第 35 卷 第 7 期
2025 年 7 月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol. 35 No. 7
Jul. 2025

多尺度 KAN 卷积与跨模态注意力的视听情绪识别

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

(1. 西南科技大学 信息工程学院,四川 绵阳 621000;
2. 四川省工业自主可控人工智能工程技术研究中心,四川 绵阳 621000;
3. 四川湖山电器股份有限公司,四川 绵阳 621000)

摘 要:针对现有视听情绪识别方法在特征提取层级别的模态互补性研究不足以及传统的视听情绪识别方法通常难以充分挖掘音频与视频模态之间的互补性等问题,提出了一种基于 Kolmogorov-Arnold Networks(KAN)卷积、多尺度特征提取和跨模态注意力机制的情绪识别模型。该模型在音视频特征提取过程中引入 KAN 卷积,通过多尺度卷积核捕捉不同层次的情绪特征,KAN 卷积通过可学习的 B 样条函数建模数据中的非线性模式,从而增强了模型对复杂情绪模式的学习能力。为了提升模态间信息的互补性,特征融合阶段采用了跨模态注意力机制。能够有效地对音视频特征进行加权融合,使得模型能够更好地捕捉音视频模态的交互关系,从而提升情绪识别的性能。在 RAVDESS 数据集上的实验结果表明,该模型的准确率和 F_1 值分别达到了 75.62% 和 77.23%,相较于传统方法取得了显著提升。研究表明,该模型在多模态情绪识别任务中表现出更强的鲁棒性和适应性,为视听情绪识别应用提供了新的有效方案。

关键词:情绪识别;视听融合;跨模态注意力;KAN 卷积;多尺度提取

中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2025)07-0100-08

doi:10.20165/j.cnki.ISSN1673-629X.2025.0043

Multiscale KAN Convolution with Cross-modal Attention for Audiovisual Emotion Recognition

LUO Zhi-xin[1],LIU Zhi-gui[1,2],TANG Rong[1],PAN Zhi-xiang[3],LI Li[1,2]

(1. School of Information Engineering,Southwest University of Science and Technology, Mianyang 621000,China;
2. Sichuan Engineering and Technology Research Center for Industrial Autonomous and Controllable Artificial Intelligence,Mianyang 621000,China;
3. Sichuan Hushan Electrical Appliance Company Limited,Mianyang 621000,China)

Abstract:To address the problems of insufficient research on modality complementarity at the feature extraction level in existing audio-visual emotion recognition methods,as well as the traditional methods´ inability to fully exploit the complementary relationship between the audio and video modalities,we propose an emotion recognition model based on Kolmogorov-Arnold Networks (KAN) convolution, multi-scale feature extraction, and cross-modal attention mechanisms. The model introduces KAN convolution during the audio and video feature extraction process,capturing emotion-related features at different levels through multi-scale convolution kernels. KAN convolution models the nonlinear patterns in the data using learnable B-spline functions,thereby enhancing the model´s ability to learn complex emotional patterns. To improve the complementarity of information between modalities,a cross-modal attention mechanism is adopted during the feature fusion stage. This mechanism effectively weights and fuses the audio and video features,allowing the model to better capture the interactive relationships between the two modalities,thus enhancing emotion recognition performance. Experiments on the RAVDESS dataset show that the proposed model achieves the accuracy of 75.62% and the F_1 score of 77.23%,significantly outperforming traditional methods. The study demonstrates that the proposed model exhibits stronger robustness and adaptability in multi-modal emotion recognition tasks,providing a new and effective solution for audio-visual emotion recognition applications.

Key words:emotion recognition;audiovisual fusion;cross-modal attention;KAN convolution;multiscale extraction

---

收稿日期:2024-11-07 **修回日期**:2025-03-10

基金项目:国家自然科学基金(U21A20157)

作者简介:罗志鑫(1997-),男,硕士研究生,通讯作者,研究方向为情绪识别与多模态融合;刘知贵(1966-),男,博士,教授,研究方向为先进控制与自动化。

0 引 言

情绪识别因其广泛的应用而吸引着越来越多的研究关注,例如情绪计算[1]、人机交互[2]和社交机器人[3]。在多模态情绪识别中,利用了对人类交流至关重要的音频和视频的模态信息,视听情绪识别尤为重要。与单模态情绪识别不同,多模态情绪识别可以利用不同模态对同一情绪的不同表征来识别情绪。通过高特征表示能力和可区分性,提高了识别准确率。

在多模态情绪识别的研究中,可以按照模态信息融合方法划分为:早期融合[4]、晚期融合[5]和模型融合[6]。早期融合是将多种模态数据提取并构建成相应的模态特征,再拼接成一个整合了各模态特征的特征集。晚期融合是找出每个模态的可信度,然后进行协调并做出联合决策。随着自然语言处理和计算机视觉任务的发展,模型融合通常采用注意力机制来实现信息交互,由于融合位置的灵活性,性能得到显著提高。在音视频情绪识别特征提取方面,Hossain 等人[7]通过卷积神经网络[8](Convolutional Neural Networks, CNN)来分别提取语音信号和视频片段中提取的关键帧,使用两个连续的极端学习机(ELM)对两个 CNN 的输出进行融合后进行情绪识别。Zhou 等人[9]探讨了音视频信息融合的方法,自注意力、关系注意力和 Transformer 注意力三种融合策略,主要用于突显重要的情绪特征。然而传统的卷积神经网络虽然在特征提取方面取得了成功,但在处理复杂非线性关系和长距离依赖时仍存在局限。CNN 的固定激活函数和线性权重矩阵限制了其对复杂语义信息的表达能力,尤其是在融合音频和视频等多模态数据时,难以有效捕捉模态间的非线性交互与关联。

