Software Engineering and Applications 软件工程与应用, 2026, 15(3), 423-432
Published Online June 2026 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2026.153040

Hans汉斯

驾驶场景下的多模态情绪识别系统设计与实现

冯建斌，魏倩楠

辽宁科技大学电子与信息工程学院，辽宁 鞍山

收稿日期：2026年5月11日；录用日期：2026年6月18日；发布日期：2026年6月30日

摘　要

在真实驾驶环境中，车内光线、道路噪声、驾驶员头部偏转等因素都会干扰情绪识别结果。单独依靠某一种数据源时，模型判断容易出现不稳定。围绕这一情况，本文设计并实现了一套驾驶场景下的多模态情绪识别系统，将视频表情、静态图像以及语音文本三类信息放入同一识别流程中。系统主要包含三个识别模块：视频流模块以ResNet-18为基础，在残差结构中加入SE注意力机制，用来加强眼部、嘴部等表情区域的响应；图像模块采用MobileNetV3，并配合旋转、裁剪、亮度扰动等数据增强方式，提高模型对不同拍摄条件的适应能力；语音文本模块分别利用Wav2Vec 2.0和BERT提取声学特征与语义特征。实验基于DMED数据集完成，融合模型准确率达到94.2%，比单模态模型提高4.1%。同时，本文使用PyQt5完成可视化界面开发，实现了数据输入、预处理、模型推理和结果展示等功能，整体上能够满足车载场景下的基本实时检测需求。

关键词

情绪识别，多模态融合，驾驶安全，ResNet-18，YOLO

Design and Implementation of a Multimodal Emotion Recognition System Specialized for Driving Scenarios

Jianbin Feng, Qiannan Wei

School of Electronic Information Engineering, University of Science and Technology Liaoning, Anshan Liaoning

Received: May 11, 2026; accepted: June 18, 2026; published: June 30, 2026

Abstract

In real driving environments, factors such as in-vehicle lighting, road noise, and driver head-pose

文章引用: 冯建斌, 魏倩楠. 驾驶场景下的多模态情绪识别系统设计与实现[J]. 软件工程与应用, 2026, 15(3): 423-432. DOI: 10.12677/sea.2026.153040

changes can interfere with emotion recognition results. When relying on only one type of data source, the model's judgment is prone to instability. To address this issue, this paper designs and implements a multimodal emotion recognition system for driving scenarios, integrating video-based facial expressions, static images, and speech-text information into a unified recognition process. The system mainly consists of three recognition modules: the video-stream module is based on ResNet-18 and incorporates an SE attention mechanism into the residual structure to enhance the response to facial regions such as the eyes and mouth; the image module adopts MobileNetV3 and uses data augmentation methods such as rotation, cropping, and brightness disturbance to improve the model's adaptability to different shooting conditions; the speech-text module uses Wav2Vec 2.0 and BERT to extract acoustic and semantic features, respectively. Experiments are conducted on the DMED dataset, and the fusion model achieves an accuracy of 94.2%, which is 4.1 percentage points higher than that of the best single-modal model. In addition, this paper develops a visualization interface using PyQt5, enabling data input, preprocessing, model inference, and result display. Overall, the system can meet the basic real-time detection requirements in vehicle-mounted scenarios.

Keywords

Emotion Recognition, Multimodal Fusion, Driving Safety, ResNet-18, YOLO

1. 引言

从公安部交管局公布的道路交通事故统计情况来看，人为因素在交通事故中占有较大比例。驾驶员在行车过程中如果出现情绪激动、注意力下降或疲劳等状态，往往会影响对路况的判断，转向、制动等操作也可能随之出现偏差，最终增加行车风险[1]。目前已有一些驾驶状态监测设备投入使用，但在复杂车内环境下，这类设备仍容易受到信息来源单一、识别结果波动以及记录不够完整等问题的影响。因此，结合视觉、语音等多源数据，对驾驶员情绪状态进行综合分析，仍然具有一定研究意义。

现有驾驶情绪识别方法中，单模态方案仍然比较常见。只使用图像时，暗光、侧脸、遮挡和车辆颠簸都会影响面部特征提取；只使用语音或文本时，又容易受到背景噪声、语义歧义和语句长度的影响。基于这些问题，本文把 YOLO、ResNet 以及时序特征处理方法结合起来，建立从数据采集、模型识别到结果保存的系统流程。系统运行时可以对驾驶员情绪变化进行实时判断，并保存识别日志，为后续安全管理或事故分析提供辅助依据。系统整体架构如图 1 所示。

