doi:10.3969/j.issn.1672-1292.2022.02.006

一种基于决策层融合的多模态情感识别方法

韩天翊[1,2],林荣恒[1,2]

(1.北京邮电大学计算机学院(国家示范性软件学院),北京 100876)
(2.北京邮电大学网络与交换技术国家重点实验室,北京 100876)

[摘要] 设计了一种软硬结合的多模态情感识别系统,使用语音和面部表情两个模态,通过梅尔频率倒谱系数与卷积神经网络对情感进行识别和分类,同时将语音情感识别迁移到神经网络计算棒以降低环境负载. 在模态融合时,采用决策层融合的方式来提高识别准确率. 实验结果表明,系统拥有较高的识别准确率,且能够在性能较差的运行环境中保持运行速度.

[关键词] 情感识别,卷积神经网络,软硬结合,多模态,决策层融合
[中图分类号]TP391 [文献标志码]A [文章编号]1672-1292(2022)02-0035-06

A Multimodal Emotion Recognition Method Based on Decision Level Fusion

Han Tianyi[1,2], Lin Rongheng[1,2]

(1.School of Computer Science(National Pilot Software Engineering School),
Beijing University of Posts and Telecommunications, Beijing 100876, China)
(2.State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: This paper designs a multimodal emotion recognition system that combines software and hardware. The system uses Mel-Frequency Cepstrum Coefficient and convolutional neural networks to recognize and classify emotions on speech and facial expressions. At the same time, emotion recognition of speech is transferred to neural network computing sticks to reduce the environmental load. In modal fusion, the method of decision-level fusion is used to improve the recognition accuracy. Experimental results show that the system has high recognition accuracy and can maintain running speed in the environment with poor performance.

Key words: emotion recognition, convolutional neural network, combination of software and hardware, multimodal, decision-level fusion

情感是一系列主观认知体验的总称,是一种心理和生理状态,是各种感觉和行为的结合. 情感通常与心情、性格等因素相互作用,也会受到激素的影响. 不同的情绪对日常行为会有着不同的导向作用,人们做的每件事都会有不同的情感表达. 了解用户使用产品时的情感,可以大大提高服务质量和服务效果.

近年来,随着科学技术的日益发达,情感识别的效率和准确性均有大幅提升. 情感识别系统能够帮助客服人员预先了解客户的情感状况,提高服务质量,进而提高服务效率和满意度. 但是,情感识别在单一模态下很难进一步提高识别准确率,又往往需要使用高性能的 GPU,这使得系统的使用范围很小,很难在现实生活中表现出应有的效果. 因此,有必要减少系统的使用条件,设计一个软硬件结合的多模态情感识别系统,并行加速计算过程,提高系统识别准确率,让系统可以在更广泛的范围内使用.

本文设计并实现了一种软硬结合的多模态情感识别方法,使用语音和面部表情两个模态,通过梅尔频率倒谱系数与卷积神经网络对情感进行识别和分类,同时将语音情感识别迁移到神经网络计算棒以降低环境负载,在模态融合时采用决策层融合的方式来提高识别准确率.

收稿日期:2021-08-31.
基金项目:江西省重点研发计划项目(20212BBE51002).
通讯作者:林荣恒,博士,副教授,研究方向:强化学习与生成对抗研究、云计算边缘计算、工业大数据分析、大数据与人工智能. E-mail: rhlin@bupt.edu.cn

1　相关工作

基于音频的情感识别出现于 1972 年. Ang 等使用决策树方法进行语音情感识别[1]. Lee 等通过结合词汇和话语来识别情感[2]. 目前广泛应用于语音情感识别的算法包括 SVM、长短期记忆网络等. Graves 等使用长短期记忆网络对音素进行分类和识别[3]. Eyben 等使用双向长短期记忆网络进行语音情感识别[4]. 陈闯等提出了一种基于改进遗传算法优化的 BP 神经网络进行识别(IAGA-BP)[5].

视频的情感识别主要依赖于使用者的面部表情. Ekman 等提出了一种通过划分面部区域的表情编码系统[6]. 此后,研究人员提出了许多识别面部表情的方法. Zhang 使用多层感知机作为分类器[7],Shan 等使用 LBP 特征进行面部情感识别[8]. 由于深度学习技术不需要在分类之前手动提取特征,深度学习技术在面部情感识别中得到了越来越广泛的应用. Ko 结合了针对单个帧的空间特征的卷积神经网络(CNN)和针对连续帧的时间特征的长短期记忆网络(LSTM)[9]. 谢非等提出了一种基于肤色增强和分块 PCA 的人脸检测及表情识别方法[10].

