软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(4):1060−1070 [doi: 10.13328/j.cnki.jos.005412]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

多文化场景下的多模态情感识别*

陈师哲, 王 帅, 金 琴

(中国人民大学 信息学院,北京 100872)

通讯作者: 金琴, E-mail: qjin@ruc.edu.cn

摘 要: 自动情感识别是一个非常具有挑战性的课题,并且有着广泛的应用价值.探讨了在多文化场景下的多模态情感识别问题.从语音声学和面部表情等模态分别提取了不同的情感特征,包括传统的手工定制特征和基于深度学习的特征,并通过多模态融合方法结合不同的模态,比较不同单模态特征和多模态特征融合的情感识别性能.在CHEAVD中文多模态情感数据集和AFEW英文多模态情感数据集进行实验,通过跨文化情感识别研究,验证了文化因素对于情感识别的重要影响,并提出3种训练策略提高在多文化场景下情感识别的性能,包括:分文化选择模型、多文化联合训练以及基于共同情感空间的多文化联合训练,其中,基于共同情感空间的多文化联合训练通过将文化影响与情感特征分离,在语音和多模态情感识别中均取得最好的识别效果.

关键词: 情感识别;多文化场景;语音情感特征;面部表情特征;多模态融合;深度卷积神经网络

中图法分类号: TP391

中文引用格式: 陈师哲,王帅,金琴.多文化场景下的多模态情感识别.软件学报,2018,29(4):1060−1070. http://www.jos.org.cn/1000-9825/5412.htm

英文引用格式: Chen SZ, Wang S, Jin Q. Multimodal emotion recognition in multi-cultural conditions. Ruan Jian Xue Bao/ Journal of Software, 2018,29(4):1060−1070 (in Chinese). http://www.jos.org.cn/1000-9825/5412.htm

Multimodal Emotion Recognition in Multi-Cultural Conditions

CHEN Shi-Zhe, WANG Shuai, JIN Qin

(School of Information, Renmin University of China, Beijing 100872, China)

Abstract: Automatic emotion recognition is a challenging task with a wide range of applications. This paper addresses the problem of emotion recognition in multi-cultural conditions. Different multi-modal features are extracted from audio and visual modalities, and the emotion recognition performance is compared between hand-crafted features and automatically learned features from deep neural networks. Multimodal feature fusion is also explored to combine different modalities. The CHEAVD Chinese multimodal emotion dataset and AFEW English multimodal emotion dataset are utilized to evaluate the proposed methods. The importance of the culture factor for emotion recognition through cross-culture emotion recognition is demonstrated, and then three different strategies, including selecting corresponding emotion model for different cultures, jointly training with multi-cultural datasets, and embedding features from multi-cultural datasets into the same emotion space, are developed to improve the emotion recognition performance in the multi-cultural environment. The embedding strategy separates the culture influence from original features and can generate more discriminative emotion features, resulting in best performance for acoustic and multimodal emotion recognition.

Key words: emotion recognition; multi-cultural condition; acoustic emotion feature; facial expression feature; multimodal fusion; deepconvolutional neural networks

* 基金项目: 国家重点研发计划(2016YFB1001200)

Foundation items: National Key Research and Development Program of China (2016YFB1001200)

本文由“多媒体大数据处理与分析”专题特约编辑赵耀教授、李波教授、华先胜研究员、文继荣教授、蒋刚毅教授、常冬霞副教授推荐.

收稿时间: 2017-04-30; 修改时间: 2017-06-26; 采用时间: 2017-10-13; jos在线出版时间: 2017-12-01

CNKI网络优先出版: 2017-12-04 06:49:15, http://kns.cnki.net/kcms/detail/11.2560.TP.20171204.0649.012.html

情感在人们的日常生活和交流中扮演着非常重要的角色,通过情感的表达,人们可以更方便地相互沟通和了解.自动情感识别能够赋予机器理解人类情感的能力,这一任务有着极其重要的应用场景[1].例如,在人机交互中,情感识别使得机器人能够根据用户的情感状态做出相对应的反馈,从而提高人机交互的质量;在医疗行业中,通过对心理疾病患者日常生活的情感识别,医生能够更有效地对病情进行诊断和治疗;在网络舆论分析中,对用户的多媒体视频进行情感检测,可以更加准确、全面地了解网民对网络事件、产品等的态度倾向.

随着全球化的发展,网络多媒体数据和产品服务对象也逐渐向多样化的文化群体扩展.心理学研究[2]表明,文化因素对于情感的表达和理解具有重要影响,尽管基本情感状态具有一定程度的表现相似性[3],但不同文化背景的人往往存在情感行为的差异性[4].例如,东方文化对于情感尤其是愤怒等负面情感的表达更加隐忍和内敛,而西方文化则更能释放情感而表现激烈.这种因文化不同导致的情绪表达差异较为一致地体现在文化群体整体中,因此在一种文化背景下得到的自动情感识别标准很有可能高度依赖于该文化背景下的特殊行为,而不适应于其他文化情感的识别.然而,现有的对于自动情感识别的研究较少地考虑到在不同文化背景下对情感识别的影响,大部分工作都是基于同一文化背景下的数据集进行情感识别.因此,在本文中,我们将探索文化因素对自动情感识别的影响,以及提高在多文化场景下自动情感识别的性能.

常用的情感识别模型是将情感状态分为离散的标签[1],例如六大基本情感:高兴、生气、难过、厌恶、恐惧和惊讶.本文采取此类情感建模方法进行离散的情感识别.人们的情感是通过多种行为信息进行表现和传递的,例如语音信号、语言内容、面部表情、肢体手势等等.其中,面部表情和语音信号被认为是最常见的情感行为信号.现有工作[5]对语音和面部表情提出了多种手工定制特征,例如统计声学特征、语音词袋特征、面部表情Dense SIFT 特征、动态的 LBP-TOP 特征等,但是目前对于不同模态下最优的情感特征并没有给出定论.在本文中,我们将进一步探索语音信号和面部表情的不同特征表示,包括传统的手工定制特征和基于深度网络模型自动学习的特征,以及不同模态特征之间的融合,并在单文化和多文化场景中分别讨论其识别和泛化性能.

