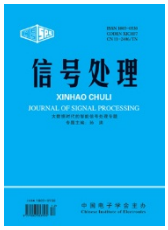

信号处理
Journal of Signal Processing
ISSN 1003-0530,CN 11-2406/TN

《信号处理》网络首发论文

题目：基于多头注意力机制的模型层融合维度情感识别方法
作者：董永峰，苏海洋，刘斌，陶建华
收稿日期：2021-01-15
网络首发日期：2021-03-30
引用格式：董永峰，苏海洋，刘斌，陶建华．基于多头注意力机制的模型层融合维度情感识别方法．信号处理.
https://kns.cnki.net/kcms/detail/11.2406.tn.20210324.0941.006.html

基于多头注意力机制的模型层融合维度情感识别方法

董永峰[1] 苏海洋[1,2] 刘 斌[2] 陶建华[1,2]
(1. 河北工业大学人工智能与数据科学学院，天津 300401；2. 中国科学院自动化研究所模式识别实验室，北京 100190)

摘 要：近年来，情感识别成为了人机交互领域的研究热点问题，而多模态维度情感识别能够检测出细微情感变化，得到了越来越多的关注多模态维度情感识别中需要考虑如何进行不同模态情感信息的有效融合。针对特征层融合存在有效特征提取和模态同步的问题、决策层融合存在不同模态特征信息的关联问题，本文采用模型层融合策略，提出了基于多头注意力机制的多模态维度情感识别方法，分别构建音频模型、视频模型和多模态融合模型对信息流进行深层特征学习，最后放入双向长短时网络中得到最终情感预测值。所提方法相比于不同基线方法在激活度和愉悦度上均取得了最佳的性能，可以在高层维度对情感信息有效捕捉，进而更好的对音视频信息进行有效融合。

关键词：维度情感识别；多模态情感融合；模型层融合；多头注意力机制
中图分类号：TP391.4 **文献标识码：**A

Model Level Fusion Dimension Emotion RecognitionMethodBased on Transformer

Dong Yongfeng[1] Su Haiyang[1,2] Liu Bin[2] Tao Jianhua[1,2]
(1. School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China; 2. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In recent years, emotion recognition had become a hot research topic in the field of human-computer interaction, and multi-modal dimensional emotion recognition could detect subtle emotional changes, which had attracted more and more attention. In multi-modal emotion recognition, it was necessary to consider how to effectively integrate different modal emotion information. Aiming at the problem of effective feature extraction and modal synchronization in feature level fusion, and the correlation problem of different modal feature information in decision level fusion, this paper adopted a model level fusion strategy and proposes a multi-modal dimension emotion recognition method based on Transformer. Respectively constructed audio model, video model and multi-modal fusion model to learn the deep features of the information flow, and finally put it into Bi-directional Long Short Term Memory to obtain the final emotional prediction value. Compared with different baseline methods, the proposed method achieves the best performance in terms of arousal and valence, and could effectively capture emotional information in high-level dimensions, and thus better effectively integrate audio and video information.

Key words: dimension emotion recognition; multimodal emotion fusion; model level fusion; Transformer

1 引言

情感识别有助于快速传达信息并理解别人的真实意图，是人机交互的关键[1]。在情感表示模型方面，

收稿日期：2021-01-15；修回日期：2021-03-08
基金项目：国家重点研发计划（2017YFB1002804）；国家自然科学基金重点项目（61831022, 61771472, 61901473，61902106）；天津市自然科学基金（19JCZDJC40000）；河北省自然科学基金（F2020202028）

主要分为离散表示模型[2]和维度表示模型[3]，离散表示模型将人类的情感划分为几种常见的情感，来反映人的基本情绪，而维度情感表示模型使用维度空间中的连续数值来描述情感状态，每个情感状态对应二维空间中的一个点[4]，坐标系横轴 arousal 代表激活度，表示情感的激昂与低迷程度，值越大表示情感越激昂，值越小表示情感越低迷；坐标轴 valence 代表愉悦度，表示情感的积极与消极程度，值越大表示情感积极程度越高，值越小表示情感消极程度越高。维度情感识别模型能够更为有效的反映交互对象的心理细微波动，对于增强交互的自然度有着重要作用，同时维度情感表示识别模型在提高情感识别的准确性和鲁棒性中也起着重要的作用。因此本文以维度情感表示模型为研究基础。

目前基于单模态的情感识别已经取得了一定的进展，Wang 等人[5]利用双向递归神经网络（Bi-RNN）对视频特征进行情感学习，但是情感是由多种模态综合表现出来的，各个模态之间也具有一定的关联，同时不同模态对于情感结果的贡献程度也不尽相同[6]。通常来说，多模态情感识别的性能要优于单模态情感识别性能，而目前主要的多模态情感融合方法是特征层融合和决策层融合。

