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：　表情识别技术综述
作者：　洪惠群，沈贵萍，黄风华
网络首发日期：　2022-04-21
引用格式：　洪惠群，沈贵萍，黄风华．表情识别技术综述[J/OL]．计算机科学与探索．
https://kns.cnki.net/kcms/detail/11.5602.TP.20220420.1147.002.html

 doi: 10.3778/j.issn.1673-9418.2110049

表情识别技术综述

洪惠群[1,2,3]，沈贵萍[1,2,3+]，黄风华[1,2,3]

1.阳光学院 人工智能学院，福州 350015

2.阳光学院 空间数据挖掘与应用福建省高校工程研究中心，福州 350015

3.阳光学院 福建省空间信息感知与智能处理重点实验室，福州 350015

+ 通信作者 E-mail:992774639@qq.com

摘 要：面部表情是判断人类情感和人机交互的重要依据，传统机器学习和深度学习的发展，给面部表情识别分析带来了许多机遇与挑战。本文首先分析了表情识别与情感分析的内在联系与区别，指出表情识别侧重于识别面部的表情及情感，接着，总结归纳了基于单模态数据集和传统机器学习方法的表情识别技术及其优缺点，介绍了基于单模态数据集与深度学习方法的表情识别技术，然后指出基于单模态数据的表情识别技术的具有一定的局限性，如：数据集在数量和质量上较为不足，识别准确率普遍不高，多停留在实验室研究阶段等，引出基于多模态数据集的表情识别及模态间融合方法，并介绍常用的多模态表情数据集，分析了基于多模态数据集的表情识别技术及模态之间的融合技术，包含特征级融合、决策级融合及混合融合三种方式，最后对表情识别分析技术进行总结与展望：考虑到数据集问题，可构建更多自然环境下的高质量表情数据集，也可结合姿势、脑电波等生理信号构建多模态数据集，利用 GAN 网络进行数据增强，关注微表情的提取，以及研究多模态融合算法等。

关键词：人脸表情识别；多模态；模态融合

文献标志码：A **中图分类号：**TP391

Summary of expression recognition technology

HONG Huiqun[1,2,3], SHEN Guiping[1,2,3+], HUANG Fenghua[1,2,3]

1.College of Artificial Intelligence, Yango University, Fuzhou 350015, China

2.Fujian University Engineering Research Center of Spatial Data Mining and Application, Yango University, Fuzhou 350015, China

3.Fujian Key Laboratory of Spatial Information Perception and Intelligent Processing, Yango University, Fuzhou 350015, China

Abstract: Facial expression is an important basis for judging human emotion and human-computer interaction. The development of traditional machine learning and deep learning has brought many opportunities and challenges to facial expression recognition and analysis. Firstly, this paper analyzes the internal relationship and difference between expression recognition and emotion analysis, and points out that expression recognition focuses on identifying facial expression and emotion. Then, it summarizes the advantages and disadvantages of the expression recognition technology based on single-mode data set and used traditional machine learning method, and introduces the expression recognition technology based on single-mode data set and used deep learning method. Then it points

基金项目：国家自然科学基金项目（41501451）；福建省自然科学基金项目（2019J01088，2019J01087）。

This work was supported by National Natural Science Foundation of China Project (41501451); Fujian Natural Science Foundation Project（2019J01088，2019J01087）.

out that the expression recognition technology based on single-mode data has certain limitations, such as insufficient quantity and quality of data sets, generally low recognition accuracy, mostly staying in the laboratory research stage. Then it introduces the expression recognition and inter mode fusion methods based on multi-mode data sets, and introduces the commonly used multi-mode expression data sets, the expression recognition technology based on multimodal data set and the fusion technology between modes are analyzed, including feature level fusion, decision level fusion and hybrid fusion. Finally, the expression recognition analysis technology is summarized and prospected: considering the problem of data set, more high-quality expression data sets in natural environment can be constructed; and multimodal data sets can also be constructed combined with physiological signals such as posture and EEG; Gan network can use to enhance data, pay attention to the extraction of micro expression, and study multimodal fusion algorithm.
Key words: Facial expression recognition; Multi-modal; Modal fusion

表情、声音、文本、姿态等，都可以用来表达人类情感，面部表情是人类情感表达的重要依据之一[1]，因此，计算机可以尝试通过分析人的面部表情来理解人的情感[2]，并在众多人机交互系统中融入，例如：各类服务型机器人、辅助检测疲劳驾驶、医疗服务、远程教育中学生学习状态监测等[3]。尽管在人们社交过程中，逐渐演化出各种的复杂的面部动作和表情来表达内心的情感[3]，但是学术界普遍研究的都是由 Friesen 和 Ekman 等心理学家提出的 6 种基本情感类别，即“高兴、愤怒、悲伤、吃惊、厌恶、恐惧”[4]。

随着计算机视觉及人工智能技术的发展[5-7]，人脸表情识别吸引着越来越多的学者进行研究。表情识别侧重于识别面部的表情及情感，而情感分析则可以根据面部表情、语音、文本、姿态、脑电信号等各种信号来进行情感分析，在情感分析的过程中，有可能没有对面部表情这一模态进行分析。因此，可以将表情识别看作情感分析的一个研究方向。本文侧重于从面部表情识别的角度去归纳总结。

在面部表情识别过程中，研究者常常会尝试结合语音、文本、姿态、脑电波等多种模态信息进行分析，根据在面部表情识别过程中所使用的数据集是单一模态的面部表情数据还是面部表情数据结合其他模态的数据进行情感识别的不同，本文将表情识别算法分为：基于单模态数据的面部表情识别[8,9]和基于多模态数据的面部表情识别。

1 基于单模态数据的面部表情识别

基于单模态数据的面部表情识别主要根据面部表情这一模态来进行分析识别，包含如图 1 所示步骤，数据集采集、图像的预处理、表情识别及判断类别等。

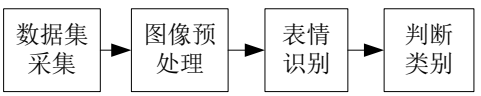

Fig.1 Main steps of multimodal facial expression recognition

图 1　单模态人脸表情识别主要步骤

1.1　数据集采集

表 1 总结了常见的表情识别数据集[10-27]的图像特点、标注类别及图像/视频数。其中：A1 代表实验室受控环境下的数据，A2 代表网站上非受控环境下的数据；B1 代表数据很少，B2 代表数据较少。

表 1 所示的数据集中，部分数据集为受控环境下的数据，数据少且皆为正面清晰人脸，标注可经过心理学专家反复确认，一般认为这些数据库标注是完全可靠的，如 CK+、JAFFE 等。部分数据集如：RAF-DB、AffectNet 等大规模数据集，是在非受控环境下取得，受标注者感知的主观性影响较大，标注质量相对比较低。因此，现有的数据集在数量和质量上均较为不足，数据量小，不足以很好地训练

目前在人脸识别任务中取得良好效果的较大深度网络结构，此外，现有的数据集缺乏具有遮挡类型和头部姿态标注的大型面部表情数据集，也会影响深度网络解决较大类内差距，学习高效表情识别能力特征的需求[28]。

