Journal of Al-Azhar University Engineering Sector

Vol. 20, No. 76, July 2025, 974-996

MULTI-CAMERA BASED INTELLIGENT VIDEO SURVEILLANCE SYSTEM USING SSD_MOBILENET_V3

M.Shaban[1,*], Marwa Elpeltagy[1], Ahmed Y. Khedr[1], A. Al-Marakbey[1]

[1]Systems and Computer Engineering Department, Faculty of Engineering, Al-Azhar University, Cairo, Egypt

***Correspondence**: mariamshaban@azhar.edu.eg

Citation:

M. Shaban, M. Elpeltagy, A. Y Khedr, A. Al-Marakbey, Multi-Camera Based Intelligent Video Surveillance System Using SSD_Mobilenet_V3. Journal of Al-Azhar University Engineering Sector, vol 20, pp. 974-996, 2025.

Received: 17 February 2025

Revised: 07 May 2025

Accepted: 31 May 2025

Doi: 10.21608/auej.2025.351357.1752

Copyright © 2025 by the authors. This article is an open access article distributed under the terms and conditions Creative Commons Attribution-Share Alike 4.0 International Public License (CC BY-SA 4.0)

ABSTRACT

Contemporary intelligent surveillance systems have shifted from passive monitoring to active threat detection through advanced behavioral analytics. This work introduces a flexible, high-speed, real-time anomaly detection framework that dynamically evaluates critical security threat indicators, including spatiotemporal event patterns, object-class kinematics, trajectory semantics, and scene-context deviations. The system detects both abnormal human activities (e.g., unauthorized zone intrusions, sudden locomotor anomalies) and anomalous vehicular events (e.g., contraflow violations, perimeter breaches) through integrated spatiotemporal analysis. A graphical user interface (GUI) allows users to define abnormal scenarios. Users can customize abnormal factors, including event time, object type, and direction of abnormal movement. They also can draw areas of interest. The deep neural network Single Shot Multi-Box Detector (SSD_MobileNet_v3) is employed to detect objects within video frames. Subsequently, a Kernelized Correlation Filter (KCF) tracker is used to monitor these objects and identify abnormal motion direction. An innovation of this system is its ability to classify human motion types by establishing a relationship between the actual dimensions of individuals and the observed distances in the video. Furthermore, a dataset for abnormal behavior detection has been created. The efficiency of the system is assessed on the authentically distorted surveillance video dataset and the obtained dataset. The proposed method achieves a recall rate of 95.83% for truck detection and 92.42% for walking intrusion detection. An F1_score of 87% for motion classification is reported. An average processing speed of 16 frames per second (fps) has been scored. The research results show that the suggested strategy outperforms existing state-of-the-art procedures.

KEYWORDS: Multi-camera Video Surveillance. Direction Detection. SSD_MobileNet_v3. Human Motion Classification.

نظام مراقبة فيديو ذكي متعدد الكاميرات باستخدام SSD_MobileNet_v3

مريم شعبان [*]، مروه سليم، أحمد يوسف خضر، أشرف عبد الرحمن المراكبي

قسم هندسة النظم و الحاسبات،، كلية الهندسة، جامعة الأزهر، القاهرة، مصر

البريد الالكتروني للباحث الرئيسي: mariamshaban@azhar.edu.eg

الملخص

تحولت أنظمة المراقبة الذكية الحديثة من الرصد السلبي إلى الكشف النشط عن التهديدات من خلال التحليلات السلوكية المتقدمة. تقدم هذه الدراسة إطار عمل مرنًا وسريعًا للكشف عن الشذوذ في الوقت الفعلي، يقوم بتقييم ديناميكي لمؤشرات التهديد الأمني الحرجة، بما في ذلك أنماط الأحداث الزمكانية،

وحركية فئة الأجسام، ودلالات المسارات، والانحرافات السياقية للمشهد. يكتشف النظام كلًا من الأنشطة البشرية غير الطبيعية (مثل التسلل إلى مناطق غير مصرح بها، أو الاضطرابات الحركية المفاجئة) والأحداث المرورية الشاذة (مثل القيادة في الاتجاه المعاكس، أو اختراق المحيط) من خلال التحليل الزمكاني المتكامل.واجهة المستخدم الرسومية (GUI) تتيح للمستخدمين تحديد السيناريوهات غير الطبيعية. حيث يمكنهم تخصيص العوامل الشاذة، بما في ذلك وقت الحدث، ونوع الجسم، واتجاه الحركة غير الطبيعية، بالإضافة إلى رسم مناطق الاهتمام. يتم استخدام شبكة العصبية العميقة " Single Shot Multi-Box Detector (SSD_MobileNet_v3)" لاكتشاف الأجسام داخل إطارات الفيديو. بعد ذلك، يتم استخدام أداة التتبع " Kernelized Correlation Filter (KCF)" لمراقبة هذه الأجسام وتحديد اتجاه الحركة غير الطبيعي.تمثل إحدى الابتكارات في هذا النظام قدرته على تصنيف أنواع الحركة البشرية من خلال إقامة علاقة بين الأبعاد الفعلية للأفراد والمسافات المرصودة في الفيديو. علاوة على ذلك، تم إنشاء مجموعة بيانات لاكتشاف السلوك الشاذ. تم تقييم كفاءة النظام على مجموعة بيانات فيديو المراقبة المشوهة حقيقيًا والمجموعة التي تم الحصول عليها. حققت الطريقة المقترحة معدل استدعاء (recall rate) بنسبة ٩٥,٨٣٪ لاكتشاف الشاحنات و ٩٢,٤٢٪ لاكتشاف التسلل سيرًا على الأقدام. كما سجلت درجة F1_score بنسبة ٨٧٪ لتصنيف الحركة، مع سرعة معالجة متوسطة تبلغ ١٦ إطارًا في الثانية (fps). تظهر نتائج البحث أن الاستراتيجية المقترحة تتفوق على الإجراءات الحديثة الحالية.

الكلمات المفتاحية : مراقبة الفيديو باستخدام كاميرات متعددة، اكتشاف الاتجاه، SSD_MobileNet_v3، تصنيف الحركة البشرية.

1. INTRODUCTION

The video surveillance systems were designed for human operators to observe protected space remotely. Watching surveillance video is a tedious task, and human observers can easily lose attention. Automation can help to reduce the manpower and increase the performance. It is very important to develop an intelligent video system that utilizes technology to automatically detect multiple abnormal behaviors.

Human activity recognition (HAR) is major for improving security in various environments, including public spaces, transportation locations, and residential areas. Intelligent road surveillance systems improve safety by providing real-time warnings and monitoring, which helps lower accident rates and enhance overall road safety. By analyzing vehicle patterns and adjusting traffic signals as needed, these systems can optimize traffic flow, reduce congestion, and shorten commute times. Surveillance systems are typically classified into three functional tiers: basic (motion detection), intermediate (object/face recognition), and advanced (behavior analysis) [1]. Our work advances the advanced tier by introducing a multi-dimensional behavior analysis system that integrates temporal, spatial, and kinematic analysis.

Our AI-driven surveillance system classifies abnormal events through five operational dimensions: (1) Temporal anomalies (e.g., unauthorized access during closed hours), (2) object-type violations (e.g., pedestrians on highways or trucks in pedestrian zones), (3) spatial breaches (e.g., fence climbing requiring both human classification and precise on-fence positioning), (4) directional deviations (e.g., wrong-way vehicle movement or counter-flow pedestrian traffic), and (5) kinematic irregularities (e.g., sudden erratic running indicative of criminal activity or emergencies). Many events depend on multiple concurrent factors, like climbing detection, which requires both object-type (human vs. animal) distinction and precise spatial (on-fence) verification. Our GUI-enabled system provides granular control over all detection parameters: configurable time ranges (e.g., nighttime monitoring), multi-region selection with distinct type-specific rules (persons/vehicles), directional constraints, and simultaneous monitoring of up to three camera feeds (each supporting multiple polygonal regions of interest).

