计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(1):337−347 [doi: 10.15888/j.cnki.csa.008907]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# 融合语音、脑电和人脸表情的多模态情绪识别①

方伟杰, 张志航, 王恒畅, 梁　艳, 潘家辉

(华南师范大学 软件学院, 佛山 528225)
通信作者: 潘家辉, E-mail: panjh82@qq.com

**摘　要**: 本文提出了一种多模态情绪识别方法, 该方法融合语音、脑电及人脸的情绪识别结果来从多个角度综合判断人的情绪, 有效地解决了过去研究中准确率低、模型鲁棒性差的问题. 对于语音信号, 本文设计了一个轻量级全卷积神经网络, 该网络能够很好地学习语音情绪特征且在轻量级方面拥有绝对的优势. 对于脑电信号, 本文提出了一个树状 LSTM 模型, 可以全面学习每个阶段的情绪特征. 对于人脸信号, 本文使用 GhostNet 进行特征学习, 并改进了 GhostNet 的结构使其性能大幅提升. 此外, 我们设计了一个最优权重分布算法来搜寻各模态识别结果的可信度来进行决策级融合, 从而得到更全面、更准确的结果. 上述方法在 EMO-DB 与 CK+数据集上分别达到了 94.36% 与 98.27% 的准确率, 且提出的融合方法在 MAHNOB-HCI 数据库的唤醒效价两个维度上分别得到了 90.25% 与 89.33% 的准确率. 我们的实验结果表明, 与使用单一模态以及传统的融合方式进行情绪识别相比, 本文提出的多模态情绪识别方法有效地提高了识别准确率.

**关键词**: 多模态情绪识别; 决策级融合; 轻量级模型; LSTM; GhostNet; 深度学习

引用格式: 方伟杰,张志航,王恒畅,梁艳,潘家辉.融合语音、脑电和人脸表情的多模态情绪识别.计算机系统应用,2023,32(1):337–347. http://www.c-s-a.org.cn/1003-3254/8907.html

## Multimodal Emotion Recognition Based on Speech, EEG and Facial Expression

FANG Wei-Jie, ZHANG Zhi-Hang, WANG Heng-Chang, LIANG Yan, PAN Jia-Hui

(School of Software, South China Normal University, Foshan 528225, China)

**Abstract**: In this study, a multimodal emotion recognition method is proposed, which combines the emotion recognition results of speech, electroencephalogram (EEG), and faces to comprehensively judge people's emotions from multiple angles and effectively solve the problems of low accuracy and poor robustness of the model in the past research. For speech signals, a lightweight fully convolutional neural network is designed, which can learn the emotional characteristics of speech well and is overwhelming at the lightweight level. For EEG signals, a tree-structured LSTM model is proposed, which can comprehensively learn the emotional characteristics of each stage. For face signals, GhostNet is used for feature learning, and the structure of GhostNet is improved to greatly promote its performance. In addition, an optimal weight distribution algorithm is designed to search for the reliability of modal recognition results for decision-level fusion and thus more comprehensive and accurate results. The above methods can achieve the accuracy of 94.36% and 98.27% on EMO-DB and CK+ datasets, respectively, and the proposed fusion method can achieve the accuracy of 90.25% and 89.33% on the MAHNOB-HCI database regarding arousal and valence, respectively. The experimental results reveal that the multimodal emotion recognition method proposed in this study effectively improves the recognition accuracy compared with the single mode and the traditional fusion methods.

**Key words**: multimodal emotion recognition; decision-level fusion; lightweight model; LSTM; GhostNet; deep learning

① 基金项目: 科技创新 2030“脑科学与类脑研究”重点项目 (2022ZD0208900); 国家自然科学基金面上项目 (62076103)
收稿时间: 2022-06-01; 修改时间: 2022-07-01; 采用时间: 2022-07-13; csa 在线出版时间: 2022-08-24
CNKI 网络首发时间: 2022-11-16

## 1 引言

### 1.1 研究背景

情绪可以理解为是人对周边事物所产生的生理反应[1], 其表达方式可大致分为两类: 一类是外在行为表现, 如人脸表情、声音等, 另一类是人的内在生理表现, 如脑电图 (electroencephalogram, EEG)、皮肤电等. 同样的, 情绪的定义也可以分为两类: 一类认为情绪是离散的, 也就是我们日常生活中所说的情绪类型, 这类定义将情绪分为有限个的情绪类型, 如快乐、愤怒、悲伤等. 另一类认为情绪的变化是连续的, 这类中最常用的是 Russel[2] 提出的唤醒效价 (arousal-valence) 二维情绪模型, 其中唤醒指的是平静或兴奋的程度, 效价指的是积极或消极的程度.

情绪的重要性促使了大量情绪识别方法出现, 然而目前大多数方法考虑到的模态不够全面, 以至于准确率以及鲁棒性无法达到理想水平. 据过去的研究发现, 人脸表情与语音在人与人之间交流中传达的情感因素分别占据 55% 和 38%[3], 两个模态对于情绪识别都具有非常高的研究价值. 在生理信号方面, 脑电信号作为中枢神经系统的信号, 比其他任何信号都能够更准确、更客观地反映人的情绪状态变化, 可以避免因被试者伪装人脸表情等非自然因素而导致的方法误判. 基于上述背景, 本文将采用语音、脑电以及人脸表情 3 个模态对情绪识别展开研究, 并在融合阶段采用唤醒效价模型对 3 个模态的识别结果进行决策级融合, 进而得到更精确、鲁棒性更高的情绪识别方法.

### 1.2 研究现状

#### 1.2.1 单模态情绪识别相关研究

人脸表情识别本质上也可以认为是一项特殊的图像分类任务, 因此有大量的研究人员使用经典的图像分类模型来进行该项任务的研究, 如 ResNet[4]、Xception[5] 等. Chowdary 等人[6] 使用在 ImageNet 数据库上进行预训练的 VGG19、ResNet50、InceptionV3 来进行面部表情识别, 结果表明这些图像分类模型在面部表情识别任务中是可行的. 但这些模型往往参数过多, 训练会消耗太多资源[6]. 2020 年华为诺亚方舟实验室[7] 提出的 GhostNet 在 ImageNet 中取得了最好的效果, 并且在提高分类准确率的前提下还大大减少了模型参数量以及计算成本, 这为我们在人脸表情的研究中提供了重要的参考价值. 对于语音情绪识别, Chen 等人[8] 设计了一个基于 3D 注意力的卷积神经网络, 使用 delta 和 delta-deltas 的梅尔频谱特征来进行语音情绪识别, 并在 EMO-DB 数据集上取得了 82.82% 的准确率. Zhang 等人[9] 提出了一种基于深度卷积神经网络和双向长短时间记忆网络的注意力模型的语音情绪识别方法, 在 EMO-DB 数据集上获得了 87.86% 的未加权平均召回率. 对于 EEG 情绪识别, Alhagry 等人[10] 提出了一种基于 LSTM 的脑电情绪识别方法, 并在 DEAP 数据集中的唤醒效价维度分别得到了 85.65% 和 85.45% 的准确率. Wu 等人[11] 提出了一种情绪相关的关键子网络选择算法, 能够利用通道之间的连接关系来获得更好的性能. 对于跨被试脑电情绪识别的研究, Li 等人[12] 提出了一种基于图注意力的自组织图神经网络, 可以构建基于通道脑电信号的图网络, 并在 SEED 数据集上达到了 86.81% 的准确率.