本文的主要贡献包括:

(1) 探索了不同模态和模态融合的情感特征的情感识别性能.基于语音声学和面部表情两个模态进行情感识别的研究,分别比较了传统语音统计声学特征、面部表情 LBP-TOP 特征与基于深度学习的语音特征 Soundnet、深度面部表情特征 FaceCNN 的情感识别性能,并通过多模态特征融合进一步提高了自动情感识别的效果,证明了不同模态特征之间的互补性.

(2) 探索了文化因素对情感识别的影响,并提出 3 种训练策略提高多文化条件下的情感识别性能.通过在单文化和跨文化场景下的情感识别性能分析证明了文化因素对于情感识别的重要影响,提出了分文化模型选择、多文化联合训练和基于共同情感空间多文化联合训练策略这 3 种训练策略,其中,最后一种训练策略大大减弱了文化因素对于情感识别的影响,提高了情感特征的区分性.

(3) 在两个不同文化的多模态情感数据集上进行实验,验证方法的有效性.在 CHEAVD 中文多模态情感数据集和 AFEW 英文多模态数据集进行实验,这两个数据集均采自于电影电视节目片段,较为贴近现实生活中的情感表达.因此,实验结果更能真实地反映我们所提方法的有效性.

本文第 1 节介绍情感识别的相关研究工作.第 2 节介绍语音声学和面部表情模态的不同情感特征.第 3 节介绍多文化场景下的自动情感识别策略.第 4 节进行情感识别实验和结果分析.最后第 5 节总结全文,并对未来的研究方向进行初步的探讨.

1 相关工作

1.1 基于语音模态的情感识别

语音情感特征的提取主要包括低层次声学特征提取和高层次声学特征转换两个步骤.低层次特征一般在语音信号较为稳定的短时间帧片段上进行提取,例如25ms的帧窗长.通常提取的低层次声学特征有3大类:韵律特征、声音质量特征和谱相关特征[6].韵律描述了说话声音的语调、音高、音长、快慢和轻重等方面的变化.

常用的韵律特征有时长(duration)、能量(energy)、语调(pitch)、基频(F0)等,其情感区分能力在语音情感识别领域研究得到了广泛认可.声音质量是用于衡量语音的纯净、清晰和辨识程度的评价指标,例如喘息、颤音、哽咽等的声学表现,能体现出情感的波动状态.常用的声学质量特征包括:共振峰频率及其带宽、频率微扰和振幅微扰、声门参数等.谱相关特征体现了声道形状变化和发声运动之间的相关性,在语音信号处理的各个领域,如语音识别、话者识别等均有成功的应用.在语音情感识别任务中,常用的谱相关特征包括:线型谱,如 LPC(linear predictor coefficient),倒谱特征,如 MFCC(mel-frequency cepstral coeffient)等.

为得到在句子层面固定维度的高层次语音情感特征,需要对提取的多个低层次帧级特征进行转换.Interspeech 2009 情感识别比赛[7]提出了使用统计函数得到全局的语音情感特征,统计指标包括极值、均值、方差等.Chen 等人[8]使用了声学词袋模型(bag-of-audio-words)和高斯超向量(GMM supervector)将句子中的帧级特征通过 *K*-means 或 GMM 等聚类方法得到声学码本,将句子层次的语音特征表示为所有帧级特征在这些码本上的分布.Xia 等人[9]提出了降噪自编码器(denoising autoencoder),将已得到的句子层次语音特征进一步提纯,去除情感中性语音特征部分,得到更加具有情感区分力的声学情感特征.Deng 等人[10]强调了跨数据集的语音情感识别问题,对多个数据集使用同一编码层进行特征编码而使用不同解码层重建原始情感特征,反映不同数据集中的区别,从而改善了跨数据集的情感识别效果.Huang 等人[11]基于语音信号转换的频谱图,利用二维深度卷积神经网络提取深度的语音情感特征.

1.2 基于面部表情的情感识别

常用的面部表情特征主要有 LBP-TOP[12]、Dense SIFT[13]、HOG[14]等多种手工定制的图像特征.LBP-TOP 特征引入时间维度,能够提取视频或图片序列的动态纹理特征;Dense SIFT 密集地在图片各个区域提取 SIFT 特征,具有位置、尺度、旋转不变性;HOG 特征是图像局部区域的梯度方向直方图分布.除了基本的图像特征外,Yao 等人[15]通过成对情感类别比较,提取面部运动单元相关表情特征.Kim 等人[16]对于视频表情识别提出一种无监督的分割方法以去除语音说话导致的面部肌肉运动影响.Chen 等人[17]使用了深度卷积神经网络提取图像的表情特征.Jung 等人[18]根据静态表情图片和动态面部特征点变化分别训练两个神经网络进行特征提取.

在面部表情分类方面,Sebe 等人[19]采用 *K* 近邻法(KNN)进行静态人脸表情图像的情感分析.Pantic 和 Rothkrantz[20]根据专家规则间建立了具有二层结构的专家系统作为情感识别模型.Ma 等人[21]利用结构性单隐层的前馈神经网络作为表情分类器,超过了普通神经网络的识别性能.Ioannou 等人[22]研究了模糊神经网络在人脸表情识别中的效果.为体现表情的动态变化,Cohen 等人[23]提出二层次的隐马尔可夫模型(HMM),其中,第 1 层分别为 6 种表情数据构建 6 个 HMM 模型,第 2 层将单个面部表情的 HMM 的状态输出信息联合作为该层 HMM 模型的输入,训练获取情感之间的转移概率,从而不仅能够识别表情还可以在视频中进行定位.

1.3 多模态情感识别

结合不同模态特征往往对于情感识别具有额外的帮助,常用的多模态情感识别的融合方法主要包括:前期融合、后期融合和基于模型的融合[24].

(1) 前期融合对不同模态的特征进行前期拼接作为情感分类器模型的输入特征,其应用非常广泛,并且取得了不错的性能[17,25],但是容易受到维度爆炸的影响,训练速度下降,且在数据规模较小时容易造成过拟合.