2. 系统架构设计

本文系统按照并行处理思路搭建，整体流程可分为数据准备、模型训练以及实时推理展示三个部分。

2.1. 视频流识别(基于改进 ResNet-18 + SE 注意力)

在视频流处理中，系统首先按照固定间隔从原始视频里抽取 5 帧表情图像，随后送入视觉识别网络。考虑到车载端设备算力有限，同时又需要保持较快响应速度，本文采用改进 ResNet-18 作为基础网络[2]，并在残差结构中加入 SE-Block。这样做的目的，是让网络在提取特征时更多关注眼周、嘴角等表情变化较明显的区域。由于驾驶视频样本数量有限，且不同采集场景之间差异较大，训练阶段加入了旋转、裁剪和亮度变化等增强方式，以减小场景变化带来的影响。多模态融合阶段分别提取 Wav2Vec 2.0 声学特

征和 BERT 文本特征，再与视觉特征合并。实验结果中，该方法在 DMED 测试集上的准确率为 94.2%，比单一模态方法提高约 4.1%。此外，PyQt5 界面完成了摄像头采集、模型推理和结果显示等功能，处理速度基本能够满足实时检测需要。视频流识别流程如图 2 所示。

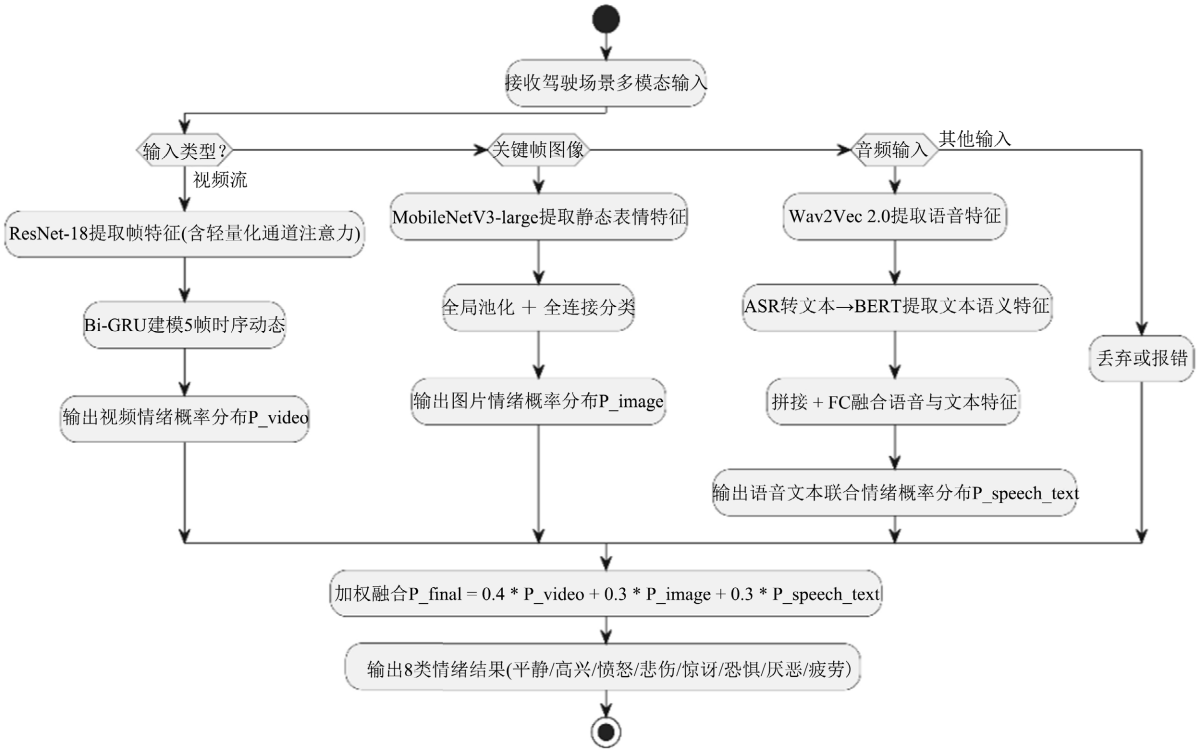

Figure 1. System composition structure diagram
图 1. 系统组成结构图

Figure 2. Video stream recognition flowchart
图 2. 视频流识别流程图

2.2. 图片表情识别(基于 MobileNetV3 + 数据增强)

