计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(9):112−119 [doi: 10.15888/j.cnki.csa.009924] [CSTR: 32024.14.csa.009924]

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

基于 KLSTM 的互信息视听情绪识别①

罗志鑫[1], 刘知贵[1,3], 唐　荣[1], 潘志祥[2], 李　理[1,3]

[1](西南科技大学 信息与控制工程学院, 绵阳 621000)
[2](四川湖山电器股份有限公司, 绵阳 621025)
[3](四川省工业自主可控人工智能工程技术研究中心, 绵阳 621010)
通信作者: 罗志鑫, E-mail: luozhixinlzx@163.com

摘　要: 针对视听情绪识别中如何高效融合音视频特征并准确提取时间依赖性情绪信息的问题, 本研究提出了一种基于 KLSTM (Kolmogorov-Arnold long short-term memory) 的互信息视听情绪识别模型. 利用互信息方法进行特征选择和自适应窗口处理, 从音频和视频信号中提取情绪相关的关键片段, 以减少信息冗余; 在特征提取中引入 KLSTM 网络, 有效捕捉视听模态信号的时间依赖特性; 在融合阶段, 通过跨模态一致性最大化确保音视频特征的协调性与互补性. 实验结果显示所提模型在 CMU-MOSI 和 CMU-MOSEI 数据集上均优于现有基准模型, 验证了其在多模态情绪识别任务中的有效性.

关键词: 情绪识别; 视听融合; 互信息

引用格式: 罗志鑫,刘知贵,唐荣,潘志祥,李理.基于 KLSTM 的互信息视听情绪识别.计算机系统应用,2025,34(9):112–119. http://www.c-s-a.org.cn/1003-3254/9924.html

KLSTM-based Mutual Information Audiovisual Emotion Recognition

LUO Zhi-Xin[1], LIU Zhi-Gui[1,3], TANG Rong[1], PAN Zhi-Xiang[2], LI Li[1,3]

[1](School of Information and Control Engineering, Southwest University of Science and Technology, Mianyang 621000, China)
[2](Sichuan Hushan Electrical Appliance Co. Ltd., Mianyang 621025, China)
[3](Sichuan Engineering and Technology Research Center for Industrial Autonomous and Controllable Artificial Intelligence, Mianyang 621010, China)

Abstract: Addressing the challenge of efficiently fusing audio and video features while accurately extracting time-dependent emotion information in audiovisual emotion recognition, a mutual information-based audiovisual emotion recognition model is proposed, incorporating Kolmogorov-Arnold long short-term memory (KLSTM). Feature selection and adaptive window processing, based on the mutual information approach, are employed to extract emotionally relevant key segments from audio and video signals, effectively reducing information redundancy. The KLSTM network is integrated into feature extraction to capture the temporal dependencies of audiovisual modal signals. In the fusion stage, cross-modal consistency maximization ensures the coordination and complementarity of audio and video features. Experimental results demonstrate that the proposed model outperforms existing benchmark models on both CMU-MOSI and CMU-MOSEI datasets, validating its effectiveness in multimodal emotion recognition tasks.

Key words: emotion recognition; audiovisual fusion; mutual information

近年来, 随着深度学习在语音和视觉信号处理中的广泛应用, 基于多模态数据的情绪识别研究取得了显著进展. 其中情绪识别系统通过结合音频和视觉特征, 更加准确地识别人类的情绪状态, 并广泛应用于人

① 基金项目: 国家自然科学基金 (U21A20157)
收稿时间: 2025-01-07; 修改时间: 2025-02-12; 采用时间: 2025-02-24; csa 在线出版时间: 2025-06-24
CNKI 网络首发时间: 2025-06-25

机交互[1]、心理健康监测[2]等领域. 音频和视觉信号在情绪表达上具有互补性, 两种模态融合有助于提高情绪识别的精度和鲁棒性. 然而, 由于情绪信号的动态特性以及不同模态信号在融合过程中的不稳定性, 难以在长时间情绪变化和复杂场景中保持稳定的视听融合效果.

1 相关工作

互信息 (mutual information, MI) 是一种用于度量两个变量之间相关性和信息共享程度的统计量, 反映了一个变量对另一个变量提供的信息量大小. 具体来说, 互信息量化了两个随机变量的联合分布与各自独立分布之间的差异. 当两个变量高度相关时, 它们的互信息值较大, 表明通过一个变量可以获得关于另一个变量的较多信息; 相反, 当两个变量独立时, 互信息值接近零, 表示一个变量无法提供关于另一个变量的有效信息. 互信息能通过最大化模态间的信息共享度, 提升融合表示的情绪相关性, 有助于构建鲁棒的情绪识别模型. 黎倩尔等[3]提出了一种基于互信息最大化和对比损失的多模态对话情绪识别模型. 模型通过在输入级和融合级上分级最大化模态之间的互信息, 使任务相关信息在融合过程中得以保存, 从而生成更丰富的多模态表示. Piho 等[4]通过基于互信息的自适应窗口方法选择脑电信号 (EEG) 中最具情绪信息的片段, 从而提高情绪识别的分类准确性和计算效率. Hacine-Gharbi 等[5]使用互信息为语音情绪识别中的特征选择提供了标准, 通过基于互信息的特征选择方法筛选出与情绪类别相关的信息, 从而在减少特征数量的同时保持较高的分类性能. Wang 等[6]基于归一化互信息 (NMI) 提出了一种 EEG 信道选择方法, 该方法用于减少信道数量以提高情绪识别的计算效率和系统的鲁棒性, 同时在选择的信道中保持高识别精度.

