Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD30296

Research on a Multimodal Unsupervised Learning Approach for Traffic Accident Detection Based on YOLOv8 Algorithm and Acoustic Feature Fusion

Boyu Qiao

Harbin Institute of Technology, Weihai, China
a2453637614@gmail.com

Abstract. The probability of traffic accidents has increased highly nowadays. So, accident detection is very important. Most existing vehicle accident detection systems are based on convolutional neural networks (CNNs) and variants such as YOLO, Faster R-CNN, RetinaNet, etc. The above methods have issues such as less diversity in the models, accuracy to be improved, and heavy dependence on a large amount of labeled data. This research primarily focused on developing a video-audio multimodal unsupervised vehicle collision detection system based on the YOLOv8 algorithm. The study first employs YOLOv8 to extract spatio-temporal features from video frames, combining them with a convolutional neural network to identify and localize key objects within accident scenes. Subsequently, short-time Fourier transform (STFT) and Mel filter banks extract time-frequency features from accident audio, yielding an acoustic representation aligned with the video embedding space. Two strategies are explored during the feature fusion stage: (1) Post-processing probability-weighted fusion;(2) Attention mechanism fusion. A sliding average probability curve is introduced to smooth instantaneous predictions to address temporal instability in accident detection, generating continuous accident risk estimates that better align with human perception. Results demonstrate that the proposed multimodal fusion method significantly outperforms unimodal models in accident detection tasks.

Keywords: YOLOv8 algorithm, Multimodal Fusion, Probability Curve, Traffic Accident Detection

1. Introduction

Traffic accidents often result in significant economic losses and casualties. Therefore, efficiently detecting and predicting traffic accidents is a critical area of research. However, traditional single-modal deep learning detection methods are prone to failure under varying lighting conditions, occlusions, or adverse weather conditions [1]. The YOLO series models are widely applied in traffic monitoring tasks due to their end-to-end detection capabilities and high inference speed. They provide a solid foundation for accident detection. However, relying solely on visual modalities still has limitations. For example, in a low-light environment or when the surveillance video itself is of low quality, it will be extremely difficult to capture by image those crucial clues which clearly

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD30296

indicate an accident has occurred. On the other hand, the audio modality can register accident-related information such as collision sounds, sudden braking noise, horn honks which happens to complement very well the limitations of visual signals. This has inspired researchers to introduce video-acoustic multimodal fusion into vehicle accident detection systems to enhance their accuracy and robustness. At the same time, existing research has demonstrated that multimodal systems exhibit significant advantages in detection tasks [2]. Based on this, this research proposes a multimodal audio-visual traffic accident detection system based on YOLOv8. The system first utilizes YOLOv8 to extract deep features of vehicles, pedestrians, and their interactions within video frames, while simultaneously extracting audio features through short-time Fourier transform and Mel filters. During the fusion stage, multiple fusion methods—including post-processing weighting [3]and attention mechanisms [4]—were designed to explore the contribution of different modal interactions to accident detection. Finally, the paper smooths the temporal prediction results using a sliding average probability curve, outputting a continuous risk curve that aligns with the evolution patterns of accidents. The study also employed unsupervised learning combined with minimal manual data annotation. This study proposes a novel solution to the issues of poor accuracy, low robustness, and heavy data dependency in intelligent transportation systems.

2. Methodology

2.1. System framework

The multimodal traffic accident detection system proposed in this study consists of three components:

(1)Visual Feature Extraction Branch: Utilizes a YOLOv8-based convolutional backbone network to extract high-dimensional spatial semantic features from sequential video frames;

(2)Audio Feature Extraction Branch: Utilizes Mel spectrograms and a convolutional neural network (CNN) to extract acoustic patterns from the traffic environment;

(3)Fusion and Discrimination Module: Integrates video and audio features through post-processing weighting and attention mechanisms, ultimately generating a continuous accident probability curve.

2.2. Video feature extraction unit

The experiment employed YOLOv8 as the feature extraction network in the visual modality. Based on its CSPDarknet and SPPF (Spatial Pyramid Pooling Fast) modules, YOLOv8 efficiently extracts spatial and local semantic features. Additionally, as the YOLOv8 algorithm inherently supports real-time detection, it effectively meets the requirements for real-time collision warnings in video scenarios.

