Emotion Recognition and Multimodal Fusion in Smart Cockpits: Advancements, Challenges, and Future Directions

Dengwu Chen [1], *Haoze Li* [2], *Zixuan Zhu* [3*]

[1]Fuzhou No.8 Middle School, Fuzhou, China
[2]School of electronics and control engineering, Chang'an University,Xi'an, China
[3]Wuhan Britain- China School, Wuhan, China

Abstract. This paper reviews the latest developments in emotion recognition and multimodal fusion technologies for smart cockpits, a critical area in the evolution of intelligent transportation systems. With the convergence of advanced sensing, machine learning, and human–computer interaction, smart cockpits have evolved from basic interfaces to comprehensive platforms capable of interpreting a driver's emotional state. The study examines the theoretical foundations—such as affective computing, multimodal data fusion, and deep learning algorithms—that underpin current systems. It provides a detailed analysis of various fusion strategies, including early, late, and hierarchical fusion, and discusses innovative models like Transformer-based methods and contrastive learning approaches that have significantly enhanced the precision of emotion recognition. Despite notable progress, challenges remain, particularly in achieving robust adaptability in dynamic driving environments and reconciling heterogeneous data from diverse modalities. The paper also highlights concerns regarding privacy, data misuse, and hardware adoption barriers in mid- to low-end vehicle models. Through a systematic review of current methodologies and system architectures, the study identifies critical gaps in existing technologies and offers insights into optimization strategies such as adaptive weight adjustment, deep learning–based optimization, and continuous learning approaches. As breakthroughs in AI and sensor technologies continue, multimodal fusion in smart cockpits is poised to deliver safer and more personalized driving experiences.

1 Introduction

At a time when the smart transportation system is booming, the smart cockpit, as the core interaction hub of intelligent networked vehicles, is undergoing profound changes. With the continuous improvement of automobile intelligence and Internet connectivity, the intelligent cockpit is no longer just a simple driving space, but a comprehensive intelligent platform that integrates a variety of advanced technologies and integrates driving control, comfort

* huaxi@ldy.edu.rs

experience, infotainment and other multi-functions. This transformation not only changes the way people interact with the car, but also brings infinite possibilities for future travel.

Highly integrated information sharing and control coordination is the core of the development of smart transportation systems. Intelligent Internet-connected vehicles, with their driverless driving capability, have become the key unit in this system, while smart cockpits are the centralized application area for all kinds of new technologies [1]. In this development process, human-computer interaction (HCI) technology has become the key support for intelligent cockpits. From the early manual control and graphical user interface to today's intelligent multimodal interaction, every advancement in human-computer interaction technology has significantly improved the level of cockpit intelligence. By acquiring multimodal spatial and temporal data, the intelligent cockpit is able to deeply understand the state and intention of the vehicle. Those inside the vehicle and external road users, and then provide multi-dimensional interaction and decision-making services to improve the accuracy of driving control, ride comfort and richness of entertainment experience.

In driving scenarios, emotions have a profound impact on driving safety and experience. The integration of emotional computing technology has injected new vitality into the development of intelligent cockpits. It enables smart cars to recognize user emotions and make decisions based on emotional dynamics, enhancing the system's understanding and responsiveness to the user [2].

This paper will focus on the emotion recognition and multimodal fusion technologies in smart cockpit. It elaborates the principles, algorithms and models of these technologies, analyzes the system architecture, optimization strategies and practical applications of these technologies in smart cockpits, analyzes the current challenges, and looks forward to the future development direction. The aim is to provide comprehensive theoretical support and practical reference for the further development of intelligent cockpit technology, and to promote greater breakthroughs in the enhancement of driving experience and safety in intelligent cockpits.

2 Relevant Theories

2.1 Theoretical Foundations

The application of emotion recognition techniques and multimodal recognition theories in smart cockpits relies on theoretical foundations in several subject areas, which provide a solid foundation for emotion recognition and drive its application and development in real-world scenarios. The following are the main theoretical foundations in this field, as shown Table 1.

Table 1 The main theoretical foundations in this field

Serial number	Theory name	Description
1	Affective computing theory	Recognizing, processing, and modeling human emotions through computer science methods and techniques.
2	Theory of multimodal data fusion	Combine data from diverse sensors or various sources in order to enhance the performance and

		resilience of the system.
3	Biological Signal Processing Theory	Bioelectrical signals (e.g., heart rate, electrical skin activity, etc.) are captured, processed, and interpreted to infer an individual's emotional state.
4	Machine Learning and Deep Learning Algorithms	Using large amounts of data to train the model and automatically learn features from the data to achieve recognition of complex patterns.
5	Human-computer interaction design principles	Designing easy-to-use and natural human-machine interfaces ensures that the technology is seamlessly integrated into the user's daily life.