特征层融合方法需要分别从多种模态信息中提取特征，构建用于识别情感的联合特征，对各模态有较高的同步要求。Chaparro 等人[7]提出基于脑电图等生理信号的多模态情感识别模型，在特征层串联融合面部表情特征和心电信号特征构成多模态特征，实验表明，多模态特征的识别率高于一种模态特征的识别率。Xu 等人[8]利用注意力机制对语音和音频文本在特征层进行融合，在交互式情绪二元运动捕捉（IEMOCAP）数据集上取得了最好的性能。在国际音视频情感识别竞赛（audio/visual emotion challenge, AVEC2017）中，Singh 等人[9]利用传统的视频纹理特征和 openXBOW 提取的音频词袋特征集（bag-of-audio-words, BoAW）进行特征层融合的情感识别。Basnet 等人[10]通过基于交互信息选择的音视频特征层融合构建情感预测模型。Aven 等人[11]对音频、视频和文本特征进行特征层融合对抑郁症相关的情感状态构建情感识别模型。特征层只是对各个模态的情感特征进行简单拼接，并没有考虑到模态之间的信息交互。决策层融合方法考虑不同模态信息对于情感识别贡献度不同，大多数多模态融合情感识别方法采用决策层融合。Poria 等人[12]利用等权重原理，在决策层加权融合音频、视频和文本的分类结果，此时等价于无加权融合。Sebastian 等人[13]在特征层上对语音和音频文本进行前期特征层融合，然后输入到网络中再与经过长短时记忆网络（Long Short Term Memory, LSTM）的文本特征结果进行决策层的后期融合。Huang 等人[14]基于 LSTM 的决策层情感识别在 AVEC2017 中取得了不错的成绩，之后又提出了端到端情感识别模型[15]。Chen 等人[16]基于 LSTM-RNN 模型提出多任务学习的多模态情感识别方法。决策层融合虽然解决了不同模态之间的时序不同步问题，但是没有考虑到不同模态的情感特征信息的关联。

本文针对以上问题，提出基于多头注意力机制的模型层融合维度情感识别方法，模型层融合既解决了不同模态时序不同步的问题，同时考虑了不同模态的情感特征信息之间的关联性。在模型层融合部分，本文利用多头注意力机制构建模型层融合模块，分别将音视频信息放入模块中进行高层维度的时序动态情感特征学习，再将其放入融合模块中进行模型层的时序动态情感特征学习，最后用双向长短时记忆网络（Bi-directionalLong Short Term Memory, BLSTM）和线性变换，得到最终情感预测值。因为维度情感数据库较小，为了解决这个的问题，本文对原始数据库进行了数据增广，然后提取音频和视频的情感特征信息。文中比较了使用相同数据库的研究人员识别方法，模型层融合方法在激活度和愉悦度上均取得了最佳的性能。实验结果表明，基于模型层的音视频维度情感识别中，可以在高层维度对情感信息有效捕捉，进而更好的对音视频信息进行有效融合。

本文在第一部分中对情感识别相关研究现状和研究内容进行了介绍，在第二部分中介绍了本文提出的基于多头注意力机制的模型层融合维度情感识别方法，在第三部分中介绍了实验结果和分析，最后在第四部分中对实验做了总结和展望。

2　基于多头注意力机制的模型层融合维度情感识别

本文所提的基于多头注意力机制的模型层融合维度情感识别方法整体框架如图 1 所示。基于音视频数据进行维度情感识别建模，主要包括音频模块、视频模块和音视频融合模块，音频和视频模块通过自注意

力机制学习各自单模态的情感时序信息，而音视频融合模块在模型层实现音视频信息的交互。整个模型是将音频模块和视频模块的高层输出转换到相同的情感语义特征空间中，然后通过音视频融合模块生成有效的情感表征，最后通过一个线性变换层输出情感预测值。具体来说，就是对原始数据进行数据增广和标签延迟优化之后，选取在长短时序列上表征较好的音视频情感特征，本文选取音频特征包括 BoAW 和 2010 届国际语音会议特征集（Interspeech 2010, IS2010）和视频特征包括视频词袋特征集（bag-of-video-words, BoVW）和人脸超分辨率序列特征集（Visual Geometry Group Face, VGGFace）作为情感特征[14]。特征提取之后把音视频情感特征进行线性变换，放入到多头注意力模型中，经过多头注意力模型的数据信息与原始情感特征进行求和与归一化，得到的数据信息再分别进行一维卷积学习情感时序信息，最后将其与上一步的数据信息一起进行求和与归一化，这样音视频的情感特征处理完毕。再将其放入到多模态融合模块，利用多头注意力机制根据音频和视频数据帧之间的相关性确定时序对齐分数，最后输出的数据信息经过 BLSTM 模型之后，线性变换得到最终的情感预测值。