Rather than relying only on detection box results, this experiment uses YOLOv8 intermediate layer output features as subsequent anomaly detection feature inputs. Because features include multi-dimensional information (scene, object, motion) comprehensiveness than final detection bounding boxes results only, intermediate layers possess sufficiently abundant information; therefore, the experiment tries to make the system robust through this assumption.

This research utilizes the YOLOv8 backbone [5]to extract intermediate features from video frames:

$$F_k \in R^{C_k x \frac{H}{k} x \frac{W}{k}} \quad (1)$$

DOI: 10.54254/2755-2721/2025.LD30296

In which C_k denotes the number of channels, and H, W represent the input image dimensions. We perform Global Average Pooling (GAP) on each feature map to reduce dimensionality while capturing global information. Finally, the video representation vector is obtained by stitching together the multi-scale features of the video.

2.3. Audio feature extraction unit

Distinct acoustic events, such as the sound of brakes, collisions, glass breaking, and tyre friction, often accompany traffic accidents. In certain situations, audio information reflects the occurrence of an accident earlier and more accurately than video. Therefore, we adopt the Mel spectrogram as the primary feature for these reasons:

(1)Compared with traditional MFCC, the Mel spectrum preserves more fine-grained frequency information.

(2)Its two-dimensional form facilitates integration with convolutional neural networks.

By integrating an audio modality module, we can enrich the information collected by the model and prevent issues arising from visual modality failures during nighttime or in occluded conditions.

In this module, a short-time Fourier transform (STFT) is first performed on each frame of the acquired audio signal. The resulting spectrum is then mapped to a Mel frequency scale using a Mel filter bank, where low frequencies have higher resolution and high frequencies have lower resolution, which corresponds to normal auditory perception [5]. Finally, the Mel energy spectrum is log-transformed to reduce the amplitude dynamic range and enhance the representation of weak signals, yielding a Mel spectrogram in two-dimensional matrix form. The final output can be directly fed into a convolutional neural network.

2.4. Cross-modal feature fusion unit

Multimodal fusion is the cornerstone of collision detection systems, enabling complementary integration of different modalities to enhance robustness and accuracy. Due to inherent differences in the dimensionality and distribution of video and audio modalities, direct fusion would result in information distortion. Therefore, the study first normalizes both modalities to facilitate subsequent fusion operations.

2.4.1. Self-attention mechanism fusion

To fully leverage the complementary nature of visual and audio modalities, this study employs a dual fusion mechanism within the cross-modal fusion module, combining self-attention fusion with post-processing weighted fusion. Self-attention mechanisms can dynamically model correlations between different modalities under unsupervised conditions, significantly reducing the computational burden on the model. Video features and audio features can be mapped linearly to form a Query, Key, Value matrix [6]:

$$Q = W_Q[F_v; F_a],\ K = W_k[F_v; F_a],\ V = W_V[F_v; F_a] \tag{2}$$

$$\mathrm{Attention}(Q, K, V) = \mathrm{softnax}(\frac{QK^c}{\sqrt{d_k}})V \tag{3}$$

In which d_k denotes the dimension of the key vector,f denotes video features, and f denotes audio features. Through equation (3), the model dynamically assigns weights to video and audio

DOI: 10.54254/2755-2721/2025.LD30296

features, enabling the modality with greater discriminative power in different scenarios to achieve higher contribution. For instance, when the video is blurred or occluded, the audio modality can be assigned a higher weight, and the same applies to other special cases. Ultimately t he model outputs the fused feature F, encoding complementary visual and audio information. This is then input into FusionNet to map out the frame-level accident probability P_{att} .

2.4.2. Probability-weighted fusion

Although attention mechanisms can effectively fuse cross-modal relationships, their robustness remains insufficient under extreme conditions such as strong audio noise or severely distorted video. This paper introduces a probabilistic weighting fusion strategy at the decision layer to enhance the system's robustness. The research sets video detectors and audio detectors to output accident occurrence probabilities P_v and P_a , respectively. The weighted fusion result is [6]:

$$P_w = \alpha P_v + (1 - \alpha) P_a \tag{4}$$

In the formula, $\alpha \in [0,1]$ represents the weight of the video modality. The weight determination methods include:

Dynamic Weighting: Automatically adjusted based on modal confidence, as shown in the following formula:

$$\alpha = \frac{P_v}{P_v + P_a}, \quad 1 - \alpha = \frac{P_a}{P_v + P_a} \tag{5}$$

This strategy ensures that the other mode can maintain overall system performance when one mode fails.

