多模态情感识别研究进展*

何 俊, 刘 跃, 何忠文
(南昌大学 信息工程学院, 南昌 330031)

摘 要: 针对多模态情感特征提取与融合的技术难点,列举了目前应用较广的多模态情感识别数据库,介绍了面部表情和语音情感这两个模态的特征提取技术,重点阐述了多模态情感融合识别技术,主要对多模态情感特征融合策略和融合方法进行了综述,对不同算法下的识别效果进行了对比。最后,对多模态情感识别研究中存在的问题进行了探讨,并对未来的研究方向进行了展望,旨在为研究此方向建立系统的知识体系,借此推动与此相关问题的进展。

关键词: 情感识别; 特征提取; 多模态融合
中图分类号: TP391.41 **文献标志码**: A **文章编号**: 1001-3695(2018)11-3201-05
doi:10.3969/j.issn.1001-3695.2018.11.001

Research progress of multimodal emotion recognition

He Jun, Liu Yue, He Zhongwen
(*College of Information & Engineering*, *Nanchang University*, *Nanchang* 330031, *China*)

Abstract: For the difficulties of multimodal emotion feature extraction and fusion, this paper listed some widely-used datasets for multimodal emotion recognition, and the methods of feature extraction based on facial expression features and speech features. It mainly introduced the multimodal fusion technology, and stated some levels or types of fusion and methods of fusion and compared them. Finally, it explored potential issues and future research directions. The purpose is to develop a systematic knowledge system for readers to study multimodal emotion recognition and promotes the progress of the related issues.

Key words: emotion recognition; feature extraction; multimodal fusion

互联网的飞速发展导致人们从人际互动转向更多的人机互动,并对情感交互技术提出了更高的要求,人们期望与之交互的机器也具有类似于人的观察、理解和生成情感特征的能力[1]。目前研究大多集中在人脸表情识别、语音识别、肢体识别等单模态领域,但单模态信息量不足且容易受到外界各种因素的影响,如面部表情容易被遮挡、语音容易受噪声干扰。鉴于各个模态之间的互补性,多模态融合的情感识别研究正日益受到重视,研究热点已经从单模态转移到实际应用场合下的多模态情感识别[2]。在文献[1]中 D'Mello 等人采用统计学的方法在不同的数据库上比较了单模态和多模态的准确性,实验中多模态表情识别要优于单一模态性能。McGurk 现象揭示了大脑在进行感知时,不同感官会被无意识地自动结合在一起对信息进行处理,任何感官信息的缺乏或不准确都将导致大脑对外界信息的理解产生偏差[3]。因此,多模态特征融合识别技术成为这两年的研究热点[4,5],吸引了包括 Google、微软、百度、华为、斯坦福大学、卡耐基梅隆大学(CMU)等知名公司和研究机构的参与,索尼公司推出了具有情感交互功能的 AIBO 机器狗,能与人进行全方位的语音、表情交流;卡耐基梅隆大学研制出一种机器人接待员 Valerie,有自己的性格和爱好,可与访问者互动;美国一些商家已经开始使用人工智能技术判断顾客在网络购物时是否开心或满意。目前,国内也在该领域取得了一定的成果,如中国科学院计算所研究的带有表情和动作的虚拟人、浙江大学的虚拟人物和情绪系统、东南大学的语音情感识别系统、中国科技大学的基于内容的交互式感性图像检索系统等。该技术的发展在教育、抑郁症治疗、临床预后评估、智能客服、网络购物等领域都有广阔的应用前景,国家"十二五规划"将智能感知与人机交互技术列为信息领域的前沿技术。

情感主要包括面部表情、语音情感、文本、肢体运动等多个模态。鉴于大多数多模态情感识别研究集中于面部表情和语音两个模态的融合识别,本文重点介绍两者的特征提取技术与融合策略,并将从多模态情感数据库、情感特征提取技术、多模态融合识别技术三个方面进行综述。

1 多模态情感识别数据库

目前国内外多模态情感数据库大多来源于网络或人为制作,对于科研领域仍是半公开或是不公开,研究人员大多使用自己建立的数据库。本章主要介绍应用较广的多模态情感数据库,这些数据库可分为自发和人为诱导两种情感类型。

1)HUMAINE 数据库 该数据库由 Douglas-Cowie 等人[6]建立,提供多模态的情感自然视频片段,它包含 50 个自然和诱导的视频片段及相关的标签。这些数据包含了不同性别和文化人的姿势、脸部、语言等线索。

