计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(12):218−225 [doi: 10.15888/j.cnki.csa.008235]

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

# 基于视听觉感知系统的多模态情感识别①

龙英潮, 丁美荣, 林桂锦, 刘鸿业, 曾碧卿

(华南师范大学 软件学院, 佛山 528225)
通讯作者: 丁美荣, E-mail: 362034935@qq.com

**摘　要**: 情绪识别作为人机交互的热门领域, 其技术已经被应用于医学、教育、安全驾驶、电子商务等领域.情绪主要由面部表情、声音、话语等进行表达, 不同情绪表达时的面部肌肉、语气、语调等特征也不相同, 使用单一模态特征确定的情绪的不准确性偏高, 考虑到情绪表达主要通过视觉和听觉进行感知, 本文提出了一种基于视听觉感知系统的多模态表情识别算法, 分别从语音和图像模态出发, 提取两种模态的情感特征, 并设计多个分类器为单特征进行情绪分类实验, 得到多个基于单特征的表情识别模型. 在语音和图像的多模态实验中, 提出了晚期融合策略进行特征融合, 考虑到不同模型间的弱依赖性, 采用加权投票法进行模型融合, 得到基于多个单特征模型的融合表情识别模型. 本文使用 AFEW 数据集进行实验, 通过对比融合表情识别模型与单特征的表情识别模型的识别结果, 验证了基于视听觉感知系统的多模态情感识别效果要优于基于单模态的识别效果.

**关键词**: 情感识别; 模型融合; 多模态; 视听觉感知系统

引用格式: 龙英潮,丁美荣,林桂锦,刘鸿业,曾碧卿.基于视听觉感知系统的多模态情感识别.计算机系统应用,2021,30(12):218–225. http://www.c-s-a.org.cn/1003-3254/8235.html

## Emotion Recognition Based on Visual and Audiovisual Perception System

LONG Ying-Chao, DING Mei-Rong, LIN Gui-Jin, LIU Hong-Ye, ZENG Bi-Qing

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

**Abstract**: As a hot spot of human-computer interaction, emotion recognition has been applied in many fields, such as medicine, education, safe driving and e-commerce. Emotions are mainly expressed by facial expression, voice, discourse and so on. Other characteristics such as facial muscles, mood and intonation vary when different kinds of emotions are expressed. Thus, the inaccuracy of emotions determined using a single modal feature is high. Considering that the expressed emotions are mainly perceived by vision and hearing, this study proposes a multimodal expression recognition algorithm based on an audiovisual perception system. Specifically, the emotion features of speech and image modalities are first extracted, and a plurality of classifiers are designed to perform emotion classification experiments for a single feature, from which multiple expression recognition models based on single features are obtained. In the multimodal experiments of speech and images, a late fusion strategy is put forward for feature fusion. Taking into account the weak dependence of different models, this work uses the weighted voting method for model fusion and obtains the integrated expression recognition model based on multiple single-feature models. The AFEW dataset is adopted for facial expression recognition in this study. The comparison of recognition results between the integrated model and the single-feature models for expression recognition verifies that the effect of multimodal emotion recognition based on the audiovisual

① 基金项目: 国家自然科学基金 (61876067); 广东省普通高校人工智能重点领域专项 (2019KZDZX1033); 广东省信息物理融合系统重点实验室建设专项 (2020B1212060069)
Foundation item: National Natural Science Foundation of China (61876067); Special Project in Key Areas of Artificial Intelligence in University of Guangdong Province (2019KZDZX1033); Construction Project of Guangdong Provincial Key Laboratory of Cyber-Physical Systems (2020B1212060069)
收稿时间: 2021-03-05; 修改时间: 2021-04-07; 采用时间: 2021-04-20

perception system is better than that of single-modal emotion recognition.

**Key words**: emotion recognition; model fusion; multimodal; audiovisual perception system

## 1 引言

随着信息处理技术、网络通信技术、大数据、人工智能等科技的迅猛发展, 计算机正逐渐地融入到人类的生活中, 并与人类协同工作. 在某些领域, 计算机甚至已经取代人类去完成各种高挑战性的工作. 为了让人与计算机能够更加高效地协同工作, 更加智能、自然地交互, 新型的人机交互 (Human Machine Interaction, HMI) 技术已经成为社会各行各业关注和研究的热点. 拟人化必然是新型人机交互技术发展的重点, 不仅要使计算机能够通过类似于人的感官系统感知周围环境、气氛, 以及使用者的意图、情感等, 还要使其能够通过学习和模仿人类的认知习惯与人类进行交流、工作等. 研究表明, 在人机交互中需要解决的相互理解的问题, 与人和人交流中相互影响的重要因素是一致的, 最关键的因素都是“情感智能”的能力[1–3]. 具有“情感智能”能力的计算机能够高效地识别使用者的情感, 从而调整与使用者的交流方式与环境, 实现更加智能、自然的交互.