Table 1 Common expression recognition data sets

表 1　常见的表情识别数据集

数据库	图像特点	标注的表情类别	图像/视频数
TFD[10]	A1、B2	7 类	112234 个图像
FER2013[11]	A2、B1	7 类	35887 个图像
NVIE[12]	A1、B1	6 类	1830 个视频
SFEW [13]	A2、B1	7 类	1766 图像
Multi-PIE[14]	A1、B2	6 类	755370 个图像
BU-3DFE[15]	A1、B1	7 类	2500 个图像
Oulu-CASIA[16]	A1、B1	6 类	2880 个图像
RaFD[17]	A1、B1	7 类	1608 个图像
KDEF[18]	A1、B1	7 类	4900 个图像
EmotioNet[19]	A2、B2	23 类基本或复合表情	100 万个图像
RAF-DB[20],[21]	A2、B1	7 类+12 复合类	29672 个图像
AffectNet[22]	A2、B2	8 类+V-A	44 万个图像
ExpW[23]	A2、B2	7 类	91793 个图像
CK+[24]	A1、B1	8 类	593 个图片
MMI[25],[26]	A1、B1	7 类	740 个图片、2900 个视频
JAFFE [27]	A1、B1	7 类	213 个图片

1.2　图像的预处理

图像预处理主要对原图像进行人脸对齐、数据增强及人脸归一化等操作。

1.3　面部表情识别

1.3.1 传统表情识别方法

传统的表情识别方法主要为浅层学习，或采用人工设计特征，需要人工较多地参与，常见的算法有：基于全局特征的提取方法[29-45]、基于局部的提取方法[46-49]、混合提取方法[50-53]的静态图像表情识别以及基于光流法[54]的动态视频的表情识别。具体方法及优缺点如表 2 所示。

1.3.2　基于深度学习表情识别方法

基于深度学习面部表情识别方法大体也可以分为基于静态图像的深度表情识别网络以及基于动态视频的深度表情识别网络。鉴于目前人脸表情数据库相对较小，直接进行深度学习网络训练，往往导致过拟合，为了缓解过拟合的问题，通常有如下几种方法：自建网络[55]、卷积网络微调[56]、分阶段微调[57]、多网络融合[58]、多通道级联[59]、生成对抗网络[60]、基于迁移学习的跨域人脸表情识别[61]等，现总结如下表 3。

基于单模态数据的表情识别准确率普遍不高，目前仍停留在实验室研究阶段，无法在实际生活中广泛运用。

2 基于多模态数据的面部表情识别

由上可知，基于单模态数据的表情识别具有一定的局限性，为了解决这些局限性，越来越多的学者们开始研究基于多模态数据的表情识别，希望能提高识别的准确率及稳定性。基于多模态数据的表情识别中，需要分别处理各模态的数据和对处理后的数据进行融合。在本文研究的多个模态中，有一个模态为面部表情数据。常见的辅助表情识别的模态有：语音、声音情绪、头部运动、手势识别、眼神交流、身体姿势、生理信号等。基于多模态数据的面部表情系统的处理框架如图 2[9]所示，该系统包含各个模态特征提取及模态信息融合。需要注意的是，单一模态数据的处理效果和多模态融合方式都很重要。在特征提取阶段，表情识别分析所采用的方法与上述基于单模态数据的面部表情的特征提取方法相同，模态融合的过程主要有三种方式：基于特征级、决策级、混合等三种融合[9]。下面，将分别总结常见的多模态数据集、多模态表情识别技术，模态融合技术等。

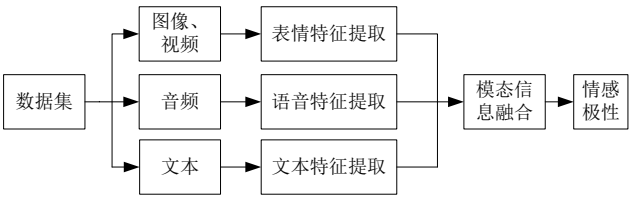

Fig.2 Framework of multimodal expression recognition

图 2　多模态表情识别的框架

2.1 多模态数据集

本文中所提到的多模态数据集应包含表情图片或视频作为其中一个模态，具体数据集总结如下表 4[62-74]。

表 4 中的多模态数据集都有表情视频或图像模态，辅以文字、音频、脑电、身体姿态等模态中的一个或多个，收集渠道有实验室录制，网上视频录制、实际环境中录制，包含有情绪或情感标签，基本都是小数据集。其中，数据模态的缩写规定如下：视频（video，V）、生理信号（physiological signal，PS）、音频（audio，A）、文字（text，T）、身体动作（body movement，BM）、面部动作（facial movements，FM）、图像（image，I）等。

2.2 基于多模态数据集的表情识别技术

现有的文献中，基于多模态数据集的表情识别技术主要根据面部表情、文本、语音以及脑电等的一个模态进行分析。文献[75-77]针对视频和音频模态进行分析，文献[78-79]针对视频和脑电模态进行分析，文献[80]针对表情视频和多模态传感器采集数据如眼动跟踪器、音频、脑电（Electroencephalogram，EEG）图、深度相机等模态进行分析，具体分析方法及优缺点如表 5 所示。文献[80]采用的视觉和非视觉传感器集成到面部表情识别的整体框图如图 3 所示。由表 5 及图 3 可知，基于多模态数据集的情感识别与融合虽然能够在一定程度上解决基于单模态表情识别的局限性，然而仍存在系统较复杂、识别准确率不够高等问题，需要进一步解决。

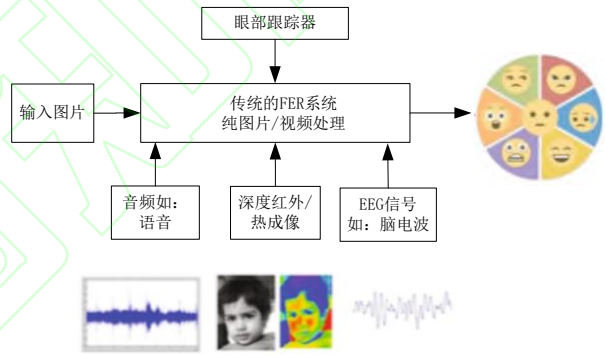

Fig.3 Integration of visual and nonvisual sensors into facial expression recognition

图 3　视觉和非视觉传感器集成到面部表情识别

Table 2 Traditional expression feature extraction method

表 2　传统表情特征提取方法

分类	主要方法		方法描述	优点	缺点
基于全局特征提取方法	1、主成分分析法（principal component analysis，PCA）[29-32]		从原始数据中提取主成分，降低特征维度，力求用较少的特征综合表达原始数据之间的关系	提取特征具有全局性	计算量大，识别率不高、无法利用训练样本中的类别信息
	2、基于纹理特征的提取方法	（1）局部二值模式（Local Binary Patterns，LBP）[33-40]	定义 3X3 的 LBP 算子，得到该窗口中心像素点的 LBP 值，并用这个值来反映该区域的纹理信息	具有旋转不变性和灰度不变性等，对光照变化不敏感	难以详细描述像素在邻域方向上的灰度值变化，难以满足不同尺寸图像纹理问题
		（2）Gabor 变换[41-45]	通过定义不同的带宽和方向对图像进行多分辨率分析，有效提取图像中的纹理特征	对光照变化不敏感，有明显方向选择性和频率选择性	当特征维度太大时，难以找到合适的参数变量
基于局部特征提取方法	1、基于几何特征的提取方法[46-49]		主动形状模型，提取人脸轮廓及眼睛、鼻子、嘴的位置	有效地提取人脸面部表情的显著特征	当人脸关键识别分类信息丢失时，会导致提取出的特征出现偏差，使得精度下降