An efficient activity recognition system for surveillance applications has been proposed [2]. A lightweight Convolutional Neural Network (CNN) is employed for feature extraction from video frames. A Dual-Stage Gated Recurrent Unit (DS-GRU) is utilized to capture temporal dependencies in the video data, allowing the model to understand how actions progress over time. This work includes many challenges like occlusions, complex interactions, environmental conditions, and motion blur. The model's performance is evaluated using UCF101, HMDB51, and custom datasets. A real-time system based on OpenPose is proposed for detecting and tracking multiple individuals in image and video streams [3]. A CNN is employed to generate confidence maps for distinguishing body parts. Part Affinity Fields (PAFs) are utilized to connect detected key points into comprehensive pose estimates, which are 2D vectors. There are several limitations in this work. Occlusion due to body parts overlap, camera perspective because of poor estimations from extreme angles, variability in body shapes, and low resolution affecting image quality are examples of these limitations. Also, fast movements cause motion blur and reduce detection accuracy. An intrusion detection method for restricted areas has been proposed [4]. The You Only Look Once (YOLO) detector is employed to determine the pedestrian's location at a certain interval. Furthermore, the

location is monitored using a Kernelized Correlation Filter (KCF) [5]. The location information for the intrusion and roving area is collected from a specified XML (Extensible Markup Language) file. The input image's region of interest (ROI) is determined using these coordinates. The tracking object's coordinate info is obtained to identify if it is inside or outside the ROI area. Intrusion is determined by multiplying PersonN (x, y) × ROI (x, y). Any intersection between the detected person and ROI will give a value and count intrusion. Thus, one of the drawbacks of this method is the difficulty of differentiating between a partial and total intrusion. Dohun Kim's method is evaluated using the KISA dataset, which is clear except for some videos that have low-light environments. A human pose-based speed detection method is introduced, utilizing key point information extracted from images [6]. The needed important points must contain at least six different points on the leg. Also, the static view from the side is necessary. The angle between the runner's legs can be obtained and utilized to determine the actual running frequency. Furthermore, the true distance among the two legs is simply calculated using ratio conversion.

There's an increasing demand for surveillance solutions that prioritize individual privacy while still maintaining public safety. In this paper, an abnormal behavior detection system is proposed to cover most abnormal situations with high performance and fewer constraints.

In the beginning, the system starts surveilling or not depending on comparing the abnormal time with the actual time. Then, all frame objects are detected using the SSD_MobileNet_v3 technique [7]. For each object, the system checks if the object type is abnormal for the regions of interest or not. It also checks if this object is inside this region or not and if it is partially or totally included. After that, it tracks the object to determine its direction employing KCF. Moreover, the abnormal human speed is detected. A relationship between the real distances (average human width in meters) and the observed distances in the video (bounding box (BB) width in pixels) is established in order to compute the speed. If the object is included in the region of interest, the system will give an alarm. Also, if the direction is abnormal or the speed is abnormal, the system will give an alarm.

This work advances intelligent surveillance systems through five key contributions: First, we present a unified framework that processes diverse detection scenarios through a single optimized system combining SSD_MobileNet_v3 for object classification with KCF tracking. The system's intuitive graphical interface allows users to: (i) define arbitrary polygonal regions by manually drawing shapes with automatic vertex extraction, (ii) configure multiple simultaneous regions of interest (ROIs) within a single camera view, each with distinct abnormal scenario parameters, and (iii) simultaneously analyze all five critical dimensions (temporal, object-type, spatial, directional, and kinematic). This integrated approach enables fast, flexible operation (processing three camera feeds in real time) and outperforms fragmented systems relying solely on human shape detection [2, 3, 6].

Second, our privacy-preserving design achieves robust performance (approximately 82% recall with 93.6% precision, equating to only 6.4% false alarms) without requiring invasive pose estimation [3, 6] or fixed human positioning or side camera views [6]. Third, the system maintains reliable performance under challenging real-world conditions (occlusions and low-light environments) where existing approaches fail due to obscured pose keypoints during occlusion events [2, 3, 6]. Fourth, the solution differentiates between partial and complete intrusions - a critical capability for minimizing false alarms in fence climbing scenarios unlike conventional approaches [4]. Finally, we evaluate system performance using a comprehensive dataset combining the UCSD Anomaly Dataset with newly collected real-world footage from YouTube and Google Images. The Authentically Distorted Surveillance Videos dataset includes challenging conditions such as defocus aberration, overexposure, sub-exposure, and combined defocus-exposure artifacts. Experimental results demonstrate strong performance, with 95.83% recall for truck detection and 92.42% for walking intrusion detection in video analysis (varying by intrusion type: prowling, walking, or running), and 89.33% recall for image analysis. These innovations collectively advance both the technical capabilities and practical implementation of modern surveillance systems.

The remainder of this paper is structured as follows: Section 2 provides an overview of abnormal video detection approaches and popular current datasets. Section 3 suggests an innovative architecture for anomalous video detection. Section 4 discusses the experimental results. Section 5 discusses the conclusion and future work.

2. RELATED WORKS

The progress of deep learning algorithms has raised the ease of abnormal action detection within video. Various approaches have been developed to detect these actions relying on either traditional methods or the advanced deep learning-based approaches. CNN, and Recurrent neural networks (RNN) are examples of powerful tools which can be employed for this purpose.

Abnormal intrusions are detected using YOLOv4 [4]. The 3D-CNN is used for fall and violence detection. The suggested technique is evaluated using KISA Datasets and an average Recall Rate of 93.8% is achieved. Building on object detection foundations, more complex behavioral analysis has been developed.

A deep learning system for detecting suspicious activity in surveillance videos is developed [8]. The framework consists of two Proceedings: CNN and RNN. VGG-16 CNN is utilized for fracture extraction from video frames. LSTM is used for order dependence learning in the video frame sequence. An accuracy of 87.15% is achieved for CAVIAR and KTH datasets. This temporal modeling approach has been further refined through adversarial learning techniques. An adversarial 3D convolutional auto-encoder for robust normal spatio-temporal pattern extraction has been proposed [9]. It consists of a 3D convolutional encoder and a 3D de-convolutional decoder that are used to encode the normal patterns in video sequences. An average accuracy rates of 90.6%, 92.7%, and 74.6% are achieved for UCSD, SUBWAY, and SHANGHAITECH, respectively. The success of autoencoder architectures has inspired further innovations in loss function design.

Furthermore, Bouindour et al. developed a 3D Convolutional Autoencoder with a novel loss function combining MSE and compactness Mahalanobis loss [10]. This loss function helped in extracting robust spatiotemporal features while minimizing the hypersphere that encompasses the target class representations. While these approaches focus on feature learning, other works have integrated multiple detection tasks. Face recognition and intruder detection are implemented using TensorFlow with AlexNet architecture [11]. Fires are detected using motion and color information. Balance point change, angles and movement distances of objects are used for Loitering detection. Finally, falls are detected using motion and acceleration features of the object. The system accuracy is 88.51% for intruder detection, 92.63% for fire detection, 80% for loitering detection, and 93.54% for fall detection.

The need for comprehensive appearance and motion analysis led to more sophisticated hybrid architectures. Video anomalies are detected using a CNN-based appearance encoding framework that processes individual frames [12]. Convolutional Long Short-Term Memory (ConvLSTM) is utilized for memorizing all previous frames to extract motion information. ConvNet and ConvLSTM are combined with Auto-Encoder to learn the regularity of appearance and motion for ordinary moments. ConvLSTM-AE accuracy rates are Avenue 77%, Ped1 75.5%, Ped2 88.1%, Subway Entrance 93.3%, and Exit 87%. While ConvLSTM-AE showed promising results for general anomaly detection, subsequent work has focused on improving feature learning mechanisms.

Moreover, Irregular frames are detected through learned regular patterns using autoencoders [13]. The process began by learning a fully connected autoencoder using conventional spatiotemporal local features. Then, in a single learning framework, a fully convolutional autoencoder is constructed to learn both the local features and the classifiers. The accuracy is 70.2% for CUHK Avenue, 81.0% for UCSD Ped1, 90.0% for UCSD Ped2, 94.3% for Subway Entrance, and 80.7% for Subway Exit.

Adversarial event prediction (AEP) is presented in [14]. AEP derives the prediction model from normal event samples so that it can identify the relationship between the current and future course of events during the training phase. Adversarial learning for both the past and future of events is introduced to acquire the prediction model. The suggested adversarial learning constrains the learning for past events representation and compels AEP to learn the representation for forecasting future events. The success of adversarial learning for temporal prediction inspired new directions in activity-specific recognition.

Furthermore, A system for abnormal human activity recognition was proposed [15]. The Bayes Classifier and VGG-16 are employed to differentiate between walking, running, punching

and tripping. The extracted features are length, width ratio, entropy, and Hu invariant moment. KTH dataset is used for evaluation. The recognition accuracy based on Bayes reached 88%, 92%, 92% and 100% for each activity. The recognition accuracy based on CNN reached 92%, 96%, 100% and 100% for each activity.

Fence Climbing is detected Using Activity Recognition [16]. Also, person motion is detected using background subtraction. The researcher concerned with two main features. These features are centroid of the blob and centroid variations. The Support Vector Machine (SVM) classifier is applied to classify walking, climbing down, and up actions.

Walking is a common human activity. Thus, if a human runs suddenly, it may indicate that an abnormal event has occurred. Also, speed detection is used to track physical fitness. Therefore, many algorithms were created to decide people's motion type.

A benchmark database with diverse scenes and ground truth annotations for human running detection was introduced [17]. The researcher developed a method based on the Farnebäck optical flow method to distinguish between running and walking. The system reached an average precision and recall rate of 67.8% and 90.1%, respectively.

