多模态情绪识别在驾驶场景下的应用

魏倩楠[1] 张志华[1] 赵英辰[1] 林嘉怡[1] 郭琛璐[1]
WEI Qiannan ZHANG Zhihua ZHAO Yingchen LIN Jiayi GUO Chenlu

摘 要 驾驶员的情绪状态是影响行车安全的关键因素之一。基于单一模态的情绪识别，在驾驶场景下存在鲁棒性不足的问题。通过融合视频面部表情、静态图片表情、语音文本等信息，文章提出一种面向驾驶场景的多模态实时情绪识别框架。首先，构建了基于改进 ResNet-18 的实时视频表情识别模型，引入轻量化注意力模块以提升对细微表情的捕捉能力；其次，针对图片表情识别，采用 MobileNetV3 作为主干网络，并结合数据增强策略以应对光照、姿态变化；最后，语音文本识别模型采用预训练的 Wav2Vec 2.0 提取语音特征，与基于 BERT 的文本情感分析模型进行特征级融合。在驾驶情绪数据集（包含 30 名被试的超过 50 h 多模态数据）上测试结果表明，所提多模态融合方法在情绪识别准确率上达到 94.2%，优于单一模态模型，可为驾驶员状态监控与智能座舱交互系统提供有效技术支持。

关键词 情绪识别；多模态融合；驾驶安全；计算机视觉；语音情感分析

doi：10.3969/j.issn.1672-9528.2026.04.032

0 引言

随着智能汽车技术的快速发展，人车交互的智能化与安全性成为研究热点。驾驶员在行车过程中的情绪波动，如愤怒、疲劳、兴奋等，会直接影响其认知判断与操作反应，是引发交通事故的重要潜在因素[1]。因此，对驾驶员情绪状态进行实时、准确的识别与监测，对于提升主动安全与驾乘体验具有重要意义。传统情绪识别研究多依赖于单一信息源，如基于计算机视觉的面部表情分析、基于语音信号的情感计算或基于自然语言处理的文本情感分析[2]。然而，在真实的驾驶场景中，单一模态信号极易受到复杂环境干扰（如光线变化、背景噪声、部分遮挡等），导致识别性能下降。多模态信息融合能够利用不同模态间的互补性，提高系统的鲁棒性与准确性。现有研究在通用场景下的多模态情感分析已取得进展，但专门针对驾驶场景的、兼顾实时性与高精度的多模态情绪识别方案仍待深入探索。本文旨在构建一个适用于车载环境的实时多模态情绪识别系统。核心贡献包括：

（1）设计并训练了分别针对实时视频流、静态抓拍图片、语音文本的三个高精度识别模型；

（2）提出一种高效的特征级与决策级融合策略，充分利用跨模态信息；

（3）整个系统设计充分考虑车载计算资源约束，模型均经过轻量化设计，满足实时性要求。

1. 辽宁科技大学电子与信息工程学院 辽宁鞍山 114051
[基金项目] 辽宁科技大学大学生创新创业训练计划项目

1 研究背景与技术基础

1.1 视觉情绪识别

视觉情绪识别主要通过对面部表情进行分析实现。早期研究方法大多采用手工设计特征（Hand-crafted Features）结合传统机器学习分类器。随着深度学习技术的发展，卷积神经网络逐渐成为该领域的主流方法。在驾驶场景中，视觉情绪识别面临光照变化、头部姿态偏移、佩戴眼镜等多重挑战[3]。为满足移动端与嵌入式设备的实时处理需求，轻量化网络结构如 MobileNet、ShuffleNet 等被广泛采用，可以在有限计算资源下维持较高的识别性能。

1.2 语音与文本情绪识别

语音情感识别通常从音频信号中提取声学特征，并输入至分类模型进行分析[4]。基于端到端学习的深度模型，如卷积神经网络与循环神经网络及其变体，能够直接处理原始波形或语谱图，在多项任务中表现出优越性能。文本情感分析则利用词向量与上下文建模技术捕捉语义中的情绪信息[5]。在驾驶场景下，语音识别易受环境噪声干扰，因此提升噪声鲁棒性成为确保识别准确性的关键。