混合提取方法	1、特定表情的局部二值模式（Expression-specific local binary pattern，es-LBP）[50]	提出了 es-LBP 特征，更好地捕获人脸在重要基点上的局部信息	es-LBP 特征优于传统的 LBP 特征	当图像受到噪声污染比较严重时，准确率下降
	2、光流法+图像梯度[51]	在 bag-of-words 设置中使用 OF 和 IG 这种以直方图形式的独特组合，使得所提出的特征在视频中独立于人脸的尺度并跟踪像素的运动，能够捕捉复杂的非刚性运动的面部组件	对皮肤色调的变化非常敏感，能克服如面部成分的非刚性运动以及不同肤色和尺度的障碍	从视频中每一张图像中提取特征的识别模型，特征的维数变得非常高，导致更高的时间和空间复杂性
	3、几何特征+纹理特征[52]	采用 Weber Local Descriptor（WLD）和 Histograms of Oriented Gradients (HOG)结合来表示局部细节	能有效提取图像纹理信息，对噪声和光照变化具有很强的鲁棒性，分类的准确率更高和所需时间更少	无法自适应权重，解决姿态和光照的变化
	4、局部 Gabor+分数次幂多项式核函数 PCA[53]	基于局部 Gabor 滤波组和分数次幂多项式核函数 PCA 的方法，并利用支持向量机对特征进行分类	降低计算复杂度，有效改善光照的影响，获得较好的识别率	
光流法[54]		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	识别模型和算法较为复杂，计算量大

Table 3 Expression recognition method based on deep learning

表 3　基于深度学习表情识别方法

分类	方法描述	优点	缺点
自建网络[55]	构建 7 层卷积神经网络，用大型人脸数据库预训练后，再用表情数据库微调训练；首次将 inception 层架构应用到跨多个数据库的表情识别	取得比传统方法更好的识别效果	由于表情数据库数据量过少，网络容易过拟合
卷积网络微调[56]	FaceNet2ExpNet 算法，先利用人脸网络的深层特征作为监督训练卷积层，待其训练完成后，加入随机初始化的全连接层，并从头开始训练	相比传统的机器学习方法和基于 VGG-16 网络方法，识别效果更好	仅在 CK+、OuluCASIA、TFD 和 SFEW 四个公开表达数据库进行对比
分阶段微调[57]	迁移了在 ImageNet 数据集预训练过的卷积神经网络（convolutional neural network，CNN），接着用 FER2013 数据集对该预训练模型进行微调，最后用目标数据库 EmotiW 对微调后的模型再进行微调	提高识别准确率	目标数据库使用的是野外静态人脸表情识别，识别准确率不超过 55.6%
多网络融合[58]	采用足够多样的子网络来提取足够多的特征，如：通过视觉分支网络负责图像序列的输入，引入低层到高层的跳转连接来考虑底层特征，通过合适的集合方法来高效融合各种子网络	能将面部大动作造成的面部特征变化综合考虑进去，在 CK+数据集上效果良好	未在更多的数据集上测试
多通道级联[59]	使用三个并行的多通道卷积网络从不同的面部区域学习融合的全局和局部特征，利用联合嵌入特征学习来探索基于融合区域的特征在嵌入空间中的身份不变和姿态感知的表达表示	在性能和鲁棒性优于现有的先进方法	仍要解决不受约束环境下表情识别准确率问题
生成对抗网络	使用反表达式剩余学习，先用 cGAN 训练生成相应输入人	能缓解身份变异问题，	网络比较复杂

（generative adversarial network，GAN）[60]	脸图像的中性人脸图像，通过学习留在生成模型中间层的沉练，建立了 DeRL 方法		处理自发表达和姿势表达的情况下风格和种族背景的不同	
基于迁移学习的跨域人脸表情识别[61]	引入稀疏重构思想获取共同投影矩阵并对积，利用两个数据库（BU-4DFE 和 BP4D+）进行预训重构系数矩阵，施加 L2,1 范式约束，引入图拉普拉斯正则化项来保留数据的局部判别结构，通过源域丰富的标签信息将样本投影到一个由标签引导的子空间中，并在 CK+、JAFFE、TFEID 等数据库中进行跨域验证		较好区分高兴和惊讶	对厌恶和生气区分度不高，同时伤心表情识别率相对较低

Table 4 Multimodal affective data set

表 4　多模态情感数据集

数据模态	数据库名称	标签类别	数据概况	数据来源
V、PS	DEAP 数据集[62]	消极到积极的九个分数	32 名男、女测试者	实验室采集
V、A、T	CH-SIMS 数据集[63]	-1（负）、0（中性）或 1（正）	2281 个长度 1-10 秒视频片段	网络视频采集
	YouTube 数据集[64]	积极、消极、中性三种标签	20 名女性和 27 名男性对产品的观点描述，包含 13 个积极、22 个中性以及 12 个消极标签的视频序列	YouTube 网采集
	CASIA 汉语自然情感视听数据库[65]	6 类情绪标签	140 分钟情感片段，由四个专业发音人录制，共 9600 句不同发音片段	电影、电视剧和脱口秀等采集
V、A、BM、FM、T 等	IEMOCAP 数据集[66]	类别标签+维度标签	12 小时的视听数据，通过表演激发情感表达	实验室采集
A、V	SAVEE 数据库[67]	7 种情感	480 段，4 名男性演员录音	实验室采集
	eNTERFACE 05 数据集[68]	6 种情感标签	42 个受试者、14 个不同国籍、1260 个视频序列	实验室采集
	ICT-MMMO 数据集[69]	正面、中立和负面三种评论	228 个正面、23 个中立和 119 个负面的多模态电影评论视频	YouTube、ExpoTV 网采集
	MOSI 数据集[70]	-3 到+3 的 7 类情感	不同年龄、不同种族的 48 名男性，41 名女性的 2～5 分钟的电影评论	YouTube 网采集
	CMU-MOSEI 数据集[71]	7 类情感标签和 6 类情绪标签	含 3228 个视频，共计 23453 个句子	YouTube 网采集
T、I、A	NewsRoverSentiment 数据集[72]	3 类的情感标签	929 个 4～15 秒的新闻视频组成	新闻视频采集
V、BM、A	AFEW 数据库[73]	7 类情感标签	截取自 54 部好莱坞电影，含各种头部姿势、遮挡及不同照明共 1809 段视频剪辑	自然环境下录制
	MED 数据集[74]	7 类情感标签	选取自电影、电视剧、直播视频等，共 1839 段视频剪辑，719 名受试者	自然环境下录制

Table 5 Multimodal emotion recognition

表 5　多模态情感识别

模态	表情模态识别方法	其他模态识别方法	融合方式	优点	缺点
表情视频+音频	局部二值模式提取视频特征，用随机森林模型进行面部情绪识别[75]	采用关联的特征和主成分分析法进行音频特征降维，应用连续混合高斯分布的隐马尔科夫模型进行音频识别	基于情绪基调对音视频识别结果进行修正，在不同情绪基调下运用线性相关性分析，进行决策层	针对单模态间情绪识别结果不一致时，融合后的识别结果不准确的问题进行改进，识别结果的准	仅在 SEMAINE 数据库进行验证

