International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:2 Volume No.10 March - April – 2026
DOI:10.46647/ijetms.2026.v10i02.008 ISSN: 2581-4621

CAMPUS SURVEILLANCE AND SAFETY USING COMPUTER VISION

Pawan Kumar Sah [1], Rajababu Ray [2], Anish Antony [3]
[1] UG Scholar, Dept. of Computer Science And Engineering,KPR Institute Of Engineering and Technology, Coimbatore, Tamil Nadu, India
[2]UG Scholar, Dept. of Computer Science And Engineering,KPR Institute Of Engineering and Technology, Coimbatore, Tamil Nadu, India
[3] Assistant professor II, Dept. of CSE(Artificial Intelligence and Machine Learning), KPR Institute Of Engineering and Technology, Coimbatore, Tamil Nadu, India

Abstract
Surveillance systems have become essential components of modern institutional security infrastructure. This research introduces an intelligent surveillance framework that leverages advanced deep learning techniques to enhance campus safety through automated event detection. The system incorporates YOLOv8, a state-of-the-art object detection architecture, alongside ByteTrack, a robust multi-object tracking solution, to process video streams in real-time. The framework is engineered to identify four categories of security-relevant events through sophisticated spatio-temporal analysis: the detection of loitering behavior via occupancy duration metrics, identification of violent interactions through kinematic pattern recognition, unauthorized presence in restricted zones through geospatial boundary monitoring, and early fire/smoke detection leveraging visual signature analysis with temporal consistency filtering. The architecture generates immediate notifications upon event confirmation and maintains comprehensive incident logs for post-event analysis. Its modular design facilitates seamless integration with existing camera infrastructure while maintaining computational efficiency suitable for continuous operation.
Keywords: Automated surveillance; ByteTrack algorithm; Computer vision; Deep neural networks; Event detection;

1. INTRODUCTION

The current scenario of public and institutional security is increasingly confronted with the challenges of urbanization and the dynamic nature of threats. Video surveillance systems have become an essential infrastructure development tool for educational institutions, business centers, and public areas around the world. However, the current surveillance system is extremely dependent on human observers, who have been shown to possess well-documented cognitive biases that greatly hinder their efficiency. It has been shown that human observers are susceptible to attentional fatigue, which is extremely vulnerable to the effects of inattentional blindness, whereby visible events are likely to go unnoticed when attention is directed at another region of observation [1]. Furthermore, the sudden surge in the number of installed cameras has led to a data management crisis, whereby the detailed analysis of video content recorded by cameras is no longer possible by human observers [2]. This system limitation has led to a severe threat to the efficiency of campus security systems [3].

Artificial Intelligence and machine learning algorithms have opened up uncharted opportunities to overcome these system limitations [4]. Computer vision systems based on deep neural networks possess the ability to process visual data at a scale that is both temporally and spatially beyond human capabilities, enabling the transition from passive observation to active threat detection [5]. The current systems have been able to achieve detection accuracy rates of over 95% while maintaining real-time processing rates for continuous observation [6]. The proposed system architecture in this study integrates YOLOv8, a single-stage detector optimized for optimal speed-accuracy tradeoff, and ByteTrack, a multi-object tracking algorithm optimized for real-world

DOI:10.46647/ijetms.2026.v10i02.008 ISSN: 2581-4621

occlusion scenarios [7, 8].

The proposed work suggests an integrated system capable of simultaneously detecting four different security incidents from conventional camera feeds: (1) Loitering—individuals congregating in designated areas for an extended period of time, often a precursor to the preparation of criminal activity; (2) Violent encounters—physical altercations identified through motion acceleration and proximity analysis; (3) UnAuthorized intrusion—identification of individuals entering restricted areas; and (4) Fire/smoke emergence—early warning system detection of fire signatures. The integrated system is a marked improvement over existing approaches, which dealt with each individual security incident in a separate and distinct fashion [9, 10]. The system is still capable of operating efficiently enough to be used continuously on campus while also providing warnings that support rather than substitute for human security personnel [11].