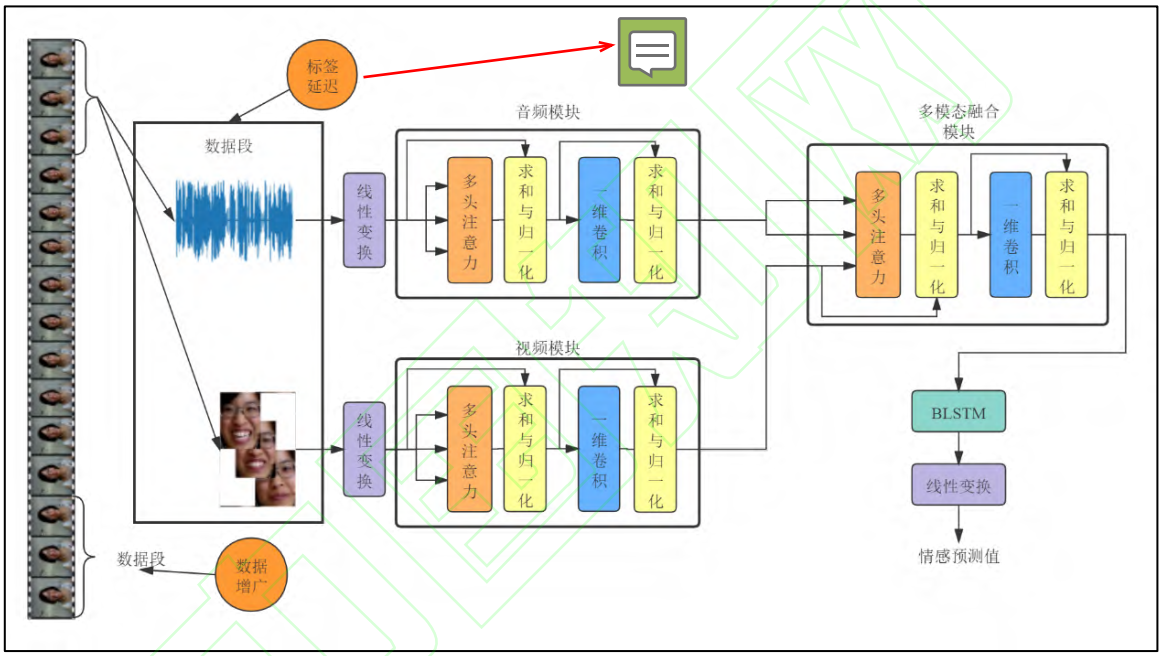

图 1　基于多头注意力机制的模型层融合维度情感识别系统框架

Fig.1　A framework of model level fusion dimension emotion recognition based on Transformer

2.1　双向长短时记忆网络

情感状态随着时间动态变化，尽管传统的 LSTM[17]可以有效的学习时间序列中的动态情感信息，但是该模型只能基于先前的信息和当前的信息来预测情感，而不能使用后续的信息。为了更充分的利用上下文信息，本文使用 BLSTM 模型对不同形式的信息进行特征学习，以更好地预测维度情感信息。在 BLSTM 的某个时刻，可以对当前时间节点的前后信息进行学习。本文中的系统框架最后使用 BLSTM 网络来训练维度情感回归模型。下面给出了时间步长 t 时 BLSTM 的公式：

$$h_t^f = sigmoid\left(u^f x_t + w_f h_{t-1} + b_f\right) \tag{1}$$

$$h_t^b = sigmoid\left(u^b x_t + w_b h_{t-1} + b_b\right) \tag{2}$$

$$y_t = sigmoid\left(v_f h_t^f + v_b h_t^b + b_f\right) \tag{3}$$

其中，u^f，u^b，w_f，w_b，v_f，v_b分别对应于前向传播和后向传播的权重向量，h_t^f，h_t^b分别表示在时

间t时刻隐藏层的前向传播和后向传播，y_t是节点在时间t时刻的输出向量，b_f，b_b，b_f表示相应的偏移量。

2.2 多头注意力机制

注意力机制（Attention）的核心是借鉴人脑在特定时间对于某一种事物的注意力关注，大脑集中在该事物的某一焦点位置，而忽略事物的其他位置。在本文中，利用多头注意力机制基于自注意力模块提取更具表现力的序列表示，并联合多个注意力表征进行情感建模。多头注意力模型扩展了传统的注意力机制，具有多个头部，而且每个头部都可以产生不同的注意力分布。这样可以有效的学习音视频信息中的长时依赖性，通过位置对计算不同位置之间的关系，可以在较长时间跨度上学习时序依赖性，而且只需计算一次即可得到变换后的表示。在本实验中，注意力机制首先分别线性变换查询Q（query）、键K（key）、值V（value）三个输入，然后计算h次放缩点积注意力（scaled dot-product attention）。每个放缩点积注意力是单独计算的，最终组合其所有的输出到另一个线性层获得最终结果，如公式（4）、（5）所示。

$$MultiHead(Q,K,V) = Concat(head_1,\dots,head_m)W^o \tag{4}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{5}$$

