International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 698-701
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0098

Multimodal Abnormal Event Detection in Public Transportation

Guhan K[1], Mohanraj N[2], Rajeshkkanna S[3], Vasanthakumar P[4], Jothi P[5]
[1,2,3,4]UG - Computer Science and Engineering, Sri Ranganathar Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India
[5]Assistant Professor, Computer Science and Engineering, Sri Ranganathar Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India
Emails: *guhanguhan540@gmail.com[1], mohanrajmohan1824@gmail.com[2], rajeshkkanna47@gmail.com[3], vk0946983@gmail.com[4], jothi@sriet.ac.in[5]*

Abstract
This project focuses on improving passenger safety in public transportation systems. As the use of public transport increases, incidents such as passenger fights, theft, vandalism, and fall accidents are also rising. To address this issue, this paper presents a multimodal abnormal event detection system using deep learning. The system uses RGB video, depth data, and audio signals to detect abnormal activities inside public transport vehicles. It is designed to work in autonomous vehicles where no driver is present. Experiments conducted on a custom dataset with events such as fighting, bag snatching, vandalism, and normal behavior show promising results, achieving an overall accuracy of 85.1%.

Keywords: *Abnormal event detection, deep learning, multimodal, public transportation*

1. Introduction
Public transportation systems play a vital role in modern society by supporting mobility and improving overall quality of life. However, the increasing demand for public transport, along with rapid urbanization, has intensified concerns regarding passenger safety and security. Although autonomous vehicles (AVs) have been primarily developed to reduce accidents caused by human driving errors such as fatigue and negligence, safety issues within the vehicle cabin remain largely underexplored [1]. In the absence of a driver or supervisory authority, abnormal events such as passenger aggression, petty theft, vandalism, and physical altercations may occur. Detecting such events using visual data alone is challenging, as abnormal behavior is highly context-dependent and often characterized by sudden and rapid movements [2][4]. To address these challenges, this work adopts a multi-pathway deep learning architecture that processes video data at different frame rates to capture both fine-grained spatial details and fast temporal dynamics. Furthermore, depth information enhances spatial awareness, while audio signals provide complementary cues for events that may not be visually observable.

2. Related Work
Research in video recognition has progressed from traditional hand-crafted feature extraction methods to advanced deep learning–based architectures. Early approaches such as 3D Convolutional Neural Networks (3D ConvNets) and the C3D model employed spatiotemporal convolutions to capture motion information from video sequences; however, their ability to model long-term and complex temporal dependencies was limited [5]. A major advancement in this field was the introduction of the SlowFast Network, which uses a dual-pathway architecture to separately capture spatial semantics and temporal dynamics at different frame rates. Despite its effectiveness, the original SlowFast model primarily relies on RGB video inputs, limiting its capability to capture non-visual contextual cues [7][8]. In the area of multimodal learning, several studies have explored the use of depth information and audio features for activity and event recognition. Depth data provides valuable spatial structure, while audio signals offer complementary cues in scenarios involving occlusion or poor visibility. However, many existing approaches treat these modalities independently rather than integrating them within a

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 698-701
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0098

unified framework. This work addresses this limitation by incorporating RGB, depth, and audio data into a unified SlowFast-based architecture for robust abnormal event detection [3][4][5].

3. System Architecture

The proposed system architecture is a multi-stream deep learning framework designed to process multimodal inputs at multiple frame rates. The architecture is inspired by the concept of pathway-based learning, where different streams are employed to effectively model the spatiotemporal characteristics of each modality. Specifically, the framework consists of a slow pathway with a large temporal stride and multiple fast pathways operating at higher frame rates to capture rapid temporal variations.

3.1. Multi-Modal Pathways

Low Frame Rate (LFR) Pathway: The LFR pathway operates at a lower temporal resolution while maintaining a higher channel capacity. This pathway is applied to the RGB modality to extract high-level spatial semantics and fine-grained visual details of the scene.

High Frame Rate (HFR) Pathways: The HFR pathways operate at higher temporal resolutions to capture fast motion and temporal dynamics. Separate fast pathways are introduced for the Depth and Audio modalities, enabling the model to detect rapid changes in spatial structure and auditory patterns that are critical for identifying abnormal events (Figure 1).

