面向电网行业的多模态心理评测算法

李华亮[1,2],梁泽权[3],黄星杰[3],刘羽中[1,2],吕建明[3,4]

(1. 广东电网有限责任公司 电力科学研究院;2. 广东电网有限责任公司 职业健康安全重点实验室,广东 广州 510062;
3. 华南理工大学 计算机科学与工程学院;4. 华南理工大学 大数据与智能机器人教育部重点实验室,广东 广州 510006)

摘 要: 为完成电网高危行业从业人员进行作业前的智能心理评测任务,提出面向电网行业员工的包含表情、声音、走姿的多模态心理评测算法。首先,构建电网行业员工数据集,从视频中抽取面部RGB图片序列、音频ComparE特征集及人体骨架关键点序列;其次,利用残差网络与双向长短时记忆网络提取面部视觉特征,在时间窗口提取音频特征,在时空图卷积网络提取步态特征,分别得到最优的单模态模型;最后,提出极性损失函数的深度学习训练方法及基于注意力机制的多模态融合算法,通过融合单模态模型输出特征获得最优多模态心理状态评测模型。实验表明,多模态融合相较于单模态系统能显著提升心理评测准度,对心理标签四分类任务的准确率达到65.66%,相较于基于面部表情、语音、步态3种单一模态的模型效果分别提升18.04%、21.22%和13.28%。

关键词: 多模态;深度学习;注意力机制;心理评测;电网行业

DOI: 10. 11907/rjdk. 231701 开放科学(资源服务)标识码(OSID):

中图分类号: TP391 **文献标识码:** A **文章编号:** 1672-7800(2024)009-0090-09

Multimodal Psychological Evaluation Algorithm for Power Grid Industry

LI Hualiang[1,2], LIANG Zequan[3], HUANG Xingjie[3], LIU Yuzhong[1,2], LYU Jianming[3,4]

(1. *Electric Power Research Institute, Guangdong Power Grid Co., Ltd*;
2. *Key Laboratory of Occupational Health and Safety, Guangdong Power Grid Co., Ltd, Guangzhou* 510062, *China*;
3. *School of Computer Science and Engineering, South China University of Technology*; 4. *Key Laboratory of Big Data and Intelligent Robot (South China University of Technology), Ministry of Education, Guangzhou* 510006, *China*)

Abstract: To solve the intelligent psychological evaluation task for high-risk industry employees in the power grid before work, a multimodal psychological evaluation algorithm is proposed for employees in the power grid industry, which includes expressions, sounds, and walking posture. Firstly, construct a dataset of employees in the power grid industry, extracting facial RGB image sequences, audio ComparE feature sets, and human skeleton keypoint sequences from videos; Secondly, residual networks and bidirectional long short-term memory networks are used to extract facial visual features, audio features are extracted in time windows, and gait features are extracted in spatiotemporal graph convolutional networks, respectively, to obtain the optimal single modal models; Finally, a deep learning training method for polarity loss function and a multimodal fusion algorithm based on attention mechanism are proposed to obtain the optimal multimodal psychological state evaluation model by fusing the output features of a single modal model. The experiment shows that multimodal fusion can significantly improve the accuracy of psychological evaluation compared with single-mode system, and the accuracy rate of the four classification tasks of psychological labels reaches 65.66%. Compared with the model based on facial expression, voice, and gait, the effect of multimodal fusion is increased by 18.04%, 21.22%, and 13.28% respectively.

Key Words: multimodality; deep learning; attention mechanism; psychological evaluation; power grid industry

收稿日期: 2023-07-23

扫描二维码阅读全文:

基金项目: 南方电网科技项目(GDKJXM20200484)

作者简介: 李华亮(1983-),男,广东电网有限责任公司电力科学研究院高级工程师,研究方向为电力环境保护及职业健康安全;梁泽权(1998-),男,华南理工大学计算机科学与工程学院硕士研究生,研究方向为深度学习与多模态融合;黄星杰(1997-),男,华南理工大学计算机科学与工程学院硕士研究生,研究方向为深度学习与人体健康安全;刘羽中(1989-),男,广东电网有限责任公司电力科学研究院工程师,研究方向为职业健康;吕建明(1981-),男,博士,华南理工大学计算机科学与工程学院教授、博士生导师,研究方向为数据挖掘、计算机视觉、机器学习、分布式计算。本文通讯作者:梁泽权。

0　引言

人类日常生活状态在很大程度上依赖当前的心理情绪条件，在一些类似电网公司的高危行业中，为了尽可能保护工作人员免受心理状态的困扰，降低工作意外发生的风险，在工作前适当进行心理评测十分必要[1,2]。在面对大量员工需要心理评测的场景，为避免人力资源消耗，可采取人工智能技术根据员工的举止现状自动评测员工心理状态，以判断此时员工是否适合上场工作。研究表明，面部表情、语音和步态 3 种模态是心理评判的重要依据[3-5]。

通过面部表情进行心理或情感识别的研究已有近 40 年历史。Mylonakis 等[6]在 20 世纪基于人脸区域提出适合表情识别的面部动作编码系统（Facial Action Coding System，FACS）。Tran 等[7]通过面部表情序列识别心理情绪并与静态表情识别进行比较。罗元等[8]针对局部导数模式（Local Derivative Pattern，LDP）[9]的人脸编码方案进行算子优化提出了 nLDP 算子，并利用传统主成分分析（Principal Components Analysis，PCA）[10]和支持向量机（Support Vector Machines，SVM）[11]技术进行表情分类。Jabid 等[9]利用隐马尔可夫模型对表情的动态序列进行研究。随着深度学习不断发展，不少利用卷积神经网络（Convolutional Neural Network，CNN）提取表情特征的算法相继被提出[12]。Tran 等[13]利用三维卷积网络（Convolutional 3D，C3D）[14]卷积分析视频的表情序列。Ekman 等[6]结合 CNN 与循环神经网络（Recurrent Neural Networks，RNN）[15]进行堆叠和训练以提取视觉特征。