行车抓拍图像通常要求模型在较短时间内完成判断，因此本模块在设计时主要考虑精度和计算量之间的平衡。本文选用 MobileNetV3-Large 作为图像特征提取网络。该网络包含深度可分离卷积(Depthwise Separable Convolution)和 SE 注意力结构，在减少参数量的同时，仍能保持较好的特征表达能力[3]。

图像处理流程相对直接。系统先把输入图像统一调整为 224 × 224 像素，并进行归一化处理；之后将图像输入 MobileNetV3 主干网络，提取面部轮廓、五官位置和表情纹理等信息。经过多层卷积后，网络会形成对应的空间特征图。随后利用全局平均池化(Global Average Pooling)对特征进行汇聚，以降低人脸位置变化对分类结果的影响。最后，展平后的特征向量进入全连接层，通过 Softmax 得到各类驾驶情绪的概率分布，并作为图像侧结果参与后续融合。

2.3. 语音文本融合识别(基于 Wav2Vec2.0 + BERT)

语音文本模块主要处理驾驶过程中的语音内容及其转写文本。本文没有把语音和文本混在一起直接判断，而是分别提取声学特征和语义特征：语音部分使用预训练 Wav2Vec 2.0，从原始波形中学习语调、节奏和情绪起伏等信息；文本部分使用 BERT 对转写内容进行语义编码。相比只看音量、语速或少量关键词的方法，这种设计能够同时考虑说话方式和表达内容，对情绪状态的判断更完整[4]。

在具体实现中，音频数据先经过重采样、降噪和分帧处理，再输入 Wav2Vec 2.0 得到声学特征向量；同步文本经过分词和编码后输入 BERT，得到文本语义向量[5]。随后将两类特征通过拼接或加权方式进行融合，并送入分类层输出情绪类别。该模块的结果不会单独作为最终判断，而是与视频流、静态图像模块的输出共同参与决策。这样可以在噪声、遮挡或语义信息不完整时，降低单一模态造成误判的可能性[6]。

3. 实验验证

为了验证系统效果，本文进行了模型测试，并搭建了面向车载场景的交互展示界面。实验使用驾驶场景多模态情绪数据集(DMED)。该数据集包含 30 名受试者在模拟驾驶舱和实车环境下采集的面部视频、音频及同步转写文本，覆盖 9 种基础情绪状态，总时长超过 50 小时。数据预处理时，训练集、验证集和测试集按 7:2:1 划分。模型训练在 NVIDIA RTX 3090 GPU 上完成，深度学习框架为 PyTorch。

3.1. 数据集特性分析

训练模型之前，本文先对 DMED 数据集的样本分布进行了统计，如图 3 所示。9 类情绪样本数量并不完全一致，其中 Happy 类样本相对更多，Sleepy 等类别样本偏少。这样的分布与真实驾驶场景比较接近，因为不同情绪在实际行车中并不会以相同频率出现。对模型训练来说，这也意味着分类器需要处理类别不均衡带来的影响。

从坐标热力图可以观察到，驾驶员面部目标大多集中在图像中心附近，目标框尺寸也呈现出较明显的聚集趋势。该现象可为 YOLO 模型的 Anchor 尺寸设置和边界框回归提供参考，使模型在检测面部区域和局部表情细节时更加稳定。

3.2. 训练收敛性与损失函数分析

如图 4 所示，记录了模型 100 个 Epoch 训练过程中的指标变化，下面结合损失和检测指标进行说明。

1) 损失函数(Loss Curves)：实验记录了定位损失(box_loss)、分类损失(cls_loss)和分布聚焦损失(dfl_loss)的变化。训练初期，训练集和验证集损失下降较明显；大约 50 个 Epoch 后，各条曲线趋于平缓。后期验证集损失没有明显反弹，说明在当前数据划分下，模型暂未出现突出的过拟合现象。

