基于长短记忆与信息注意的视频–脑电交互协同情感识别

刘嘉敏 [1] 苏远歧 [2] 魏 平 [1] 刘跃虎 [1,3]

摘 要 基于视频–脑电信号交互协同的情感识别是人机交互重要而具有挑战性的研究问题. 本文提出了基于长短记忆神经网络 (Long-short term memory, LSTM) 和注意机制 (Attention mechanism) 的视频–脑电信号交互协同的情感识别模型. 模型的输入是实验参与人员观看情感诱导视频时采集到的人脸视频与脑电信号, 输出是实验参与人员的情感识别结果. 该模型在每一个时间点上同时提取基于卷积神经网络 (Convolution neural network, CNN) 的人脸视频特征与对应的脑电信号特征, 通过 LSTM 进行融合并预测下一个时间点上的关键情感信号帧, 直至最后一个时间点上计算出情感识别结果. 在这一过程中, 该模型通过空域频带注意机制计算脑电信号 α 波, β 波与 θ 波的重要度, 从而更加有效地利用脑电信号的空域关键信息; 通过时域注意机制, 预测下一时间点上的关键信号帧, 从而更加有效地利用情感数据的时域关键信息. 本文在 MAHNOB-HCI 和 DEAP 两个典型数据集上测试了所提出的方法和模型, 取得了良好的识别效果. 实验结果表明本文的工作为视频–脑电信号交互协同的情感识别问题提供了一种有效的解决方法.

关键词 情感识别, 长短记忆神经网络, 时–空注意机制, 多模态信号融合

引用格式 刘嘉敏, 苏远歧, 魏平, 刘跃虎. 基于长短记忆与信息注意的视频–脑电交互协同情感识别. 自动化学报, 2020, **46**(10): 2137−2147

DOI 10.16383/j.aas.c180107

Video-EEG Based Collaborative Emotion Recognition Using LSTM and Information-Attention

LIU Jia-Min[1] SU Yuan-Qi[2] WEI Ping[1] LIU Yue-Hu[1,3]

Abstract Video-EEG based collaborative emotion recognition is an important yet challenging problem in research of human-computer interaction. In this paper, we propose a novel model for video-EEG based collaborative emotion recognition by virtue of long-short term memory neural network (LSTM) and attention mechanism. The inputs of this model are the facial videos and EEG signals collected from a participant who is watching video clips for emotional inducement. The output is the participant′s emotion states. At each time step, the model employs convolution neural network (CNN) to extract features from video frames and corresponding EEG slices. Then it employs LSTM to iteratively fuse the multi-modal features and predict the next key-emotion frame until it yields the emotion state at the last time step. Within the process, the model computes the importance of different frequency-band EEG signals, i.e. α wave, β wave, and θ wave, through spatial band attention, in order to effectively use the key information of EEG signals. With the temporal attention, it predicts the next key emotion frame in order to take advantage of the temporal key information of emotional data. Experiments on MAHNOB-HCI dataset and DEAP dataset show encouraging results and demonstrate the strength of our model. The results show that the proposed method presents a different perspective for effective collaborative emotion recognition.

Key words Emotion recognition, long-short term memory neural network (LSTM), temporal-spatial attention, multi-modal fusion

Citation Liu Jia-Min, Su Yuan-Qi, Wei Ping, Liu Yue-Hu. Video-EEG based collaborative emotion recognition using LSTM and information-attention. *Acta Automatica Sinica*, 2020, **46**(10): 2137−2147

情感识别是人机交互的重要研究问题之一, 其研究目的是建立可识别人类情感并做出正确反馈的机器人系统, 使人机交互过程更加友好, 自然与智能. 本文采用二维情感表示理论[1], 将人类情感表示为 “激活度 (Arousal) −效价值 (Valence)” 二维空间中的坐标点 (图 1). 其中, 激活度用于表现人类情

收稿日期 2018-02-26 录用日期 2018-10-06
Manuscript received February 26, 2018; accepted October 6, 2018
国家自然科学基金 (91520301) 资助
Supported by National Natural Science Foundation of China (91520301)
本文责任编委 张道强
Recommended by Associate Editor ZHANG Dao-Qiang
1. 西安交通大学人工智能与机器人研究所 西安 710049 2. 西安交通大学计算机科学与技术系 西安 710049 3. 陕西省数字技术与智能系统重点实验室 西安 710049
1. Institute of Artificial Intelligence and Robotics, Xi′an Jiaotong University, Xi′an 710049 2. Department of Computer Science and Technology, Xi′an Jiaotong University, Xi′an 710049 3. Shaanxi Key Laboratory of Digital Technology and Intelligent System, Xi′an 710049

感激励程度的大小, 效价值用于表现人类对情感状态评价的好坏. 如人类高兴时, 情感的激活度与效价值的数值均较高. 二维情感表示模型可更加充分地表达和量化人类情感状态, 是多数情感识别模型使用的情感表示方法. 图 1 表示通过图片、视频等外部刺激可诱导出人类不同的情感状态, 进而通过多种传感器采集人类情感的多模态信号用于情感识别.

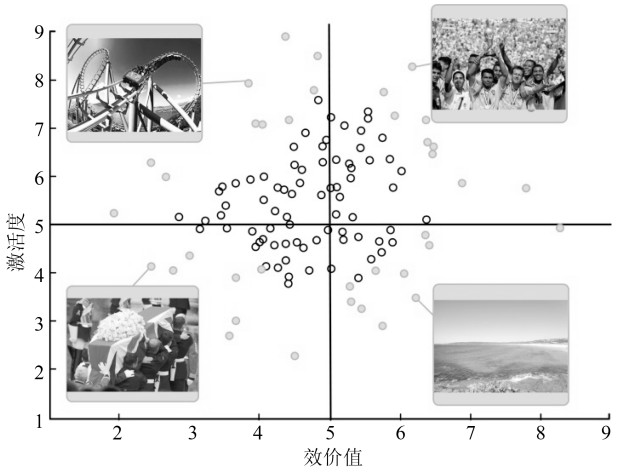


图 1 二维情感表示模型

Fig. 1 The two-dimensional expression of emotion

本文所针对的人脸视频是普通摄像头、深度相机等采集到的人类面部表情信号, 可帮助分析人类直观与外在的情感状态. 脑电信号是脑活动产生的微弱生物电于头皮处收集放大的信号, 是大脑内亿万神经元活动在大脑皮层的综合反映, 可帮助分析人类深层与内在的情感状态. 使用人脸视频和脑电信号两种模态的情感识别模型可综合人类外在与内在的情感信息, 从而给出更准确的识别结果. 通过分析实验参与人员的人脸视频与脑电信号, 识别整段情感信号中实验参与人员的情感激活度和效价值.

传统多模态情感识别方法的基本思路是手动设计提取各模态的特征, 然后进行多模态信号的融合, 最后利用标记数据集训练模式分类器[2−4]. 然而, 这类方法在处理较大规模的人类情感数据时效率较低. 近年流行的深度学习方法具有强大的特征表达能力 (例如 LSTM (Long-short term memory neural network) 在处理时序信号时可达到良好的效果). 目前多数基于脑电信号与人脸视频的情感识别方法将两个模态的信号视作时间序列, 对两个模态分别构建 LSTM 情感识别模型学习得到各个序列的识别结果, 最终将识别结果进行决策层融合[5−6]. 这些方法在识别效果上优于传统方法.