1.3 多模态情绪识别

多模态情绪识别通过融合视觉、图片、语音文本等多种信息来源提升识别效果[6]。常见的融合层次包括数据级、特征级与决策级融合，其中特征级与决策级融合方法应用更为广泛[7]。在车载环境中，需综合考虑融合性能与系统计算开销，寻求适合于实时情绪分析的多模态融合策略。

2 模型构建与训练

本文提出的多模态情绪识别框架包含视频表情识别、图片表情识别、语音文本识别三个子模型，以及一个多模态融合模块。系统实时接收车载摄像头采集的视频流，并从视频流中抽取关键帧图像作为图片表情识别模型的输入，同时，系统接收麦克风采集的语音信号，语音信号经自动语音识别（automatic speech recognition, ASR）转写为文本后，基于文本语义特征进行情绪识别。最终输出驾驶员的平静、高兴、愤怒、悲伤、惊讶、恐惧、厌恶、疲劳等情绪类别。

2.1 基于改进 ResNet-18 的实时视频表情识别模型

为平衡精度与速度，本文选择 ResNet-18 作为主干网络，ResNet-18 具有结构清晰、参数规模适中、计算效率高等优点，适合部署于车载等资源受限的实时场景 [8]，其网络结构如图 1 所示。

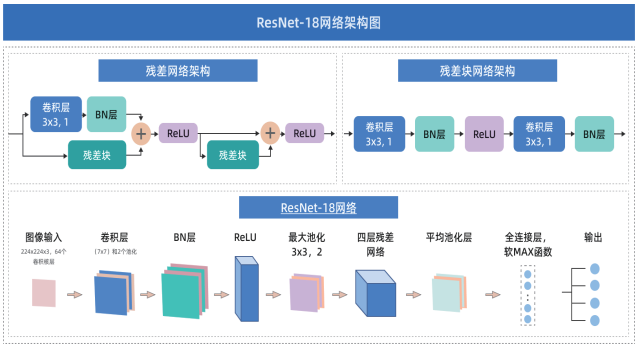

图 1 Res-Net 网络架构图

ResNet-18 以输入尺寸为 224 px×224 px 的人脸图像作为输入，首先通过一个 7×7 卷积层进行初步特征提取，随后经过批归一化（BN）层和 ReLU 激活函数，并通过最大池化层降低特征图尺寸、减少计算量。

在此基础上，网络主体由四个残差阶段（residual block stages）组成，每个阶段包含若干个残差块。如图 1 中残差块结构所示，每个残差块由两个 3×3 卷积层构成，并通过捷径连接（shortcut Connection）将输入特征与卷积输出相加，有效缓解了深层网络训练过程中的梯度消失问题，提高了特征表达能力。

经过四个残差阶段后，网络采用全局平均池化（global average pooling）对空间特征进行压缩，得到高层语义特征表示。最后，通过全连接层并结合 Softmax 函数输出各类表情的概率分布，实现对单帧人脸表情的分类。

通过上述结构，ResNet-18 能够有效提取人脸图像中的空间表情特征，为后续的视频时序建模与多模态融合提供可靠的特征基础。

2.2 基于 MobileNetV3 的图片表情识别模型

对于从视频流中截取的关键帧（如检测到显著表情变化时）或单独输入的图片，采用轻量化的 MobileNetV3-large 作为特征提取器。

（1）网络适配：将 MobileNetV3 最后的分类层替换为适应 8 类情绪的全连接层，如图 2 所示。

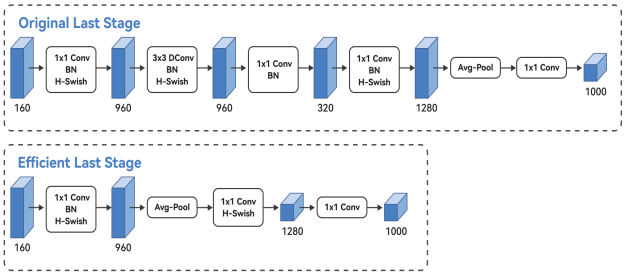

图 2 MobileNetV3-large 网络最后阶段结构改进对比图

图 2 中上半部分为原始最后阶段结构，由多层卷积操作组成，包括 1×1 卷积、3×3 深度可分离卷积、批归一化以及 h-swish（hard swish）激活，并通过平均池化和 1×1 卷积输出分类特征。