人的声音和表情可以直接反映人的情感. 在早期的研究中,情感识别往往使用单一模态数据进行,导致识别精度通常较低. 为了解决单数据识别精度差的问题,Kim 等提出了一种同时使用语音和视频两种模式的情感识别系统,将不同的模态数据一起使用可以提高系统识别的准确性[11]. Hossain 等提出了一种多方向回归的视听情感识别系统[12],使用 MDR 和基于脊波变换的特征,用极限学习机进行分类,获得的准确率为 83.06%. 闫静杰等提出了一种基于稀疏典型相关分析的特征融合方法,使用表情和语音数据获得了 60.09%的准确率[13]. Jiang 等提出了一种使用语音和面部特征来识别情绪的系统,使用三态流 DBN 模型作为分类器,在 eNTERFACE 数据库中,准确率为 66.54%[14]. 在 Kaya 等的研究中,语音信号中的音频特征、图像帧中的密集特征以及基于卷积神经网络的图像帧特征在评分级别上被融合以识别情绪[15].

2　基于硬件的语音情感识别方法

2.1　梅尔频率倒谱系数提取

梅尔频率倒谱系数是目前应用于情感识别中重要的声学特征. 考虑到人类的听觉特征,即人耳的灵敏度对频率的感知并不是线性的,将普通频率转换为梅尔频率,定义梅尔频率如下:

$$\mathrm{mel}(f)=2\ 595\times\log_{10}\left(1+\frac{f}{700}\right).$$

梅尔频率倒谱系数(Mel-frequency cepstral coefficients,MFCC)是在 Mel 标度频率域提取出来的倒谱参数,图 1 所示为梅尔频率倒谱系数提取流程.

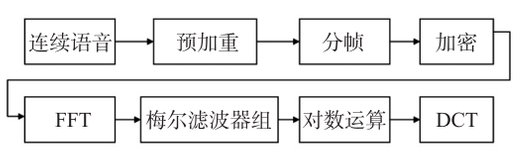

图 1　梅尔频率倒谱系数提取流程

Fig. 1　Extraction process of Mel spectrum

预加重是为了提高语言的高频部分,使频谱变得更加平滑,一般将语音信号通过一个高通滤波器:

$$H(z)=1-\mu z^{-1},$$

式中,$H(z)$为传递函数;μ 为预加重系数,取 $\mu=0.97$. 为了平滑信号,减弱频谱泄露,使用汉明窗对信号进行加窗处理:

$$W(n,a)=(1-a)-a\times\cos\left[\frac{2\pi n}{(N-1)}\right],0\leqslant n\leqslant N-1.$$

式中,a 的值不同会产生不同的汉明窗,一般情况下取 $a=0.46$.

经过一系列步骤,最终得到 MFCC 计算公式为:

$$C_{\mathrm{MFCC}}(j)=\sqrt{\frac{2}{N}\sum_{l=1}^{L}m_l\cos\left((l-0.5)\frac{\mathrm{j}\pi}{l}\right)},$$

式中,m_l 表示滤波器输出的对数;L 是滤波器的个数.

PyAudio 库可在系统中录制音频. 在录音过程中,需要设置每个缓冲区的帧数、采样位数、声道模式及采样频率. 在使用过程中,首先实例化 PyAudio 库,打开音频数据流并分别传入上述 4 个参数,同时调用 Wave 库,不断从音频数据流中读取 Chunk 写入文件流,最终得到一个完整的声音文件. Wavio 是一个

python 库,可以读取 wav 文件,在打开采集到的音频后,依次对音频对象进行预加重、分帧、FFT、梅尔滤波等操作,就可以获得音频数据的梅尔频率倒谱系数特征值.

2.2 卷积神经网络模型迁移

可使用 AlexNet 卷积神经网络进行音频的情感识别[16]. AlexNet 的网络结构包括 5 个卷积层、3 个池化层和 3 个全连接层. 多个卷积内核可以提取数据源中的不同特征. 在前两个卷积层后是最大池化层,之后第三、第四和第五卷积层直接相连,在第五个卷积层后是重叠的最大池化层,同时输出将进入全连接层. 全连接层可以为分类器提供标签.