近年来, 情感识别技术逐渐被应用在医学、教育、安全驾驶、电子商务等领域. 例如, 在教育领域, 智能教育系统通过分析学生们的学习情绪, 反馈学习数据, 老师可以根据相关数据调整教学模式, 以满足学生个性化学习的需求, 提高学习效率与效果; 在安全驾驶领域, 计算机使用情感识别技术分析驾驶者的情绪, 可以根据驾驶者的情绪变化调整车速上限、规划路线等, 从而避免危险的发生, 保证驾驶者的安全. 随着人机交互领域的不断扩张和情感识别领域的不断发展, 情感识别技术的应用也越来越广.

## 2 情感识别研究现状

美国心理学家梅拉比安认为, 情感表达=55% 面部表情+38% 声音+7% 其它[4,5]. 人的情感主要通过面部状态、声音以及文字等方式进行表达. 从生物角度来看, 人类主要通过视听觉感知系统来进行情感识别, 即是主要在语音和图像两种模态上进行情感识别. 在语音和图像多模态情感识别的研究中, 许多学者已经取得了一定的研究成果.

在语音模态上, 曹鹏等使用 Mallat 塔式算法与小波变换奇异点检测算法相结合进行基音频率参数提取, 并通过实验证实了该算法的有效性[3]. 屠彬彬等提出了一种基于样本熵与 Mel 频率倒谱系数融合的语音情感识别方法, 得到了较高的识别率[6]. 姚增伟等通过提取 Mel 频率倒谱系数作为输入, 分别使用卷积神经网络和长短时记忆网络进行特征提取, 并且在 IEMOCAP 语音情感语料库中获得 51.7% 的准确率[7].

在面部图像模态上, 邹元彬等在 JAFFE 数据集上分别提取面部图像的局部二值模式 LBP 和局部相位量 LPQ 特征, 并使用支持向量机 SVM 作为分类器进行实验, 得到了 90.57% 的识别率[8]. 陈津徽等提出了一种基于改进 VGG19 网络的人脸表情识别算法, 并在 FER2013 数据集上得到了 72.69% 的准确率[9].

在语音和面部图像的多模态研究中, 朱晨岗基于视听觉感知系统, 分别使用基于 Mel 尺度小波包分解的子带能量特征基于光流法提取的运动特征等, 并用循环神经网络作为分类器进行多模态情感分类实验[10]. 贺奇基于语音和图像进行多模态情感识别研究, 分别使用 92 维语音情感特征和基于序列图像脸部特征点提取方法提取的表情图像特征进行实验, 并验证了基于语音和图像的多模态情感识别比单一模态的识别效果更好[11]. 袁亮通过深度学习技术进行情感识别研究, 分别提出了一种基于卷积神经网络和循环神经网络的面部表情识别方法和一种基于长短时间记忆网络和卷积神经网络的语音情感识别方法, 并通过决策融合算法融合面部表情和语音模态的特征进行实验, 同样验证了多模态的情感识别效果要优于单模态的情感识别效果[12]. 因此, 多模态情感识别研究具有可行性, 同时从以上文献成果可以发现, 多模态融合实现的方法和实验选择是比较灵活的. 融合的目的就是将各单模态下能反应情感的特征数据合并成一个性能更优的数据结果. 所以, 可以基于相同实验数据, 尝试对两种模态进行早期融合或晚期融合, 还可以通过调整其融合权重, 灵活选择实验测试方法, 以达到更加精确的识别率.

## 3 研究原理和实现方法

本文对于情感识别的研究主要也是在语音和图

像两种主流模态上进行, 首先将视频样本切分为语音和图像数据, 然后分别提取两种模态的情感特征, 并使用多个分类器进行实验, 得到多个基于单特征的表情识别模型, 最后采用晚期融策略进行模型融合, 得到最优的集成表情识别模型, 实验的主要流程如图 1 所示.