语音情感内容对频谱能量在各个频谱区间的分布有着明显影响，并在语音识别领域广泛使用，其他特征显示形式包括共振峰、梅尔频率倒谱系数（Mel Frequency Cepstral Coefficients，MFCC）、Teager 能量算子。Sato 等[16]通过分析语音的 MMFC 特征识别情感，根据 MMFC 统计学整理出更高阶的 eGeMAPS[17]和 ComparE[18]特征集。在深度学习中，通常利用较为广泛的长短时记忆网络（Long short-term memory，LSTM）[19]和双向 BiLSTM[20]等识别语音序列。Huang 等[21]利用 CNN 从音频中提取特征。Trigeorgis 等[22]将 CNN 与 LSTM 结合的端到端模型运用于情感识别情景，在 RECOLA 数据集中取得了不错的准确率。Badshah 等[23]提取语谱图作为神经网络输入，以在 Emo-DB 数据集中提升模型的识别准确率。

虽然人体姿态识别心理状态的研究较为分散，但可借鉴以往关于动作识别的研究。Yan 等[24]利用图卷积结合时空特征（Spatial Temporal Graph Convolutional Networks，ST-GCN）识别人体骨架，利用了人体姿态和步态识别。Bhattacharya 等[25]基于 ST-GCN 搭建的对抗生成网络所产生的数据集 Gait，提出步态情感识别的半监督分类神经网络。

目前，关于心理评测研究主要存在以下缺陷：①仅限于面部表情或语音的单模态，未综合考虑多模态的互补性；②简单提取与拼接模态特征，忽略了注意力机制的重要性；③忽略了不同心理标签间差异和下游分类任务的情景特殊性。

为此，本文主要面向电网行业场景进行多模态心理评测的算法探讨，所作贡献如下：①提出多模态心理评测方法框架，基于 CNN 与 BiLSTM 提取表情、声音 ComparE 特征，采用 ST-GCN 提取步态特征，通过融合 3 个不同的模态特征有效评测员工心理；②混合深度网络信息提出了基于注意力机制的多模态融合算法，并在此基础上增强了对心理分析结果有利的数据通道；③提出极性损失函数，在心理上拉开主观性质不同的标签组，使模型在电网心理数据集上收敛效果更优；④收集并整理 558 个不同员工的视频数据集（演讲视频和走姿视频），为心理评测算法奠定数据基础。

1　相关工作

近几年，多模态融合技术是深度学习的热点，在多个领域有着广泛研究与应用。多模态技术主要包括特征层融合、决策层融合、混合多模态融合。其中，特征层融合指在早期将声音、图像等多模态串联得到总的特征向量后输入模型；决策层融合分别将不同模态输入对应的分类器中得到结果，再把不同的决策进行融合；混合多模态融合则结合前两者。Wollmer 等[26]通过 BiLSTM 对提取的音频视频进行特征层融合，然后依据 ASR 系统将 MFCC 生成 BoW/BoNG 的语言特征，最后基于 SVM 进行分类以进行决策层融合。Boxuan 等[27]提出基于生理和面部表情的多模态情感识别框架提高了模型识别性能。Hossain 等[28]首先在频域处理语音得到梅尔图，再结合两个连续的极端学习机融合两个 CNN 的输出，使该系统得到了不错的效果。

目前，多模态融合应用在心理评测的研究主要集中在多模态融合技术上的多种尝试与在模态选择上的泛化，在面部表情、语音、文字、生理性状等方面的模态特征，或在模态缺失、模态相互影响与融合策略等方向寻求突破。早在 20 世纪就有学者提出注意力机制的相关概念，即生成注意力权重决定模型关注数据的哪些部分，以增强对结果的增益部分，抑制负面效果的数据区域，不仅能提升模型性能，还可作为空间或时序上的可解释性表现。例如，Mnih 等[29]在深度学习的视觉领域中运用注意力，在 RNN 模型上通过注意力机制对图像进行分类。Woo 等[30]提出卷积块注意力模块（Convolutional Block Attention Module，CBAM），在 CNN 框架的时间、空间维度引入注意力机制。

2 电网行业数据集预处理

2.1 数据集

本文数据集一共采集了558名员工信息，每位员工包括一个3 min带音频的演讲视频和一个不定长的走姿视频，包含沉稳型、冲动型、快而准确型、慢而不准确型4种心理状态之一。其中，心理标签是在相继拍摄完演讲视频和走姿视频后，对员工进行问卷调查的得分判断而来。因此，员工演讲时在表情、语音中的表现及随后的步态表现都与心理标签存在一定联系。

在每段演讲视频中，员工面对摄像头脱稿自由演讲，演讲内容包括自我介绍、家乡介绍、工作报告等。在走姿视频中，员工面向或背向摄像头往前后方向作出重复走路的动作，由此可从演讲视频获得每位员工的面部表情、音频两个对齐的模态信息，从走姿视频获得其步态信息。图1(彩图扫OSID可见，下同)分别为演讲视频和走姿视频的某一帧图像。

Fig. 1 Example of speech and walking posture

图1 演讲与走姿示例

2.2 数据预处理

2.2.1 面部表情

Kazemi等[31]提出级联残差回归树，搭建级联回归树实现人脸区域对齐算法，以训练出可精确沿人脸器官的重要轮廓标记出关键点的模型结构。该结构的预训练模型被收录于C++的机器学习开源库dlib中，本文将直接使用此预训练模型，每隔3 s抽取演讲视频人脸区域，并标记68个关键点(见图2)，实际提取结果如图3所示。

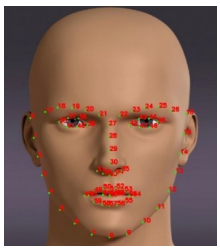

Fig. 2 68 face markers

图2 人脸68个标记点

将演讲视频批处理到面部表情数据后，每位员工将有60个256×256大小的面部表情图像，并由每张图像产生出68个关键点坐标，其中RGB图像为后续深度神经网络的输入数据特征。假设人脸区域在原图的像素长宽分别为H、W，则每个关键点的坐标集合可记为