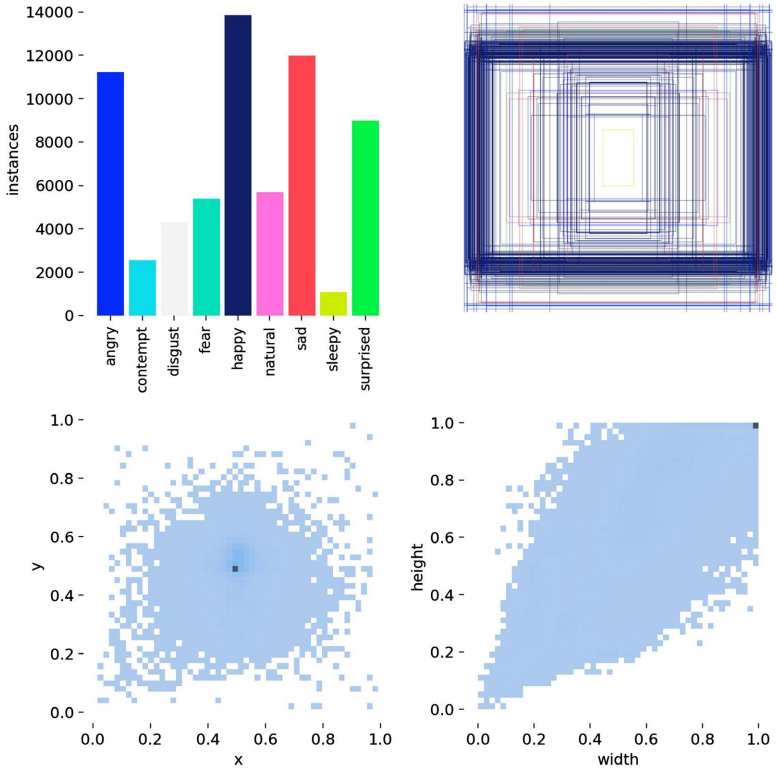

Figure 3. Statistical analysis of the DMED dataset
图 3. DMED 数据集统计分析图组

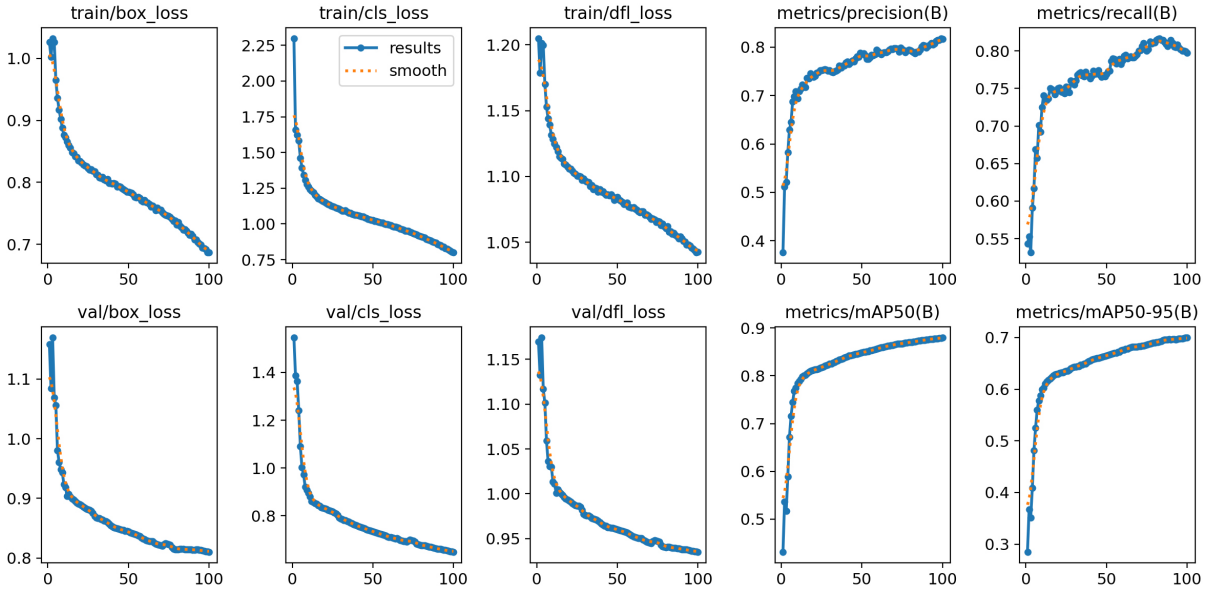

Figure 4. Training procedure and loss change curves of the ambivalent emotion recognition model
图 4. YOLO 情绪识别模型训练过程及损失变化曲线

2) 核心性能指标(Metrics)：随着训练轮次增加，精确率、召回率和 mAP 整体呈上升趋势。训练结束时，mAP50 稳定在 0.88 左右，mAP50-95 约为 0.70。这说明模型已经能够较好地区分不同情绪类别，在存在背景干扰的情况下也能保持一定检测稳定性。