其中，参数矩阵$W_i^Q \in R^{d_{model}\times d_k}, W_i^K \in R^{d_{model}\times d_k}, W_i^V \in R^{d_{model}\times d_k}, W_i^O \in R^{md_v\times d_{model}}$。

2.3 评价指标

维度情感的性能的评价标准是一个不断探索的问题，早期的文献一般采用均方误差（Root Mean square err, RMSE）来度量估计的性能。设$\hat{\theta}$是估计的标签，θ是真实的标签，n是样本数目，$\sigma_{\hat{\theta}}^2$，σ_{θ}^2分别是$\hat{\theta}$和θ的方差，$\mu_{\hat{\theta}}$，μ_{θ}分别是$\hat{\theta}$和θ的期望，则 RMSE 的定义为：

$$\mathrm{R}MSE = \sqrt{\frac{1}{n}\sum_{f=1}^{n}\left(\hat{\theta}(f)-\theta(f)\right)^2} \tag{6}$$

RMSE 描述了预测与真值的偏差，但 RMSE 对于异常值敏感，且无法对θ和$\hat{\theta}$的相对变化趋势进行描述，因此不能很好地描述与真值的吻合度。鉴于 RMSE 的确定皮尔逊系数（Pearson correlation coefficient，PCC）被用来作为维度情感预测的评价指标，其定义为：

$$\rho = \frac{\frac{1}{n}\sum_{f=1}^{n}[(\hat{\theta}(f)-\mu_{\hat{\theta}})(\theta(f)-\mu_{\theta})]}{\sigma_{\hat{\theta}}\sigma_{\theta}} = \frac{E[(\hat{\theta}-\mu_{\hat{\theta}})(\theta-\mu_{\theta})]}{\sigma_{\hat{\theta}}\sigma_{\theta}} \tag{7}$$

PCC 的取值范围是[-1,1],它反映了预测与真值具有线性关系的紧密程度。PCC 能够很好地反映预测与真值的协同变化关系。但是，由于 PCC 对预测的幅值不敏感，无法对θ和$\hat{\theta}$的偏差进行度量，因此仍不能很好地描述与真值的吻合程度。为了更好地描述预测与真值的吻合程度，一致性相关系数（Concordance correlation coefficient， CCC）作为预测性能的评价指标，其定义为：

$$\rho_c = \frac{2\rho\sigma_{\theta}\sigma_{\hat{\theta}}}{\sigma_{\theta}^2+\sigma_{\hat{\theta}}^2+(\mu_{\hat{\theta}}-\mu_{\theta})^2} \tag{8}$$

CCC 结合了 PCC 与 RMSE 的优点，既反应了预测与真值的协同变化关系又反应了预测与真值的温和程度，是目前广泛使用的维度情感预测性能评价指标，所以本文实验均以 CCC 为最终评价标准。

2.4 数据增广

为了解决维度情感数据集的数据量较小，在情感识别模型训练过程中会出现过拟合的问题，本文提出了针对多模态数据集的数据增广方法，具体如下：

（1）音频数据：对数据集进行数据预处理之后，对每个数据集中的视频以音频为主进行数据分段，在这个过程中，分段的大小和不同识别模型进行学习的性能会有所差异，经过多次探索，数据段大小为 6s 的时候使用本文提出的模型，识别性能最好，同时记录分段时间戳。

