Multimodal fusion framework: comprehensive information utilization and performance optimization in driver style classification

Yuqing Lin[a], Zhen Wang*[a], Xiangmo Zhao[a], Dingrui Xue[b,a]
[a]School of Information Engineering, Chang'an University, Xi'an 710064, Shaanxi, China;
[b]Northwest Institute of Mechanical and Electrical Engineering, Xianyang 712099, Shaanxi, China

ABSTRACT

The study proposes a multimodal fusion framework termed Vision-Motion Multimodal Classification (VMMC), with the objective of addressing the inherent complexities in the classification of driver styles. This framework amalgamates visual and motion data, leveraging the harmonious interplay between the visual modality feature extraction module and the motion feature extraction module, complemented by the integration of a cross-modal attention mechanism, to achieve precise classification of driver driving styles. Through meticulous experimental evaluation, the VMMC framework demonstrates substantial advantages across metrics such as precision, recall, and F1 score, thus validating the superiority of the VMMC framework. These research findings not only provide novel perspectives on the application of multimodal fusion in driver style classification but also offer invaluable insights for a deeper understanding of driving style patterns.

Keywords: Driving style classification, multimodal fusion, cross-modal attention, vision-motion integration

1. INTRODUCTION

The examination of driving styles stands as a cornerstone in the realm of intelligent transportation and driver assistance systems, where a nuanced understanding of drivers' behaviors and risk tendencies holds significant implications for enhancing driving safety and advancing autonomous vehicle technologies[1,2]. However, traditional methodologies often fall short in capturing the complex interplay between human operators and vehicular systems, thus necessitating a paradigm shift towards holistic approaches that integrate diverse sensory modalities to unveil the intricate semantic layers embedded within driving behaviors[3-7].

Despite recent strides in multimodal fusion techniques, current methodologies frequently rely on rudimentary strategies such as feature concatenation or cascading, inadvertently constraining the full exploration of synergistic relationships and complementarities among disparate modalities[8]. Consequently, there exists a compelling need to delve deeper into the intrinsic correlations among sensory inputs, unlocking their latent potential to enrich feature representations and bolster classification accuracies.

In response to this exigency, we introduce an innovative Multimodal Driving Style Classification (VMMC) framework designed to seamlessly integrate visual and motion modalities, thereby harnessing their synergistic information to achieve refined driving style classification[9-11]. The VMMC framework transcends conventional approaches by offering several pioneering advancements:

Comprehensive Multimodal Integration: By fusing visual and kinetic signals, the VMMC framework transcends the limitations of unimodal analyses, enabling a holistic interpretation of driving behaviors that encompasses both spatial and temporal dynamics.

Dynamic Cross-Modal Attention Mechanisms: Leveraging sophisticated attention mechanisms, the VMMC framework dynamically models intermodal correlations, enabling adaptive feature selection and refinement to capture salient aspects of driving style across modalities.

Streamlined Unimodal Encoding Networks: Through tailored encoding networks for visual and kinetic modalities, the VMMC framework optimally extracts discriminative features from each modality, facilitating a synergistic fusion that enhances classification performance.

*zhenwang@chd.edu.cn

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 1339516 · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3048698

The forthcoming chapters will meticulously explore the theoretical underpinnings and technical background of our study, offering a comprehensive review of existing methodologies for driving style analysis and multimodal learning techniques (Chapter Two). Subsequently, we will delve into the theoretical principles and computational models underpinning the VMMC framework (Chapter Three), followed by a rigorous evaluation of its classification performance on publicly available datasets to substantiate its efficacy (Chapter Four). Finally, the concluding chapter will synthesize our findings, offering insights into future research trajectories and practical applications, thereby contributing to the advancement of knowledge in the domains of intelligent transportation systems and beyond.

Our work represents a significant stride towards a deeper understanding of driving behaviors and paves the way for the development of personalized driver assistance systems and the refinement of autonomous driving capabilities, with profound implications for both theory and practice in intelligent transportation.

2. RELATED WORKS

In recent years, the study of driving styles has garnered considerable attention due to its significance in enhancing road safety and the development of intelligent transportation systems[12-15]. Research efforts have focused on discerning and classifying driving behaviors with meticulous attention[16-18], a pursuit integral to refining decision-making mechanisms in autonomous vehicles and advanced driver-assistance systems.

Traditional approaches to scrutinizing driving styles have predominantly relied on intricate feature extraction techniques, often integrated into conventional machine learning paradigms such as decision trees and support vector machines. While effective to a certain extent, these methodologies exhibit limitations in generalization capabilities and necessitate significant domain expertise and human resources for optimal performance[19].