Furthermore, a human pose-based system is designed to extract key-point information from images for speed measurement [6]. The needed important points must contain at least six different points on the leg (left_knee, left_ankle, left_hip, right_knee, right_hip, right_ankle). Also, A static point of view from the side is required. to extract the key points. The angle that separates runner's legs could be determined. employing the key point information obtained. The angle information is used to establish the actual running frequency. The ratio conversion may then be used to simply determine the actual distance separating the two legs. Using the MPII data set, experiments show that in some circumstances, the detection accuracy of the Simple Baselines approach is 89.3%. Additionally, the speed sequences' average relative error of 4.89% satisfies the realistic detection criteria.

CNN-based speed detection (walking/running) is implemented using wrist-worn wearable sensors with high precision [18]. Data from 15 participants is collected while they are walking/running at different speeds on a treadmill. Accelerometer output and gyroscope sensory output from the wrist-worn device are the inputs for CNN individually. The max pooling outputs of accelerometer and gyroscope sensory data are concatenated and fed into the dense layer, then the output layer for speed classification.

Anomaly detection is formulated as a non-linear dynamical system through fusion of MobileNetV3 feature extraction and Bi-LSTM temporal modeling, enabling quantitative behavior instability assessment [19]. Cross-camera tests on UCF-Crime show 94.1% attack detection at a 0.1% false alarm rate, outperforming transformer-based approaches in low-light scenarios by 11.3% mAP.

Moreover, Generative Adversarial Nets (GANs) are a wonderful way to solve classification issues because they can identify important features in the frames without the need for predefined anomaly categories. GAN-based framework demonstrates robust abnormal event detection, achieving 97.4%, 93.5%, and 99% accuracy on UCSD Ped1, Ped2, and UMN datasets respectively [20]. The Deep Spatiotemporal Translation Network (DSTN) framework combines deep convolutional neural networks with GAN-based Edge Wrapping to enhance anomaly localization, achieving state-of-the-art accuracies of 98.5% (UCSD Ped1), 95.5% (UCSD Ped2), and 99.6% (UMN) [21]. Additionally, A surveillance framework for robust outdoor object detection in dynamically changing complex environments was proposed [22]. They achieved this using a modified Gaussian Mixture Model (GMM) and adaptive thresholding, effectively addressing shadows and partial occlusions.

Supervised and unsupervised machine learning approaches for suspicious behavior recognition in surveillance systems have been systematically analyze [23]. The study systematically categorizes existing approaches, highlighting their strengths and limitations in detecting anomalies. It discusses feature extraction methods, model architectures, and benchmark datasets used in behavior analysis. The paper serves as a valuable reference for understanding the evolution of machine learning in intelligent surveillance applications.

A hierarchical autoencoder architecture for unsupervised spatiotemporal anomaly detection has been developed [24]. Their three-stage framework achieves state-of-the-art performance of 89.2% AUC on ShanghaiTech Campus, with particular success in occlusion handling (83.4% recall vs. 71.2% in comparable works). The authors demonstrate that their temporal attention mechanism reduces false positives by 19% compared to conventional autoencoders. While computationally efficient (23% faster than 3D-CNN alternatives), the paper acknowledges remaining challenges in adapting to extreme illumination changes.

A traffic anomaly detection framework (MEDAVET) utilizing spatiotemporal modeling was developed for complex highway environments [25]. It employs bipartite graphs for vehicle tracking, the Convex Hull algorithm to define movement zones, and QuadTree structures to handle occlusions and stationary objects. Evaluated on the UA-DETRAC and Track 4 benchmarks, MEDAVET achieved an F1 score of 85.71%. While MEDAVET focuses on anomaly detection through computer vision, other AI approaches target traffic optimization through adaptive control systems, Saif et al. developed an artificial intelligence-driven traffic control system utilizing reinforcement learning (RL), graph convolutional long short-term memory (GCN-LSTM), and genetic algorithms (GA), achieving an 84.8% reduction in vehicle waiting times while maintaining emergency vehicle prioritization [26]. Their model adapts dynamically to spatio-temporal traffic patterns. Beyond traffic signal optimization, reinforcement learning is also proving effective for autonomous vehicle control. A reinforcement learning-based system that integrates adaptive cruise control with lane-keeping functionality, significantly improving autonomous vehicle stability has been developed [27]. Their approach demonstrated 37% smoother lane-centering and 28% faster collision avoidance in simulations compared to conventional methods. While these studies address traffic monitoring and control optimization, autonomous navigation requires equally advanced perception capabilities. A high-precision lane detection system for autonomous vehicles that combines geometric modeling through second-order polynomial curve fitting with YOLO-based object detection has been developed [28]. Their solution achieved 98.64% accuracy on Tusimple and 96.92% on KITTI datasets, enabling real-time road geometry analysis and trajectory prediction. This dual approach significantly enhances AV situational awareness in diverse driving conditions.

Our framework advances surveillance systems by addressing critical gaps: (1) Unlike rigid systems [4, 8, 16], we provide unmatched user flexibility through an intuitive GUI enabling custom abnormal scenario definition with multi-ROI configurations and unique rules per zone. (2) Where [2] offers single-tier analysis, we integrate temporal, spatial, and kinematic dimensions. (3) While [3] fails under occlusion (67.8% precision), our system maintains 93.6% precision in real-world conditions. (4) We solve [4]'s partial intrusion limitation through pixel-precise ROI analysis. (5) Unlike [6]'s fixed side-view requirement, our method works across perspectives (87% F1-score). (6) Hybrid models [9, 14] improve spatiotemporal modeling yet lack interpretability and adaptability to custom scenarios.

3. THE PROPOSED SYSTEM

The proposed algorithm framework is displayed in **Fig. 1**. As the object type (persons, trucks, vehicles, persons& trucks, or all) is the main essential factor that decides if the motion is normal or not. For example, in fence climbing, if it happens by a cat, it is normal, but if it happens by humans, it is abnormal. There are two main categories of object type detection methods: single-stage algorithms based on regression, like YOLO [29] and SSD [30], and two-stage algorithms based on proposed regions, such as Faster R-CNN [31].

This work employs the SSD_MobileNet_v3 architecture for object detection tasks [7], as shown in **Fig. 1,** because it is faster than YOLO while the detection performance is approximately the same. SSD employs a pyramid feature layer-based detection technique on feature maps of various sizes, performing both location regression and SoftMax classification.

SSD also uses Faster R-CNN's anchor concept with variable scales, aspect ratios, and previous frames. This will make it more accurate in identifying and localizing objects of varying sizes. SSD is a robust object detection technique based on a feed-forward CNN. A set of bounding boxes and associated scores are generated, indicating the presence of object class instances in these boxes. The network in the frontend and an extra feature extraction layer at the backend compose

the SSD network structure. To address the high computational cost and large parameter size of VGG-16, the more efficient MobileNet-V3 architecture is implemented [32].

Moreover, the proposed algorithm has the ability to ascertain the type of motion of the individuals (i.e., running or walking) and its directionality. Walking is a common human activity. So, if a human runs suddenly, it is an indication that an abnormal event has occurred. Abnormal event detection sometimes depends on the motion direction (up, down, left, right, up and right, and stop). For example, it is not allowed for cars to move in the opposite direction on the road.

As KCF has gained widespread acceptance as a tracking algorithm. It is suitable for raw pixel values in addition to the histogram of oriented gradient features. Consequently, KCF tracker is used to track the objects to determine their motion type and direction due to its high speed and accuracy. KCF is a variant of correlation filter in which the correlation between two samples is computed. The correlation scores the highest value when these samples match. The proposed algorithm provides a multi-camera surveillance service. The system can cover 3 cameras, each having all the abilities we described above.

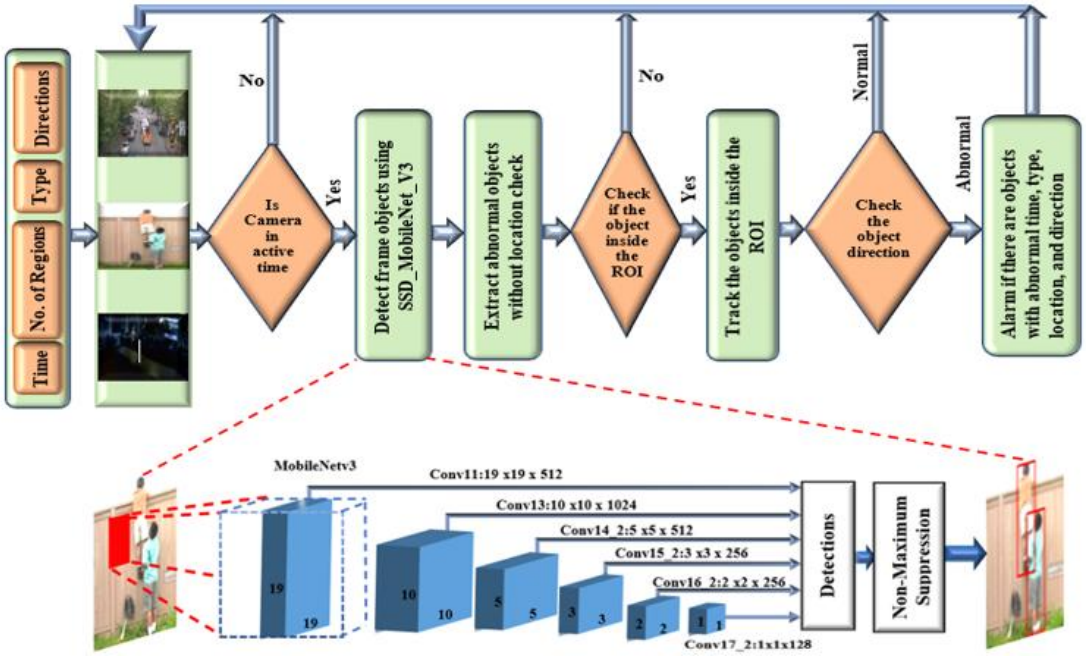


Fig. 1: System Flow chart with SSD_MobileNet_V3.