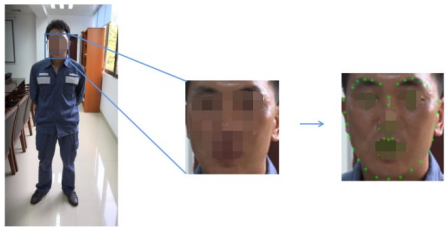

Fig. 3 Facial expression preprocessing results

图3 面部表情预处理结果

$\{(x_i, y_i) | i \in \{0, 1, ..., 67\}\}$，位于鼻头的关键点可作为中心关键点对其他关键点坐标进行归一化与标准化，每个关键点P_i坐标转换如式(1)所示。

$$P_i = \left(\frac{x_i - x_{30}}{W}, \frac{y_i - y_{30}}{H}\right) \tag{1}$$

将所有归一化坐标组成的向量记作$\boldsymbol{F}$，作为后续传统机器学习的输入数据特征。由于人脸轮廓与五官比例因人而异，可将坐标信息转化为距离和角度信息，计算关键点间距离占比和不同区域间角度排除人脸外貌差异带来的干扰[32]。然后，将所有转化后的距离和角度组成的向量记作$\boldsymbol{F}'$，作为后续传统机器学习的输入数据特征。

2.2.2 语音数据

本文从每段演讲视频中获取3 min的语音，先把这段语音切割成共60个3 s的语音小段，再对每小段语音进行信息预处理。在每小段语音中，最直观的两个特点是随时间变化的整体音量与音高，即振幅与频率，但更多的语音信息往往被蕴含在不同频率区域的不同能量表现中。由于无论多复杂的语音波形都可通过不同频率、振幅的正弦波叠加而来。因此，不少语音往往都通过快速傅里叶变换(Fast Fourier Transform，FFT)，将时域语音波形拆解成不同频率区域中的声波能量，从而将声音的时域信息转化为频域信息。

为此，本文采取Python的音频处理库librosa处理每小段3 s的语音，得到一张496×369大小的声谱图，如图4所示。其中，水平方向为时间轴，竖直方向为频率轴，在声谱图中的某个像素点(t，w)的颜色表示在时刻t中w Hz频率域上的音量，如图4右边颜色条所示，颜色越深表示该点音量越高。然后，使用CNN模型对声谱图进行特征提取，作为后续深度神经网络的输入数据特征[33]。

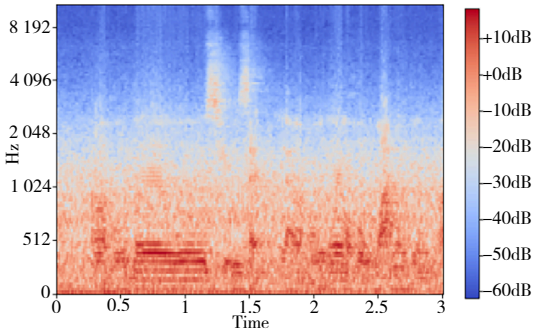

Fig. 4 Spectrogram

图4 声谱图

除了声谱图外，本文还借用opensmile工具分别提取每个小段语音的eGeMAPS和ComparE的特征集。这两个特征集都是根据各种数据统计方法，从语音的MFCC特征分析中提取得到的经典手工语音特征集，包括MFCC的几个重要系数及其共振峰中心和方差等[34]。其中，每小段语音的eGeMAPS特征为88维向量，ComparE特征为6 373维向量，eGeMAPS向量记作$\boldsymbol{V}$，ComparE向量记作$\boldsymbol{V}'$，作为后续传统机器学习的输入数据特征。

2.2.3　走姿骨架

对于步态视频，可通过开源库Openpose[35]中的姿态关键点标记模型对每帧图片的人体进行姿态识别，得到25个关键点坐标，如图5所示。

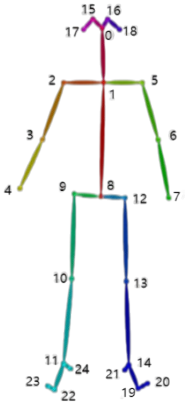

Fig. 5　Body skeleton

图5　人体骨架

假设第i个点坐标为(x_i, y_i)，记两点间的欧式距离为$d(i,j)$，基于椎骨距离$d(1,8)$和臀部点进行坐标归一化，归一化后的25个关键点M_i坐标如式(2)所示。

$$M_i = \left(\frac{x_i - x_8}{d(1,8)}, \frac{y_i - y_8}{d(1,8)}\right) \tag{2}$$

然后，将所有归一化坐标组成的向量记作$\boldsymbol{P}$，作为后续传统机器学习以及深度神经网络的输入数据特征。由于步态特征蕴含的心理状态信息，可能较为集中在四肢做类似圆摆运动的频率信息上，其中上肢主要以颈部为支点在运动，下肢则围绕臀部运动。为此，基于这两个关键支点，根据式(3)—式(6)提取出其他关键点与支点之间的距离信息D_i、角度信息A_i，作为圆摆运动的代表特征，从而将骨架中的25个关键点数据从二维向量(x_i, y_i)转换为另一种二维向量(D_i, A_i)。

$$D_i = \frac{d(i,1)}{d(1,8)} (i \in B_1) \tag{3}$$

$$D_i = \frac{d(i,8)}{d(1,8)} (i \in B_2) \tag{4}$$

$$A_i = \begin{cases} \arctan\dfrac{y_i - y_1}{x_i - x_1} (x_i \neq x_1) \\ \dfrac{\pi}{2} (x_i = x_1, y_i > y_1) \\ -\dfrac{\pi}{2} (x_i = x_1, y_i < y_1)(i \in B_1) \end{cases} \tag{5}$$

$$A_i = \begin{cases} \arctan\dfrac{y_i - y_8}{x_i - x_8} (x_i \neq x_8) \\ \dfrac{\pi}{2} (x_i = x_8, y_i > y_8) \\ -\dfrac{\pi}{2} (x_i = x_8, y_i < y_8)(i \in B_2) \end{cases} \tag{6}$$

假设归一化后的坐标集合为$\boldsymbol{M}$，转化后的向量集合为$\boldsymbol{M}'$。B1、B2区域所包含的关键点如表1所示，分别对上肢B1和下肢B2采取不同的坐标转化，所有转化后的距离和角度组成的向量记作$\boldsymbol{P}'$，作为后续传统机器学习以及深度神经网络的输入数据特征。

Table 1　Key point collection in the upper and lower limb regions

表1　上、下肢区域关键点集合

上下肢区域	关键点集合
B1	0-7，15-18
B2	8-14，19-24

3　心理评测算法设计

3.1　单模态心理评测算法

3.1.1　基于传统机器学习的心理评测算法

预处理数据集后可整理得到数据特征集合类型(见表2)，所有算法与实验将基于此表中的预处理数据集。其中，$\boldsymbol{S} \in \{\boldsymbol{F}, \boldsymbol{F}', \boldsymbol{V}, \boldsymbol{V}', \boldsymbol{P}, \boldsymbol{P}'\}$表示不同的数据特征集合。根据上述数据流类型，基于MFE-PCA-SVM的心理评测算法步骤如下：