图 2 中下半部分为改进后的最后阶段结构，保留 1×1 卷积、平均池化和 h-Swish 激活等操作，减少了中间卷积层数量，最终通过 1×1 卷积完成特征映射。利用其内置的 SE（squeeze-and-excitation）模块和 h-swish 激活函数，在低计算成本下保持较强表征能力 [9]。

（2）针对驾驶场景的增强：考虑到驾驶场景中驾驶员头部姿态变化频繁、视角偏移明显等实际情况，本文通过引入头部偏移、仿射变换等数据增强方式，模拟驾驶过程中常见的姿态变化与成像差异。该类增强操作使模型在训练阶段能够接触到更加多样化的人脸外观，从而引导识别模型学习对姿态变化和空间变形具有鲁棒性的判别特征，有效提升模型在真实驾驶场景下的人脸表情识别性能与泛化能力。

2.3 语音文本识别模型

该模块分为语音特征提取和文本情感分析两个子模块，并实现特征级融合。

（1）语音情感特征提取：采用在大量无标签音频上预训练的 Wav2Vec 2.0 Base 模型 [10]。对输入音频，提取其上下文表征序列，然后通过一个时间池化层（如注意力池化）聚合为固定维度的语音情感特征向量。

（2）文本情感分析：语音通过 ASR 引擎转换为文本。采用预训练的 BERT-base（bidirectional encoder representations from transformers）模型获取文本的 [CLS] 标记向量作为文本语义特征 [11]。针对驾驶场景常见短语在数据集上对 BERT 进行领域自适应微调。

（3）特征融合：将语音特征和文本特征进行拼接，并通过一个全连接层进行融合与降维，得到联合特征，再输入分类器 [12]。这种方式允许模型在 ASR 出现误差时，仍能从语音的声学特性中捕捉情绪线索。

3 实验验证

3.1 模型训练与性能评估

本研究的核心目标是开发一个适用于车载环境的驾驶员情绪识别模型。为此，构建了驾驶场景多模态情绪数据集（DMED），其中包含 30 名不同性别、年龄的驾驶员在模拟驾驶舱及实车环境下产生的面部视频、图像、音频及转写文本数据，涵盖 8 种基本情绪，总时长超过 50 h。数据按 7:2:1 的比例划分为训练集、验证集和测试集。实验在 NVIDIA RTX 3090 GPU 上使用 PyTorch 框架进行模型训练，最终在 Jetson AGX Xavier 嵌入式平台上完成推理测试，以模拟真实车载部署环境。

（1）超参数设置与训练过程

整个模型训练共进行 20 轮。采用动态调整的学习率策略，以优化收敛过程。由于各子模型训练过程具有相似的收敛特性，本文以图像情绪识别子模型为例展示其训练与验证过程的损失及准确率变化情况，如图 3 所示。训练过程在 20 个轮次内呈现出良好的收敛趋势。训练集准确率与损失函数随轮次增加稳步优化，验证集指标亦同步改善，表明未出现过拟合现象。最终，模型在验证集上取得了 79.32% 的最佳准确率，在独立测试集上的准确率为 74.32%。此外，视频情绪识别子模型与语音文本情绪识别子模型在相同训练策略下亦表现出稳定的收敛过程，其中视频模型在测试集上的准确率与加权 F_1 值均保持在较高水平，语音文本模型虽受模态特性影响整体性能略低，但仍取得了具有参考价值的识别效果。

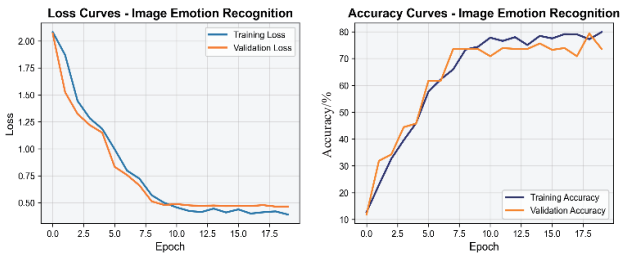

图 3 图像情绪识别模型在训练与验证集上的损失与准确率变化曲线

（2）学习率调度策略

为进一步提升模型性能，本文采用了分阶段的学习率调度策略。在训练初期（第 0~8 轮），设置较高的初始学习率 0.001，以加快模型的搜索与收敛速度；随后在中期阶段将学习率降低至 0.000 5，以实现更加稳定的训练过程；在训练后期（接近第 20 轮），进一步将学习率衰减至 0.000 3，以便对模型参数进行精细化调整。该学习率调度策略与前述 20 轮训练设置相配合，有助于在保证收敛速度的同时提升模型的最终性能，如图 4 所示。