针对 CNN 的这些问题,KAN(Kolmogorov Ar-nold Networks)实现了一种更加高效的特征提取与转换方式。KAN 不仅能够帮助捕捉复杂的非线性关系,还能生成更加结构化且可解释的输入数据表示,从而更有效地表达音频和视频数据中的模态间交互与关联特征。Shen 等人[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,实现图像分类任务中准确性和泛化能力的显著提升。Hoang 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。

尽管在视听情绪识别特征提取方面已有一定进展,但在视听情绪识别的模态融合中依然面临许多挑战。虽然现有的视听情绪融合研究致力于挖掘和促进不同模态信息之间的互补性,但在充分利用这些互补信息的效率和精度上仍有很大的提升空间,尤其是在捕捉复杂情绪特征时。互补信息能够显著提升融合模块的性能,但许多现有方法依然存在冗余问题。一些模型只有在拼接后才能输出最终结果,而拼接后的特征常常包含重复的表示,因此在拼接前对特征信息进行过滤是减少冗余的一个重要步骤。此外,当前的方法在特征学习过程中难以保证信息的完整性,常常在模内和模间信息学习时丢失部分语义信息,影响最终表现。

以上研究发现,基于 KAN 的音频建模能够更好地捕捉复杂的情感特征,在图像处理中,KAN 也展现出了处理面部特征的优势。此外,视听跨模态注意力机在此前的研究中也相对较少。基于以上研究,该文提出了一种多尺度 KAN 卷积与跨模态注意力的视听情绪识别模型(Multiscale KAN Convolution with Cross-Modal Attention for Audiovisual Emotion Recognition Models, MKCC)。目标是解决现有的视听情绪识别方法的局限性,建立一个端到端的模型,并在中间层进行融合。该文的贡献可归纳如下:

(1)多尺度 KAN 卷积特征提取模块。为了增强模型对多模态情绪特征的提取能力,引入了基于多尺度 KAN 卷积的特征提取方法。与传统的卷积层相比,多尺度 KAN 卷积通过学习非线性函数,能够更高效地捕捉跨模态特征之间的复杂关系,并通过多尺度卷积核在不同尺度上提取丰富的情绪特征信息。

(2)视听跨模态注意力机制模块。为了实现音频和视频特征的有效融合,引入了跨模态注意力机制,确保不同模态信息之间的交互性和互补性。通过该机制,音频和视频特征得以充分融合,最终提升了识别精度。

1 相关工作

随着深度学习的不断发展,使用机器学习和深度学习算法进行情绪识别逐渐成为研究热点之一。Liam 等人[12]使用 CNN 处理视觉线索,并与音频信息结合,经过微调的 FaceNet 模型来学习面部情感识别特征,用于进一步的情感识别。Praveen 等人[13]将 CNN 用于音频和视觉两个分支。视觉部分使用 3D-CNN 来捕捉视频序列中的时空特征。音频部分使用 CNN 处理频谱图,从中提取与讲话者情感状态相关的特征。Cai 等人[14]将 CNN 用于从视频的面部表情帧中提取特征,识别出面部图像中代表情绪识别的空间模式。CNN 的核心是多层感知机(Multilayer Perceptron, MLP),已经被证明在从简单回归到复杂图像分类的任务中是有效的。MLP 以其近似大范围函数的能力而闻名,这一性质得到了普遍近似定理的证实[15]。然而,深度学习架构是不断发展的,这是由对

改进性能、可解释性和效率的不断渴望所驱动的。最近推出的 KAN[16] 是挑战 MLP 主导地位的新兴范例之一。KAN 从 Kolmogorov-Arnold 表示定理中获得灵感,该定理断言任何连续的多元函数都可以表示为单变量函数的组合。这个定理是 KANs 结构的基础,其中激活函数不是在神经元节点上训练和应用,而是直接在网络图的边缘上应用。这改变了网络内信息流的动态。与传统 MLP 相比,KAN 增强了模型容量,能更好地处理数据中的复杂依赖关系,以及改进的学习表征的可解释性。通过将激活函数分散到边缘,KAN 促进了一种更模块化的特征提取和转换方法,可能产生更结构化和可解释的输入数据表示。

把 KAN 引入卷积层中,增强了网络处理空间信息的能力。在 MNIST 和 Fashion MNIST 数据集上的实验结果表明,与小型 CNN 相比,KAN 卷积模型显示出更高的准确性,而性能略低于中型 CNN。这表明 KAN 卷积在图像识别任务中的竞争优势,同时通过减少对完全连接层的需求来保持低参数复杂性[17]。在这些进步的推动下,该文探讨了 KAN 在情绪识别任务中的应用。情绪识别任务是通过神经网络架构进行的无监督学习,其任务是通过编码器-解码器框架学习输入数据的有效情绪表示。传统的卷积神经网络(CNN)在涉及图像数据的任务中占据主导地位,因为它们能够捕获空间层次和平移不变性[18]。与 CNN 相比,基于 KAN 卷积利用基于边缘的激活来潜在地捕获图像中更细微的关系和依赖关系。这为如何构建和优化神经网络以完成情绪识别任务提供了一种新颖的视角。该文旨在探讨 KAN 卷积是否可以在视听情绪识别任务上取得良好的性能。