（2）视频数据：根据音频的分段时间戳，对视频进行分段，因为数据集的视频规格为 50 帧/秒，所以分段之后的视频信息为每段 300 帧。

（3）数据段：按照时间戳对数据进行分段之后，每个原始视频都会得到多个数据段，再将数据段用于相应的情感特征提取。

2.5 标签延迟

表 1 不同特征标签延迟的 CCC

Tab.1 CCC with different feature annotation delays

	BoAW		IS2010		BoVW		VGGFace	
	激活度	愉悦度	激活度	愉悦度	激活度	愉悦度	激活度	愉悦度
0.0	0.314	0.409	0.332	0.356	0.471	0.499	0.458	0.476
0.2	0.289	0.347	0.338	0.352	0.480	0.504	0.457	0.479
0.4	0.277	0.340	0.331	0.354	0.492	0.509	0.462	0.483
0.6	0.320	0.331	0.318	0.335	0.501	0.512	0.469	0.486
0.8	0.324	0.328	0.314	0.330	0.521	0.513	0.473	0.488
1.0	0.344	0.368	0.305	0.327	0.520	0.522	0.477	0.453
1.2	0.358	0.376	0.303	0.322	**0.537**	0.529	0.483	0.462
1.4	0.372	0.369	0.292	0.318	0.527	0.533	0.486	0.465
1.6	0.391	0.408	0.278	0.320	0.510	0.538	0.490	**0.480**
1.8	0.410	0.436	0.322	0.338	0.491	**0.542**	0.520	0.476
2.0	0.424	0.447	0.347	0.352	0.473	0.540	0.510	0.471
2.2	0.432	0.457	0.342	0.359	0.461	0.537	**0.513**	0.469
2.4	0.441	0.485	0.360	0.367	0.461	0.538	0.507	0.467
2.6	0.451	0.482	0.366	0.375	0.458	0.532	0.499	0.462
2.8	0.459	0.484	**0.371**	**0.388**	0.452	0.530	0.487	0.452
3.0	**0.461**	**0.486**	0.368	0.370	0.447	0.527	0.488	0.450

对于数据集的情感标签来说，不同的标注人员在标注时会有不同的反应时间，而且某一时刻标注的情感状态会受到前面音视频数据的影响，从而导致情感标签和真实的情感状态会有不同程度的偏差。因此，本文首先对数据集的情感标签进行延迟优化实验。

如表 1 所示，首先对音频特征（BoAW 和 IS2010）进行 0.0s-3.0s 的标签延迟实验（标签延迟步长为 0.2s）。BoAW 特征在 3.0s 标签延迟下的 CCC 性能达到最大，分别为 0.461 和 0.486。IS10 特征的激活度和愉悦度在 2.8s 标签延迟下的 CCC 性能达到最大，分别为 0.371 和 0.388。然后对视频特征（BoVW 和 VGGFace）进行 0.0s～3.0s 的标签延迟实验。BoVW 特征的激活度在 1.2s 标签延迟下的 CCC 性能达到最大，为 0.537。愉悦度在 1.8s 标签延迟下的 CCC 性能达到最大，为 0.542。VGGFace 特征的激活度在 2.2s 标签延迟下的 CCC 性能达到最大，为 0.513。愉悦度在 1.6s 标签延迟下的 CCC 性能达到最大，为 0.480。音频特征和视频特征的激活度、愉悦度的最佳标签延迟都比较接近，说明所选择的情感特征在维度情感识别上都较为稳定[14]。

3 实验结果与分析

3.1 数据集介绍

本文中使用 AVEC2017 数据集上的多态维情感数据集进行相关实验。AVEC2017 数据集使用的语料库是 SEW A 数据集的子集[18]。该数据集是由对象的自然行为组成的音频和视频数据集。使用网络摄像头和麦克风在受试者家中的计算机上收集所有数据。对象的年龄在 18 至 60 岁之间，并且每个记录中只有一个人的数据。因此，实验数据减少了对音频情感标签的干扰。在数据集的注释过程中，注释者根据激活和效价两种情感维度对生活，工作或其他讨论内容进行注释。整个注释过程由 6 位注释者（3 位女性，3 位男性）进行，他们年龄在 20-24 岁之间，并且会说德语，每隔 0.1s 进行一次标签注释。在该实验中，共有 48 个音频和 48 个对应的视频，每个视频的录制时间范围从 46 秒到 3 分钟不等，为了使实验结果更加准确，在实验开始前，对数据集进行了三种分类方式，分别记作标签 1、标签 2 和标签 3，每种标签选择音视频的根据完全随机，然后根据每种标签的不同分类做了 3 次实验，其中每种标签的 34 个音频和视频用作训练集，7 个音频和视频用作验证集，另外 7 个音视和视频用作测试集。

