STM JOURNALS

International Journal of
Optical Innovations & Research

Volume 3, Issue 1, 2025
January–June
DOI (Journal): 10.37591/IJOIR

https://journals.stmjournals.com/ijoir

Research IJOIR

Automated Suspicious Activity Detection in Video Surveillance Using Deep Learning: A Review

Prati Dubey[1,*], Rakesh Kumar Mittan[2]

Abstract

In the current era of advanced security systems, video surveillance plays an essential role in ensuring safety by detecting suspicious activities. With the increase in real-time data, manual monitoring has become impractical, paving the way for automated surveillance systems utilizing machine learning (ML) and artificial intelligence (AI) technologies. This study explores the integration of ML and AI models, specifically convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, for suspicious human activity detection in video streams. The proposed system involves video data collection, preprocessing, feature extraction, and model training to identify abnormal behavior. We have reviewed existing literature on human activity detection, discussing several models and techniques for video anomaly detection and object tracking. The study demonstrates how these intelligent systems can be applied in various environments, providing real-time insights and proactive security measures. The study also highlights the challenges related to deep learning, such as overfitting, computational requirements, and the need for extensive labelled data for effective model training.

Keywords: Suspicious activity detection, video surveillance, machine learning, convolutional neural networks (CNN), long short-term memory (LSTM) networks

INTRODUCTION

Nowadays, with the growing applications of advanced indoor and outdoor security systems, video surveillance shows its significant role; for these suspicious human activities of objects are detected using image processing tools or more advance artificial intelligence (AI) tools such as machine learning (ML). The ML models are effective in identifying suspicious actions, such as fighting, stealing, and trespassing, etc. [1–3]. Advanced surveillance systems play a critical role for safety and security purposes in daily life [4, 5]. The increased volume of real-time data makes manual monitoring time-consuming and impractical task. To remove these issues, automated surveillance systems are rapidly advancing for analyzing abnormal events within image/video frames.

***Author for Correspondence**
Prati Dubey
E-mail: 1dubeyprati245@gmail.com

[1]Research Scholar, Department of Computer Science and Engineering, Rabindranath Tagore University Bhopal, Madhya Pradesh, India
[2]Associate Professor, Department of Computer Science and Engineering, Rabindranath Tagore University Bhopal, Madhya Pradesh, India

Received Date: January 07, 2025
Accepted Date: March 27, 2025
Published Date: April 08, 2025

Citation: Prati Dubey, Rakesh Kumar Mittan. Automated Suspicious Activity Detection in Video Surveillance Using Deep Learning: A Review. International Journal of Optical Innovations & Research. 2025; 3(1): 20–27p.

Suspicious activity detection in videos typically involves several general steps commonly used by researchers:

- *Foreground object detection*: Background subtraction is a robust method used to identify changes in a sequence of frames and extract foreground objects, isolating them from the background.
- *Object detection:* Object detection within video frames can be achieved through non-tracking based or tracking-based approaches. Tracking-based methods help create the trajectory of an object over time by locating its position in each frame of the video.

- *Feature extraction*: Shape and motion-based features of objects are extracted using various algorithms to aid in object identification. In some cases, the feature vector in the process of suspicious object detection involves a systematic series of steps:
- *Input data acquisition*: Video/image data is supplied as input from the surveillance cameras or imaging devices.
- *Preprocessing*: The raw input is preprocessed to improve data quality and eliminate noise.
- *Object classification*: Objects are classified as humans, vehicles, or something else, using various techniques, such as Support Vector Machine, Haar-classifier, Bayesian, K-Nearest Neighbor, Skin color detection, and Face detection.
- *Object analysis*: Upon classification, activity analysis is accomplished through various threshold values in order to isolate abnormal activity [6]. Classic objective detection preprocessing performed includes: noise filtering, image enhancement, and normalization, thus improving data quality and filtering out irrelevant information. The next step after preprocessing is feature selection and extraction, where distinctive features critical for object detection and recognition are identified and isolated from the preprocessed data. Such features are central in discriminating objects of interest against the background or other irrelevant aspects. After their identification, a feature network is constructed for subsequent analysis [7].

The feature network construction involves the positioning of desirable characteristics to form a basis for a recognition stage. With recognition, training algorithms based on the developed feature networks determine their characteristics and class. Subsequently, the feature network developed helps make informed decisions regarding the nature of the objects within the given input data. This outputs results in verifying the presence of suspected threats in the examined video/image as regards the conclusion of the detection process [8].