			融合	确率有一定的提升	
	采用 VGGNet-19 网络进行面部表情特征提取[76]	基于先验知识对音频进行特征提取	采用特征直接级联结合 PCA 降维，并用双向长短时间记忆网络（Long Short-Term Memory，LSTM）建模	在 AViD-Corpus 及 SEMAINE 数据库进行验证，结果有一定改善	结果改善不多
	基于深度学习算法和 Gabor 变换相结合的面部表情连续情感识别[77]	使用梅尔频率倒谱系数提取语音情感特征，利用迁移学习，使用预训练后的神经网络模型对语音情感状态进行学习	考虑模态间的互补性，分析比较了多元线性回归及卡尔曼滤波两种决策层融合算法	与单模态比，提升识别准确性	无法解决伪表情的识别问题
表情视频+脑电信号	使用直方图均衡化进行预处理，再用均匀模式局部二值模式算法提取人脸表情特征，并用于 JAFFE 数据库进行验证[78]	采用小波阈值去噪进行脑电信号预处理，用分型维数和多尺度熵算法提取脑电信号特征，用支持向量机对 DEAP 数据库中的脑电信号进行情绪分类	通过支持向量机分类器进行情绪分类	两个模态分别验证	没有在脑电和人脸表情的双模态情绪数据库上进行情绪识别分类。
	基于双线性卷积网络（Bilinear Convolution network，BCN）提取面部表情特征[79]	将脑电信号转换为三组频带图像序列，利用 BCN 融合图像特征，得到表情脑电信号的多模态情感特征	设计了一种三层双向 LSTM 结构的特征融合网络，融合表情和脑电特征	有助于提高情感识别的准确性	仅在 matlab 上仿真，暂无法在实际环境中应用。
图像+眼动、音频、脑电等	采用深度卷积生成对抗网络进行数据增强，采用 Faceness-net 算法检测人脸，使用提取增强特征图的深度卷积层组成的注意感知动作单元和全连接层单元的双增强胶囊网络，利用压缩函数对人脸表情进行识别[80]	采用多模态传感器采集数据如眼动跟踪器、音频、脑电图、深度相机等，并集成到面部表情识别中。	特征融合	取得较好的结果	模型复杂，传感器采集的数据庞大，冗余信息多

2.3 多模态数据的融合方式

在基于多模态数据的表情识别中，除了各个模态的特征识别外，模态融合也是十分重要的。因此选择合适的模态融合方式可以提高识别的准确性及稳定性[81]，融合是从不同模态中提取信息集成多模态特征[82]。常见的融合方式有：特征级的融合、决策级的融合和混合融合等[83]。

2.3.1 特征级的融合

特征级的融合属于中间层级的融合[84]，通常需要从原始信息中提取有效的特征，然后对这些特征信息进行分析和处理[84]。特征级的融合对信息压缩有利，提取的特征与决策分析直接相关，因此，特征级的融合结果能为决策分析提供所需的特征信息[84]，但是当不考虑模态间的关联性，直接将各模态的特征进行级联时，且当过多模态融合时，其产生的特征向量可能产生维度灾难。其融合框图如图 4 所示：

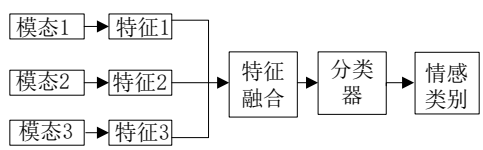

Fig.4 Fusion block diagram of feature level

图 4　特征级的融合框图

多模态情感识别方法中，研究者大量使用基

于特征级的融合方法，但大多研究是将不同模态的特征直接级联，鲜少考虑模态间的信息互补关联。文献[85]利用开源软件 OpenEAR、计算机表情识别工具箱进行语音和面部的情感特征的提取，删除视频中出现频率低的单词，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，并使用特征级融合法将三种特征融合，利用支持向量机分析得到情感极性。具体实现过程如下图 5[85]所示。

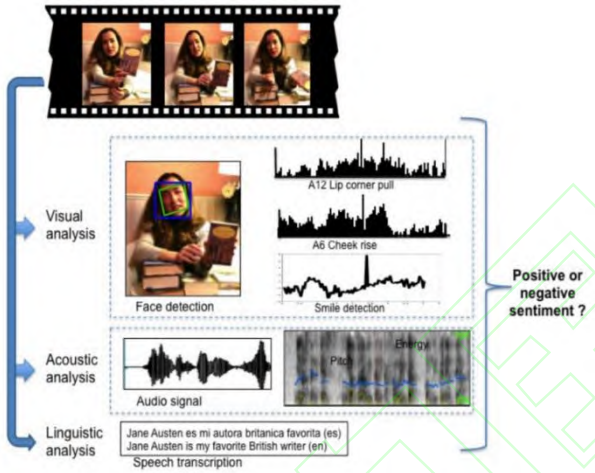

Fig.5 Multimodal feature extraction

图 5　多模态特征提取

文献[86]中通过挖掘话语前后视频页面的关系，提出了基于 LSTM 的情感分析模型[86]，进行特征提取时，先用 text-CNN、3D-CNN 和 openSMILE 分别对单模态文本、图像、语言数据进行特征提取，这提取的是上下文无关的特征向量，然后将这些特征喂入 LSTM 网络捕捉上下文之间的关系，最后进行特征融合得到判断的结果。具体实现过程如图 6[86]所示：其中：Contextual LSTM 的实现过程是，首先将数据输入到 LSTM 中，得到了一个上下文有关的特征，然后再经过全连接层得到一个预测结果，再进行一个 Softmax 得到预测概率。具体实现过程如图 7[86]所示：

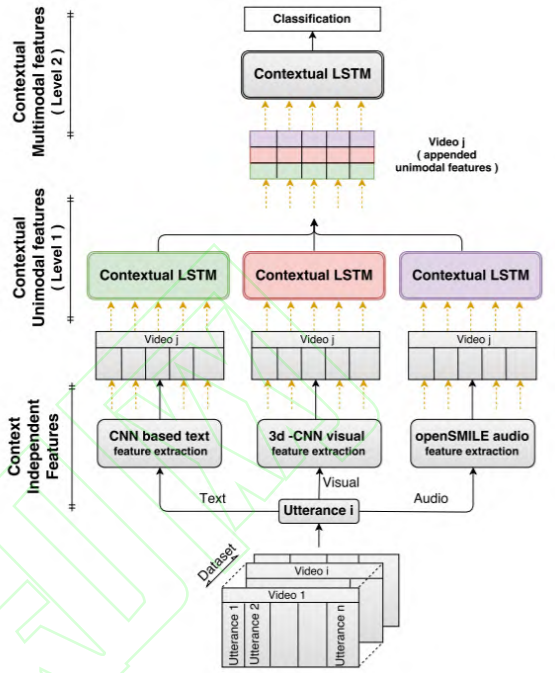

Fig.6 Hierarchical architecture for extraction context dependent multimodal utterance features

图 6　提取上下文相关多模态话语特征的层次结构

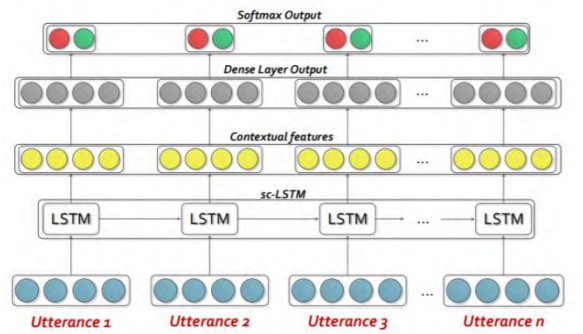

Fig.7 Contextual LSTM network

图 7　Contextual LSTM 网络

文献[87]提出了能识别面部表情、姿态、身体动作和声音的多模态情感识别框架，利用级联三维卷积神经网络以及深度置信网络得到新的深度时空特征，对视频和音频等呈现的时空信息进行有效建模实现情感识别，并且提出了一种基于双线性池理论的新的音视频特征级融合算法，在多模态情感数据集 eNTERFACE 以及 FABO 中，都取得不错的结果。