2.4.3. Fusion result combination and temporal smoothing

This study ultimately employs a joint fusion mechanism to further integrate the attention fusion results with the weighted fusion results, yielding the final prediction probability:

$$P_{final} = \omega P_{att} + (1 - \omega) p_w \tag{6}$$

In the formula, $\omega \in [0,1]$ is a hyperparameter that adjusts the contribution of the two fusion methods.

Taking into account the temporal continuity of the accident detection task, this paper introduces a sliding average strategy to the output probability to obtain the final smooth accident probability curve [7].

In summary, the dual fusion mechanism proposed in this study achieves deep interaction of cross-modal features through a self-attention mechanism, enhances robustness under special conditions via probabilistic weighting fusion, and ensures continuity in the temporal dimension by incorporating a temporal smoothing strategy. This design effectively leverages the complementary advantages of multimodal information, providing theoretical support for the performance improvements demonstrated in the subsequent experimental section.

DOI: 10.54254/2755-2721/2025.LD30296

3. Experiment

3.1. Dataset and preprocessing

The study was conducted on the following data sources: Public Transportation Dataset (Dashcam perspective for video modality, covering diverse traffic scenarios including normal driving, minor collisions, and severe accidents.)

Video frame rate is standardized to 25 FPS, with resolution scaled to 1280x720. The audio sampling rate is 16 kHz. Time-frequency maps are generated using the short-time Fourier transform (STFT) and Mel filter banks. The experiment employs an unsupervised learning strategy, requiring no manual annotations during training or inference. Incident labels are referenced solely during qualitative analysis to validate model performance.

3.2. Experimental environment

This experiment was conducted on MacOs 14.6.1 (23G93) with the following hardware configuration: Apple M3 chip (8-core CPU, 10-core GPU, 16GB unified memory).

3.3. Experimental steps

3.3.1. Feature extraction

Video Modality: Intermediate layer features extracted using YOLOv8, projected into low-dimensional embeddings via VideoHead.

Audio Modality: Mel-spectrogram input to AudioCNN yields acoustic vector representations.

3.3.2. Intermediate layer self-attention fusion

Fuse video and audio at the intermediate layer. Dynamically allocate modal weights through a self-attention mechanism to enable cross-modal feature interaction. Ultimately, output the fused feature F.

3.3.3. Fully connected predictive networks for unsupervised training

Input F into FusionNet to map it to the frame-level accident probability P_{att} .The training strategy employs an anomaly detection objective: the probability of normal traffic frames remains low, while the probability of abnormal (accident) frames increases significantly [8].

3.3.4. Decision-layer probability weighted fusion

The video branch and audio branch features are input into FusionNet, producing probability outputs P_v and P_a respectively. The final frame-level probability is obtained through weighted fusion.

3.3.5. Probability curve smoothing and visualization

Continuously smooth probability curves are generated for the output frame probability P_{final} using a moving average. These smoothed curves are synchronized with video frames for output, visually illustrating the evolution of accident risk over time.

DOI: 10.54254/2755-2721/2025.LD30296

4. Results

This research developed a video-audio multimodal unsupervised learning vehicle accident detection system. It extracts visual features from the intermediate layers of YOLOv8, combines them with audio Mel-spectrogram embeddings, and performs intermediate layer fusion through a self-attention mechanism. At the decision layer, probabilistic weighted fusion is employed to output a continuous accident probability curve ultimately. It obtained the following experimental results, as illustrated in the example figure.