然而, 该研究内容中仍有两个关键问题亟待解决. 一是如何以交互协同的方式融合人类情感的异构多模态信号, 进而给出更加准确的情感识别结果. 二是如何从包含大量冗余信息的多模态信号中迅速定位情感关键信息, 从而提升模型的效率和准确率. 以一段 2 分钟的人脸视频为例, 它记录了一位实验参与人员在观看喜剧影像时的表情信息. 此视频中, 实验参与人员只有约 10 秒开怀大笑的表情, 其余视频帧对情感识别是冗余的.

针对以上两个问题, 本文提出了视频–脑电信号交互协同的 LSTM 情感识别模型, 同时引入了空域频带注意机制和时域注意机制. 其结构如图 2 所示, 该模型包括特征提取与交互协同两个相互耦合关联的阶段, 以 “选择性聚焦” 的方式分析人类情感的各模态时间序列, 进而给出情感识别结果. 在特征提取阶段, 首先将原始脑电信号可视化为 α 波, β 波与 θ 波的图像序列以保留脑电信号的时域–空域信息, 从而令两个模态更高效地交互协同工作; 然后提取基于卷积神经网络 (Convolution neural network, CNN)[7] 的人脸视频帧与对应的可视化脑电图像的

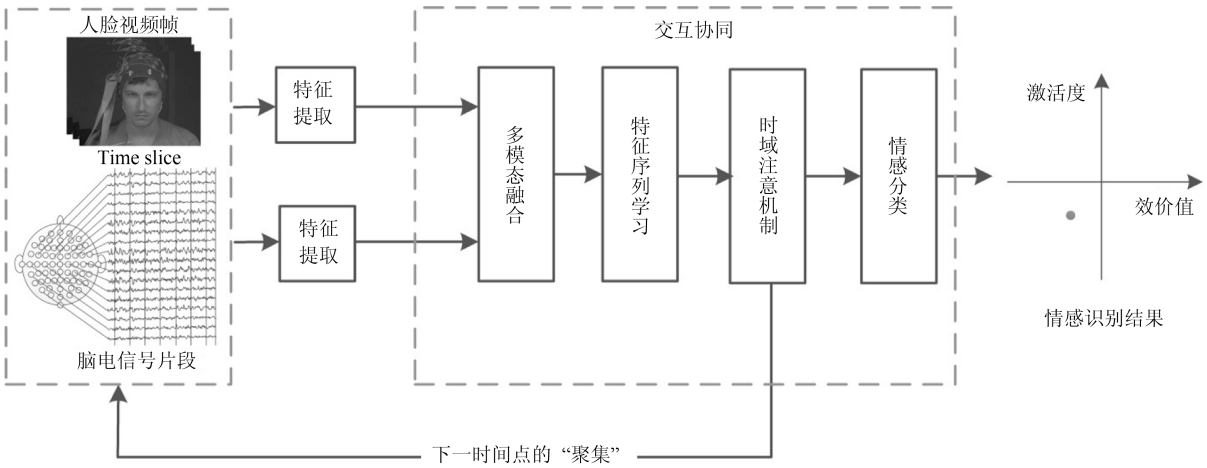


图 2 视频–脑电信号交互协同的情感识别模型

Fig. 2 The overall architecture of multi-modal emotion

特征. 在交互协同阶段, 首先使用 LSTM[8] 融合两个模态的特征并对该特征序列进行学习; 接下来预测下一时间点上 "聚焦" 的关键信号帧的时间信息, 将预测反馈至特征提取阶段; 重复上述过程直至序列结束, 最终计算出整段信号的情感识别结果. 这一过程中, 通过空域频带注意机制, 模型对脑电信号的 α 波, β 波与 θ 波可视化图像进行重要度计算, 从而有效利用脑电信号空域关键信息; 通过时域注意机制, 预测下一时间点的关键信号帧时间信息, 高效利用情感数据的时域关键信息.

基于上述思想, 本文的章节安排如下: 第 1 节综述情感识别的国内外研究现状; 第 2 节阐述了基于长短记忆与信息注意的视频–脑电信号交互协同情感识别模型; 第 3 节描述了模型的训练过程; 第 4 节给出了模型的计算实验评价结果; 第 5 节对本文工作进行总结.

1 相关工作

深度神经网络 (Deep neural network, DNN) 具有强大的挖掘数据深层、潜在表达特征的能力, 近年来被广泛应用于多模态情感识别领域[5−6, 9]. He 等[9] 首先手动设计提取人脸视频、音频信号与心电信号的特征, 然后提出了基于双向长短记忆神经网络 (Bi-directional LSTM, BLSTM) 的情感识别模型, 其中第一层三组 BLSTM 分别用于识别三个模态的信号所表达的情感状态, 第二层一组 BLSTM 用于将三个模态上的识别结果进行决策级融合, 进而给出情感状态的最终识别结果. Koelstra 等[10] 和 Huang 等[11] 对脑电信号和人脸视频分别提取情感特征、构建情感识别模型并基于决策级融合取得良好的情感识别效果. 其中, 提取有效的脑电信号情感特征是问题研究的关键, Koelstra 等和 Huang 等使用的是脑电信号空域–频域特征. 神经科学研究表明, 脑电的 δ 波 ($1\sim4$ Hz)、θ 波 ($4\sim8$ Hz)、α 波 ($8\sim13$ Hz)、β 波 ($13\sim30$ Hz) 和 γ 波 ($36\sim44$ Hz) 等 5 个频段与人类情感状态密切相关. 因此, 在提取脑电频域特征时, 多数学者将原始脑电信号映射在 5 个频段, 再从其中提取各头皮电极相关的频域特性, 进而组合为特征向量. 以上方法需直接分析包含大量冗余信息的各模态情感信号. 为了剔除冗余信息, Zhalehpour 等[12] 首先利用 Maximum dissimilarity-based 模型从人脸视频中提取关键信号帧, 然后设计情感识别模型分别分析人脸视频中的关键帧和音频信号, 最后进行决策级融合计算出情感识别的结果. 但这种方法依旧需要分析整段人脸视频, 会降低模型效率.

以上提到的方法在多模态情感识别任务中均可取得良好的识别效果. 本文在情感识别当前进展的基础上进行了两点改进: 一是将多模态信号以交互协同的方式进行融合; 二是设计可有效且迅速定位关键信息的情感识别模型.

2 视频–脑电信号交互协同的情感识别模型

本文将情感识别视作一个以 "选择性聚焦" 方式分析人类情感各模态的时间序列信号的过程. 该过程受启发于人类视觉系统的注意机制[13]. 人类观察场景时, 并非一次性理解整个场景, 而是动态地 "聚焦" 视觉空间中的多个局部获取信息, 再将获取的信息综合以理解当前的场景. 同理, 所提的情感识别模型接收到人类情感的各模态信号时, 对每一时间点的信号进行学习并预测出下一时间点将要 "聚焦" 的关键信号帧, 反复进行分析与预测, 直到获取充足的信息进而给出情感识别结果.