(2) 后期融合消除了前期融合维度爆炸的缺点,它将不同模态的情感输出作为输入,为情感识别训练一个第 2 层的基于不同模态的输出情感识别模型[17],但却忽略了特征之间的关系.Zhang 等人[1]利用后期融合的各种策略对多模态融合进行实验与对比,结果显示,不同的后期融合策略对整体的识别会有所影响.

(3) 以模型为基础的多模态融合综合了前两者的优点,但这类模型取决于具体的情感分类器.例如,神经网络模型融合可拼接不同模态的网络隐层,对于核模型,可以通过核融合结合多模态特征.由于对于不同数据,不同模态对情感识别的重要性不一致,Chen 等人[26]提出多模态注意力选择机制以动态地调整不同模态融合的权重.Peng 等人[27]基于深度神经网络提出层次化的结构用于跨模态的关联学习.Wu 等人[28]通过点击图学习不同模态之间的共同特征表示.

1.4 多文化的情感识别研究

心理学方面对多文化情感识别采用了心理学相关的实验和分析研究.Ekman 等人[3]通过给不同人种观看不同情感的面部表情照片,提出了基本情感(basic emotions)在跨文化情形下具有相似的面部表情表现.Tickle 等人[4]在英语和日本语之间的语音情感表达比较中发现语音情感信号也具有一定程度跨文化的相似性.Elfebein 等人[2]的跨文化情感分析表明,尽管情感表现具有相似性,但当情感都是由相同文化背景群体的人表达和进行理解时,情感的识别准确率最高,说明了不同文化群体之间在情感表达和理解上存在差异.

对于自动情感识别研究,Hozjan 等人[29]在多种不同语言的语音情感识别实验中发现,与韵律相关的声学信号对于不同语言的语音情感识别具有较好的泛化性能.Elbarougy 等人[30]探索了跨语言场景下的维度情感识别(dimensional emotion recognition),提出使用分层的情感识别模型模拟人类对于不同文化的语音情感处理分析过程,包括底层声学特征、中层的情感原语(semantic primitives)和高层的情感维度.Sagha 等人[31]首先使用语言识别模型预测语音的语言类型,然后基于预测的语言类型选择对应语言的情感识别分类器进行多语言的语音情感识别,取得较好的多语言情感识别效果.Abdelwahab 等人[32]提出使用模型自适应(model adaptation)方法,利用少量同领域的数据对跨领域(cross domain)的情感模型进行自适应调整,使其在该领域数据集上的情感识别性能有所提升.这些跨文化的自动情感识别工作主要集中在语音情感识别上,而本文中将讨论语音、面部表情及多模态融合情感识别进行多文化下的情感识别,并提出新的训练策略以提高多文化下的情感识别效果.

2 多模态情感特征和融合

2.1 语音情感特征

2.1.1 统计声学特征

统计声学特征在语音情感识别任务中的应用广泛且性能优异.我们使用开源工具 OpenSmile[33]提取在 InterSpeech 2009 Paralinguistic Challenge 中所使用的统计声学特征 IS09[7].首先,以 25ms 为窗长、10ms 为窗移从原始声音信号中提取低层次声学特征,包括能量(energy)、基频(F0)、跨零率(zcr)、梅尔倒谱系数(MFCC)等常见帧级特征及其与相邻特征的相对变化量;然后,基于 12 种不同的统计函数,将语音句子中的多个帧级特征转换为高层次的统计声学特征,例如均值(mean)、方差(standard deviation)、偏度系数(skewness)等;为了使不同特征维度的范围基本保持一致,我们针对训练集计算得到每个特征维度的均值和方差,对所有特征数据进行零归正(z-norm),最终得到 384 维的 IS09 统计声学特征.

2.1.2 深度声学特征

随着深度学习的发展,越来越多的研究开始直接从原始的数据中自动学习最佳特征表示,而非传统专家设计的特征(hand-crafted features).因此,在本文中,我们将探讨基于深度学习自动学习的语音特征在语音情感识别中的性能.基于深度一维卷积神经网络(1-D CNN)模型 Soundnet[34],直接以原始的音频信号为输入提取语音特征.Soundnet 的网络结构如图 1 所示,由 7 个一维的全卷积层(full convolutional layer)和相邻的最大池化层(max pooling layer)构成.由于语音情感数据的稀缺难以训练具有大规模参数的 Soundnet 网络,我们使用 Aytar 等人[34]基于大规模无标注的视频数据预训练的模型参数.该模型使用现在较为成熟的视觉方面的识别模型作为老师,预测得到视频中物体场景等的概率分布(即图 1 中 output 层 1 401 个输出概率,对应 ImageNet 和 PlacesNet 中 1 401 个不同的语义类别)作为知识传授给学生 Soundnet 模型.训练得到的模型的底层特征具有良好的泛化性能,可应用于多种不同的语音识别任务.因此,我们直接利用该预训练的模型抽取深度语音特征.在本文中,我们提取了 Soundnet 网络 conv6 层的 512 维的语音特征,并通过平均池化(mean pooling)得到句子层次的语音特征,最后使用 L2 标准化(L2-normalization)对特征进行规范化处理.图 1 中,粗边框的方框为全卷积层,内部数字为该卷积层的卷积核个数;斜线方框为最大池化层.Soundnet 以原始音频信号作为输入,视觉模型预测的 1 401 维概率为目标进行训练.

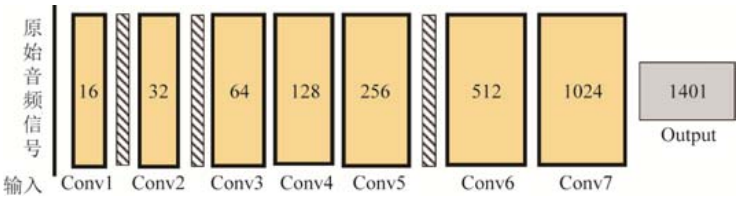


Fig.1 The architecture of Soundnet to extract deep acoustic features

图 1 深度声学模型 Soundnet 的网络结构

2.2 面部表情情感特征

2.2.1 人脸检测与预处理

我们使用开源工具 SeetaFace[35]进行人脸检测,其采用漏斗结构(funnel-structured cascade,简称 FuSt)由粗到细级联进行多角度人脸检测,从而达到较高的准确度和检测速度.我们从视频中每 5 帧检测 1 次人脸并进行面部特征点检测.在提取 LBP-TOP 特征时,对人脸图片根据面部特征点进行对齐.