With the advent of deep learning methodologies, researchers have increasingly turned to automated feature learning approaches rooted in deep neural networks[20-23]. These approaches aim to extract discriminative feature representations directly from raw data in an end-to-end fashion, thus bypassing the need for manual feature engineering. For instance, Wüurtz and Göhringer utilized Long Short-Term Memory (LSTM) networks to encode driving style features from GPS data, showcasing the potential of recurrent neural networks in capturing temporal dependencies in driving behavior[24].

Building upon this trend, Mou et al. introduced an attention-enhanced Convolutional Neural Network (CNN) and LSTM model that integrates eye-tracking, vehicle telemetry, and environmental data to detect drivers' stress levels[25]. This multi-modal approach reflects the growing recognition of the diverse sources of information that contribute to understanding driving behavior, emphasizing the importance of holistic data fusion techniques.

Furthermore, recent studies have highlighted the significant inter-individual variability in the perception of road conditions and vehicle dynamics, underscoring the need for personalized analysis and classification of drivers. For example, Hirose et al. employed cluster analysis on time-series data including acceleration, torque, and steering wheel angle to group drivers based on their distinct driving styles, illustrating the potential for data-driven segmentation of driver populations[26].

Despite the advancements enabled by deep learning models, a singular modality may not fully capture the complexity of driving styles[27]. To address this limitation, researchers have explored multimodal fusion approaches to integrate heterogeneous information from diverse sources. Wang et al. proposed a cascaded CNN and LSTM architecture for fusing in-vehicle sensor data with video streams to detect driving styles[28], while Vaitkus.V, et al. introduced an attention-based framework for fusing video and CAN bus data, demonstrating the efficacy of leveraging complementary modalities for enhanced classification performance[29].

Nevertheless, existing methodologies are not without their constraints and opportunities for refinement. Common fusion strategies often rely on simplistic concatenation or junction approaches, potentially overlooking nuanced cross-modal correlations. Moreover, optimizing feature extraction efficiency remains a critical challenge[30], particularly for visual and kinetic modalities. To address these limitations, the proposed VMMC framework introduces cross-modal attention mechanisms and efficient feature encoding techniques, aiming to achieve accurate and robust classification of driving styles.

In summary, the landscape of driving style analysis and classification is evolving rapidly, driven by advancements in deep learning, multimodal fusion, and personalized modeling approaches. By synthesizing insights from diverse research endeavors, this paper contributes to the ongoing pursuit of enhancing traffic safety and advancing the capabilities of intelligent transportation systems.

3. METHOD

We propose a comprehensive multimodal fusion framework named VMMC, tailored for precise classification of driving styles. This framework seamlessly integrates visual and motion information, facilitating a holistic understanding of driving behaviors. Comprising a visual modality feature extraction module and a motion feature extraction module, VMMC leverages the complementary nature of visual and motion data to enrich feature representation.

For our driving style classification task, the input data encompasses two primary modalities: a video sequence $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_T] \in \mathbb{R}^{T\times H\times W\times D}$ and a vehicle motion trajectory sequence $\boldsymbol{M} = [\boldsymbol{m}_1, \boldsymbol{m}_2, \dots, \boldsymbol{m}_T] \in \mathbb{R}^{T\times d_m}$. The video sequence $\boldsymbol{V}$ captures the visual dynamics of driving, where T denotes the temporal duration, $H \times W$ represents the resolution, and D signifies the number of channels. On the other hand, the vehicle motion trajectory sequence $\boldsymbol{M}$ encapsulates features such as velocity, acceleration, and other d_m-dimensional attributes, providing insights into the dynamic behavior of the vehicle.

The architecture of the VMMC framework, depicted in Figure 1, orchestrates the seamless integration of visual and motion modalities. The visual modality feature extraction module processes the video sequence $\boldsymbol{V}$, while the motion feature extraction module analyzes the vehicle motion trajectory sequence $\boldsymbol{M}$. Finally, by employing cross-modal attention mechanisms, VMMC fuses the extracted visual and motion features to obtain a more enriched feature representation, facilitating the rational classification of driving styles.

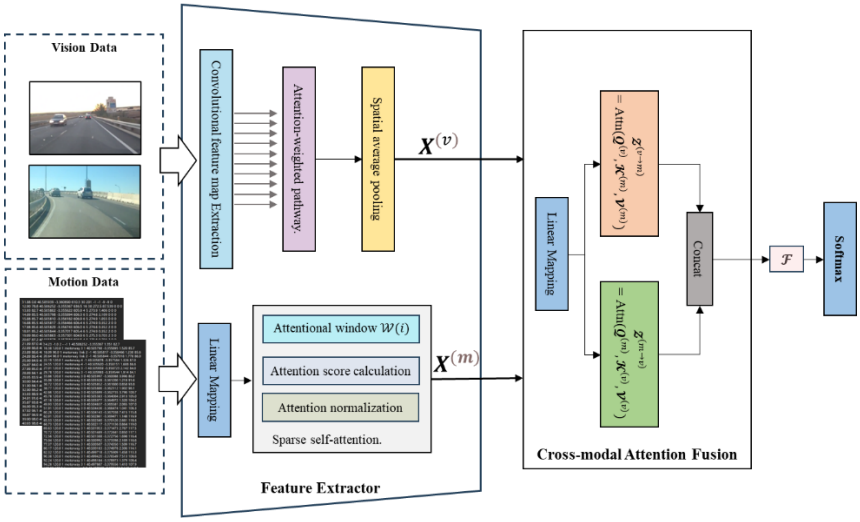

Figure 1. The architecture of the VMMC framework.