文献[88]提出了一种基于深度置信网络的多模情绪识别方法。如图 8[88]，首先对语音和表达式信号进行预处理和特征提取，获得单模信号的高级特征；然后，利用双模态深度置信网络融合高级语音特征和表达特征，得到用于分类的多模态融合特征，并去除模态之间的冗余信息，最后，利用 LIBSVM 软件对多模态融合特征进行分类，实现最终的情感识别。在多模态特征融合阶段，采用 3 个隐藏层的多模态融合深度置信网络（Deep confidence network，DBN）结构。在初始阶段，两个 DBN 网络分别训练。当训练到第三隐含层时，将第三层的两个特征值结合起来输入到后面反向传播（back propagation，BP）层[88]。在微调阶段，根据分类器的实际输出对第三隐藏层进行微调。从第三隐含层到两个 DBN 各自的隐含层，进行微调。最后，提出了一种基于 DBN 的多模态融合情感识别模型。DBN 训练后，确定其权重和偏差。对于训练样本和测试样本，输入 DBN，通过第三隐藏层提取的特征值为多模态融合后的特征值。然后进入 LIBSVM 分类器进行情感分类。但数据集采用的是《老友记》十季的视频片段，同一个人的脸部细节发生了变化，给表情识别带来了更多的困难。

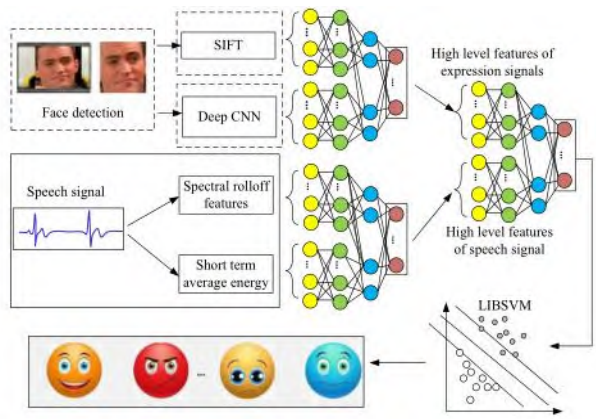

Fig.8 Overall architecture of multimodal emotion recognition model

图 8　多模态情感识别模型总体架构

2.3.2　决策级的融合

决策级的融合通常是指对单模态的信息进行逐个预处理及特征处理，然后经过分类器，得到各自的分类结果后，再将各自的分类结果，按照某种形式进行融合，得到最终的情感分类结果[89-91]。由于各个模态的分类结果的量纲等通常是一致的，决策级的融合相较于特征级融合更为简单，但是，决策级融合往往只是对单模态的情感识别结果进行二次加工，并没有对数据本身的特点进行充分挖掘，产生结果容易受到某一模态的情感识别效果的影响。决策级的融合框图如图 9 所示。

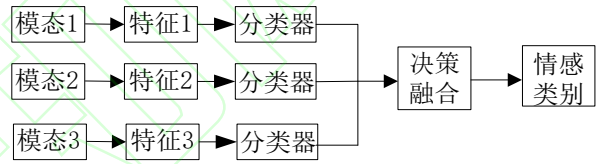

Fig.9 Fusion block diagram of decision level

图 9　决策级的融合框图

文献[89]利用了三个模态（视频、音频、文本）的组合特征向量来训练一个基于多核学习的分类器，同时提出了一种并行决策级数据融合方法，能更快得到结果[89]，但是准确率有待进一步提高。

文献[90]提出了一种融合面部表情以及血容量脉冲 BVP 生理信号的多模态情感识别法[90]。该方法一方面通过提取测试样本的面部视频特征来自局部二进制模式-3 维正交平面（local binary patterns from three orthogonal planes，LBPTOP）、梯度方向直方图-3 维正交平面（Gradient direction histogram - 3D orthogonal plane，HOG-TOP）两种时空表情特征后，送入 BP 分类器进行模型训练；另一方面，利用视频颜色放大技术获取血容量脉冲（blood volume pulse，BVP）信号，并提取生理信号情感特征，将特征送入 BP 分类器进行模型训练，最后将分类器得到的结果用模糊积分进行决策级融合，并得出识别结果。具体实现流程如图 10 所示，但是生理信号情感判别的准确率还是偏低。

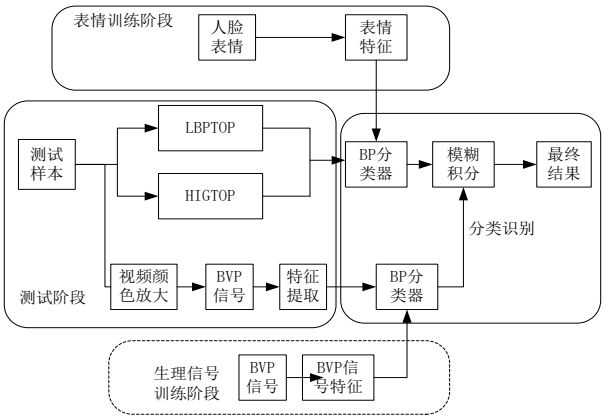

Fig.10 The flow chart of dual-modality emotion recognition

图 10　双模态情感识别系统流程图

2.3.3　混合融合方法

混合融合是指将特征级的融合和决策级的融合相结合，比如，某个分类器可以对面部模态和身体手势模态进行特征级的融合，另一个分类器对语音模态、生理信号模态进行特征级融合，这两个分类器上有另外的决策级分类器可以处理两个特征级分类器的结果，并最终得到情感标签[91]。混合融合的模型难度和复杂度比较高，能结合特征级的融合和决策级的融合的优点[9],混合融合框图如图 11 所示，但实用性较差。

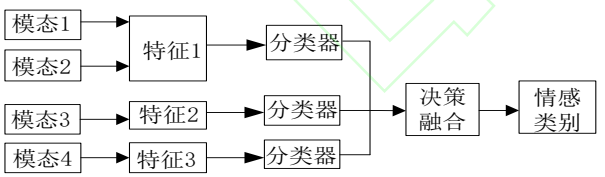

Fig.11 Hybrid fusion block diagram

图 11　混合融合框图

文献[91]引入了面部表情、皮肤电反应、脑电图等模态进行多模态识别与融合，采用基于混合融合的多模态情感分析，其中，采用 CNNF 模型进行训练面部表情信号,采用 CNN_V 模型和 CNN_A 模型训练 EEG 和皮肤电反应(galvanic skin response, GSR)信号，加权单元分别计算 CNN_V 模型和 CNN_A 模型输出的化合价和加权和，然后将结果送到距离计算器计算情感距离，并与 CNN_F 模型得到的面部识别结果一起送到决策树进行决策融合得到情感类别。文献[91]提出一种多模态情感识别的混合融合方法，采用潜在空间特征级融合方法，保持各模式之间的统计相关性，寻找共同的潜在空间来融合音频和视频信号，采用基于 Dempster-Shafer（DS）理论的证据融合方法来融合视听相关空间和文本模态，该方法解决了声像信息的冗余和冲突的问题，兼顾了特征级和决策级的融合，但存在 DS 融合方法的证据冲突问题。

3　总结与展望

随着计算机处理能力的不断提升，深度学习网络及融合算法的不断改进，基于多模态的数据的表情识别将得到快速的发展，本文通过总结基于单一模态数据的传统面部表情特征提取方法、基于单一模态的深度学习算法、基于多模态数据的表情识别与融合算法，将面临的挑战和发展趋势归纳如下：