为了满足情感识别过程的环境要求且并行加速计算过程,让系统可以在更广泛的范围内使用,需要将训练后的卷积神经网络运行在 Intel 神经网络计算棒上. 为此,需进行一些格式转换,使得系统可以成功运行. 神经网络计算棒的数据输入格式仅接受由 OpenVINO 工具套件的模型优化器生成的网络中间表示 IR,而模型优化器可以接收 ONNX 格式的神经网络,所以首先需将 AlexNet 神经网络生成对应的 ONNX 格式模型,之后通过模型优化器将 ONNX 格式转换为网络中间表示 IR. 在转换模型后,可在神经网络计算棒中使用推理引擎来预测收集的数据. 此外,可将多线程和异步操作用于预测,以提高神经网络计算棒的性能.

3 基于卷积神经网络的视频情感识别

3.1 基于 Haar 特征的人脸检测

OpenCV 中的 VideoCapture 类可用来从摄像头捕获视频或读取图像序列. 视频数据从摄像头拍摄后,每隔几帧将视频流中的图片进行人脸识别,使用 Haar 特征与级联分类器可以从图片中切割人脸表情[17].

Haar 特征是基于“块”的特征,一般分为 3 类:边缘特征、线性特征、中心特征和对角线特征. 特征模板内有白色和黑色两种矩形,并定义该模板的特征值为白色矩形像素和减去黑色矩形像素和,反映了图像的灰度变化情况. Haar 特征值的计算方法是将灰度化的图像分为黑色和白色两个区域,并计算白色区域 W 与黑色区域 B 的像素值之和的差值,乘以相应的权重系数 T,得到 i 区域 Haar 特征值:

$$C_{\text{Haar}}(i)=\left[\int_{W}p(x,y)\,\mathrm{d}x\mathrm{d}y-\iint_{B}p(x,y)\,\mathrm{d}x\mathrm{d}y\right]\times T_i.$$

积分图是一种只遍历一次图像就可以求出图像中所有区域像素和的快速算法,能够提高图像特征值计算的效率. 对于灰度图像中的任意一点,其积分图定义为:

$$\text{sum}(X,Y)=\sum_{x<X,y<Y}\text{image}(x-1,y-1),$$

其中,第 0 行和第 0 列为 0,image(x,y) 为点(x,y)处的原始灰度图.

Adaboost 是一种迭代算法,会挑选出一些最能代表人脸的矩形特征,按照加权投票的方式将多个弱分类器构造为强分类器. 将训练得到的强分类器串联可以组成一个级联结构的层叠分类器来提高分类器的检测速度. 得到识别出的人脸表情后,将切割到的人脸表情图片中最大的一个经如下转换后可以作为卷积神经网络的输入:

(1)从视频流找到的最大的人脸图片格式为 ndarray,首先将其转化成 PIL Image 类型;

(2)重新设定大小,将输入的 PIL Image 大小设置为 48;

(3)在 PIL Image 中心进行剪裁;

(4)随机水平翻转给定的 PIL Image,翻转概率为 0.5;

(5)将 PIL Image 格式转成 Tensor 格式,大小范围为[0,1].

在得到 Tensor 格式的数据后,可以将其导入卷积神经网络,对表情进行情感识别.

3.2 卷积神经网络结构

本文利用 TensorFlow 处理进行视频数据的情感识别. 视频识别的网络由 10 个卷积层、4 个池化层及 2 个全连接层组成. 前 4 个卷积层后是池化层,第五、六卷积层是直接相连的,之后紧跟一个池化层,与之类似,七、八卷积层与九、十卷积层后都紧跟一个池化层. 全连接层可以给分类器提供标签. 图 2 所示为视频模型的网络结构.

在预测当前小片段的类别时,可能识别结果与前序结果完全相反. 由于情感一般是连续和平稳的,在

大多数情况下不会发生突变,系统在识别情感时需要结合之前的识别结果.