在视听情绪识别中, 动态变化的情绪信号对模型的时序特征捕捉能力提出了更高要求. 递归神经网络 (recurrent neural network, RNN)[7]和长短期记忆网络 (long short-term memory, LSTM)[8]虽在时间序列处理上取得进展, 但在长时间序列下易出现梯度消失或爆炸, 导致情绪信息有效捕捉不足. 近年来, Kolmogorov-Arnold network (KAN)[9] 作为一种基于 KAN 表示定理的新型神经网络架构被提出, 通过组合一维函数实现多维函数的逼近, 具有优越的非线性映射能力. Shen 等[10]通过将基于 KAN 的模型应用于图像处理领域, 证明了该方法在捕捉图像数据中复杂非线性关系方面的优越性, 实现了图像分类任务中准确性和泛化能力的显著提升. Vo-thanh 等人[11]将 KAN 应用于音频处理, 提出了一种双域融合模型, 能够更加精准地建模音频驱动的面部表情变化, 表明 KAN 在音频模态建模中的有效性. 然而, 标准 KAN 结构缺乏处理时间依赖的机制, 难以直接应用于情绪信号的动态特征提取中.

本文基于上述研究现状, 提出基于 KLSTM 的互信息视听情绪识别模型 (mutual information audiovisual emotion recognition model based on KLSTM, MIKL).

2 方法

在基于 KLSTM 的互信息视听情绪识别模型中引入互信息方法, 优化视听情绪特征的选择与融合, 通过互信息特征选择和自适应窗口方法提取情绪相关的高信息量特征, 并通过分层互信息最大化增强跨模态一致性, 从而在融合过程中保留了不同模态的关键情绪信息, 提高了情绪识别的准确性和鲁棒性. 设计了基于 KLSTM 的特征提取模块, 该模块结合了 Kolmogorov-Arnold 网络的非线性表示能力与 LSTM 的长短期记忆管理特性, 能够有效捕捉情绪信号的复杂时间依赖, 增强模型对动态情绪特征的感知和建模能力.

在基于 KLSTM 的互信息视听情绪识别模型中音视频信号首先经过特征选择模块. 该模块利用互信息筛选出与情绪高度相关的音频和视觉特征, 从而去除冗余信息, 确保输入特征的高信息量. 互信息自适应窗口模块选择包含情绪信息最强的时间片段, 以便在时序处理阶段聚焦于最具代表性的情绪特征片段, 提升模型的计算效率和准确性. 经过自适应窗口处理后的特征进入 KLSTM 网络, KLSTM 模块用于捕捉时间序列中的短期和长期依赖特征. 模型通过跨模态一致性最大化模块 (基于互信息) 来增强音频与视觉特征之间的信息共享度. 此模块通过最大化音频和视觉模态之间的互信息, 确保视听数据在融合时保持较高的一致性. 本文所提的 MIKL 模型如图 1 所示.

2.1 互信息特征选择

特征选择旨在从音频和视觉数据中提取与情绪标签密切相关的特征, 以去除冗余或无关的信息. 互信息 (MI) 是一种用于度量两个随机变量之间依赖关系的统计量, 能够反映特征与情绪标签之间的关联程度. 在视听情绪识别中, 通过互信息筛选特征可以提高模型的计算效率和泛化性能.