This methodological approach not only capitalizes on the inherent characteristics of visual and motion data but also emphasizes the importance of cross-modal interactions in capturing the nuanced aspects of driving behaviors. Through the VMMC framework, we aim to advance the state-of-the-art in driving style classification, contributing to enhanced traffic safety and the development of intelligent transportation systems.

(a) Visual features extraction

ECANet epitomizes a lightweight channel attention mechanism. Through seamless integration, it autonomously discerns correlations among distinct channels within feature maps, thereby enhancing network performance. Inspired by its principles, we seamlessly integrate channel attention mechanisms into our visual modality encoding process to enrich the model's representational capacity. This strategic integration amplifies the network's emphasis on pivotal features while upholding a modest level of computational complexity.

We start by assigning an attention weight α_c to each channel within the convolutional feature map. The sum of these attention weights for all channels constitutes the attention vector $\boldsymbol{\alpha}$.

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}' \odot \boldsymbol{\beta} \tag{1}$$

where the symbol $\odot$ signifies element-wise multiplication. The initial attention vector $\boldsymbol{\alpha}' \in \mathbb{R}^C$ is acquired through a

linear transformation followed by an activation function. The vector $\boldsymbol{\beta} \in \mathbb{R}^C$ represents the interaction encoding weights, derived by applying a linear transformation and activation function to the descriptor $\mathbf{v}^g$, obtained through global average pooling across each channel group.

$$\begin{cases} \boldsymbol{\alpha}' = \sigma(\boldsymbol{W}_1\boldsymbol{d} + \boldsymbol{b}_1), \mathbf{d} = \frac{1}{H \times W}\sum_{i,j} \boldsymbol{V}_{:,i,j} \\ \boldsymbol{\beta} = \sigma(\mathbf{W}_2\mathbf{v} + \boldsymbol{b}_2),\ \mathbf{v} = \text{Concat}(\mathbf{v}^g \text{ for } g = 1 \text{ to } k) \end{cases} \tag{2}$$

where, σ symbolizes the sigmoid activation function, where $\boldsymbol{W}_1$ and $\boldsymbol{b}_1$ respectively denote the weight matrix and bias term. $\boldsymbol{d}$ signifies the global descriptor, representing the average value of each channel in the spatial dimension. $\boldsymbol{W}_2$ and $\boldsymbol{b}_2$ represent the weight matrix and bias term respectively, while k indicates the number of channel groups. $\mathbf{v}^g$ denotes the global average pooling descriptor for the g-th channel group, articulated as follows:

$$\mathbf{v}^g = \frac{1}{C/k}\sum_{c=\frac{(g-1)C}{k}+1}^{\frac{gC}{k}} \frac{1}{HW}\sum_{i,j} \boldsymbol{V}_{c,i,j} \tag{3}$$

Continuing, we acquire the weighted output feature map through attention weighting:

$$\boldsymbol{Y}_c = \alpha_c \cdot \boldsymbol{V}_c, c = 1,2,\cdots,C \tag{4}$$

where $\boldsymbol{V}_c$ represents the c^{th} channel of the original input, and $\boldsymbol{Y}_c$ denotes the c^{th} channel of the weighted output feature map.

Finally, we perform average pooling on the spatial dimensions $H' \times W'$, compressing the visual features into a sequential form, obtaining the input $\boldsymbol{X}^{(v)}$ for subsequent cross-modal fusion:

$$\boldsymbol{X}^{(v)} = \frac{1}{H'W'}\sum_{h=1}^{H'} \sum_{w=1}^{W'} \boldsymbol{Y}_{:,:,h,w} \in \mathbb{R}^{T \times C} \tag{5}$$

(b) Motion features extraction

For the series of vehicle motion trajectories $\boldsymbol{M} = [\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_T] \in \mathbb{R}^{T \times d_m}$, where T denotes the sequence length and d_m represents the feature dimension, we apply a linear mapping to the input sequence $\boldsymbol{M}$ to derive the query matrix $\boldsymbol{Q}$, the key matrix $\boldsymbol{K}$, and the value matrix $\boldsymbol{V}$.

$$\boldsymbol{Q} = \boldsymbol{M}(\boldsymbol{W}_Q^{(m)})^T, \boldsymbol{K} = \boldsymbol{M}(\boldsymbol{W}_K^{(m)})^T, \boldsymbol{V} = \boldsymbol{M}(\boldsymbol{W}_V^{(m)})^T \tag{6}$$

where $\boldsymbol{W}_Q^{(m)}, \boldsymbol{W}_K^{(m)}, \boldsymbol{W}_V^{(m)} \in \mathbb{R}^{d_a \times d_m}$ represent the learnable linear transformation parameters, with d_a denoting the attention dimensionality.

