Multi-modal Real-time Hazard Detection for Autonomous Driving Systems

Md Mahinur Alam, Kanita Jerin Tanha, Minsoo Kim, and Taesoo Jun
Pervasive Intelligent Computing Laboratory, Department of IT Convergence Engineering,
Kumoh National Institute of Technology, Gumi, South Korea
(mahinuralam213, kanitajerin17, kms991022, and taesoo.jun)@kumoh.ac.kr

***Abstract*—Traditional autonomous driving systems often struggle with detecting out-of-label hazards and unexpected obstacles that fall outside predefined categories, limiting their effectiveness in dynamic real-world environments. This paper presents a novel multi-modal real-time hazard detection framework that integrates dashcam video analysis with speed and sound anomaly detection to enhance autonomous driving safety. Our approach leverages unsupervised learning techniques to identify driver reactions through speed variations and sound pattern analysis, providing critical behavioral indicators for hazard presence. The system employs a set of heuristic rules as weak classifiers for hazard detection, which are combined using an ensemble method to improve robustness and mitigate overconfidence in scenarios lacking labeled data. Additionally, we incorporate Vision-Language Models (VLMs) to generate descriptive captions for detected hazards, enabling semantic understanding and interpretation of complex driving scenarios. Experimental results demonstrate that our complete multi-modal system achieves 85.24% mAP and 82.24% AUC, significantly outperforming video-only baselines (45.95% mAP, 42.55% AUC) and validating the effectiveness of multi-modal fusion for real-time hazard detection.**

***Index Terms*—Autonomous driving, hazard detection, multi-modal, vision language model (VLM).**

I. Introduction

Autonomous driving systems hold great potential to transform transportation, but reliable perception and hazard detection in unpredictable environments remain major challenges. Traditional single-modal perception models struggle to identify out-of-distribution obstacles such as debris, animals, or unusual behaviors. To address this, multi-modal perception integrates complementary sensory data, visual, audio, and kinematic, improving hazard recognition [1], [2].

Vision excels at spatial understanding, audio captures temporal cues like honking or braking, and vehicle dynamics reveal driver reactions. Incorporating Vision-Language Models (VLMs) further enhances semantic interpretation by describing detected hazards contextually. Real-time performance is also essential, demanding efficient algorithms that balance latency and accuracy [3], [4].

The main contributions of this work are: (1) A multi-modal real-time hazard detection framework that effectively combines video, speed, and audio data for enhanced perception robustness; (2) An unsupervised approach for detecting driver reactions through speed and sound anomaly analysis; (3) Integration of VLM for generating descriptive hazard captions that provide semantic context; and (4) Comprehensive experimental validation demonstrating significant performance improvements over single-modal approaches, with our complete system achieving 85.24% mAP.

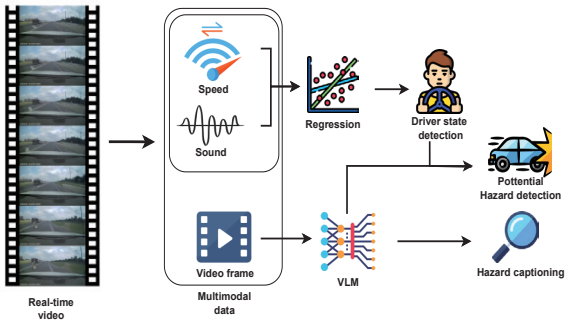

Fig. 1: Proposed multi-modal real-time autonomous driving hazard detection system.

II. Proposed System

Figure 1 illustrates the proposed real-time multi-modal hazard detection system. The system ingests synchronized dashcam frames I_t (RGB image at time t), vehicle speed s_t (telemetry), and short audio windows a_t and outputs per-region hazard flags and concise natural-language captions.

a) Preprocessing.: Video frames are resized and normalized to the detector input; audio is converted to short-time mel-spectrograms $A_t = \mathrm{Mel}(a_t)$; speed is smoothed using an exponential moving average (EMA):

$$\hat{s}_t = \alpha s_t + (1-\alpha)\hat{s}_{t-1}, \quad \alpha \in (0,1].$$

Short temporal buffers of length W (approximately 0.5–1.5 s) are maintained for short-term cues.

b) Visual stream.: A real-time object detector yields region proposals (regions of interest, ROI) $\mathcal{B}_t = \{b_{t,i}\}$. Each ROI is mapped to an appearance embedding $\phi^V_{t,i}$ (lightweight convolutional backbone) and, when available, simple motion proxies (centroid shift, area change). These produce a normalized visual hazard score $h^V_{t,i} \in [0,1]$.

c) Audio stream.: The audio module encodes A_t and produces an anomaly score h^S_t via reconstruction error (autoencoder) or a lightweight event classifier. Audio triggers are time-aligned to nearby ROIs to boost their hazard likelihood.

d) Driver-reaction signal.: Short-window regression on the EMA-smoothed speed estimates local slope a_t (deceleration). A normalized driver-reaction score h^D_t is derived

TABLE I: Experimental evaluation of the proposed multi-modal approach.

