Analysis. *Proceedings of the* 51*st Annual Meeting of the Association for Computational Linguistics*, Volume 1, 973-982.

[4] Wollmer, M., Knaup, T., *et al*. (2013) Youtube Movie Reviews: Sentiment Analysis in an Audio-Visual Context. *IEEE Intelligent Systems*, **28**, 46-53. https://doi.org/10.1109/MIS.2013.34

[5] Park, S.S., *et al*. (2016) Multimodal Analysis and Prediction of Persuasiveness in Online Social Multimedia. *ACM Transactions on Interactive Intelligent Systems*, **6**, Article 25. https://doi.org/10.1145/2897739

[6] Morency, L.-P., *et al*. (2016) MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos.

[7] Zadeh, A.A.B., Poria, S., Cambria, E. and Morency, L.P. (2018) Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. *Proceedings of the* 56*th Annual Meeting of the Association for Computational Linguistics*, Volume 1, 2236-2246. https://doi.org/10.18653/v1/P18-1208

[8] Yu, W., *et al*. (2020) CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-Grained Annotation of Modality. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, July 2020, 3718-3727. https://doi.org/10.18653/v1/2020.acl-main.343

[9] Dash, A.K., Rout, J.K. and Jena, S.K. (2016) Harnessing Twitter for Automatic Sentiment Identification Using Machine Learning Techniques. *Proceedings of* 3*rd International Conference on Advanced Computing*, *Networking and Informatics*, Vol. 44, 507-514. https://doi.org/10.1007/978-81-322-2529-4_53

[10] Vinodhini, G. and Chandrasekaran, R.M. (2019) A Comparative Performance Evaluation of a Neural Network-Based Approach for Sentiment Classification of Online Reviews. *Journal of King Saud University of Computer and Information Sciences*, **28**, 2-12. https://doi.org/10.1016/j.jksuci.2014.03.024

[11] Kaibi, I. and Nfaoui, E.H. (2019) A Comparative Evaluation of Word Embeddings Techniques for Twitter Sentiment Analysis. 2019 *International Conference on Wireless Technologies*, *Embedded and Intelligent Systems* (*WITS*), Fez, 3-4 April 2019, 1-4. https://doi.org/10.1109/WITS.2019.8723864

[12] Ahuja, R., Chug, A., Kohli, S., Gupta, S. and Ahuja, P. (2019) The Impact of Features Extraction on the Sentiment Analysis. *Procedia Computer Science*, **152**, 341-348. https://doi.org/10.1016/j.procs.2019.05.008

[13] Mohey, D. (2016) Enhancement Bag-of-Words Model for Solving the Challenges of Sentiment Analysis. *International Journal of Advanced Computer Science and Applications*, **7**, 244-252. https://doi.org/10.14569/IJACSA.2016.070134

[14] Poria, S., Cambria, E., Hussain, A. and Huang, G.-B. (2015) Towards an Intelligent Framework for Multimodal Effective Data Analysis. *Neural Networks*, **63**, 104-116. https://doi.org/10.1016/j.neunet.2014.10.005

[15] Piana, S., Staglianó, A., Odone, F., Verri, A. and Camurri, A. (2014) Real-Time Automatic Emotion Recognition from Body Gestures.

[16] Noroozi, F., Corneanu, C.A., Kaminska, D., Sapinski, T., Escalera, S. and Anbarjafari, G. (2018) Survey on Emotional Body Gesture Recognition.

[17] Yakaew, A., Dailey, M. and Racharak, T. (2021) Multimodal Sentiment Analysis on Video Streams Using Lightweight Deep Neural Networks. *Proceedings of the 10th International Conference on Pattern Recognition Applications and Methods*, 4-6 February 2021, 442-451. https://doi.org/10.5220/0010304404420451

[18] Song, K., Yao, T., Ling, Q., *et al*. (2018) Boosting Image Sentiment Analysis with Visual Attention. *Neurocomputing*, **312**, 218-228. https://doi.org/10.1016/j.neucom.2018.05.104

[19] 王仁武, 孟现茹. 图片情感分析研究综述[J]. 图书情报知识, 2020(3): 119-127.

[20] 朱雪林. 基于注意力机制的图片文本联合情感分析研究[D]: [硕士学位论文]. 南京: 东南大学, 2019.