The proposed scheme can be summarized in the following steps:

A. At first, the user provides the system with the required information through a flexible GUI. The user should provide a working time interval, the number of regions he needs in each camera, and a list of abnormal object types for each region. Additionally, the object's motion type and motion direction should be provided. The user can draw a polygon for each region of interest as shown in Fig. 2, the system extracts the polygon vertices and saves them in a three-dimensional list. The camera index is represented using one of these dimensions, while the other two dimensions represent the region index and the polygon vertices in that region.

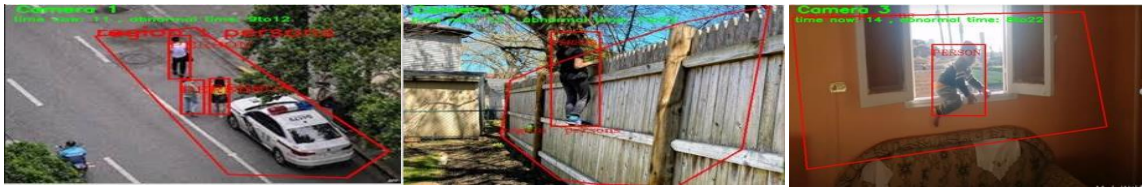
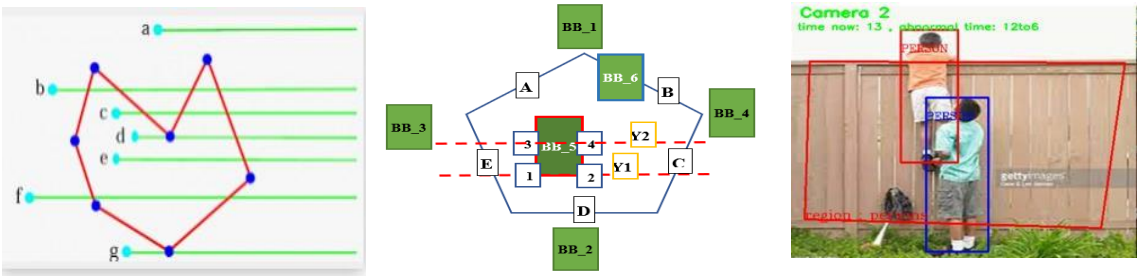


Fig. 2: Detect certain abnormality in a restricted region.

B. Then, SSD is employed to detect object types.

C. Using the information in the previous steps, the system can check if the object's bounding box is outside the ROI or inside the ROI partially or totally as in **Fig. 3a**. If all the BB points of the object are outside the polygon, the object is considered outside the ROI and is excluded (do nothing). While if all the BB points of the object are inside the polygon, the object is considered inside the ROI and is included for further investigation (alarm) as BB_5 in **Fig. 3b**. If some points of the BB are outside and other points are inside, like BB_6 in **Fig. 3c**, the object is partially inside (warning).



a. **b.** **c.**

Fig. 3: **Determine BB position corresponding to ROI.**

D. KCF tracker is employed to track these objects to determine their motion type and direction.

E. To recognize the motion type for human objects, a novel method is employed. The main idea of the method is to establish a relationship between the observed distances in the video and the real distances. Most adults have a pacing distance between 61 and 91.4 cm [33]. **Fig. 4a**, and an average body breadth is between 58 and 65 cm [33], as shown in **Fig. 4b**. Assuming the average human width is 0.65 meters for both the front view and the pacing distance for the side view. Consequently, we assume that the human width is one meter after wrapping BB around him. The SSD_MobileNet_V3 detector and KCF tracker provide bounding box (BB) locations and dimensions for consecutive frames, which are utilized to compute motion characteristics. For each pair of consecutive frames ($\Delta t = 0.05s$ at 20fps), two bounding boxes are analyzed: BB_1 (first frame) and BB_2 (second frame). The average BB width is calculated in pixels , as shown in Equation 1:

$$D_{1\_pixels} = (width_1 + width_2)/2 \quad (1)$$

This averaging process mitigates potential minor variations in BB dimensions that might occur due to distance changes from the camera, while the high frame rate ensures minimal inter-frame perspective alterations. The calculated value D_{1_pixels} serves as a stable reference for subsequent speed computations between frames.

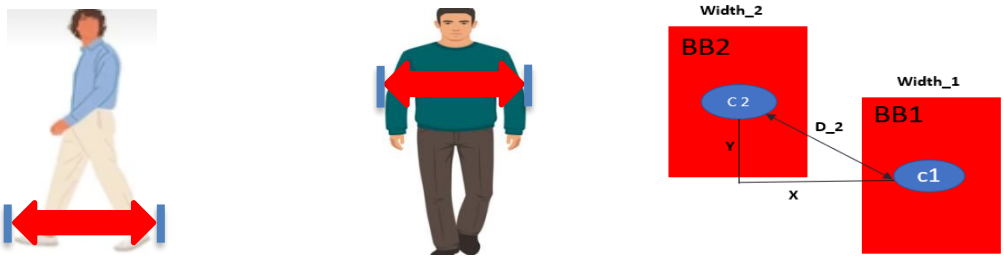
Then, the distance between the centers of BB1 and BB2, which represents the distance he ran in pixels, is computed , as shown in Equation 2:

$$D_{2\_pixels} = \sqrt{(X^2 + Y^2)} \quad (2)$$

As shown in **Fig. 4c**. If we assumed that D1_meter equals 1 meter, we could get, $D_{2\ meter}$, as shown in Equation 3:

$$D_{2_{meter}} = \frac{D_{2_{pixels}} * D_{1_{meter}}}{D_{1_{pixels}}} \quad (3)$$

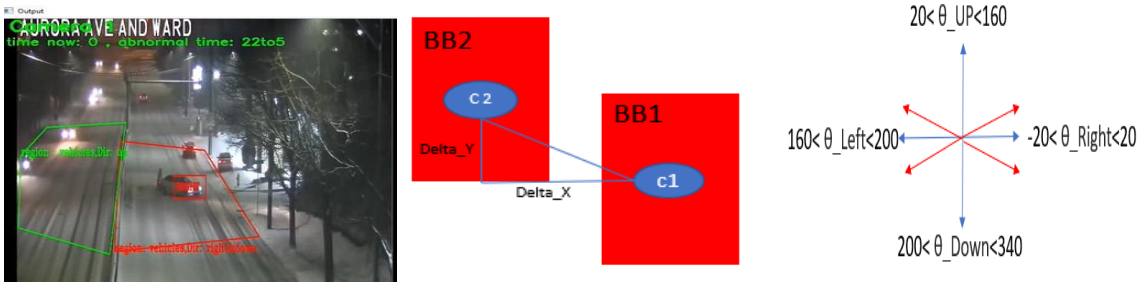
The moving distance time is equal to the time between two consecutive frames. Finally, with the distance traveled in meters known ($D_{2_{meter}}$) and the time interval between consecutive frames established ($\Delta t = 0.05$ s), we can calculate the speed of the individual Human speed $= \frac{D_{2_{meter}}}{\Delta t}$. The average human walking speed is 1.4 m/s [34]. So, any speed above 3 m/s is considered running.