步骤1：利用特征提取层(Manual Feature Extractor，MFE)提取在时间窗口上如均值、方差、偏度、峰度、皮尔相关系数、傅里叶变换等特征，然后拼接得到长度为s的手工特征向量。

步骤2：利用PCA算法将手工特征向量降到适合长度$k<s$的PCA特征向量，以降低数据冗余。

步骤3：利用机器学习的支持向量机(Support vector machine，SVM)模型对PCA特征向量进行分类训练，其中SVM选择径向基函数作为核函数，并选择适当的惩罚函数训练SVM。具体算法流程如图6所示。

Table 2　Types of data characteristics after preprocessing

表2　预处理后的数据特征类型

模态	特征类型
面部表情	图像、归一化坐标向量$\boldsymbol{F}$、转化向量$\boldsymbol{F}'$
语音	图像、eGeMAPS向量$\boldsymbol{V}$、ComparE向量$\boldsymbol{V}'$
步态	归一化坐标向量$\boldsymbol{P}$、转化向量$\boldsymbol{P}'$

3.1.2　基于深度神经网络的心理评测算法

由于面部表情和声谱图RGB图可通过残差网络(Residual Network，ResNet)进行图像识别，而步态则是一种稀疏的拓扑图，适合使用ST-GCN进行特征识别。其中，基于RGB图像集合进行深度神经网络的心理评测算法步骤为：

步骤1：图像像素归一化。假设面部表情或语音RGB

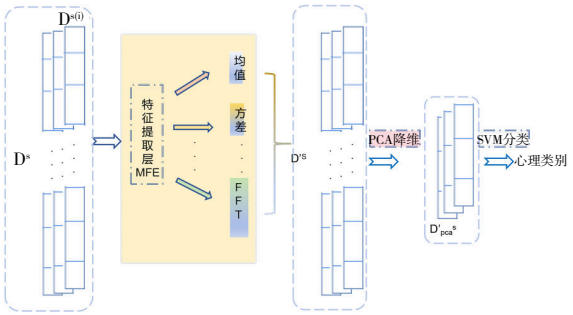

Fig. 6 MFE-PCA-SVM psychological evaluation algorithm based on traditional machine learning

图6 基于传统机器学习的MFE-PCA-SVM心理评测算法

图像序列为 $\boldsymbol{L}^{S(i)}$，$\boldsymbol{L}^{S(i)}\in\boldsymbol{R}^{t\times c\times w\times h}$，其中 $S(i)$ 表示第 i 个 S 模态的样本，t 为序列长度，w、h 分别为图像的宽长，c 为通道数，本文 c=3。每个通道特征值在0-255范围内，在输入网络前都除以255进行归一化。

步骤2：ResNet图像识别模型的训练。记训练集为 $\boldsymbol{L}^{S}$，拆开 $\boldsymbol{L}^{S}$ 每一个图像序列 $\boldsymbol{L}^{S(i)}$ 的每一帧，记第 j 帧的单张图片为 $L^{S(i)}_{j}$，将属于训练集 L^{S} 中任意的 i，j 图片随机打乱并组合成一个新的训练图片集 $\boldsymbol{M}$，M_i 表示第 i 张图片，$M_i\in\boldsymbol{R}^{c\times w\times h}$，基于普通的ResNet18卷积深度学习 $\boldsymbol{M}$，其中ResNet18后两层为自定义的全连接层和softmax分类层。

步骤3：引入注意力（Attention）机制。本文借鉴CBAM模型将在网络最后一个卷积层后加入Attention机制，分别为基于空间的注意力和基于通道的注意力[30]。然后，假设最后一个卷积层输出的特征为 $F\in\boldsymbol{R}^{c\times w\times h}$，$F$ 可看作由 c 个 $w\times h$ 的特征图的组合。

在基于空间的注意力机制中，首先对所有特征图的同一个位置分别进行最大池化和平均池化，得到两张特征图 $\boldsymbol{F}_{max}$，$\boldsymbol{F}_{avg}\in\boldsymbol{R}^{1\times w\times h}$，两者合成为 $\boldsymbol{R}^{2\times w\times h}$ 并使用卷积层和Sigmoid函数进行处理，以学习到一张空间注意力过滤图 $\boldsymbol{\alpha}_{spatial}\in\boldsymbol{R}^{w\times h}$；然后用 $\boldsymbol{\alpha}_{spatial}$ 中每个位置的注意力值与原本特征 F 中每张特征图上的对应位置值相乘，得到经过空间注意力处理完的 $\boldsymbol{F}_{spatial}$。具体计算过程如式（7）、图7所示。

$$\begin{aligned}\boldsymbol{\alpha}_{\text{spatial}}&=\text{Sig}(\text{Conv}(\text{Maxpool}(\boldsymbol{F}),\text{Avgpool}(\boldsymbol{F})))\\ \boldsymbol{F}_{\text{spatial}}&=\boldsymbol{\alpha}_{\text{spatial}}\otimes\boldsymbol{F}\end{aligned}\tag{7}$$

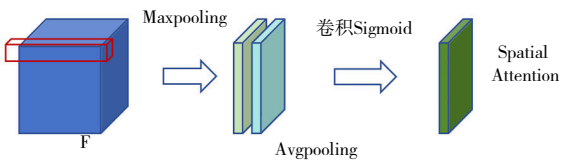

Fig. 7 Spatial attention

图7 空间注意力

在基于通道的注意力机制中，首先对每一张特征图进行最大池化和平均池化，得到 $\boldsymbol{F}_{max}$，$\boldsymbol{F}_{avg}\in\boldsymbol{R}^{c\times1\times1}$，压缩到 $\boldsymbol{R}^{c}$ 后将其分别输入多层感知机（Multiple Layer Perceptron，MLP）中学习输出 $\boldsymbol{F'}_{max}$，$\boldsymbol{F'}_{avg}\in\boldsymbol{R}^{c}$。然后，将两个输出沿着通道与对应的值相加并输入Sigmoid函数，形成通道注意力向量 $\boldsymbol{\alpha}_{channel}\in\boldsymbol{R}^{c}$；最后，将原本特征 $\boldsymbol{F}$ 中的每张特征图上的所有值与 $\boldsymbol{\alpha}_{channel}$ 对应通道中的注意力值相乘，得到经过通道注意力处理完的 $\boldsymbol{F}_{channel}$。具体计算过程如式（8）、图8所示。

$$\begin{aligned}\boldsymbol{\alpha}_{\text{channel}}&=\text{Sigmoid}(\text{MLP}(\text{Maxpool}(\boldsymbol{F})),\text{MLP}(\text{Avgpool}(\boldsymbol{FF})))\\ \boldsymbol{F}_{\text{channel}}&=\boldsymbol{\alpha}_{\text{channel}}\otimes\boldsymbol{F}\end{aligned}\tag{8}$$