其框架如图 2 所示, 本文提出的脑电信号与人脸视频交互协同的 LSTM 情感识别模型主要包括特征提取与交互协同两个阶段. 在特征提取阶段, 首先选取需要 "聚焦" 的关键信号帧进行数据预处理, 然后提取出表达与泛化能力较强的特征; 在交互协同阶段, 首先将两个模态的特征融合并进行学习. 特别之处在于, 本文还会通过空域频带注意机制对脑电信号中 α 波、β 波与 θ 波的可视化图像进行重要度计算; 通过时域注意机制进行强化学习 (Reinforcement learning, RL)[14], 计算下一时间点需要 "聚焦" 的关键信号帧时间信息并反馈至特征提取阶段. 最终, 利用情感分类器输出情感识别结果. 在该模型下, 输入信号和模型行动之间构成一个闭环 — 一个有选择地反复 "聚焦" 人类情感多模态的信号, 进行情感识别的过程.

在上述基本模型的基础上, 以下章节将展开说明人脸视频–脑电信号交互协同情感识别方法的实现过程.

2.1 基于 CNN 的特征提取过程

本文输入信号为实验参与人员观看情感诱导视频时采集到的人脸视频与脑电信号. 其中, 人脸视频是普通摄像机采集到的实验参与人员的面部活动信号, 属于视觉信号. 脑电信号 (EEG) 是指按照时间顺序, 在头皮表层记录下的由大脑神经元自发性、节律性运动而产生的电位[15], 属于生理信号. 脑电信号的采集方式是让实验参与人员在观看情感诱导视频时佩戴电极脑电帽, 从而得到人类大脑皮层上 32 个不同位置的脑电信号. 两个异构的信号难以直接融合, 为此本文提出提取表达能力与泛化能力较强的特征, 同时令两模态的特征有效地交互协同. 针对人脸视频, 基于 CNN 提取面部表情特征; 与传统特征提取方法相比, CNN 具有更强大的挖掘数据深层

潜在的分布式表达特征的能力. 针对脑电信号, 本文首先将脑电信号转化为三组频带的图像序列, 这种可视化处理保留脑电信号的时域–空域特征的同时将两个模态的信号统一为图像. 然后基于 CNN 与空域频带注意机制提取脑电图像的特征.

如图 3 所示, 人脸视频的特征提取过程为: 首先, 利用 Faster-RCNN 模型[16] 检测出视频帧中人脸区域; 然后, 利用 CNN 对人脸区域提取特征; 最后, 利用全连接层处理特征输出最终特征向量 $\boldsymbol{x}_{v,n}$. 图 4 显示的是 VGG-16 三个卷积层输出的特征图.

如图 3 所示, 脑电信号的特征提取则较为复杂: 首先, 原始的脑电信号通过小波软阈值算法去除伪迹[17], 从而得到相对纯净的信号; 然后, 借鉴[18] 中数据处理方法将脑电信号划分为每段持续时长为 T 的片段 ($1/T$ 对应于人脸视频的帧率); 接下来, 在 t^{th} 段数据内提取 α 波、β 波与 θ 波三个脑电波频带的频谱能量信息并可视化至相应的电极帽 32 个电极上得到三个频带的脑电图像 (图 5), 可以看出随着人类情感激活度的上升 β 波在前额出会明显增强; 最后, 利用 CNN 对三个频带的脑电图像分别提取层特征 $\boldsymbol{e}_{\alpha,n}$, $\boldsymbol{e}_{\beta,n}$ 和 $\boldsymbol{e}_{\theta,n}$ 进行融合, 如式 (1) 与式 (2) 所示.

计算中利用空域频带注意机制计算三组特征的重要度 $\boldsymbol{e}'_n$, 最后利用全连接层 (Fully-connected layer) 处理 $\boldsymbol{e}'_n$ 输出特征向量 $\boldsymbol{x}_{e,n}$.

$$\boldsymbol{e'}_n = \boldsymbol{e}_{\alpha,n}\theta_{en,1} + \boldsymbol{e}_{\beta,n}\theta_{en,2} + \boldsymbol{e}_{\theta,n}\theta_{en,3} \tag{1}$$

式中, $\theta_{en,1}$, $\theta_{en,2}$, $\theta_{en,3}$ 分别表示分配给 $\boldsymbol{e}_{\alpha,n}$, $\boldsymbol{e}_{\beta,n}$, $\boldsymbol{e}_{\theta,n}$ 的重要度:

$$\theta_{en,i} = \frac{\exp(\boldsymbol{W}_{h,i}\boldsymbol{h}_{n-1} + b_{n,i})}{\sum\limits_{j=1}^{3}\exp(\boldsymbol{W}_{h,j}\boldsymbol{h}_{n-1} + b_{n,j})}, \quad i = 1,2,3 \tag{2}$$