LITERATURE REVIEW

Gawande *et al*. I have created a dataset that captures student behaviors in an educational environment, including incidents such as cheating, theft of laboratory equipment, fighting, and threatening situations [1]. The data set marked the general manner in identifying individuals from a consistent and standardized approach, making it possible for those individuals studying to correct, track, and analyze their behavior. The researcher validated their contribution for effectiveness with their own dataset and publicly available benchmark datasets. The state-of-the-art detection accuracy achieved an impressive 96.12%, with, amongst other things, an overall error rate being 6.68%, surpassing the existing methods. Empirical results show a remarkable advancement in anomaly detection. The paper is concluded with a summary of the major findings along with directions on future research plans in this domain.

Tutar *et al*. proposed a hybrid model in video anomaly detection which combines several machine learning techniques with pixel-based video anomaly detection and frame-based video anomaly detection models [2]. The PBVAD model is based on the motion influence map (MIM) algorithm which takes into account the spatiotemporal factors. The FBVAD model uses k-nearest neighbors (kNN) and support vector machine (SVM) machine learning algorithms in a hybrid way. High-performance anomaly detection on the UCF-Crime dataset was obtained, averaging AUC performance metrics of 98.0% for FBVAD-kNN and average success rates of 8741.7% for PBVAD-MIM. Important contributions of the study pertain towards the real-time detection of anomalies in video data and the prevention thereof. Selvakumar presented the Human Abnormal Behavior Recognition and Tracking (HABRT) model which consisted of techniques like gathering video frames, labeling the actions as normal or abnormal, and monitoring object extraction and abnormality classification [3]. They employed Multiscale Dilated-assisted Residual Attention Network (MD-RAN) for abnormal behavior classification; hyperparameter optimization went through Modified Random Parameter based Chimp Optimization Algorithm (MRP-ChOA); activities in tracking went through Adaptively Modified You Only Look Once (YOLO) V3 (AM-YOLO V3); and MD-RAN was used for classification of abnormalities. They compared them with other algorithms and achieved very high accuracy on

dataset 1, denoting its capability for extended results in abnormal recognition and tracking. Alruwais *et al*. developed deep learning based algorithms to continuously track a student's mood in an online environment [4]. They recognized seven emotions: anger, contempt, happiness, sadness, fear, and surprise in a pseudo-normal ratio of 99% accuracy. Their approach utilized various CNN model-based convolutional layers, dropout regularization, and batch normalization to augment their predictions and to limit overfitting. Their goal was to improve student engagement and learning outcomes, which in turn would benefit educators in understanding students' behaviors and providing personalized tutoring. Finally, improvement of performance and assessment of student activity were anticipated. Wang *et al*. presented the Quantized Object Recognition Model (QORM), a deep learning-based approach for image categorization that aims to reduce the impact of pattern fluctuations [5]. QORM employs quality equalizers for segment-wise analysis and differentiation of patterns. It compares training inputs with quantized segments based on characteristics like saturation and direction to identify objects and individuals. Separate training is conducted to address non-classifiable images resulting from pattern variations leading to errors. This study used the Home Object 06 dataset, achieving significant improvements in accuracy (8.06%), F1-Score (9.01%), and sensitivity (14.82%) for object and individual categorization in surveillance systems, offering a potential solution to the challenges associated with object recognition in such scenarios. Daud *et al*. developed a platform combining software-defined radio technology and machine-learning algorithms to identify and classify suspicious or non-suspicious liquids [6]. They fine-grained the samples from OFDM methods to acquire channel state information of liquids operating at frequencies of 900 MHz and 2.45 GHz. Machine learning algorithms employing the dielectric property of liquids were implemented for the classification. The system correctly categorized more than 95% of both suspicious and non-suspicious liquids, establishing its efficacy for liquid classification. This is a versatile portable, modular, and cost/investment-efficient solution. Talib *et al*. proposed an improved YOLOv8 model known as YOLOv8-CAB for object detection, with special consideration towards small objects and with varying image types [7]. They improved feature extraction without increasing the complexity of the model by tuning the C2F block. SA was also tuned in order to have faster detection performance.