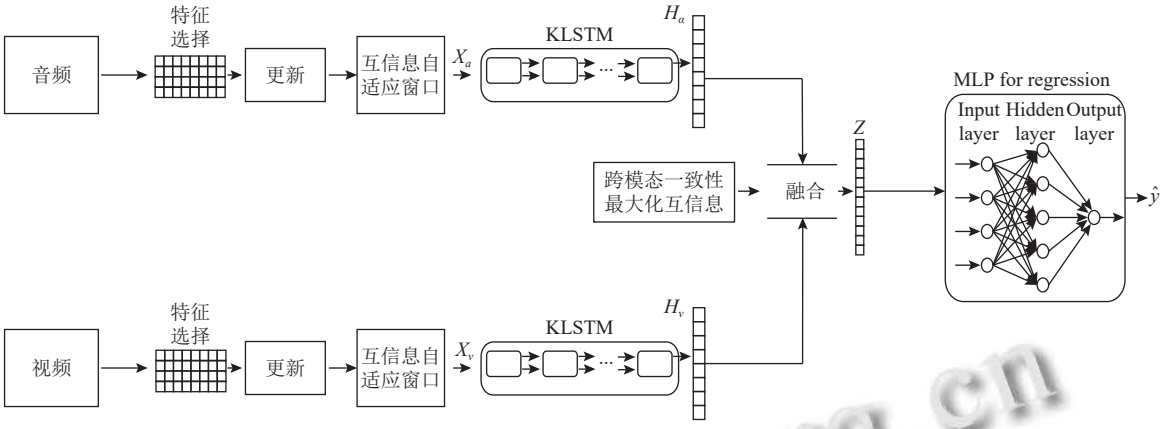

图 1 MIKL 模型的总体架构

首先, 计算特征与情绪标签的联合概率分布. 假设特征 X_i 和情绪标签 Y 的联合概率 $p(x,y)$ 定义为:

$$p(x,y) = P(X_i = x, Y = y) \tag{1}$$

其中, $p(x,y)$ 表示特征值 $X_i = x$ 和情绪标签 $Y = y$ 同时出现的概率. 联合概率分布可以通过样本数据中 X_i 和 Y 的联合观测来估计.

接下来, 计算特征和情绪标签的边缘概率. 边缘概率分布 $p(x)$ 和 $p(y)$ 定义为特征和情绪标签单独出现的概率, 分别为:

$$\begin{cases} p(x) = \sum\limits_{y} p(x,y) \\ p(y) = \sum\limits_{x} p(x,y) \end{cases} \tag{2}$$

边缘概率可以通过在联合概率中对一个变量求和得到, 表示特征或情绪标签独立存在的概率.

最后, 将联合概率和边缘概率代入互信息公式, 用于衡量特征 X_i 中包含的情绪标签 Y 的信息量. 互信息 $I(X_i;Y)$ 的定义如下:

$$I(X_i;Y) = \sum_{x\in X_i}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{3}$$

其中, $\dfrac{p(x,y)}{p(x)p(y)}$ 为联合概率与边缘概率之比, 用于衡量特征值 X_i 和标签 Y 的关联强度.

若 X_i 和 Y 独立, 联合概率等于边缘概率乘积, 则 $I(X_i;Y) = 0$; 当互信息值较大时, 表示特征 X_i 对情绪标签 Y 的预测能力较强.

通过逐一计算每个特征与情绪标签的互信息值, 可以筛选出互信息值较大的特征, 用于情绪识别模型的输入.

2.2 互信息自适应窗口

动态片段处理的目的是通过计算不同时间片段和情绪标签之间的互信息值, 从音频和视觉信号中提取最具情绪表达的信息片段. 自适应窗口互信息方法可以有效识别出含有较强情绪信息的时间片段, 从而提升情绪识别的准确性和计算效率.

首先, 将音频或视觉信号划分为多个长度为 k 的滑动窗口 W_k, 每个窗口包含一段信号子集 $\{s_t, s_{t+1}, \cdots, s_{t+k-1}\}$. 窗口的位置可以在信号序列上以滑动或不重叠的方式选取. 假设每个窗口片段 W_k 包含情绪信息的分布特征, 则可以计算片段 W_k 与情绪标签 Y 的联合概率分布. 窗口 W_k 和情绪标签 Y 的联合概率分布 $p(w,y)$ 表示片段 W_k 中信号与情绪标签 Y 的共同分布, 可定义为:

$$p(w,y) = P(W_k = w, Y = y) \tag{4}$$

其中, $p(w,y)$ 通过观测每个时间窗口 W_k 与情绪标签的共现频率估计, 反映了该时间片段中包含的情绪信息量.

然后, 对片段 W_k 的边缘概率 $p(w)$ 和情绪标签的边缘概率 $p(y)$ 定义为: 它们分别表示片段或情绪标签的独立出现概率. 边缘概率的计算公式如下:

$$\begin{cases} p(w) = \sum\limits_{y} p(w,y) \\ p(y) = \sum\limits_{w} p(w,y) \end{cases} \tag{5}$$

其中, $p(w)$ 是片段 W_k 的边缘分布, 表示片段独立于情绪标签时的概率, 而 $p(y)$ 表示情绪标签的边缘概率.

最后, 通过联合概率和边缘概率计算窗口片段 W_k 与情绪标签 Y 的互信息. 互信息 $I(W_k;Y)$ 量化了片段 W_k 与情绪标签之间的依赖关系, 定义如下:

$$I(W_k;Y)=\sum_{w\in W_k}\sum_{y\in Y}p(w,y)\log\frac{p(w,y)}{p(w)p(y)} \tag{6}$$

其中, $\frac{p(w,y)}{p(w)p(y)}$表示片段与标签间信息关联的强度. 当该比值大于 1 时, 表示片段 W_k含有与情绪标签密切相关的信息; 比值越高, 依赖关系越强, 表明片段包含更多情绪信息.

为找到最佳片段, 通过比较不同时间窗口的互信息值, 选取互信息值最大的窗口片段$W_k^*=\arg\max I(W_k;Y)$. 该片段保留了最具情绪表达的信息, 可以显著提升情绪分类的准确性, 同时降低模型计算复杂度.