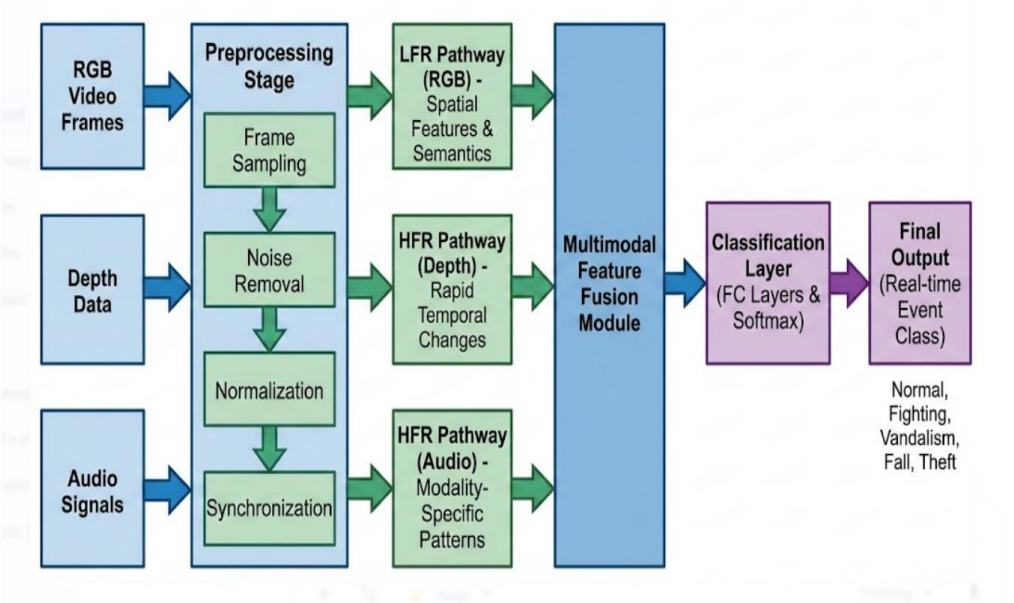

Figure 1 Multimodal Abnormal Event Detection Framework

The block diagram of the proposed system illustrates the complete workflow of the multimodal abnormal event detection framework. The system takes RGB video frames, depth data, and audio signals as inputs, which are first passed through a preprocessing stage. This stage includes frame sampling at different rates, noise removal, normalization, and synchronization across modalities. The pre-processed inputs are then fed into their respective pathways. The RGB data is processed through the Low Frame Rate (LFR) pathway, which focuses on capturing rich spatial features and scene-level semantics.

4. Methodology

4.1. Data Preprocessing

The input data from all modalities undergoes preprocessing to ensure uniformity and efficient learning. Video frames from the RGB and depth modalities are resized to 224 × 224 pixels to maintain consistency across inputs. Normalization is applied using the standard deviation as a scaling factor to improve training stability [4][5][6].

RGB Modality: Standard color video frames captured from the camera are used as input.

Depth Modality: Depth information is converted

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 698-701
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0098

into heatmap representations using a color gradient to enhance spatial feature extraction.

Audio Modality: Audio signals are transformed into Mel spectrograms, which effectively capture time–frequency characteristics relevant to abnormal event detection [3].

4.2. Multi-Modal Fusion

Features extracted from the Low Frame Rate (LFR) and High Frame Rate (HFR) pathways are integrated using lateral fusion connections. The fusion operation is mathematically defined as:

$$F_{\text{merged}} = F_{\text{slow}} + \text{pool}\left(F_{\text{HFR}}^{\text{RGB+Depth+Audio}}\right)$$

This fusion strategy enables a comprehensive representation by jointly leveraging detailed spatial semantics and fast temporal dynamics, thereby improving the robustness of abnormal event detection [9].

5. Simulation Results

5.1. Dataset and Experimental Setup

The proposed system was evaluated using a custom multimodal dataset consisting of approximately **45 hours of video data**. The dataset includes five event classes: *bag snatching, fall down, fighting, vandalism,* and *normal behavior*. The model was trained for **250 epochs** using stochastic gradient descent with a **momentum of 0.9** and a **weight decay of** $10^{-4}$to prevent overfitting [10]-[13].