The YOLOv8-CAB had better performance than the conventional YOLO models, achieving a mean average precision of 97%, which is an increase of 1% on previous models, especially in the detection of smaller objects. This enhancement opens new directions for real-time object detection techniques. Sun *et al*. introduced a flying bird object detection algorithm based on motion information (FBOD-BMI) to mitigate the problems of unclear features and sizes of the objects in low-SNR (Signal-to-Noise Ratio) from surveillance videos [8]. They produced Adaptive Spatiotemporal Cubes (ASt-Cubes) for flying bird objects with the intent of increasing SNR and retaining the pertinent environmental information adaptively. They also designed a lightweight U-shape net (LW-USN) based on ASt-Cubes to detect flying bird objects effectively while rejecting false detections. Experimental results using surveillance video datasets demonstrated the algorithm's effectiveness in detecting flying bird objects in unattended traction substations. Gautam *et al*. discuss the importance of object tracking systems for surveillance, particularly in detecting suspicious abandoned objects in various areas like airports, railway stations, parking lots, and public transport to prevent terrorism and related incidents [9]. They present a multi-object tracking model along with abandoned baggage detection in real-time environments. The paper analyzes track-by-track methods, including the YOLO track and the SORT algorithm track, using a dataset of images annotated with YOLO for six specific categories. The goal is to analyze and understand the application of these methods, particularly in identifying vehicles or pedestrians in continuous videos for congestion monitoring. Ahmed *et al*. propose a novel Detection Transformer (DETR) framework for detecting and classifying highly cluttered suspicious items, especially in X-ray scans [10]. Their framework involves extracting features from a CNN backbone using object proposals based on coherent contour maps. These weights are passed to the CNN model in the DETR to enhance feature extraction. The transformer encoder-decoder is fed with representative features for predicting bounding boxes of cluttered and concealed prohibited items. The framework is evaluated on a large dataset of X-ray scans from the PIDray dataset, outperforming state-of-the-art

schemes in terms of mean average precision for easy, hard, and hidden subsets of the dataset. Pullakandam *et al*. proposed a real-time weapon detection system using the YOLOv8 model, which they claim to be faster, more accurate, and superior to YOLOv5 [11]. They quantized the weights of YOLOv8 for faster performance and evaluated both YOLOv8 and YOLOv5 for weapon detection. YOLOv8 achieved a mean Average Precision (mAP) of 90.1%, outperforming YOLOv5 with an mAP of 89.1%. Weight quantization further reduced inference time by 15% compared to the original YOLOv8 configuration.

HUMAN ACTIVITY DETECTION

The proposed human activity detection system comprises a multi-stage process aimed at achieving accurate and robust results, as illustrated in Figure 1. The initial phase involves Video Data Collection, where a diverse and representative set of videos capturing various human activities is gathered.

This extensive dataset enables to generalization of AI/ML models across scenarios and activities, is what it proves to be: vast amalgamation. Actual video footage worked upon comprise the Video Dataset, giving the impetus for the entire system. The Data Annotation phase involves the proper labelling of every video that considers and annotates important contextual information about the human activities portrayed in them.