Figure 1. Multimodal probability curve

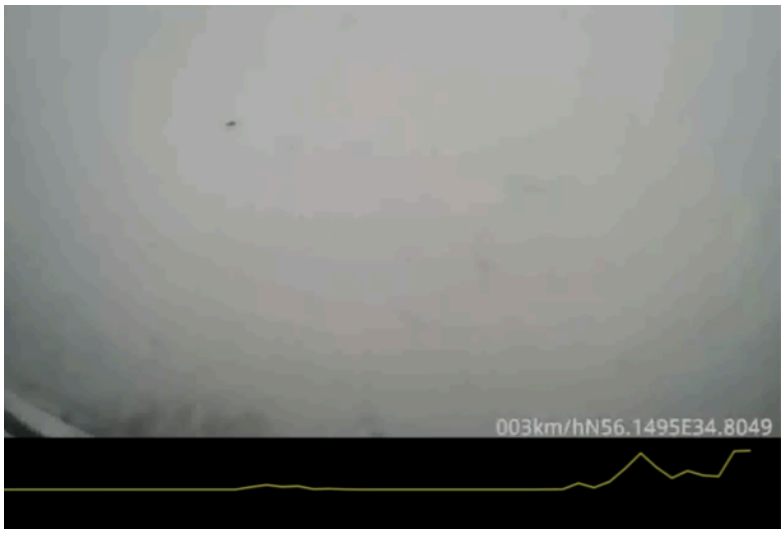

Figure 2. Video single-modal probability curve

As shown above(Figure 1&2), multimodal prediction models provide real-time and accurate probability prediction curves. However, in single-modal models, misjudgments may occur due to visual occlusion. Therefore, the visual-audio multimodal accident detection system developed by this research institute significantly enhances the accuracy of car crash detection.

DOI: 10.54254/2755-2721/2025.LD30296

5. Conclusion

By comparing the above results, we can conclude that: The multimodal system can significantly increase the probability of abnormal events in traffic videos (such as collisions, abnormal braking, pedestrians darting out, etc.) without utilizing accident annotations. Moreover, The multimodal system maintains low probability values for normal traffic frames, aligning with the unsupervised anomaly detection design objective. Furthermore, the fused probability curve exhibits temporal smoothness and continuity, more closely approximating human perception of accident risk fluctuations. Compared to single-modal systems, it demonstrates significant improvements in stability and accuracy. At the same time, Compared to single-modal models, multimodal approaches maintain stable predictions under varying lighting conditions, occlusions, or audio noise interference while significantly reducing reliance on large amounts of labeled data.

However, this approach still has certain limitations: The model only integrates video and audio modalities, without utilizing other onboard sensor data.

Overall, this approach achieves high robustness, multimodal fusion, and interpretable accident risk prediction in an unsupervised environment, providing an effective solution for traffic accident detection.

Future work in this study shall comprise efforts to enhance multi-modality and make the whole setup realizable in real life. Sensors like IMUs, radar, or any telematics data onboard will be the next step toward better perception of our system about complicated traffic scenarios and minor incidents. The cross-modal anomaly detection and generalization capabilities may be further taken to a level by self-supervised learning or even optimized through intermediate-layer feature extraction.

References

[1] Nassar, D. H., & Al-Tuwaijari, J. M. (2024). A review of vehicle accident detection and notification systems based on machine learning techniques. Iraqi Journal of Science, 65(8): 1–15.

[2] Brousmiche, M., Rouat, J., Dupont, S. (2022). Multimodal Attentive Fusion Network for audio-visual event recognition. Information Fusion, 85: 52–59.

[3] Choi, J., Kong, C. W., Kim, G., Lim, S. (2021) Car crash detection using ensemble deep learning and multimodal data from dashboard cameras. Expert Systems with Applications, 183: 115400–115400

[4] Lin, M., Chen, Q., Yan, S. (2013) Network in network. arXiv preprint arXiv: 1312.4400.

[5] Jocher, Glenn, Chaurasia, Ayush, & Qiu, Jie. (2023). Ultralytics YOLOv8: Next-Generation Object Detection. https: //docs.ultralytics.com/modes/train/

[6] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 357–366.

[7] Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. Multimedia Systems, 16(6): 345-379.

[8] Yao, Y., Wang, X., Xu, M., Pu, Z., Atkins, E., & Crandall, D. (2022). DoTA: Unsupervised detection of traffic anomaly in driving videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7), 1–15