2.3 KLSTM 特征提取网络

在视听情绪识别的特征提取中 KAN 结合了和 LSTM 模块, 以有效捕捉情绪信号序列中的长短期依赖关系. KLSTM 的设计充分利用了 Kolmogorov-Arnold 网络的非线性映射能力和 LSTM 的记忆管理机制, 为情绪特征的提取提供了强大支持.

首先, KLSTM 层通过 KAN 递归核捕捉时间序列中的短期依赖性, 并在层内嵌入记忆管理模块, 使网络能够动态保持每一时刻的状态信息. 对于每一时刻t, 隐藏状态 h_t的更新遵循以下公式:

$$h_t=f(W_{hh}\cdot h_{t-1}+W_{hx}\cdot x_t+b_h) \tag{7}$$

其中, W_{hh}和W_{hx}分别是前一时刻隐藏状态h_{t-1}和当前输入x_t的权重矩阵, b是偏置项, f为非线性激活函数. 此公式表示每一时刻的隐藏状态通过前一时刻的状态和当前输入共同确定, 反映了短期记忆的动态更新机制.

KLSTM 层在每个时间步t应用时间依赖的激活函数$\varphi_{l,j,i,t}$, 该函数结合了输入的当前状态和历史信息, 具体定义为:

$$x_{l+1,j}(t)=\sum_{i=1}^{n_l}\varphi_{l,j,i,t}(x_{l,i}(t),h_{l,i}(t)) \tag{8}$$

其中, $x_{l+1,j}(t)$表示第$l+1$层的第j个节点在时间t上的输出, $h_{l,i}(t)$是第 l层第i个节点的历史状态. 每个节点的历史状态更新为:

$$h_{l,i}(t)=W_{hh}\cdot h_{l,i}(t-1)+W_{hx}\cdot x_{l,i}(t) \tag{9}$$

该公式通过引入权重矩阵 W_{hh}和W_{hx}对先前隐藏状态和当前输入的相对重要性进行加权, 从而捕捉时间依赖性.

其次, KLSTM 通过结合 LSTM 模块实现对长时记忆的有效管理. LSTM 模块的遗忘门f_t控制历史信息的保留:

$$f_t=\sigma(W_f\cdot x_t+U_f\cdot h_{t-1}+b_f) \tag{10}$$

其中, W_f和U_f是当前输入x_t和前一时刻隐藏状态h_{t-1}的权重矩阵, σ为 Sigmoid 激活函数. 此门控机制决定了前一状态信息的保留程度.

输入门i_t控制新信息的写入, 公式如下:

$$i_t=\sigma(W_i\cdot x_t+U_i\cdot h_{t-1}+b_i) \tag{11}$$

输出门o_t控制当前状态信息的输出, 公式如下:

$$o_t=\sigma(KAN(x,t)) \tag{12}$$

其中, $KAN(x,t)$是通过 KAN 提取的非线性特征. 在 LSTM 模块中, 记忆单元c_t保存更新后的长期信息, 公式为:

$$c_t=f_t\odot c_{t-1}+i_t\odot\tilde{c}_t \tag{13}$$

其中, $\tilde{c}_t=\tanh(W_c\cdot x_t+U_c\cdot h_{t-1}+b_c)$是当前输入的候选记忆, 遗忘门和输入门共同决定记忆单元c_t中存储的信息.

最终的输出隐藏状态h_t代表当前时间步的特征输出, 公式为:

$$h_t=o_t\odot\tanh(c_t) \tag{14}$$

这一输出既包含了 KAN 的短期信息, 也结合了 LSTM 模块的长期记忆, 适用于视听情绪信号中复杂时间依赖的特征提取任务. 通过 KAN 层的非线性映射能力与 LSTM 的门控记忆机制相结合, KLSTM 架构能够有效地捕捉视听情绪数据中的复杂时间序列特征, 增强情绪识别的准确性和鲁棒性. KLSTM 特征提取块的结构如图 2 所示.

2.4 跨模态一致性最大化互信息

跨模态一致性最大化旨在通过最大化音频和视觉特征之间的互信息, 使融合表示能够保留各模态的关键情绪信息, 确保视听数据在融合后仍能有效反映情绪特征. 这一方法通过模态间互信息和融合表示互信息的分层计算实现视听数据的深度融合, 最终提升情绪识别的准确性和鲁棒性.

首先, 音频特征X_a和视觉特征X_v之间的模态间互信息$I(X_a;X_v)$用于衡量两个模态特征之间的信息共享程度, 以确保音频和视觉特征的一致性. 模态间互信息依赖于音频和视觉特征的联合概率分布$p(x_a,x_v)$和边

缘概率分布$p(x_a)$和$p(x_v)$公式如下:

$$\begin{cases} p(x_a,x_v)=P(X_a=x_a,X_v=x_v) \\ p(x_a)=\sum\limits_{x_v\in X_v}p(x_a,x_v) \\ p(x_v)=\sum\limits_{x_a\in X_a}p(x_a,x_v) \end{cases} \tag{15}$$

其中, 联合概率 $p(x_a,x_v)$表示音频和视觉特征的共同出现频率, 边缘概率$p(x_a)$和$p(x_v)$表示音频和视觉特征的独立出现概率.