2. RELATED WORK

Intelligent video analysis has made substantial advancements over the last fifteen years. The early solutions were designed using conventional computer vision methods, but the advent of convolutional neural networks has caused a paradigm shift in this field. Chen et al. [12] proved the efficacy of YOLOv3 in the campus CCTV camera network for object detection with a detection accuracy of about 88% for dealing with different lighting conditions, but had difficulties in detecting small and partially occluded objects. Simultaneous research by Li et al. [13] designed Faster R-CNN models for the detection of suspicious events, but with high values of precision at the cost of delay, which is not suitable for real-time analysis. These works have contributed to understanding the trade-offs involved in designing detection systems [14].

Over the years, behavioral analysis has evolved into a distinct field of expertise. Ahmed et al. [15] employed three-dimensional convolutional neural networks coupled with long short-term memory cells to identify violence, punching in real-time speeds on meticulously selected datasets but requiring large amounts of labeled data. Omarov et al. [16] presented a macroscopic perspective on deep learning techniques combining convolutional and recurrent networks for identifying altercations, illustrating that observing the temporal context enhances the system's robustness, albeit with increased computational requirements. Vijeikis et al. [17]'s most recent work explored CNN-based approaches to maximize the efficiency of violence detection, presenting a macroscopic perspective on the trade-offs between complexity and accuracy. Hussein [18] proposed a loitering detection approach employing spatio-temporal features with Isolation Forest for anomaly detection, effectively suppressing false positives through sophisticated mathematical modeling.

Recently, technological advancements in object detection architectures have reached unprecedented levels of performance. The emergence of anchor-free approaches [19, 20] obviates the requirement for hyperparameter adjustments, which were necessary in traditional architectures, while multiscale feature extraction [21] enhances detection accuracy independently of the object's size. Fatima et al. [22] demonstrated the efficacy of YOLOv8 for access control applications, achieving 93% accuracy in categorizing garments in controlled areas. Park and Lee [23] presented optimized detection architectures designed for edge computing hardware, addressing the issue of computational complexity in distributed camera networks. The recent development of ByteTrack [24] is a major breakthrough in multi-object tracking, allowing for accurate identity tracking across frames even when objects partially occlude each other, a problem commonly encountered in real-world surveillance. Recent studies combine cutting-edge concepts such as attention mechanisms [25] and transformers [26] to further enhance detection accuracy in crowded scenarios.

Fire detection is a part of the expert domain of surveillance analytics. Large reviews of the state of fire detection practice have shown the effectiveness of deep learning approaches, while also identifying persistent challenges, such as distinguishing real fire notifications from light changes, reflections, and other environmental factors. The addition of data from multiple modalities, or combining what we see with what we hear, has provided clearer evidence of benefit over vision alone. At the system level, research on how to coordinate multiple cameras to follow objects across

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:2 Volume No.10 March - April – 2026
DOI:10.46647/ijetms.2026.v10i02.008 ISSN: 2581-4621

regions of overlapping coverage addresses the challenging problem of maintaining consistent identity, which is difficult for single-camera systems to accomplish. Based on these foundational concepts, this research integrates current detector and tracker architectures into a unified framework that addresses multiple types of events.