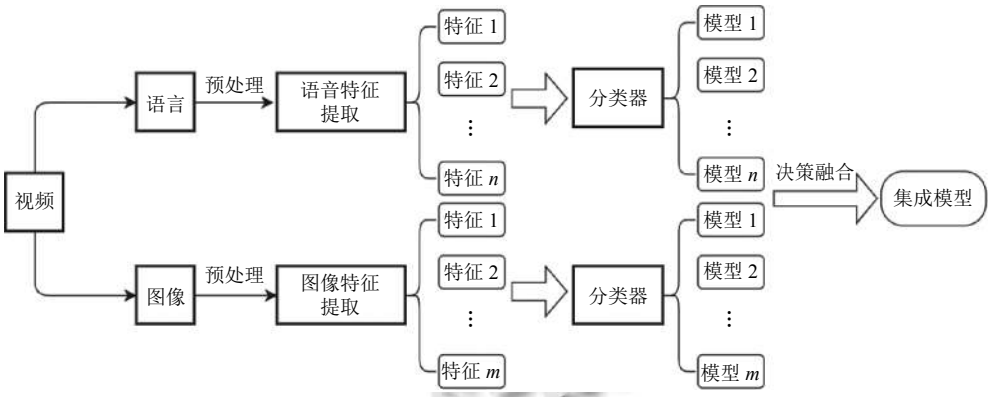


图 1 实验流程图

### 3.1 语音特征提取

语音模态的特征主要包括 Mel 频率倒谱系数 MFCC、SoundNet 卷积神经网络提取的特征以及 IS09、IS11、IS13 等帧级特征, 其中 IS09、IS11、IS13 等帧级特征使用 openSMILE 工具提取.

(1) Mel 倒谱系数 MFCC

Mel 频率倒谱系数 MFCC 的提取过程[13–15] 如下:

首先, 对采样得到的一帧离散语音序列 $x(n)$ 作快速傅里叶变换 FFT, 快速傅里叶变换的公式如下:

$$X(k)=\sum_{n=0}^{N-1}x(n)\mathrm{e}^{-j\frac{2\pi}{N}nk},k=0,1,2,\cdots,N-1 \tag{1}$$

其中, $N$ 为帧长.

其次, 配置 Mel 滤波器组并计算滤波输出, Mel 滤波器的频率响应 $H_m(k)$ 为:

$$H_m(k)=\begin{cases}0, & k<f(m-1)\\ \dfrac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))}, & f(m-1)\le k\le f(m)\\ \dfrac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m)-f(m-1))}, & f(m)\le k\le f(m+1)\\ 0, & k\ge f(m+1)\end{cases} \tag{2}$$

其中, $f(m)$ 为滤波器的中心频率.

然后, 计算每个滤波器组输出的对数能量 $S(m)$.

$$S(m)=\ln\left(\sum_{k=0}^{N-1}|X_a(k)|^2H_m(k)\right),\ 0\le m\le M \tag{3}$$

其中, $M$ 为滤波器的个数.

最后, 经离散余弦变换 DCT 可得到 MFCC 系数 $C(n)$, 公式描述如下:

$$C(n)=\sum_{m=0}^{N-1}S(m)\cos\left(\frac{\pi n(m-0.5)}{M}\right),\ n=1,2,\cdots,L \tag{4}$$

其中, $L$ 为 MFCC 系数的阶数.

(2) SoundNet 卷积神经网络

SoundNet 网络是一种具有较高语音信息学习能力的深度卷积神经网络[16], 其实现的基本原理如下:

首先将视频切割音频和 RGB 图像帧两部分, RGB 图像帧部分分别使用了图像类卷积神经网络 ImageNet CNN 和场景类神经网络 Places CNN 进行识别分类, 并将 RGB 图像帧分类的结果作为 SoundNet 网络的监督信息, 从而可以学习得到语音的相关信息. SoundNet 网络由 8 层卷积层和 3 层池化层组成, 损失函数为 KL 散度. 图 2 为 SounNet 网络结构图, 其中 conv $n$ 代表第 $n$ 层卷积层, pool $n$ 代表第 $n$ 层池化层, 下同.

### 3.2 图像特征提取

图像模态的特征主要包括使用 DenseNet、VGG 等卷积神经网络提取的特征, 以及 LBP-TOP 特征描述子.

(1) DenseNet 网络

DenseNet 网络采用了一种密集连接的模式, 不需要重新学习冗余的特征映射, 具有减轻梯度消失、加强特征的传递以及高效利用特征等优点. 本文实验中

使用的是 DenseNet 网络中的一个特殊网络 DenseNet-BC 网络.