2.2.2 LBP-TOP

LBP-TOP(local binary patterns of three orthogonal planes)[12]是在面部表情识别领域应用最广泛且性能较好的特征之一.传统的二维图片纹理特征(local binary pattern,简称 LBP)根据图片中每个像素点与其周围相邻像素点的大小关系进行编码,最终将所得到的整个图片在其编码空间的分布作为特征.LBP-TOP 针对 LBP 在视频的时间维度上作了扩展,使其能够描述表情中动态的图片纹理,并基于区域块进行特征提取以着重强调面部表情中局部区域的动态特性.给定视频中面部表情序列,根据二维图片的长轴 X、宽轴 Y 和视频的时间轴 T 形成 XY、XY 和 XZ 这 3 个相互正交的平面.LBP-TOP 在 XY、XY 和 XZ 平面分别计算 LBP 特征,然后将 3 个平面的特征串联得到与时间空间相关的纹理特征.在本文中,我们将图片分为 4×4 个区域块,在每个区域块提取 LBP-TOP 特征进行拼接,最终得到 2 832 维视频层次的 LBP-TOP 特征.

2.2.3 深度表情特征

深度卷积神经网络(convolutional neural network,简称 CNN)在物体识别、场景检测等多种视觉相关任务中都取得了一流的识别性能.通过层次化的网络结构,CNN 能够提取不同抽象层次的图像特征,例如低层次的不同方向的线段和高层次的物体轮廓等;而通过图片中不同区域的参数共享机制,显著降低了参数规模,提高了 CNN 的计算效率和泛化性能.本文从头开始训练了 CNN 模型,用于提取面部表情特征.我们使用 FER+面部表情数据集[36]作为 CNN 的训练数据.FER+数据集包括 35 538 张人脸图片,每张图片由 10 个标注者分别标注为 8 类情感之一(中性、高兴、惊讶、悲伤、生气、厌恶、恐惧和轻蔑),因此可以得到每张图片中面部表情的情感概率分布.我们使用的 CNN 网络架构如图 2 所示,采用 VGGNET 架构[37],网络包括 12 个全卷积层和 4 个最大池化层.图中,粗边框的方框为全卷积层,内部数字为该卷积层的卷积核个数;斜线方框为最大池化层.该模型以原始图片像素作为输出,输出 8 类情感类别.我们将图片的情感概率分布作为优化目标,而非独热编码(one-hot vector),优化网络预测的情感概率与目标情感概率之间的交叉熵(cross-entropy): $L=-\sum_{i=1}^{N}\sum_{k=1}^{8}p_k^i\log q_k^i$. 其中,$p$ 和 q 分别为目标情感和预测情感概率分布,为数据集图片个数.训练好 CNN 模型后,再提取 conv5 层作为本文使用的深度表情特征,并使用平均池化(mean pooling)得到视频层次特征.

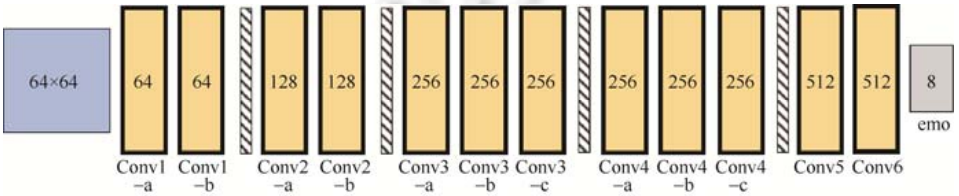


Fig.2 The architecture of our CNN model to extract facial expression features

图 2 面部表情特征模型 CNN 网络结构

由于所提取特征维度相对数据规模较低,并不会造成严重过拟合,因此,我们采用前期融合策略进行不同特征的融合,即将不同的特征进行拼接作为分类模型的输入.

3 多文化情感识别

为方便叙述,假设我们有 *A*,*B* 两种文化的多模态情感数据集,本文提出的模型可扩展到多种文化.我们提出3种不同的识别策略,将文化因素对情感识别的影响考虑在情感识别模型中.

3.1 分文化模型选择

由于不同文化背景群体的情感产生、表现和理解各有不同,因此,一个简单的策略是根据预测的文化类别选择专门在该文化的情感数据下训练的情感识别模型.首先,我们根据提取的情感特征分别对 *A* 和 *B* 文化训练专门的情感识别模型;对于待分类的数据,首先利用一个文化识别模型预测其文化类别,并以此选择对应文化的情感识别模型进行情感预测,其流程示意图如图 3 所示.优点是使用了纯净文化背景下的数据进行情感识别模型训练,无需考虑文化不同带来的情感表达差异,但其性能受文化识别模型的影响,容易造成错误累加,且在训练单文化情感识别模型时可能存在数据稀缺问题,不能利用其他文化背景的数据进行辅助学习.

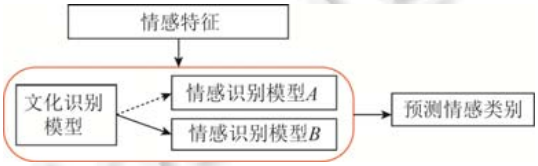


Fig.3 The frame work of culture selection strategy for multi-cultural emotion recognition

图 3 分文化模型选择策略示意图

3.2 多文化联合训练

另一种直观的策略是不考虑文化背景的差异,直接联合所有的文化背景下的数据进行情感识别模型的训练.其优点是可以增加情感识别训练的数据集,能够补充每类单文化场景下较为稀少的情感数据资源,但是,由于忽略了文化背景的差异,情感识别性能可能会受到影响.

3.3 基于共同情感空间的多文化联合训练

综合前两种策略的优缺点,一方面我们希望能在情感识别模型中去除文化因素的影响以提高情感识别效果,另一方面又能增多情感识别的训练数据,尤其是对于数据规模较小的文化而言,更多的训练数据是提高性能的很大突破点.因此,我们提出基于共同情感空间的多文化联合训练策略,其基本思想是将提取的特征分解为与情感相关和文化相关的两部分,将不同文化下的特征映射到同一情感空间中,然后用所有的数据进行情感模型的训练.我们采用去噪自编码器(denoising autoencoder)作为基本模型,其模型结构如图 4 所示.