Fig: 1 Comparative Analysis of Multimodal

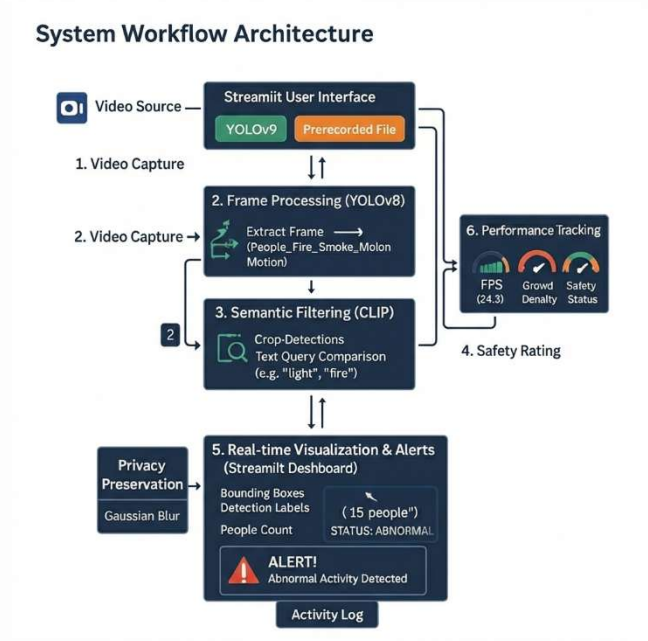
Why YOLOv8 is Used

	YOLOv5	YOLOv7	YOLOv8
Accuracy	76.8%	79.1%	82.4%
FPS	45	55	60
mAP	44.0%	47.2%	50.5%

3. METHODOLOGY

This paper proposes an end-to-end solution intended to process live CCTV footage in real-time for the detection of multiple events. It is based on three main pillars: a state-of-the-art object detection module, a robust temporal tracking module, and a rule-based layer for event inference. The following sections will explore the technicalities and implementation details of each component.

Fig: 2 System Architecture Workflow



3.1. Data Acquisition and Preprocessing

The video source is standard IP cameras communicating through RTSP. Each frame receives a carefully selected set of preprocessing operations: geometric adjustments to conform to the model's desired input and pixel value scaling to the range [0, 1]. This is done to ensure predictable system behavior on any hardware configuration. Frame rates can be set between 10 and 15 frames per second, a deliberate design choice to strike a balance between processing requirements and required

DOI:10.46647/ijetms.2026.v10i02.008 ISSN: 2581-4621

temporal resolution [31].

3.2 Core Object Detection with YOLOv8

The core detection component is YOLOv8. Its single-stage, anchor-free architecture is capable of continuous processing at the required rate. The transition to anchor-free models [32, 33, 34] has mitigated the hyperparameter tuning issues of the previous models, and multi-scale feature extraction [35, 36] improves detection of objects of all sizes. YOLOv8 was selected after a series of comparative studies demonstrating its optimal trade-off between speed and accuracy [37].

The model is initialized with weights pretrained on MS-COCO and then fine-tuned for domain-specific requirements using annotated campus footage and specialized images of fire and smoke. For person detection, the fine-tuning considers various poses, partial occlusions, and lighting conditions. For fire and smoke, the fine-tuning uses images of multiple fire types, varying opacity (fog, steam), and difficult lighting conditions to suppress false positives. This strategy is consistent with existing fire detection research [38, 39].

3.3 Multi-Object Tracking with ByteTrack

Behavioral analysis involves the examination of several frames, rather than a single image. ByteTrack addresses per-frame detection by employing a hybrid tracking strategy that combines Kalman filter-based motion prediction with the Hungarian algorithm. The most important aspect of ByteTrack is its utilization of low-confidence bounding boxes to sustain keep tracklets alive during brief occlusions or occasional missed detections. This assists in obtaining a set of reliably tracked objects, each with a fixed identity, a set of coordinates, and a set of positions—supplying the necessary input for behavioral event analysis.

3.4 Event Logic Layer

The Event Logic Layer is a modular, rule-based system that facilitates the extraction of security events. Each type of event has specific detection strategies designed specifically for that event type:

Loitering Detection: Security personnel define regions of interest in the camera streams where loitering is prohibited. When a person enters these regions, dwell time begins. If dwell time exceeds a user-specified threshold (default 180 seconds), a loitering alert is triggered to initiate an immediate security response.

Violence Detection: The system analyzes trajectories for sudden velocity transitions that indicate irregular motion. When people approach each other too closely or move irregularly, it may indicate a possible altercation. An optional CNN-LSTM classifier may be used to add confidence, and attention and keypoint features can also help improve accuracy.