5.2. Performance Metrics

The performance of the proposed multimodal framework was evaluated using accuracy, precision, recall, and F1-score. The multimodal model integrating **RGB, Depth, and Audio** modalities achieved an overall accuracy of **85.1%**, significantly outperforming the single-modality RGB-based approach, which achieved an accuracy of 77.5% (Table 1).

Table 1 Performance Evaluation of the Proposed Multimodal Abnormal Event Detection System

Class	Accuracy (%)	Precision	Recall	F1-score
Fighting	83.8	0.836	0.848	0.825
Normal	87.6	0.875	0.884	0.867
Vandalism	84.6	0.844	0.855	0.834
Overall	**85.1**	–	–	–

5.3. Visualization

To interpret the model's decision-making process, Gradient-weighted Class Activation Mapping (Grad-CAM) was employed. The generated heatmaps highlight regions of interest relevant to each event class, confirming that the model effectively focuses on critical spatial and temporal areas associated with abnormal activities [14]-[18].

Conclusion

This project presents a robust multimodal abnormal event detection system aimed at improving passenger safety in public transportation environments. By integrating RGB, depth, and audio modalities within a multi-pathway deep learning architecture, the proposed system effectively captures both spatial and temporal features essential for detecting abnormal in-cabin events. Experimental results demonstrate an overall detection accuracy of 85.1%, confirming the effectiveness of the proposed approach for real-time monitoring in autonomous public transport vehicles. Future work will focus on expanding real-world data collection to improve model generalization and further optimizing the architecture for deployment on resource-constrained edge devices, enabling scalable and efficient real-time surveillance solutions.

References

[1]. L. E. Olsson, T. Gärling, D. Ettema, M. Friman, and S. Fujii, "Happiness and satisfaction with work commute," *Social Indicators Research*, vol. 111, no. 1, pp. 255–263, Mar. 2013.

[2]. World Health Organization, "Road traffic injuries," Jul. 2023. [Online]. Available: https://www.who.int/newsroom/fact-sheets/detail/road-traffic-injuries

[3]. S. Kwon, H. Kim, G. S. Kim, and E. Cho, "Fatigue and poor sleep are associated with driving risk among Korean occupational drivers," *Journal of Transport & Health*, vol. 14, Sep. 2019, Art. no. 100572.

[4]. D. Q. Nguyen-Phuoc, O. Oviedo-Trespalacios, T. Nguyen, and D. N. Su, "The effects of unhealthy lifestyle behaviours on risky riding behaviours—A study on app-based motorcycle taxi riders in Vietnam," *Journal of Transport & Health*, vol. 16, Mar. 2020, Art. no. 100666.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 698-701
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0098

[5]. S. E. Shladover, "Connected and automated vehicle systems: Introduction and overview," *Journal of Intelligent Transportation Systems*, vol. 22, no. 3, pp. 190–200, May 2018.

[6]. D. Tsiktsiris *et al.*, "An efficient method for addressing COVID-19 proximity-related issues in autonomous shuttles public transportation," in *Proc. IFIP Int. Conf. Artificial Intelligence Applications and Innovations*, Jun. 2022, pp. 170–179.

[7]. D. Tsiktsiris *et al.*, "Enhanced security framework for enabling facial recognition in autonomous shuttles public transportation during COVID-19," in *Proc. IFIP Int. Conf. Artificial Intelligence Applications and Innovations*, 2022.

[8]. D. Tsiktsiris *et al.*, "Real-time abnormal event detection for enhanced security in autonomous shuttles mobility infrastructures," *Sensors*, vol. 20, no. 17, p. 4943, Sep. 2020.

[9]. D. Tsiktsiris *et al.*, "A novel image and audio-based artificial intelligence service for security applications in autonomous vehicles," *Transportation Research Procedia*, vol. 62, pp. 294–301, Jan. 2022.

[10]. C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, Oct. 2019, pp. 6201–6210.

[11]. C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 1933–1941.

[12]. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 2005, pp. 886–893.

[13]. I. Laptev *et al.*, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2008, pp. 1–8.

[14]. S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[15]. D. Tran *et al.*, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, Dec. 2015, pp. 4489–4497.

[16]. K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, vol. 27, 2014, pp. 568–576.

[17]. D. Kondratyuk *et al.*, "MoViNets: Mobile video networks for efficient video recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 16015–16025.

[18]. Y. Liu *et al.*, "Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–38, Jul. 2024.