（1）人脸图片的影响因素有很多，如角度旋转、遮挡、模糊、光线、分辨率、头部姿势、个体属性差别等，这些数据的处理技术不成熟，影响表情识别的进展。

（2）基于多模态的数据集偏少，大部分数据集大多是由视觉、文本、语音等模态的数据组成，姿势、脑电波及其他生理信号等模态的数据少。

（3）数据集中的数据分布不均衡，常见的高兴、伤心的表情多且容易识别，愤怒、蔑视等表情少且难捕获。

（4）现有的模态融合技术往往没有深入挖掘模态之间的相关性，以提高表情识别的准确性。

（5）算法大多十分复杂，在多模态数据分析过程中，如果选用的模态过多，则融合的算法就十分复杂，如果选太少，可能无法提高识别准确率。

（6）基于单模态数据的处理方法及各模态间的融合算法的选择是影响识别准确率的关键因素之一。各个步骤算法的选择都很重要。

针对上述观点，今后可以在如下几个方面做进一步的研究。

（1）构建更多自然环境下高质量的表情数据集或 3D 人脸表情数据集，进一步解决角度旋转、遮挡、光线、头部姿势及个体属性差异等复杂情况下的表情识别准确率不高的问题。如：加入智能传感器用于解决诸如照明变化、主体依赖和头部姿势等重大挑战。

（2）构建基于含姿势、脑电波及其他生理信号等模态的多模态数据集，并研究多模态之间的模态相关性，以提高模型的泛化能力。

（3）未来与来自三维人脸模型、神经科学、认知科学、红外图像和生理数据的深度信息相结合，可以成为一个很好的未来研究方向。

（4）改进现有的表情识别技术，利用 GAN 网络提高对表情数据增强，解决表情数据量不平衡的问题。

（5）如何确定自然欺骗性面部表情的正确情绪状态也是未来研究方面，随着微表情在心理学领域的发展，可将现有的技术应用于微表情的提取，制作微表情方面的数据集。

（6）改进模态融合时的权值问题，对不同环境下，给不同模态不同的权值分配也是模态融合重点研究方向之一。

（7）为了让机器更全面、更有效地感知周围的世界，需要赋予它理解、推理和融合多模态信息的能力，如：语音、图像、气味和生理信号等。利用多模态融合特征提高跨媒体分析的性能，如视频分类、事件检测、情感分析、跨模态翻译等也是研究方向之一。同时，多模态信息融合所产生的特征冗余、缺少关键特征等问题仍有待解决。

（8）最后，基于多模态数据和深度学习网络的表情识别技术需要大量的优质数据集及计算力，如何将复杂的基于多模态数据的算法部署在计算资源有限的机器人终端上，研究如何对神经网络进行剪枝及轻量化，也是未来的研究方向之一。

4 结语

本文对现有的面部表情识别领域的研究成果进行总结，归纳出基于单模态数据集和传统机器学习的表情识别技术，基于单模态数据集和深度学习的表情识别技术、以及基于多模态数据集表情识别技术及模态融合技术等领域的成果，概要地介绍了多模态数据库，最后对当前表情识别存在的问题与挑战进行总结和展望，指出后续表情识别的一些研究方向，如：非正面人脸表情识别、微表情、多模态情感分析，轻量级神经网络等。

参考文献

[1] 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.
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.

[2] PENG X J,QIAO Y. Advances and challenges in facial expression analysis[J].Journal of Image and Graphics, 2020, 25(11):2337-2348.
彭小江,乔宇.面部表情分析进展和挑战[J].中国图象图形学报,25(11):2337-2348.

[3] LI S, DENG W H. Deep facial expression recognition: a survey[J].Journal of Image and Graphics, 2020, 25(11): 2306-2320.
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.

[4] EKMAN P, ROSENBERG E L. What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS)[M].New York: Oxford University Press,1997.

[5] LIU Y, GUO YY, FANG J,et al. Survey of research on deep learning image-text cross-modal retrieval[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(3): 489-511.
刘颖,郭莹莹,房杰,等. 深度学习跨模态图文检索研究综述[J]. 计算机科学与探索, 2022, 16(3): 489-511.

[6] MA Y J,XU X D,ZHANG R,et al. Generative adversarial network and its research progress in image generation[J]. Journal of Frontiers of Computer Science and Technology, 2021,15(10):1795-1811.
马永杰,徐小冬,张茹,等.生成式对抗网络及其在图像生成中的研究进展[J].计算机科学与探索，2021，15(10): 1795-1811.

[7] LIU Y, ZHANG Y X, SHE J C, et al. Review of new face occlusion inpainting technology research[J].Journal of Frontiers of Computer Science and Technology, 2021, 15(10): 1773-1794.

刘颖,张艺轩,佘建初.人脸去遮挡新技术研究综述[J].计算机科学与探索,2021,15(10):1173-1794.

[8] MO H W, FU Z J. Unsupervised cross-domain expression recognition based on transfer learning[J]. Journal of Intelligent Systems, 2021, 16(3): 397–406.
莫宏伟,傅智杰.基于迁移学习的无监督跨域人脸表情识别[J].智能系统学报, 2021, 16(3): 397–406.

[9] LIU J M, ZHANG P S, LIU Y, et al. Summary of multi-modal sentiment analysis technology[J]. Journal of Frontiers of Computer Science and Technology, 2021,15(7):1165-1182.
刘继明,张培翔,刘颖,等.多模态的情感分析技术综述[J].计算机科学与探索,2021,15(7):1165-1182.

[10] SUSSKIND J M, ANDERSON A K, HINTON G E. The toronto face database[D].Department of Computer Science, University of Toronto, Toronto, ON, Canada, Tech. Rep, vol. 3, 2010.

[11] GOODFELLOW I J, Erhan D, Carrier P L, et al. Challenges in representation learning: A report on three machine learning contests[C]//20thInternational Conference on Neural Information Processing, Daegu, Korea, 3-7 November, 2013. UK, Neural Networks, 2015, 64: 59-63.

[12] WANG S F, LIU Z L, LV S L, et al. A natural visible and infrared facial expression database for expression recognition and emotion inference[J]. IEEE Transactions on Multimedia, 2010, 12(7): 682-691.

[13] DHALL A, GOECKE R, LUCEY S, et al. Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark[C]//Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference, Barcelona, Spain, 6-13 November 2011. Piscataway, IEEE, 2011: 2106-2112.

[14] GROSS R, MATTHEWS I, COHN J, et al. Multipie[J]. Image and Vision Computing,2010.28(5): 807-813.

[15] YIN L, WEI X, SUN Y, et al. A 3d facial expression database for facial behavior research[C]//Automatic face and gesture recognition, FGR 2006 7th international conference, Southampton, UK,10-12 April 2006. Piscataway, IEEE, 2006 :211-216.

[16] ZHAO G, HUANG X, TAINI M, et al. Facial expression recognition from near-infrared videos[J]. Image and Vision Computing, 2011, 2(9): 607-619.

[17] LANGNER O, DOTSCH R, BIJLSTRA G, et al. Presentation and validation of the radboud faces database[J]. Cognition and Emotion, 2010, 24(8): 1377-1388.

[18] LUNDQVIST D, FLYKT A, OHMAN A.The karolinska directed emotional faces (kdef)[M/CD].CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, Sweden,1998.

[19] BENITEZ-QUIROZ C F,SRINIVASAN R, MARTINEZ A M. Emotionet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild[C]//Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR), Las Vegas, NV, USA, June 26th-July 1st, 2016. Piscataway, IEEE, 2016 :5562-5570.

[20] LI S,DENG W, DU J. Reliable crowdsourcing and deep locality preserving learning for expression recognition in the wild[C]//in IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Venice, Italy, 21-16 July,2017. Piscataway, IEEE, 2017, 2584-2593.

[21] LI S,DENG W. Reliable crowdsourcing and deep locality preserving learning for unconstrained facial expression recognition[J]. IEEE Transactions on Image Processing, 2018.

[22] MOLLAHOSSEINI A, HASANI B, MAHOOR M H, Affectnet: A database for facial expression, valence, and arousal computing in the wild[J]. IEEE Transactions on Affective Computing,2017, 99: 1-1.