2)Belfast 数据库 该数据库由 Douglas-Cowie 等人[7]在 2000 年建立,它来源于电视访谈节目和宗教节目,包含了 125 位测试者(由 31 名男性和 94 名女性组成)和 298 个视频片段,每个测试者都包含 1 个中立情感和 1 个强烈情感片段(高兴、气愤、厌恶、恐惧、伤心、惊讶等其中之一)。

3)SEMAINE 数据库 该数据库是由 McKeown 等人[8]建立的大型视听数据库,总共有 150 位参加者 959 段对话,每段对话 5 min。SEMAINE 数据库是一个面向自然人机交互和人工智能研究的数据库,数据录制在人机交互的场景下进行,20

收稿日期: 2017-09-01; **修回日期**: 2017-11-13 **基金项目**: 国家自然科学基金资助项目(61463034)

作者简介:何俊(1969-),男,江西东乡人,教授,硕导,博士,主要研究方向为数据挖掘、人机交互技术、模式识别等(boxhejun@tom.com);刘跃(1992-),男,安徽亳州人,硕士研究生,主要研究方向为深度学习、模式识别、人机交互技术;何忠文(1986-),男,江西九江人,硕士,主要研究方向为数据挖掘、人机交互技术、模式识别等.

位用户(22 ~ 60 岁,8 男 12 女)被要求与性格迥异的四个机器角色进行交谈。每个片段的标注工作由 6 ~ 8 位参与者在 valence、activation、power、expectation 和 intensity 这五个情感维度上进行。

4)IEMOCAP(interactive emotional dyadic motion capture)数据库 它是由 Busso 等人[9]录制的情感数据库,包含约 12 h 的视听数据(视频、音频、语音文本、面部表情等)。10 名专业演员(5 男 5 女)在根据台词或即兴的场景下,特意引导出情感表达,之后人工将每一段对话切分成单句,每句至少包含开心、伤心、生气、中性等情感中的一种,每一句话至少由 3 位标注员进行类别标注。

5)eNTERFACE 数据库 它是由 Martin 等人[10]制作的视听数据库,总共有 42 位测试者,分别来自 14 个不同的国家。该数据库作为测试和评估视频、音频或者视听情感识别算法的参考数据库。

2 多模态情感特征提取

不同人物之间表达情感的方式不同,如果一个人趋向于用语言表达情感,那么音频数据则可能包含较多的情感线索;如果趋向于用表情来表达其情感状态,则更多的情感线索集中在脸部表情特征上[9]。因此,特征提取是多模态情感识别的重要步骤,视频中获取的语音、表情等信息直接影响情感识别的准确程度。由于情感的表达大多表现在面部表情和声音上,本章将介绍关于面部表情和语音特征的提取技术。

2.1 面部表情特征提取

面部表情是理解情感的主要线索,脸部的每个部分都是进行识别的重要信息[11]。Ekman 等人[12]将人脸划分为 44 个相互独立又相互联系的运动单元(action units),在此基础上提出了面部表情编码系统(facial action coding system,FACS)。结合心理学家 Ekman 将人类基本情感分为高兴(happy)、惊讶(surprise)、愤怒(anger)、恐惧(fear)、悲伤(sad)、厌恶(disgut)六种,以此奠定了当今表情识别的基础。本文把面部表情特征提取方法归纳为二维图像法、三维模型法、视频法三类。

2.1.1 二维图像法

二维图像法可分为三种,即基于脸部各器官和脸部凸起的位置特征的几何法、基于脸部纹理特征的像素法以及两者结合的混合法。

典型的几何法包括主动形状模型(ASM)[13]和点分布模型(PDM)[14]。几何法的优点是特征语义清晰、实时性好;缺点是几何特征只代表了表情特征的一个子集,以至于识别率不高。

像素法包括 Gabor 小波[15]、局部二进制模式(LBP)、尺度不变特征变换(SIFT)、主成分分析(principal component analysis,PCA)和线性判别分析(LDA)、光流法(optical flow models)[16]、边缘检测等方法。像素法的优点是特征信息完整;缺点是特征维数高且受姿态的影响。为了降低这一缺点的影响,研究者们运用卡尔曼滤波和 PCA 对特征进行降维处理以提高识别准确性[17,18]。

混合法将前两种方法集合起来,以求最大程度地利用人脸表情特征来提高识别率,但时间代价高且缺乏实时性。基于随机梯度下降的多维形变模型[19]和主动外观模型(active appearance model,AAM)[20]充分利用了脸部形状和纹理的特征进行表情识别。