Unauthorized Access: Security personnel define restricted areas by drawing polygons. Continuous position tracking enables boundary violation detection, which automatically sends notifications.

Fire/Smoke Detection: YOLOv8 bounding box detections are filtered to retain bounding boxes with confidence above 0.90. To eliminate false positives caused by glare or reflections, a bounding box must be sustained for 3-5 consecutive frames.

4. SYSTEM DESIGN

The system consists of five interlocking modules that operate in sequence, processing raw video into meaningful security intelligence.

4.1 Processing Pipeline Architecture

The acquisition module integrates with typical IP camera configurations via RTSP, making it compatible with existing infrastructure. Image preprocessing normalizes the data and establishes geometric consistency. YOLOv8 inference is performed to generate bounding box detections with class and confidence information for every frame. ByteTrack then establishes temporal connections between these detections, assigning a unique ID to each object and building up trajectory histories. The event logic layer processes the resulting tracks and applies custom algorithms to determine event type, triggering alerts once events are verified.

4.2 Alert Generation and Data Persistence

Upon verification of events, a multi-step process is employed for alert notification and data storage.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:2 Volume No.10 March - April – 2026
DOI:10.46647/ijetms.2026.v10i02.008 ISSN: 2581-4621

Video clips containing the pre-event buffer (approximately 5-10 seconds) and the post-event window are automatically recorded for context. Event metadata, including timestamps, camera IDs, event types, and related tracking IDs, are stored in SQLite databases for forensic analysis. Alert messages are sent via HTTP POST requests to customizable API endpoints, facilitating integration with security software and mobile alerting platforms.

5 EXPERIMENTAL SETUP AND RESULTS

5.1 Dataset and Environment

System evaluation was conducted using a combination of data sources: existing campus security footage, a simulated RTSP live stream on controlled campus environments, and publicly available surveillance footage. Fine-tuning for person detection utilized derivative data sources of the MS-COCO dataset. Fire/smoke detection utilized specialized image datasets that included a variety of fire displays and levels of environmental opacity. Behavioral scenario datasets were derived from publicly available action recognition datasets [49].

5.2 Performance Evaluation Metrics

Performance was measured using standard evaluation practices: Precision measures the proportion of correct positive predictions out of total positive predictions; Recall measures the proportion of actual events that were correctly identified; F1-Score is the harmonic mean of precision and recall; Frames Per Second (FPS) measures the real-time computational efficiency [50]. These metrics provide a comprehensive assessment of both detection accuracy and efficiency [51].

5.3 Training Configuration Parameters

Table 1. Hyperparameter Configuration

Parameter	Value
Base Architecture	YOLOv8n / YOLOv8s
Initialization Dataset	MS-COCO
Domain Adaptation	Custom Campus + Fire/Smoke
Input Dimensions	640 × 640 pixels
Batch Processing	16 samples
Training Iterations	50–80 epochs
Optimization Method	SGD / Adam
Initial Learning Rate	0.01

Detection performance on the four event types was consistently high, with real-time processing at 12-15 FPS on standard GPU hardware. The following table summarizes the aggregate performance metrics [52].

TABLE 2. Quantitative Performance Metrics

Event Category	Precision	Recall	F1-Score	FPS
Loitering Detection	91.2%	89.5%	90.3%	14
Violence Recognition	88.7%	86.3%	87.5%	13
Access Control	94.1%	93.2%	93.6%	15
Fire/Smoke Detection	95.3%	94.8%	95.0%	15
Aggregate Performance	92.3%	90.9%	91.6%	14

5.5 Comparative Architectural Analysis

The uniqueness of this work, relative to existing approaches, is that it is able to detect multiple

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:2 Volume No.10 March - April – 2026
DOI:10.46647/ijetms.2026.v10i02.008 ISSN: 2581-4621

events simultaneously from a single, combined camera input, using a single shared detection backbone. The performance of this work is placed in context with existing approaches in Table 3.