To optimize the management of extensive sequence inputs while upholding minimal computational demands, we drew inspiration from the sparse self-attention mechanism introduced in Longformer. This approach involves selectively retaining attention scores solely between designated positions within the sequence. We define a variable-sized attention window $\mathcal{W}(i)$:

$$\mathcal{W}(i) = \{j \in [1,n]: |i-j| < w\} \cup \{p_1(i), p_2(i), \ldots, p_m(i)\} \tag{7}$$

where, w denotes the dimensionality of the attention window. $p_k(i)$ represents the indices of an additional set of m global positions that are height-wise associated with position i.

When computing the attention score matrix $\boldsymbol{A}$, we only compute the score $\boldsymbol{A}_{ij}$ for positions (i,j) where $j \in \mathcal{W}(i)$.

$$\boldsymbol{A}_{ij} = \begin{cases} \frac{(\boldsymbol{Q}_i \boldsymbol{K}_j^T)}{\sqrt{d}}, j \in \mathcal{W}(i) \\ -\infty, \quad j \notin \mathcal{W}(i) \end{cases} \tag{8}$$

where $\boldsymbol{Q} \in \mathbb{R}^{T \times d_a}$ and $\boldsymbol{K} \in \mathbb{R}^{T \times d_a}$.

The attention score matrix $\boldsymbol{A} \in \mathbb{R}^{T \times T}$ is obtained through computation, and after applying the softmax operation, we obtain the normalized attention matrix $\bar{A}_{ij}$:

$$\bar{A}_{ij} = \frac{e^{A_{ij}}}{\sum_{k=1}^{T} e^{A_{ik}}} \tag{9}$$

The ultimate encoded sequence is illustrated as follows:

$$\boldsymbol{X}_{ij}^{(m)} = \sum_{k=1}^{T} \bar{A}_{ik} \cdot V_{kj} \tag{10}$$

where $\boldsymbol{V} \in \mathbb{R}^{T \times d_a}$ denotes a numerical matrix, $\boldsymbol{X}_{ij}^{(m)}$ signifies the element located at the i^{th} row and j^{th} column of the sequence post encoding through motion modality.

(c) Cross-modal fusion

The visual modality input feature sequence is denoted as $\boldsymbol{X}^{(v)} = \{x_1^{(v)}, x_2^{(v)}, \ldots, x_T^{(v)}\} \in \mathbb{R}^{d_v}$, and the motion modality input feature sequence is denoted as $\boldsymbol{X}^{(m)} = \{x_1^{(m)}, x_2^{(m)}, \ldots, x_T^{(m)}\} \in \mathbb{R}^{d_m}$, where T is the sequence length, and d_v and d_m are the dimensions of the visual and motion features respectively.

To optimize multimodal fusion, we utilize linear mapping on the input feature sequence to generate query $\boldsymbol{Q}^{(\cdot)}$, key $\boldsymbol{\mathcal{K}}^{(\cdot)}$, and value $\boldsymbol{\mathcal{V}}^{(\cdot)}$ sequences:

$$\begin{cases} \boldsymbol{Q}^{(v)} = \boldsymbol{X}^{(v)}(\boldsymbol{W}_Q^{(v)})^T & \in \mathbb{R}^{T \times d_q} \\ \boldsymbol{\mathcal{K}}^{(v)} = \boldsymbol{X}^{(v)}(\boldsymbol{W}_K^{(v)})^T & \in \mathbb{R}^{T \times d_k} \\ \boldsymbol{\mathcal{V}}^{(v)} = \boldsymbol{X}^{(v)}(\boldsymbol{W}_V^{(v)})^T & \in \mathbb{R}^{T \times d_v} \\ \boldsymbol{Q}^{(m)} = \boldsymbol{X}^{(m)}(\boldsymbol{W}_Q^{(m)})^T & \in \mathbb{R}^{T \times d_q} \\ \boldsymbol{\mathcal{K}}^{(m)} = \boldsymbol{X}^{(m)}(\boldsymbol{W}_K^{(m)})^T & \in \mathbb{R}^{T \times d_k} \\ \boldsymbol{\mathcal{V}}^{(m)} = \boldsymbol{X}^{(m)}(\boldsymbol{W}_V^{(m)})^T & \in \mathbb{R}^{T \times d_v} \end{cases} \tag{11}$$

where, $\boldsymbol{W}_Q^{(\cdot)}, \boldsymbol{W}_K^{(\cdot)}$, and $\boldsymbol{W}_V^{(\cdot)}$ are trainable parameters for linear transformations in $\mathbb{R}^{d_\cdot \times d_\cdot}$, where d_q and d_k represent the dimensions of the query and key, respectively, and d_v represents the dimension of the value.