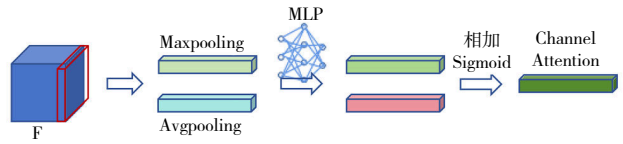

Fig. 8 Channel attention

图8 通道注意力

图9为数据流经过ResNet和Attention机制的总体过程，每个图片样本将从 $\boldsymbol{M}_i\in\boldsymbol{R}^{c\times w\times h}$ 变为一维特征向量 $\boldsymbol{M'}_i\in\boldsymbol{R}^{c''}$。

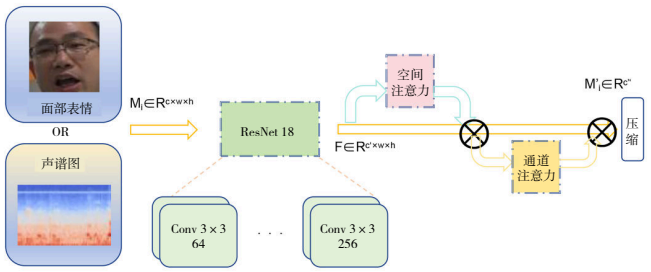

Fig. 9 ResNet-Attention algorithm

图9 ResNet-Attention算法

步骤4：重新定义极性损失函数 $loss_p$。在训练ResNet的过程中，通常在最后的softmax层中使用交叉熵损失函数作为模型收敛的指标，原因为心理评测任务具有较强的主观性。本文将基于标签的差异对损失函数进行修改与添加，实验发现在心理评测标签分类中沉稳型与冲动型、快而准确型与慢而不准确型两组心理状态，在所有组合中主观差异最明显。若预测标签和真实标签的比较属于这两组则应引入权值 α 更新损失函数，并将新的损失函数定义为极性损失函数 $loss_p$，如式（9）所示。

$$loss_p=-\frac{1}{N}\sum_{i=1}^{N}(1+\alpha(g(\overline{y}_i,y_i)))\sum_{c=1}^{C}1_{[c=y_i]}\log_2 p_{i,c}\tag{9}$$

$$g(\overline{y}_i,y_i)=\begin{cases}1 & (\overline{y}_i,y_i)\in\{(0,1),(1,0),(2,3),(3,2)\}\\ 0 & otherwise\end{cases}$$

式中：C 表示类别数量；N 表示训练样本数量；$p_{i,c}$ 表示第 i 个样本预测结果属于 c 类别的概率。

本文在图9的向量压缩后接上分类层，按新定义的 $loss_p$ 训练完ResNet18，取训练集和预测集的准确率综合最好的模型作为图片特征的提取模型，作为后面时序分析的输入。

步骤5：Bi-LSTM时序学习模型的训练。将每个视频中某个模态的图片帧特征 $\boldsymbol{M'}_i\in\boldsymbol{R}^{c'}$ 的序列组合记为 $\boldsymbol{T}_j\in\boldsymbol{R}^{t\times c'}$，表示第 j 个视频的某个模态特征序列，其中 c' 为ResNet去

除分类层后输出的图像特征的长度。经过双层 Bi-LSTM 处理，将最后一次的输出 $\boldsymbol{T'}_j^t \in \boldsymbol{R}^{c'}$ 作为特征序列在时间上的总体特征，压缩成一维向量后再进行 softmax 层分类。其中，Bi-LSTM 利用 $loss_p$ 进行训练。在面部表情和声谱图两个模态中基于深度神经网络的心理评测算法进行，总体上也是 ResNet-Attention-Bi_LSTM 的流程，如图 10 所示。

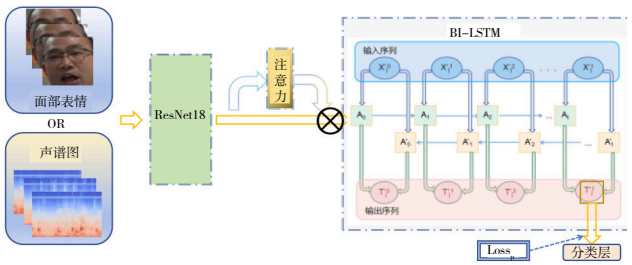

Fig. 10　ResNet-Attention-Bi_LSTM algorithm

图 10　ResNet-Attention-Bi_LSTM 算法

首先，记步态数据流 $\boldsymbol{P} \in \boldsymbol{R}^{c \times t \times v}$，$t$ 为时序长度，v 为步态图关键点的个数，c 为每个点包含的数值个数，取 c=2，表示点坐标或距离角度组合的有序数对；其次，在步态图上建立邻接矩阵 $\boldsymbol{A}$，将 $\boldsymbol{P}$、$\boldsymbol{A}$ 输入 ST-GCN。其中，ST-GCN 具有多层基本模块，在每层模块中 $\boldsymbol{P}$、$\boldsymbol{A}$ 经过 Attention 层，利用其参数在步态躯干的边权重上训练出注意力值与 $\boldsymbol{A}$ 相乘得到 $\boldsymbol{A}'$；再次，经过 GCN 层对 P 上的点邻居集合，按给定的子集划分方式进行卷积后乘 $\boldsymbol{A}'$；最后，经过 TCN 层在时间维度上对每个点进行卷积。