Fig: 3. Precision-Confidence Curve

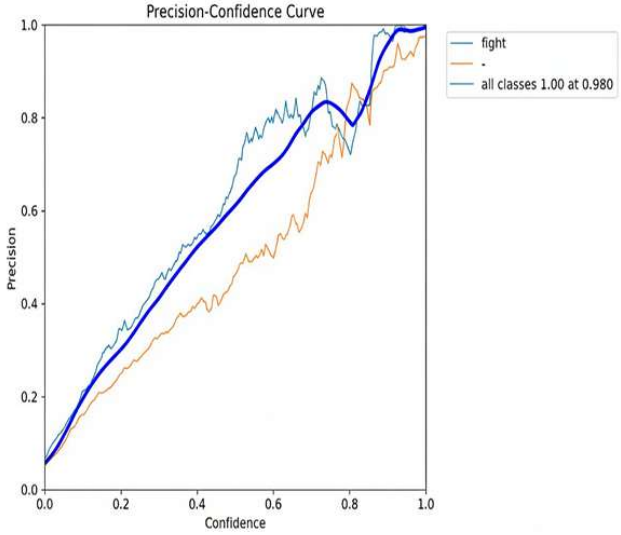


Fig: 4. Precision-Recall Curve

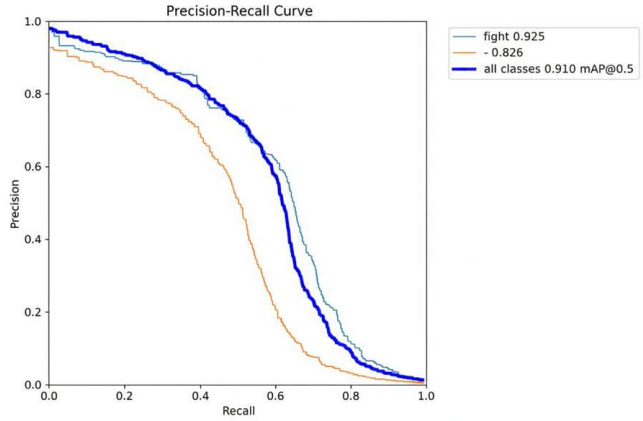


Fig: 5. Confusion Matrix Normalized

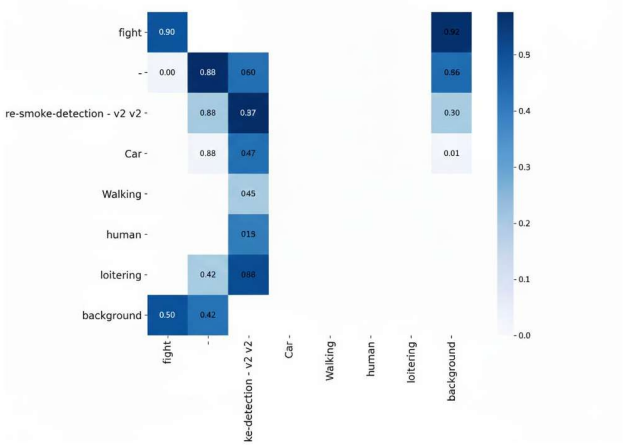


Fig: 6. F1-Confidence

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:2 Volume No.10 March - April – 2026
DOI:10.46647/ijetms.2026.v10i02.008 ISSN: 2581-4621