原始输入特征 x 加噪后的特征 $\tilde{x}$ 通过映射得到与情感相关的特征 z_e 和与文化相关的特征 z_c,转换公式如下:

$$\text{情感相关隐层特征:}\quad z_e=\varphi(W_{ze}\tilde{x}+b_{ze}) \tag{1}$$

$$\text{文化相关隐层特征:}\quad z_c=\varphi(W_{zc}\tilde{x}+b_{zc}) \tag{2}$$

情感特征 z_e 和文化特征 z_c 既要满足能更好地预测对应的情感或文化的条件,同时也能重建原始特征 x,如下所示:

$$\text{情感预测概率分布:}\quad p_e=softmax(W_{pe}z_e+b_{pe}) \tag{3}$$

$$\text{文化预测概率分布:}\quad p_c=softmax(W_{pc}z_c+b_{pc}) \tag{4}$$

$$\text{情感相关特征重建:}\quad r_e=\varphi(W_{re}z_e+b_{re}) \tag{5}$$

$$\text{文化相关特征重建:}\quad r_c=\varphi(W_{rc}z_c+b_{rc}) \tag{6}$$

$$\text{原始特征重建:}\quad r_x=r_e\oplus r_c \tag{7}$$

其中,W_{**},b_{**} 均为模型参数,$\varphi(x)=\max(x,0)$ 为 RELU 激活函数.我们同时优化多个目标函数,包括情感和文化的

识别性能、原始特征重建的性能以及情感特征与文化特征之间尽量满足正交关系,从而使得从原始输入特征分解出更具情感区分力的特征:

情感预测损失函数: $L_e = cross\_entropy(p_e, \hat{p}_e)$ (8)

文化预测损失函数: $L_c = cross\_entropy(p_c, \hat{p}_c)$ (9)

原始特征重建损失函数: $L_r = r_x - x_2^2$ (10)

情感-文化重建特征相似度损失函数: $S_c = r_e, r_c$ (11)

其中, $\hat{p}_e$ 和 $\hat{p}_c$ 分别为真实的情感和文化标签.因此,整体的模型优化目标为最小化:

$$L = L_e + L_c + L_r + S_c \quad (12)$$

最终,得到的情感特征 z_e 为去除了文化因素影响的情感特征,我们将 z_e 作为新的特征表示,在所有文化的数据集上训练得到多文化联合的情感分类模型.

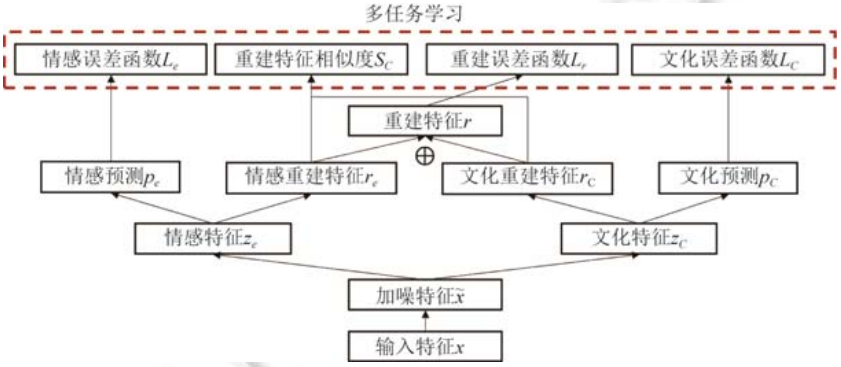


Fig.4 Training process of the multi-cultural emotion recognition model which separates the common emotion space from original features

图 4 基于共同情感空间的多文化联合训练模型结构

4 数据集

4.1 中文多模态情感数据集CHEAVD

CHEAVD(Chinese natural emotional audio-visual database)[38]是由中国科学院自动化研究所构建的中文多模态情感数据集,包括 140 分钟从中文影视剧和电视节目中所截取的音视频片段,这些音视频片段的背景复杂多样,接近于真实生活中的情感数据.每个音视频片段的时长从 1s~19s 不等,平均时长为 3.3s,分别标注为 7 类常见情感(生气、高兴、悲伤、惊讶、厌恶、担心、焦虑)及中性情感中的一种.整个数据集被分为 3 部分:1 981 个视频为训练集、243 个视频为验证集以及 628 个视频为测试集.然而,该数据集的情感类别分布非常不均匀,其中主要情感类别(高兴、生气、悲伤和中性)占所有数据的 80%左右.因此,在我们的工作中,仅选用每个集合中的高兴、生气、悲伤以及 60%的中性情感类别的视频进行训练和测试,使得情感类别分布更加均匀.这 4 类情感类别的视频数量见表 1 上半部分.

Table 1 Number of video clips for the four emotion classes in the CHEAVD and AFEW datasets

表 1 CHEAVD 和 AFEW 数据集中 4 类情感的视频数量分布

数据	集合	中性	生气	高兴	悲伤	总和
CHEAVD	训练集	435	399	307	256	1397
中文情感	验证集	54	49	38	31	172
数据集	测试集	85	73	56	75	289
AFEW	训练集	123	114	128	100	465
英文情感	验证集	21	19	22	17	79
数据集	测试集	63	64	63	61	251

4.2 英文多模态情感数据集AFEW

AFEW(acted facial expression in the wild database)[39]由从英文电影和电视节目中截取的音视频片段组成,因而能较为接近真实生活的情感表达.每个音视频片段被标注为 6 类基本情感(生气、高兴、悲伤、惊讶、厌恶、恐惧)和中性情感之一.数据集包括训练集、验证集和测试集 3 部分,然而测试集的标签并不公开.因此,我们从训练集中随机选出 15%的音视频片段作为验证集,原数据集的验证集作为我们实验的测试集.为了与 CHEAVD 数据集进行比较,我们也只选取了生气、高兴、悲伤和中性情感的音视频数据.实验中,AFEW 数据的情感分布见表 1 的下半部分.