在视听情绪识别中利用注意力机制优化视听融合情绪识别模型已成为近年来情绪识别研究的一个热点方向。情绪识别注意力机制是一种能够对不同模态的数据进行对齐和加权的机制。其核心思想是通过自适应地选择每个模态中的关键特征,强调与任务相关的特征,同时弱化不重要的信息,从而有效融合情绪识别特征。对于情绪识别任务,情绪识别注意力机制可以同时关注语音、文本和视觉信息,动态调整这些模态的信息权重,捕捉更细腻的情绪线索。

2018 年,Huang 等人[19] 利用交叉注意力模块在其交叉路径上收集所有像素的上下文信息。通过进一步的循环操作,每个像素最终都可以捕获完整的图像依赖关系。2023 年,Liu 等人[20] 使用交叉注意力模块来融合不同视图之间的信息。Yang 等人[21] 使用两个编码器分别计算 3D 点云和 2D 多视图图像的自参与特征。解码器实现交错的 2D-3D 交叉注意力,并进行隐式 2D 和 3D 特征融合。2024 年,Jian 等人[22] 提出一个差异信息注入模块(Discrepancy Information Injection Module,DIIM)和两个替代共同信息注入模块(Alternate Common Information Injection Modules,ACIIM)。DIIM 通过修改基本的交叉注意力机制设计,可以促进源图像差异信息的提取。同时,ACIIM 通过交替使用基本的交叉注意力机制设计,能够充分挖掘共同信息并整合长期依赖关系。

鉴于 KAN 卷积与注意力机制在情绪识别中的诸多优秀表现,该文针对视听融合情绪识别任务,在利用深度神经网络提取视频与音频特征的基础上,引入 KAN 卷积与跨模态注意力机制对模型进行改进,使得网络模型能够更有效地关注最重要且与情绪识别最相关的多模态信息,从而提高情绪识别的准确性。

2 模型构建

该文提出的模型主要分为三个模块:特征提取模块、多尺度 KAN 卷积特征提取模块和跨模态融合模块。特征提取模块用于从视觉和音频输入中提取基础特征;多尺度 KAN 卷积特征提取模块则负责在各模态内部对特征进行进一步的卷积和池化处理,并实现模态间的初步信息交互;融合模块用于在跨模态块中进行深度的特征融合,并最终通过分类器实现输出分类。模型的总体框架如图 1 所示。

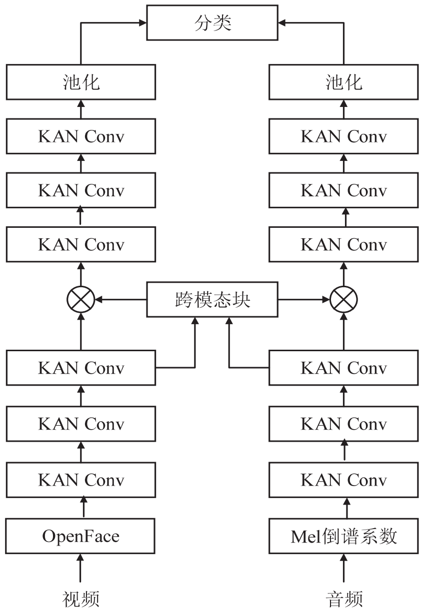

图 1 MKCC 模型总体架构

模型的情绪识别过程如下:首先,视觉输入通过 OpenFace[23] 进行特征提取,音频输入提取 Mel 倒谱系数后进行特征提取。这些特征分别通过多尺度 KAN 卷积进行处理,并在跨模态块中实现第一次多模态信息的交互和融合,增强各模态的表示能力。其次,各模态的增强特征经过池化层进行降维,减少冗余信息,并通过进一步的卷积操作提取深度特征。最后,将融合

后的特征输入到分类器中,完成情绪识别任务的分类输出。

2.1　特征提取模块

第一部分是单个视频帧的视觉特征提取和音频特征提取,然后是整个视频序列的联合表示学习。为了实现能够从原始视频中学习端到端可训练模型,该文将特征提取作为模型的一部分,并与视听融合模块一起对其进行优化。该文使用了 OpenFace 工具包从视频帧中提取动作单元(Action Unit,AU)。这个 API 提供了 18 个不同动作单元的存在和强度检测。如果 AU 显示,则 AU 的存在是二进制编码为 1,否则为 0。相比之下,强度是一个连续的变量,范围从 0 到 5。这两种预测都来自两个独立的网络,它们遵循了相同的预处理阶段。首先,对人脸进行对齐,以计算几何和基于外观的特征;其次,将特征送到两个不相连的支持向量机模型,分别返回 AU 在每一帧中的存在和唤醒。默认情况下,视频级别的特征在假设许多帧代表中性情绪的情况下被归一化。该文也应用了这种默认配置。对于音频特征提取部分,该文主要使用梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficients,MFCCs)作为特征,预加重滤波器的权重取 0.95,分帧的帧长和帧移设置为 256 和 128 个样本,使用的窗函数是汉明窗,Mel 频率的转换公式如下:

$$m = 2\ 595\log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

其中,f 是线性频率,m 是 Mel 频率。对每个滤波器的输出取对数,得到对数能量谱。再将对数能量谱进行离散余弦变换,将其转换为倒谱系数。

2.2　多尺度 KAN 卷积特征提取层

文中模型使用多尺度 KAN 卷积进行特征处理。多尺度 KAN 卷积在视觉和音频两种模态上独立应用,能够处理复杂的模态特征,并通过层层递进的卷积网络逐步提取出深层的情绪相关信息。多尺度 KAN 卷积相比于 CNN 的优势在于其使用可学习的非线性函数代替固定的激活函数和线性权重矩阵,能够更灵活地捕捉复杂的非线性关系。这种特性使得多尺度 KAN 卷积在保持相似准确度的情况下,显著减少了参数量,提升了模型的计算效率和泛化性能。