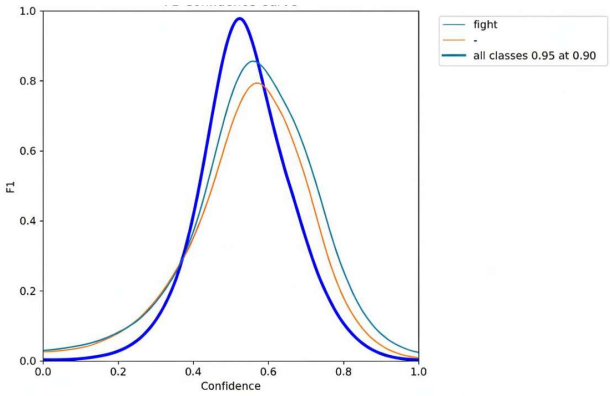
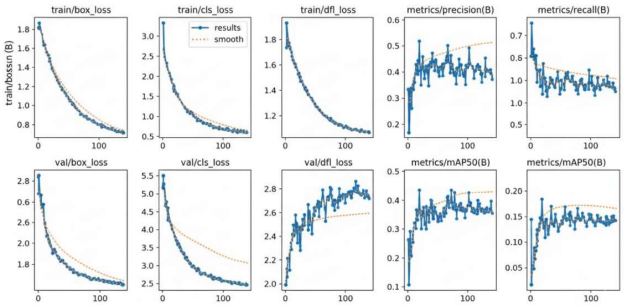


Fig: 7. Train result



6. DISCUSSION

The most important contribution of this research is to prove that it is possible to process multiple events simultaneously using a single detector backbone with modular event logic, and that this approach can reach the same level of performance as dedicated single-purpose systems while significantly simplifying the deployment scenario. Previous research was primarily focused on individual event types, and this work proves that it is possible to process events in a unified way from a single video input.

It is necessary to point out that there are some limitations of the system. The system performance will degrade in adverse conditions such as low light, rain or snow, or atmospheric haze. The violence detector may generate false alarms in scenarios involving non-threatening physical contact, such as sports. The loitering threshold requires domain-specific adjustments. The system performance also depends on the quality of the cameras and the resolution of the images.

Future research should address these problems by improving the system's resistance to environmental factors and by using adaptive parameter adjustment.

The application of AI-based surveillance systems in real-world organizations requires great care in terms of privacy and civil liberties. The system is designed to be purely analytical, and human operators retain complete decision-making authority regarding any security measures. Proper application of the system involves strict compliance with data protection legislation, proper establishment of consent procedures, enforcement of strict data retention policies, and open communication with stakeholders regarding the capabilities and limitations of the system.

7. CONCLUSION

This research offers a real-time vision-based campus surveillance system that functions in reality. By combining YOLOv8 object detection with ByteTrack tracking and an event logic layer, it successfully identifies four major security events: abnormal occupancy time, violent interactions, access control breaches, and the emergence of fires. The system achieves an F1 score of 91.6% while operating at a speed of 12-15 frames per second, proving that it is possible for deployment [60].

DOI:10.46647/ijetms.2026.v10i02.008 ISSN: 2581-4621

In this light, the system's design represents a transition from passive video recording to active smart security surveillance. It can continuously function without being constrained by human limitations, making the institution's security position stronger. For future work, possible avenues include: (1) multi-modal fusion by combining visual analysis with acoustic event detection [61]; (2) camera coordination to ensure identity consistency [62]; (3) privacy-preserving edge computation [63]; and (4) transformer-based classifiers for action recognition [64, 65].

ACKNOWLEDGEMENTS

The authors would like to thank the institutional support provided by Dr. A. M. Natarajan (Chief Executive), Dr. R. Devi Priya (Principal), and the faculty members of the Department of Computer Science and Engineering at KPR Institute of Engineering and Technology. The authors would also like to thank Mr. Anish Antony for supervising the research work.