[21] You, Q.Z., Jin, H.L. and Luo, J.B. (2017) Visual Sentiment Analysis by Attending on Local Image Regions. *Thirty-First AAAI Conference on Artificial Intelligence*, **31**, 231-237. https://doi.org/10.1609/aaai.v31i1.10501

[22] Mittal, N., Sharma, D., Joshi, M.L., *et al*. (2018) Image Sentiment Analysis Using Deep Learning. In: *Proceedings of the* 2018 *IEEE/WIC/ACM International Conference on Web Intelligence*, IEEE, Piscataway, 684-687. https://doi.org/10.1109/WI.2018.00-11

[23] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 5998-6008.

[24] Andayani, F., Theng, L.B., Tsun, M.T.K. and Chua, C. (2022) Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files. *IEEE Access*, **10**, 36018-36027. https://doi.org/10.1109/ACCESS.2022.3163856

[25] Heusser, V., Freymuth, N., Constantin, S. and Waibel, A. (2019) Bimodal Speech Emotion Recognition Using Pre-Trained Language Models.

[26] Jing, D., Manting, T. and Li, Z. (2021) Transformer-Like Model with Linear Attention for Speech Emotion Recognition. *Journal of Southeast University*, **37**, 164-170.

[27] Sakatani, Y. (2021) Combining RNN with Transformer for Modeling Multi-Leg Trips. *ACM WSDM WebTour* 2021, Jerusalem, 12 March 2021, 50-52.

[28] Monkaresi, H., Hussain, M.S. and Calvo, R.A. (2012) Classification of Affects Using Head Movement, Skin Color Features and Physiological Signals. 2012 *IEEE International Conference on Systems*, *Man*, *and Cybernetics* (*SMC*), Seoul, 14-17 October 2012, 2664-2669. https://doi.org/10.1109/ICSMC.2012.6378149

[29] Cai, G. and Xia, B. (2015) Convolutional Neural Networks for Multimedia Sentiment Analysis. In: Li, J., Ji, H., Zhao, D. and Feng, Y., Eds., *Natural Language Processing and Chinese Computing*, Vol. 9362, Springer International Publishing, Nanchang, 159-167. https://doi.org/10.1007/978-3-319-25207-0_14

[30] Dobrisek, S., Gajsek, R., Mihelic, F., Pavesic, N. and Struc, V. (2013) Towards Efficient Multi-Modal Emotion Recognition. *International Journal of Advanced Robotic Systems*, **10**, 53. https://doi.org/10.5772/54002

[31] Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K. and Morency, L.-P. (2013) YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context. *IEEE Intelligent Systems*, **28**, 46-53. https://doi.org/10.1109/MIS.2013.34

[32] Siddiquie, B., Chisholm, D. and Divakaran, A. (2015) Exploiting Multimodal Affect and Semantics to Identify Politically Persuasive Web Videos.

[33] Mansoorizadeh, M. and Charkari, M. (2014) Multimodal Information Fusion Application to Human Emotion Recognition from Face and Speech. *Multimedia Tools and Applications*, **49**, 277-297. https://doi.org/10.1007/s11042-009-0344-2

[34] Lin, J.-C., Wu, C.-H. and Wei, W.-L. (2012) Error Weighted Semi-Coupled Hidden Markov Model for Audio-Visual Emotion Recognition. *IEEE Transactions on Multimedia*, **14**, 142-156. https://doi.org/10.1109/TMM.2011.2171334

[35] Zeng, Z., Hu, Y., Liu, M., Fu, Y. and Huang, T.S. (2006) The Training Combination Strategy of Multi-Stream Fused Hidden Markov Model for Audio-Visual Affect Recognition. *Proceedings of the* 14*th Annual ACM International Conference on Multimedia*, Santa Barbara, 23-27 October 2006, 65. https://doi.org/10.1145/1180639.1180661

[36] Sebe, N., Cohen, I., Gevers, T. and Huang, T.S. (2006) Emotion Recognition Based on Joint Visual and Audio Cues. 18*th International Conference on Pattern Recognition* (*ICPR*'06), Hong Kong, 20-24 August 2006, 1136-1139. https://doi.org/10.1109/ICPR.2006.489

[37] Song, M., Bu, J., Chen, C. and Li, N. (2004) Audio-Visual-Based Emotion Recognition—A New Approach. *Proceedings of the* 2004 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, 1020-1025. https://doi.org/10.1109/CVPR.2004.1315276

[38] Al-Azani, S. and El-Alfy, E.-S.M. (2020) Enhanced Video Analytics for Sentiment Analysis Based on Fusing Textual. *Auditory and Visual Information*, **8**, 15. https://doi.org/10.1109/ACCESS.2020.3011977

[39] Corradini, A., Mehta, M., Bernsen, N.O., Martin, J.C. and Abrilian, S. (2005) Multimodal Input Fusion in Human-Computer Interaction. In: *Data Fusion for Situation Monitoring*, *Incident Detection*, *Alert and Response Management*, IOS Press, Tsakhkadzor, 223-234.