2.1.2 三维模型法

三维模型法主要是为了解决人脸姿态对识别率的影响,但其建模较为困难且时间代价较高。文献[21]提出了多状态的脸部成分模型,用于提取持续性和暂时性特征。3D 局部约束模型(constrained local model,CLM-Z)[22]是一种非刚性脸部追踪模型,它可以在不同的姿势下利用深度和灰度信息追踪脸部特征。

2.1.3 视频法

视频法是将时间信息、动作单元和以时长、内容、效价为单位的特征用于视频中的情感识别[23,24]。基于视频的方法的关键是在视频序列中保持准确追踪。肌肉模型[25]、3D 线框模型[26]、几何形状模型[27,28]已经用于视频中脸部特征的提取。目前形变模型在脸部追踪方面取得了较好的效果[29]。

此外,随着深度学习的发展,其分类效果比现有分类算法(如光流法、ASM、AAM)要优秀得多。鉴于深度学习在监督和非监督模式下提取鲁棒性特征的潜力,大量用于情感特征提取的深度学习算法被提出[30~33],该算法的优势在于不需要视觉情感识别领域的知识也可以很好地进行情感识别。文献[34]提出了基于时空特征的深度 3D 卷积网络(deep 3D convolutional network,C3D),该结构利用最简单分类器,其分类效果要优于目前最新水平的算法。为了处理在情感识别中越来越多的大数据和噪声,文献[35]提出了一种基于卷积神经网络(convolutional neutral network,CNN)的改进策略,该算法通过微调深度神经网络对有噪声的训练数据滤波和使用域迁移学习来提高性能。复杂的深度学习网络如 3DCNN、递归神经网络(recurrent neutral network,RNN)已经被应用于解决时间问题。Poria 等人在文献[36]中提出将 CNN 与 RNN 堆叠和训练提取视觉特征。

2.2 语音情感特征提取

当前,语音情感识别的声学特征可归纳为韵律特征[37,38]、基于谱的相关特征和音质特征等。早期的调查主要集中于口语中语音和声学的音频特征提取,这些特征大多以帧为单位进行提取,一般以独立的语句或单词形式参与感情识别[39,40]。韵律学特征又被称为超音段特征,其中最为常用的韵律特征有时长、基频、能量等。文献[41]建立了一种全局控制 Elman 模型,有效融合全局统计特征和语段时序特征,使系统识别性能达到最优的最佳识别段长。语音中的情感内容对频谱能量在各个频谱区间的分布有着明显的影响,其区分能力已得到语音情感识别领域的广泛认可[42]。其他语音特征也被用于特征提取[43,44],包括共振峰、梅尔频率倒谱系数(MFCC)、Teager 能量算子、对数频率功率系数(LFPC)、线性预测倒谱系数(LPCC)。Lugger 等人[45]在研究中提取共振峰特征作为音质的特征用于语音情感识别。广泛应用于语音情感识别的算法有 GMM、HMM、SVM(support vector machine)、长短期记忆网络(long short-term memory,LSTM)[46]和 bi-directional LSTM[47]等。深度学习算法在语音情感识别中不断得到关注[48~50],文献[49,50]利用 CNN 从音频中提取特征,并将提取的特征送入分类器进行情感识别。

3 多模态特征融合

多模态融合是在进行分析和识别任务时处理不同形式的数据的过程。多模态数据的融合可以为决策提供更多的信息,从而提高了决策总体结果的准确率[3]。

3.1 多模态融合策略

目前多模态特征融合策略主要有特征层融合[36,51~55]和决策层融合[56~61]两种,此外,还有混合多模态融合[62~64]、模型层融合[46,65~68]。

3.1.1 特征层融合

特征层融合[52]也称早期融合,如图 1 所示。首先把情感声学特征、文本信息和人脸表情特征信息提取出来,然后将提

取出来的语音、文本和表情等特征串联成一个总的特征向量用于情感识别。该融合方法利用不同模态相互之间的联系,但没有考虑到各情感特征的差异性,同时该融合策略很难表示不同模态之间的时间同步性。随着融合模态的增多,会使得学习多种模态特征之间的相关性变得更加困难。