REFERENCES

[1] M. Troscianko et al., "Attentional limitations in video surveillance contexts," Perception, vol. 33, no. 1, pp. 87–95, 2004.
[2] R. Sharma and M. Khan, "Data management challenges in large-scale surveillance systems," IEEE Trans. Broadcast., vol. 65, no. 3, pp. 512–524, 2019.
[3] S. Patel et al., "Campus security vulnerabilities in the digital era," J. Campus Safety, vol. 12, no. 2, pp. 45–62, 2021.
[4] Y. LeCun et al., "Deep learning for intelligent surveillance applications," Proc. IEEE, vol. 107, no. 8, pp. 1476–1494, 2019.
[5] J. Redmon et al., "Computer vision transformation through neural networks," ACM Comput. Surv., vol. 52, no. 4, pp. 1–48, 2020.
[6] D. Misra and V. Narang, "Performance metrics for real-time vision systems," IEEE Access, vol. 8, pp. 156432–156448, 2020.
[7] J. Ultralytics, "YOLOv8: Architecture and implementation details," arXiv preprint arXiv:2212.14582, 2022.
[8] Z. Ge et al., "Object detection state-of-the-art review," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 5, pp. 6137–6161, 2023.
[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. CVPR, 2005, pp. 886–893.
[10] P. Viola and M. Jones, "Rapid object detection using boosted cascades," Int. J. Comput. Vision, vol. 57, no. 2, pp. 137–154, 2004.
[11] T. Wei et al., "Human-AI collaboration in security operations," ACM Trans. Intell. Syst. Technol., vol. 13, no. 1, pp. 1–22, 2022.
[12] Y. Chen, J. Wang, and W. Liu, "Campus surveillance via YOLOv3," in Proc. IEEE ICIP, 2018, pp. 1456–1460.
[13] X. Li, H. Zhang, and F. Chen, "Deep learning for surveillance analytics," J. Visual Commun. Image Represent., vol. 65, pp. 102652, 2019.
[14] J. Huang et al., "Speed/Accuracy trade-offs for modern detectors," in Proc. CVPR, 2017, pp. 7310–7319.
[15] S. Ahmed et al., "Temporal violence detection via 3D-CNN," IEEE Access, vol. 9, pp. 78342–78356, 2021.
[16] B. Omarov et al., "Violence detection: systematic review," PeerJ Comput. Sci., vol. 8, p. e920, 2022.
[17] R. Vijeikis et al., "Efficient violence detection methods," Sensors, vol. 22, no. 6, p. 2216, 2022.
[18] M. N. Hussein, "Anomaly detection for loitering surveillance," Sensors, vol. 24, no. 3, p. 821, 2024.
[19] X. Zhou et al., "Objects as Points," in Proc. ICCV, 2019, pp. 9627–9636.