a. Person pacing zone **b.** Average body breadth **c.** Distance calculation

Fig. 4: Human motion speed determination

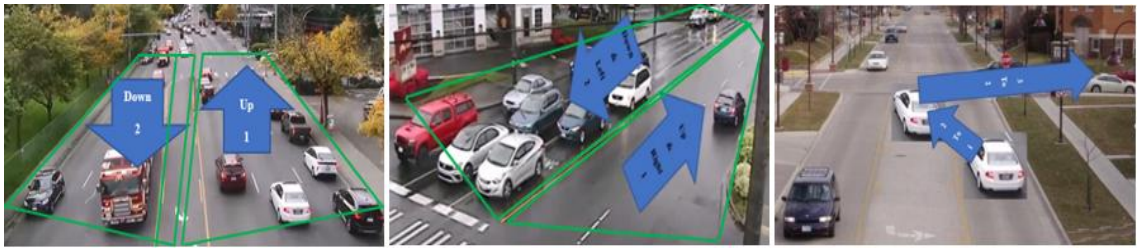
F. The system identifies abnormal motion directionality by analyzing the displacement between consecutive BB centroids. **Fig. 5 a.** For each tracked object, the planar displacement components (Δx, Δy) are computed from the BB centroid coordinates (**Fig. 5 b.**), and the motion direction angle $\theta = \arctan(\Delta y/\Delta x)$ is derived (**Fig. 5 c**). The definition of natural versus unnatural motion is context-dependent, determined by the camera's perspective relative to the road geometry as shown in **Fig. 5 d,e.** For instance, as shown in **Fig. 5 d**., in Area 1, upward motion is classified as natural, while downward motion is unnatural, with the opposite convention applied to Area 2. Empirical testing revealed that an initial angular threshold of 45° resulted in false positives during lane changes, as shown in **Fig. 5 f.,** as such maneuvers typically occur at approximately 45° without representing true directional changes. Consequently, the threshold was refined to ±20° (40° total range), which effectively distinguishes between lane deviations ($\theta \approx 45°$) and genuine directional changes ($\theta \approx 90°$). Moreover, to avoid the common effect of camera vibrations in outdoor settings, we determine the direction only after it has stabilized for four consecutive frames.



a. Vehicle's wrong direction detection

b. Distances calculations

c. Direction calculation



d. Up/Down Bidirectional Road

e. Up - Right/Down - Left Bidirectional Road

f. Lane Changes vs. Direction Changes

Fig. 5: Direction detection

4. EXPERIMENTAL RESULTS

This section displays the dataset details, evaluation criteria, and the conducted experiments as follows.

4.1. Dataset

The system performance is evaluated on both the newly collected dataset and the Authentically Distorted Surveillance Videos dataset [35]. The newly collected dataset consists of 75 video clips for outdoors, which are gathered from YouTube in order to test our algorithm in terms of vehicle directionality and the presence of objects in prohibited locations. For example, this includes instances where a truck is located in a lane designated for passenger vehicles, pedestrians on roadways designated for vehicles, and vehicles traveling in the wrong direction. The dataset is authentic and encompasses the challenges typically faced by surveillance cameras, such as vibrations and changes in lighting conditions over time. 45 videos of them are abnormal events, and the rest are normal. In addition, there are 150 images that are collected from Google images and UCSD_Anomaly_Dataset for different events.

1000 video clips from the Authentically Distorted Surveillance Videos dataset are included for indoor events. Defocus aberration (defocus), overexposure, subexposure (exposure), and a combination of defocus and exposure are the genuine distortions affecting these recorded videos. After that, the merged dataset is divided into two parts: the image dataset and the video dataset.

The Video Dataset: There are 330 videos for Person Running Intrusion (PR) and 330 videos for Person Walking Intrusion (WL). Moreover, the Person Prowl intrusion (PW) class has 340 videos in which the person makes suspicious movements. In addition, the newly collected dataset is composed of 24 videos for truck detection and 51 videos for vehicle's wrong-direction detection. **Fig. 6** depicts samples from the newly collected dataset.

Fig. 6: Videos dataset samples

The Image Dataset: The collected images dataset comprises diverse anomalous scenarios, including 20 images for tampering with antiquities in museums, 40 images for pedestrians walking on highways, 50 images for vehicles on pedestrian walkways, 20 images for infants climbing window or balcony fences, and 20 images for unauthorized person intrusion. These images are characterized by events that require high precision in determining the bounding box (BB) location concerning the area of interest. These images were used to assess the efficacy of our algorithm. **Fig. 7** presents samples from the image dataset.

Fig. 7: Images dataset samples

4.2. System GUI

To facilitate the interaction between the user and the proposed system, a GUI is designed as in Fig. 8. The user can enter all abnormal conditions for each camera. The GUI screen is divided into five columns. The user can enter the surveillance time range in the first and second columns. In the third column, the user can provide the number of regions he needs to construct each camera along with a list containing abnormal object types for each region in the fourth column. Finally, in the fifth column, the user can provide a list containing the abnormal motion direction for each region.

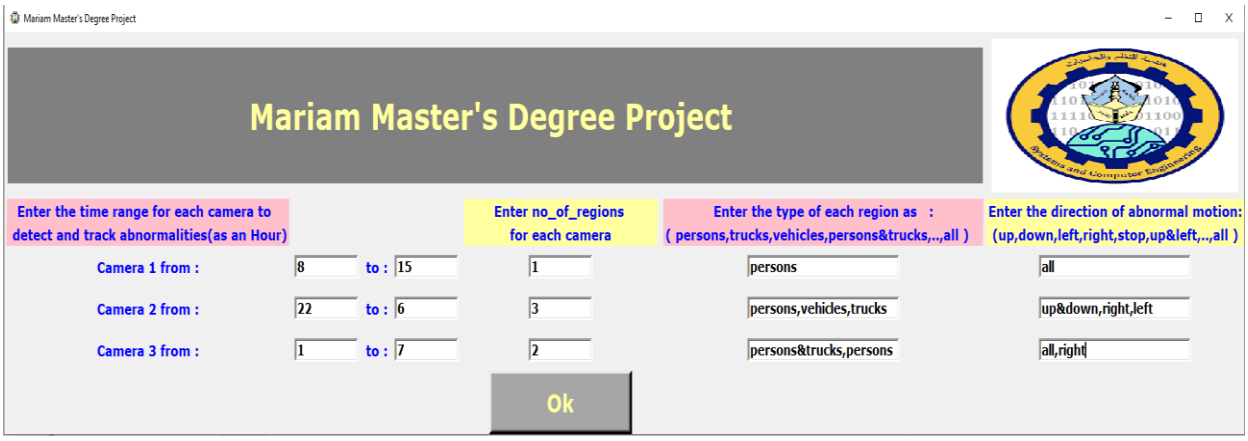


Fig. 8: Abnormal behavior detection GUI

4.3. Evaluation Measures(Metrics)

The accuracy, recall, precision, and F1 score are the prominent evaluation metrics that are used to evaluate the effectiveness of the proposed abnormal event detection method. Speed is also an essential evaluation criterion in such applications. The evaluation measures' Equations 4, 5, 6, and 7 have been defined as follows:

$$\text{Accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{total number of samples}} \tag{4}$$

$$\text{Recall} = \frac{\text{number of true positives}}{\text{number of false negatives} + \text{number of true positives}} \quad (5)$$

$$\text{Precision} = \frac{\text{number of true positives}}{\text{number of false positives} + \text{number of true positives}} \quad (6)$$

$$\text{F1\_score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (7)$$

4.4. Experimental Results and Discussion

The surveillance system efficiency depends on two factors: the system speed, and the abnormal event detection accuracy. These factors will be discussed in the following sections.

4.4.1 System speed

The proposed system implements an optimized surveillance architecture that intelligently manages monitoring requirements through two complementary approaches: first, for events occurring within the same field of view but under different conditions, the system employs multi-region analysis within a single camera feed, establishing distinct ROIs with customized detection rules for each zone (e.g., a roadway with bidirectional lanes and pedestrian zones). This approach proves more computationally efficient than using multiple cameras for spatially co-located events. Our experiments demonstrate that multi-zone single-camera analysis introduces negligible latency per additional ROI; second, for events requiring monitoring across separate physical locations, the framework supports simultaneous processing of up to three camera feeds. The computational performance of the surveillance system is influenced by several key factors: the choice of detector (SSD_MobileNet_V3 vs. YOLOv3) as benchmarked in **Table 1**, the number of active cameras and tracked objects, and the input frame resolution as analyzed in **Fig. 9** and its description.

As quantitatively demonstrated in **Table 1**, our SSD_MobileNet_v3 implementation achieves significantly faster processing times compared to YOLOv3, with experimental results showing a 2.5× speed advantage. While existing surveillance systems have employed YOLOv3 architectures [36,37], our selection of SSD_MobileNet_v3 prioritizes computational efficiency for edge deployment scenarios. Notably, the performance gap between these architectures becomes more pronounced at higher resolutions, as SSD_MobileNet_v3 demonstrates superior scalability in resolution-sensitive applications. As shown in **Table 1**, the time decrease percentage (Equation 8) further supports this observation.

$$\text{Time decreased percentage} = \frac{RE solution\ 2\ time - resolution\ 1\ time}{resolution\ 1\ time} * 100 \quad (8)$$

Table 1: SSD_MobileNet_V3 vs. YOLO_V3 Time Efficiency

Model	Resolution :580*326	Resolution :1280*720	Average frame analysis time	Average no. of frame analysed/sec	Time decreased percentage
YOLO_V3[23, 36, 37]	0.174	0.211	0.1925	5.19	21.264%
SSD_MobileNet_V3	0.0731	0.080	0.07655	13.06	9.986%

Another study is carried out to detect the effect on the analysis time if there is an increase in the number of cameras, the video resolution, and finally the number of tracked objects. The study is conducted using SSD_MobileNet_V3 for detection and KCF for tracking. The research is conducted on two videos with different resolutions. The first resolution is twice the second. The first video includes only one person, while the second video includes eight persons.