3.3. 模型情绪识别效果分析

3.3.1. 基于 P-R 曲线的识别性能分析

由图 5 的 P-R 曲线可以看出，模型在 9 类驾驶情绪上的整体表现较稳定。全类别平均精度 mAP@0.5 为 0.880。该结果说明，在光照变化、侧脸和局部遮挡等驾驶舱常见干扰下，模型仍然能够完成较可靠的情绪检测。

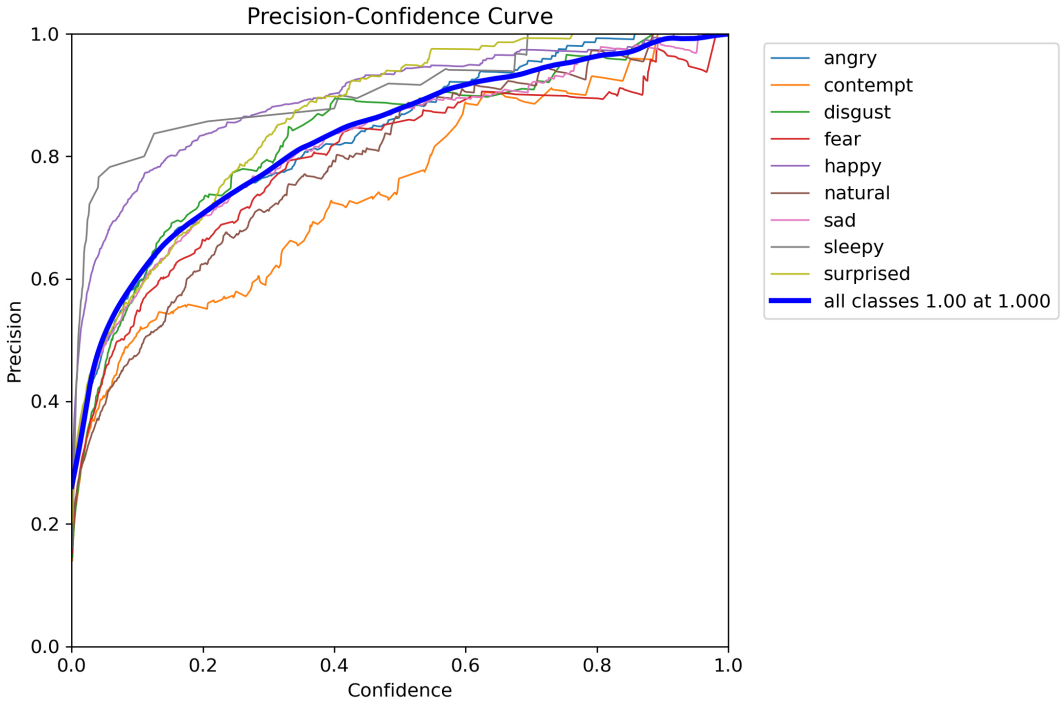

Figure 5. Precision-Recall (P-R) curves for multi-class emotion recognition
图 5. 多类情绪识别的精确率－召回率(P-R)曲线

从具体类别看，Happy 和 Sleepy 的识别效果更好。这两类情绪通常有比较明显的外部表现，例如嘴角上扬、眼睑闭合或打哈欠等，模型更容易捕捉到相关特征。相比之下，Contempt (0.766)和 Natural (0.787)的识别难度更高，主要原因是两类表情变化幅度较小，与其他类别之间的边界不够清楚。整体来看，当前模型在计算量相对可控的情况下，已经能够满足车载端情绪识别的基本需求。

3.3.2. 置信度阈值与识别性能分析

为了观察置信度阈值变化对识别结果的影响，本文进一步分析了图 6 所示的 F1-Confidence 曲线。F1 分数同时考虑精确率和召回率，因此可以用来衡量实时监测系统在误报和漏报之间的平衡情况。

曲线结果表明，全类别 F1 值随着置信度阈值升高呈现先升后降的变化。当阈值约为 0.347 时，模型综合表现较好，此时 F1 均值为 0.81。实际部署时，可将阈值设置在 0.35 左右。这样既可以减少低置信度结果带来的误报，也能尽量保留对潜在风险情绪的识别。