[23] ZHANG Z, LUO P,CHEN C L, et al..From facial expression recognition to interpersonal relation prediction[J]. International Journal of Computer Vision, 2018.126(5): 1-20.

[24] LUCEY P, COHN J F, KANADE T, et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression[C]// Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. San Francisco, 2010. Los Alamitos, IEEE: 94-101.

[25]PANTIC M,VALSTAR M ,RADEMAKER R, et al.. Web-based database for facial expression analysis[C]// in Multimedia and Expo, Amsterdam, the Netherlands 06-06 July 2005.Piscataway, IEEE, 2005:5-9.

[26]VALSTAR M, PANTIC M,Induced disgust, happiness and surprise: an addition to the mmi facial expression database[C]//in Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, 2010, 65-70.

[27] LYONS M, AKAMATSU S, KAMACHI M,et al..Coding facial expressions with gabor wavelets[C]// Third IEEE International Conference onAutomatic Face and Gesture Recognition,Nara Japan, April 14-16 1998. Los Alamitos,IEEE Computer Society ,1998:200-205.

[28] LI S,DENG W H. Deep facial expression recognition: a survey[J]. IEEE Transactions on Affective Computing, 2018:1-1.

[29] WOLD S, ESBENSEN K, GELADI P. Principal component analysis[J]. Chemometrics and Intelligent Laboratory Systems,1987, 2(1/2/3): 37-52.

[30] NIU Z G, QIU X H. Facial expression recognition based on weighted principal component analysis and support

vector machines[C]//2010 3rd International Conference on Advanced Computer Theroy and Engineering (ICACTE), Chengdu, China, 20-22 August, 2010. Piscataway, IEEE, 2010:174-178.

[31] ZHU Y N, LI X X, WU G H. Face expression recognition based on equable principal component analysis and linear regression classification[C]//2016 3rd International Conference on Systems and Informatics. Shanghai, China, 19-21 Nov. 2016, Piscataway, IEEE, 2017:876-880.

[32] ZHOU S R, LIANG X M, ZHU C, et al.. Facial expression recognition based on independent component analysis and hidden Markov model[J]. Journal of Image and Graphics, 2008(12):76-83.
周书仁,梁昔明,朱灿,等.基于 ICA 与 HMM 的表情识别[J]. 中国图象图形学报, 2008(12): 76-83.

[33] OJALA T, PIETIKAINEN M, MAENPAA T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987.

[34] LIAO S C, ZHU X X, LEI Z, et al. Learning multi-scale block local binary patterns for face recognition[C]// Proceedings of the 2007 International Conference on Advances in Biometrics. Seoul, Korea, 27–29 August 2007, DBLP, 2007: 823-827.

[35] KABIR H, JABID T, CHAE O. A local directional pattern variance (LDPV) based face descriptor for human facial expression recognition[C]// Proceedings of 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, Boston, MA, USA, 29 Aug.-1 Sept. 2010.Piscataway,IEEE, 2010:526-532.

[36] JABID T, KABIR M H, CHAE O. Robust facial expression recognition based on local directional pattern[J]. ETRI Journal,2010, 32(5): 784-794.

[37] LI H, LI G M. Research on facial expression recognition based on LBP and deep learning[C]//Proceedings of 2019 International Conference on Robots & Intelligent System (ICRIS).Haikou, China,15-16 June 2019. Piscataway, IEEE, 2019: 94-97.

[38] NAGARAJA S, PRABHAKAR C J, KUMAR P P. Complete local binary pattern for representation of facial expression based on curvelet transform[C]// Proceedings of International Conference on Multimedia Processing, Communication and Information Technology (MPCIT). [S.l.]: Association of Computer Electronics and Electrical Engineers, 2013: 48-56.

[39] WU X L, QIU Q C, LIU Z, et al.. Hyphae detection in fungal keratitis images with adaptive robust binary pattern[J].IEEE Access, 2018, 6:13449-13460.

[40] RUBEL A S, CHOWDHURY A A, KABIR M H. Facial expression recognition using adaptive robust local complete pattern[C]//Proceedings of 2019 IEEE International Conference on Image Processing (ICIP). Taipei, China, 22-25 Sept. 2019.Piscataway, IEEE, 2019: 41-45.

[41] ZHANG Z Y, MU X M, GAO L. Recognizing facial expressions based on Gabor filter selection[C]// Proceedings of International Congress on Image and Signal Processing. Shanghai, China, 15-17 Oct. 2011, Piscataway, IEEE, 2011: 1544 -1548.

[42] BASHYAL S, VENAYAGAMOORTHY G K. Recognition of facial expressions using Gabor wavelets and learning vector quantization[J]. Engineering Applications of Artificial Intelligence, 2008, 21(7): 1056-1064.

[43] ABBOUD B, DAVOINE F, DANG M. Expressive face recognition and synthesis[C]// 2003 Conference on Computer Vision and Pattern Recognition Workshop. Madison, WI, USA, 16-22 June 2003. Los Alamitos, IEEE Computer Society, 2003:1-6.

[44] DENG H B, JIN L W. Facial expression recognition based on local Gabor filter bank and PCA+LDA[J]. Journal of Image and Graphics, 2007, 12(2): 322-329.
邓洪波,金连文.一种基于局部 Gabor 滤波器组及 PCA + LDA 的人脸表情识别方法[J]. 中国图象图形学报,2007, 12(2):322-329.

[45] YAO W, SUN Z, ZHANG Y. Optimal Gabor feature for facial expression recognition[J]. Computer Graphics, 2008,22(1): 79-84.
姚伟,孙正,张岩.面向脸部表情识别的 Gabor 特征选择方法[J]. 计算机辅助设计与图形学学报, 2008, 22(1): 79-84.

[46] PANTIC M, BARTLETT M S. Face recognition[M]. [S.l.]:I-Tech Education and Publishing, 2007.

[47] COOTES T F, TAYLOR C J, COOPER D H, et al..Active shape models-their training and application [J].Computer Vision and Image Understanding, 1995, 61(1): 38-59.

[48] MATTHEWS I, BAKER S. Active appearance models revisited[J]. International Journal of Computer Vision, 2004, 60(2):135-164.

[49] BARMAN A, DUTTA P. Facial expression recognition using distance and shape signature features[J]. Pattern Recognition Letters, 2017:1-11.

[50] CHAO W L, DING J J, LIU J Z. Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection[J]. Signal Processing, 2015,117:1-10.

[51] AGARWAL S, MUKHERJEE D P. Facial expression recognition through adaptive learning of local motion descriptor[J]. Multimedia Tools and Applications, 2017, 76(1):1073-1099.

[52] WANG X H, JIN C, LIU W, et al.. Feature fusion of HOG and WLD for facial expression recognition[C]// Proceedings of the 2013 IEEE/SICE International Symposium on System Integration (SII). Kobe, Japan, 15-17 Dec. 2013. Piscataway: IEEE, 2013:227-232.

[53] LIU S S, TIAN Y T. Facial expression recognition method based on Gabor wavelet features and fractional power polynomial kernel PCA[C]//Proceedings of International Symposium on Neural Networks. Berlin, Heidelberg: Shanghai, China, June 6-9, 2010. Cham, Switzerland: Springer, 2010: 144-151.

[54] SHAO H, WANG Y, WANG Y J.Dynamic sequence [J].expression recognition based on AAM and optical flow method[J]. Computer engineering and design, 2017,38(6): 1642-1646+1656.
邵虹,王洋,王昳昀.基于 AAM 和光流法的动态序列表情识别[J].计算机工程与设计，2017，38(6)：1642-1646+1656.