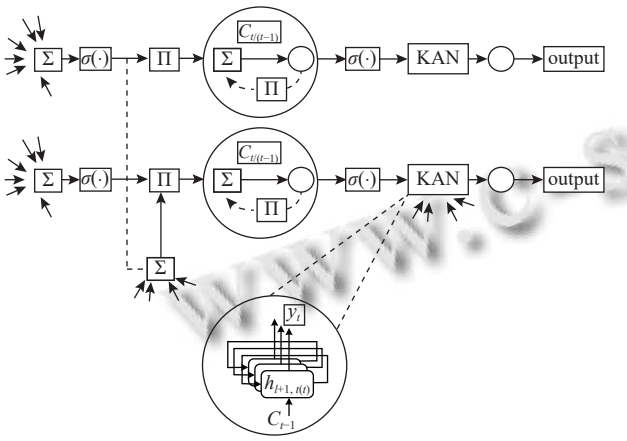

图 2 KLSTM 特征提取块

基于这些概率分布, 模态间互信息$I(X_a;X_v)$可定义为:

$$I(X_a;X_v)=\sum_{x_a\in X_a}\sum_{x_v\in X_v}p(x_a,x_v)\log\frac{p(x_a,x_v)}{p(x_a)p(x_v)} \tag{16}$$

其中, $\dfrac{p(x_a,x_v)}{p(x_a)p(x_v)}$度量音频和视觉特征的依赖性. 当该比值大于 1 时, 表示音频和视觉特征之间存在显著的关联, 互信息值越高, 表明两模态在情绪特征表达上越一致.

其次, 融合表示与单模态特征之间的互信息$I(Z;X_a)$和 $I(Z;X_v)$用于确保融合表示 Z 能够保留音频和视觉的情绪特征信息. 设 Z 为通过融合网络生成的视听表示, 其与音频特征的联合概率分布 $p(z,x_a)$和边缘概率分布$p(z)$与$p(x_a)$可表示为:

$$\begin{cases} p(z,x_a)=P(Z=z,X_a=x_a) \\ p(z)=\sum\limits_{x_a\in X_a}p(z,x_a),\quad p(x_a)=\sum\limits_{z\in Z}p(z,x_a) \end{cases} \tag{17}$$

联合概率$p(z,x_a)$表示融合表示和音频特征的共现频率, 边缘概率$p(z)$和$p(x_a)$表示融合表示和音频特征的独立出现概率.

由以上分布定义, 可将融合表示与音频的互信息$I(Z;X_a)$计算为:

$$I(Z;X_a)=\sum_{z\in Z}\sum_{x_a\in X_a}p(z,x_a)\log\frac{p(z,x_a)}{p(z)p(x_a)} \tag{18}$$

同理, 融合表示与视觉特征的互信息 $I(Z;X_v)$可定义为:

$$I(Z;X_v)=\sum_{z\in Z}\sum_{x_v\in X_v}p(z,x_v)\log\frac{p(z,x_v)}{p(z)p(x_v)} \tag{19}$$

其中, $\dfrac{p(z,x_a)}{p(z)p(x_a)}$ 和 $\dfrac{p(z,x_v)}{p(z)p(x_v)}$分别表示融合表示与音频、视觉特征的关联程度, 互信息值越大, 说明融合表示更好地保留了各单模态中的情绪信息. 通过最大化$I(X_a;X_v)$、$I(Z;X_a)$和$I(Z;X_v)$, 可以实现视听特征间的一致性和有效融合, 使得最终的融合表示能够全面且准确地反映音频和视觉特征中的情绪信息, 从而提升情绪识别的整体性能.

3 实验分析

3.1 数据集与评价指标

CMU-MOSI 和 CMU-MOSEI 是卡内基梅隆大学开发的多模态情绪识别基准数据集, 广泛用于多模态情绪和情感分析研究. CMU-MOSI 数据集包含 2 199 个视频片段, 提供了音频、视频和文本数据, 并标注了情感强度 ([−3, 3]) 用于识别积极和消极情绪. CMU-MOSEI 是 CMU-MOSI 的扩展版, 包含 23 500 个视频片段, 支持 6 种情绪类别和情感强度标注. 两个数据集为多模态情绪识别模型提供了丰富的音频、视觉和文本信息, 成为多模态情绪分析的重要测试基准.

模型采用 4 个评价指标, 包括: *MAE* (mean absolute error)、*Acc*-7 (七分类准确率)、*Acc*-2 (二分类准确率) 和 *F*1 (*F*1 分数), 从不同的角度评估模型的表现.

MAE 是一种用于回归任务的误差度量, 表示模型预测值和真实值之间的平均绝对误差.