2.2 Algorithms and Models

Algorithms and models for emotion recognition and multimodal fusion are current hot research directions in artificial intelligence and affective computing. By integrating information from multiple modalities (e.g., text, speech, image, etc.), multimodal emotion recognition can capture the complexity and diversity of human emotions more comprehensively, thereby enhancing the precision of emotion recognition. The following is a detailed analysis from the aspects of algorithms and models.
Multimodal fusion:

Algorithms and models for emotion recognition and multimodal fusion are current hot research directions in artificial intelligence and affective computing. By integrating information from multiple modalities (e.g., text, speech, image, etc.), multimodal emotion recognition can capture the complexity and diversity of human emotions more comprehensively, so it leads to an improvement in the accuracy of emotion recognition. The following is a detailed analysis from the aspects of algorithms and models.

Early fusion: Features from different modalities are directly spliced or fused at the feature extraction stage. For example, speech and text features are extracted using MFCC features and Glove word vectors, respectively, and then fused via a bidirectional gated recurrent unit (BiGRU) [3].

Late fusion: The prediction results of different modalities are weighted or voted in the decision-making stage. For example, proposes a multiplication-based fusion method that dynamically adjusts the weights of each modality in response to noise [4].

Hierarchical Fusion: Fusion is performed at multiple levels, combining the advantages of early and late fusion. For example, proposes a approach of combined modal fusion and graph comparative learning that captures both global contextual features and local unimodal features [5].

Multion emotion reognition model: In recent years, deep learning and the Transformer architecture have made significant progress in multimodal emotion recognition:

Transformer: Transformer excels in multimodal sentiment recognition with its powerful feature extraction capabilities and parallel computing advantages. For example, provides a

comprehensive review of Transformer-based multimodal sentiment recognition methods, pointing out its advantages in capturing cross-modal dependencies [6].
Contrastive learning: Contrast learning improves the feature representation of a model by maximizing the distance between positive sample pairs and minimizing negative sample pairs. For example, proposes a joint modal fusion and graph contrast learning approach that effectively mitigates the oversmoothing problem [5].

3 System Analysis and Application Research

3.1 Algorithm System Analysis

The algorithmic system of emotion recognition and multimodal fusion theory has made remarkable progress in recent years and has become one of the core research trends in affective computing. Among them, the application of deep learning in affective computing, effective multimodal information fusion, and contextual information perception constitute its algorithmic system, which is analyzed as follows:

Multi-task Learning and Attention Mechanism: The Multi-modal Fusion Emotion Recognition Method (MTL-BAM) based on Multi-task Learning and Attention Mechanism (MTL-BAM) achieves more efficient emotion recognition by combining emotion labels from multiple modalities.This approach can fully utilize the complementary information between different modalities to improve recognition accuracy and robustness [7].

Transformer-based Multimodal Fusion Algorithm: for example, the TMFER (Transformer-based Multimodal Fusion Emotion Recognition) algorithm, which fuses a variety of information, such as speech, and images, employs the BERT model for pre-training processing, as well as MFCCfeature extraction and ResNet network, thus improving the accuracy and information richness of emotion recognition [8].

Hybrid fusion methods: for example, certain studies have fused audio and visual modalities based on linear mapping in hidden space, followed by multi-task network training, by which the accuracy and generalization of emotion recognition have been improved [9].

3.2 System Optimization Strategies

The core of multimodal fusion lies in integrating information from different modalities to provide a more comprehensive and accurate analysis for emotion recognition. For example, a person's emotional state can be better determined by combining facial expressions and voice signals, and this fusion is more reliable and sensitive than single-modal information processing. Therefore, research on multimodal fusion aims to compensate for the shortcomings of single-modal analysis and provide higher accuracy and robustness for emotion recognition. Optimizing a multimodal emotion recognition system involves many challenges, including synergy between data modalities, effective fusion of information, and tracking of emotion changes in dynamic environments. To address these challenges, multiple studies have proposed different solutions.

Modal synergy and weight adjustment: The use of hybrid optimally adapted multilayer residual attention network can effectively adjust the weights for different signals to ensure that the contribution of each modality in the fusion process is maximized [10].

Deep learning based optimization: In recent years, deep learning models have been increasingly used in emotion recognition. The Attention mechanism is widely used in multimodal fusion to recognize emotions efficiently and minimize information loss [11].

Continuous learning and adaptive strategies: Self-Supervised Learning and Adaptive Adjustment Strategies Show Potential to Improve Model Flexibility and Environmental Adaptation in Emotion Recognition [12].

3.3 Practical Applications in Smart Cockpits

Emotion recognition systems have demonstrated real-world benefits in a variety of applications, including human-computer interaction, autonomous driving and driver monitoring, etc.