然而情绪作为一个复杂的生理现象, 基于单个模态得到的识别结果往往是不可靠的. 一个优秀的情绪识别方法应该是能够从多个模态进行综合判断, 从而得到更精确的识别结果. 因此近些年来也有少部分采用双模态进行的情绪识别研究, 这在一定程度上提高了模型的准确率以及鲁棒性.

#### 1.2.2 多模态情绪识别相关研究

Zhou 等人[13] 提出了一种基于自适应分解双线性池的多模态融合注意力网络用于视听情绪识别, 并在 IEMOCAP 数据集上达到了 75.49% 的最佳准确率. 为增强人脸表情和语音之间的情绪相关映射, Ma 等人[14] 提出了一种用于视听情感识别的深度加权融合方法, 并使用深度信念网络对多模态情感特征进行高度非线性融合, 在 RML 视听情感数据集上取得了 82.38% 的准确率. 以上都是对外部行为信号融合的研究, 但近年来的一些研究发现, 生理信号与外部信号融合可以取得更好的效果. Li 等人[15] 提出了一种决策级融合方法, 他们将人脸表情和脑电信号结合起来进行实时识别人的情绪, 并在 MAHNOB-HCI 和 DEAP 数据集上验证了其方法的可行性. Wang 等人[16] 提出了一种基于最大权重法的脑电图和人脸表情信息融合的情绪识别方法, 有效提高了情绪识别的准确率.

事实上, 目前大多数已提出的多模态情绪识别方法仅采用两种模态进行情绪识别, 如语音与人脸表情的融合, EmotiW、AVEC 等国际视听情感识别大赛引起了不少研究人员使用这两种模态进行情绪识别研究.

尽管语音与人脸表情在人与人之间的交流过程中共占据了 93% 的情感因素[3], 但正如上文所述, 外部行为表现并不具有客观性, 仅使用外部行为表现来判断人真实的情绪是不可靠的, 一个优秀的情绪识别方法还需要与情绪最相关的生理信号-脑电信号来支持. 因此, 将语音、脑电以及人脸结合起来, 从理论上讲可以构建一个更为全面的多模态情绪识别方法.

在本文中, 我们为 3 种模态分别设计了一个深度学习模型. 对于语音模态, 设计了一个轻量级的全卷积神经网络 (lightweight full convolutional neural network, LFCNN) 模型, 该模型仅有少量参数但能很好的学习到情绪特征达到较高的准确率. 针对脑电模态, 设计了一个树状 LSTM (tree-like LSTM, tLSTM) 模型, 可以充分学习训练过程中各阶段的脑电情绪特征. 针对人脸模态, 首次将 GhostNet 应用于人脸表情识别, 并通过改进 GhostNet 的结构, 使改进后的模型在人脸表情识别领域达到了先进水平. 最后, 设计了一个最优权重分布算法来搜寻各模态的可信度, 从而进行决策级融合以获得更全面、准确的结果.

## 2 基于深度学习的多模态情绪识别算法

### 2.1 基于语音的情绪识别方法

#### 2.1.1 语音数据预处理与特征提取

语音的梅尔频谱 (Mel-spectrogram) 特征在过去语音情绪识别领域表现出了优异的效果[17]. 因此, 在本文的研究中也将使用 Mel-spectrogram 特征进行语音情绪识别. Mel-spectrogram 是将频率转换为梅尔尺度的频谱图. 我们首先通过短时傅里叶变换 (short-term Fourier transform, STFT) 提取原语音的时频特征, 得到时频特征之后再通过 Mel 滤波器组即可得到 Mel-spectrogram. 在转换过程中涉及 Mel 频率的转换, 我们设原始频率为$f$, 转换后的 Mel 频率为$F_{\text{Mel}}$那么它们之间的转换关系可以用式 (1) 来表示:

$$F_{\text{Mel}} = 2595\lg\left(1+\frac{f}{100}\right) \tag{1}$$

#### 2.1.2 语音情绪特征学习模型

本文提出的 LFCNN 将被用于学习语音情绪特征, 该模型主要由并行卷积部分、残差结构部分以及串行卷积部分组成, LFCNN 完整的结构如图 1 所示.

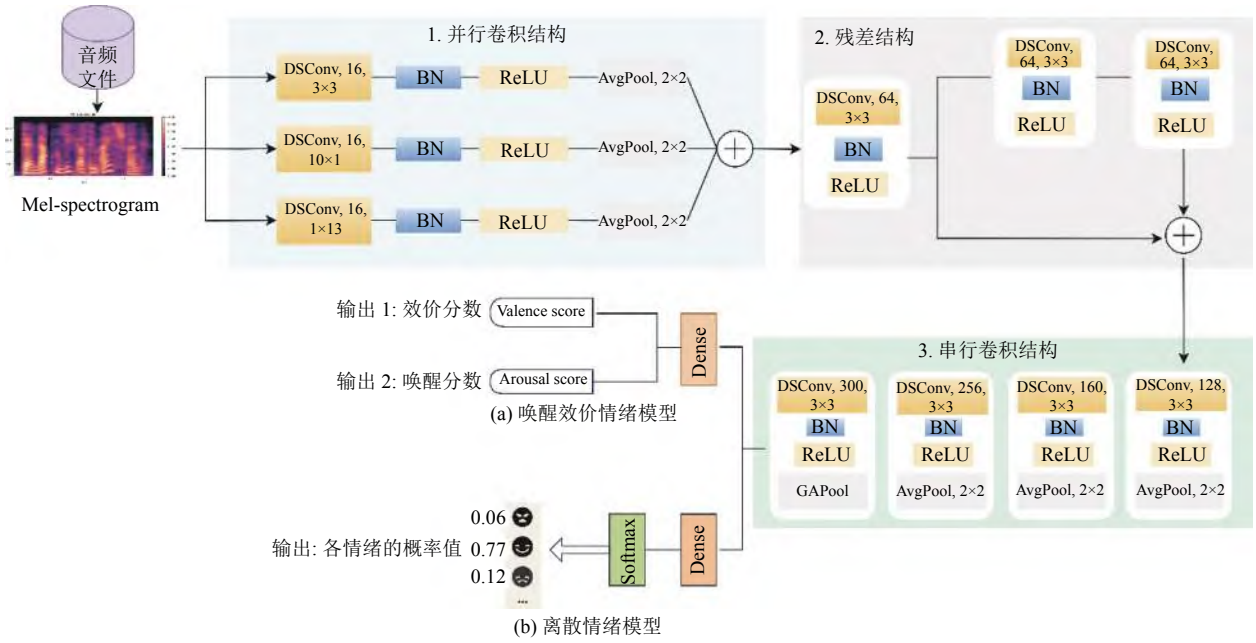



图 1 提出的用于语音情绪识别的 LFCNN 结构

我们在对过去有关卷积层的研究中发现, 深度可分离卷积 (depth separable convolution, DSConv) 的参数量相比传统卷积要少, 并且 Xception 的成功证明了深度可分离卷积相比传统卷积的优越性, 因此本文将使用它来设计我们提出的 LFCNN. 要注意的是后面提到的卷积层均为深度可分离卷积.