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{20}$$

其中, y_i表示真实值, $\hat{y}_i$表示模型预测值, n是样本总数. 对于每个样本, 计算预测值和真实值的绝对差值, 再求其平均. 情绪识别任务中, *MAE* 可以衡量模型对情绪强度 (如[−3, 3]) 的预测准确性, *MAE* 值越小, 说明模型的预测误差越小.

Acc-7 是用于多分类情绪识别的准确率, 表示模型在 7 类情绪标签上的分类准确性.

$$Acc\text{-}7=\frac{\sum_{i=1}^{n}\mathbb{I}(y_i=\hat{y}_i)}{n} \tag{21}$$

其中, y_i为真实类别, $\hat{y}_i$为预测类别, $\mathbb{I}$为指示函数, 当 $y_i=\hat{y}_i$时取 1, 否则取 0, n为样本总数. *Acc*-7 值越高, 说明模型对多类别情绪的分类能力越强.

Acc-2 是二分类任务的准确率, 适用于识别情绪的积极/消极二分情况.

$$Acc\text{-}2=\frac{TP+TN}{TP+TN+FP+FN} \tag{22}$$

其中, *TP* 为真正类数, *TN* 为真负类数, *FP* 为假正类数, *FN* 为假负类数. *Acc*-2 值越高, 表示模型在情绪二分类任务中越准确.

*F*1 分数是一种综合精确率 (*Precision*) 和召回率 (*Recall*) 的分类评价指标, 特别适用于不平衡数据集.

$$F1=2\times\frac{Precision\times Recall}{Precision+Recall} \tag{23}$$

其中, $Precision=\frac{TP}{TP+FP}$, $Recall=\frac{TP}{TP+FN}$. 情绪识别中, *F*1 分数用于评估模型对特定情绪类别的分类性能, 尤其适用于正负类样本不平衡的情况. 高 *F*1 分数表示模型在保证高准确率的同时, 具有良好的检出能力.

3.2 对比实验

在 CMU-MOSI 和 CMU-MOSEI 数据集上进行对比实验.

在 CMU-MOSI 和 CMU-MOSEI 数据集上的实验结果表明, 本文提出的 MIKL 模型在视听情绪识别任务中表现优异. 如表 1 和表 2 所示, 相较于其他基准模型, MIKL 的 *MAE* 值显著降低, 并在 *Acc*-7、*Acc*-2 和 *F*1 分数上均优于其他模型性能, 体现了其在情绪识别任务中的强大性能和鲁棒性. 表 3 的模型复杂度对比实验显示, MIKL 的参数量和训练时间虽高于其他模型, 需要更多的计算资源, 但其性能的显著提升验证了模型设计的有效性.

这一优势主要源于本文的两大创新点: 首先, 通过 KLSTM 结合特征提取模块, 模型能够有效捕捉视听情绪信号中的复杂时间依赖性, 并增强对动态情绪特征的感知能力; 其次, 基于互信息的特征选择、自适应窗口和跨模态一致性最大化方法, 能够筛选出与情绪高度相关的特征片段, 并实现视听特征的深度融合. 尽管模型复杂度较高, 但这两项创新显著提升了 MIKL 在处理长时间序列和跨模态融合任务中的表现, 使其在不同数据集上均展现出卓越的泛化能力, 进一步验证了其在视听情绪识别领域的优越性. 图 3 是 MIKL 模型在 CMU-MOSEI 数据集上得到的 7 分类混淆矩阵.

表 1 在 CMU-MOSI 的对比实验

模型	*MAE*	*Acc*-7 (%)	*Acc*-2 (%)	*F*1 (%)
TFN[12]	0.901	34.9	80.8	80.7
LMF[13]	0.917	33.2	82.5	82.4
MFM[14]	0.877	39.0	83.0	83.0
MulT[15]	0.861	42.63	84.1	83.9
MISA[16]	0.804	43.26	82.10	82.03
LNLN[17]	0.713	45.79	85.98	85.95
MIKL	0.700	46.65	86.06	85.98

表 2 在 CMU-MOSEI 的对比实验

模型	*MAE*	*Acc*-7 (%)	*Acc*-2 (%)	*F*1 (%)
TFN	0.593	50.2	82.5	82.1
LMF	0.623	48.0	82.0	82.1
MFM	0.568	51.3	84.4	84.3
MulT	0.580	51.5	82.5	82.3
MISA	0.568	52, 67	85.2	85.1
LNLN	0.530	53.46	85.17	85.30
MIKL	0.526	54.24	85.97	85.94

表 3 在 CMU-MOSI 的模型复杂度对比实验

模型	参数量 (M)	每轮训练时间
TFN	2.43	27' 13"
LMF	1.26	18' 07"
MFM	2.14	22' 37"
MulT	2.57	32' 23"
MISA	3.10	37' 42"
LNLN	3.32	42' 25"
MIKL	3.58	47' 35"