As shown in **Fig. 9**, the orange dots represent the system's speed when operating with a single camera, while the yellow dots indicate its performance with two cameras. It is observed that the speed decreases by an average of 3 fps under the same conditions. For instance, in the case of low-resolution video while tracking one object, using a single camera yields a speed of approximately 21.4 fps, whereas using two cameras results in 19.6 fps under identical conditions. Similarly, when the system operates with three cameras instead of two, the speed further decreases by an additional 3 fps, reducing the system's performance to approximately 16.4 fps under the same previously mentioned conditions.

This figure also illustrates the impact of video resolution on the system's speed. It is evident that when using a single camera with low-resolution video, the speed reaches about 21.4 fps. However, doubling the resolution reduces the speed to 16.4 fps.

Additionally, the effect of the number of tracked objects on speed was examined. When using a single camera with low-resolution video to track one object, the speed is 21.4 fps. However, if the number of tracked objects increases to four under the same conditions, the speed drops to 17.8 fps. Finally, when tracking eight objects under identical settings, the speed further declines to 13.8 fps.

The system frame rate depends on the detection rate for high-resolution videos (Authentically Distorted Surveillance Videos dataset). The system frame rate reaches 14 fps for the detection rate D/3F, 16 fps for the detection rate D/15F, and 17 fps for the detection rate D/30F, as shown in **Table 2**.

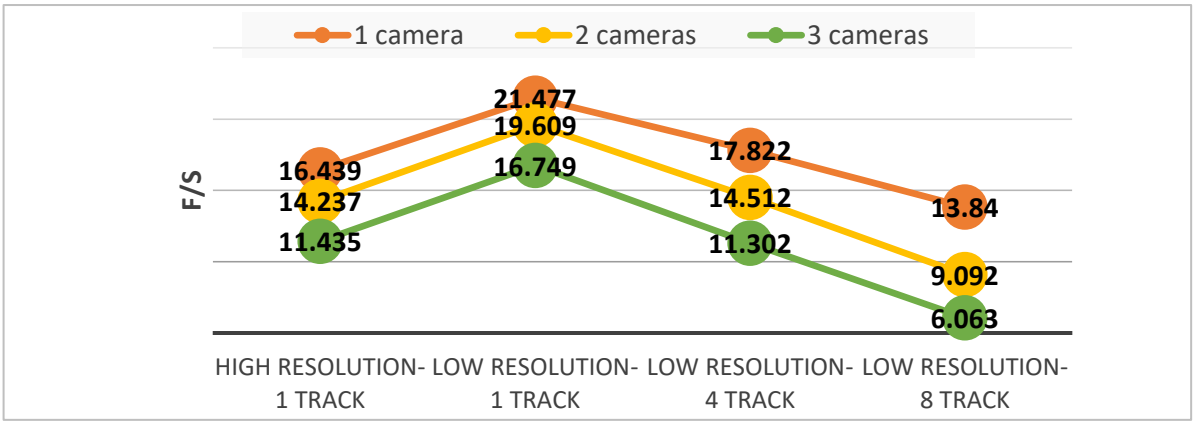


Fig. 9: System speed affected by the number of cameras, the video resolution, and number of tracking objects

Table 2: System Speed Affected by Detection Rate for High Resolution Dataset.

Detection rate	Frame rate (fps)
D/3F	14
D/15F	16
D/30F	17

4.4.2 Event detection recall rates

There are many types of intrusion motion, and each of them makes the human shape differ in the image. As we can see, there are three types of intrusion motion: walking, prowling, and running. Each of them has a different detection accuracy. Walking intrusion detection reaches the highest event detection rate with the distorted dataset compared with other intrusions due to the normal human shape in this case. Moreover, Prowl Intrusion has the lowest event detection rate compared to other intrusions due to the abnormal human shape in this case. Also, the video distortion causes loss in object detection in many frames. Run intrusion detection accuracy is between walking and prowling due to slightly abnormal human shape and speed. In addition, people may enter and exit without being detected in the small regions.

Roads have many abnormal events. Many roads accept only small vehicles and reject trucks either all the time or at certain times. A vehicle driving in the wrong direction is another abnormal situation. The wrong direction detection depends on a specific area in the road because it is normal to drive up on half of the road while it is abnormal to drive down on the same half. Thus, multi-regions, each with different abnormality conditions, are used. All these situations have been experimentally covered.

At the beginning, the surveillance system looks at the event time to decide if it is normal time or not. In the case of abnormal time, the system checks if the object is inside the zone or not depending on the object type, shape, and direction.

In the case of the image dataset, the conducted experimental results proved that the proposed algorithm scored a recall rate of 89.33%. According to the video dataset, **Table 3** illustrates the accuracy rates at different resolution levels. The recorded results for detection rate of D/15F show that the proposed system achieves a recall rate of 95.83%, 92.42%, 88.23%, 85.75%, and 75.29% for truck detection, walking intrusion, wrong direction detection, running intrusion, and prowl intrusion, respectively.

The direction of movement was identified with 100% precision in the videos where vehicles were detected. Furthermore, for all detected intrusions in the videos, it is identified if they are totally, partially, or not inside the ROI with 100% precision. As indicated in **Table 3**, the results for the detection rate D/15F are comparable to those of D/3F; however, as explained in **Table** 2, D/15F is faster than D/3F. Consequently, we have opted to utilize the detection rate D/15F. The details regarding the recall, precision, and F1 score results associated with the detection rate D/15F are presented in **Table 4**.

Table 3: Videos Dataset Recall Results

Detection rate / Event	D/3F	D/15F	D/30F
Trucks detection	95.83%	95.83%	95.83%
Wrong direction detection	92.15%	88.23%	88.23%
Prowl intrusion	83.14%	75.29%	64.57%
Walking intrusion	94.08%	92.42%	86.60%
Running intrusion	86.73%	85.75%	83.02%

Table 4: Comparative Analysis

Authors	Dataset		Accuracy	Precision	Recall	F1_Score
Kim et al. [4]	KISA Datasets		_	_	93.8%	_
Zhao et al. [6]	MPII data set		89.3%	_	_	_
Sun. et al. [9]	UCSD		90.6%	_	_	_
	SUBWAY		92.7%	_	_	_
	SHANGHAITECH		74.6%	_	_	_
Lao et al. [17]	standard benchmark dataset		_	67.8%	90.1%	_
Mishra et al. [24]	UCSD Peds1		86.4%	_	_	_
	Avenue dataset		88.9%	_	_	_
A. Ullah [2]	UCF-101,		_	86.398%	83.542%	82.120%
	UCF-50,		_	91.294%	89.647%	88.637%
	HMDB51,		_	64.982%	61.952%	60.734%
	Hollywood2, and		_	68.219%	66.846%	65.797%
	YouTube Actions		_	92.637%	91.541%	91.436%
The proposed method	Authentically Distorted Surveillance Videos dataset.	Motion classification	87.71%	93.6%	82%	87%
		Walking person detection	_	88.4%	92.42%	90.37%
		Running person detection	_	79.27%	85.75 %	82.38%
		Prowling person detection	_	60.37%	75.29%	67.04%
	YouTube Street camera footage videos vehicles	Wrong direction	_	100%	88.23%	88.23%
		Truck detection	_	85.18%	95.83%	90.19%
	The images dataset including UCSD		_	87.01%	89.33 %	88.14%

4.4.3 Running detection accuracy rates

The run detection technique is applied to automatically classify the human motion type. The technique's results depend on the tracking algorithm. Many tracking methods have been applied, as shown in **Fig. 10** and **Fig. 11**. According to the experimental results, KCF is the best one compared to Median Flow, CSRT, and Boosting. Boosting has the highest recall rate, reaching 96%. However, it has a very low precision rate of 54.7%, which means half of the alarms are wrong. Consequently, we are concerned with the F1_Score, which is a combination of precision and recall rates. KCF has the highest F1 F1_Score reaching 87%. The recall rate of applying KCF with a detection rate of 15 for run detection reaches approximately 82%. The precision rate reached 93.6%, which means only 6.4% of alarms are wrong using KCF.

To summarize raw counts of TP, FP, TN, and FN at the optimal threshold. We have added confusion matrices to quantify trade-offs between false positives (FP) and false negatives (FN). At a detection rate using the KCF tracker and with a detection rate of D/15F as shown in **Fig. 12**, the model achieves 249 TP and 55 FP, with a precision of 93.61%. Detection rate D/30F reduces FP but increases FN, as shown in **Fig. 13**.