LFCNN 的第 1 部分为并行卷积结构, 它包含 3 个并行的卷积层, 3 个卷积层的卷积核的数量均设置为

16, 不同之处在于它们的卷积核的大小分别为 3×3、13×1 和 1×10, 它们的输出将合并在一起送至模型的第 2 部分. 模型的第 2 部分采用残差结构思想, 主干边包含有两个卷积层, 每个卷积层为 64 个大小为 3×3 的卷积核. 第 3 部分是 4 个连续的卷积层, 其内核大小均为 3×3, 内核数量依次为 128、160、256、300. 需要注意的是, 除了第 3 部分所有的卷积层后面都接有批归一化层 (batch normalization, BN)、线性整流函数 ReLU 激活层以及池化层. 对于具体池化方法的选择, 除模型的第 3 部分结束时采用了全局平均池化 (GlobalAverage-Pooling, GAPool) 外, 所有的池化方法都采用平均池化 (AveragePooling, AvgPool). 对于模型的最后一部分, 我们可以根据训练样本的标签类型灵活设计. 当数据集用于使用维度模型来描述情绪时, 这意味着 LFCNN 需要被设置为一个多任务回归模型, 我们需要使用多个全连接层 (也称稠密层, Dense) 来输出多个维度的分数. 当使用离散模型时, LFCNN 被设置为一个分类任务模型, 每种情绪的概率通过 Softmax 层输出得到.

### 2.2 基于脑电的情绪识别方法

#### 2.2.1 脑电信号预处理与特征提取

在过去的大量研究中, 频域中的功率谱密度 (power spectral density, PSD) 特征被广泛认为是最适合 EEG 情绪分析的特征[11], 因此本文选择使用 PSD 特征进行情绪分析. 本文提取 PSD 的方法采用小波变换, 可以看作是 STFT 的改进方法, 能够更好地处理时频信号. EEG 具有不同频带的 PSD 特性, 因此使用可以支持多个频带的 Daubechies 小波变换系数[18] 进行频域转换. 具体地, 令$\epsilon$为开尺度参数, $\tau$为移位置参数, 则一维连续小波变换$W_f(\epsilon,\tau)$程可表示为:

$$W_f(\epsilon,\tau)=\int_{-\infty}^{+\infty} f(t)\psi_{\epsilon,\tau}(t)dt \tag{2}$$

其中, $\psi$表示一维母小波函数, 计算公式为:

$$\psi_{\epsilon,\tau}(t)=\frac{1}{\sqrt{\epsilon}}\psi\left(\frac{t-\tau}{\epsilon}\right) \tag{3}$$

此外, 连续小波的逆变换定义为:

$$f(t)=\frac{1}{C_\psi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{W_f(\epsilon,\tau)\psi_{\epsilon,\tau}(t)}{\epsilon^2}d\epsilon d\tau \tag{4}$$

其中, $C_\psi$的计算公式表示为:

$$C_\psi=\int_{-\infty}^{+\infty}\left(\frac{|\hat{\psi}(u)|^2}{|u|}\right)du \tag{5}$$

其中, $\hat{\psi}(u)$是$\psi(t)$的傅里叶变换.

本文选取了 5 个频带的 PSD 特征: theta (4 Hz < $f$ < 8 Hz)、slow alpha (8 Hz < $f$ < 10 Hz)、alpha (8 Hz < $f$ < 12 Hz)、beta (12 Hz < $f$ < 30 Hz) 和 gamma (30 Hz < $f$ < 64 Hz). 这意味着假设共使用 $N$ 个电极通道, 总共可以获得 5×$N$ 个特征. 我们在过去的研究中发现并不是所有电极都能传达出情绪相关的信息, 且不同频带能够传达情绪信息的通道也不同[19], 因此我们有针对性地选择了 14 个电极 (FP1、FP2、F8、FC2、FC6、T7、CZ、C4、T8、CP1、CP2、CP6、PO4、OZ) 作为我们的研究对象, 这些电极的组合能够很好地反映出不同频带上的情感信息[20], 其分布如图 2 所示, 蓝色代表我们选择的电极. 此外, 我们还选择了 3 对对称电极 (FP1-FP2、T7-T8、CP1-CP2) 来扩展特征数量, 因此特征总数为 5×(14+3) = 85. 我们使用 10 s的时间窗从 5 个频带中提取 PSD 特征, 并采用 50% 的重叠率采样来扩展数据集.

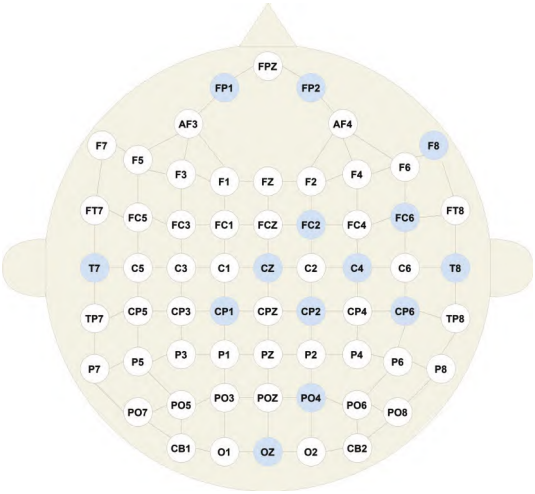



图 2 国际 10-20 系统的电极位置

#### 2.2.2 脑电情绪特征学习模型

我们提出的 tLSTM 结构如图 3 所示. 对于树状部分, 叶子节点处的 LSTM 单元都有相同数量的神经元, 这是为了保证它们的输出形状一致. 此外为保证模型大小与性能之间的平衡, 我们提出的树形结构由 4 个层级组成, 不同层级代表不同学习阶段, 叶子节点处于不同的层级, 这样能够将较浅和较深的特征进行初步融合并送入后续结构中学习更深层次的特征. 树状部分的 LSTM 单元输出是整个序列的输出, 而之后的 LSTM 单元的输出是最后一个隐藏层的输出.

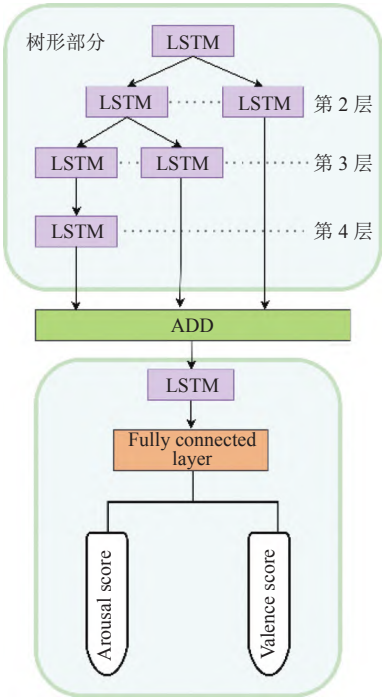



图 3 提出的用于脑电情绪识别的 tLSTM 结构

### 2.3 基于人脸的情绪识别方法

#### 2.3.1 人脸图像预处理

我们捕捉到的人脸图像往往会包含自然环境中的其他物体, 这些干扰因素会影响到后续的特征学习, 因此我们需要从中提取有效的部分, 即人脸部分. 这部分过程在数据预处理中称之为降噪提纯, 在该任务中也可以理解为人脸检测.