5 实验和结果分析

5.1 实验设置

我们使用了支持向量机 SVM(super vector machine)和随机森林(random foreset)作为情感分类器,并通过格搜索(grid search)选择在验证集上性能最好的超参数作为最终模型.对于 SVM,我们搜索了线性核(linear kernel)和 RBF 核(RBF kernel),并在 2^{-7}~2^{10} 之间优化代价超参数;对于随机森林,我们从 100~600 以 100 为步长优化树的棵树,从 2~20 以步长为 2 优化树的深度.在基于共同情感空间的多文化联合模型中,我们设置情感相关隐层特征和文化相关隐层特征的维度均为 300.使用准确率(ACC)、宏查全率(MAR)、宏查准率(MAP)和宏 F1(MF1)评价情感分类性能的指标.

5.2 单模态和多模态特征情感识别

首先,我们对不同模态的传统特征和深度特征进行比较.表 2 展示了在 CHEAVD 和 AFEW 这两种不同文化数据集中不同模态特征的识别性能.对于语音情感特征,在 CHEAVD 数据集上,深度语音特征 Soundnet 比传统专家特征 IS09 有绝对数值 4%的显著提高,但在 AFEW 数据集上,IS09 与 Soundnet 的性能基本相当.整体上,我们认为基于深度学习的特征 Soundnet 优于传统统计声学特征 IS09.对于面部表情特征,深度表情特征 FaceCNN 在两个数据集上与 LBP-TOP 相比都取得了非常显著的提升,说明了在图像上深度学习特征的有效性.

我们进一步探索不同模态之间是否存在互补性,表 2 最后一行展示了融合语音特征 Soundnet 和面部表情特征 FaceCNN 的情感分类结果,多模态特征的融合大大提高了单模态的识别准确率,在不同文化和不同指标上都有一致的大幅度提升.我们也可以看到,面部表情相对语音声学的情感识别性能更强,但两者之间存在非常强的互补性,结合语音和面部表情对情感识别是非常有益的.同时,AFEW 数据集的单模态和多模态识别结果都优于 CHEAVD 数据集的结果,这可能也反映了不同文化背景下情感识别的难度,CHEAVD 对应的东方文化情感比 AFEW 所对应的西方文化情感更难以识别,可能源于东方文化情感表达的内敛性.

Table 2 Performance comparison of different modality features and multimodal fusion on CHEAVD and AFEW

表 2 CHEAVD 和 AFEW 数据集中不同模态和多模态融合的性能比较

特征	CHEAVD				AFEW			
	ACC	MAR	MAP	MF1	ACC	MAR	MAP	MF1
IS09	43.25	40.97	44.50	42.66	48.61	48.23	48.90	48.56
Soundnet	46.02	45.10	49.05	46.99	48.21	47.91	48.45	48.18
LBP-TOP	37.37	37.59	40.39	38.94	54.58	54.29	53.62	53.96
FaceCNN	53.98	54.01	60.19	56.93	64.54	64.34	65.70	65.01
Soundnet-FaceCNN	62.28	62.36	66.47	64.35	68.53	68.26	70.59	69.41

5.3 多文化条件下情感识别

为探究文化因素对情感识别性能的影响,我们展示了在跨文化情况下情感识别的效果,即我们用文化 *A* 训练得到的情感识别模型在文化 *B* 测试数据上进行测试,实验结果见表 3.可以看到,在跨文化场景下,情感识别性能显著下降,尤其是用 AFEW 所对应的西方文化情感模型预测 CHEAVD 所对应的东方文化数据.这也反映了不同文化之间情感表现的差异带来的多文化情境下情感识别的挑战.与语音特征相比,面部表情特征的性能下降得更剧烈,这说明语音特征的跨文化表达比面部表情更为相似.最后,使用多模态特征在跨文化背景下所进行的

实验结果表明,其效果仍然优于单模态特征.

Table 3 Cross culture performance comparison of different modality features on CHEAVD and AFEW

表 3 CHEAVD 和 AFEW 数据集中不同模态特征在单文化和跨文化条件下的情感识别性能比较

特征	模型	CHEAVD				AFEW			
		ACC	MAR	MAP	MF1	ACC	MAR	MAP	MF1
Soundnet	同文化训练数据	46.02	45.10	49.05	46.99	48.21	47.91	48.45	48.18
	跨文化训练数据	36.68	36.58	36.94	36.76	39.04	39.01	43.62	41.19
FaceCNN	同文化训练数据	53.98	54.01	60.19	56.93	64.54	64.34	65.70	65.01
	跨文化训练数据	40.14	43.33	47.40	45.27	47.01	46.50	58.84	51.95
Soundnet-FaceCNN	同文化训练数据	62.28	62.36	66.47	64.35	68.53	68.26	70.59	69.41
	跨文化训练数据	42.91	45.67	51.23	48.29	57.77	57.65	64.63	60.94

为了提高情感识别在多文化背景下的效果,我们比较了基于 3 种不同策略的模型在 CHEAVD 和 AFEW 联合的多文化数据集中进行训练测试的情感识别性能.实验结果见表 4,对于语音情感识别,由于跨文化的差异性较小,多文化联合训练比单文化选择训练和单文化训练有少许提升,而基于同一情感空间的联合训练由于进一步从情感特征中分离了文化因素的影响,进一步提高了在多文化语音情感识别的效果.对于面部表情特征,单文化训练优于 3 种多文化训练策略,这说明了不同文化之间面部表情的差异性仍然较大,但基于共同情感空间的联合训练与多文化联合训练相比,在性能上仍然有一定的提高,说明我们提出的模型能够减少文化因素的影响.最后,在多模态特征的多文化情感识别中,基于共同情感空间的多文化联合训练策略也取得了最好的识别性能.