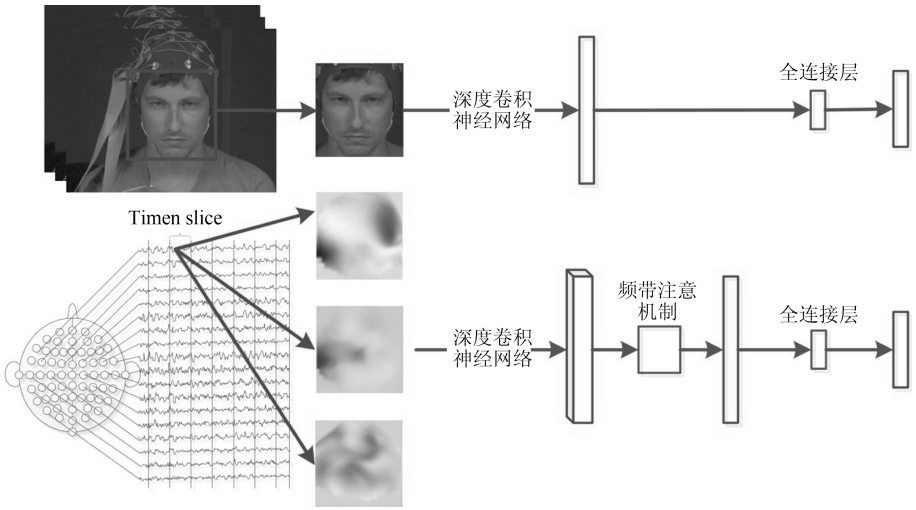


图 3 人脸视频与脑电信号的特征提取过程

Fig. 3 The process of bi-modal feature extraction

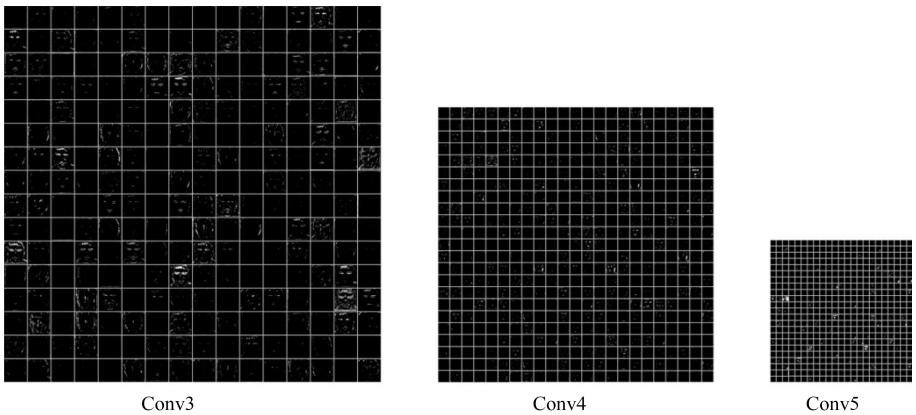


图 4 人脸视频帧 (Frame78) 的卷积层特征图

Fig. 4 The feature maps of three convolution layers on Frame78

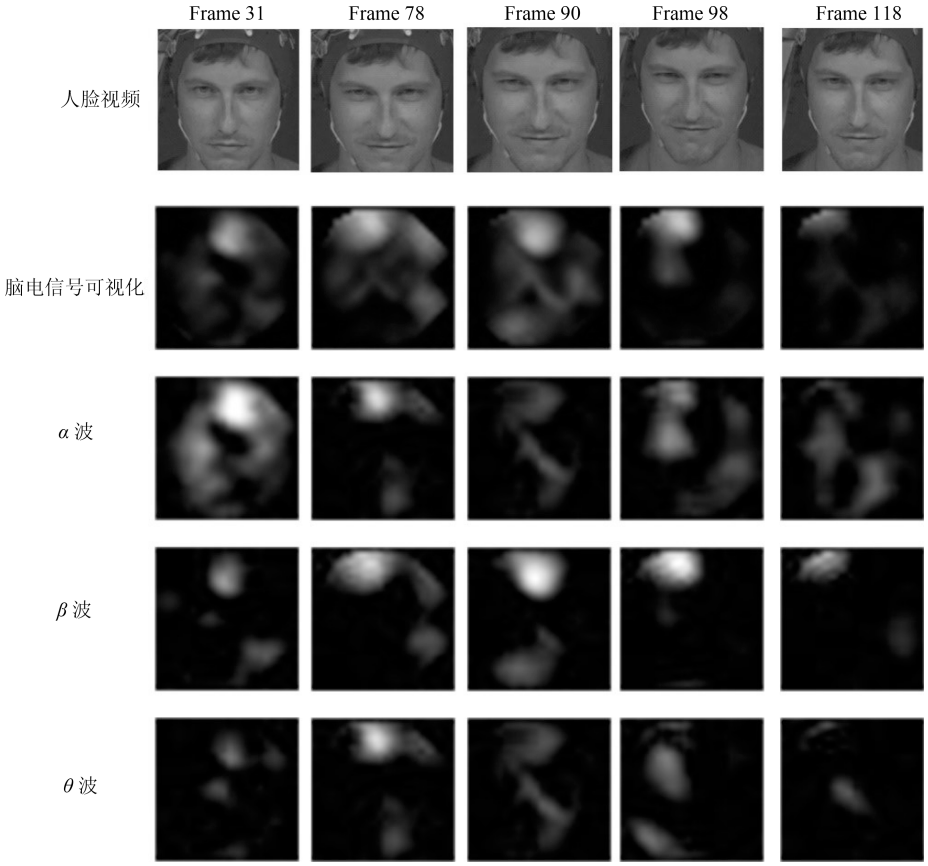


图 5　人脑电信号可视化示意图 (从上到下: 人脸视频帧; 对应的脑电信号可视化图; α 波可视化图; β 波可视化图; θ 波可视化图. 从左到右: 情感信号第 31 帧; 第 78 帧; 第 90 帧; 第 98 帧; 第 118 帧)

Fig. 5　The visualization of EEG signals (From top to down: video frames; the visualization of corresponding EEG signals; the visualization of α wave; the visualization of β wave; the visualization of θ wave. From left to right: the 31st frame; the 78th frame; the 90th frame; the 98th frame; the 118th frame in the emotion data)

式中, $\boldsymbol{W}_{h,i}$ 表示待学习的权重矩阵, $b_{n,i}$ 表示偏差. $\boldsymbol{h}_{n-1}$ 表示多层 LSTM 上一个时间点 $n-1$ 的隐状态.

2.2　基于 LSTM 与注意机制的交互协同过程

交互协同过程如图 6 所示, 本文使用一个两层 LSTM[8] (其中, 第一层包括两个共享参数的 LSTM) 对两个模态的特征序列进行融合与学习. LSTM 擅长处理时间序列, 同时可避免传统循环神经网络的长距离依赖问题. 本文还引入时域注意机制以强化学习的方式学习预测下一时间点需要 "聚焦" 的信号帧, 最后基于 Softmax 分类器[19] 完成情感识别功能.

本文以硬注意机制 (Hard attention)[20−21] 为理论基础, 提出了时间注意机制. 该机制工作流程主要分为 4 个部分: 观察 (Glimpse) 部分、核心 (Core) 部分、行为 (Action) 部分和奖励 (Reward) 部分. 给定一段长度为 T 的人脸视频和脑电信号, 将行为序列最大长度预设值为 $N_{\max}$, 则在时间点 n:

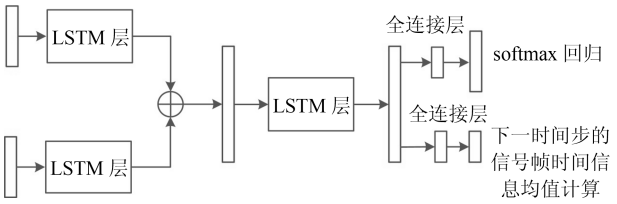


图 6　基于 LSTM 与注意机制的交互协同过程

Fig. 6　The process of emotion recognition based on LSTM and attention mechanism

1) 观察 (Glimpse) 部分: 该部分首先接收聚焦位置 f_n 以及该位置两个模态的特征向量 $\boldsymbol{x}_{v,n}$、$\boldsymbol{x}_{e,n}$, 第一层 LSTM 将两个模态的特征与上一时间点的状态处理为两组隐状态, 并拼接成一个特征向量, 从而实现多模态信号融合.

2) 核心 (Core) 部分: 该部分由第二层 LSTM 层组成. 将观察 (Glimpse) 部分输出的情感特征及上一时间点的 LSTM 隐藏层状态 $\boldsymbol{h}_{n-1}$ 作为输入, 并输出新的隐藏层状态 $\boldsymbol{h}_n$. 该部分包含对历史聚焦的情感信息的整合.

3) 行为 (Action) 部分: 该部分用来预测下一时间点上的关键信号帧时间位置 f_{n+1}, 最终在最后一个时间点使用 Softmax 分类器输出情感识别结果 $\boldsymbol{p}=(p_1,\cdots,p_C)^{\mathrm{T}}$. $p_k=p(C_k|\boldsymbol{h}_N),\ k=1,\cdots,C$ 表示情感状态属于 C_k 类的概率. 其中, 预测过程的终止条件为: 下一时间点上的关键信号帧时间位置为给定情感信号的最后一帧, 即 $f_{n+1}=T$; 或行为序列长度达到最大设定值, 即 $N=N_{\max}$.

4) 奖励 (Reward) 部分: 在每一次采样分析后, 都反馈一个奖励信息. 在时间注意机制的作用下, 情感识别模型工作过程是一个强化学习的过程. 文中模型无法一次性完整地观察到环境的, 即每次采样所得是两个模态情感信号的局部信息. 在该条件下, 模型应自主学习策略 $\boldsymbol{\pi}(\boldsymbol{a}_n|\boldsymbol{s}_{1:n};\boldsymbol{\theta})$. 其中, $\boldsymbol{\alpha}_n$ 表示在时间点 n 情感识别模型在该策略下的行为, 即计算下一时刻需要 "聚焦" 的信号帧的时间信息 f_{n+1}. $\boldsymbol{s}_{1:n}$ 表示历史状态 (包括当前时间点), 即时间注意机制部分的输入和输出时间序列. 所以, 参数为 $\boldsymbol{\theta}$ 的策略 $\boldsymbol{\pi}$ (即 $\boldsymbol{\pi}_\theta$) 就是根据当前输入和历史观察分析结果, 计算出下一时间步 "聚焦" 的关键信号帧的策略. 我们的目标是希望能找到某一策略, 从而得到最大化的奖励信息时间累积和. 奖励的累积和具有延时性, 即 $R_N=\sum_{n=1}^{N}r_n$, 其中 R_N 是指在 N 个时间点内进行一次情感识别后得到的总奖励, r_n 则是一次识别中每个聚焦分析行为得到的奖励, 在本文中与整个行为序列结束后的奖励一致.