Human-computer interaction: multimodal emotion recognition significantly improves the naivety and friendliness of human-computer interaction. For example, in applications such as intelligent assistants and social robots, more humanized services can be provided by recognizing the user's emotional state.

Autonomous Driving and Driver Monitoring: In autonomous driving and intelligent driver assistance systems, multimodal emotion recognition techniques are used to supervise the driver's emotional and attentional condition to enhance driving process.

4 Challenges

4.1 Enhancing Synergy in Complex Scenarios

In complex real-world environments, human behavior exhibits a high degree of variability and unpredictability. Even when similar behaviors are observed, they can stem from entirely different underlying intentions. For example, a driver’s slight frown may indicate concentration, frustration, or even confusion, depending on the situational context. Current algorithms and technological frameworks, however, often struggle to differentiate these subtle nuances accurately. The inherent challenge lies in the fact that human emotions and decision-making processes are deeply influenced by an interplay of contextual factors, personal experiences, and individual habits. These factors are not static; they evolve over time as people learn, adapt, and change their preferences in response to new experiences and external influences.

One of the major limitations of existing emotion recognition systems is their reliance on static models that do not fully account for the dynamic nature of human behavior. Many current systems are trained on datasets that capture snapshots of emotional expressions under controlled conditions. While these datasets provide valuable insights into common facial expressions, they do not always reflect the fluid and context-dependent emotional states that occur in real-world driving scenarios. Moreover, the algorithms deployed in these systems often use fixed rules or pre-defined models that lack the flexibility to adjust to rapid changes in driver behavior. For instance, when faced with a sudden change in road conditions or unexpected events, a driver’s emotional response might differ significantly from the responses captured in training datasets.

To address these challenges, it is crucial to promote the flexibility of interaction models in smart cockpits. Future research must focus on developing adaptive algorithms that can continuously learn and update themselves based on new data and real-time feedback. Techniques such as reinforcement learning and online learning could enable systems to adjust their predictions as they encounter new behavioral patterns. Additionally, the integration of contextual data—such as vehicle speed, road type, and weather conditions—can provide valuable insights that help interpret the subtleties of driver emotions more accurately. Achieving a higher level of synergy between these diverse data streams and the emotion recognition model is essential to ensure that the system not only captures immediate facial expressions but also understands the broader context in which these expressions occur. This would enable smart cockpit systems to predict driver intentions more accurately and provide timely interventions, thereby enhancing overall driving safety and comfort. However, promoting such flexibility and synergy remains a formidable challenge that requires significant advancements in both algorithm design and data integration techniques [1].

4.2 Privacy and Data Misuse Concerns

Another critical challenge in the development of emotion recognition systems for smart cockpits is the protection of user privacy and the prevention of data misuse. Emotion recognition inherently involves the collection and analysis of highly sensitive personal data, including facial images, voice recordings, and biometric signals. The potential misuse of this information poses serious risks, ranging from unauthorized surveillance to the illegal sale of personal data. In an era where data breaches and privacy violations are increasingly common, ensuring the security and proper use of sensitive information is paramount.

Current systems often rely on centralized data processing, where large volumes of personal data are transmitted to and processed in remote servers. This approach not only increases the risk of data exposure but also raises concerns about consent and the ownership of personal information. For instance, if an automobile company were to record and store detailed emotional data of drivers without proper safeguards, it could lead to scenarios where this data is accessed by unauthorized parties or sold to third parties for profit. Such actions would not only undermine public trust but could also result in severe legal and ethical repercussions.

To address these concerns, future research must prioritize the implementation of robust privacy-preserving mechanisms in emotion recognition systems. Techniques such as federated learning, where data is processed locally on the device rather than being sent to a central server, can significantly reduce the risk of data breaches. Additionally, differential privacy techniques can be incorporated to ensure that individual data points cannot be reverse-engineered from aggregated results. Leveraging the powerful inference frameworks of large-scale neural network models while maintaining strict privacy controls is a new and challenging research frontier. By integrating these privacy-enhancing technologies, researchers can drive the model to better understand vehicle users' emotions while simultaneously defending user privacy. This dual objective is essential not only for regulatory compliance but also for fostering public confidence in smart cockpit systems [1].

Moreover, establishing clear and transparent data governance policies is crucial. Users should be fully informed about what data is being collected, how it is used, and the measures in place to protect it. Consent mechanisms need to be robust and easily understandable, ensuring that users retain control over their personal information. Addressing privacy and data misuse concerns is not only a technical challenge but also a social and ethical imperative that must be tackled holistically by combining advanced technological solutions with comprehensive policy frameworks.