$\boldsymbol{P}$、$\boldsymbol{A}$ 依次经过 ST-GCN 中堆叠的基本模块，输出最终的 c' 层特征图 $\boldsymbol{P'} \in \boldsymbol{R}^{c' \times 1 \times v}$，压缩成一维向量后再进行 softmax 层的分类也利用 $loss_p$ 进行训练。在步态的模态中，基于深度神经网络的心理评测算法总体上是 ST-GCN 的流程，如图 11 所示。

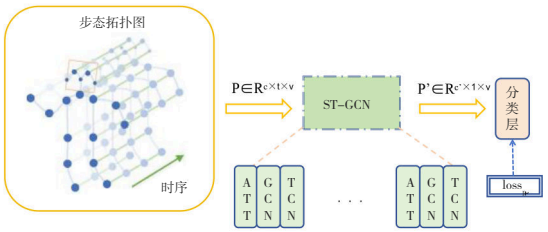

Fig. 11　Gait ST-GCN algorithm

图 11　步态 ST-GCN 算法

3.2　多模态融合心理评测算法

3.2.1　多模态决策层融合的心理评测算法

每个模态在各自的单模态算法中都对员工作出了心理评测，在此基础上综合 3 种评测分数，在最后的决策层决定总体分数最高的类别为员工最终评测出的心理状态。本文取各单模态中心理评测算法效果最好的模型，将每个模态数据分别输入对应的最好模型 S 中得到各类预测结果 $p^S{}_c$，即 c 类别心理的预测概率。定义 $p^{S(i)}{}_{c(j)}$ 为第 i 种模态的第 j 类心理预测概率，则在决策层的融合规则如式（10）所示。

$$c = argmax(\{\frac{1}{3}\sum_{i=0}^{3} p_{c(j)}^{S(i)} | j \in \{0,\ 1,\ 2,\ 3\}\}) \tag{10}$$

首先平均每个模态的预测概率分类，再取其平均概率的最大值，最大值所属的心理类别 c 即为多模态决策层融合的心理评测结果，如图 12 所示。

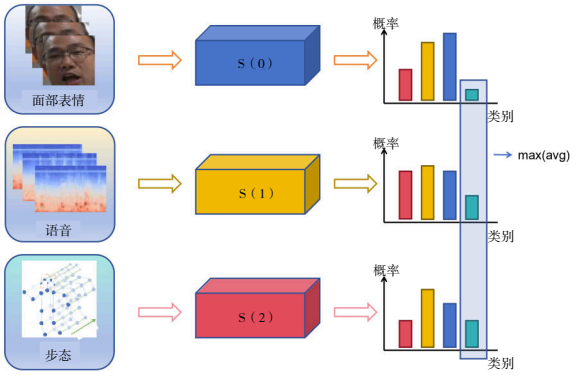

Fig. 12　Multimodal decision layer fusion algorithm

图 12　多模态决策层融合算法

3.2.2　多模态特征层融合的心理评测算法

本文基于注意力机制的多模态特征层融合算法，除了等待每个模态模型给出评测分数再与决策层综合判断外，也可在单模态的心理评测算法过程中提取中间某一步模态特征，与其他模态特征进行融合或交互，再输入总体模型得到最终的心理状态类别。为了保持各模态心理评测算法模型的完整性与连贯性，本文将在各模态算法最后的 softmax 分类层前参考注意力机制对 3 个一维向量特征进行融合。

记模型 $S(i)$ 的一维向量为 $\boldsymbol{Z}^{S(i)} \in \boldsymbol{R}^{c(i)}$，其中 $c(i)$ 为各自模态的向量长度，直接将 3 个 $\boldsymbol{Z}^{S(i)}$ 前后拼接得到 $\boldsymbol{Z} \in \boldsymbol{R}^{c}$，$c$ 为 $c(i)$ 的和。基于拼接向量 $\boldsymbol{Z}$ 分别作 3 种不同的 MLP_i 得到 3 个对应 $c(i)$ 长度的注意力向量 $\boldsymbol{\beta}(i) \in \boldsymbol{R}^{c(i)}$，再分别与 $\boldsymbol{Z}^{S(i)}$ 乘积得到 $\boldsymbol{Z'}^{S(i)} \in \boldsymbol{R}^{c(i)}$，最后前后拼接 3 个 $\boldsymbol{Z'}^{S(i)}$ 得到经 attention 层的总体一维向量特征 $\boldsymbol{Z'} \in \boldsymbol{R}^{c}$。如式（11）、图 13 所示。

$$Z' = concat(\{Z^{S(i)} \otimes (MLP_i(Z)) | i \in \{0, 1, 2\}\}) \tag{11}$$

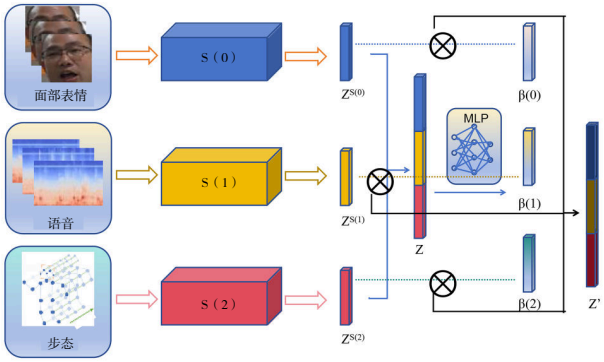

Fig. 13　Multimodal feature layer fusion algorithm

图 13　多模态特征层融合算法

在融合3个模态的特征层后，将Z'进行softmax层分类得到最终结果。

4 实验结果与分析

4.1 实验数据集与环境

实验数据为上述从电网公司员工身上采集到的视频数据，每个员工为一个样本，样本数据类型包括员工标签、演讲视频、步态视频、心理状态标签。两个短视频信息通预处理后可分别得到3种模态的数据信息，包括面部表情图像、语音音频、步态关键点序列。表情和语音从3 min的演讲视频中间隔3 s划分而得；步态关键点从步态视频中识别得到。具体数据集属性如表3所示，样本数量为有效的无模态缺失的数据数量。

实验在Linux显卡服务器环境中进行，算法设计主要使用Python语言、sklearn和pytorch框架实现，深度学习优化器为Adam，学习率Lr=0.001，L2正则化权重0.000 1，batch=64，训练批次为100，取测试集准确率最优结果。

Table 3 Properties of grid datasets

表3 电网数据集的属性

属性	具体数值
视频样本数	558
原始数据	3分钟演讲视频+不定长走姿视频
心理标签	4(沉稳、冲动、快而准确、慢而不准确)
模态数	3(表情、语音、步态)