3.2 参数设置

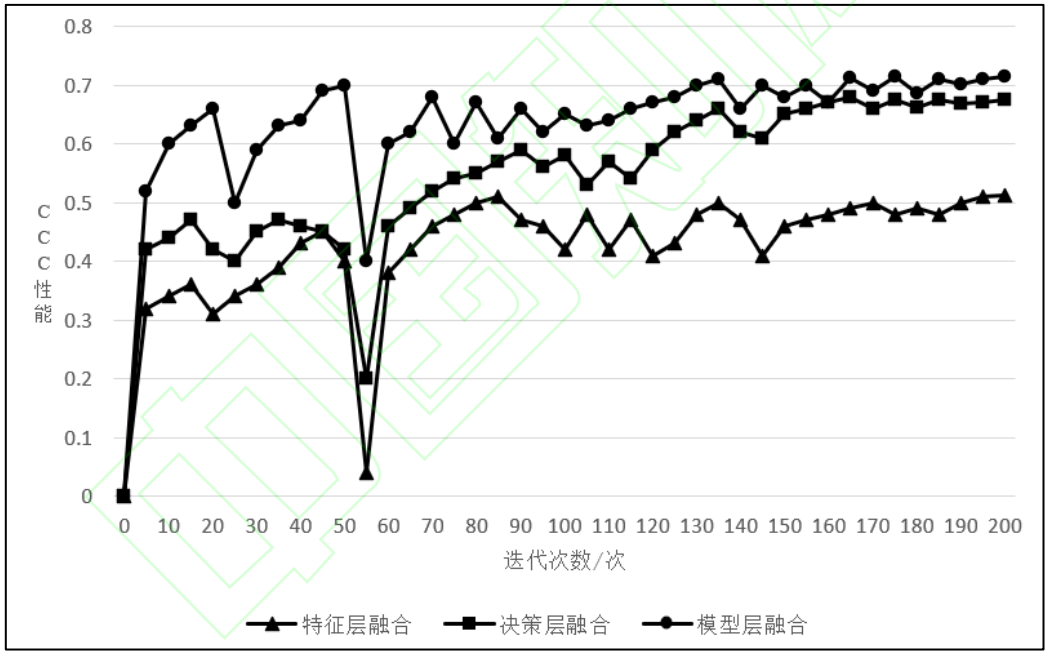

图 2　三种融合方法的迭代次数和 CCC 性能可视化结果

Fig.2　Visualization of epochs and CCC of three fusion methods

基于音视频的模型层融合实验中由两个四头注意力层组成，并且每两个注意力层有一个残差连接和层归一化，注意力的隐藏节点数和一维时序层的输出节点数均为 64。BLSTM 中的层数设置为 3，卷积核均为$3\times3\times3$，隐藏层对不同的输入节点特征信息进行了优化。使用的优化器是 Adam，损失函数使用 crossentropyloss，学习率初始化为 0.01，每 40 个周期减少一半，最大迭代次数为 200，样本批次大小为 128。如图 2 所示为三种融合方法的 epochs 和 CCC 性能的可视化结果，从中可以看出模型层融合的稳定性和 CCC 性能都要比其他两种模型要好。

3.3 实验结果与分析

表 2 单模态和三种多模态融合方法的实验结果

Tab.2 Experiment result of single-model and three multi-model fusion methods

	激活度			愉悦度		
	RMSE	PCC	CCC	RMSE	PCC	CCC
单模态音频	—	—	0.380	—	—	0.409
单模态视频	—	—	0.472	—	—	0.510
特征层融合	0.134	0.592	0.512	0.138	0.522	0.481
决策层融合	0.075	0.697	0.676	0.091	0.637	0.624
模型层融合	0.069	0.737	**0.715**	0.080	0.760	**0.759**

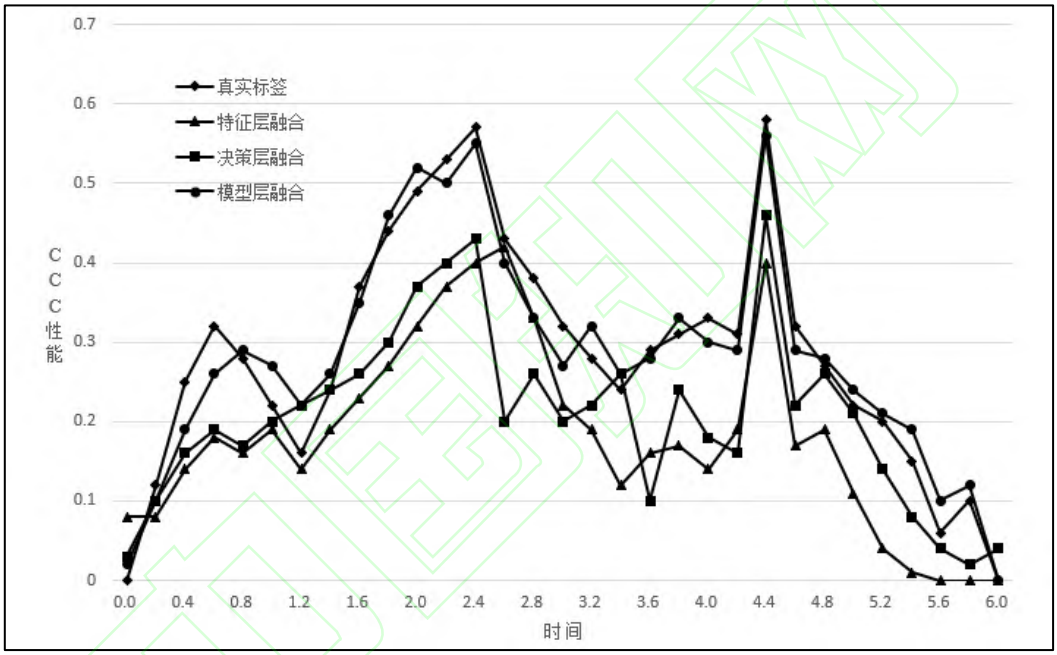

图 3 三种融合方法的情感预测值和真实值可视化结果

Fig.3 Visualization of sentiment prediction and true value of three fusion methods