DenseNet-BC 网络是包含了 Bottleneck layer 瓶颈层和 Transition layer 过渡层的特殊 DenseNet 网络结构, 其中, 过渡层即由一层卷积层和一层池化层组成的网络层. DenseNet-BC 网络包含了 3 个 Dense Block 和 2 层过渡层. 图 3 为 DenseNet-BC 网络结构图, 其中 Dense block *n* 代表第 *n* 个密集块.

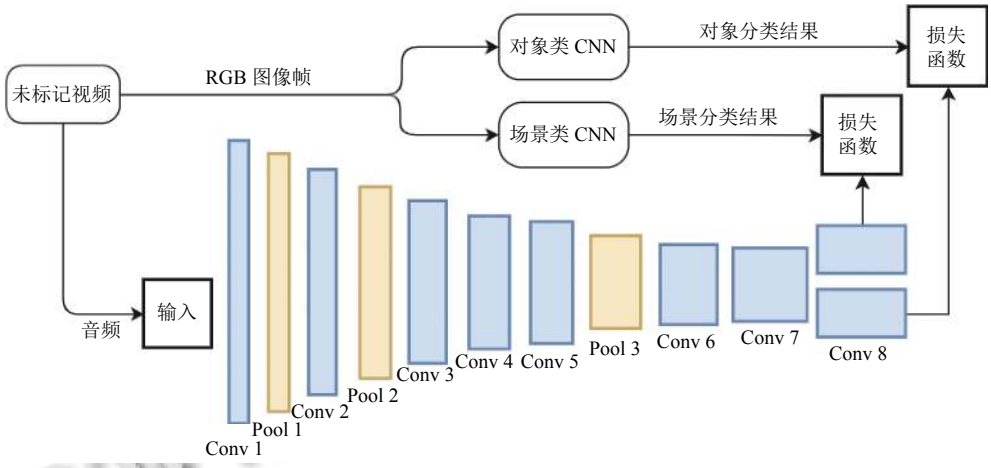


图 2 SoundNet 网络结构及实现原理图

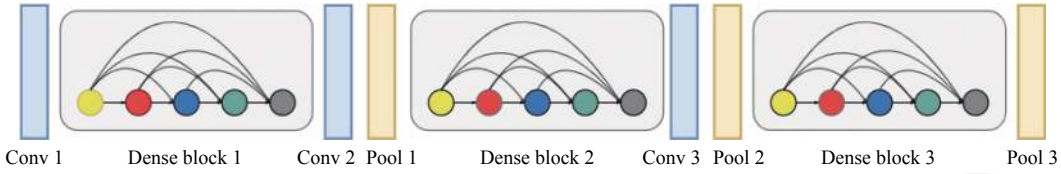


图 3 DenseNet-BC 网络结构图

(2) VGG 网络

VGG 网络是使用 3×3 小卷积核和 2×2 最大池化层的深度卷积神经网络, 并且极大地提升了网络的深度, 其独特的结构特点在很大程度上提高了神经网络的学习能力. 本文实验中使用的是 VGG 系列网络中的 VGG-16 网络. VGG-16 网络具有 13 个卷积层、5 个池化层和 3 个全连接层. 图 4 为 VGG-16 网络结构图, 其中 Fc *n* 代表第 *n* 层全连接层.

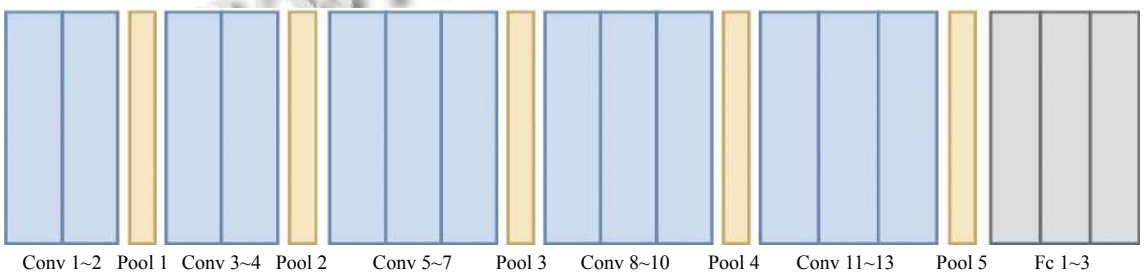


图 4 VGG-16 网络结构图

### 3.3 分类器的选择与设计

在分类器方面, 本文选择了多种分类器进行实验, 包括支持向量机 SVM 和随机森林 RF 等经典分类器, 同时, 考虑到在将视频样本切分为图像样本时, 得到的是长序列的图像帧, 而长短期记忆网络 LSTM 在处理长序列数据具有较显著的优势, 所以设计了一个基于

LSTM 的分类器用以实验.