We calculate cross-modal attention:

$$\begin{cases} \boldsymbol{\mathcal{Z}}^{(v \to m)} = \mathrm{Attn}(\boldsymbol{Q}^{(v)}, \boldsymbol{\mathcal{K}}^{(m)}, \boldsymbol{\mathcal{V}}^{(m)}) \in \mathbb{R}^{T \times d_v} \\ \boldsymbol{\mathcal{Z}}^{(m \to v)} = \mathrm{Attn}(\boldsymbol{Q}^{(m)}, \boldsymbol{\mathcal{K}}^{(v)}, \boldsymbol{\mathcal{V}}^{(v)}) \in \mathbb{R}^{T \times d_q} \end{cases} \tag{12}$$

where, $\boldsymbol{\mathcal{Z}}^{(v \to m)}$ designates the attentional representation from the visual modality to the motion modality, while $\boldsymbol{\mathcal{Z}}^{(v \to m)}$ designates the attentional representation from the motion modality to the visual modality. The term $\mathrm{Attn}(\cdot)$ refers to the scaled dot-product attention function, defined as:

$$\begin{aligned} \mathrm{Attn}\left(\boldsymbol{Q}^{(\cdot)}, \boldsymbol{\mathcal{K}}^{(\cdot)}, \boldsymbol{\mathcal{V}}^{(\cdot)}\right) &= \mathrm{softmax}\left(\frac{\boldsymbol{Q}^{(\cdot)}{\boldsymbol{\mathcal{K}}^{(\cdot)}}^T}{\sqrt{d_k}}\right)\boldsymbol{\mathcal{V}}^{(\cdot)} \\ &= \sum_{j=1}^{T} \gamma_{i,j}\mathbf{v}_j \end{aligned} \tag{13}$$

In the equation, $\gamma_{i,j} = \frac{\exp(q_i \cdot k_j)}{\sum_{l=1}^{T} \exp(q_i \cdot k_l)}$, where q_i, k_j, and v_j denote the i^{th} row and j^{th} column of $\boldsymbol{Q}^{(\cdot)}, \boldsymbol{\mathcal{K}}^{(\cdot)}$, and $\boldsymbol{\mathcal{V}}^{(\cdot)}$ respectively.

The final multimodal fusion feature representation $\boldsymbol{\mathcal{F}}$ is:

$$\boldsymbol{\mathcal{F}} = \mathrm{Concat}\left(\boldsymbol{X}^{(v)}, \boldsymbol{\mathcal{Z}}^{(m \to v)}, \boldsymbol{X}^{(m)}, \boldsymbol{\mathcal{Z}}^{(v \to m)}\right) \in \mathbb{R}^{T \times (d_v + d_q + d_m + d_v)} \tag{14}$$

(d) Loss function

The classifier predicts the driving style $\hat{\boldsymbol{y}}$ based on the multimodal fusion representation $\boldsymbol{\mathcal{F}}$.

$$\hat{\boldsymbol{y}} = \frac{\exp(\boldsymbol{\mathcal{F}}_{ij})}{\sum_{k=1}^{N} \exp(\boldsymbol{\mathcal{F}}_{ik})} \tag{15}$$

where $i \in \{1,2,\ldots,T\}$, $j \in \{1,2,\ldots,(d_v + d_q + d_m + d_v)\}$.

We define the total loss function $\mathcal{L}$ as:

$$\mathcal{L} = \mathcal{L}_{\text{style}} + \lambda_a \mathcal{L}_{\text{att}} + \lambda_r \mathcal{L}_{\text{reg}} \tag{16}$$

in this specified context, $\mathcal{L}_{\text{style}}$ represents the cross-entropy loss utilized in the classification of driving styles: $\mathcal{L}_{\text{style}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$. $\mathcal{L}_{\text{att}}$ embodies the entropy regularization term applied to attention: $\mathcal{L}_{\text{att}} = -\sum_{i=1}^{T}\sum_{j=1}^{T} \alpha_{i,j} \log \alpha_{i,j}$. Furthermore, $\mathcal{L}_{\text{reg}}$ indicates the regularization term associated with the kernel norm: $\mathcal{L}_{\text{reg}} = \lambda_r |\boldsymbol{\Theta}|_F^2$. In this instance, N denotes the number of style categories, $\boldsymbol{\Theta}$ encompasses the complete set of model parameters, and λ_a and λ_r serve as the weighting coefficients for each loss term.

4. EXPERIMENT

In this section, we will offer a comprehensive elucidation of our implementation particulars and carry out experiments on a publicly accessible dataset to evaluate the effectiveness of our proposed VMMC framework.