3 情感识别模型训练

3.1 损失函数

本文使用标准反向传播 (Backpropagation through time, BPTT)[8] 训练 $\boldsymbol{p}$. 该模型目标是最小化损失函数, 该损失函数由交叉熵函数和正则项组成. 正则项是为了防止三个频带的脑电信号的重要度差距过大.

$$L=-\sum_{k=1}^{C}(y_k\log p_k+(1-y_k)\log(1-p_k))+\mu\sum_{j=1}^{3}\left(\frac{1}{3}-\frac{\sum_{n=1}^{N}\theta_{en,j}}{N}\right)^2 \tag{3}$$

式中, $\boldsymbol{y}=(y_1,\cdots,y_C)^{\mathrm{T}},\ k=1,\cdots,C$ 表示 Ground truth, 是一个 One-hot 编码向量. p_k 表示给定信号的情感状态属于第 k 类的概率. μ 为平衡系数, 本文设置为 0.02.

在空域频带注意机制的作用下, 情感识别模型会随着时间点的增长而忽略某些频带, 但是这些频带的信息对情感识别结果同样起着一定作用. 因此本文设计了如式 (3) 所示的正则项, 目的是限制模型对三个频带的脑电波特征分配尽可能均衡的重要度量.

3.2 奖励函数

由于 f_{n+1} 具有不可微的性质, 因此本文使用基于策略梯度 (Policy gradient)[14] 的强化学习进行训练. 给定序列空间 $\boldsymbol{A}$, $p_\theta(\boldsymbol{\tau})$ 表示 $\boldsymbol{A}$ 上参数为 $\boldsymbol{\theta}$ 的分布, 其中 $\boldsymbol{\tau}\in\boldsymbol{A}$ 是一组状态行为序列. 强化学习的目标函数为:

$$J(\boldsymbol{\theta})=\sum_{\boldsymbol{\tau}\in\boldsymbol{A}}p_\theta(\boldsymbol{\tau})r(\boldsymbol{\tau}) \tag{4}$$

式中, $r(\boldsymbol{\tau})$ 表示每种可能发生的序列带来的奖励; $J(\boldsymbol{\theta})$ 表示可能发生的序列分布下的期望奖励. 本文希望学习网络参数 $\boldsymbol{\theta}$, 以最大化 f_{n+1} 序列的期望奖励.

该目标函数的梯度表示为:

$$\nabla J(\boldsymbol{\theta})=\sum_{\boldsymbol{\tau}\in\boldsymbol{A}}p_\theta(\boldsymbol{\tau})\nabla\log p_\theta(\boldsymbol{\tau})r(\boldsymbol{\tau}) \tag{5}$$

一般情况下模型无关的强化学习任务 (Model-free reinforcement learning) 中, 策略梯度通过采样进行估计. 本文使用蒙特卡罗策略梯度方法[14], 该方法基本思想是持续探索, 即令模型探索环境, 根据当前策略生成一个从起始状态到终止状态的状态–动作序列.

利用蒙特卡罗法采样[22] 和近似估算, 即根据当前策略随机采样得到 M 个序列:

$$\nabla J(\boldsymbol{\theta})\approx\frac{1}{M}\sum_{m=1}^{M}\nabla\log p_\theta(\boldsymbol{\tau})r(\boldsymbol{\tau}) \tag{6}$$

假设第 m 条序列为 $\boldsymbol{\tau}^m=\{\boldsymbol{s}_1^m,\boldsymbol{a}_1^m,\cdots,\boldsymbol{s}_N^m,\boldsymbol{a}_N^m\}$, 其似然概率为:

$$p_\theta(\boldsymbol{\tau}^m)=\prod_{n=1}^{N}P(\boldsymbol{s}_{n+1}^m|\boldsymbol{s}_n^m,\boldsymbol{a}_n^m)\boldsymbol{\pi}_\theta(\boldsymbol{a}_n^m|\boldsymbol{s}_n^m) \tag{7}$$

式中, P 表示状态转移概率; $\boldsymbol{\pi}_\theta$ 表示行为策略, 本文在训练过程中使用的高斯策略. 在时间点 n, 第 m 个行为序列下, $\boldsymbol{s}_{n+1}^m$ 表示该策略的下一时间点的状态; $\boldsymbol{a}_n^m$ 表示该策略的当前行为 (即 f_{n+1}); $\boldsymbol{s}_n^m$ 表示该策略的状态.

因此, 蒙特卡洛策略梯度表达式如下:

$$\nabla J(\boldsymbol{\theta})=\frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N}\nabla\log\boldsymbol{\pi}_\theta(\boldsymbol{a}_n^m|\boldsymbol{s}_n^m)R^m \tag{8}$$

式中, R^m 表示第 m 个序列下获得的奖励.

本文仅考虑整组行为序列完成后的奖励, 如式 (9) 所示:

$$R' = \begin{cases} \lambda_{tp}, & \text{若正检} \\ \lambda_{fp}, & \text{若误检} \end{cases} \tag{9}$$

式中, λ_{tp} (>0), λ_{fp} (<0) 分别表示每一时间点的正检和误检的奖励数值, 本文方法会重点惩罚误检项.

同时考虑到有效情感信息的稀疏性, 还加入如式 (10) 所示的稀疏性约束项 $\lambda_{\text{sparse}}N<0$, 从而令模型观察尽可能少的信号同时获得尽可能高的准确率.

$$R = \lambda_r R' + \lambda_{\text{sparse}} N \tag{10}$$

式中, λ_r 表示奖励因子, 数值大于零; λ_{sparse} 表示稀疏性因子, 数值小于零; N 表示行为序列的长度.

策略迭代的基本思路为:

$$\boldsymbol{\theta} = \boldsymbol{\theta} + \varepsilon \nabla J(\boldsymbol{\theta}) \tag{11}$$

式中, ε 表示步长因子, 即算法的学习率.

4　实验结果与分析

4.1　实验数据与评价指标

为了验证本文方法的有效性, 本节在 MAHNOB-HCI 数据集[15] 与 DEAP 数据集[23] 上进行实验, 主要针对情感的激活度和效价值进行识别, 并采用识别准确率 (Classification rate) 和 F1-*score* 作为识别效果的评价指标.

MAHNOB-HCI 数据集是一个多模态情感识别及隐性标注 (Implicit tagging) 数据集, 包括采集自 27 位实验参与人员观看 20 段视频时的 527 组原始人脸视频、音频和脑电信号. 在看完每段视频后, 实验参与人员使用 (Self-assessment manikin, SAM)[1] 标定情感的激活度, 效价值, 分为 9 个级别 (分别为 1～9). 同时使用离散情感标签标定情感, 该数据集根据标签将实验人员情感的激活度和效价值各分为三类.