(1) 基于 LSTM 的分类器设计

基于 LSTM 设计的分类器的结构如图 5 所示, 输入序列 $X$ 为不同时间的特征, 在输入层后添加一层批标准化层, 多个 LSTM 结点组成的 LSTM 阵列进行特征信息的捕获, 通过平均池化层对不同时间的特征信息平均并输出到 Softmax 层进行分类.

### 3.4 决策层融合

在多模态情绪识别领域, 加权投票法、加权平均法是较为常见的决策融合方法, 其中, 投票法更适用于决策融合中的各模型相互独立的情况. 考虑到本文中各个模型的训练都相互独立, 并不存在强依赖关系, 采用加权投票法进行决策融合可能会带来一定的提升. 加权投票法具体实现如下:

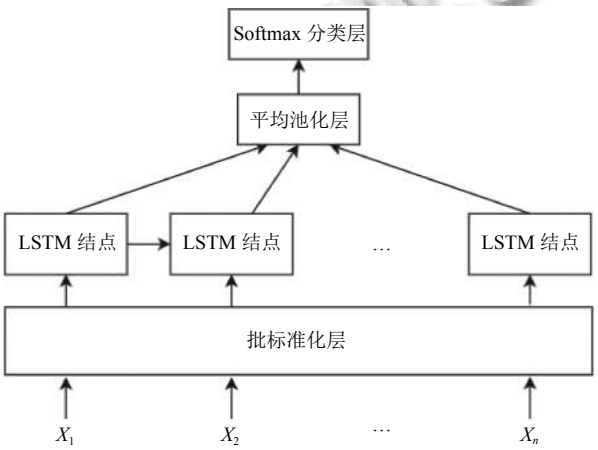


图 5 基于 LSTM 的分类器结构图

设表情类别数目为 $M$, 模型的数量为 $L$, $h_i$ 为第 $i$ 个情感识别模型, $w_i$ 第 $i$ 个模型对融合模型决策结果的贡献权重, 其中, $w_i$ 的约束为:

$$\begin{cases} \sum_{i=1}^{L} w_i = 1 \\ w_i \in [0,1],\ i = 1,2,\cdots,L \end{cases} \tag{5}$$

对于样本 $x$, 设 $f(x)$ 是基于加权投票法得到的各种表情类别的加权投票值的集合, $y(x)$ 是表情类别的决策结果, 则有:

$$f_j(x) = \sum_{i=1}^{L} w_i I(h_i(x), j) \tag{6}$$

$$y(x) = \arg\max f_j(x) \tag{7}$$

其中, $j$=1, 2, ⋯, $M$, 指示函数 $I$ 的定义为:

$$I(a,b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases} \tag{8}$$

权重的学习使用基于个体分类模型相对优势的投票权重学习方法[17].

## 4 实验过程及结果

本文主要基于 AFEW 数据集[18] 来进行多模态情感识别研究. 实验首先将 AFEW 数据集的视频数据切分为音频数据和图片数据, 分别进行语音、图像模态的情感特征提取.

### 4.1 音频数据处理和特征提取

在提取语音特征前, 需要对音频文件进行重采样、分帧和加窗 3 个预处理操作, 其中帧长为 25 ms, 帧移为 10 ms, 窗函数为汉明窗, 然后提取 MFCC、IS09、IS11、IS13 等段级特征, 其中段长度为整个话语段的长度, IS09、IS11、IS13 是基于 openSMILE 工具包提取的. 在使用 SoundNet 卷积神经网络提取特征时, 把音频文件的原始数据作为输入, 提取后的特征标记为 SoundNet.

### 4.2 图像数据处理和特征提取

在提取图像特征时, 首先要对图像进行人脸检测和人脸的校正裁剪两个预处理操作. 由于 AFEW 数据集的作者已经提供了大部分已经裁剪好的人脸灰度图像, 未提供的图片数据仅为 Train 训练集下的 17 个视频和 Val 验证集下的 12 个视频. 因此, 我们仅对未提供的图片数据进行预处理操作, 在成功提取人脸灰度图像后仍需进行直方图均衡化处理, 以减轻灯光对图像的影响.