在 0.8~1.0 的高置信度区间内，Sleepy 和 Natural 等类别的曲线变化相对平稳。即使把阈值提高到 0.6 以上，模型仍能保持一定识别能力。这说明模型对部分关键驾驶状态具有较好的判别稳定性。

3.3.3. 基于混淆矩阵的分类结果分析

为进一步查看模型在测试集上的分类情况，本文绘制归一化混淆矩阵，结果如图 7 所示。

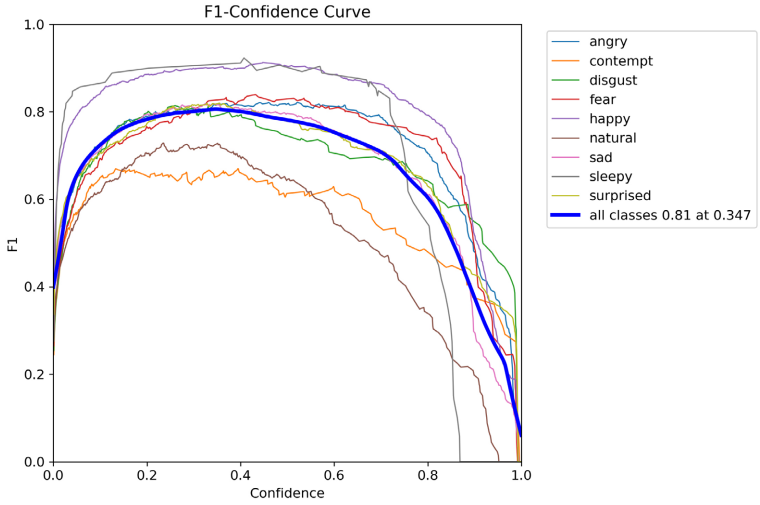

Figure 6. Relationship curve between F1 score and confidence in driver emotion recognition task
图 6. 驾驶员情绪识别任务的 F1 分数与置信度关系曲线