(a) Dataset and setup

The UAH-Driveset stands as a prominent multimodal driving dataset that has garnered significant attention from the research community[31,32]. Painstakingly curated by the research team at the University of Alcalá (UAH) in Spain, this dataset has earned acclaim for its rich array of driving scenarios and detailed style annotations. Encompassing various real-world driving situations, including urban roads and highways, the UAH-Driveset provides a wealth of multimodal information, ranging from drivers' biometric characteristics to vehicle state data, video recordings, and vehicle sensor data. These datasets serve as invaluable resources for delving into the intricacies of driving style patterns and driving styles.

In this dataset, we elected to extract features pertaining to vehicle velocity, acceleration, rate of acceleration change, steering dynamics, vehicle positioning, temporal intervals, and traffic volume, resulting in 15 feature subsets. The objective was to capture a diverse array of driving style patterns and nuances in driving style. We opted for a batch size of 16, accompanied by 500 epochs, and implemented 6-fold cross-validation. The Adam optimizer was employed, with the learning rate set to 1×10^{-4}. Following data preprocessing, we partitioned the sequential data with a 60% overlap rate and extracted features from each segment. The sample allocation ratio was 6:2:2 for training, validation, and testing, respectively.

(b) Performance and analysis

We evaluated our approach using the Precision (P), Recall (R), and F1 metrics[33,34]:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}, \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}, \mathrm{F1} = 2 \times \frac{\mathrm{P \times R}}{\mathrm{P+R}} \tag{17}$$

where TP represents the quantity of true positives, FP denotes the quantity of false positives, and FN signifies the quantity of false negatives.

We compared the VMMC method with established classical approaches, namely CNN, LSTM, CNN-LSTM, FCN-LSTM, and GRU.

Table 1. The effectiveness of varied methodologies in the classification of driving styles.

Method	P	R	F1
CNN	0.6125	0.6100	0.6112
LSTM	0.8894	0.8911	0.8902
GRU	0.9097	0.9125	0.9111
CNN-LSTM	0.9568	0.9597	0.9582
FCN-LSTM	0.9541	0.9601	0.9571
VMMC	0.9705	0.9762	0.9733

Table 1 illustrates the performance of six distinct algorithmic methodologies in the classification of driving styles. Conventional algorithms such as CNN, LSTM, and GRU rely solely on visual or motion cues, thereby constraining their ability to comprehensively exploit data. Consequently, their performance notably lags behind more integrated approaches such as CNN-LSTM, FCN-LSTM, and VMMC. Particularly noteworthy is VMMC, which, leveraging its multimodal

fusion framework, integrates visual and motion cues, thus enhancing the model's capacity to capture the diversity of driving styles. Moreover, VMMC introduces a cross-modal attention mechanism, facilitating the seamless integration of visual and motion features. Discerning their interdependencies and dynamically balancing the significance of various modalities, enhances the effectiveness of feature fusion and representation learning. Conversely, algorithms lacking such mechanisms may struggle to effectively integrate heterogeneous information sources, potentially underscoring the superiority of VMMC over CNN-LSTM, FCN-LSTM, and analogous algorithms.

Benefiting from the aforementioned advantages, VMMC outperforms other algorithms across all metrics, achieving precision, recall, and F1 scores of 0.9705, 0.9762, and 0.9733 respectively. These scores represent improvements of 1.72%, 1.68%, and 1.69% over the second-best method.

(c) Ablation experiment

To ascertain the effectiveness of each component, we established three control groups for experimentation, as shown in Figure 2, and specifically:

VMMC-NoV: Omitting the visual feature extraction module from the original framework aimed to underscore the importance of visual information in driving style classification.

VMMC-NoM: Removing the motion feature extraction module from the original framework aimed to underscore the significance of visual information in driving style classification.

VMMC-NoMu: Eliminating the cross-modal attention mechanism resulted in straightforward processing of visual and motion features before concatenation, with the objective of validating the efficacy of the cross-modal attention mechanism.

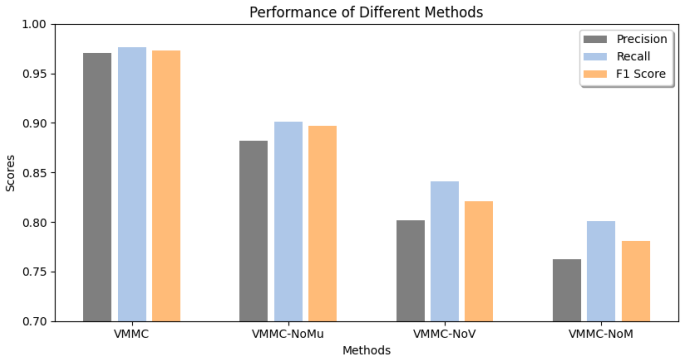

Figure 2. The Performance of VMMC, VMMC-NoMu, VMMC-NoV, and VMMC-NoM.