完成预处理操作后, 我们将预训练后的 DenseNet-BC 和 VGG16 卷积神经网络模型在 FRE2013 数据集上微调, 然后将预处理后的图像作为微调后的模型的输入来提取图像特征. 使用 DenseNet-BC 卷积神经网络提取特征时, 将 DenseNet-BC 网络的最后一个平均池化层的输出作为特征, 该特征被标记为 DenseNet-pooling3. 使用 VGG 卷积神经网络提取特征时, 将 VGG-16 网络的第 13 层卷积层和第 1 层全连接层的输出作为特征, 分别被标记为 VGG-conv13、VGG-fc1.

基于 LBP-TOP 特征描述子提取的特征已经被 AFEW 数据集的作者提供, 将该特征标记为 LBP-TOP.

### 4.3 实验结果与分析

在完成语音和图像模态的特征提取后, 使用支持

向量机 SVM、随机森林 RF 以及基于 LSTM 的分类器进行表情分类, 得到多个基于音频、图像单特征表情识别模型.

(1) 基于语音单特征模型的分类结果, 如表 1 所示.

表 1 基于音频单特征表情识别模型及其准确率

| 模型编号 | 特征 | 分类器 | 准确率 (%) |
|---|---|---|---|
| 1 | MFCC | **LSTM** | **24.80** |
| 2 | MFCC | SVM | 21.40 |
| 3 | IS09 | RF | 32.11 |
| 4 | IS11 | RF | 22.98 |
| 5 | IS13 | RF | 20.89 |
| 6 | SoundNet | RF | 25.07 |
| 7 | SoundNet | **LSTM** | **31.33** |

通过分析实验结果数据, 可以得出以下几点结论:

① LSTM 分类器在语音特征 MFCC、SoundNet 上相较于支持向量机 SVM、随机森林 RF 等分类器有着 3.4%~6.26% 准确率提升;

② 在语音单特征模型中, 基于 IS09 特征的模型取得最高准确率为 32.11%.

(2) 基于图像单特征模型的分类结果, 如表 2 所示.

表 2 基于图像单特征表情识别模型及其准确率

| 模型编号 | 特征 | 分类器 | 准确率 (%) |
|---|---|---|---|
| 8 | DenseNet-pooling3 | **LSTM** | **41.78** |
| 9 | DenseNet-pooling3 | SVM | 38.12 |
| 10 | VGG-conv13 | **LSTM** | **42.56** |
| 11 | VGG-conv13 | SVM | 38.64 |
| 12 | VGG-fc1 | **LSTM** | **40.73** |
| 13 | VGG-fc1 | SVM | 34.46 |
| 14 | LBP-TOP | SVM | 36.55 |

通过分析实验结果, 可以发现以下几点:

① 基于 LSTM 的分类器在图像特征 VGG-conv13、VGG-fc1、DenseNet-pooling3 上相较于分类器支持向量机 SVM 有着 3.92%~6.27% 准确率提升;

② 在图像单特征模型中, 基于 VGG-conv13 特征的模型取得最高准确率为 42.56%;

③ 基于图像单特征的最优模型比基于语音单特征的最优模型的准确率高 11.23%.

(3) 基于融合模型的分类结果, 如表 3 所示.

在进一步实验中, 使用加权投票法对多个单特征模型进行决策融合, 分别得到基于语音模态、图像模态以及语音和图像双模态的 3 种融合模型.

通过对比 3 种融合模型的分类结果, 可以发现基于音频和图像双模态融合模型的准确率达到 50.13%, 此准确率高于单模态融合模型的准确率. 该双模态融合模型在愤怒 Angry、害怕 Fear、高兴 Happy、中性 Neutral 等几种情绪上的分类准确率均达到 60% 以上, 而在厌恶 Disgust 和惊讶 Surprise 两种情绪上的识别效果较差, 其混淆矩阵数据如图 6 所示.