DEAP 是一个多模态情感识别数据集, 包括采集自 32 位实验参与人员观看 40 段音乐视频时的人脸视频、外部生理信号和脑电信号. 其中, 10 位实验参与人员的数据中不包括人脸视频. 在看完每段视频后, 实验参与人员使用 (SAM)[1] 标定情感的激活度, 效价值 (数值为 1～9). 如表 1 所示, 该数据集根据数值大小将情感的激活度和效价值分别分为 3 个级别.

本文使用识别准确率和 F1-*score* 两个指标对模型识别结果进行评价. 识别准确率 (Classification rate, CR) 表示测试集中正确分类的样本数与测试集样本总量的百分比 (式 (12)). F1-*score* 是统计学中用来衡量多分类模型精确度的一种指标, 可看作是模型精确率 (Precision) 和召回率 (Recall) 的一种加权平均, 可兼顾模型的精确率和召回率 (式 (13)).

$$CR = \frac{N_{TP}}{N_{\text{data}}} \tag{12}$$

$$CR = \frac{2N_{TP}}{2N_{TP} + N_{FP} + N_{FN}} \tag{13}$$

式中, N_{data} 表示测试集中情感数据的样本总量; N_{TP}、N_{FP} 与 N_{FN} 表示所有测试样本中的正检总量、误检总量与漏检总量.

表 1　激活度和效价值的三分类

Table 1　Valence and arousal class with range

	激活度	效价值
Low	1～4.5	1～4.5
Medium	4.5～5.5	4.5～5.5
High	5.5～9	5.5～9

4.2　实现细节

本文使用 MAHNOB-HCI 数据集进行模型训练, 使用 MAHNOB-HCI 的测试集与 DEAP 数据集进行模型测试. 将 MAHNOB-HCI 数据集中 27 位实验参与人员的数据以 5:1:1 的比例分为训练集 A, 验证集 A′ 和测试集 B. 在数据预处理过程中, 将数据集的人脸视频降采样为 8 fps. 同时检测并裁减出视频中人脸图像, 重缩放图像尺寸为 227 × 227. 在训练过程中, 本文使用 Adam 方法[24] 来更新参数. 每次更新使用的样本集是通过经验回放机制从训练集 A 中抽取 mini-batch = 12 个样本得到. 为了防止模型过拟合, 本文将 dropout 的数值设置为 0.5. 将最大时间步 $N_{\max}$ 的数值设置 30. 另外, 本文中使用的所有经过 fine-tune 的 VGG-16 网络[25] 被固定参数, 仅用来提取特征.

为了有效地训练模型, 本文将两层 LSTM 分为 3 个模块分别进行训练: 在训练第一层用来处理人脸视频特征的 LSTM 时, 先去掉该层用来处理脑电信号特征的 LSTM, 并将第二层 LSTM 的神经元数目设置为 1 024 个. 同理地, 在训练第一层用来处理脑电信号特征的 LSTM 时, 先去掉该层用来处理人脸视频特征的 LSTM, 并将第二层 LSTM 的神经元数目设置为 1 024 个. 在训练第二层 LSTM 时, 将已预训练的第一层 LSTM 进行参数固定.

4.3　多种情感识别方法结果对比

本文首先将提出的模型与其他经典模型的识别效果进行对比 (表 2), 本文提出的模型对激活度与

效价值的识别率和 F1-*score* 相对其他方法均有明显提升. 在 MAHNOB-HCI 数据集上与当前识别效果最好的方法相比, 情感激活度识别准确率和 F1-*score* 分别提升了0.6 %和 0.014, 情感效价值识别准确率和 F1-*score* 分别提升了 1.5 % 和 0.012. 识别效果提升的原因在于其余方法均需直接分析包含大量冗余信息的多模态情感信号, 而本文提出的模型则引入了信息注意机制, 从而压缩了冗余信息并提升了准确率. 同时从表 2 可以看出, 对情感效价值的识别效果优于对情感激活度的识别效果. 这是因为情感激活度用于表现情感激励程度的大小, 情感效价值用于表现人类对情感状态评价的好坏, 相比之下效价值更容易被直观地分析和理解. 特别地, 在 MAHNOB-HCI 测试集 B 上基线方法对整段情感信号的脑电情感特征序列和眨眼特征序列进行融合并基于 SVM 分类器进行情感识别, 在效价值识别上取得了较好的效果, 这是因为眨眼特征对效价值识别做出了一定贡献, 但对激活度的识别却无明显效果. 总体来说, 本文提出的模型具有更好的效果, 且仅需分析 10 % 的信号.

本文还将模型对 MAHNOB-HCI 数据集中三组数据样本的识别效果可视化至图 7, 可以看出对模型可准确识别出整段情感信号中实验参与人员的情感激活度和效价值. 第一组数据样本和第三组数据样本分别为高激活度低效价值 (情绪紧张) 样本和低激活度低效价值 (情绪悲伤) 样本, 本文模型对两组数据的识别效果均较为准确; 第二组数据样本为中激活度高效价值 (情绪高兴) 样本, 本文模型对该数据样本的效价值识别效果准确而激活度识别有偏差, 原因是模型对情感 “高兴” 程度的认知可能与实际有偏差.

4.4 注意机制可视化

本文从 MAHNOB-HCI 测试集 B 中选定样本进行单组情感识别测试, 并将每一时间步上的情感关键信息可视化为图 8 与图 9. 其中, 在可视化时域

表 2 不同方法在 MAHNOB-HCI 数据集与 DEAP 数据集上的识别效果

Table 2 The recognition result of different methods on MAHNOB-HCI dataset and DEAP dataset

	激活度		效价值	
	CR (%)	F1-*score*	CR (%)	F1-*score*
Baseline[15](MAHNOB-HCI)	67.7	0.620	**76.1**	**0.740**
Koelstra et al.[10] (MAHNOB-HCI)	72.5	0.709	73.0	0.718
Huang et al.[11] (MAHNOB-HCI)	63.2		66.3	
VGG-16+ 本文模型 (MAHNOB-HCI)	**73.1**	**0.723**	74.5	0.730
VGG-16+ 本文模型 (DEAP)	**85.8**		**84.3**	

图 7 本文模型在 MANNOB-HCI 数据集上的可视化识别结果 (从上到下分别为三组情感数据中的人脸视频. 从左到右分别为情感数据; Groundtruth 与本文模型的识别结果)

Fig. 7 The visualization of results of the proposed model on MAHNOB-HCI dataset (From up to down: three groups of emotion data. From left to right: emotion data;the groundtruth and results of the proposed model)

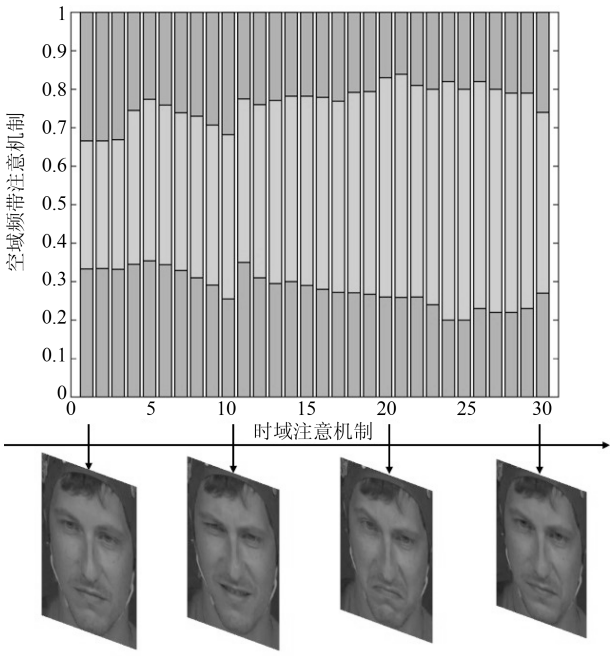


图 8　高激活度数据样本注意机制可视化结果

Fig. 8　The presentation of the band attention weights on EEG signals and the temporal attention policy for a "nervous" man with high arousal