The VMMC demonstrates superior performance across all assessment metrics, indicating the effectiveness of the comprehensive model in seamlessly integrating visual and motion data while augmenting representational capacity through cross-modal attention mechanisms, thereby significantly enhancing the precision of driving style classification. Despite a slight decline in performance upon the removal of the cross-modal attention mechanism, it still surpasses the results of other ablative experiments. This further underscores the importance of both visual and motion attributes in driving style classification. The removal of the visual feature extraction module, relying solely on motion features for classification, leads to a noticeable decrease in performance compared to the complete model. However, it surpasses the performance of removing the motion feature extraction module, suggesting that for the task of driving style classification, the information provided by our motion feature extraction module exceeds that of the visual feature extraction module.

5. CONCLUSION

This paper introduces the VMMC framework, which seeks to achieve a precise classification of driving styles. Through the integration of visual and motion information and the enhancement of feature representation via cross-modal attention mechanisms, this framework presents an efficient solution for the classification of driver style.

In the VMMC framework, we initiated the extraction of features from both visual and motion modalities. The visual

aspect employed a lightweight channel attention mechanism, while the motion component utilized a sparse self-attention mechanism to manage lengthy sequence inputs. Subsequently, through multimodal fusion, we adeptly integrated visual and motion data to achieve a more comprehensive feature representation. Of particular significance, we introduced a cross-modal attention mechanism to dynamically discern and reconcile the interconnections among diverse modalities, thereby enhancing the effectiveness of feature fusion.

The experimental findings illustrate that the VMMC framework displays notable advantages in the realm of driver-style classification. In contrast to conventional methodologies and alternative comprehensive approaches, VMMC attains superior performance metrics encompassing precision, recall, and F1 score.

ACKNOWLEDGEMENTS

This work was partially supported by the National Key R&D Program of China (No. 2021YFB2501204), National Natural Science Foundation of China (No. 52202488, U23A20682), Key Research and Development Plan of Shaanxi Province (No. 2022GY-309), Young Talent fund of University Association for Science and Technology in Shaanxi, China (No. 20220111), Young science and technology Talent fund of Shaanxi, China (No. 2024ZC-KJXX-025), Fundamental Research Funds for the Central Universities, CHD (300102244201), Taishan industry-leading talents project.

REFERENCES

[1] Deng, Z., Chu, D., Wu, C., et al., "A probabilistic model for driving-style-recognition-enabled driver steering behaviors," IEEE Transactions on Systems, Man, and Cybernetics: Systems, 52(3), 1838-1851 (2020).
[2] Martinelli, F., Mercaldo, F., Orlando, A., et al., "Human behavior characterization for driving style recognition in vehicle system," Computers & Electrical Engineering, 83, 102504 (2020).
[3] Zhang, C., Wang, W., Chen, Z., et al., "Shareable driving style learning and analysis with a hierarchical latent model," IEEE Transactions on Intelligent Transportation Systems, (2024).
[4] Shahverdy, M., Fathy, M., Berangi, R., et al., "Driver behavior detection and classification using deep convolutional neural networks," Expert Systems with Applications, 149, 113240 (2020).
[5] Gao, B., Cai, K., Qu, T., et al., "Personalized adaptive cruise control based on online driving style recognition technology and model predictive control," IEEE Transactions on Vehicular Technology, 69(11), 12482-12496 (2020).
[6] McDonald, A. D., Ferris, T. K. and Wiener, T. A., "Classification of driver distraction: A comprehensive analysis of feature generation, machine learning, and input measures," Human Factors, 62(6), 1019-1035 (2020).
[7] Azadani, M. N. and Boukerche, A., "Driving behavior analysis guidelines for intelligent transportation systems," IEEE Transactions on Intelligent Transportation Systems, 23(7), 6027-6045 (2021).
[8] Jia, S., Hui, F., Li, S., et al., "Long short-term memory and convolutional neural network for abnormal driving behaviour recognition," IET Intelligent Transport Systems, 14(5), 306-312 (2020).
[9] Song, X., Yin, Y., Cao, H., et al., "The mediating effect of driver characteristics on risky driving behaviors moderated by gender, and the classification model of driver's driving risk," Accident Analysis & Prevention, 153, 106038 (2021).
[10] Xu, J., Pan, S., Sun, P. Z. H., et al., "Human-factors-in-driving-loop: Driver identification and verification via a deep learning approach using psychological behavioral data," IEEE Transactions on Intelligent Transportation Systems, 24(3), 3383-3394 (2022).
[11] Chu, H., Zhuang, H., Wang, W., et al., "A review of driving style recognition methods from short-term and long-term perspectives," IEEE Transactions on Intelligent Vehicles, (2023).
[12] Chen, Y., Wang, K. and Lu, J. J., "Feature selection for driving style and skill clustering using naturalistic driving data and driving behavior questionnaire," Accident Analysis & Prevention, 185, 107022 (2022).
[13] Zhang, C., Wang, W., Chen, Z., et al., "Shareable driving style learning and analysis with a hierarchical latent model," IEEE Transactions on Intelligent Transportation Systems, (2024).
[14] Wang, J., Li, W., Li, F., et al., "100-driver: a large-scale, diverse dataset for distracted driver classification,"