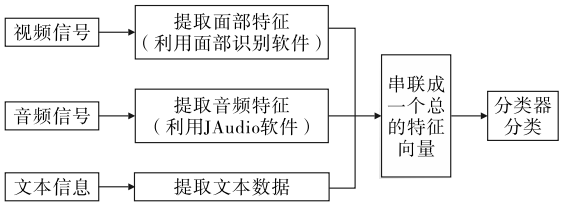

图1　特征层融合

3.1.2　决策层融合

决策层融合[42]也称后期融合,如图 2 所示。首先对语音和面部表情分别提取特征,并将其送入各自的分类器中,再依据某种原则将各个分类器的结果进行融合决策,获得最终的识别结果。决策层的融合较特征层融合更容易进行,该方法充分考虑了语音和表情特征的差异性,语音和表情特征可以选择各自最合适的分类器进行分类,但没有考虑到情感特征之间的联系,学习过程会变得冗长耗时。

3.1.3　混合多模态融合

混合多模态融合法[62]结合特征层融合和决策层融合两种方法,它不仅继承了两种方法的优点,还克服了彼此的缺点。Wollmer 等人[62]提出了一种混合法(图 3),首先利用 BLSTM 网络对提取的音/视频数据进行特征层融合、分类,再利用 ASR 系统对提取的 MFCC 特征生成 BoW/BoNG 的语言特征,并利用线性支持向量机进行分类,将两种结果进行决策融合。

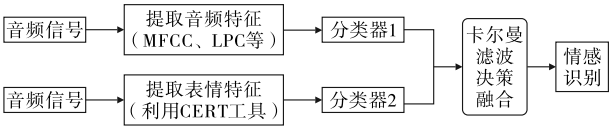

图2　决策层融合

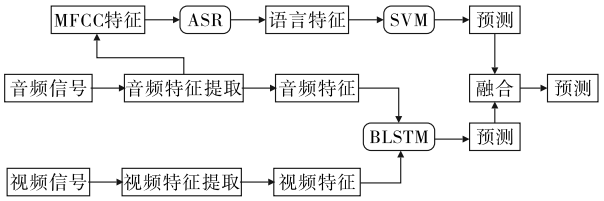

图3　混合多模态融合

3.1.4　模型层融合

模型级融合利用各模态信息流之间的关联信息,根据实验要求建立模型的多模态融合方法。Song 等人[67]基于音/视频和最大互信息原理在不同流上建立一个最优的连接,不同视频流建立三重隐马尔可夫模型。Zeng 等人[68]提出一种多流融合隐马尔可夫模型(MFHMM),该模型基于最大熵准则加权组合法,该方法可以利用多种学习算法取得一个鲁棒性的多流融合结果。

表 1 列举了不同的多模态识别算法利用不同融合策略的识别率。

表 1　对情感识别研究算法进行对比

文献	数据库	特征提取方法	分类算法	模式	融合策略	实验结果
[61]	IEMOCAP(10 折交叉验证)	高斯混合模型	SVM	视频 + 音频	决策层融合	75.45%(平均准确率)
[36]	IEMOCAP(10 折交叉验证)	CNN/RNN openSMILE	多核 SVM	视频 + 音频 + 文本	特征层融合	76.85% (准确率)
[65]	IEMOCAP(10 折交叉验证)	Praat 语音处理工具 MoCap 面部标记坐标	BLSTM	视频 + 音频	模型层融合	71.83% (未加权准确率)
[57]	eNTERFACE(5 折交叉验证)	通用背景模型 + 最大后验概率法 + openSMILE 特征提取器	SVM	视频 + 音频	决策层融合	77.50%(准确率)
[52]	*1	Luxand 软件(面部) JAudio 软件(语音) EmoSenticSpace(文本)	ELM(视觉) SVM(文本、语音)	视频 + 音频 + 文本	特征层融合	87.95%(准确率)
[63]	eNTERFACE/TMU-EMODB (10 折交叉验证)*2	Praat 语音处理工具 Hager 的有效区域跟踪算法 + LDA/PCA(特征提取与降维)	自适应多分类器	音频 + 视频	特征层融合	70.00%/66.00% (平均识别率)
					决策层融合	38.00%/65.00% (平均识别率)
					混合融合法	71.00%/77.00% (平均识别率)

*1:训练集视觉模式采用 CK ++ 数据库,语音模式采用 ISEAR 数据库,文本模式采用 eNTERFACE 数据库,测试集统一采用 eNTERFACE 数据库