对于一张包含人脸的图片, 第一步是检测人脸的 68 个关键点, 然后将人脸和人脸的关键点进行旋转对齐. 两眼中心连线作为旋转图像的水平线, 左右眼中心坐标分别为$(x_1,y_1)$和$(x_2,y_2)$, 设要旋转的角度为 $\theta$, 则计算公式可以表示为式 (6). 最后再根据关键点裁剪图像的人脸部分. 在本文中, 我们将灰度图像的大小调整为 48×48×1 作为模型的输入形状.

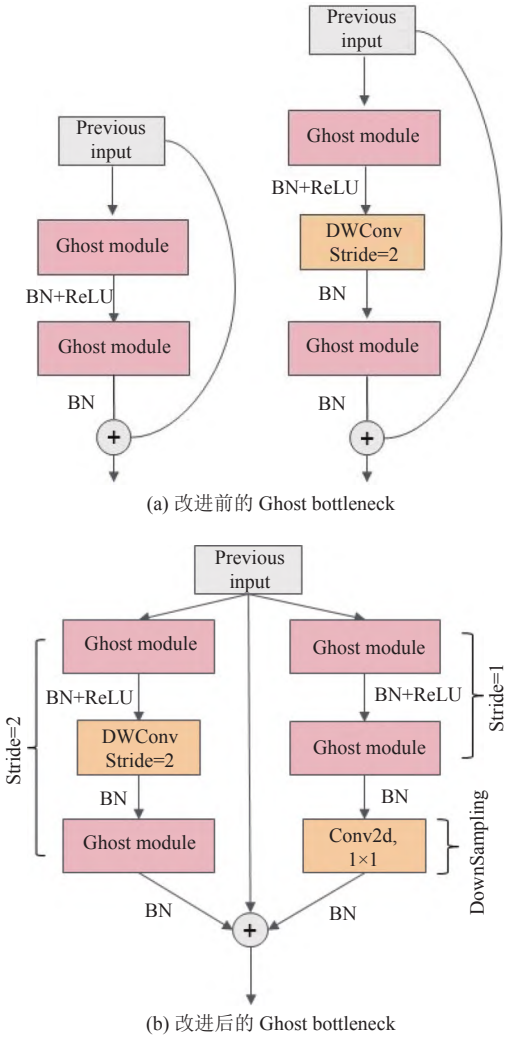
$$\theta = 180° \frac{\arctan\left(\dfrac{y_2 - y_1}{x_2 - x_1}\right)}{\pi} \tag{6}$$

#### 2.3.2 人脸表情特征学习模型

GhostNet 作为华为诺亚方舟实验室 2020 年提出的新图像分类模型, 已被一些研究人员在工作中使用并取得了不错的效果[21]. 因此, 本文将使用 GhostNet 进行人脸表情识别进行初步研究探索, 有望获得较好的结果.

GhostNet 主要由多个 Ghost bottleneck 组成, 其中 Ghost bottleneck 核心又由 Ghost module 组成. Ghost module 相比与传统的卷积层, 它的核心思想是通过一些廉价的操作来生成更多的特征图, 这也是 GhostNet 参数少而性能高的主要原因之一.

本文对 GhostNet 的改进主要集中在对 Ghost bottleneck 结构的改进中. 原始的 Ghost bottleneck 分为步长 (Stride) 为 1 和步长为 2 两种情况, 当 Stride = 1 时是两个连续的 Ghost module 结构, 而当 Stride = 2 时两个 Ghost module 之间插入了一个步长为 2 的深度可分离卷积层, 具体结构如图 4(a) 所示. 而我们改进的 Ghost bottleneck 结合了这两种操作的特点, 因此能够学习到更全面的特征. 改进后的 Ghost bottleneck 结构如图 4(b) 所示. 当我们输入的数据形状为 48×48×1 时, 整个改进过后的 GhostNet 结构如表 1 所示, 其中 EXP 表示扩展因子; OUT 表示输出通道数; SE 表示是否使用 SE 模块.




(a) 改进前的 Ghost bottleneck

(b) 改进后的 Ghost bottleneck

图 4 改进前后的 Ghost bottleneck 对比

表 1 改进后的 GhostNet 结构

| 操作 | 输出 | Ghost bottleneck设置 | | |
|---|---|---|---|---|
| | | EXP | OUT | SE |
| Conv2d, 16, 3×3 | (batch, 24, 24, 16) | — | — | — |
| Ghost bottleneck | (batch, 12, 12, 40) | 120 | 40 | True |
| Dropout, 0.3 | (batch, 12, 12, 40) | — | — | — |
| Ghost bottleneck | (batch, 6, 6, 80) | 240 | 80 | False |
| Dropout, 0.3 | (batch, 6, 6, 80) | — | — | — |
| Ghost bottleneck | (batch, 3, 3, 160) | 672 | 160 | True |
| Dropout, 0.3 | (batch, 3, 3, 160) | — | — | — |
| Ghost bottleneck | (batch, 2, 2, 160) | 960 | 160 | False |
| Dropout, 0.3 | (batch, 2, 2, 160) | — | — | — |
| Conv2d, 256, 1×1 | (batch, 2, 2, 256) | — | — | — |
| Dropout, 0.3 | (batch, 2, 2, 256) | — | — | — |
| GAVPool | (batch, 1, 1, 256) | — | — | — |
| Conv2d, 512, 1×1 | (batch, 1, 1, 512) | — | — | — |
| Dense, Softmax | (batch, 7) | — | — | — |

### 2.4 基于决策级融合的多模态情绪识别方法

融合方法对于多模态任务来说也是重要且不可缺少的部分. 在得到各模态的情绪识别结果之后, 我们设计了一种决策级融合方法, 以得到更精确、全面的情绪识别结果.

#### 2.4.1 决策级融合

决策级融合是多模态融合方法当中常用的方法之一, 我们也可以称之为后期融合, 即在得到多个模态的识别结果后, 对结果进行加权融合, 进而得到更全面的结果. 传统的决策级融合方式是直接将结果进行等权融合[22], 这种做法虽然考虑到了多个模态的情绪识别结果, 但却忽略了各个模态的可信度, 从而导致鲁棒性提升不大. 针对上述缺陷, 我们提出了一种能够针对各模态的可信度进行判断的最优权重分布算法, 该方法能够赋予可信度高的模态更高的权重, 而可信度较低的模态则赋予低权重, 从而能够使得融合结果精确率更高、鲁棒性更好.

#### 2.4.2 最优权重分布算法

以搜寻各模态唤醒分数的最优权重为例, 假设有 $n$ 个模态对应 $n$ 个回归模型, 共 $T$ 次试验用于预测, 第 $k$ 个模型中试验 $t$ 的预测平均唤醒评分为 $A_{tk}$, $k\in\{1,2,3,\cdots,n\}, t\in\{1,2,3,\cdots,T\}$. 设权重集$\varpi$为{0.00, 0.01, 0.02, ⋯, 0.98, 0.99, 1.00}, 即一个从 0.00 开始到 1.00 结束, 步长为 0.01 的数组. 以均方根误差 (root mean-square error, RMSE) 作为衡量指标来评价当前权重分布的性能, 当各模态处于性能最好的权重分布时, RMSE 应该是最小的, 记为 $RS_{\min}$. 我们提出的最优权重分布搜索算法流程如图 5 所示, 主要可分为 3 个步骤, 具体过程如下.