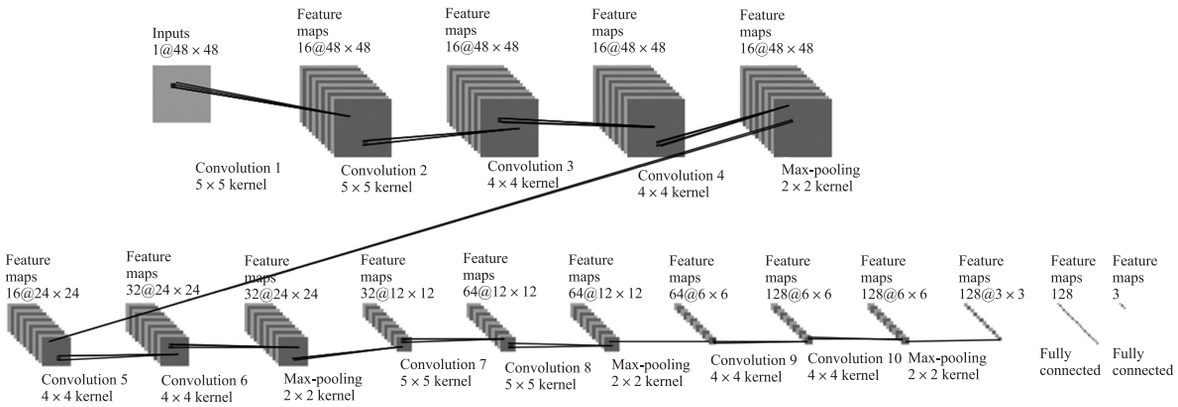

图 2　视频模型网络结构

Fig. 2　Network structure of video model

4　基于决策层融合的多模态情感识别

数据采集需要分别处理系统输入的语音流数据与视频帧数据. 从语音流中提取的梅尔频率倒谱系数首先被用于卷积神经网络进行训练,将训练后的网络迁移至神经网络计算棒,后续通过 OpenVINO 推理引擎进行情感识别. 对于视频帧数据,系统要通过 OpenCV 对帧进行人脸检测,将检测到的人脸输入卷积神经网络来识别情感. 系统的两种模式会同时进行工作,图 3 所示为情感识别系统整体的结构框图.

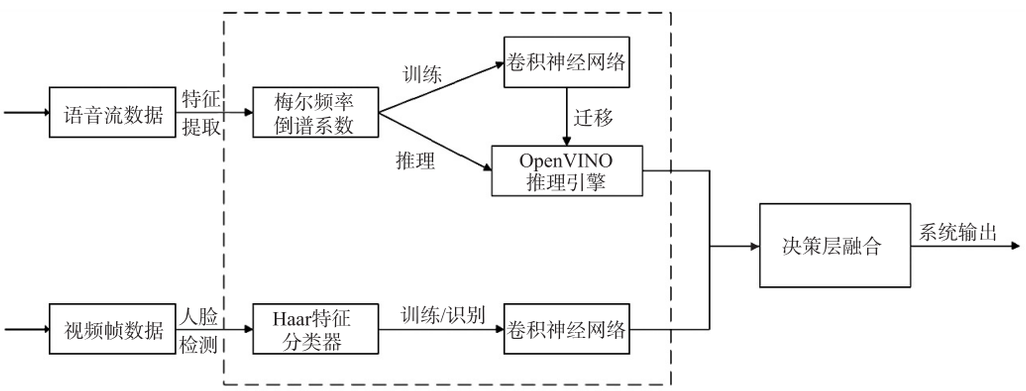

图 3　情感识别系统结构框图

Fig. 3　Module structure of the multimodal emotion recognition system

在单模态情感识别结束后,可得到分别由语音和视频识别的结果. 为了提高系统总体的准确率,需融合两种模态. 多模态融合通过不同模态之间的关联性,将多个角度的数据结合,来提高总体识别准确率. 决策层融合[18]是在提取不同模态的数据后,各自独立地进行情感分类,最后利用两种模态的结果得到多模态的识别结果. 本文中决策层融合采用加权求和的方式:

$$p = w_0 \cdot p^{\text{audio}} + w_1 \cdot p^{\text{video}},$$

式中,w_0 和 w_1 分别为语音和视频情感识别的权重,且 $w_0 + w_1 = 1$.

考虑到同一种感情不同人说话会有不同的表现形式,而人的表情并不会因为人的不同而相差甚远,因此对于视频数据情感识别的权重要高于语音数据. 图 4 显示了决策层融合的基本流程.

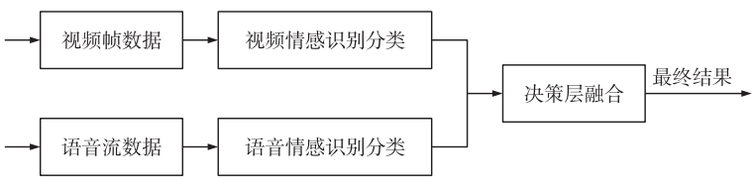

图 4　决策层融合基本流程

Fig. 4　The basic process of decision-level fusion

5 实验与结果

为了测试系统的有效性,多模态情感识别系统需检查每个模块的输入和输出,包括功能测试和非功能测试. 计算机的配置为 i7-8550u, mx150, 16G. 系统的 Python 开发环境为 Python3.7.3,系统版本为 Windows10 1903. 计算机中装有 OpenVINO R2020.2 套件,并已进行配置. 在此基础上设计了系统的测试用例,测试用例涵盖了每个模块的功能. 经测试显示,所有结果均与预期相同.