*2:实验在两个数据集上进行,识别率与"/"两侧数据库相对应

3.2　多模态融合方法

本节主要介绍多模态融合方法。根据这些方法的基本特性主要分为基于规则的融合法、基于分类融合法、基于估计的融合法以及基于深度学习的融合方法。

3.2.1　基于规则的融合法

基于规则的融合法[23,27]是基于统计学方法的多模态信息融合方法,如线性加权融合[57]、多数同意规则等。线性加权融合利用加法和乘法运算融合不同模态的信息,但此法易受到离群值的影响。多数同意规则是基于多数分类器决策的,它是加权组合的一种特殊情况,大多数分类器得到的相似决策就是最终的结果。

3.2.2　基于分类融合法

基于分类融合法[69]是利用一系列的分类算法将多模信息分成预定义组。在该体系下包括支持向量机、贝叶斯推论、Dempster-Shafer(DS)证据理论、动态贝叶斯网络、神经网络、最大熵模型。支持向量机是分类任务应用最为广泛的监督学习算法,在算法中,输入数据被分成预定义学习组用于解决基于多模态融合模式分类问题。这种方法通常用于决策层融合和混合多模态融合。贝叶斯推论基于概率理论融合来自不同模态的数据或者不同分类器的决策,得到联合概率。该方法被广泛应用于时间序列数据。DS 理论、动态贝叶斯网络都是贝叶斯推论的推广。

3.2.3　基于估计的融合法

基于估计的融合法,该体系包括 Kalman 滤波[58]、粒子滤波。Kalman 滤波用于实时动态低维数据,适用于线性系统;扩展 Kalman 滤波用于非线性系统。Glodek 等人[58]利用 Kalman 滤波融合来自音/视频两个通道分类器的结果。粒子滤波也就是知名的连续蒙特卡洛方法(sequential Monte Carlo method,SMC),是一种基于仿真的成熟模型估计技术,用于获得非高斯和非线性状态空间的状态分布。

3.3 基于深度学习的融合方法

近年来深度信念网络(deep belief network,DBN)、卷积神经网络、递归神经网络等多种深度学习网络被提出。传统的多模态特征融合方法一般是线性融合,深度学习模拟人脑的认知过程,建立多层结构,对样本实现从低层到高层的自主逐层提取特征,可利用该方法来形成不同类型数据的联合特征表示。深度典范相关分析[70]、多模态深度学习[4]以及多模态玻尔兹曼机[5]等这些方法的基本思路是通过不同的深层模型对不同模式的数据进行逐层学习,将学习得到的结果进行合并,以得到多模态联合特征表示。文献[5]提出一种多模态玻尔兹曼机模型(图4),分别将图像和文本特征各自训练一个受限玻尔兹曼机(RBM),然后将两个 RBM 输出组合成一个新的融合特征,送入分类模型进行识别。表情特征和语音情感特征属于非线性关系,人脸表情和语音情感等特征的表现形式差异大,属于交叉模态融合,深度学习在此类融合方面表现出其他机器学习算法都不具备的优异性能。

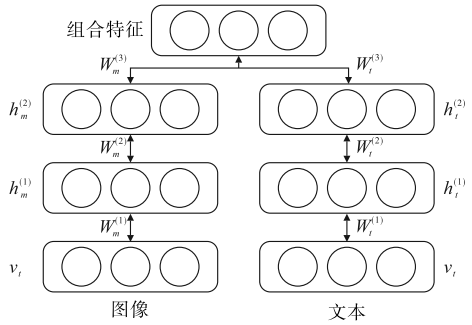

图4 多模态玻尔兹曼机

4 结束语

本文重点阐述了面部表情和语音的情感特征提取技术及融合策略与方法,并对目前较流行的数据库进行了介绍。随着深度学习和大量融合算法的兴起,多模态情感识别得到了快速的发展,但多模态情感识别仍面临着诸多挑战。首先是缺乏标准数据库,目前的数据库是存在争议的,主要问题是数据能否准确地复制自发表情的自然特征。例如在大多数现实场合中,常常通过微笑表达尴尬。大多数据库使用的是诱发表情而不是自发表情,自发表情不易分类,而诱发表情易分类;在此基础上还需要标准的多模态数据库。其次,多模态系统是建立在单模态基础上,单模态通道的噪声、帧的同步性、测试者的发音、大数据的降维以及融合算法的实时性等都是多模态情感识别亟待解决的问题。最后,尽管各种融合算法验证了多模态的有效性,但这些方法均未考虑生理、心理层面的因素。因此,横跨多学科的融合策略是未来情感识别研究的一个发展方向。

参考文献:

[1] D'Mello S K, Kory J. A review and meta-analysis of multimodal affect detection systems[J]. ACM Computing Surveys, 2015, 47(3):1-36.