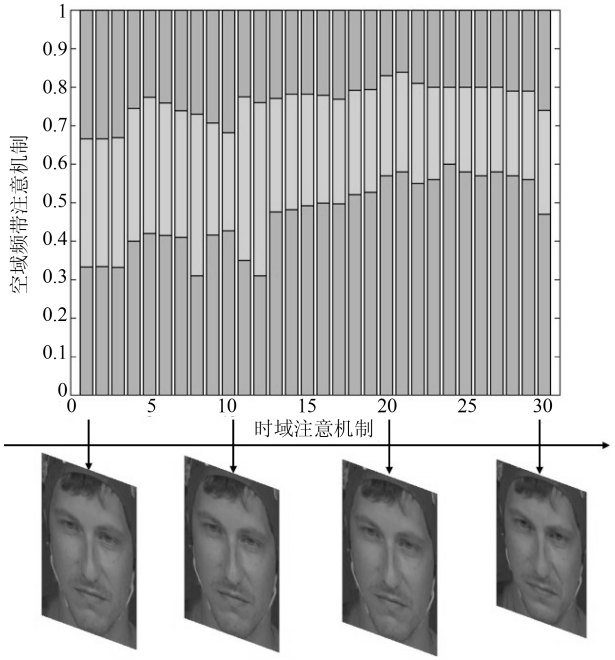


图 9　低激活度数据样本注意机制可视化结果

Fig. 9　The presentation of the band attention weights on EEG signals and the temporal attention policy for a "nervous" man with low arousal

注意机制时仅选取 4 个时间步显示. 图中上侧条形从下到上分别代表每个时间上脑电信号 α, β, θ 三个频带的重要度. 从图示可以直观看出, 本文提出的模型可迅速且精确地定位有效信息, 并在有效信息的区域更加频繁地分析数据以得到更准确的识别结果. 此图中有一个有趣的现象, 当该名实验参与人员越来越紧张 (情感激活度越来越高) 时, 脑电 β 波会逐渐占据主导地位. 这与生理学中脑电 β 波会在人类处于紧张、焦虑、恐慌等情感状态时占据主导地位的结论一致. 随着脑电 β 波变强, 人类身体会越来越处于紧张的状态, 这种情况下人类身心能量快速消耗, 容易感受到压迫与疲倦. 而图 9 中, 当人类的情感状态保持稳定的低激活度状态时, α 波占据主导地位. 生理学研究表明, 当人类脑电波主要频率处于 α 波时, 人类处于意识清醒且身心放松的状态, 也是人类思考的最佳状态.

4.5　模型增量化研究

为了量化本文提出的情感识别模型中不同部分的效果, 本文在 MAHNOB-HCI 测试集 B 与 DEAP 数据集上设置了一组模型增量实验. 经过测试得到如表 3 与表 4 的测试结果, 其中, w/o band and temp 表示在本文提出的模型基础上, 去掉空域频带注意和时域注意机制; w/o band 表示在本文提出的模型基础上, 去掉空域频带注意机制; w/o temporal 表示在本文提出的模型基础上, 去掉时域注意机制; vis-EEG-LSTM 表示本文提出的模型. 表 3 与表 4 显示, 空域频带注意和时域注意机制的引入能提升激活度和效价值的识别率和 F1-*score*. 这是因为空域频带注意机制的引入有效利用了脑电信号的空域–时域–频域信息, 并且结合了脑电信号在不同情感状态下具有不同主导作用的生理学现象; 时域注意机制的引入则解决了情感识别中存在的噪声干扰、计算冗余等问题. 此外, 通过对比可看出时域注意机制的引入对识别效果的提升具有更大的作用, 这是因为包含大量冗余情感信息的人脸视频和脑电信号会大大降低情感识别的准确率, 而本文使用的时域注意机制可有效地辅助模型 "聚焦" 信号中的关键情感信息, 从而进行更准确可靠的情感识别.

表 3　本文提出的情感识别模型的识别准确率和 F1-*score* (MAHNOB-HCI 数据集)

Table 3　The classification rate and F1-*score* of ablation studies on MAHNOB-HCI dataset

	激活度		效价值	
	CR (%)	F1-*score*	CR (%)	F1-*score*
w/o band and temp	66.4	0.650	68.9	0.678
w/o band	70.9	0.690	73.0	0.711
w/o temporal	69.7	0.680	70.4	0.695
vis-EEG-LSTM	**73.1**	**0.723**	**74.5**	**0.730**

表 4 本文提出的情感识别模型的识别准确率和 F1-*score* (DEAP 数据集)

Table 4 The classification rate and F1-*score* of ablation studies on DEAP dataset

	激活度		效价值	
	CR (%)	F1-*score*	CR (%)	F1-*score*
w/o band and temp	79.1	0.774	78.5	0.770
w/o band	83.1	0.816	82.5	0.809
w/o temporal	78.1	0.754	81.4	0.805
vis-EEG-LSTM	**85.8**	**0.837**	**84.3**	**0.831**

4.6 单模态与双模态情感识别对比

为了对比人脸视频和脑电信号在情感识别任务中发挥的作用, 本文在 MAHNOB-HCI 测试集 B 与 DEAP 数据集上使用本文提出的模型, 针对人脸视频和脑电信号两个模态分别进行情感识别实验. 测试结果如表 5 与表 6 所示, 在 MAHNOB-HCI 数据集上使用人脸视频的识别效果要好于使用脑电信号的识别效果. 其原因是 MAHNOB-HCI 数据集的人脸视频中实验参与人员面部表情变化明显, 更容易提取有效的表情信息. 而脑电信号的变化则比较复杂, 相比面部表情较难区分. 而在 DEAP 数据集上使用脑电信号的识别效果要好于使用人脸视频的识别效果. 其原因是该数据集采集到的人脸视频中人类面部表情变化非常细微, 较难分析. 同时实验结果均显示令两个模态交互协同可提升情感识别效果. 这是因为在情感表达过程中人脸表情与脑电信号尽管是相互分离的两个模态, 但是本质上具有相关性.

表 5 两种单模态情感识别与多模态情感识别的识别准确率和 F1-*score* (MAHNOB-HCI 数据集)

Table 5 The classification rate and F1-*score* of uni-modal and bi-modal emotion recognition on MAHNOB-HCI dataset

	激活度		效价值	
	CR (%)	F1-*score*	CR (%)	F1-*score*
人脸视频	70.8	0.691	72.9	0.711
脑电信号	69.9	0.673	73.3	0.720
人脸视频 + 脑电信号	**73.1**	**0.723**	**74.5**	**0.730**

表 6 两种单模态情感识别与多模态情感识别的识别准确率和 F1-*score* (DEAP 数据集)

Table 6 The classification rate and F1-*score* of uni-modal and bi-modal emotion recognition on DEAP dataset

	激活度		效价值	
	CR (%)	F1-*score*	CR (%)	F1-*score*
人脸视频	67.1	0.653	66.3	0.650
脑电信号	84.7	0.815	83.4	0.819
人脸视频 + 脑电信号	**85.8**	**0.837**	**84.3**	**0.831**

合理利用多模态的信号进行情感识别可综合各个模态的优势, 从而令识别结果更加准确可靠.