[55] MOLLAHOSSEINI A, CHAN D, MAHOOR M H. Going deeper in facial expression recognition using deep neural networks[C]//Proceedings of 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Placid, NY, USA, 3-10 March, 2016. Piscataway: IEEE, 2016: 1-10.

[56] DING H, ZHOU S K, CHELLAPPA R. Facenet2expnet: regularizing a deep face recognition net for expression recognition[C]//Proceedings of the 12th IEEE International Conference on Automatic Face and Gesture Recognition. Washington, 29 June 2017. Piscataway: IEEE, 2017.118-126.

[57] NG H W, NGUYEN V D, VONIKAKIS V, et al.. Deep learning for emotion recognition on small datasets using transfer learning[C]//Proceedings of the 17th ACM International Conference on Multimodal Interaction. Seattle, USA. 9-13 November, 2015. NewYork: ACM, 2015: 443-449.

[58] VERMA M, KOBORI H, NAKASHIMA Y, et al.. Facial expression recognition with skip-connection to leverage low-level features[C]//Proceedings of 2019 IEEE International Conference on Image Processing (ICIP). Taipei, China, 22-25 Sep,2019.Piscataway: IEEE, 2019: 51-55.

[59] LIU Y Y, DAI W, FANG F, et al. Dynamic multi-channel metric network for joint pose-aware and identity-invariant facial expression recognition[J]. Information Sciences ,2021(578):195-213.

[60] YANG H Y,CIFTCI U,YIN L J.Facial expression recognition by de-expression residue learning[C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, USA,18-25 June 2018. Piscataway: IEEE, 2018: 2168-2177.

[61] ZHANG W J, SONG P, CHEN D L, et al.. Cross-domain facial expression recognition based on sparse subspace transfer learning[J]. Journal of Data Acquisiton and Processing.2021,36(1):113-121.
张雯婧,宋鹏,陈栋梁,等.基于稀疏子空间迁移学习的跨域人脸表情识别[J].数据采集与处理,2021，36(1)：113-121.

[62] 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.

[63] 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, 6-10 July ,2020. Stroudsburg: ACL, 2020: 3718-3727.

[64] 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.

[65] LI Y,TAO J H,CHAO L L, et al.. CHEAVD: a Chinese natural emotional audio-visual database[J]. Journal of Ambient Intelligence and Humanized Computing, 2017, 8(6):913-924.

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

[67] WU M , SU W J, CHEN L F, et al.. Two-stage fuzzy fusion based-convolution neural network for dynamic emotion recognition [J]. IEEE Transactions on Affective Computing,2020.

[68] MARTIN O, KOTSIA I, MACQ B, et al. The eNTERFACE' 05 audio-visual emotion database[C]// 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA, 2006:8-8.

[69] WOLLMER 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.

[70] 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.

[71] 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]// meeting of the association for computational linguistics, Melbourne, Australia, 15-20 July. 2018. Stroudsburg: ACL, 2236-

2246.
[72] 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.
[73] DHALL A,GOECKE R,LUCEY S,et al..Collecting large,richly annotated facial-expression databases from movies[J].IEEE Multi Media,2012,19(3):34-41.
[74] CHEN J,WANG K J,ZHAO C,et al..MED: multimodal emotion dataset in the wild[J].Journal of Image and Graphics,25(11):2349-2360.
陈静,王科俊,赵聪,等.真实环境下的多模态情感数据集 MED[J].中国图象图形学报，2020，25(11)：2349-2360.
[75] WEI F G, ZHANG S D, FU X H. audio-visual bimodal emotion recognition based on emotional tone[J]. Computer Applications and Software, 2018, 35(8): 238-242.
卫飞高,张树东,付晓慧.基于情绪基调的音视频双模态情绪识别算法[J].计算机应用与软件，2018，35(8)：238-242.
[76] SONG G J, ZHANG S D, WEI F G. Research on audio-visual dual-modal emotion recognition fusion frame-work[J]. Computer Engineering and Applications, 2020, 56(6):140-146.
宋冠军,张树东,卫飞高.音视频双模态情感识别融合框架研究[J].计算机工程与应用,2020,56(6):140-146.
[77]ZHANG L. Multimodal emotion recognition based on face and speech and the application in reasoning of robot service tasks[D].ShanDong University,2021.
张龙.基于表情和语音的多模态情感识别及其在机器人服务任务推理中的应用[D].山东大学,2021.
[78] SHEN J. Bimodal emotion recognition system based on EEG and facial expression[D]. Nanjing: Nanjing University of Posts and telecommunications,2020.
沈健.基于脑电和人脸表情的双模态情绪识别系统[D].南京:南京邮电大学,2020.
[79] ZHAO Y F, CHEN D Y. Expression EEG multimodal emotion recognition method based on the bidirectional LSTM and attention mechanism[J]. Computational and Mathematical Methods in Medicine, 2021 (2021): 9967592, 1-12.
[80] ULLAH A, WANG J, ANWAR M S , et al.. Empirical investigation of multimodal sensors in novel deep facial expression recognition in-the-wild[J]. Journal of Sensors, 2021(2021): 8893661,1-13.
[81] HE J,ZHANG C Q,LI X Z,et al..Survey of research on multimodal fusion technology for deep learning[J]. Computer Engineering,2020,46(5):1-11.
何俊,张彩庆,李小珍,张德海.面向深度学习的多模态融合技术研究综述[J].计算机工程,2020,46(5):1-11.
[82]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.
[83]NEMATI S, ROHANI R, BASIRI M E,et al.. A hybrid latent space data fusion method for multimodal emotion recognition[J]. IEEE Access,2019(7): 172948 -172964.
[84] CHEN W Q.Research of multi-modal emotion recognition based on deep learning[D]. JiNan: Shandong University,2020.
陈炜青,基于深度学习的多模态情感识别研究[D].济南:山东大学,2020.
[85] PEREZ-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.
[86] 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.
[87] NGUYEN D, NGUYEN K, SRIDHARAN S. Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition[J]. Computer Vision and Image Understanding, 2018, 174(15): 33-42.
[88] LIU D,CHEN L X,WANG Z Y, et al.. Speech expression multimodal emotion recognition based on deep belief network[J] Grid Computing ,2021,19: 22.
[89]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.
[90]REN F J, YU M L, HU M,et al.. Dual-modality video emotion recognition based on facial expression and BVP physiological signal[J].Journal of Image and Graphics, 2018,23(5): 0688-0697.
任福继,于曼丽,胡敏,等.融合表情和 BVP 生理信号的双模态视频情感识别[J].中国图象图形学报，2018，23(5):0688-0697.
[91]YUCEL C, ERHAN E, SEYMA C O. Cross-subject Multimodal Emotion Recognition Based on Hybrid Fusion[J].IEEE ACCESS,2020, (8):168865-168878.

洪惠群（1984-），女，福建南安人，硕士，讲师、工程师，从事图像处理、计算机视觉、表情识别等研究，E-mail：honghuiqun@qq.com，**CCF 会员**。

HONG Huiqun, born in 1984, engineer, lecturer, Her research interests include image processing, computer vision, expression recognition.etc.

沈贵萍（1990-），女，博士，副教授，主要从事模式识别与智能系统、表情识别、图像恢复、视频表达、音频处理等研究。E-mail:992774639@qq.com

SHEN Guiping, born in 1990, Ph.D., associate professor, Her research interests include pattern recognition and intelligent system, expression recognition, image restoration, video expression, audio processing, etc.

黄风华（1982-），男，福建莆田人，博士，教授，硕士生导师，主要从事机器学习、数据挖掘、遥感等，Email：fenghuait@sina.com，**CCF 会员。**

HUANG Fenghua, born in 1982,Ph.D., professor. His research interests include machine learning, data mining, remote sensing.etc.