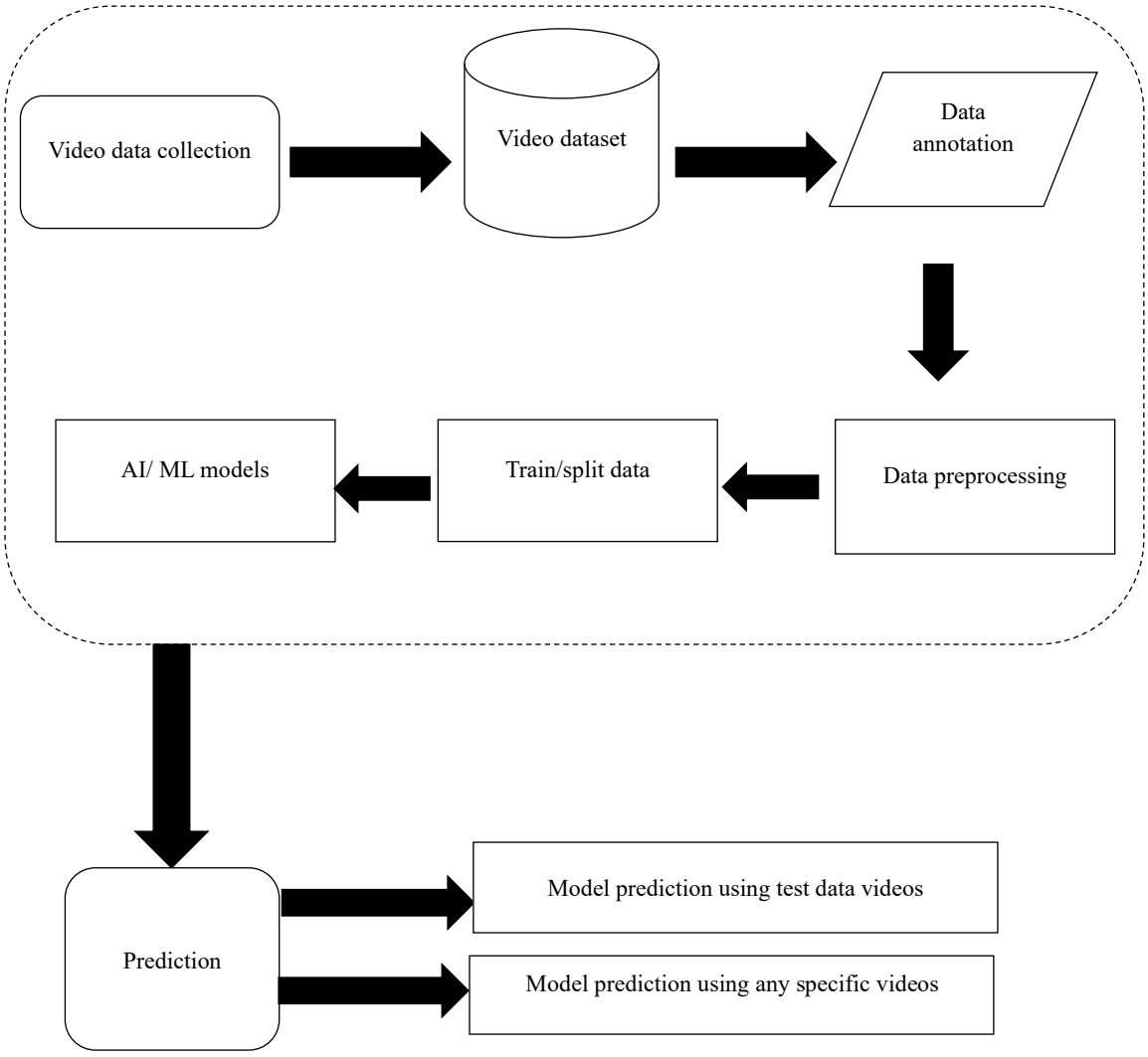

Figure 1. Human activity detection.

The proposed system for human activity detection is a multi-step process:

- *Video data collection*: A representative dataset of videos capturing different human activities is gathered for generalizability of the model.
- *Data annotation*: Videos are annotated with labels and context information, so AI/ML models can perform supervised learning.
- *Data preprocessing*: Videos are cleaned, normalized, and formatted before allowing uniformity and quality in the datasets and tackling problems like inconsistency in video quality and lighting.
- *Feature extraction*: Characterization of the preprocessed data to extract high-fidelity information, including measures related to space and time, as well as motion involving objects. These are the inputs for AI/ML models.
- *AI/ML models*: With the use of the training set of labelled and preprocessed data, the models were trained to learn the appropriate trends and correlations present among them with optimization to minimize errors and enhance accuracy.
- *Prediction*: Predictions can be divided into two phases:
- *Model predictions on specific videos*: This uses videos from the training dataset to test model predictions, testing whether it learned the nuances of these activities.
- *Model predictions on testing data videos*: To determine the degree of generalization, these videos have not been previously seen by the model and provide insights regarding how it performs on real-world data.

SUSPICIOUS OBJECT DETECTION AND ITS RELATIONSHIP WITH ML/AI

The efficiency of Machine Learning (ML) in detecting suspicious activities such as those involving human behavior depends on the nature of the training data combined with the architecture of the ML model. When the model is trained on a dataset with instances of suspicious human activities, it has the potential to identify such activities and discern the underlying patterns and behaviors indicative of them. This implies that the accuracy and scope of detection will actually rely on how equitable the datasets are in representing various suspicious behaviors and how intricate the model is when it comes to pattern interpretation.

Broadly speaking, the ability of the model to detect suspicious activities depends mainly on the number of such examples the ML algorithm was exposed to during the training, implying the importance of comprehensive and diverse datasets in any ML applications [1]. The deployment of ML for the detection of suspicious activities also extends to the recognition of individual behaviors such as fighting, stealing, trespassing actions, and various acts of suspicious or dangerous acts. Such capacity is, of course, especially applicable within the context of video surveillance systems, which crop up in all kinds of settings from the indoor to outdoor.