[2] Poria S, Cambria E, Bajpai R, *et al*. A review of affective computing: from unimodal analysis to multimodal fusion[J]. Information Fusion, 2017, 37(9):98-125.

[3] McGurk H, MacDonald J. Hearing lips and seeing voices[J]. Nature, 1976, 264(5588):746-748.

[4] Noda K, Arie H, Suga Y, *et al*. Multimodal integration learning of robot behavior using deep neural networks[J]. Robotics and Autonomous Systems, 2014, 62(6):721-736.

[5] Srivastava N, Salakhutdinov R. Multimodal learning with deep Boltzmann machines[J]. Journal of Machine Learning Research, 2014, 15(8):2949-2980.

[6] Douglas-Cowie E, Cowie R, Sneddon I, *et al*. The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data[C]//Proc of the 2nd International Conference on Affective Computing and Intelligent Interaction. Berlin: Springer-Verlag, 2007: 488-500.

[7] Douglas-Cowie E, Cowie R, Schröder M. A new emotion database: considerations, sources and scope[C]//Proc of ISCA Workshop on Speech and Emotion. 2000:39-44.

[8] McKeown G, Valstar M, Cowie R, *et al*. The SEMAINE database: annotated multimodal records of emotionally colored conversations between a person and a limited agent[J]. IEEE Trans on Affective Computing, 2012, 3(1):5-17.

[9] Busso C, Bulut M, Lee C C, *et al*. IEMOCAP: interactive emotional dyadic motion capture database[J]. Language Resources and Evaluation, 2008, 42(4):335-359.

[10] Martin O, Kotsia I, Macq B, *et al*. The eNTERFACE'05 audio-visual emotion database[C]//Proc of the 22nd International Conference on Data Engineering Workshops. Washington DC: IEEE Computer Society, 2006.

[11] Emambakhsh M, Evans A. Nasal patches and curves for an expression-robust 3D face recognition[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2016, 39(5):995-1007.

[12] Ekman P, Friesen W V. Facial action coding system[M]//Investigator's Guide. [S. l.]: Palo Consulting Psychologists Press, 1978.

[13] Cootes T F, Taylor C J, Cooper D H. Active shape models: their training and application[J]. Journal Computer Vision and Image Understanding, 1995, 61(1):38-59.

[14] Lanitis A, Taylor C J, Cootes T F. Automatic face identification system using flexible appearance models[J]. Image and Vision Computing, 1995, 13(5):393-401.

[15] Zhang Zhengyou. Feature-based facial expression recognition: sensitivity analysis and experiments with a multilayer perceptron[J]. International Journal of Pattern Recognition & Artificial Intelligence, 1999, 13(6):893-911.

[16] Yacoob Y, Davis L. Computing spatio-temporal representations of human faces[C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1994:70-75.

[17] Haro A, Flickner M, Essa I. Detecting and tracking eyes by using their physiological properties, dynamics, and appearance[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2000:163-168.

[18] 应自炉,唐京海,李景文,等. 支持向量鉴别分析及在人脸表情识别中的应用[J]. 电子学报,2008,36(4):725-730.

[19] Morency L P, Whitehill J, Movellan J. Generalized adaptive view-based appearance model: integrated framework for monocular head pose estimation[C]//Proc of the 8th IEEE International Conference on Automatic Face & Gesture Recognition. 2008:1-8.

[20] 王磊,邹北骥,彭小宁,等. 一种改进的提取人脸面部特征点的 AAM 拟合算法[J]. 电子学报,2006,34(8):1424-1427.

[21] Tian Yingli, Kanade T, Cohn J F. Recognizing action units for facial expression analysis[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2001, 23(2):97-115.

[22] Morency L, Baltrušaitis T, Robinson P. 3D constrained local model for rigid and non-rigid facial tracking[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2012:2610-2617.

[23] Lien J J J, Kanade T, Cohn P D J F, *et al*. Detection, tracking, and classification of action units in facial expression[J]. Robotics and Autonomous Systems, 2000, 31(3):131-146.

[24] Kring A M, Sloan D M. The facial expression coding system (FACES): development, validation, and utility[J]. Psychological Assessment, 2007, 19(2):210-224.