步骤 1. 在$\varpi$中循环枚举 $n$ 个模态的权重. 设第 $k$ 个模态的权重为$\omega_k$, 当所有权重之和满足式 (7) 时进入步骤 2. 当循环枚举结束时保存最优权重分布, 算法结束.

$$\sum_{k=1}^{n}\omega_k = 1 \tag{7}$$

步骤 2. 计算当前权重分布下的多模态融合得到的预测唤醒分数. 假设试验 $t$ 的预测唤醒分数为$\hat{y}_t$, 则计算公式可表示为:

$$\hat{y}_t = \sum_{k=1}^{n}\omega_k A_{tk} \tag{8}$$

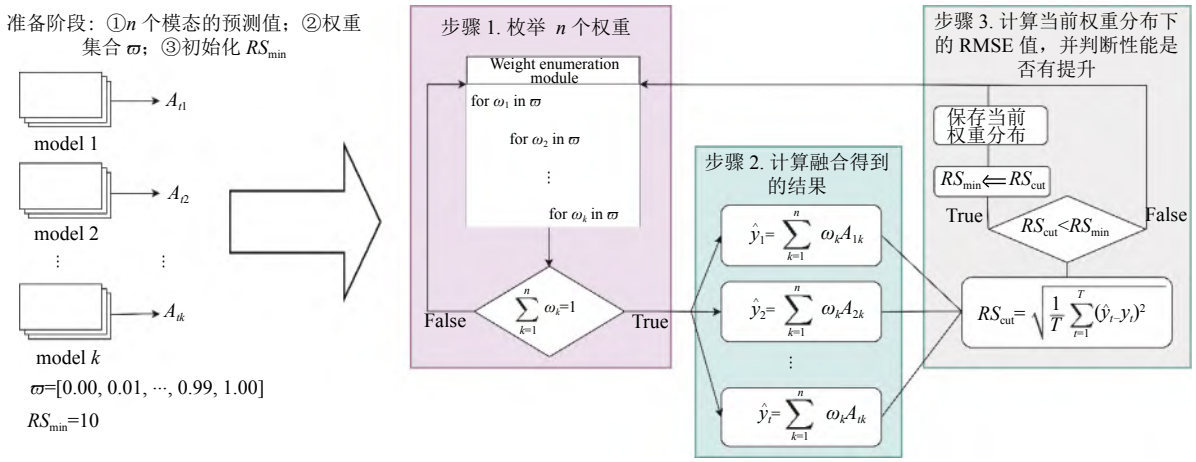



图 5 最优权重分布算法步骤示意图

步骤 3. 计算当前权重分布下 $T$ 次试验的 RMSE, 记为 $RS_{\text{cut}}$, 计算公式为式 (9), 其中 $y_t$ 为试验 $t$ 的真实唤醒分数. 通过比较 $RS_{\text{cut}}$ 和 $RS_{\text{min}}$ 的大小关系来判断当前权重分布是否拥有更好的性能, 当 $RS_{\text{cut}}<RS_{\text{min}}$ 时, 认为当前权重分布有更好的性能, 所以将 $RS_{\text{min}}$ 更新为 $RS_{\text{cut}}$, 保存当前权重分布. 当 $RS_{\text{cut}}\geqslant RS_{\text{min}}$ 时, 认为当前的权重分布没有表现出更好的性能, 不需要对 $RS_{\text{min}}$ 进行更新. 但无论大小关系如何, 都要再次执行步骤 1 枚举下一组权重分布.

$$RS_{\text{cut}}=\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t-y_t)^2} \tag{9}$$

## 3 实验与结果

### 3.1 实验环境与数据集

所有模型均基于 Python 语言及 TensorFlow 2.3 深度学习框架设计实现, 并在 Windows 10 操作系统下的 NVIDIA GTX1050 GPU 上进行训练.

实验用到的数据集包括: EMO-DB、CK+、MAHNOB-HCI 以及 FER2013. 其中 EMO-DB 是一个语音情绪识别数据集; CK+是一个人脸表情识别数据集, MAHNOB-HCI 是一个多模态情绪识别数据集, 同时包含了语音、脑电以及人脸数据; Fer2013 是一个大型人脸表情数据集.

### 3.2 EMO-DB 实验

#### 3.2.1 实验步骤

对于 EMO-DB 上的实验, 首先使用 noisereduce 库去除原音频文件中的噪声, 由于速率的变化不会影响语音情感信息, 因此可以通过对原文件进行变速操作扩充数据集. 经过上述操作之后再使用 librosa 库来提取语音的 Mel-spectrogram 特征并将数据保存在一个 numpy 数组中, 供我们用作模型的训练与验证. 我们在一个训练过程中有 300 个 epochs, batchsize 设置为 64, 使用初始学习率为 $10^{-4}$ 的 Adam 优化器, 且从第 150 个 epoch 开始每 10 个 epoch 下降 $e^{-0.10}$. 为了验证我们提出的方法足够可靠, 实验采用 10 折交叉验证方法进行.

#### 3.2.2 实验结果

对 EMO-DB 的分类结果达到了 94.36% 的平均准确率和 94.38% 的 $F1$ 值, 图 6 展示了具体结果的混淆矩阵. 我们提出的模型的大小仅为 2.28 MB, 参数量如表 2 所示. 表 3 展示了我们的工作与几年来的一些工作之间的比较, 表 3 对比了预测精度和模型的大小. 从表 3 可以看出, 本文提出的 LFCNN 在轻量级方面拥有绝对的优势, 并且有着更高的准确率, 足以证明我们方法的优越性.

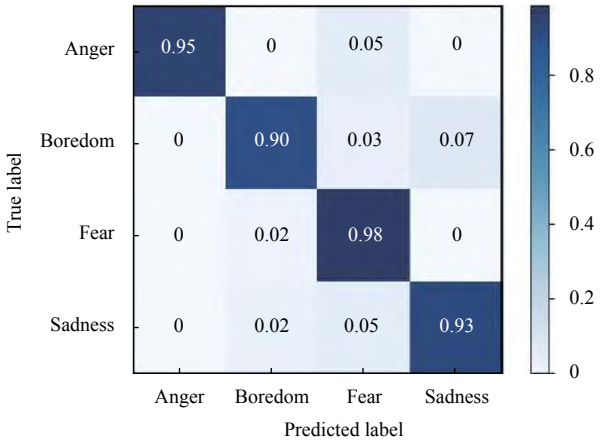



图 6 LFCNN 在 EMO-DB 上实验得到的混淆矩阵

表 2 LFCNN 的参数量统计

| 图2所示模块 | 参数量 |
|---|---|
| 并行卷积结构 | 752 |
| 残差结构 | 20 752 |
| 串行卷积结构 | 153 232 |
| 其他 | 4 884 |
| 总参数量 | 179 620 |