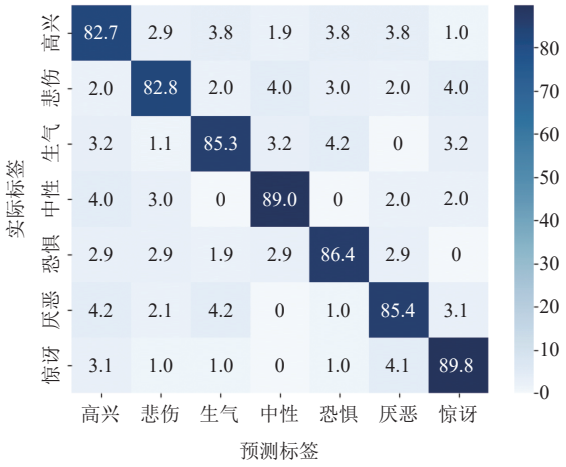

图 3 MIKL 在 CMU-MOSEI 得到的混淆矩阵

3.3 消融实验

在 CMU-MOSI 数据集上进行消融实验, 具体的消融实验设计如表 4.

表 4 在 CMU-MOSI 的消融实验

设置	*MAE*	*Acc*-7 (%)	*Acc*-2 (%)	*F*1 (%)
标准 LSTM	0.945	53.85	85.05	84.95
–互信息特征选择	0.836	53.53	85.39	85.47
–互信息自适应窗口	0.834	54.11	85.61	85.64
–跨模态一致性最大	0.733	53.4	84.94	84.95
MIKL	0.700	46.65	86.06	85.98

消融实验结果表明, 使用标准 LSTM 替代 KLSTM 后, *MAE* 增加至 0.945, *Acc*-7 和 *Acc*-2 也有所下降, 说明 KLSTM 的非线性特征提取能力和对时间依赖性的建模在情绪识别任务中起到了关键作用. 去除互信息特征选择模块后, *MAE* 降低至 0.836, *Acc*-7 和 *F*1 分数下降明显, 表明该模块能够有效筛选出与情绪高度相关的特征, 减少冗余信息干扰, 从而提高情绪识别的分类准确性. 去除互信息自适应窗口模块后的 *MAE* 为 0.834, 相较于完整模型略有下降, 但 *Acc*-7 和 *F*1 分数同样降低, 说明自适应窗口模块在选择情绪信息最强的时间片段上发挥了重要作用, 能够提升模型的精确度和鲁棒性. 去除跨模态一致性最大化模块后, *MAE* 也有所上升, *Acc*-7、*Acc*-2 和 *F*1 分数均下降, 表明该模块在视听特征融合中的有效性. 通过增强音频和视频特征之间的一致性, 跨模态一致性最大化模块确保了视听情绪识别的整体性能.

3.4 鲁棒性实验

在 CMU-MOSI 数据集上进行鲁棒性实验, 同时为扩展至更复杂或噪声更大的场景, 在数据集 IEMOCAP 下验证泛化能力. 鲁棒性实验通过对输入的音频和视频信息进行加权融合以考察模型在不同模态权重组合下的性能表现. 设定音频输入为a, 视频输入为$1-a$, 其中a的取值范围为 [0, 1]. 可以通过不同的a值观察模型的鲁棒性, 即在偏重不同模态信息时的情绪识别性能.

表 5 在 CMU-MOSI 数据集的鲁棒性实验结果表明, 模型在不同的音视频权重组合下表现出显著的性能差异. 当音频权重和视频权重均为 0.5 时, 模型的性能达到最佳, 说明音视频信息在情绪识别中能够互为补充, 综合提高了模型的识别能力和鲁棒性. 相比之下, 当仅依赖单一模态时, 模型性能有所下降, 表明音视频模态融合在情绪识别中的重要性.

表 5 在 CMU-MOSI 的鲁棒性实验

音频权重	视频权重	*MAE*	*Acc*-7 (%)	*Acc*-2 (%)	*F*1 (%)
0.0	1.0	0.820	50.5	83.0	82.5
0.2	0.8	0.780	52.5	84.5	84.0
0.4	0.6	0.750	53.5	85.2	84.8
0.5	0.5	0.730	54.0	85.5	85.0
0.6	0.4	0.770	52.8	84.8	84.3
0.8	0.2	0.800	51.5	83.8	83.2
1.0	0.0	0.840	49.8	82.5	82.0

表 6 在 IEMOCAP 数据集上鲁棒性实验结果表明, 当音频权重和视频权重均为 0.5 时, 模型在悲伤、中性、愤怒、兴奋情绪中取得最优性能, 快乐、沮丧中的性能也接近最优性能. 实验验证了模型的泛化能力, 同时表明情绪识别需针对具体类别优化模态权重以提升整体鲁棒性.