4.2 单模态心理评测算法的实验分析

4.2.1 基于传统机器学习的算法

基于表2部分预处理数据特征集合$S\in\{F, F', V, V', P, P'\}$，采用基于传统机器学习的MFE-PCA-SVM心理评测算法分别对S中的每个集合进行实验，通过调整PCA压缩的维数k得到相应的预测集准确率，如表4所示。

Table 4 Dimensionality reduction length k and the prediction set accuracy of each single mode on traditional machine learning algorithms

表4 降维长度k和各单模态在传统机器学习算法上的预测集准确率 (%)

维度k / 模态		16	32	48	60	原维度s
表情	F	34.92	39.68	38.10	41.27	25.40
	F'	36.51	28.57	30.16	30.16	34.92
语音	V	33.33	39.68	42.86	44.44	44.44
	V'	38.10	34.92	39.68	31.75	36.51
步态	P	34.92	28.57	25.40	22.22	38.10
	P'	46.03	44.44	42.86	46.03	34.92

由表4可知，面部表情的归一化坐标向量F、步态的转化向量P'在经过PCA处理后相较于原长度的结果明显更好，但其他类型PCA效果一般。在面部表情模态上，未经过人脸区域转化的归一化坐标向量F相较于转化后的F'表现更好。在语音模态上，原维度更高的ComparE特征集相较于eGeMaps特征集表现更好。在步态模态上，经过距离—角度转化后的向量P'的效果相较于未转化的归一化坐标向量P更好。实验发现，每个模态基于传统机器学习效果最好的算法依次为面部表情基于归一化坐标向量F的60维PCA算法、语音基于ComparE向量V的60维PCA算法、步态基于距离—角度转化向量P'的16或60维PCA算法。

4.2.2 基于深度神经网络的算法

首先，基于预处理得到的面部表情、声谱图集合，采用基于深度神经网络的心理评测算法ResNet-Attention-Bi_LSTM，使用$loss_p$损失函数训练ResNet18和Bi-LSTM，并通过调整α值得到相应预测集的准确率。然后，基于预处理得到的步态特征$S\in\{P, P'\}$，使用ST-GCN和$loss_p$方法进行算法训练。实验结果如表5所示。

由表5可知，在交叉熵损失函数基础上采用改进的极性损失函数$loss_p$的训练效果更好，α在0.25~0.75之间的效果相对较好；对于步态模态结果而言，距离—角度转化向量P'相较于归一化坐标向量P的整体效果更好。每个模态基于深度神经网络的效果最好的算法依次为α=0.75的面部表情ResNet-Attention-Bi_LSTM算法、α=0.50的语音ResNet-Attention-Bi_LSTM算法、α=0.50的步态模态基于距离—角度转化向量P'的ST-GCN算法。

Table 5 Polarity parameter α and prediction set accuracy of each single modal on the deep neural network algorithm

表5 极性参数α和各单模态在深度神经网络算法上的预测集准确率 (%)

极性α / 模态		0.00	0.25	0.50	0.75	1.00
表情	图片	41.27	46.03	44.44	47.62	46.03
语音	图片	31.75	31.75	42.86	38.10	36.51
步态	P	38.10	44.44	44.44	39.68	39.68
	P'	47.62	49.21	52.38	44.44	42.86

4.3 多模态融合心理评测算法分析

4.3.1 多模态融合算法

本文综合单模态算法实验得到针对各模态效果最好的模型依次为面部表情极性参数α=0.75的ResNet-Attention-Bi_LSTM算法、语音基于ComparE向量V的降为60维MFE-PCA-SVM算法、步态基于距离—角度转化向量P'=0.50的ResNet-Attention-Bi_LSTM算法。在多模态融合算法实验中，首先采取决策层融合模型提取各心理标签分类

的概率分布；其次对同类别的概率取平均再整体取最大值；再次采取特征层融合，根据拼接向量 $\boldsymbol{Z}$ 和 MLP 生成注意力与原本各自的特征向量相乘后拼接，作为分类层的输入；最后经过调整各种超参数及不同模态之间的融合实验，得到不同模态融合下的预测集准确率结果，如表 6 所示。

Table 6　Experimental results of psychological evaluation algorithm based on multimodal fusion

表 6　多模态融合的心理评测算法实验结果

算法	面部表情	语音	步态	准确率/%
决策层融合	√	√		52.38
	√		√	52.38
		√	√	54.09
	√	√	√	56.76
特征层融合	√	√		62.30
	√		√	60.66
		√	√	57.38
	√	√	√	65.66

由表 6 可知，3 个模态的融合效果相较于两个模态更好，在特征层上融合效果相较于在决策层上融合效果更好。由此可得，特征层能更充分融合不同模态之间的内在信息；决策层仅能融合各模态的最终结果，但可解决模态缺失的问题。

4.3.2　多模态与单模态算法

本文挑选各模态分别在机器学习和深度学习方法中效果最好的实验结果，与 LBP[6]、C3D[31]、MFCC-SVM[32]等多模态融合算法进行比较，具体数据如表 7 所示。由此可知，多模态融合相对于单模态而言能显著提升心理评测准度，且基于注意力机制的特征层融合效果最好，对心理标签四分类的准确率能达到 65.66%，相较于面部表情、语音、步态 3 种模态的模型效果提升 18.04%、21.22%、13.28%。

Table 7　Experimental results of psychological evaluation algorithms

表 7　心理评测算法实验结果

算法	模态	准确率/%
MFE-PCA-SVM	面部表情	41.27
MFE-PCA-SVM	语音	44.44
MFE-PCA-SVM	步态	46.03
ResNet-Attention-Bi_LSTM	面部表情	47.62
ResNet-Attention-Bi_LSTM	语音	42.86
ST-GCN	步态	52.38
LBP/MFCC-SVM[3][33]	面部表情+语音	39.39
C3D[32]	面部表情+语音	39.39
决策层融合	面部表情+语音+步态	56.76
特征层注意力融合	面部表情+语音+步态	65.66

5　结语

本文提出融合表情、声音、步态 3 个模态的多模态心理评测模型，并在真实数据集上进行了性能分析。首先基于注意力机制的多模态融合方法增加模型的了泛化性；其次引入极性损失函数智能区分不同心理状态的界限。实验表明，该多模态融合算法相较于单模态模型而言，在心理评测任务上具有显著优势。

在电网行业的心理评测场景中的应用中，本文模型的效果也优于其他经典算法，证明了其在该领域的先进性与适用性。未来，将对所提模型的应用场景进行泛化，使其在真实作业场景下也能准确评测人员的生理及心理状态。

参考文献：

[1] LI X J, MENG H, LI S Y, et al. Influence of electric field in transformer area of 500 kV substation on human body[J]. Guangdong Electric Power, 2022,35(2):121-127.
李小娟，孟欢，李韶瑜，等. 500 kV 变电站变压器区域电磁场对人体的影响[J]. 广东电力，2022,35(2):121-127.