表 3 与近几年在 EMO-DB 上的研究对比

| 文献 | 验证方案 | 准确率 (%) | 模型大小 (MB) |
|---|---|---|---|
| [8] | 10-fold | 82.82 | 323.46 |
| [23] | 5-fold | 85.57 | 128.00 |
| [17] | 5-fold | 88.70 | 31.20 |
| [24] | 5-fold | 93.00 | 14.40 |
| [25] | 10-fold | 85.55 | — |
| 本文 | 10-fold | **94.36** | **2.28** |

### 3.3 CK+实验

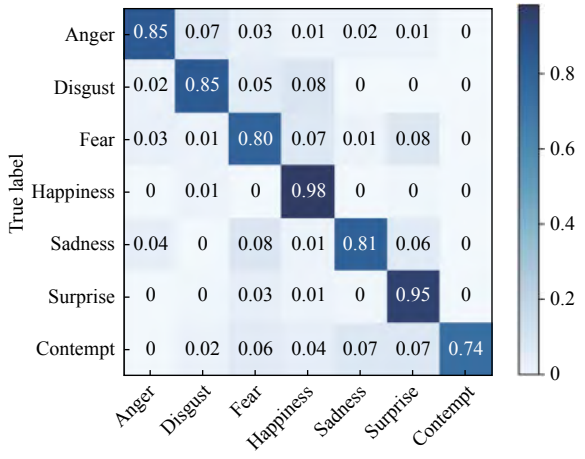
#### 3.3.1 实验步骤

为验证改进后的 GhostNet 在人脸表情识别方面的优越性, 我们使用改进前后的 GhostNet 网络在 CK+人脸表情数据集上进行了验证. CK+数据集中只有 327 个有效序列 (即包含情绪标签的数据), 若一个序列仅提取一张图片, 那么数据量将会太少, 不利于我们训练模型, 因此我们提取了每个视频的最后 3 帧以扩大数据量. 所以在 CK+数据集中总共提取了 981 张

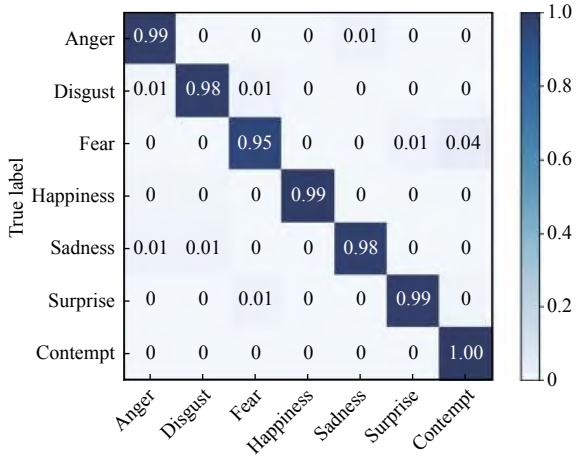
有效人脸图像, 每张图像都是大小为 48×48 的灰度图像. 一次训练过程包含 250 个 epochs, batchsize 设置为 64, 使用固定学习率为 $10^{-3}$ 的 Adam 优化器训练. 在 10 折交叉验证中, 记录每折训练结束后 7 种情绪在测试集上的预测正确的样本数量, 并在 10 折交叉验证结束后得到由这 7 种情绪组成的混淆矩阵.

3.3.2 实验结果

GhostNet 和我们改进的 GhostNet 在 CK+人脸表情识别数据集上得到的 7 种情绪的混淆矩阵如图 7(a) 和图 7(b) 所示. 实验结果表明, 使用原始的 GhostNet 训练得到的结果平均准确率只能达到 90.21%, 而改进后的 GhostNet 达到了平均 98.27% 的准确率, 这足以说明我们改进后的方法是有效的.




(a) 改进前的 GhostNet




(b) 改进后的 GhostNet

图 7 改进前后的 GhostNet 在 CK+数据集上得到的混淆矩阵

此外, 我们发现 GhostNet 网络在训练过程中过拟合现象非常严重. 图 8(a) 展示了一次训练过程中训练集和验证集的准确率和损失值随 epoch 的变化曲线. 为缓解过拟合现象, 我们在修改 Ghost bottleneck 结构的同时引入了多个 Dropout 层来缓解过拟合现象, 图 8(b) 显示了改进后的 GhostNet 在训练过程中准确率和损失值随着 epoch 变化的曲线, 相比改进之前大大地缓解了过拟合现象. 改进后的 GhostNet 平均准确率达到了 98.27% 的准确率, 但恐惧表达的准确率只有 95%, 这可能与恐惧的数据量较少以及恐惧表情的特征和蔑视具有相似性有关. 尽管如此, 我们提出的方法在最近的研究中也取得了先进的成果, 表 4 显示了与最近的一些研究的比较, 从表中可以看出我们提出的改进的 GhostNet 优于其他经典分类模型, 这充分证明了我们提出的方法的优越性.

表 4 与近几年在 CK+数据集上的研究对比

| 文献 | 模型 | 验证方案 (折) | 准确率 (%) |
|---|---|---|---|
| [5] | Xception | 10-fold | 98.20 |
| [6] | Inceptionv3 | — | 94.20 |
| [26] | MobileNet | 10-fold | 96.00 |
| [4] | ResNet50 | 5-fold | 89.80 |
| 本文工作1 | 改进前* | 10-fold | 90.21 |
| 本文工作2 | 改进后* | **10-fold** | **98.27** |

注: *指GhostNet.

### 3.4 MAHNOB-HCI 实验

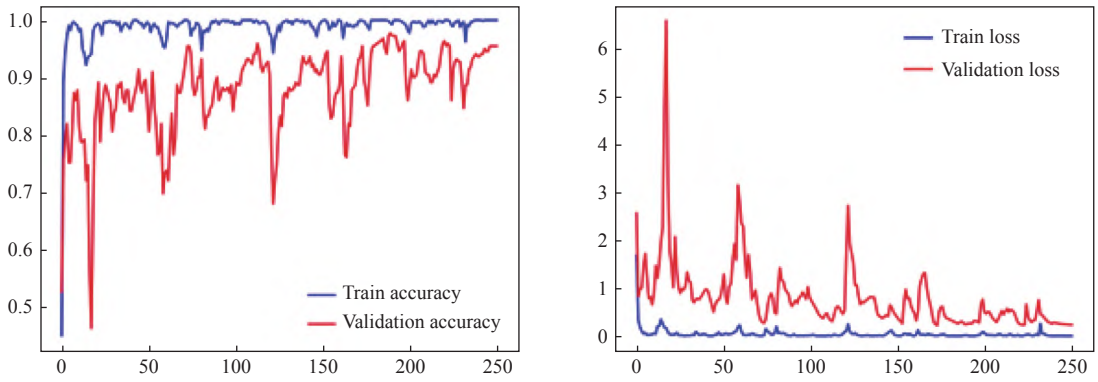
我们在 MAHNOB-HCI 数据集上验证了本文决策级融合方法. 实验采用了留一试交叉验证法, 即每个受试者保留一次试验数据作为测试集, 其他试验的数据用作训练集.