5.1 模型测试

系统对开心、生气与正常 3 种表情进行识别. 语音识别模型使用营业厅通话录音作为数据集. 数据集按照 8:2 的比例随机划分为训练集与测试集. 使用测试集对语音模型测试,其分类准确度为 74.5%. 同时,为了测试神经网络计算棒对系统性能的影响,选择了不同长度的音频文件进行测试.

表 1 显示了对于相同时长的音频在不同硬件上的运行时间,可以发现,神经网络计算棒的运行时间与高性能 CPU 相似,可以在性能较差的运行环境中保持运行速度.

视频模型的数据集结合了 Fer2013、CK+和 GENKI 数据集,训练集和测试集按 8:2 进行划分,使用支持向量机和深度神经网络作为基线算法. 实验结果如表 2 所示,卷积神经网络识别准确率为 78.62%,召回率 77.98%,优于支持向量机和深度神经网络的识别准确率. 图 5 显示了面部表情的混淆矩阵. 取 $w_0=0.3$, $w_1=0.7$,系统经过决策层融合,准确率提升了 3.4%. 系统测试效果如图 6 所示. 相比于文献[13]和[19],以及多流隐马尔可夫模型[14]和异步 DBN 模型[14],本文方法可以得到相似或更高的准确率.

表 1 不同硬件所需时间的对比

Table 1 Comparison table of time for different hardware

Duration of audio	Performance testing	
	CPU	Neural network computing stick
18 s	2 s	1 s
2 min 33 s	4 s	5 s
4 min	5 s	6 s

表 2 不同方法对情感的识别结果

Table 2 Recognition results of different methods

Method	Accuracy	Loss
SVM	53.3%	—
DNN	70%	0.738 6
CNN	78.62%	0.352 0

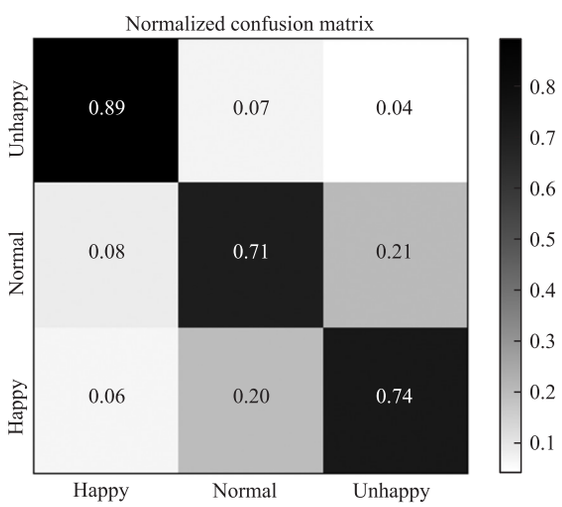

图 5 面部表情识别结果

Fig. 5 Recognition results of facial expressions

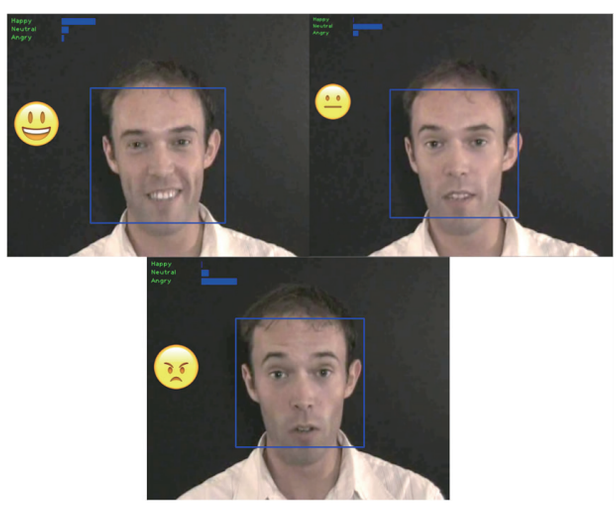

图 6 系统效果示意图

Fig. 6 The schematic diagram of the system

5.2 功能与非功能测试

为了测试系统的有效性,多模式情感识别系统需要检查每个模块的输入和输出,包括功能测试和非功能测试. 该系统基于浏览器/服务器架构模式,系统非功能测试主要对兼容性和连接速度进行测试. 兼容性测试是指系统对浏览器的支持. 经测试,该系统功能无明显异常,在主流浏览器中具有良好的兼容性,系统界面可以顺利打开,没有明显的延迟和滞后.