Table 4 Performance comparison of different training strategies for multi-cultural emotion recognition

表 4 在多文化情感识别中 3 种不同训练测试策略的情感识别性能比较

特征	模型	CHEAVD				AFEW			
		ACC	MAR	MAP	MF1	ACC	MAR	MAP	MF1
Soundnet	单文化训练	46.02	45.10	49.05	46.99	48.21	47.91	48.45	48.18
	分文化模型选择	46.02	43.69	45.86	44.75	46.22	45.97	48.65	47.27
	多文化联合训练	49.13	46.54	47.15	46.84	48.61	48.34	51.16	49.71
	基于共同情感空间联合训练	50.17	48.44	50.07	49.24	51.00	50.90	52.33	51.60
FaceCNN	单文化训练	53.98	54.01	60.19	56.93	64.54	64.34	65.70	65.01
	分文化模型选择	53.63	53.61	59.85	56.56	63.75	63.54	65.27	64.39
	多文化联合训练	52.25	52.64	56.42	54.46	59.36	59.13	65.26	62.04
	基于共同情感空间联合训练	52.94	52.95	59.02	55.82	60.96	60.63	67.16	63.73
Soundnet-FaceCNN	单文化训练	62.28	62.36	66.47	64.35	68.53	68.26	70.59	69.41
	分文化模型选择	62.28	62.33	66.39	64.30	67.73	67.53	69.41	68.45
	多文化联合训练	60.55	61.06	64.24	62.61	65.74	65.49	70.12	67.72
	基于共同情感空间联合训练	64.36	64.59	66.63	65.60	69.32	69.14	72.12	70.60

6 结论和未来工作展望

在全球化经济和文化交流日益频繁的背景下,文化因素对于情感识别的影响不容忽视.在本文中,我们对多文化场景下的多模态情感识别进行了探讨.在语音和面部表情模态方面分别比较了基于深度学习和传统手工定制的特征,实验结果表明,基于深度学习的特征具有更强的情感识别性能,且语音和面部表情特征具有较高的互补性,多模态融合对单文化和多文化情感识别均有帮助.我们提出 3 种不同策略应对多文化条件下情感识别模型训练的问题,包括分文化模型选择、多文化联合训练以及基于共同情感空间的多文化联合训练,其中,基于共同情感空间的多文化联合训练通过多任务的训练学习,进一步将文化影响的特征与情感特征分离,在语音和多模态情感识别中均取得最好的识别效果.在未来的工作中,我们将进一步探讨在多文化背景下情感识别的模型以及在更多的数据上验证本文得到的结论的一致性.

References:

[1] Zhang S. Reasearch on emotion recognition based on speech and facial expression [Ph.D. Thesis]. Chengdu: University of Electronic Science and Technology of China, 2012 (in Chinese with English abstract).

[2] Elfenbein HA, Ambady N. On the universality and cultural specificity of emotion recognition: A meta-analysis. Psychological Bulletin, 2002,128(2):203. [doi: 10.1037/0033-2909.128.2.203]

[3] Darwin C, Ekman P, Prodger P. The Expression of the Emotions in Man and Animals. New York: Oxford University Press, 1998.

[4] Tickle A. English and Japanese speakers' emotion vocalisation and recognition: A comparison highlighting vowel quality. In: Proc. of the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion. 2000. 104–109. http://www.isca-speech.org/archive_open/archive_papers/speech_emotion/spem_104.pdf

[5] Zeng Z, Pantic M, Roisman GI, Huang TS. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2009,31(1):39–58. [doi: 10.1109/TPAMI.2008.52]

[6] Han WJ, Li HF, Ruan HB, Ma L. Review on speech emotion recognition. Ruan Jian Xue Bao/Journal of Software, 2014,25(1): 37–50 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4497.htm [doi: 10.13328/j.cnki.jos.004497]

[7] Schuller B, Batliner A, Steidl S, Seppi D. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, 2011,53(9):1062–1087. [doi: 10.1016/j.specom.2011.01.011]

[8] Chen S, Jin Q, Li X, Yang G, Xu J. Speech emotion classification using acoustic features. In: Proc. of the 9th Int'l Symp. on Chinese Spoken Language Processing (ISCSLP). IEEE, 2014. 579–583. [doi: 10.1109/ISCSLP.2014.6936664]

[9] Xia R, Liu Y. Using denoising autoencoder for emotion recognition. In: Proc. of the Interspeech. 2013. 2886–2889. http://isca-speech.org/archive/archive_papers/interspeech_2013/i13_2886.pdf

[10] Deng J, Xia R, Zhang Z, Liu Y, Schuller B. Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition. In: Proc. of the 2014 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014. 4818–4822. [doi: 10.1109/ICASSP.2014.6854517]

[11] Huang Z, Dong M, Mao Q, Zhan Y. Speech emotion recognition using CNN. In: Proc. of the 22nd ACM Int'l Conf. on Multimedia. ACM, 2014. 801–804. [doi: 10.1145/2647868.2654984]

[12] Zhao G, Pietikainen M. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2007,29(6). [doi: 10.1109/TPAMI.2007.1110]

[13] Yang J, Yu K, Gong Y, Huang T. Linear spatial pyramid matching using sparse coding for image classification. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2009). IEEE, 2009. 1794–1801. [doi: 10.1109/CVPR.2009.5206757]

[14] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 2005). 2005. 886–893. [doi: 10.1109/CVPR.2005.177]

[15] Yao A, Shao J, Ma N, Chen Y. Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In: Proc. of the 2015 ACM on Int'l Conf. on Multimodal Interaction. ACM, 2015. 451–458. [doi: 10.1145/2818346.2830585]

[16] Kim Y, Mower PE. Say cheese vs. smile: Reducing speech-related variability for facial emotion recognition. In: Proc. of the 22nd ACM Int'l Conf. on Multimedia. ACM, 2014. 27–36. [doi: 10.1145/2647868.2654934]

[17] Chen S, Li X, Jin Q, Zhang S, Qin Y. Video emotion recognition in the wild based on fusion of multimodal features. In: Proc. of the 18th ACM Int'l Conf. on Multimodal Interaction. ACM, 2016. 494–500. [doi: 10.1145/2993148.2997629]

[18] Jung H, Lee S, Yim J, Park S, Kim J. Joint fine-tuning in deep neural networks for facial expression recognition. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 2983–2991. [doi: 10.1109/ICCV.2015.341]