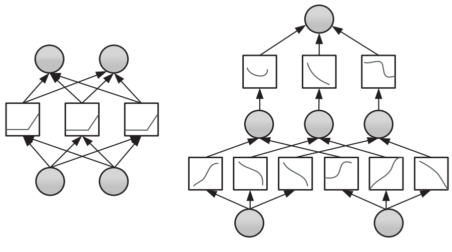

图 2　KAN 与 MLP 的区别

KAN 的核心在于其独特的架构,与传统的在节点上使用固定激活函数的 MLP 不同,如图 2 所示。KAN 在网络边缘实现可学习的激活函数。这种从静态节点函数到动态节点函数的关键转变涉及到用自适应样条函数取代传统的线性权重矩阵,这些函数在训练过程中被参数化和优化。这允许更灵活和响应更快的模型架构,可以动态地适应复杂的数据模式。

Kolmogorov-Arnold 表示定理假定一个多元函数 $f(x_1,x_2,\cdots,x_n)$ 可以表示为:

$$f(x_1,x_2,\cdots,x_n) = \sum_{q=1}^{2n+1} \Phi_q\left(\sum_{p=1}^{n} \varphi_{q,p}(x_p)\right) \tag{2}$$

式中,$\varphi_{q,p}$ 是映射每个输入变量 (x_p) 的单变量函数,例如 $\varphi_{q,p}:[0,1] \to R$ 和 $\Phi_q:R \to R$ 。

KAN 将每层结构表示为这些可学习的 1D 函数的矩阵。

$$\Phi = \{\varphi_{q,p}\},p = 1,2,\cdots,n_{\text{in}},q = 1,2,\cdots,n_{\text{out}} \tag{3}$$

特别地,每个函数 $\varphi_{q,p}$ 都可以定义为 b 样条,这是一种由基样条的线性组合定义的样条函数,增强了网络学习复杂数据表示的能力。这里 n_{in} 表示特定层的输入特征的数量,而 n_{out} 表示该层产生的输出特征的数量,反映了网络层之间的维数转换。该矩阵中的激活函数 $\varphi_{l,j,i}$ 为可学习样条函数,表示为:

$$f_{\text{spline}}(x) = \sum c_i B_i(x) \tag{4}$$

式中,$f_{\text{spline}}(x)$ 为可学习样条函数,c_i 为可训练系数。

公式 4 中允许每个 $\varphi_{l,j,i}$ 根据数据调整其形状,为网络如何模拟输入之间的相互作用提供了前所未有的灵活性。KAN 的整体结构类似于 MLP 中的堆叠层,但增强了使用复杂的函数映射而不是简单的线性变换和非线性激活。

$$f_{\text{KAN}}(x) = (\Phi_{L-1} \circ \Phi_{L-2} \circ \cdots \circ \Phi_0)(x) \tag{5}$$

式中,$f_{\text{KAN}}(x)$ 表示 KAN。

每一层的变换 Φ_l 作用于输入 x_l 产生下一层的输入 x_{l+1} ,描述为:

$$x_{l+1} = \Phi_l(x_l) = \begin{pmatrix} \varphi_{l,1,1}(\cdot) & \cdots & \varphi_{l,1,n_l}(\cdot) \\ \vdots & \ddots & \vdots \\ \varphi_{l,n_{l+1},1}(\cdot) & \cdots & \varphi_{l,n_{l+1},n_l}(\cdot) \end{pmatrix} x_l \tag{6}$$

式中,每个激活函数 $\varphi_{l,j,i}$ 是一个样条,为模型输入提供了一个丰富的、可适应的响应面。

在深度学习中使用 KAN 结构的动机是其在边缘上具有可学习的激活函数,增强了它们的表达能力和效率。通过用样条函数代替线性权矩阵,KAN 减少了实现高精度所需的参数数量,从而实现了更快的收敛和更好的泛化。卷积 Kolmogorov-Arnold 网络类似于 CNN。不同之处在于,卷积层被 KAN 卷积层取代,在

平坦化之后,可以有 KAN 或 MLP。与其他架构相比,卷积 KAN 的主要优点是它需要的参数要少得多。这是由该网络的构造给出的,因为 b 样条能够平滑地表示任意激活函数。

在计算机视觉中,卷积通常与卷积神经网络中的数学运算交替使用。该操作包括在输入中传递一个核或过滤器,并计算每个位置的点积。在 KAN 卷积中,主要思想是利用 Kolmogorov-Arnold 网络的方法提出这种数学运算的替代实现。KAN 卷积与 CNN 中使用的卷积的主要区别在于内核。在 CNN 中,它是由权重组成的,而在 KAN 中,内核的每个元素 φ 是一个可学习的非线性函数,利用 b 样条。形式上,每个元素被定义为:

$$f = w_1 \times f_{\text{spline}}(x) + w_2 \times f_{\text{silu}}(x) \tag{7}$$

式中,$f_{\text{silu}}(x)$ 表示 silu 函数。

在 KAN 卷积中,内核在图像上滑动,并将相应的激活函数 φ_{ij} 应用于相应的像素 a_{kl} 并将输出像素计算为 $\varphi_{ij}(a_{kl})$ 的和。设 K 为 KAN 核 $\in R^{N \times M}$ 图像为矩阵:

$$\boldsymbol{M}_{\text{Image}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mp} \end{bmatrix} \tag{8}$$