[25] Ohta H, Saji H, Nakatani H. Recognition of facial expressions using muscle-based feature models[C]//Proc of the 14th International Conference on Pattern Recognition. 1998:1379-1381.

[26] Cohen I, Sebe N, Cozman F G, *et al*. Learning Bayesian network classifiers for facial expression recognition using both labeled and unlabeled data[C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2003:595-601.

[27] Verma R, Davatzikos C, Loughead J, *et al*. Quantification of facial expressions using high-dimensional shape transformations[J]. Journal of Neuroscience Methods,2005,141(1):61-73.
[28] Davatzikos C. Measuring biological shape using geometry-based shape transformations[J]. Image and Vision Computing,2001,19(1-2):63-74.
[29] Wen Zhen, Huang T S. Capturing subtle facial motions in 3D face tracking[C]//Proc of the 9th IEEE International Conference on Computer Vision. Washington DC:IEEE Computer Society,2003:1343-1350.
[30] 周飞燕,金林鹏,董军. 卷积神经网络研究综述[J]. 计算机学报,2017,40(6):1229-1251.
[31] 孙晓,潘汀. 基于兴趣区域深度神经网络的静态面部表情识别[J]. 电子学报,2017,45(5):1189-1197.
[32] Xu Baohan, Fu Yanwei, Jiang Yugang, *et al*. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization[J]. IEEE Trans on Affective Computing,2015,9(2):255-270.
[33] Xu Can, Cetintas S, Lee K C, *et al*. Visual sentiment prediction with deep convolutional neural networks[J]. arXiv 1411.5731,2014.
[34] Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks[C]//Proc of IEEE International Conference on Computer Vision. 2015:4489-4497.
[35] You Quanzeng, Luo Jiebo, Jin Hailin, *et al*. Robust image sentiment analysis using progressively trained and domain transferred deep networks[C]//Proc of the 29th AAAI Conference on Artificial Intelligence. [S.l.]:AAAI Press,2015:381-388.
[36] Poria S, Chaturvedi I, Cambria E, *et al*. Convolutional MKL based multimodal emotion recognition and sentiment analysis[C]//Proc of the 16th IEEE International Conference on Data Mining. 2017:439-448.
[37] Iyengar G, Nock H J, Neti C. Audio-visual synchrony for detection of monologues in video archives[C]//Proc of IEEE International Conference on Multimedia and Expo. 2003:329-332.
[38] 韩文静,李海峰. 基于韵律语段的语音情感识别方法研究[J]. 清华大学学报:自然科学版,2009,49(S1):1363-1368.
[39] 韩文静,李海峰,阮华斌,等. 语音情感识别研究进展综述[J]. 软件学报,2014,25(1):37-50.
[40] 刘振焘,徐建平,吴敏,等. 语音情感特征提取及其降维方法综述[J/OL]. 计算机学报,(2017-08-13). http://kns.cnki.net/kcms/detail/11.1826.TP.20170813.1200.006.html.
[41] 韩文静,李海峰,韩纪庆. 基于长短时特征融合的语音情感识别方法[J]. 清华大学学报:自然科学版,2008,48(S1):708-714.
[42] Lee C M, Narayanan S S. Towards detecting emotions in spoken dialogs[J]. IEEE Trans on Speech and Audio Processing,2005,13(2):293-303.
[43] 陶华伟,查诚,梁瑞宇,等. 面向语音情感识别的语谱图特征提取算法[J]. 东南大学学报:自然科学版,2015,45(5):817-821.
[44] 赵力,钱向民,邹采荣,等. 语音信号中的情感识别研究[J]. 软件学报,2001,12(7):1050-1055.
[45] Lugger M, Janoir M E, Yang Bin. Combining classifiers with diverse feature sets for robust speaker independent emotion recognition[C]//Proc of the 17th European Signal Processing Conference. Piscataway, NJ:IEEE Press,2011:1225-1229.
[46] Graves A, Fernández S, Schmidhuber J. Bidirectional LSTM networks for improved phoneme classification and recognition[C]//Proc of the 15th International Conference on Artificial Neural Networks: Formal Models and Their Applications. Berlin: Springer-Verlag, 2005:799-804.
[47] Eyben F, Wöllmer M, Graves A, *et al*. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues[J]. Journal on Multimodal User Interfaces,2010,3(1-2):7-19.
[48] 朱小燕,王昱,徐伟. 基于循环神经网络的语音识别模型[J]. 计算机学报,2001,24(2):213-218.
[49] Huang Zhengwei, Dong Ming, Mao Qirong, *et al*. Speech emotion recognition using CNN[C]//Proc of the 22nd ACM International Conference on Multimedia. New York:ACM Press,2014:801-804.
[50] Anand N, Verma P. Convoluted feelings convolutional and recurrent nets for detecting emotion from audio data[R]. Stanford: Stanford University,2015.