综上所述, 本文提出的基于长短记忆与信息注意的视频–脑电信号交互协同情感识别方法可综合人类内在与外在的情感信息, 更准确地给出识别结果.

5 结论

本文提出了一种基于长短记忆与信息注意的视频–脑电信号交互协同情感识别方法. 该方法具有两个模态信号综合作用、相互补充的优势, 可准确识别人类的情感状态. 为了更有效地利用脑电信号的空域关键信息, 所提出方法将脑电信号转换为图像序列, 并利用空域频带注意机制对 α, β, θ 三个频带的脑电信号进行重要度计算. 为了有效利用情感数据的时域关键信息, 引入时域注意机制自动定位情感数据中的关键信号帧. 在两个数据集的实验结果表明, 所提出的情感识别模型能够实现更准确的识别效果. 然而, 自然场景下得人类情感状态不同于特定数据集, 会随时间发生变化. 在保证情感识别效果的前提下, 如何识别一段情感信号中的不同情感状态仍然是未来需要研究的重点问题.

References

1 Bynion T M, Feldner M T. *Self-Assessment Manikin*. Berlin: Springer International Publishing, 2017. 1−3

2 Lin J C, Wu C H, Wei W L. Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition. *IEEE Transactions on Multimedia*, 2012, **14**(1): 142−156

3 Jiang D, Cui Y, Zhang X, Fan P, Ganzale I, Sahli H. Audio visual emotion recognition based on triple-stream dynamic bayesian network models. In: Proceedings of the 2011 International Conference on Affective Computing and Intelligent Interaction. Berlin, GER: Springer-Verlag, 2011. 609−618

4 Xie Z, Guan L. Multimodal information fusion of audio emotion recognition based on kernel entropy component analysis. *International Journal of Semantic Computing*, 2013, **7**(1): 25−42

5 Khorrami P, Le Paine T, Brady K. How deep neural networks can improve emotion recognition on video data. In: Proceedings of the 2016 IEEE International Conference on Image Processing. New York, USA: IEEE, 2016. 619−623

6 Liu J, Su, Y, Liu, Y. Multi-modal emotion recognition with temporal-band attention based on lstm-rnn. In: Proceedings of the 2017 Pacific Rim Conference on Multimedia. Berlin, GER: Springer, 2017. 194−204

7 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of the 2012 Annual Conference on Neural Information Processing Systems. Massachusetts, USA: MIT Press, 2012. 1097−1105

8 Sak H, Senior A, Beaufays F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. arXiv preprint arXiv: 1402.1128, 2014.

9 He L, Jiang D, Yang L, Pei E, Wu P, Sahli H. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In: Proceedings of the 2015 International Workshop on Audio/visual Emotion Challenge. New York, USA: ACM, 2015. 73−80

10 Koelstra S, Patras I. Fusion of facial expressions and EEG for implicit affective tagging. *Image and Vision Computing*, 2013, **31**(2): 164−174

11 Huang X, Kortelainen J, Zhao G, Li X, Moilanen A, Seppanen T, Pietikainen M. Multi-modal emotion analysis from facial expressions and electroencephalogram. *Computer Vision and Image Understanding*, 2016, **147**: 114−124

12 Zhalehpour S, Akhtar Z, Erdem C E. Multimodal emotion recognition with automatic peak frame selection. In: Proceedings of the 2014 IEEE International Symposium on Innovations in Intelligent Systems and Applications. New York, USA: IEEE, 2014. 116−121

13 Xu K, Ba J L, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R S, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the 2015 International Conference on Machine Learning. New York, USA: ACM, 2015. 2048−2057

14 Liu Chang, Liu Qin-Rang. Using reinforce learning to train multi attention model. *Acta Automatica Sinica*, 2017, **43**(9): 1563−1570
(刘畅, 刘勤让. 使用增强学习训练多焦点聚焦模型. 自动化学报, 2017, **43**(9): 1563−1570)

15 Soleymani M, Lichtenauer J, Pun T, Pantic M. A multimodal affective database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 2012, **3**(1): 42−55

16 Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of the 2015 Advances in Neural Information Processing Systems. Massachusetts, USA: MIT Press, 2015. 91−99

17 Mowla M R, Ng S C, Zilany M S A, Paramesran R. Artifacts-matched blind source separation and wavelet transform for multichannel EEG denoising. *Biomedical Signal Processing and Control*, 2015, **22**(3): 111−118

18 Bashivan P, Rish I, Yeasin M, Codella N. Learning representations from EEG with deep recurrent-convolutional neural networks. In: Proceedings of the 2016 International Conference on Learning Representation. San Juan, Puerto Rico: ICLR, 2016.

19 Anzai Y. *Pattern Recognition and Machine Learning*. Elsevier, 2012.

20 Lei T, Barzilay R, Jaakkola T. Rationalizing neural predictions. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. British Columbia, Canada: ACL, 2016. 107−117

21 Yu A W, Lee H, Le Q V. Learning to skim text. arXiv preprint arXiv: 1704.06877, 2017.

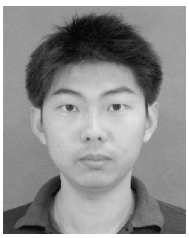
22 Rubinstein R Y, Kroese D P. *Simulation and the Monte Carlo Method*. John Wiley & Sons, 2008. 167−168

23 Koelstra S, Muhl C, Soleymani M, Lee S, Yazdani A, Ebrahimi T, Pun T, Nijholt A, Patras I. Deap: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 2012, **3**(1): 18−31

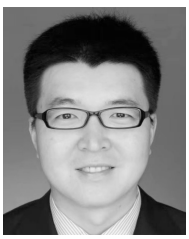
24 Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv: 1412.6980, 2014.

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

刘嘉敏 西安交通大学硕士研究生. 主要研究方向为人机交互, 多模态情感识别, 增强学习.
E-mail: ljm.168@stu.xjtu.edu.cn
(**LIU Jia-Min** Master student at Xi′an Jiaotong University. Her research interest covers human-computer interaction, multi-modal emotion recognition, and reinforcement learning.)



苏远歧 西安交通大学讲师. 主要研究方向为图像处理, 计算机视觉, 计算机图形学. 本文通信作者.
E-mail: yuanqisu@mail.xjtu.edu.cn
(**SU Yuan-Qi** Lecturer at Xi′an Jiaotong University. His research interest covers image processing, computer vision, and computer graphics. Corresponding author of this paper.)



魏 平 西安交通大学副教授. 主要研究方向为计算机视觉, 机器学习, 认知计算.
E-mail: pingwei@xjtu.edu.cn
(**WEI Ping** Associate professor at Xi′an Jiaotong University. His research interest covers computer vision, machine learning, and computational cognition.)

刘跃虎 西安交通大学教授. 主要研究方向为计算机视觉, 人机交互, 增强现实与仿真测试.
E-mail: liuyh@mail.xjtu.edu.cn
(**LIU Yue-Hu** Professor at Xi′an Jiaotong University. His research interest covers computer vision, human-computer interaction, augmented reality and simulation testing.)