式中,$\boldsymbol{M}_{\text{Image}}$ 表示图像矩阵。

那么,KAN 卷积的定义如下:

$$(\boldsymbol{M}_{\text{Image}} * K)_{i,j} = \sum_{k=1}^{N} \sum_{l=1}^{M} \varphi_{kl}(a_{i+k,j+l}) \tag{9}$$

式 9 使得 KAN 卷积在视听情绪识别任务中能够有效地提取输入数据中的情绪特征。

KAN 卷积能够在单一尺度上捕捉数据的复杂特征,但单一尺度的特征提取往往无法全面表达情绪的多样性与复杂性。情绪表达通常包含丰富的细节和整体的特征,在不同的尺度上表现出不同的特征。因此,仅使用单一尺度的卷积核可能会导致信息的缺失,无法充分表达情绪特征。为了解决这一问题,在 KAN 卷积的基础上引入多尺度特征提取模块。通过在特征提取过程中应用不同尺寸的卷积核,可以在局部和全局范围内捕捉情绪特征,从而使模型具有更全面的特征表达能力。为了在不同尺度上捕捉情绪特征,该文在 KAN 卷积的基础上增加了多尺度特征提取。具体而言,在特征提取过程中使用不同尺寸的卷积核(如 3×3、5×5 和 7×7)对输入特征进行处理,从而在细节和整体层面上获取情绪相关的特征信息。

设输入特征图为 $\boldsymbol{X} \in R^{H \times W \times C_{in}}$,不同卷积核大小的 KAN 卷积操作分别记为 O_{KAN_3}、O_{KAN_5}、O_{KAN_7},其中下标表示卷积核的大小。多尺度特征提取的计算公式为:

$$Y = O_{\text{Concat}}(O_{\text{KAN}_3}(X), O_{\text{KAN}_5}(X), O_{\text{KAN}_7}(X)) \tag{10}$$

其中,O_{Concat} 表示在通道维度上的特征级联操作。具体地,$O_{\text{KAN}_3}(X)$ 使用 3×3 的卷积核提取局部细节特征,$O_{\text{KAN}_5}(X)$ 使用 5×5 的卷积核覆盖更大的区域提取中尺度特征,而 $O_{\text{KAN}_7}(X)$ 利用 7×7 的卷积核获取全局特征。通过该多尺度 KAN 卷积特征提取模块,模型能够在不同尺度上充分表达情绪特征,增强了情绪识别的准确性。

无论是视频还是音频都由 KAN 卷积块组成。每个块由 KAN 卷积层,批量归一化和 MaxPooling 组成。在 KAN 卷积层中有独特的样条卷积,其中重要的参数是样条函数的阶数(order, o)和样条卷积网格的大小(grid size, g)。其他一般参数是输出通道数(output dim, d)、卷积核(kernel, k)、步幅(stride, s)。在文中模型中的视频分支和音频分支的架构见表 1 和表 2。

表 1 视频分支的架构

	视频
第一部分	KAN1D [o =3, g =5, k =3, d =64, s=1]+BN1D KAN1D [o =3, g =5, k =5, d =64, s =1]+BN1D KAN1D [o =3, g =5, k =7, d =64, s =1]+BN1D
第二部分	KAN1D [o =3, g =5, k =3, d =128, s =1]+BN1D KAN1D [o =3, g =5, k =5, d =128, s =1]+BN1D KAN1D [o =3, g =5, k =7, d =128, s =1]+BN1D
预测	全局平均池化+线性

表 2 音频分支的架构

	音频
第一部分	KAN1D [o =3, g =5, k =5, d =64, s =1] + BN1D KAN1D [o =3, g =5, k =3, d =64, s =1] + BN1D KAN1D [o =3, g =5, k =7, d =64, s =1] + BN1D
第二部分	KAN1D [o =3, g =5, k =3, d =128, s =1] + BN1D KAN1D [o =3, g =5, k =5, d =128, s =1] + BN1D KAN1D [o =3, g =5, k =7, d =128, s =1] + BN1D
预测	全局平均池化+线性

2.3 特征融合层

MKCC 视听融合情绪识别模型不仅考虑了视频和音频内部的特征信息,还考虑了视听模态之间的跨模态关联,实现了多模态情绪特征的有效挖掘和融合。

为进一步捕获与情绪识别相关的关键跨模态信息,该文在前述模型的基础上引入了跨模态注意力机制,对其进行改进,提出了融合跨模态注意力机制的 MKCC 模型。该文使用的跨模态注意力机制的结构如图 3 所示。现有模态融合方法大多基于特征拼接或简单权重加和,通常忽略了不同模态之间的交互关系以及模态特征的重要性差异,容易导致信息冗余或特征弱化。跨模态注意力机制通过动态计算音视频模态之间的相关性权重,精确建模音频和视频模态间的交

互特性,从而有效捕捉关键的情绪信息,实现更精细的模态融合。

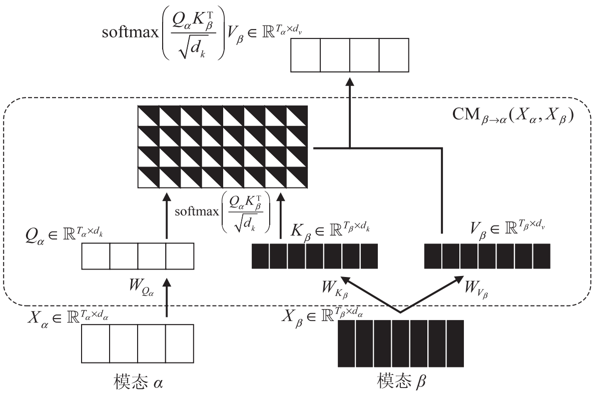

图 3　跨模态注意力机制的结构

在该跨模态注意力机制中考虑两种模态音频 α 和视频 β,每个模态的两个序列分别记为 $X_\alpha \in R^{T_\alpha \times d_\alpha}$ 和 $X_\beta \in R^{T_\beta \times d_\beta}$,分别表示序列长度和特征维度。在其余部分,$T_{(\cdot)}$ 和 $d_{(\cdot)}$ 分别代表序列长度和特征维度。假设一种融合跨模态信息的好方法是提供跨模态的潜在适应,即将 β 转换为 α。