Using ML in surveillance systems, the capability for automatic detection of abnormal or suspicious behavior is improving these systems' effectiveness, thereby enhancing the security and safety of these areas. The recognition of human behavior through ML shows its application not only in security but also in several other areas like analysis of shopping behavior, showing its flexibility. The integration of ML into surveillance systems shows a substantial source of progress in security, thereby proving that security cameras will be a critical and ubiquitous means of increasing safety, both public and private [2].

On the other hand, in the current scenario of surveillance, it has become utterly impractical due to multiple data littered across the cameras to monitor the events by humans. Further, the chase of looking for a particular incident within the footage is a time-consuming process. Automated video surveillance systems are gaining momentum in this regard, particularly when it comes to the analysis of abnormal events.

AI, ML, and DNNs, in particular, are revolutionizing the field of image and video surveillance. AI allows computers to imitate human thought processes; ML is the process of using the training data in

making a prediction on future data. The recent introduction of GPUs and massively big datasets has given a major boost to the use of ML in processing and analyzing video content. DNNs are driving this technological push and have become the go-to architecture for the efficient analysis of automatic video data due to their ability to perform complex learning tasks.

The rise in convergence between AI, ML, and DNN technologies is contributing to smarter surveillance systems that can detect unusual incidents and analyze them automatically, potentially redressing the inefficiencies of observed monitoring practices [3]. At their best, ML models emphasize raising CNN and LSTM networks to analyze and process image and video information taken from CCTV footage in a sophisticated manner.

The models basically facilitate feature extraction and build high-level representations of the data, allowing us to leverage interesting characteristics since the feature extraction procedure is made automatic. The way CNNs handle is worth mention in that they learn the visual pattern directly from the image pixels; accordingly, they are specialized for the analysis of still images or individual video frames.

LSTMs are suited to the learning and retention of long-term dependencies, which is a major characteristic for inferring the context in and the flow of the events in the videos since they operate on dynamic sequences such as video streams that view motion over time. The commingling of CNN and LSTM models into a surveillance system allows for the automatic monitoring of human behavior and investigations of CCTV footage to detect suspicious events.

The proposed model sets in motion a pre-emptive security measure with the ability to alert authorities or stakeholders upon detecting potentially questionable behavior. The automation of feature extraction and the ability to build generalized models from data are why, themselves, these learning machines, CNNs and LSTMs, are used to the advantage of enhancing the effectiveness and efficiency of surveillance systems and hence offer a more holistic attention to security monitoring [4].

Intelligent video monitoring systems, particularly those intended for event detection and recognition of human behavior, depend on numerous components for such success. The capability of any of these systems to detect suspicious activities is dependent largely on the availability of good quality training data, architecture of the model, and the nature of the deployment environment. Such systems, to keep performing effectively and stay relevant, require continued monitoring, upgrades, and a culture of continual improvement.

Through this approach, it is possible to retain some consistency in such systems, i.e., to keep them updated and the ability to recognize new and evolving models of suspicious behavior. Another fascinating branch of ML is deep learning, specifically this branch does an exemplary job of learning various advanced patterns and representations right from the raw input data itself. This wonderfully cuts down on the need for the process of feature extraction, which is usually a cumbersome and error-prone task. One of the major highlights of deep learning is that it allows unhindered access to raw data and learns very well from that very data without excessively normalizing and going through preprocessing. This is very different from classical ML techniques that strongly depend on such preprocessing efforts.

With regards to feature learning, deep learning models carry out a great job of automatically learning higher-level features from raw input that enable them to learn complex hierarchical relationships existing within the data. In applications such as human activity recognition (HAR), deep learning models perform astonishingly well, learning spatial and temporal dependencies and scale-invariant features without the need for prior feature engineering. Such potential allows these models to easily generalize upon discriminative patterns of human activities, thus giving them an expansive edge regarding surveillance and monitoring tasks [5].

In spite of their transformative power across a range of complex fields-from automated real-time video surveillance to human behavior recognition, deep learning models have a number of requirements that need to be met to fully leverage their capabilities. One significant challenge in the deployment of deep learning models, particularly neural networks, is their dependency on substantial amounts of training data. This is owing to the complex architectures of neural networks that tend to have a high number of parameters.