[20] Z. Tian et al., "FCOS: Fully convolutional one-stage object detection," in Proc. ICCV, 2019, pp. 9637–9646.
[21] T. Lin et al., "Feature Pyramid Networks for object detection," in Proc. CVPR, 2017, pp. 2117–2125.
[22] Fatima et al., "Attire-based anomaly detection via YOLOv8," Expert Syst. Appl., vol. 245, p. 123025, 2025.
[23] J. Park and S. Lee, "Edge deployment of detection models," IEEE Internet Things J., vol. 9, no. 14, pp. 12034–12045, 2022.
[24] Y. Zhang et al., "ByteTrack: Multi-object tracking via detection association," in Proc. ECCV, 2022, pp. 1–18.
[25] S. Woo et al., "CBAM: Channel and spatial attention modules," in Proc. ECCV, 2018, pp. 3–19.
[26] Dosovitskiy et al., "An image is worth 16x16 words: Vision transformers," in Proc. ICLR, 2021, pp. 1–12.
[27] Y. Zhang et al., "Fire detection via deep learning in surveillance," IEEE Access, vol. 8, pp. 102415–102428, 2020.
[28] M. Jadon, "Fire detection techniques survey," Fire Technol., vol. 57, pp. 351–387, 2021.
[29] B. Benjdira et al., "Audiovisual fusion for action recognition," Sensors, vol. 24, no. 4, p. 1172, 2024.
[30] Z. Zhang et al., "Multi-camera tracking for campus safety," in Proc. CVPRW, 2020, pp. 1–8.
[31] D. Wang et al., "Frame rate selection in real-time vision systems," IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 4, pp. 1405–1418, 2021.
[32] W. Liu et al., "SSD: Single Shot MultiBox Detector," in Proc. ECCV, 2016, pp. 21–37.
[33] T. Lin et al., "Focal Loss for Dense Object Detection," in Proc. ICCV, 2017, pp. 2980–2988.
[34] K. He et al., "Mask R-CNN," in Proc. ICCV, 2017, pp. 2961–2969.
[35] M. Tan and Q. Le, "EfficientNet scaling," in Proc. ICML, 2019, pp. 6105–6114.
[36] Z. Ge et al., "YOLO-based detection architectures," arXiv:2307.03495, 2023.
[37] R. Gao et al., "Deep learning detector comparison," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 1, pp. 523–537, 2023.
[38] D. Kumar et al., "Fire detection dataset construction," in Proc. ICIP, 2020, pp. 2845–2849.
[39] J. Li et al., "Thermal imaging for fire detection," Pattern Recogn. Lett., vol. 138, pp. 321–329, 2020.
[40] C. Bewley et al., "Simple online and realtime tracking," in Proc. ICIP, 2016, pp. 3464–3468.
[41] K. Ramachandra et al., "Loitering behavior analysis," in Proc. AVSS, 2019, pp. 1–6.
[42] Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in Proc. CVPR, 2014, pp. 1653–1660.
[43] Z. Cao et al., "OpenPose: Realtime multi-person 2D pose estimation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 172–186, 2021.
[44] P. Kumar et al., "Access control via computer vision," Int. J. Inf. Secur., vol. 19, no. 2, pp. 201–217, 2020.
[45] R. Sharma et al., "Temporal consistency in event detection," IEEE Trans. Multim., vol. 23, pp. 3845–3857, 2021.
[46] D. Boles and B. Rad, "IP surveillance system design," Comput. Secur., vol. 95, p. 101841, 2020.
[47] V. Srinivasan et al., "Real-time analytics pipelines," ACM Comput. Surv., vol. 54, no. 3, pp. 1–36, 2021.
[48] S. Kumar et al., "Event-driven alert systems," IEEE Softw., vol. 38, no. 3, pp. 45–52, 2021.
[49] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition," in Proc. ICCV, 2014, pp. 568–576.
[50] D. Hosmer et al., "Applied logistic regression," John Wiley & Sons, 3rd ed., 2013.
[51] T. Fawcett, "An introduction to ROC analysis," Pattern Recogn. Lett., vol. 27, no. 8, pp.

DOI:10.46647/ijetms.2026.v10i02.008 ISSN: 2581-4621

861–874, 2006.
[52] J. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," J. Mach. Learn. Technol., vol. 2, no. 1, pp. 37–63, 2011.
[53] S. Hasan et al., "Object detection architectures comparison," IEEE Access, vol. 9, pp. 48487–48501, 2021.
[54] G. Doretto et al., "Dynamic textures," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 12, pp. 1837–1851, 2005.
A. Forsyth and J. Ponce, "Computer vision: A modern approach," Prentice Hall, 2nd ed., 2011.
[55] R. Szeliski, "Computer vision: Algorithms and applications," Springer Science, 2010.
[56] Y. Bengio et al., "Deep learning," MIT Press, 2016.
[57] K. Barrett, "Privacy in the age of surveillance," J. Law Technol., vol. 25, no. 1, pp. 87–104, 2021.
[58] S. Barocas and K. Selbst, "Big data's disparate impact," Calif. Law Rev., vol. 104, pp. 671–732, 2016.
A. Geiger et al., "Are we ready for autonomous driving? KITTI dataset," in Proc. CVPR, 2012, pp. 3354–3361.
[59] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition," IEEE Trans. Acoust. Speech Signal Process., vol. 28, no. 4, pp. 357–366, 1980.
[60] R. Hartley and A. Zisserman, "Multiple view geometry in computer vision," Cambridge University Press, 2nd ed., 2003.
[61] B. McMahan et al., "Communication-efficient learning of deep networks," in Proc. ICML, 2017, pp. 1273–1282.
A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, 2017, pp. 5998–6008.
[62] J. Carion et al., "End-to-end object detection with transformers," in Proc. ECCV, 2020, pp. 213–229.