[40] Iyengar, G., Nock, H.J. and Neti, C. (2003) Audio-Visual Synchrony for Detection of Monologues in Video Archives. *IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, Hong Kong, 6-10 April 2003, V-772. https://doi.org/10.1109/ICME.2003.1220921

[41] 刘兵. 情感分析: 挖掘观点、情感和情绪[M]. 北京: 机械工业出版社, 2019: 149-156.

[42] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G. and Hassabis, D. (2016) Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature*, **529**, 484-489. https://doi.org/10.1038/nature16961

[43] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. and Manzagol, P.-A. (2010) Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, **11**, 3371-3408.

[44] Sak, H., Senior, A. and Beaufays, F. (2014) Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. *Neural and Evolutionary Computing*, **1**, 1-5. https://doi.org/10.21437/Interspeech.2014-80

[45] Pal, S., Ghosh, S. and Nag, A. (2018) Sentiment Analysis in the Light of LSTM Recurrent Neural Networks. *International Journal of Synthetic Emotions*, **9**, 33-39. https://doi.org/10.4018/IJSE.2018010103

[46] Tang, D., Qin, B. and Feng, X. (2016) Effective LSTMs for Target-Dependent Sentiment Classification. *Proceedings of COLING* 2016, *the* 26*th International Conference on Computational Linguistics*: *Technical Papers*, Osaka, December 2016, 3298-3307.

[47] Basiri, M.E., Nemati, S. and Abdar, M. (2020) An Attention-Based Bidirectional CNN-RNN Deep Model for Sentiment Analysis. *Future Generation Computer Systems*, **115**, 279-294. https://doi.org/10.1016/j.future.2020.08.005

[48] Letarte G., Paradis, F. and Giguere, P. (2018) Importance of Self-Attention for Sentiment Analysis. *Proceedings of the* 2018 *EMNLP Workshop BlackboxNLP*: *Analyzing and Interpreting Neural Networks for NLP*, Brussels, November 2018, 267-275. https://doi.org/10.18653/v1/W18-5429

[49] Li, W., Qi, F. and Tang, M. (2020) Bidirectional LSTM with Self-Attention Mechanism and Multi-Channel Features for Sentiment Classification. *Neurocomputing*, **387**, 63-77. https://doi.org/10.1016/j.neucom.2020.01.006

[50] Xu, Q., Zhu, L. and Dai, T. (2020) Aspect-Based Sentiment Classification with Multiattention Network. *Neurocomputing*, **388**, 135-143. https://doi.org/10.1016/j.neucom.2020.01.024

[51] Cao, R., Ye, C. and Zhou, H. (2021) Multimodel Sentiment Analysis with Self-Attention. *Proceedings of the Future Technologies Conference* (*FTC*), Volume 1, 16-26. https://doi.org/10.1007/978-3-030-63128-4_2

[52] Shenoy, A. and Sardana, A. (2020) Multilogue-Net: A Context Aware RNN for Multi-Modal Emotion Detection and Sentiment Analysis in Conversation. *The* 58*th Annual Meeting of the Association for Computational Linguistics*, Seattle, 5-10 July 2020, 19-28. https://doi.org/10.18653/v1/2020.challengehml-1.3

[53] Mai, S., Hu, H. and Xing, S. (2019) Conquer and Combine: Hierarchical Feature Fusion Network with Local and Global Perspectives for Multimodal Affective Computing. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, July2019, 481-492. https://doi.org/10.18653/v1/P19-1046

[54] Chauhany, D., Poria, S., Ekbaly, A., *et al*. (2017) Contextual Inter-Modal Attention for Multi-Modal Sentiment Analysis. *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing*, Brussels, October-November 2018, 3454-3466.

[55] Kim, T. (2020) Multi-Attention Multimodal Sentiment Analysis. *ICMR*'20 *Proceedings of the* 2020 *International Conference on Multimedia Retrieval*, Dublin, 8-11 June 2020, 436-441.

[56] Liangy, P.P., Kolteryz, J.Z., Morencyy, L.P., *et al*. (2019) Multimodal Transformer for Unaligned Multimodal Language Sequences. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, 28 July-2 August 2019, 6558-6569.