3.4.1 实验步骤

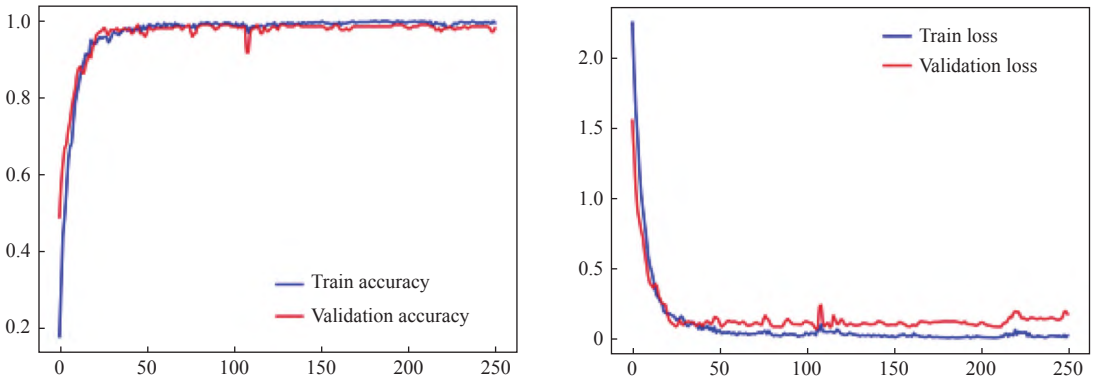
对于 EEG 数据, 我们使用 MNE 库来提取原始的 EEG 信号特征. 如第 2.2.1 节所述, 我们使用 10 s 的时间窗从 5 个频带中提取 PSD 特征, 每个样本有 85 个特征. 对于 tLSTM 模型, 树状部分叶子节点的 LSTM 神经元数量设置为 96, 其他节点的数量设置为 128, 所有 LSTM 单元隐藏层之间的 Dropout 均设置为 0.5. 对于语音数据, 处理过程与训练设置与 EMO-DB 上的实验基本相同. 对于人脸数据, 我们使用 OpenCV 库每 10 帧捕获一次图像并将图像转换为灰度图像, 然后根据第 2.3.1 节描述的方法, 在图像上进行人脸检测和旋转对齐, 最后将图像大小调整为 48×48 并保存在 numpy 数组中用于模型训练. 此外, 对于人脸表情识别的模型训练, 我们将首先使用 FER2013 数据集进行预

训练, 然后再使用 MAHNOB-HCI 的人脸数据进行模型微调.

值得注意的是, 在 MAHNOB-HCI 上的实验需要将模型的输出改为两个分数, 模型训练的损失函数也需要变化. MAHNOB-HCI 情绪的描述是使用唤醒效价二维模型, 分数区间为 1–9, 我们将分数分为高 (≥5) 和低 (<5) 两个类别来评估我们模型的性能, 这也是在对唤醒效价情感模型研究中使用最广泛的方法[27]. 3 个模态的预测分数得到之后, 分别使用传统的融合方法以及我们提出的最优权重分布算法进行加权融合, 最后根据分数将结果划分为高或者低两类, 并与真实结果进行比较, 得到最终的多模态情绪识别准确率.




(a) 改进前的 GhostNet




(b) 改进后的 GhostNet

图 8 改进前后的 GhostNet 在 CK+数据集上训练得到的准确率与损失值随 epoch 变化曲线 (横坐标: epoch, 纵坐标 (左): 准确率, 纵坐标 (右): 损失值)

### 3.4.2 实验结果

在 MAHNOB-HCI 数据集上获得的实验结果表明, tLSTM 模型在脑电情绪识别以及我们提出的最优权重分布算法在决策级融合中均取得了较好的结果. 图 9 显示了被试 1 号到被试 15 号各方法的验证结果, 表 5 展示了各方法得到的平均准确率. 从中可以得知我们提出的融合方法在唤醒维度和效价维度上都达到了很高的精度, 并且我们提出的最优权重分布算法相比于传统的融合算法提高了分类准确率. 需要注意的是, 融合结果不一定比某一模态更准确, 例如被试 2 号和被试 13 号的人脸表情识别准确率高于融合后的识别结果, 这是因为融合结果综合考虑了多种模态的识别结果, 能够适应更多情形. 多模态情绪识别方法的意义不仅是为了提高识别准确率, 还要考虑考虑到方法的鲁棒性. 例如, 当受试者表达与真实情绪不同的人脸表情时, 多模态融合得到的结果不会与真实情绪有太大的偏差, 因为受试者的脑电图仍然代表了他们真实的情绪状态. 此外, 语音情绪识别在 MAHNOB-HCI 中是一项非常具有挑战性的任务, 因为数据集中的语音信号不仅包括被试者的声音, 还包括大量刺激材料的声音, 以至于难以对被试者的声音进行提纯, 这使得我们很难在该数据集的语音方面中实现高识别率.

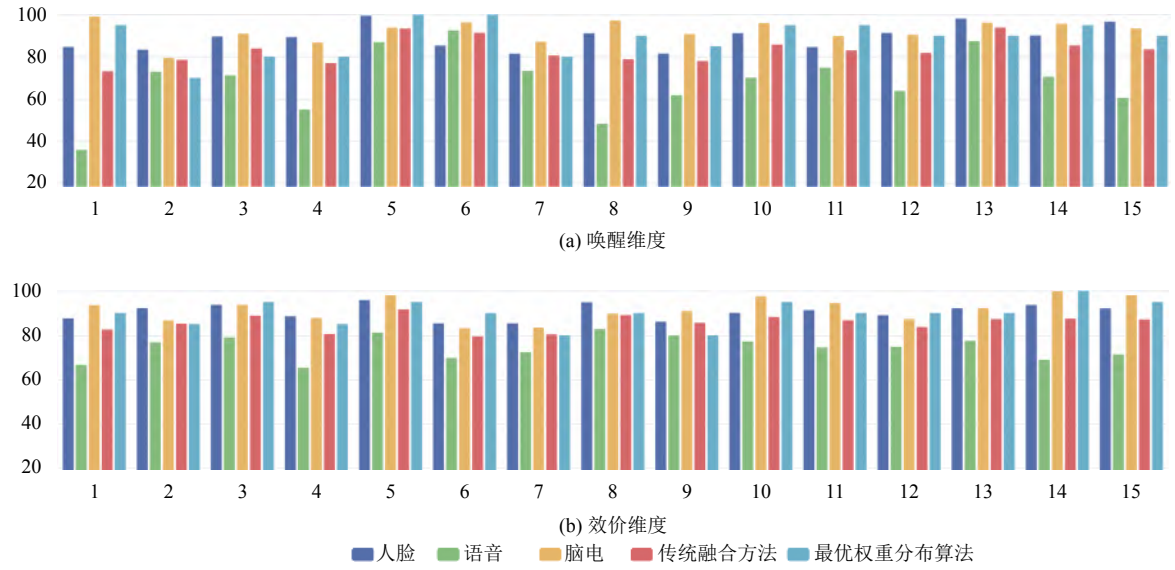



图 9 所提出的方法在 MAHNOB-HCI 中对唤醒效价两个维度的分类准确率情况 (横坐标: 被试者编号, 纵坐标: 分类准确率 (%))

表 5 MAHNOB-HCI 数据集上各方法得到的准确率 (%)

| 维度 | 人脸 | 语音 | 脑电 | 传统融合 | 最优权重分布算法 |
|---|---|---|---|---|---|
| 唤醒 | 89.17 | 70.15 | 92.19 | 84.68 | **90.25** |
| 效价 | 90.51 | 74.53 | 91.74 | 82.19 | **89.33** |

## 4 总结与展望

在本文所介绍的工作中, 我们使用深度学习技术分别为语音、脑电以及人脸表情设计了 3 个情绪识别模型. 对于语音信号, 我们设计的 LFCNN 在保证高识别准确率的前提下, 大大地减小了模型参数量. 针对脑电信号, 设计了一个 tLSTM 模型, 可以更好地学习每个阶段的情绪特征. 对于人脸表情识别, 我们将 GhostNet 应用到该领域中来, 并对 GhostNet 进行了改进, 有效解决了模型过拟合的问题. 针对多模态融合方式, 我们设计了一种最优权重分布搜索算法来搜寻每种模态的可靠性并实现决策级融合.