[51] Sarkar C, Bhatia S, Agarwal A, *et al*. Feature analysis for computational personality recognition using YouTube personality data set[C]//Proc of ACM Multi Media on Workshop on Computational Personality Recognition. New York:ACM Press,2014:11-14.
[52] Poria S, Cambria E, Hussain A, *et al*. Towards an intelligent framework for multimodal affective data analysis[J]. Journal Neural Networks,2015,63(3):104-116.
[53] Zhang Shiqing, Zhang Shiliang, Huang Tiejun, *et al*. Multimodal deep convolutional neural network for audio-visual emotion recognition[C]//Proc of ACM International Conference on Multimedia Retrieval. New York:ACM Press,2016:281-284.
[54] Poriaa S, Cambriab E, Howardc N, *et al*. Fusing audio, visual and textual clues for sentiment analysis from multimodal content[J]. Neurocomputing,2016,174(1):56-59.
[55] Castellano G, Kessous L, Caridakis G. Emotion recognition through multiple modalities: face, body gesture, speech[M]//Affect and Emotion in Human-Computer Interaction. Berlin: Springer,2008:92-103.
[56] Sahoo S, Routray A. Emotion recognition from audio-visual data using rule based decision level fusion[C]//Proc of IEEE Students' Technology Symposium. Piscataway,NJ:IEEE Press,2016:7-12.
[57] Dobrišek S, Gajšek R, Mihelič F, *et al*. Towards efficient multi-modal emotion recognition[J]. International Journal of Advanced Robotic Systems,2013,10(1):257-271.
[58] Glodek M, Reuter S, Schels M, *et al*. Kalman filter based classifier fusion for affective state recognition[C]//Proc of International Workshop on Multiple Classifier Systems. 2013:85-94.
[59] Yamasaki T, Fukushima Y, Furuta R, *et al*. Prediction of user ratings of oral presentations using label relations[C]//Proc of the 1st International Workshop on Affect & Sentiment in Multimedia. New York:ACM Press,2015:33-38.
[60] Gharavian D, Bejani M, Sheikhan M. Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks[J]. Multimedia Tools and Applications,2017,76(2):2331-2352.
[61] Metallinou A, Lee S, Narayanan S. Audio-visual emotion recognition using Gaussian mixture models for face and voice[C]//Proc of the 10th IEEE International Symposium on Multimedia. 2008:250-257.
[62] 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.
[63] Mansoorizadeh M, Charkari N M. Multimodal information fusion application to human emotion recognition from face and speech[J]. Multimedia Tools and Applications,2010,49(2):277-297.
[64] Poria S, Cambria E, Gelbukh A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis[C]//Proc of Conference on Empirical Methods in Natural Language Processing. 2015:2539-2544.
[65] Eyben F, Katsamanis A, Wollmer M, *et al*. Context-sensitive learning for enhanced audiovisual emotion classification[J]. IEEE Trans on Affective Computing,2012,3(2):184-198.
[66] Lu Kun, Jia Yunde. Audio-visual emotion recognition with boosted coupled HMM[C]//Proc of the 21st International Conference on Pattern Recognition. Piscataway,NJ:IEEE Press,2012:1148-1151.
[67] Song Mingli, Bu Jiajun, Chen Chun, *et al*. Audio-visual based emotion recognition:a new approach[C]//Proc of IEEE Computer Society Computer Vision and Pattern Recognition. 2004:1020-1025.
[68] Zeng Zhihong, Hu Yuxiao, Liu Ming, *et al*. Training combination strategy of multi-stream fused hidden Markov model for audio-visual affect recognition[C]//Proc of the 14th ACM International Conference on Multimedia. New York:ACM Press,2006:65-68.
[69] Nefian A V, Liang Luhong, Pi Xiaobo, *et al*. Dynamic Bayesian networks for audio-visual speech recognition[J]. EURASIP Journal on Advances in Signal Processing,2002,2002(11):783042.
[70] Andrew G, Arora R, Bilmes J, *et al*. Deep canonical correlation analysis[C]//Proc of the 30th International Conference on Machine Learning. 2013:1247-1255.