如表 2 所示，本文首先使用 BLSTM 模型对音频和视频进行了单模态情感识别，在 BLSTM 模型中，IS2010 作为短时音频特征、BoAW 作为长时音频特征在长短时记忆网络中学习效果较好。在视频特征中，深度学习网络的特征提取，可以对视频的时空信息进行有效建模，尤其是 VGGFace 网络相比于其他深度学习网络更适合用于情感特征提取，特别是对于短时的情感识别模型中，而基于 openXBOW 的 BoVW 特征集，适合在长时的情感识别建模中使用。结果显示，视频在激活度和愉悦度上的 CCC 性能都要比音频高，音频情感信息相比于视频情感信息较弱，验证了视频情感特征的性能强于音频情感特征的性能[9]。

本文进一步使用 BLSTM 模型分别进行了基于特征层融合和决策层融合的维度情感识别实验。特征层融合在激活度上比单模态维度情感识别更有效，为 0.512，它的主要的优势在于基于深度学习的情感特征可以更好的进行学习，但是在实验结果中，愉悦度上的 CCC 性能并没有提升，这是由于特征层融合使得不同模态信号之间的相互作用很难在不同时间从不同但耦合的模态中提取有效的情感特征。此外实验过程中过多的维度使得实验性能下降，也印证了特征层融合面临着不同模态之间同步的问题，容易出现维度灾难的问题。决策层融合在激活度和愉悦度上的 CCC 性能都要高于特征层融合，分别为 0.676 和 0.624，性能提升很大，它主要是利用了不同模态的情感识别高层表征，对最终情感进行了非线性决策，但是在实验

结果中，不同的模态组合的性能相差较大，支配度上的性能提升也并不明显，这主要是因为决策层融合需要多模态信息流的同步，而且没有考虑到不同模态情感特征信息的关联。

模型层融合的 CCC 性能在激活度和愉悦度上的结果比特征层融合和决策层融合的结果都要好，分别为 0.715 和 0.759。结果表明，模型层融合进一步解决了多模态信息流的同步和不同模态情感特征信息关联的问题，因此模型层融合使视频模态可以更好的接收音频模态的情感信息，并实现了音视频的有效融合。如图 3 所示为实验过程中单个样本的三种融合方法情感预测值和真实值的可视化结果，模型层融合的预测值与真实值较为吻合，也证明了模型层融合方法相对于特征层和决策层来说，在 CCC 性能上表现更好。最后，本文比较了使用相同数据集的研究人员的方法，如表 3 所示，本文提出的模型层融合方法在激活度和愉悦度上表现最佳。这表明多头注意力机制可以有效的生成情感特征表征，进一步使得 BLSTM 模型可以更好地进行维度情感识别建模。

表 3　实验结果和其他论文方法比较

Tab.3　Comparison of our experiment with other paper methods

性能	激活度		愉悦度	
	RMSE	CCC	RMSE	CCC
基线方法[9]	—	0.361	—	0.437
CNN+特征层融合[10]	—	—	0.036	0.367
RNN+特征层融合[11]	—	0.565	—	0.499
Bi-RNN+单模态[5]	0.371	0.532	0.356	0.696
LSTM+决策层融合[15]	—	0.583	—	0.654
ConvLSTM+决策层融合[14]	0.093	0.599	0.085	0.721
LSTM-RNN+决策层融合[16]	—	0.672	—	0.756
本文方法 1（特征层融合）	0.134	0.512	0.138	0.481
本文方法 2（决策层融合）	0.075	0.676	0.091	0.624
本文方法 3（模型层融合）	0.069	**0.715**	0.080	**0.759**

4　结论

本文在基于 BLSTM 模型的特征层和决策层融合维度情感识别实验的基础上，通过借鉴注意力机制模型，构建了基于多头注意力机制的模型层融合维度情感识别模型，并进行了基于音视频的模型层融合实验，本文提出的基于多头注意力机制的模型层荣和维度情感识别方法在激活度上的 CCC 性能表现较为突出，比第二名高出了 6.40%，说明了在维度情感识别领域，模型层融合对于情感的兴奋水平识别较为明显，而在愉悦度上的 CCC 性能仅比基线方法高出了 0.40%，说明在模型层融合中，对于情感的正负状态识别性能略低，这主要是由于对立情感在维度上的表现相近造成的，比如高兴和悲伤，两种情感的维度绝对值都较为相近。但是本文的方法在激活度和愉悦度上都表现出了最好的性能，说明本文所提方法可以融合模型中的高层输出使得音频和视频信息有效融合，进一步提高情感识别的 CCC 性能。多模态情感融合目前常用方法是模型层融合，但是模型层融合在性能和鲁棒性方面还存在一些问题。因此，考虑缺失信息的多模态模型层的情感融合将是下一步的研究重点之一。

参考文献

[1] TAO Jianhua, TAN Tieniu. Affective computing: A review[J]//Affective Computing and Intelligent Interaction. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005: 981-995.