Thus, the complexity of the architectures makes the models overfit, or, in novice speech, works perfectly on training data but is built poorly on new or unseen data. Overfitting is a critical problem because it undermines the behavior of the model on real data that is not available during the training since it does not cater to all features of such data. Thus, because of highly convoluted and sophisticated architectures, a deep neural network requires numerous examples of training data to learn very flexibly and generalize quite robustly.

The demand for vast copious datasets currently poses the biggest challenge in deep learning applications, where methods might not easily or economically afford to collect massive amounts of labeled data. Another challenge is the sheer computational intensity of deep learning. Deep learning models require incredibly daunting computing power as their training involves running many more iterations and simultaneous computations. Such procedures cannot be run easily using ordinary and simple machines; they need powerful machines to achieve efficiency.

Dedicated hardware like Graphics Processing Units and Tensor Processing Units is often put in place to respond to meet these demands, due to their ease in speeding up the training of neural networks, to summarize, on the one hand; but on the other, deep learning provides significant advantages in feature-learning automation and improving systems that are quite capable of detecting and analyzing human behavior-depending greatly on large datasets and substantial computing power. The emergent need is to underscore a balance of deep learning potential benefits with the practical needs of successful deployment in practice [6].

CONCLUSION

The study suggests the importance of using machine learning and AI for automated video surveillance systems for suspicious human activity detection. The fusion of CNN and LSTM models has been proven to be effective in modeling spatial and temporal characteristics of activities, thereby helping to ensure better performance for security purposes. The application of DL models in this process acknowledges menial automation to some extent for feature extraction, although there are certain problems remaining to deal with, such as overfitting and needing high data volume/processing power to get our modelling correct. Future works would involve exploring hybrid models and optimizing architectures for improved real-time performance and generalization to unseen data. Constant technological evolution in the video surveillance systems makes this surveillance modality one of the most viable approaches to offering adequate solutions for the generally accepted public safety.

REFERENCES

1. Gawande U, Hajari K, Golhar Y. Novel person detection and suspicious activity recognition using enhanced YOLOv5 and motion feature map. Artif Intell Rev. 2024 Jan 18; 57(2): 16.
2. Tutar H, Güneş A, Zontul M, Aslan Z. A hybrid approach to improve the video anomaly detection performance of pixel-and frame-based techniques using machine learning algorithms. Computation. 2024 Jan 24; 12(2): 19.
3. Selvakumar S. An effective framework of human abnormal behaviour recognition and tracking using multiscale dilated assisted residual attention network. Expert Syst Appl. 2024 Aug 1; 247: 123264.
4. Alruwais NM, Zakariah M. Student Recognition and Activity Monitoring in E-Classes Using Deep Learning in Higher Education. IEEE Access. 2024 Jan 17; 12: 66110–28.

5. Wang J, Hu F, Abbas G, Albekairi M, Rashid N. Enhancing image categorization with the quantized object recognition model in surveillance systems. Expert Syst Appl. 2024 Mar 15; 238: 122240.
6. Daud A, Khan MB, Khattak AB, Tanoli SA, Mustafa A, Rehman M, López OL. Next-Generation Security: Detecting Suspicious Liquids Through Software Defined Radio Frequency Sensing and Machine Learning. IEEE Sens J. 2024 Jan 15; 24(5): 7140–52.
7. Talib M, Al-Noori AH, Suad J. YOLOv8-CAB: Improved YOLOv8 for Real-time object detection. Karbala Int J Mod Sci. 2024; 10(1): 5.
8. Sun ZW, Hua ZX, Li HC, Zhong HY. Flying bird object detection algorithm in surveillance video based on motion information. IEEE Trans Instrum Meas. 2023 Nov 28; 73: 1–15.
9. Gautam D, Gupta H, Shekhar H, Kumar M, Nasser SJ, Fouad L. Suspicious Object Tracking with Yolov3 with Python Using Open-CV. In 2023 IEEE 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). 2023 May 12; 1772–1775.
10. Ahmed A, Alansari M, Alnuaimi K, Velayudhan D, Hassan T, Werghi N. Detection transformer framework for recognition of heavily occluded suspicious objects. In 2023 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA). 2023 Jun 12; 1–6.
11. Pullakandam M, Loya K, Salota P, Yanamala RM, Javvaji PK. Weapon object detection using quantized yolov8. In 2023 IEEE 5th international conference on energy, power and environment: towards flexible green energy technologies (ICEPE). 2023 Jun 15; 1–5.