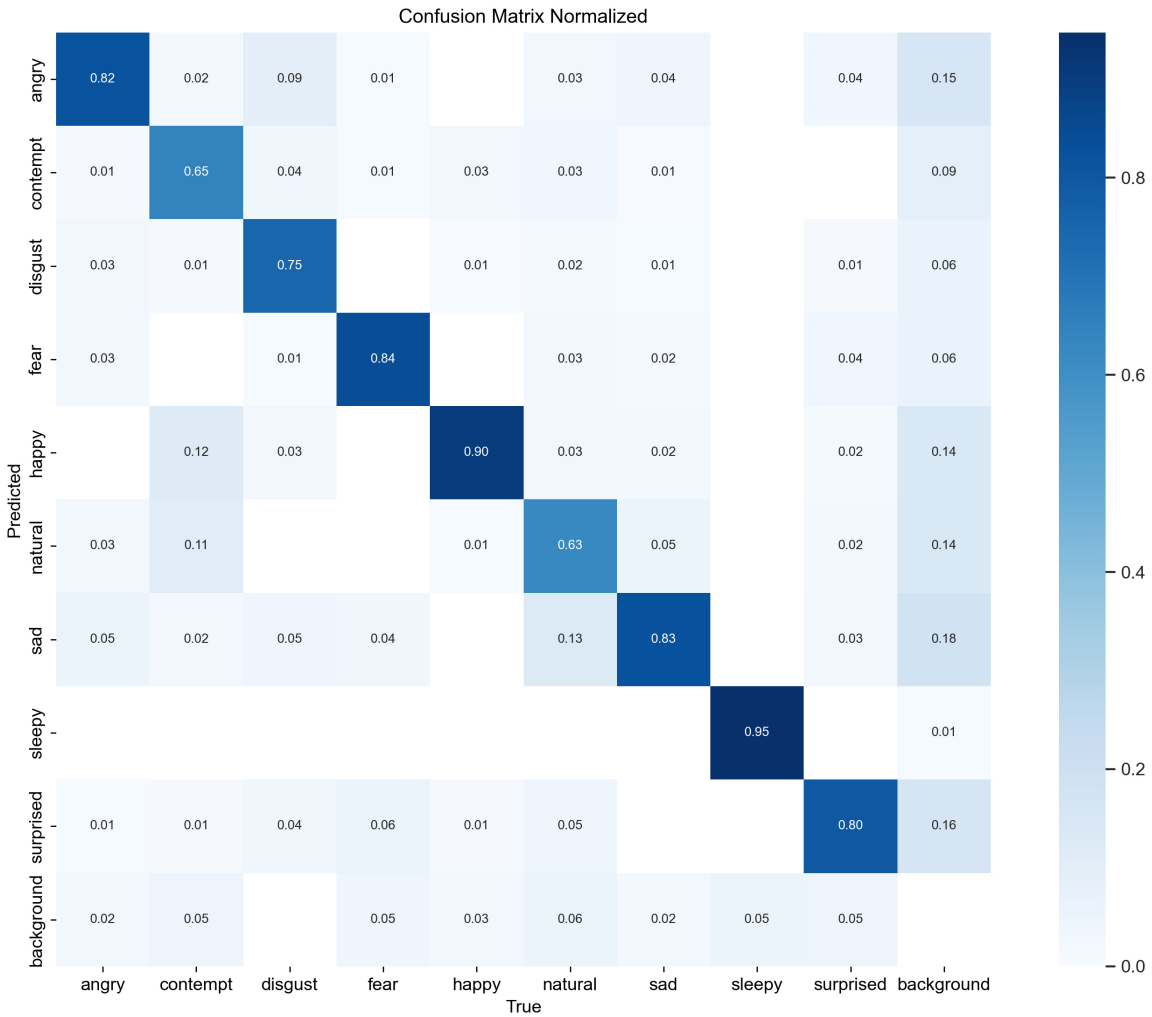

Figure 7. Normalized confusion matrix for 9-class driving emotion recognition
图 7. 9 类驾驶情绪识别的归一化混淆矩阵

从混淆矩阵可以看出，Sleepy 类别的识别准确率最高，为 0.95。该结果与困倦状态的视觉表现有关，例如眼睑闭合、打哈欠等特征较容易被模型捕捉。Happy 类别准确率为 0.90，说明视觉特征与语音特征在该类别上能够形成一定互补。Fear 和 Angry 的准确率分别为 0.84 和 0.82，整体识别结果也较为稳定。

同时，Natural 与 Sad 之间仍存在一定混淆。原因可能是轻微低落情绪和自然状态在面部表情上差异较小，如果再受到光照变化或口罩遮挡影响，视觉特征会进一步变弱。加入语音和文本信息后，系统可以补充一部分视觉信息不足，从而降低部分误判。

3.4. 与公开多模态驾驶情绪识别方法的横向比较

由表 1 可见，公开研究普遍表明多模态融合能够提升驾驶情绪识别性能。与 Xiang 等的面部视频 + 生理信号融合、MDEmoNet 的面部视频 + 驾驶行为融合不同，本文系统面向普通车载摄像头和车内语音输入条件，采用视频、图像和语音文本三路信息进行融合。在 DMED 相同数据划分下，本文融合模型准确率达到 94.2%，F1 分数为 94.2%，高于视频模型(准确率 88.5%，F1 为 87.9%)、图片模型(准确率 90.1%，F1 为 89.7%)和语音文本模型(准确率 82.3%，F1 为 81.5%)。其中，融合模型较最佳单模态图片模型准确率提高 4.1 个百分点，说明所设计的多模态融合流程能够在不额外引入生理采集设备或车辆行为传感器的情况下，提高驾驶场景情绪识别的稳定性。

Table 1. Comparison of multimodal driving emotion recognition methods
表 1. 公开多模态驾驶情绪识别方法与本文方法对比

方法	数据集/对比方式	输入模态	公开或本文结果	说明
Xiang 等[7]	MMDE 公开文献结果	面部视频 + 生理信号	多模态融合较单独面部视频提高 11.28%，较单独生理信号提高 6.83%	证明驾驶场景中多源信息融合优于单一模态；但其生理信号模态与本文数据不完全一致
MDEmoNet [8]	PPB-Emo 公开文献结果	面部视频 + 驾驶行为	Accuracy 67.92%, Macro-F1 0.6780	采用多任务学习与决策层融合，是智能座舱驾驶情绪识别中的公开多模态方法
本文视频模型	DMED 本文同一划分	视频表情	Accuracy 88.5%, F1 87.9%	单模态基线
本文图片模型	DMED 本文同一划分	静态图像	Accuracy 90.1%, F1 89.7%	最佳单模态基线
本文语音文本模型	DMED 本文同一划分	语音 + 文本	Accuracy 82.3%, F1 81.5%	单模态基线
本文融合模型	DMED 本文同一划分	视频 + 图像 + 语音文本	Accuracy 94.2%, F1 94.2%	较最佳单模态模型提高 4.1 个百分点