[2] PENG P, XUE D. Research on design of cloud side collaborative IoT framework based on edge agent and its application[J]. Guangdong Electric Power, 2023,36(5):18-26.
彭鹏，薛东. 基于边缘代理的云边协同物联网框架设计及其应用研究[J]. 广东电力，2023,36(5):18-26.

[3] RUDOKAITE J, ERTUGRUL I O, ONG S. Predicting vasovagal reactions to needles from facial action units[J]. Journal of Clinical Medicine, 2023, 12(4):1644.

[4] BACHOROWSKI J A. Vocal expression and perception of emotion[J]. Current Directions in Psychological Science, 1999, 8(2): 53-57.

[5] ROETHER C L, OMLOR L, CHRISTENSEN A, et al. Critical features for the perception of emotion from gait[J]. Journal of Vision, 2009, 9(6): 15.

[6] EKMAN P, FRIESEN W V. Facial action coding system (FACS): a technique for the measurement of facial actions[J]. Rivista Di Psichiatria, 1978, 47(2):126-138.

[7] MYLONAKIS C M, MYLONAKIS Z D. A novel three-dimensional direction-of-arrival estimation approach using a deep convolutional neural network[J]. IEEE Open Journal of Vehicular Technology, 2024,5:643-657.

[8] LUO Y, ZHANG T, ZHANG Y. An improved LDP facial expression feature extraction method[J]. Semiconductor Optoelectronics, 2016, 37(1): 122-125.
罗元，张天，张毅. 一种改进的 LDP 面部表情特征提取方法[J]. 半导体光电，2016, 37(1):122-125.

[9] JABID T, KABIR M H, CHAE O. Local directional pattern (LDP) for face recognition[C]// 2010 Digest of Technical Papers International Conference on Consumer Electronics, 2010: 329-330.

[10] ROWEIS S. EM algorithms for PCA and SPCA[EB/OL]. https://cs.nyu.edu/~roweis/papers/empca.pdf.

[11] HEARST M A, DUMAIS S T, OSUNA E, et al. Support vector machines[J]. IEEE Intelligent Systems and Their Applications, 1998, 13(4): 18-28.

[12] LECUN Y, KAVUKCUOGLU K, FARABET C. Convolutional networks and applications in vision[C]// Proceedings of 2010 IEEE International Symposium on Circuits and Systems, 2010: 253-256.

[13] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[DB/OL]. https://arxiv. org/abs/1412. 0767.

[14] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997,9(8):1735-1780.

[15] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks [J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.

[16] SATO N, OBUCHI Y. Emotion recognition using mel-frequency cepstral coefficients [J]. Information and Media Technologies, 2007, 2(3): 835-848.

[17] PALAZ D, MAGIMAIDOSS M, COLLOBERT R. Analysis of CNN-based speech recognition system using raw speech as input [EB/OL]. https://ronan. collobert. com/pub/2015_cnnspeech_interspeech. pdf.

[18] EYBEN, F, SCHERER K R, SCHULLER B, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing[J]. IEEE Transactions on Affective Computing, 2016, 7(2):190-202.

[19] JI S, XU W, YANG M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 221-231.

[20] GRAVES A , FERNÁNDEZS, SCHMIDHUBER J . Bidirectional LSTM networks for improved phoneme classification and recognition[C]// Artificial Neural Networks: Formal Models and Their Applications, 2005: 799-804.

[21] HUANG Z , MING D , MAO Q , et al. Speech emotion recognition using CNN [C]// 2021 International Conference on Artificial Intelligence and Smart Systems,2021:1176-1181.

[22] TRIGEORGIS G, RINGEVAL F, BRUECKNER R, et al. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network [C]// IEEE International Conference on Acoustics, 2016: 5200-5204.

[23] BADSHAH A M, AHMAD J, RAHIM N, et al. Speech emotion recognition from spectrograms with deep convolutional neural network[C]// 2017 International Conference on Platform Technology and Service, 2017: 1-5.

[24] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[DB/OL]. https://arxiv. org/abs/1801. 07455.

[25] BHATTACHARYA U, RONCAL C, MITTAL T, et al. Take an emotion walk: perceiving emotions from gaits using hierarchical attention pooling and affective mapping[DB/OL]. https://arxiv. org/abs/1911. 08708.

[26] 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.

[27] BOXUAN Z , ZIKUN Q, SHUO Y. Emotion recognition with facial expressions and physiological signals[C]// 2017 IEEE Symposium Series on Computational Intelligence,2017:978-982.

[28] HOSSAIN M S, MUHAMMAD G. Emotion recognition using deep learning approach from audio-visual emotional big data[J]. Information Fusion, 2019,6(49):69-78.

[29] MNIH V, HEESS N, GRAVES A. Recurrent models of visual attention [DB/OL]. https://arxiv. org/abs/1406. 6247.

[30] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]// Proceedings of the European Conference on Computer Vision,2018: 3-19.

[31] KAZEMI V, SULLIVAN J. One millisecond face alignment with an ensemble of regression trees[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014:1867-1874.

[32] NOROOZI F, MARJANOVIC M, NJEGUS A, et al. Audio-visual emotion recognition in video clips affective computing[J]. IEEE Transactions on, 2017(10):60-75.

[33] SCHULLER B, STEIDL S, BATLINER A, et al. The interspeech 2016 computational paralinguistics challenge: deception, sincerity and native language [EB/OL]. https://www5. informatik. uni-erlangen. de/Forschung/Publikationen/2016/Schuller-TI2. pdf.

[34] TIWARI T V. MFCC and its applications in speaker recognition[J]. International Journal on Emerging Technologies,2010,1(1): 19-22.

(责任编辑:刘嘉文)