Method	mAP (%)	AUC (%)	AUC-Frame (%)	AP (%)
Video	45.95	42.55	43.36	41.63
Video + Speed	56.08	54.51	53.58	54.41
Video + Sound	61.25	60.24	60.01	55.21
Video + Speed + Sound	**85.24**	**82.24**	**78.91**	**80.21**

from negative slopes (braking intensity) and associated with contemporaneous ROIs.

e) Fusion and decision.: Per-ROI modality scores are combined into a single, interpretable fused score

$$H_{t,i} = w_V\, h^V_{t,i} + w_S\, h^S_t + w_D\, h^D_t,$$

where weights w_V, w_S, w_D are calibrated on validation data (or learned when labeled data permit). A binary hazard decision is issued by thresholding with temporal persistence:

$$\hat{y}_{t,i} = \begin{cases} 1 & \text{if } H_{t,i} \geq \tau \text{ for at least } k \text{ consecutive frames,} \\ 0 & \text{otherwise,} \end{cases}$$

with τ and persistence k tuned on validation.

f) Captioning and runtime design.: Flagged ROIs are cropped and asynchronously captioned by a VLM to produce short, safety-focused descriptions $c_{t,i}$. The pipeline is modular and parallelized (visual and audio processing run continuously; VLM runs at lower frequency or asynchronously) to meet low-latency constraints.

g) Training and evaluation.: The detector is fine-tuned on available bounding boxes; the audio autoencoder is trained on normal driving audio; fusion weights (or a small meta-classifier) are calibrated on validation splits. Evaluation reports localization-aware mean Average Precision (mAP) at Intersection over Union (IoU) thresholds, Area Under the Curve (AUC) for scoring, and F1-score (F1) for reaction detection; caption quality is assessed qualitatively or with standard language metrics.

III. Performance Analysis

The COOOL dataset contains over 200 dashcam videos annotated for hazard detection in autonomous driving. It includes a wide range of hazards, both common (vehicles, pedestrians) and rare (animals, debris), demonstrating real-world driving scenarios.

The proposed multi-modal hazard detection system was evaluated on the COOOL dataset using several input modality combinations. Table I shows that the video-only baseline achieves 45.95% mAP and 42.55% AUC. When speed is incorporated (Video + Speed), detection improves to 56.08% mAP and 54.51% AUC. Adding sound alongside video (Video + Sound) further advances results to 61.25% mAP and 60.24% AUC. The full system combining video, speed, and sound achieves the highest performance: 85.24% mAP, 82.24% AUC, 78.91% AUC-Frame, and 80.21% AP.

These results clearly demonstrate the benefit of multi-modal fusion, where integrating behavioral and auditory cues with visual data substantially boosts hazard detection accuracy and robustness.

Fig. 2: Visualization of multi-modal real-time hazard detection with VLM-based captioning.

Fig. 2 provides qualitative examples from the test set, illustrating the temporal alignment of sound triggers and driver reactions (left), as well as real-time detection of hazards in dashcam footage (right). The system's Vision-Language Model generates descriptive captions for each detected hazard, increasing interpretability and aiding downstream decision-making in complex driving scenarios.

IV. Conclusion

This paper presented a real-time, multi-modal hazard detection framework for autonomous driving, integrating video, speed, and sound data along with semantic captioning. The proposed system demonstrated robust performance on challenging real-world scenarios, significantly outperforming single-modality baselines. These results highlight the effectiveness of multi-modal fusion for enhancing both the accuracy and interpretability of hazard detection in autonomous vehicles.

Acknowledgment

This research was funded by the Innovative Human Resource Development for Local Intellectualization Program (IITP-2025-RS-2020-II201612, 34%) through IITP under MSIT, the Basic Science Research Program (2018R1A6A1A03024003, 33%) through NRF, and ITRC Program (IITP-2025-RS-2024-00438430, 33%) funded by MSIT through IITP.

References

[1] M. M. Alam, M. A. Dini, D.-S. Kim, and T. Jun, "Tmnet: Transformer-fused multimodal framework for emotion recognition via eeg and speech," *ICT Express*, 2025.

[2] N. Kamod, "Weighted anomaly scoring for vision-based driving hazard prediction and identification," in *Proceedings of the Winter Conference on Applications of Computer Vision*, 2025, pp. 638–643.

[3] M. J. A. Shanto, M. M. Alam, M. Golam, D.-S. Kim, and T. Jun, "A blockchain-based framework for distributing and managing cnn-derived brain tumor detection models," , pp. 1900–1901, 2024.

[4] M. M. Alam, G. Mohtasin, M. R. Subhan, D.-S. Kim, and T. Jun, "Federated semi-supervised digital twin for enhanced human-machine interaction in industry 5.0," in *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2024, pp. 1270–1275.