6　结论

针对传统情感识别系统准确率较低和往往需要高性能环境的问题,本文设计和实现了一种软硬结合的多模态情感识别系统以降低环境负载. 语音情感识别首先对每句语音进行语音信号预处理,提取语音的梅尔频率倒谱系数特征,将卷积神经网络迁移至神经网络计算棒对其进行情感识别分类. 视频情感识别使用 Haar 特征与级联分类器从图片中切割人脸表情,通过卷积神经网络对表情进行识别和分类. 实验结果表明,本研究采用的软硬结合的多模态情感识别方法具有较高的识别准确率,能够在性能较差的运行环境中保持运行速度.

[参考文献](References)

[1] ANG J,DHILLON R,KRUPSKI A,et al. Prosody-based automatic detection of annoyance and frustration in human-computer dialog[C]//Seventh International Conference on Spoken Language Processing. Denver,USA:DBLP,2002.

[2] LEE C M,NARAYANAN S S,PIERACCINI R. Combining acoustic and language information for emotion recognition[C]//Seventh International Conference on Spoken Language Processing. Denver,USA:DBLP,2002.

[3] GRAVES A, FERNÁNDEZ S, SCHMIDHUBER J. Bidirectional LSTM networks for improved phoneme classification and recognition[C]//International Conference on Artificial Neural Networks. Berlin,Germany:Springer,2005.

[4] 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,2010,3:7-19.

[5] 陈闯,CHELLALI R,邢尹. 改进遗传算法优化 BP 神经网络的语音情感识别[J]. 计算机应用研究,2019,36(2):344-346,361.

[6] EKMAN P,FRIESEN W V. Manual for the Facial Action Coding System[M]. Palo Alto:Consulting Psychologists Press,1978.

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

[8] SHAN C F,GONG S G,MCOWAN P W. Robust facial expression recognition using local binary patterns[C]//IEEE International Conference on Image Processing 2005. Genova,Italy:IEEE,2005.

[9] KO B C. A brief review of facial emotion recognition based on visual information[J]. Sensors,2018,18(2):401.

[10] 谢非,龚俊,王元祥,等. 基于肤色增强和分块 PCA 的人脸表情识别方法[J]. 南京师范大学学报(工程技术版),2017,17(2):49-56.

[11] KIM Y,LEE H,PROVOST E M. Deep learning for robust feature generation in audiovisual emotion recognition[C]//2013 IEEE International Conference on Acoustics,Speech and Signal Processing. Vancouver,Canada:IEEE,2013.

[12] HOSSAIN M S,MUHAMMAD G. Audio-visual emotion recognition using multi-directional regression and Ridgelet transform[J]. Journal on Multimodal User Interfaces,2016,10:325-333.

[13] 闫静杰,卢官明,李海波,等. 基于人脸表情和语音的双模态情感识别[J]. 南京邮电大学学报(自然科学版),2018,38(1):60-65.

[14] JIANG D M,CUI Y L,ZHANG X J,et al. Audio visual emotion recognition based on triple-stream dynamic bayesian network models[C]//International Conference on Affective Computing and Intelligent Interaction. Berlin,Germany:Springer,2011.

[15] KAYA H,GÜRPINAR F,SALAH A A. Video-based emotion recognition in the wild using deep transfer learning and score fusion[J]. Image and Vision Computing,2017,65:66-75.

[16] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM,2017,6(6):84-90.

[17] LIENHART R,MAYDT J. An extended set of Haar-like features for rapid object detection[C]//Proceedings of the International Conference on Image Processing 2002. Rochester,USA:IEEE,2002.

[18] LEE C M,NARAYANAN S S. Toward detecting emotions in spoken dialogs[J]. IEEE Transactions on Speech and Audio Processing,2005,13(2):293-303.

[19] DATCU D,ROTHKRANTZ L. Multimodal recognition of emotions in car environments[C]//Proceedings of the Driver Car Internation & Interface 2009. Prague,Czech:DCI&I,2009.

[责任编辑:严海琳]