4.3 Adoption Barriers in Low-End and Mid-Range Models

For researcher, improve the hardware compatibility is a big problem, present technical assistance and artisan can not produce the efficient hardware to solute, for example, it is difficult to make sure the screen can normal display in the process of update and before upgrade. To programmer, they also need to transform the software along with the adjustment of hardware. One of them fall behind the current algorithm, the adoption Barriers in Low-End and Mid-Range Models will influence the deeper manufacture and the feeling of users such as the fault phenomenon of man-machine interaction. The relevant scientific payoffs is still evolving.

5 Conclusion

This paper has comprehensively examined the advancements in emotion recognition and multimodal fusion within smart cockpits, highlighting both the promising achievements and the significant challenges that remain. Our review has shown that the integration of affective computing theories, sophisticated deep learning models, and innovative fusion strategies has enabled smart cockpit systems to better understand and respond to driver emotional states. Advanced methodologies—ranging from early and late fusion techniques to Transformer-based and contrastive learning approaches—demonstrate considerable potential in capturing the complex dynamics of human emotion. However, several limitations persist. Computational intensity and the “black box” nature of current deep learning models hinder real-time application and reduce interpretability, which is critical for safety-critical environments. Privacy concerns and the high cost of data acquisition further complicate widespread adoption, especially in mid- to low-end vehicle models.

Future research must focus on optimizing model efficiency and enhancing system transparency through explainable AI techniques. There is also a pressing need to develop scalable multimodal fusion frameworks that can dynamically adjust to varying environmental conditions and individual differences among drivers. Continuous learning and adaptive strategies will be essential for maintaining system performance in ever-changing scenarios. By addressing these challenges, the next generation of smart cockpits can achieve a higher degree of personalization and safety, ultimately transforming the driving experience and contributing to more intelligent and responsive transportation systems.

Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. Gao, F., Ge, X., Li, J., Fan, Y., Li, Y., Zhao, R.: 'Intelligent cockpits for connected vehicles: Taxonomy, architecture, interaction technologies, and future directions', Sensors, 2024, 24, (16), pp. 5172
2. Zepf, S., Hernandez, J., Schmitt, A., Minker, W., Picard, R.W.: 'Driver emotion recognition for intelligent vehicles: A survey', ACM Comput. Surv., 2020, 53, (3), pp. 1-30.
3. Zhu, X., Liu, Z., Cambria, E., Yu, X., Fan, X., Chen, H., Wang, R.: 'A client–server based recognition system: Non-contact single/multiple emotional and behavioral state assessment methods', Comput. Methods Programs Biomed., 2025, 260, pp. 108564.
4. Wang, R., Zhu, J., Wang, S., Wang, T., Huang, J., Zhu, X.: 'Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking', Int. J. Multimed. Inf. Retr., 2024, 13, (4), pp. 39.
5. Wang, F., Ju, M., Zhu, X., Zhu, Q., Wang, H., Qian, C., Wang, R.: 'A Geometric algebra-enhanced network for skin lesion detection with diagnostic prior', J. Supercomput., 2025, 81, (1), pp. 1-24.

6. Zhao, Z., Zhu, X., Wei, X., Wang, X., Zuo, J.: 'Application of Workflow Technology in the Integrated Management Platform of Smart Park', IEEE Adv. Inf. Manag., Communicates, Electron. Automat. Control Conf., 2021, 4, pp. 1433-1437.
7. Zhang, Y., Zhao, H., Zhu, X., Zhao, Z., Zuo, J.: 'Strain Measurement Quantization Technology based on DAS System', IEEE Adv. Inf. Manag., Communicates, Electron. Automat. Control Conf., 2019, pp. 214-218.
8. Qin, Z., Chen, X., Li, G., Cui, J.: 'TMFER: Multimodal fusion emotion recognition algorithm based on transformer', Proc. 4th Int. Conf. Intell. Comput. Hum.-Comput. Interact. (ICHCI), 2023, pp. 1-8.
9. Nemati, S., Rohani, R., Basiri, M.E., Abdar, M., Yen, N.Y., Makarenkov, V.: 'A Hybrid Latent Space Data Fusion Method for Multimodal Emotion Recognition'. IEEE Access, 7, pp. 172948–172964.
10. Subbaiah, B., Murugesan, K., Saravanan, P., et al.: 'An efficient multimodal sentiment analysis in social media using hybrid optimal multi-scale residual attention network', Artif. Intell. Rev., 2024, 57, (2). DOI: 10.1007/s10462-023-10645-7
11. Zhang, F., Li, X.-C., Lim, C.P., Hua, Q., Dong, C.-R., Zhai, J.-H.: 'Deep emotional arousal network for multimodal sentiment analysis and emotion recognition', Inf. Fusion, 2022, 88, pp. 296-304
12. Yin, L., Lin, S., Sun, Z., Wang, S., Li, R., He, Y.: 'PriMonitor: An adaptive tuning privacy-preserving approach for multimodal emotion detection', World Wide Web, 2024, 27, (2).