[2] ANTONETTI P, VALOR C. A theorisation of discrete emotion spillovers: an empirical test for anger[J]. Journal of Marketing Management, 2020(1):1-27.

[3] LIU Meng, INCE R A A, CHEN Chaona, et al. Emotion categories are represented by a 2-dimensional valence-arousal space[J].

Journal of Vision, 2020, 20(11): 1224.

[4] COWIE R, DOUGLAS-COWIE E, TSAPATSOULIS N, et al. Emotion recognition in human-computer interaction[J]. IEEE Signal Processing Magazine, 2001, 18(1): 32-80.

[5] WANG XIAOHUA, PENG MUZI, PAN LIJUAN, et al. Two-level attention with two-stage multi-task learning for facial emotion recognition[J]. Vis Commun Image R, 2019: 217-225.

[6] CHEN L, WU Min, PEDRYCZ W, et al. Emotion recognition and understanding for emotional human-robot interaction systems[M]. Cham: Springer International Publishing, 2021.

[7] KWAK Y, KONG K, SONG W J, et al. Multilevel feature fusion with 3D convolutional neural network for EEG-based workload estimation[J]. IEEE Access, 2020, 8: 16009-16021.

[8] XU Haiyang, ZHANG Hui, HAN Kun, et al. Learning alignment for multimodal emotion recognition from speech[C]//Interspeech 2019. ISCA: ISCA, 2019: 3569-3573.

[9] Singh N, Dhall A. Continuous Multimodal Emotion Recognition Approach for AVEC 2017[J]. ArXiv Preprint ArXiv:1709.05861, 2017.

[10] BASNET R, ISLAM M T, HOWLADER T, et al. Statistical selection of CNN-based audiovisual features for instantaneous estimation of human emotional states[C]//2017 International Conference on New Trends in Computing Sciences (ICTCS). Amman. IEEE, 2017: 50-54.

[11] SAMAREH A, JIN Yan, WANG Zhangyang, et al. Detect depression from communication: How computer vision, signal processing, and sentiment analysis join forces[J]. IISE Transactions on Healthcare Systems Engineering, 2018, 8(3): 196-208.

[12] PORIA S, CAMBRIA E, HOWARD N, et al. Fusing audio, visual and textual clues for sentiment analysis from multimodal content[J]. Neurocomputing, 2016, 174: 50-59.

[13] SEBASTIAN J, PIERUCCI P. Fusion techniques for utterance-level emotion recognition combining speech and transcripts[C]//Interspeech 2019. ISCA: ISCA, 2019: 51-55.

[14] HUANG Jian, LI Ya, TAO Jianhua, et al. Continuous multimodal emotion prediction based on long short term memory recurrent neural network[C]//AVEC '17: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. 2017: 11-18.

[15] HUANG Jian, LI Ya, TAO Jianhua, et al. End-to-end continuous emotion recognition from video using 3D convlstm networks[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada. IEEE, 2018: 6837-6841.

[16] CHEN Shizhe, JIN Qin, ZHAO Jinming, et al. Multimodal multi-task learning for dimensional and continuous emotion recognition[C]//AVEC '17: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. 2017: 19-26.[LinkOut]

[17] ZHANG Su, GUAN Cuntai. Emotion recognition with refined labels for deep learning[C]//2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). Montreal, QC, Canada. IEEE, 2020: 108-111.

[18] RINGEVAL F, SCHULLER B, VALSTAR M, et al. AVEC 2017: Real-life depression, and affect recognition workshop and challenge[C]//AVEC '17: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. 2017: 3-9.

作者简介

董永峰 男，1977 年生，河北人。河北工业大学人工智能与数据科学学院教授。研究方向为智能信息处理、知识图谱。手机号：13820865534。

E-mail：dongy@hebut.edu.cn

苏海洋 男，1994 年生，河北人。河北工业大学人工智能与数据科学学院，硕士研究生。研究方向为情感计算、语音识别。手机号：18812669876。

E-mail：haiyang3565@qq.com

刘斌 男，1984 年生，内蒙古人。中国科学院自动化研究所副研究员。研究方向为语音信号处理、语音增强与语音编码。手机号：113522681618。

E-mail：liubn@nlpr.ia.ac.cn

陶建华 男，1972 年生，江苏人。中国科学院自动化研究所研究员，博士生导师。研究方向为语音识别与合成、人机交互。手机号：15620128852。

E-mail：jtao@nlpr.ia.ac.cn