跨模态注意力机制包括以下三个模块,输入特征编码将音频和视频模态的原始特征分别编码为固定维度的向量表示 X_a 和 X_β。注意力计算通过查询-键值机制(Q_a,K_β,V_β),计算音频模态对视频模态的注意力分布(即权重矩阵),突出高相关性的模态特征。根据注意力权重进行加权融合,将视频模态特征 V_β 的高相关性部分融入音频模态中,生成最终的融合特征 Y_a。将查询定义为 $Q_\alpha = X_\alpha W_{Q_\alpha}$,键定义为 $K_\beta = X_\beta W_{K_\beta}$,值定义为 $V_\beta = X_\beta W_{V_\beta}$,其中 $W_{Q_\alpha} \in \mathrm{R}^{d_\alpha \times d_k}$,$W_{K_\beta} \in R^{d_\beta \times d_k}$ 和 $W_{V_\beta} \in R^{d_\beta \times d_v}$ 是权重。β 到 α 潜在适应性表示为跨模态注意力 $Y_\alpha = \mathrm{CM}_{\beta\to\alpha}(X_\alpha, X_B) \in R^{T_\alpha \times d_v}$。

$$Y_\alpha = \mathrm{CM}_{\beta\to\alpha}(X_\alpha, X_\beta) = f_{\mathrm{softmax}}(\frac{Q_\alpha K_\beta^{\mathrm{T}}}{\sqrt{d_k}})V_\beta =$$

$$f_{\mathrm{softmax}}(\frac{X_\alpha W_{Q_\alpha} W_{K_\beta}^{\mathrm{T}} X_\beta^{\mathrm{T}}}{\sqrt{d_k}})X_\beta W_{V_\beta} \tag{11}$$

式中,Y_α 与 Q_α 具有相同的长度,并且在 V_β 的特征空间中表示。因此,方程中缩放(由 $\sqrt{d_k}$ 缩放)的 softmax 计算得分矩阵 $f_{\mathrm{softmax}}(\cdot) \in R^{T_\alpha \times T_\beta}$。因此,第 i 个时间步的 Y_α 是 V_β 的加权摘要,其权重由 $f_{\mathrm{softmax}}(\cdot)$ 的第 i 行决定,将该方程称为单头跨模态注意力。

跨模态注意力机制在设计上具有新颖性,不同于现有方法依赖于固定加权规则或简单点积公式,该文创新性地引入了模态间交互映射矩阵,通过学习到的权重自适应调整模态间相关性,动态捕捉关键特征。使用缩放因子 $\sqrt{d_k}$ 提高了数值稳定性,进一步优化了权重分布的计算,使得机制能够更高效地捕捉跨模态信息。

3　实验验证

对提出的多尺度 KAN 卷积与跨模态注意力的视听情绪识别模型在公开数据集 RAVDESS 上进行了实验,并采用精确率(Accuracy, V_{Accuracy})和 F_1 值作为评估指标评价了该方法的性能。

3.1　实验设置

RAVDESS 数据集中每个演员录制 6 个视频序列,将它们裁剪或补零到 3.6 秒,这是平均序列长度。对于音频处理,提取了 10 个 Mel 频率倒谱系数进行进一步处理。对于视觉数据,从 3.6 秒的视频中选择 15 个均匀分布的帧,并使用人脸检测算法来裁剪演员的脸。图像大小调整为 224×224 像素。该文在原始的 15 帧视频上训练模型。该文将在 AffectNet[24] 数据集上预先训练 OpenFace 的权重转移。将数据分为训练集、验证集和测试集,确保参与者的身份不会跨集重复。具体地说,使用了 4 个参与者进行测试,4 个参与者进行验证,16 个参与者进行培训,并报告了平均超过 5 次的结果。视频被缩放到[0,1]尺度,并使用随机水平翻转和随机旋转进行数据增强。所有模型都使用 SGD 进行 100 个 epochs 的训练,学习率为 0.04,动量为 0.9,权重衰减为 1e-3。

3.2　评估指标

该文采用准确率(Accuracy, V_{Accuracy})和 F_1 值作为评估指标评价了 MKCC 方法的性能。计算公式如下:

$$V_{\mathrm{Accuracy}} = \frac{N_{\mathrm{TP}} + N_{\mathrm{TN}}}{N_{\mathrm{TP}} + N_{\mathrm{TN}} + N_{\mathrm{FP}} + N_{\mathrm{FN}}} \tag{12}$$

$$F_1 = \frac{2 \times (P \times R)}{P + R} \tag{13}$$

式中,N_{TP}、N_{TN}、N_{FP}、N_{FN} 分别为 TP、TN、FP、FN 的数量,P 表示精确率,R 表示召回率。

3.3　对比实验

在视听情绪识别任务上,对提出的 MKCC 模型进行了多种对比实验以评估其性能。

表 3　RAVDESS 数据集对比实验　%

模型	Accuracy	F_1
3DRexNent50	62.99	63.02
1D CNN	56.53	58.34
Averaging	68.82	68.57
Concat + FC[25]	71.04	72.56
MMTM[26]	73.12	73.37
MSAF[27]	74.86	76.95
MERC[28]	75.57	76.89
MKCC	75.62	77.23