表 6 在 IEMOCAP 的鲁棒性实验 (%)

音频权重	视频权重	快乐		悲伤		中性		愤怒		兴奋		沮丧	
		Acc-2	*F*1	*Acc*-2	*F*1	*Acc*-2	*F*1	*Acc*-2	*F*1	*Acc*-2	*F*1	*Acc*-2	*F*1
0.0	1.0	**30.56**	35.63	56.73	62.90	57.55	53.00	59.41	59.24	52.84	58.85	65.88	59.41
0.2	0.8	25.00	30.38	55.92	62.41	52.86	52.39	61.76	59.83	55.52	60.25	71.13	60.69
0.4	0.6	22.22	29.91	58.78	64.57	**62.76**	57.38	64.71	63.04	58.86	63.42	67.19	60.81
0.5	0.5	25.69	33.18	**75.10**	**78.80**	58.59	**59.21**	**64.75**	**65.28**	**80.27**	**71.86**	61.15	58.91
0.6	0.4	28.56	**36.99**	55.15	63.91	59.71	51.27	61.15	60.36	53.47	60.46	67.11	61.21
0.8	0.2	23.34	31.45	56.77	64.14	53.76	53.04	62.96	61.47	56.48	61.88	**72.81**	**62.61**
1.0	0.0	23.03	30.47	59.25	65.25	64.47	59.12	63.02	62.87	59.85	65.12	68.57	59.06

4 结束语

本研究通过理论分析和实验验证, 提出了一种基于 KLSTM 特征提取和互信息优化的视听情绪识别模型. 实验结果表明, MIKL 模型在情绪识别的准确性和鲁棒性上均取得显著提升. 本文引入 KLSTM 网络, 结合互信息的特征选择、自适应窗口和跨模态一致性优化策略, 成功实现了对情绪信号中动态特征的高效提取与融合, 增强了模型对情绪信息的捕捉能力和跨模态信息的一致性. 未来的研究可以探索如何更有效地利用视听信息的互补性, 进一步提升模型在各种复杂

场景下的适应能力.

参考文献

1 黄贤盛. 基于生理信号情绪识别的手部康复人机交互设备研究 [硕士学位论文]. 杭州: 杭州电子科技大学, 2024.

2 万欣, 袁海曼. 在线健康社区用户情绪识别与分析. 医学信息学杂志, 2023, 44(12): 15–19, 39. [doi: 10.3969/j.issn.1673-6036.2023.12.003]

3 黎倩尔, 黄沛杰, 陈佳炜, 等. 基于互信息最大化和对比损失的多模态情绪识别模型. 中文信息学报, 2024, 38(7): 137–146. [doi: 10.3969/j.issn.1003-0077.2024.07.014]

4 Piho L, Tjahjadi T. A mutual information based adaptive windowing of informative EEG for emotion recognition. IEEE Transactions on Affective Computing, 2020, 11(4): 722–735. [doi: 10.1109/TAFFC.2018.2840973]

5 Hacine-Gharbi A, Ravier P. On the optimal number estimation of selected features using joint histogram based mutual information for speech emotion recognition. Journal of King Saud University—Computer and Information Sciences, 2021, 33(9): 1074–1083. [doi: 10.1016/j.jksuci.2019.07.008]

6 Wang ZM, Hu SY, Song H. Channel selection method for EEG emotion recognition using normalized mutual information. IEEE Access, 2019, 7: 143303–143311. [doi: 10.1109/ACCESS.2019.2944273]

7 Medsker LR, Jain L. Recurrent Neural Networks: Design and Applications. Boca Raton: CRC Press, 2001. 2.

8 Graves A. Long short-term memory. Supervised Sequence Labelling with Recurrent Neural Networks. Berlin Heidelberg: Springer, 2012. 37–45.

9 Liu ZM, Wang YX, Vaidya S, *et al*. KAN: Kolmogorov-Arnold networks. arXiv:2404.19756, 2024.

10 Shen S, Younes R. Reimagining linear probing: Kolmogorov-Arnold networks in transfer learning. arXiv:2409.07763, 2024.

11 Vo-thanh HS, Nguyen QV, Kim SH. KAN-based fusion of dual-domain for audio-driven facial landmarks generation. arXiv:2409.05330, 2024.

12 Chen MH, Wang S, Liang PP, *et al*. Multimodal sentiment analysis with word-level fusion and reinforcement learning. Proceedings of the 19th ACM International Conference on Multimodal Interaction. Glasgow: ACM, 2017. 163–171.

13 Liu Z, Shen Y, Lakshminarasimhan VB, *et al*. Efficient low-rank multimodal fusion with modality-specific factors. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne: ACL, 2018. 2247–2256.

14 Tsai YHH, Liang PP, Zadeh A, *et al*. Learning factorized multimodal representations. Proceedings of the 7th International Conference on Learning Representations. New Orleans: OpenReview.net, 2019.

15 Tsai YHH, Bai S, Liang PP, *et al*. Multimodal Transformer for unaligned multimodal language sequences. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 6558–6569.

16 Zuo HL, Liu R, Zhao JM, *et al*. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island: IEEE, 2023. 1–5.

17 Zhang HY, Wang WB, Yu TS. Towards robust multimodal sentiment analysis with incomplete data. Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, 2024.

(校对责编: 张重毅)