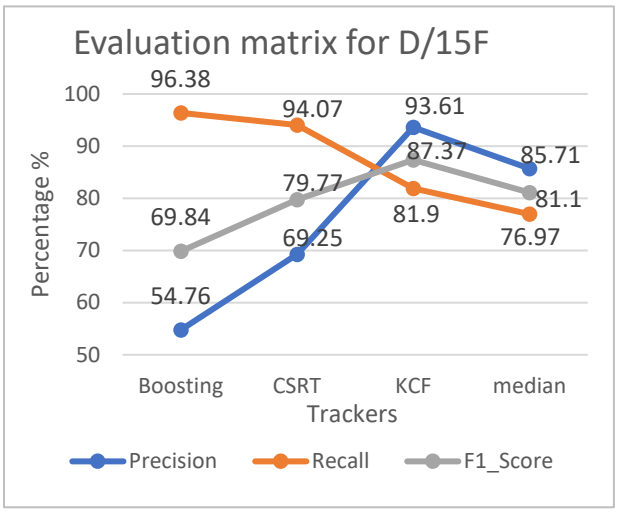


Fig. 10: Evaluation matrix for different trackers with D/15F

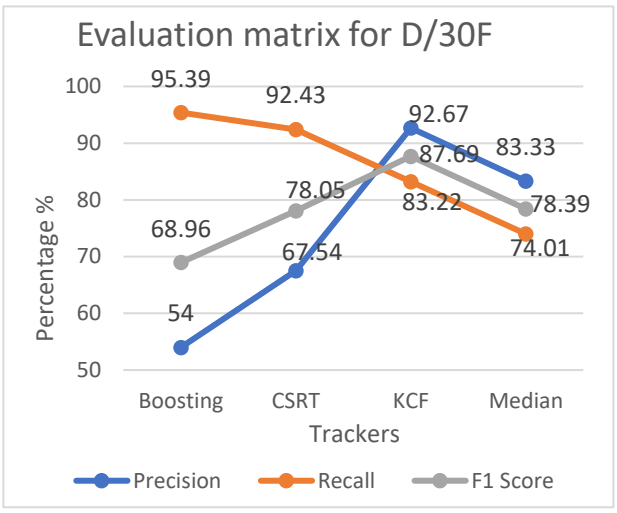


Fig. 11: Evaluation matrix for different trackers with D/30F

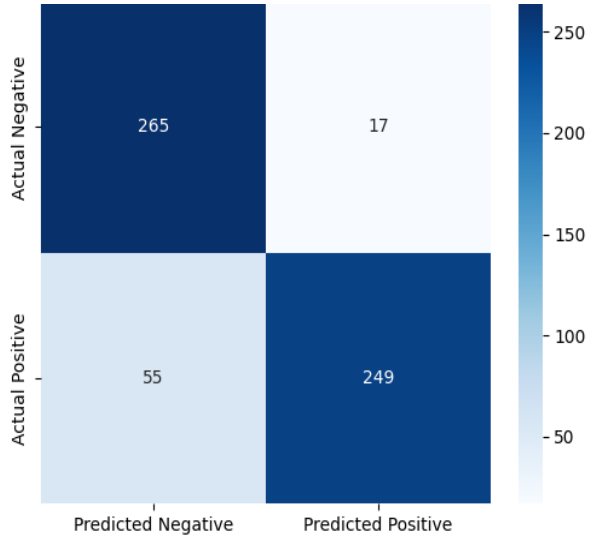


Fig. 12: Confuion matrix for Detection rate D/15F

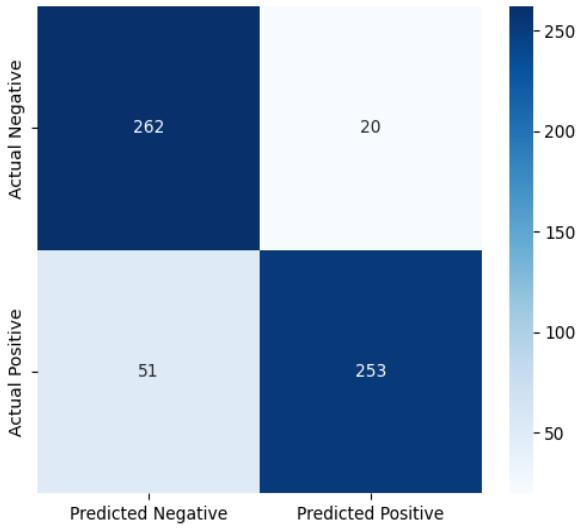


Fig. 13: Confuion matrix for Detection rate D/30F

4.4.4 Comparative analysis:

This section compares the results of the proposed scheme with state-of-the-art methods. The proposed system's multidimensional detection capabilities enable deployment across critical security domains. For critical infrastructure protection, it simultaneously monitors temporal violations (unauthorized after-hours access), critical infrastructure (fence climbing with human/animal differentiation), and kinematic threats (sudden erratic movements). Smart transportation systems benefit from directional analysis (wrong-way vehicle detection) and object-type enforcement (pedestrian-free zones). Urban security applications leverage configurable multi-ROI monitoring, combining temporal rules (park curfews) with spatial-object constraints (vehicle-free pedestrian plazas). The GUI's granular control allows rapid adaptation to diverse scenarios, from private sites requiring precise zone-specific rules to public spaces needing combined temporal-kinematic analysis.

The reported results prove that the proposed method outperforms other state-of-the-art methods. As illustrated in **Table 4**: For motion classification tasks, the proposed system demonstrates superior robustness on distorted surveillance footage compared to existing approaches. While Ullah [2] achieves strong performance on the UCF-50 dataset (91.29% precision, 89.65% recall, 88.64% F1-score), their method degrades significantly on the more challenging HMDB51 dataset (64.98% precision, 61.95% recall, 60.73% F1-score). In contrast, our system maintains robust performance (93.6% precision, 82% recall, 87% F1-score) on the Authentically Distorted Surveillance Videos dataset. This represents a 25.62 percentage point improvement in precision over [17] (67.8% precision for running detection) and comparable accuracy to [6] (89.3% on the MPII dataset), while operating without their stringent constraints (e.g., fixed human positioning and only human side views).

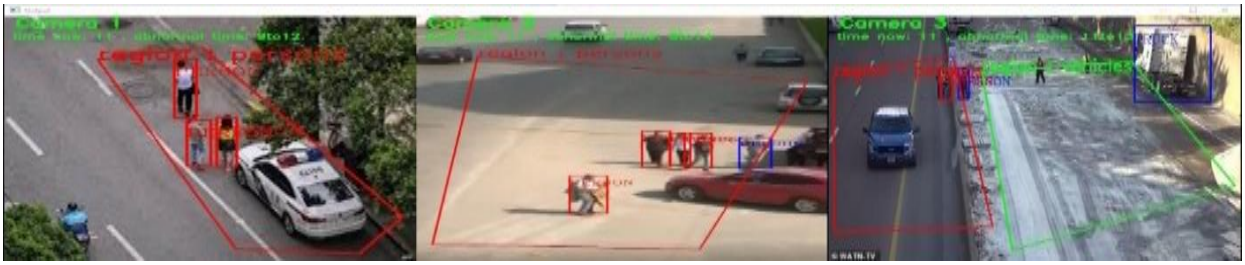
For unauthorized pedestrian intrusion detection, our method achieves 92.42% recall on distorted footage while additionally distinguishing between partial and complete intrusions, a capability absent in Kim's [4] approach (93.8% recall on pristine KISA datasets). Furthermore, our system outperforms (86.4% accuracy) on the UCSD dataset, achieving 87.01% precision, 89.33% recall, and 88.14% F1-score [24]. These results demonstrate our framework's superior reliability in real-world surveillance conditions characterized by occlusion, motion blur, and variable lighting. The direction of movement was identified with 100% precision in the videos where vehicles were detected. Furthermore, for all detected intrusions in the videos, it is identified if they are totally, partially, or not inside the ROI with 100% precision , a critical capability for security applications.

In conclusion, this work presents an intelligent surveillance system with four key innovations: (1) a unified detection framework combining SSD_MobileNet_v3 classification and KCF tracking that simultaneously achieves 95.8% vehicle recall and 92.4% pedestrian recall; (2) a privacy-conscious motion classification method obtaining 82% recall and 93.6% precision without pose estimation; (3) reliable operation under occlusion and low-light conditions; and (4) an advanced intrusion classification system that substantially decreases false alarms. The integrated architecture processes multiple video feeds in real-time while outperforming conventional surveillance approaches. Comprehensive evaluation on enhanced real-world datasets demonstrates the system's robustness across diverse challenging scenarios. Notably, the solution maintains computational efficiency despite its multi-object detection capabilities. These contributions collectively advance both the theoretical foundations and practical implementation of modern surveillance technologies.

Fig. 14 demonstrates a series of image dataset results from an automated surveillance system, highlighting distinct security and safety scenarios with corresponding threat-level responses: (a) tampering with museum antiquities, where the system triggers a red alarm if a person enters the restricted region of interest (ROI), issues a warning for proximity to the ROI, and remains inactive when a safe distance is maintained; (b) pedestrians illegally walking on a highway; (c) vehicles improperly occupying pedestrian walkways; and (d) a child climbing a window or balcony fence, where the system escalates from a warning (near the hazardous zone) to a red alarm upon entry into the unsafe region. These cases exemplify the system’s capability to dynamically assess risks and prioritize threats in real-world environments.