基于表 3 可以得出,文中网络模型相比传统的神

经网络 3D RexNent50 和 1D CNN 准确率分别提高了 12.63 百分点和 19.09 百分点,与早期融合方法相比,文中方法的准确率提高了 4.58 百分点,与晚期融合方法相比提高了 6.80 百分点,与 MMTM 相比提高了 2.50 百分点,与 MSFA 相比提高了 0.76 百分点,与 MERC 相比提高了 0.05 百分点。F_1 值相比其他方法也有所提高。

实验证明,多尺度 KAN 卷积通过增强对视听模态中时空特征的提取能力,有效捕捉了音频与视觉之间更为细致的交互特征。在音频方面,多尺度 KAN 卷积能够精准提取语音中的情感和节奏变化,帮助模型更好地理解语音驱动的面部表情变化;在视觉方面,它自适应地聚焦于不同尺度的面部特征,尤其是与声音相关的局部区域,从而提升了整体的识别精度。

此外,跨模态注意力机制能够有效处理情绪表达在视听模态中的时间异步性问题。情绪在面部表情上的体现通常只持续一小段时间,而语音的情感表达则较为持续。通过跨模态注意力机制,模型能够在多尺度特征中提取那些时间不完全同步但情感特征一致的信息,从而更精准地捕捉情绪的整体表达。

3.4 消融实验

MKCC 是对传统基于 CNN 网络的改进。为了验证改进模块对模型性能的影响,设计消融实验比较改进模块之后的效果,如表 4 所示。

表 4 在 RAVDESS 数据集上的消融实验结果

模型	卷积核组合	Accuracy/%	F_1 /%
CNN		56.53	58.34
KAN	--	72.21	74.24
KAN(单尺度)+Attention	3×3	72.35	74.40
KAN(单尺度)+Attention	5×5	72.40	74.50
KAN(单尺度)+Attention	7×7	72.45	74.60
KAN(双尺度)+Attention	3×3, 5×5	74.10	75.80
KAN(双尺度)+Attention	3×3, 7×7	74.20	75.90
KAN(双尺度)+Attention	5×5, 7×7	74.25	76.00
KAN(多尺度)+Attention	3×3,5×5, 7×7	75.62	77.23

从表 4 实验结果可以看出,该文提出的基于多尺度 KAN 卷积和跨模态注意力机制的情绪识别模型在 RAVDESS 数据集上取得了显著的性能提升。具体而言,相较于传统 CNN 模型,加入 KAN 卷积后模型的准确率从 56.53% 提升至 72.21%, F_1 值从 58.34% 提升至 74.24%,表明 KAN 卷积在非线性特征提取方面的优势。进一步引入多尺度特征提取后,不同卷积核组合有效增强了模型对情绪特征的捕捉能力,其中 3×3、5×5 和 7×7 卷积核的组合取得了最佳效果,使准确率和 F_1 值分别达到 75.62% 和 77.23%。此外,跨模态注意力机制显著提升了音视频模态间的交互效果,进一步提高了特征融合的有效性。这些结果验证了多尺度 KAN 卷积结合跨模态注意力机制在多模态情绪识别中的有效性。

3.5 鲁棒性实验

在鲁棒性实验中,通过对音频和视频特征赋予不同权重组合,观察模型在不同模态信息不平衡情况下的表现,以分析其在视听情绪识别任务中的适应性和稳定性。在输入层分别对音频和视频特征赋予权重,设音频权重为 a,视频权重为 $1-a$。输入特征定义如下:

$$x = a \times y + (1 - a) \times z \tag{14}$$

式中,x 表示输入特征,y 表示音频特征,z 表示视频特征。权重 a 的取值范围设置为 0、0.25、0.5、0.75 和 1,以模拟单模态(当 $a=0$ 或 $a=1$ 时)和视听融合(当 $a=0.25$,0.5,0.75 时)的情况。在每个权重值下进行测试,记录模型的 Accuracy 和 F_1 值,以评估模型在不同音视频模态权重组合下的鲁棒性。

表 5 在 RAVDESS 数据集上的鲁棒实验结果

权重	音频权重	视频权重	Accuracy/%	F_1 /%
0	0	1.00	73.50	75.52
0.25	0.25	0.75	74.10	76.01
0.50	0.50	0.50	75.62	77.23
0.75	0.75	0.25	74.80	76.80
1.00	1.00	0	72.80	74.85

从表 5 实验结果可以看出,文中模型在不同音视频模态权重组合下表现出较高的鲁棒性和适应性。具体而言,在单模态输入的情况下(仅音频或仅视频),模型的准确率和 F_1 值略有下降,但仍保持较好的识别效果;而在多模态融合的权重设置中,特别是音视频特征均衡输入(权重为 0.5)时,模型表现最佳,达到 75.62% 的准确率和 77.23% 的 F_1 值。这表明,文中模型能够在模态信息不均衡的情况下依然有效融合情绪特征,跨模态注意力机制和多尺度 KAN 卷积的设计增强了模型对多模态情绪识别任务的鲁棒性和稳定性。

4 结束语

该文提出了一种多尺度 KAN 卷积与跨模态注意力的视听情绪识别模型(MKCC)。该模型通过设计的跨模态块,充分考虑了不同模态之间的互补信息,不仅能够在模态之间传递关键信息,还能在每个模态内部进行有效的特征交互,从而增强语义特征的表达与传递能力。多尺度 KAN 卷积的引入确保了在信息交互过程中保持较高的效率和信息传递的完整性,避免了传统方法中信息损失或冗余的问题。在 RAVDES