（3）模型分类性能分析

为评估模型在测试集上的细粒度分类能力，绘制了归一化混淆矩阵。如图 5 所示，模型在八类情绪上的预测结果主要集中于对角线位置，表明模型对各类情绪有较好的识别能力。与此同时，个别情绪类别之间仍存在少量误分类现象，主要分布于相邻或情绪表现相近的类别之间，整体混淆程度较低。上述结果说明，所提出的多模态情绪识别模型在保持较高分类准确性的同时，具备较好的类别区分能力与识别稳定性。

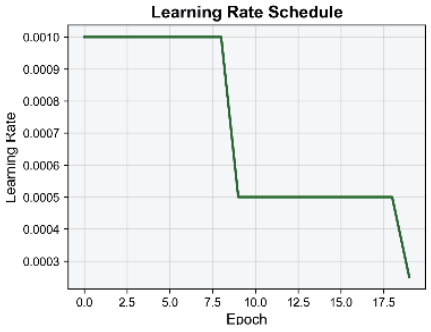

图 4 多模态情绪识别模型学习率调度

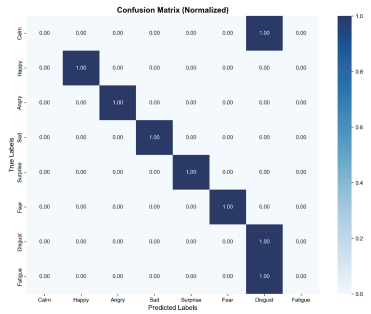

图 5 多模态情绪识别模型归一化混淆矩阵

（4）特征空间可视化

如图 6 所示，为进一步探究模型所学特征的区分性，利用 t-SNE 算法将高维特征降维至二维平面进行可视化。

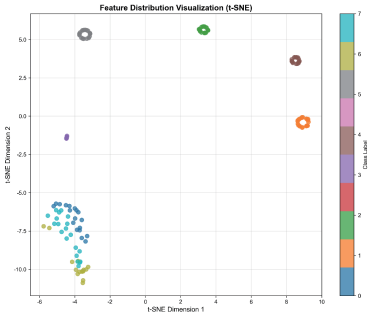

图 6 多模态情绪特征判别分布的 t-SNE 可视化

可以观察到，大多数同类样本在特征空间中呈现出相对集中的分布趋势，形成若干具有代表性的簇状结构。同时，部分情绪类别在低维嵌入空间中存在一定程度的重叠，造成该现象的主要原因在于不同情绪类别之间在情感语义和表达方式上具有一定的相似性。例如，“高兴”“激动”等情绪在情感极性、语义含义以及语音语调、面部表情等多模态特

征上存在较强的共性，使得模型在特征表示时难以形成完全清晰的边界，从而在特征空间中表现为相邻甚至局部重叠的分布。总体而言，该特征分布仍体现出良好的类内聚合与类间可分性，从特征层面验证了多模态融合模型的判别能力。

3.2 单模态模型测试结果与分析

三种单模态模型在 DMED 测试集上的性能对比如表 1 所示。可以看出，在受控相对较好的条件下，图片模型取得了较高准确率和加权 F_1 值。视频模型因利用了时序信息，在整体性能上略低于图片模型，但仍保持较高的识别精度。语音文本模型整体性能低于视觉模型，主要受驾驶舱噪声环境对语音信号质量的影响，同时其较高的模型复杂度在单模态条件下限制了复杂情绪特征的建模效果。但语音模态能够从语调、语速等非视觉线索中补充驾驶员状态信息，在多模态融合框架中仍具有重要的辅助价值。