表 3 3 种融合模型及其准确率

| 融合模型 | 添加融合的单特征模型 | 准确率 (%) |
|---|---|---|
| 语音模态融合模型 | 1、3、4、5、7 | 35.51 |
| 图像模态融合模型 | 8、10、12、14 | 46.48 |
| 双模态融合模型 | 1、3、4、5、7、8、10、12、14 | **50.13** |

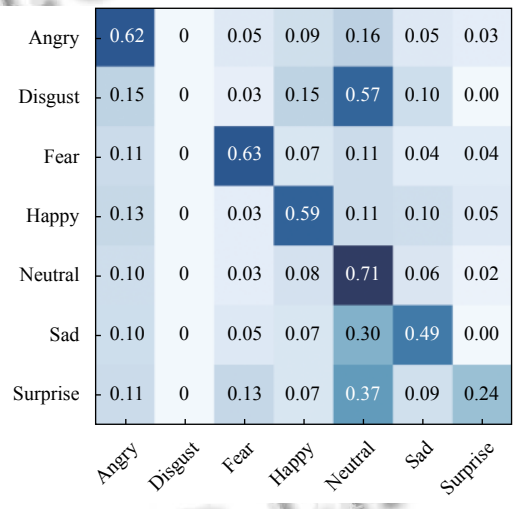


图 6 融合模型的混淆结果矩阵

### 4.4 双模态融合模型的实验结果对比与分析

情绪识别的相关研究有很多. 本文提出的一种基于视听觉感知系统的多模态表情识别算法, 在 AFEW 数据集进行实验得到了 50.13% 的准确率.

如表 4 所示, AFEW 数据集是 EmotiW 比赛的官方数据集, 该数据集的准确率基线 Baseline 为 41.07%, 在该比赛中, 参赛者的平均准确率 50% 左右, 最高准确率为 61.87%. 虽然本文的方案在准确率上低于最高的准确率, 但也保持在平均水平之上, 仍然具有一定的竞争力.

## 5 结语

本文提出了一种基于视听觉感知系统的多模态表情识别算法, 分别提取语音和图像两种模态的情感特征, 并设计多个分类器为单个情感特征进行情绪分类实验, 得到多个基于单特征的表情识别模型. 最后使用

晚期融合策略进行特征融合, 得到基于多个单特征模型的融合表情识别模型, 并通过对比实验证明语音和图像双模态融合表情识别模型的有效性.

本文使用 AFEW 数据集进行表情识别实验, 首先将 AFEW 数据集的视频数据切分为音频数据和图片数据, 分别进行语音、图像模态的情感特征提取. 在语音数据上, 提取的情感特征包括 MFCC、IS09、IS11、IS13 等段级特征, 以及使用卷积神经网络 SounNet 提取的特征. 在图片数据上, 提取的情感特征包括使用深度卷积神经网络 VGG-16 和 DenseNet 提取的特征, 以及基于 LBP-TOP 特征描述子提取的特征. 然后, 使用了多个分类器对单个特征进行了情绪分类实验, 并证明了使用基于 LSTM 分类器相较于支持向量机 SVM、随机森林 RF 等分类器对实验效果有着小幅度的提高. 最后使用加权投票法进行模态融合实验, 通过对比基于语音模态、图像模态以及语音和图像双模态的 3 种融合模型的分类结果, 证明了基于语音和图像双模态融合模型的效果要优于基于单模态融合模型的识别效果.

表 4 EmotiW 比赛: 音视频情绪分类的部分数据[19]

| 排名 | 队伍 | 准确率 (%) |
|---|---|---|
| 1 | SituTech[20] | 61.87 |
| 2 | E-HKU[21] | 61.10 |
| 3 | AIPL[22] | 60.64 |
| 3 | OL_UC | 60.64 |
| 5 | UoT | 60.49 |
| 6 | NLPR | 60.34 |
| 7 | INHA[19] | 59.72 |
| 8 | SIAT | 58.04 |
| 9 | TsinghuaUniversity | 57.12 |
| 10 | AIIS-LAB | 56.51 |
| | **Baseline** | **41.07** |

但本文仍然存在许多不足之处: 情感的体现过程一般为: 开始——高潮——结束, 情感主要体现在高潮部分, 而在音频模态实验中, 提取的特征是基于整段语音样本的, 其中包含过多冗余数据, 影响了识别的准确性, 考虑将语音样本分段或许可以有效地降低数据的冗余; 在特征融合阶段可以尝试采用特征层融合策略进行实验对比, 甚至根据应用场景或应用群体的需求, 可以考虑基于文本、声音、图像、视频等多种模态融合的情感识别实践研究.