数据集的实验验证表明,MKCC 方法达到了当前最先进的水平,取得了 75.76% 的准确率。这表明该方法在视听情绪识别任务中具有显著的优势。在未来的研究中,计划进一步扩展该模块,探索更多模态之间的高效互动,尤其是对涉及更多复杂模态交互的任务。此外,将考虑优化跨模态注意力机制,使其更适应不同任务的需求,从而在更广泛的多模态任务场景中应用该方法,提升不同领域的模型性能。

参考文献:

[1] 张国锋,李祖枢. 人工生命行为选择情绪机制研究进展[J]. 小型微型计算机系统,2012,33(8):74-80.

[2] 吴江照,李　伟,张其隆,等. 多模态情感对话技术:研究综述与发展趋势[J]. 人工智能,2024(3):45-56.

[3] 陈晓婷,李　实. 对话情绪识别综述[J]. 计算机工程与应用,2023,59(3):33-48.

[4] 潘家辉,何志鹏,李自娜,等. 多模态情绪识别研究综述[J]. 智能系统学报,2020,15(4):33-45.

[5] 任泽裕,王振超,柯尊旺,等. 多模态数据融合综述[J]. 计算机工程与应用,2021,57(18):49-64.

[6] 祁铧颖,贺　萍. 跨模态数据融合综述[J]. 软件工程,2022,25(10):1-7.

[7] HOSSAIN M S,MUHAMMAD G. Emotion recognition using deep learning approach from audio - visual emotional big data[J]. Information Fusion,2019,49:69-78.

[8] LECUN Y,KAVUKCUOGLU K,FARABET C. Convolutional networks and applications in vision[C]//Proceedings of 2010 IEEE international symposium on circuits and systems. Paris:IEEE,2010:253-256.

[9] ZHOU H,MENG D,ZHANG Y,et al. Exploring emotion features and fusion strategies for audio-video emotion recognition[C]//2019 international conference on multimodal interaction. Suzhou:ACM,2019:562-566.

[10] SHEN S,YOUNES R. Reimagining linear probing:Kolmogorov-Arnold networks in transfer learning[J]. arXiv:2409.07763,2024.

[11] VO-THANH H S,NGUYEN Q V,KIM S H. KAN-based fusion of dual-domain for audio-driven facial landmarks generation[J]. arXiv:2409.05330,2024.

[12] SCHONEVELD L,OTHMANI A,ABDELKAWY H. Leveraging recent advances in deep learning for audio-visual emotion recognition[J]. Pattern Recognition Letters,2021,146:1-7.

[13] PRAVEEN R G,GRANGER E,CARDINAL P. Cross attentional audio-visual fusion for dimensional emotion recognition[C]//2021 16th IEEE international conference on automatic face and gesture recognition (FG 2021). Lugano:IEEE,2021:1-8.

[14] CAI Y,ZHENG W,ZHANG T,et al. Video based emotion recognition using CNN and BRNN[C]//Pattern recognition:7th Chinese conference,CCPR 2016. Chengdu:Springer,2016:679-691.

[15] BAKER M R,PATIL R B. Universal approximation theorem for interval neural networks[J]. Reliable Computing,1998,4:235-239.

[16] LIU Z,WANG Y,VAIDYA S,et al. Kan:Kolmogorov-Arnold networks[J]. arXiv:2404.19756,2024.

[17] BODNER A D,TEPSICH A S,SPOLSKI J N,et al. Convolutional Kolmogorov - Arnold networks[J]. arXiv:2406.13155,2024.

[18] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM,2017,60(6):84-90.

[19] HUANG Z,WANG X,HUANG L,et al. Ccnet:criss-cross attention for semantic segmentation[C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul:IEEE,2019:603-612.

[20] LIU Y,ONG N,PENG K,et al. Mmvit:multiscale multiview vision transformers[J]. arXiv:2305.00104,2023.

[21] YANG C K,CHEN M H,CHUANG Y Y,et al. 2D-3D interlaced transformer for point cloud segmentation with scene-level supervision[C]//Proceedings of the IEEE/CVF international conference on computer vision. Paris:IEEE,2023:977-987.

[22] JIAN L,XIONG S,YAN H,et al. Rethinking cross-attention for infrared and visible image fusion[J]. arXiv:2401.11675,2024.

[23] BALTRUŠAITIS T,ROBINSON P,MORENCY L P. Openface:an open source facial behavior analysis toolkit[C]//2016 IEEE winter conference on applications of computer vision. Lake Placid:IEEE,2016:1-10.

[24] MOLLAHOSSEINI A,HASANI B,MAHOOR M H. Affectnet:a database for facial expression,valence,and arousal computing in the wild[J]. IEEE Transactions on Affective Computing,2017,10(1):18-31.

[25] MIDDYA A I,NAG B,ROY S. Deep learning based multimodal emotion recognition using model-level fusion of audio - visual modalities[J]. Knowledge-Based Systems,2022,244:108580.

[26] NAGRANI A,YANG S,ARNAB A,et al. Attention bottlenecks for multimodal fusion[J]. Advances in Neural Information Processing Systems,2021,34:14200-14213.

[27] SU L,HU C,LI G,et al. Msaf:multimodal split attention fusion[J]. arXiv:2012.07175,2020.

[28] MOCANU B,TAPU R,ZAHARIA T. Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning[J]. Image and Vision Computing,2023,133:104676.