[19] Sebe N, Lew MS, Sun Y, Cohen I, Gevers T, Huang TS. Authentic facial expression analysis. Image and Vision Computing, 2007, 25(12):1856–1863. [doi: 10.1016/j.imavis.2005.12.021]

[20] Pantic M, Rothkrantz LJM. Expert system for automatic analysis of facial expressions. Image and Vision Computing, 2000,18(11): 881–905. [doi: 10.1016/S0262-8856(00)00034-2]

[21] Ma L, Khorasani K. Facial expression recognition using constructive feed forward neural networks. IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics, 2004,34(3):1588–1595. [doi: 10.1109/TSMCB.2004.825930]

[22] Ioannou SV, Raouzaiou AT, Tzouvaras VA, Mailis TP, Karpouzis KC, Kollias SD. Emotion recognition through facialexpression analysis based on a neurofuzzy network. Neural Networks, 2005,18(4):423–435. [doi: 10.1016/j.neunet.2005.03.004]

[23] Cohen I, Sebe N, Garg A, Chen LS, Huang TS. Facial expression recognition from video sequences: Temporaland static modeling. Computer Vision and Image Understanding, 2003,91(1):160–187. [doi: 10.1016/S1077-3142(03)00081-X]

[24] Cao TY. Researchon multi-modal fusion emotion recognition [Ph.D. Thesis]. Tianjin: Tianjin University, 2012 (in Chinese with English abstract).

[25] Chen S, Dian Y, Li X, Lin XZ, Jin Q, Liu HB, Lu L. Emotion recognition in videos via fusing multimodal features. In: Proc. of the Chinese Conf. on Pattern Recognition. Singapore: Springer-Verlag, 2016. 632–644. [doi: 10.1007/978-981-10-3005-5_52]

[26] Chen S, Jin Q. Multi-Modal conditional attention fusion for dimensional emotion prediction. In: Proc. of the 2016 ACM on Multimedia Conf. ACM, 2016. 571–575. [doi: 10.1145/2964284.2967286]

[27] Peng Y, Huang X, Qi J. Cross-Media shared representation by hierarchical learning with multiple deep networks. In: Proc. of the IJCAI. 2016. 3846–3853. http://www.ijcai.org/Proceedings/16/Papers/541.pdf

[28] Wu F, Lu X, Song J, Yan S, Zhang ZM, Rui Y, Zhuang Y. Learning of multimodal representations with random walks on the click graph. IEEE Trans. on Image Processing, 2016,25(2):630–642. [doi: 10.1109/TIP.2015.2507401]

[29] Hozjan V, Kačič Z. Context-Independent multilingual emotion recognition from speech signals. Int'l Journal of Speech Technology, 2003,6(3):311–320. [doi: 10.1023/A:1023426522496]

[30] Elbarougy R, Akagi M. Cross-Lingual speech emotion recognition system based on a three-layer model for human perception. In: Proc. of the Signal and Information Processing Association Annual Summit and Conf. (APSIPA). IEEE, 2013. 1–10. [doi: 10.1109/APSIPA.2013.6694137]

[31] Sagha H, Matejka P, Gavryukova M, Povolny F, Schuller B. Enhancing multilingual recognition of emotion in speech by language identification. In: Proc. of the Interspeech. 2016. 2949–2953. http://www.isca-speech.org/archive/Interspeech_2016/abstracts/0333.html

[32] Abdelwahab M, Busso C. Supervised domain adaptation for emotion recognition from speech. In: Proc. of the 2015 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015. 5058–5062. [doi: 10.1109/ICASSP.2015.7178934]

[33] Eyben F, Wöllmer M, Schuller B. Opensmile: The munich versatile and fast open-source audio feature extractor. In: Proc. of the 18th ACM Int'l Conf. on Multimedia. ACM, 2010. 1459–1462. [doi: 10.1145/1873951.1874246]

[34] Aytar Y, Vondrick C, Torralba A. Soundnet: Learning sound representations from unlabeled video. In: Advances in Neural Information Processing Systems. Miami Beach: Curran Associates, Inc., 2016. 892–900. http://papers.nips.cc/paper/6146-soundnet-learning-sound-representations-from-unlabeled-video.pdf

[35] Wu SZ, Kan M, He ZL, Shan SG, Chen X. Funnel-Structured cascade for multi-view face detection with alignment-awareness. Neurocomputing, 2017,221(C):138–145. [doi: 10.1016/j.neucom.2016.09.072]

[36] Barsoum E, Zhang C, Ferrer CC, Zhang Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proc. of the 18th ACM Int'l Conf. on Multimodal Interaction. New York: ACM, 2016. 279–283. https://dl.acm.org/citation.cfm?doid=2993148.2993165

[37] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014.

[38] Li Y, Tao J, Schuller B, Jia J. MEC 2016: The multimodal emotion recognition challenge of CCPR 2016. In: Proc. of the Chinese Conf. on Pattern Recognition. Singapore: Springer-Verlag, 2016. 667–678. [doi: 10.1007/978-981-10-3005-5_55]

[39] Dhall A, Goecke R, Joshi J, Hoey J, Gedeon T. Emotiw 2016: Video and group-level emotion recognition challenges. In: Proc. of the 18th ACM Int'l Conf. on Multimodal Interaction. ACM, 2016. 427–432. [doi: 10.1145/2993148.2997638]

附中文参考文献:

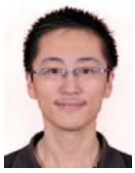
[1] 张石清.基于语音和人脸的情感识别研究[博士学位论文].成都:电子科技大学,2012.

[6] 韩文静,李海峰,阮华斌,马琳.语音情感识别研究进展综述.软件学报,2014,25(1):37–50. http://www.jos.org.cn/1000-9825/4497.htm [doi: 10.13328/j.cnki.jos.004497]

[24] 曹田熠.多模态融合的情感识别研究[博士学位论文].天津:天津大学,2012.

陈师哲(1994－),女,湖南邵阳人,博士生,CCF 学生会员,主要研究领域为多媒体语义内容分析.

金琴(1972－),女,博士,博士生导师,CCF 专业会员,主要研究领域为多媒体计算.



王帅(1993－),男,硕士生,CCF 学生会员,主要研究领域为多模态情感计算.