IEEE Transactions on Intelligent Transportation Systems, (2023).
[15] Huang, Y., Du, J., Yang, Z., et al., "A survey on trajectory-prediction methods for autonomous driving," IEEE Transactions on Intelligent Vehicles, 7(3), 652-674 (2022).
[16] Ma, Z. and Zhang, Y., "Driver-automated vehicle interaction in mixed traffic: Types of interaction and drivers' driving styles," Human Factors, 66(2), 544-561 (2024).
[17] Wang, W., Qie, T., Yang, C., et al., "An intelligent lane-changing behavior prediction and decision-making strategy for an autonomous vehicle," IEEE Transactions on Industrial Electronics, 69(3), 2927-2937 (2021).
[18] Li, G., Chen, Y., Cao, D., et al., "Extraction of descriptive driving patterns from driving data using unsupervised algorithms," Mechanical Systems and Signal Processing, 156, 107589 (2021).
[19] Xing, Y., Lv, C., Wang, H., et al., "An ensemble deep learning approach for driver lane change intention inference," Transportation Research Part C: Emerging Technologies, 115, 102615 (2020).
[20] Zhang, C., Wang, W., Chen, Z., et al., "Shareable driving style learning and analysis with a hierarchical latent model," IEEE Transactions on Intelligent Transportation Systems, (2024).
[21] Wang, K., Qu, D., Yang, Y., et al., "Risk-quantification method for car-following behavior considering driving-style propensity," Applied Sciences, 14(5), 1746 (2024).
[22] Mohammed, K., Abdelhafid, M., Kamal, K., et al., "Intelligent driver monitoring system: An internet of things-based system for tracking and identifying the driving behavior," Computer Standards & Interfaces, 84, 103704 (2023).
[23] Elander, J., West, R. and French, D., "Styleal correlates of individual differences in road-traffic crash risk: An examination of methods and findings," Psychol. Bull., 113, 279 (1993).
[24] W¨urtz, S. and G¨ohner, U., "Driving style analysis using recurrent neural networks with LSTM cells," J Adv Inf Technol, 1–9 (2020). https://doi.org/10.12720/jait.11.1.1-9
[25] Mou, L., Zhou, C., Zhao, P., et al, "Driver stress detection via multimodal fusion using attention-based CNN-LSTM," Expert Syst Appl, 173(114), 693 (2021). https://doi.org/10/gkxx56
[26] Hirose, T., Oguchi, Y., and Sawada, T., "Framework of tailormade driving support systems and neural network driver model," IATSS Res., 28(1), 108–114, (2004).
[27] Moreira-Matias, L. and Farah, H., "On developing a driver identification methodology using in-vehicle data recorders," IEEE Trans Intell Transp Syst, 18(9), 2387–2396 (2017). https://doi.org/10/gbwkgg
[28] Wang, W., Xi, J., Chong, A., et al., "Driving style classification using a semisupervised support vector machine," IEEE Transactions on Human-Machine Systems, 47(5), 650-660 (2017).
[29] Vaitkus, V., Lengvenis, P. and Žylius, G., "Driving style classification using long-term accelerometer information," 2014 19th International Conference on Methods and Models in Automation and Robotics (MMAR), IEEE, 641-644 (2014).
[30] Queiroz, R., Sharma, D., Caldas, R., et al., "A driver-vehicle model for ADS scenario-based testing," IEEE Transactions on Intelligent Transportation Systems, (2024).
[31] Romera, E., Bergasa, L. M., Arroyo, R., "A real-time multi-scale vehicle detection and tracking approach for smartphones," ITSC, 1298-1303 (2015).
[32] Bergasa, L. M., Almería, D., Almazán, J., et al., "Drivesafe: An app for alerting inattentive drivers and scoring driving behaviors," 2014 IEEE Intelligent Vehicles symposium proceedings. IEEE, 240-245 (2014).
[33] Khodairy, M. A. and Abosamra, G., "Driving behavior classification based on oversampled signals of smartphone embedded sensors using an optimized stacked-LSTM neural networks," IEEE Access, 9, 4957–4972 (2021). https://doi.org/10/gmv3r7
[34] Xie, J., Hu, K., Li, G., et al., "CNN-based driving maneuver classification using multi-sliding window fusion," Expert Syst Appl, 169, 114,442 (2021). https://doi.org/10/gmwv7j