注：Xiang 等[7]和 MDEmoNet [8]的结果来自公开文献；本文单模态模型和融合模型结果均来自 DMED 测试集相同划分下的实验结果。由于公开方法所用数据集和输入模态与本文不完全一致，表中公开方法主要用于方法层面的横向参照，不作为同一数据集下的直接数值排名。

4. PyQt5 GUI 可视化界面实现

本系统采用 Python 进行开发，将目标检测算法与图形界面结合起来，用于展示驾驶情绪识别结果。

4.1. 系统开发环境与架构

在系统架构上，本文采用分层设计，将底层数据处理和前端界面交互相对分开。这样做便于后期维护，也方便继续扩展功能。底层识别部分集成 Ultralytics YOLO 框架，主要完成驾驶员面部区域检测和情绪类别识别；前端界面基于 PyQt5 编写，支持图像、视频和实时摄像头三种输入方式。系统还加入 CUDA 自动检测逻辑，如果设备具备 GPU 环境，则优先使用 GPU 推理，以减少模型推理延迟。

4.2. 功能模块实现

推理引擎模块 emotion_detector.py 主要负责三类工作：加载 YOLO 权重、读取数据集配置，以及封装推理接口。系统通过 predict_image 和 predict_frame 接口，把输入图像或视频帧转换为包含 BBox 坐标、置信度和类别标签的检测结果。由于模型加载和视频检测都比较耗时，程序采用多线程方式处理这些任务，避免界面在运行过程中长时间无响应。

加载线程：主要用于异步读取模型权重，减少系统启动时的界面卡顿。

检测线程：主要用于处理实时视频流，并结合跳帧采样策略，将处理速率控制在约 10 帧/秒。该设置可以保留主要情绪变化信息，同时降低 CPU 计算压力，使界面交互更加稳定。

此外，系统调用 Matplotlib 绘制情绪概率分布柱状图，并将历史检测结果保存为结构化文本，便于后续查看、对比和分析。

4.3. 界面展示与交互流程

用户在控制面板中选择输入源后，系统会先进行依赖检查，然后启动检测流程。界面右侧设置了实时画面展示区、统计信息区和文本详情区，用于显示当前情绪类别、置信度以及相关检测记录。

5. 结束语

本文针对驾驶场景下的情绪监测需求，完成了一套多模态情绪识别系统的设计与实现。系统将视频微表情、静态图像和语音文本信息结合起来，使模型在复杂环境下的识别结果更加稳定。实验结果表明，融合模型在 DMED 数据集上的准确率为 94.2%，相较单模态方法有所提升。基于 PyQt5 的交互界面能够完成实时推理和结果展示，多线程处理方式也保证了界面运行的基本流畅性。

当然，系统还存在一些不足。首先，在强光、暗光和大面积遮挡等极端情况下，模型的泛化能力还需要依靠更多真实道路数据继续验证。其次，当前多模态融合仍以特征拼接和决策融合为主，对不同模态贡献度的动态调整还不够充分。后续研究可以进一步引入自适应注意力融合方法，减少冗余信息干扰，提高复杂驾驶场景下的识别可靠性。

参考文献

[1] 王海涌, 田爱爱, 张丹. 基于多特征融合的列车司机疲劳驾驶检测[J/OL]. 计算机应用与软件: 1-10. https://link.cnki.net/urlid/31.1260.tp.20260205.1351.004, 2026-06-18.

[2] 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+37.

[3] 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.

[4] 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.

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

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

[7] Xiang, G., Yao, S., Deng, H., Wu, X., Wang, X., Xu, Q., *et al*. (2024) A Multi-Modal Driver Emotion Dataset and Study: Including Facial Expressions and Synchronized Physiological Signals. *Engineering Applications of Artificial Intelligence*, **130**, Article ID: 107772. https://doi.org/10.1016/j.engappai.2023.107772

[8] Hu, C., Gu, S., Yang, M., Han, G., Lai, C.S., Gao, M., *et al*. (2024) MDEmoNet: A Multimodal Driver Emotion Recognition Network for Smart Cockpit. 2024 *IEEE International Conference on Consumer Electronics* (*ICCE*), Las Vegas, 6-8 January 2024, 1-6. https://doi.org/10.1109/icce59016.2024.10444365