表 1 单模态模型在 DMED 测试集上的性能对比

模型	准确率 /%	加权 F_1 分数 /%	参数量 /10^6	推理速度（单样本）/ms
视频模型 (ResNet-18+GRU)	88.5	87.9	13.2	35
图片模型 (MobileNetV3)	90.1	89.7	5.4	12
语音文本模型 (Wav2Vec2+BERT)	82.3	81.5	约 220	120

3.3 多模态融合结果与分析

通过在验证集上对不同权重组合进行对比实验，以准确率和 F_1 分数作为评价指标，最终确定最优融合权重为 α=0.4，β=0.3，γ=0.3，其中 α 表示视频模态权重，β 表示图片模态权重，γ 表示语音文本模态权重，且满足 $\alpha+\beta+\gamma=1$。多模态融合模型在测试集上的结果如表 2 所示。融合模型在各项指标上均显著超越最佳单模态模型（图片模型），准确率提升 4.1%，F_1 分数提升 4.5%。充分证明了多模态融合的有效性。

表 2 多模态融合模型与单模态模型性能对比

单位：%

模型	准确率	精确率	召回率	F_1 分数
仅视频模型	88.5	88.1	88.0	87.9
仅图片模型	90.1	89.8	89.6	89.7
仅语音文本模型	82.3	81.8	81.5	81.5
多模态融合 (本文)	94.2	94.0	94.1	94.2

4 结语

本文针对驾驶场景下的情绪识别任务，提出了一种融合视频、图片、语音与文本的多模态实时识别框架。设计并训练了三个轻量化、高精度的单模态识别模型，采用基于动态权重的决策级融合方法，有效提升了复杂驾驶环境下情绪识别的准确率与鲁棒性。实验验证表明，多模态融合方法相比最优单模态方法，准确率提升了 4.1%。本文工作为车载驾驶员状态监控系统提供了可行的技术方案。

未来工作将从以下几个方面展开：首先，探索更高效的低层特征融合方法，以进一步挖掘模态间的深层关联；其次，考虑引入更多生理信号模态（如心率、皮电），构建更全面的驾驶员状态感知系统；最后，将模型部署到真实车载平台，进行大规模、长周期的实车测试，以验证其在实际复杂环境下的稳定性与实用性。

参考文献：

[1] 丁诚 . 车联网场景下面向驾驶安全的情绪识别研究 [D]. 南京 : 南京邮电大学 , 2023.

[2] 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):633-645.

[3] 焦蕊 . 基于深度学习的情绪识别技术研究 [D]. 北京 : 中央民族大学 ,2022.

[4] 廖志成 . 基于声音特征和脑电的情绪识别研究 [D]. 成都 : 电子科技大学 ,2024.

[5] 侯米潇 . 基于判别语义学习的情绪识别方法研究 [D]. 哈尔滨 : 哈尔滨工业大学 ,2024.

[6] 方伟杰 , 张志航 , 王恒畅 , 等 . 融合语音、脑电和人脸表情的多模态情绪识别 [J]. 计算机系统应用 , 2023, 32(1): 337-347.

[7] 刘勇 . 基于多模态生物电信号的情绪识别方法研究 [D]. 长春 : 长春理工大学 ,2021.

[8] 段函作 , 潘溢洲 , 寇嘉铭 , 等 . 基于改进 ResNet18 模型的驾驶员面部表情识别方法 [J]. 传感器与微系统 , 2025, 44(6): 29-32.

[9] 岑承瑞 , 李海侠 . 基于 MobileNetV3 的人脸微表情识别系统研究 [J]. 现代信息科技 , 2025, 9(24): 77-82.

[10] 曹荣贺 , 吴晓龙 , 冯畅 , 等 . 基于 Wav2vec2.0 与语境情感信息补偿的对话语音情感识别 [J]. 信号处理 , 2023, 39(4): 698-707.

[11] 张栋 . 基于多模态特征的情感分析方法研究 [D]. 苏州 : 苏州大学 , 2020.

[12] 朱清扬 . 基于脑电、外周生理信号和面部表情的多模态情感识别 [D]. 南京 : 南京邮电大学 ,2021.

【作者简介】

魏倩楠（2004—），女，河南郑州人，本科在读，研究方向：深度学习。

张志华（1979—），女，辽宁鞍山人，博士，研究方向：计算智能、机器学习，email: aszzh@sina.com。

（收稿日期：2026-01-26 修回日期：2026-04-01）