本文方法均在开源数据集中得到了验证, 并取得了先进的效果. 目前, 包括本文在内的研究提出的决策级融合方法大多都为每种模态设置了固定的权重, 在极端情况下可能会影响最终的识别结果. 因此, 有必要探索一种可以为每种模态动态分配权重的方法, 以提高算法的整体鲁棒性. 此外, 考虑到在未来应用当中脑电采集设备一般都会采用便携式设备, 而便携式脑电采集设备可供选择的通道有限, 因此需要研究在使用少量通道的前提下尽可能地提高情绪识别准确率的方法.

## 参考文献

1 潘家辉, 何志鹏, 李自娜, 等. 多模态情绪识别研究综述. 智能系统学报, 2020, 15(4): 633–645. [doi: 10.11992/tis.202001032]

2 Russell JA. Affective space is bipolar. Journal of Personality and Social Psychology, 1979, 37(3): 345–356. [doi: 10.1037/0022-3514.37.3.345]

3 Mehrabian A. Communication without words. Communication Theory. Routledge, 2017: 193–200.

4 Mishra S, Joshi B, Paudyal R, *et al*. Deep residual learning for facial emotion recognition. In: Shakya S, Bestak R, Palanisamy R, *et al*., eds. Mobile Computing and Sustainable Informatics. Singapore: Springer, 2022. 301–313. [doi: 10.1007/978-981-16-1866-6_22]

5 Nasri MA, Hmani MA, Mtibaa A, *et al*. Face emotion recognition from static image based on convolution neural networks. Proceedings of the 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). Sousse: IEEE, 2020. 1–6. [doi: 10.1109/atsip49331.2020.9231537]

6 Chowdary MK, Nguyen TN, Hemanth DJ. Deep learning-based facial emotion recognition for human-computer interaction applications. Neural Computing and Applications, 2021: 1–18. [doi: 10.1007/s00521-021-06012-8.]

7 Han K, Wang Y, Tian Q, *et al*. GhostNet: More features from cheap operations. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 1580–1589. [doi: 10.48550/arXiv.1911.11907]

8 Chen MY, He XJ, Yang J, *et al*. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Processing Letters, 2018, 25(10): 1440–1444. [doi: 10.1109/lsp.2018.2860246]

9 Zhang H, Gou RY, Shang JL, *et al*. Pre-trained deep convolution neural network model with attention for speech emotion recognition. Frontiers in Physiology, 2021, 12: 643202. [doi: 10.3389/fphys.2021.643202]

10 Alhagry S, Fahmy AA, El-Khoribi RA. Emotion recognition based on EEG using LSTM recurrent neural network. Emotion, 2017, 8(10): 355–358. [doi: 10.14569/ijacsa.2017.081046]

11 Wu X, Zheng WL, Li ZY, *et al*. Investigating EEG-based functional connectivity patterns for multimodal emotion recognition. Journal of Neural Engineering, 2022, 19(1): 016012. [doi: 10.1088/1741-2552/ac49a7]

12 Li JC, Li SQ, Pan JH, *et al*. Cross-subject EEG emotion recognition with self-organized graph neural network. Frontiers in Neuroscience, 2021, 689: 611653. [doi: 10.3389/fnins.2021.611653]

13 Zhou HS, Du J, Zhang YY, *et al*. Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 2617–2629. [doi: 10.1109/taslp.2021.3096037]

14 Ma YX, Hao YX, Chen M, *et al*. Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. Information Fusion, 2019, 46: 184–192. [doi: 10.1016/j.inffus.2018.06.003]

15 Li RX, Liang Y, Liu XJ, *et al*. MindLink-Eumpy: An open-source Python toolbox for multimodal emotion recognition. Frontiers in Human Neuroscience, 2021, 15: 621493. [doi: 10.3389/fnhum.2021.621493]

16 Wang M, Huang ZY, Li YC, *et al*. Maximum weight multi-modal information fusion algorithm of electroencephalographs and face images for emotion recognition. Computers & Electrical Engineering, 2021, 94: 107319. [doi: 10.1016/j.compeleceng.2021.107319]

17 Muppidi A, Radfar M. Speech emotion recognition using quaternion convolutional neural networks. Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021. 6309–6313. [doi: 10.1109/icassp39728.2021.9414248]

18 Bhatnagar G, Wu QMJ, Raman B. A new fractional random wavelet transform for fingerprint security. IEEE Transactions on Systems, Man, and Cybernetics —Part A: Systems and Humans, 2012, 42(1): 262–275. [doi: 10.1109/tsmca.2011.2147307]

19 Jenke R, Peer A, Buss M. Feature extraction and selection for emotion recognition from EEG. IEEE Transactions on Affective Computing, 2014, 5(3): 327–339. [doi: 10.1109/taffc.2014.2339834]

20 Huang HY, Xie QY, Pan JH, *et al*. An EEG-based brain computer interface for emotion recognition and its application in patients with disorder of consciousness. IEEE Transactions on Affective Computing, 2021, 12(4): 832–842. [doi: 10.1109/TAFFC.2019.2901456]

21 Paoletti ME, Haut JM, Pereira NS, *et al*. GhostNet for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(12): 10378–10393. [doi: 10.1109/tgrs.2021.3050257]

22 Huang XH, Kortelainen J, Zhao GY, *et al*. Multi-modal emotion analysis from facial expressions and electroencephalogram. Computer Vision and Image Understanding, 2016, 147: 114–124. [doi: 10.1016/j.cviu.2015.09.015]

23 Mustaqeem, Sajjad M, Kwon S. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE Access, 2020, 8: 79861–79875. [doi: 10.1109/access.2020.2990405]

24 Mustaqeem, Kwon S. Att-Net: Enhanced emotion recognition system using lightweight self-attention module. Applied Soft Computing, 2021, 102: 107101. [doi: 10.1016/j.asoc.2021.107101]

25 Andayani F, Theng LB, Tsun MT, *et al*. Hybrid LSTM-transformer model for emotion recognition from speech audio files. IEEE Access, 2022, 10: 36018–36027. [doi: 10.1109/ACCESS.2022.3163856]

26 Priya RNB, Hanmandlu M, Vasikarla S. Emotion recognition using deep learning. Proceedings of the 2021 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). Washington: IEEE, 2021. 1–5. [doi: 10.1109/AIPR52630.2021.9762207]

27 Li DH, Yang ZY, Hou FZ, *et al*. EEG-based emotion recognition with haptic vibration by a feature fusion method. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 2504111. [doi: 10.1109/tim.2022.3147882]

(校对责编: 孙君艳)