## 参考文献


1 张会云. 语音情感识别研究综述. 信息通信, 2019, (11): 58–60. [doi: 10.3969/j.issn.1673-1131.2019.11.027]

2 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]

3 曹鹏. 语音情感识别技术的研究与实现 [ 硕士学位论文 ]. 镇江: 江苏大学, 2005. [doi: 10.7666/d.y827197]

4 林记明. 体态语言的功能及其应用. 西安外国语学院学报, 2001, 9(4): 47–51. [doi: 10.3969/j.issn.1673-9876.2001.04.013]

5 Soleymani M, Pantic M, Pun T. Multimodal emotion recognition in response to videos. IEEE Transactions on Affective Computing, 2012, 3(2): 211–223. [doi: 10.1109/T-AFFC.2011.37]

6 屠彬彬, 于凤芹. 基于样本熵与 MFCC 融合的语音情感识别. 计算机工程, 2012, 38(7): 142–144. [doi: 10.3969/j.issn.1000-3428.2012.07.047]

7 姚增伟, 刘炜煌, 王梓豪, 等. 基于卷积神经网络和长短时记忆神经网络的非特定人语音情感识别算法. 新型工业化, 2018, 8(2): 68–74. [doi: 10.19335/j.cnki.2095-6649.2018.2.009]

8 邹元彬, 乐思琦, 廖清霖, 等. 基于 LBP 和 LPQ 的面部表情识别. 信息技术与信息化, 2020, (9): 199–205. [doi: 10.3969/j.issn.1672-9528.2020.09.064]

9 陈津徽, 张元良, 尹泽睿. 基于改进的 VGG19 网络的面部表情识别. 电脑知识与技术, 2020, 16(29): 187–188. [doi: 10.14004/j.cnki.ckt.2020.3328]

10 朱晨岗. 基于视听觉感知系统的情感识别技术研究 [ 硕士学位论文 ]. 天津: 天津理工大学, 2018.

11 贺奇. 基于语音和图像的多模态情感识别研究 [ 硕士学位论文 ]. 哈尔滨: 哈尔滨工业大学, 2017.

12 袁亮. 基于深度学习的双模态情感识别 [ 硕士学位论文 ]. 南京: 南京邮电大学, 2018.

13 张钰莎, 蒋盛益. 基于 MFCC 特征提取和改进 SVM 的语音情感数据挖掘分类识别方法研究. 计算机应用与软件, 2020, 37(8): 160–165, 212. [doi: 10.3969/j.issn.1000-386x.2020.08.028]

14 郭卉, 姜囡, 任杰. 基于 MFCC 和 GFCC 混合特征的语音情感识别研究. 光电技术应用, 2019, 34(6): 34–39. [doi: 10.3969/j.issn.1673-1255.2019.06.008]

15 罗相林, 秦雪佩, 贾年. 基于 MFCC 及其一阶差分特征的语音情感识别研究. 现代计算机, 2019, (11): 20–24. [doi: 10.3969/j.issn.1007-1423.2019.11.004]

16 Aytar Y, Vondrick C, Torralba A. Soundnet: Learning sound representations from unlabeled video. arXiv: 1610.09001, 2016.

17 李宏菲, 李庆, 周莉. 基于多视觉描述子及音频特征的动态序列人脸表情识别. 电子学报, 2019, 47(8): 1643–1653. [doi: 10.3969/j.issn.0372-2112.2019.08.006]

18 Dhall A, Goecke R, Lucey S, *et al*. Collecting large, richly annotated facial-expression databases from movies. IEEE MultiMedia, 2012, 19(3): 34–41. [doi: 10.1109/MMUL.2012.26]

19 Dhall A, Kaur A, Goecke R, *et al*. EmotiW 2018: Audio-video, student engagement and group-level affect prediction. Proceedings of the 20th ACM International Conference on Multimodal Interaction. Boulder: ACM, 2018. 653–656. [doi: 10.1145/3242969.3264993]

20 Liu CH, Tang TH, Lv K, *et al*. Multi-feature based emotion recognition for video clips. Proceedings of the 20th ACM International Conference on Multimodal Interaction. Boulder: ACM, 2018. 630–634. [doi: 10.1145/3242969.3264989]

21 Fan YR, Lam JCK, Li VOK. Video-based emotion recognition using deeply-supervised neural networks. Proceedings of the 20th ACM International Conference on Multimodal Interaction. Boulder: ACM, 2018. 584–588. [doi: 10.1145/3242969.3264978]

22 Lu C, Zheng WM, Li CL, *et al*. Multiple spatio-temporal feature learning for video-based emotion recognition in the wild. Proceedings of the 20th ACM International Conference on Multimodal Interaction. Boulder: ACM, 2018. 646–652. [doi: 10.1145/3242969.3264992]