Fig. 15 presents empirical results from our intelligent video surveillance system, demonstrating multi-threat detection capabilities across five critical security scenarios: (a) truck detection, identifying large vehicle presence in restricted zones with 95.83% recall; (b) wrong-direction vehicle detection, where the system accurately flags a vehicle moving against prescribed traffic flow (indicated by a red bounding box and alarm trigger in subfigure b); (c) prowl intrusion detection, capturing slow, deliberate movements; (d) walking intrusion detection in secured perimeters; and (e) run intrusion detection for rapid security breaches.

a.

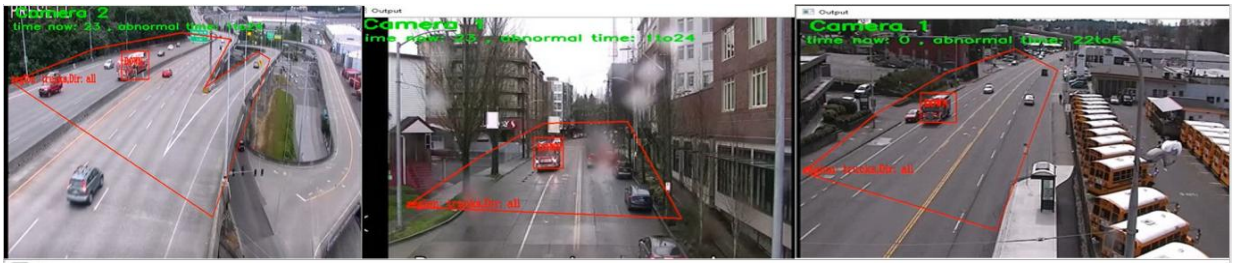


b.

c.

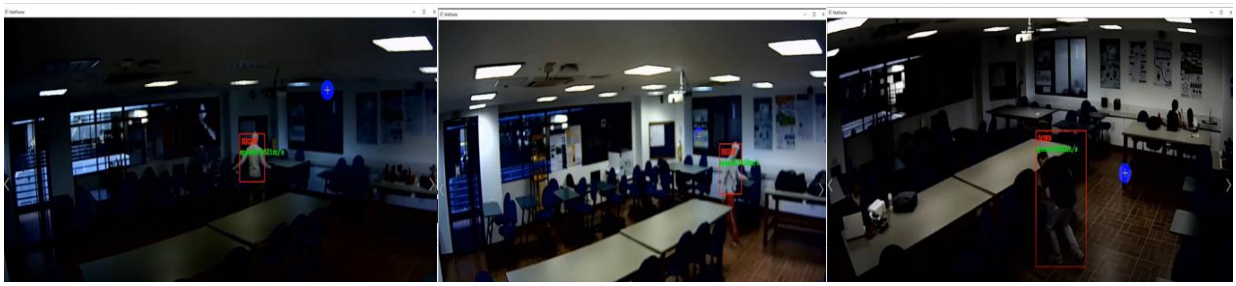
d.

Fig. 14: Image dataset results a. Tamper with antiquities in Museums b. Persons Walk on the highway c. Vehicles on pedestrian walkway d. Baby climbing window or balcony fence.

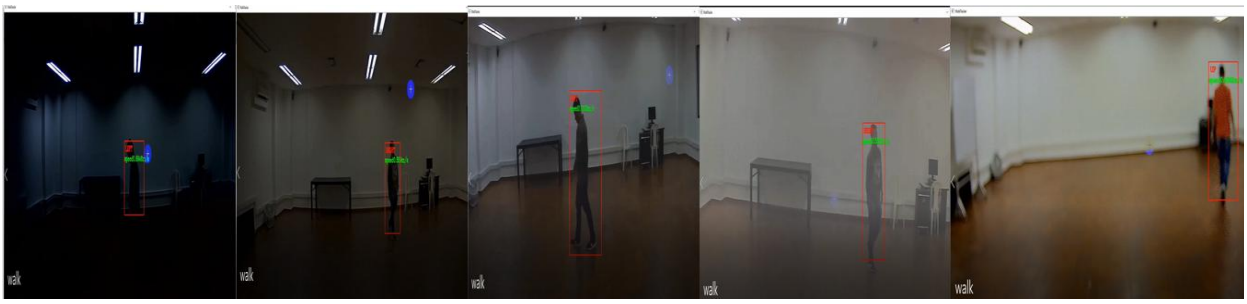


a.

b.



c.



d.

e.

Fig. 15: Videos dataset results a. Truck detection b. Wrong Direction detection c. Prowl intrusion detection d. Walking intrusion detection e. Run intrusion detection

4.4.5 System Pseudocode

This subsection formalizes the algorithmic implementation of the proposed multi-camera surveillance framework. The system pseudocode encapsulates the core computational workflow as shown in **Fig. 16**, comprising real-time object detection via SSD_MobileNet_v3, multi-target tracking, and rule-based anomaly alert generation. The modular architecture integrates a main execution script with six specialized functions Multi_cam_regions_types , SSD_Function , track, get_time , check_ROI, and concatenate.

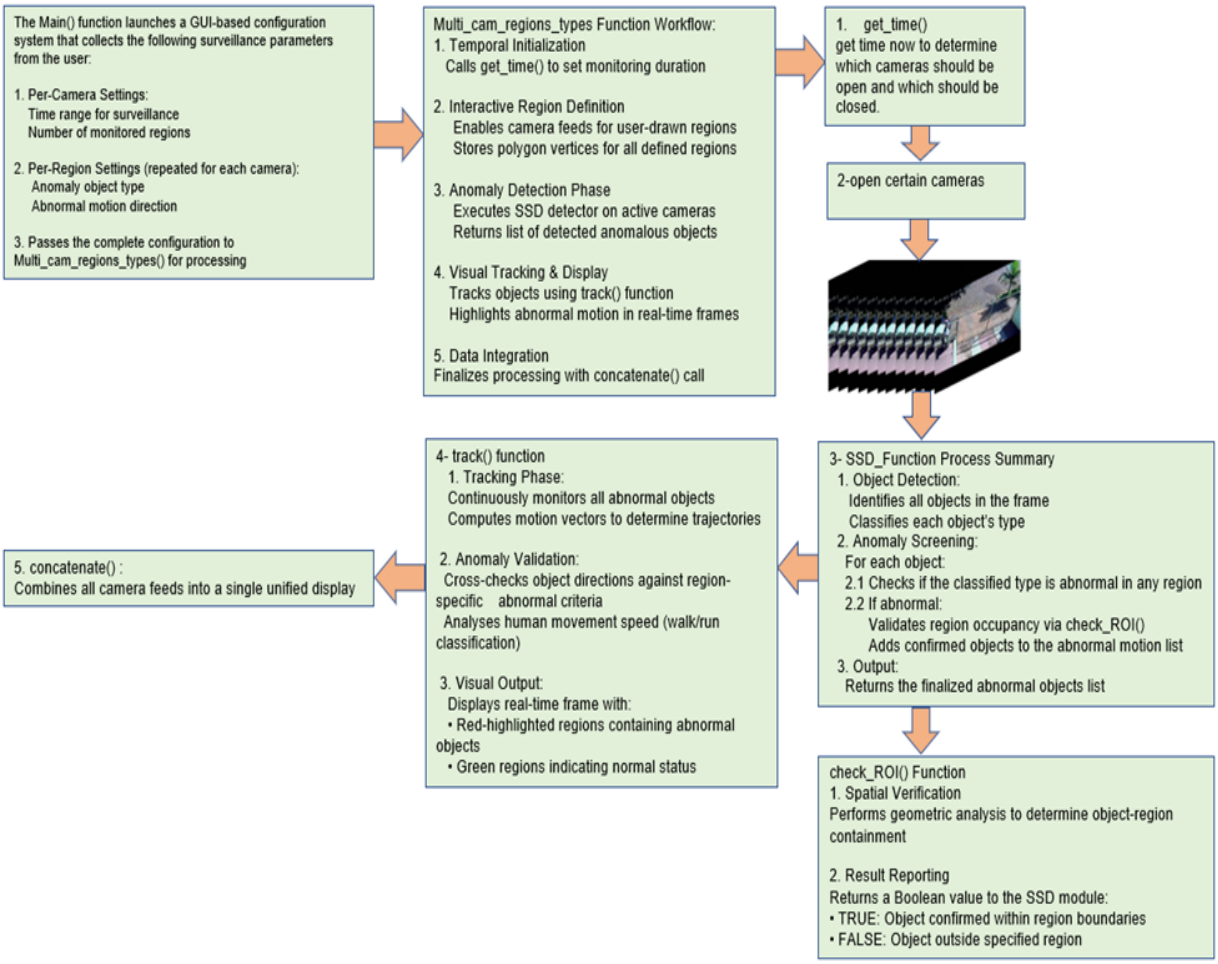


Fig. 16: Pseudocode of the Proposed Intelligent Surveillance System

4.4.6 System setups

The technical specifications and implementation details are summarized in **Table 5** below:

Table 5: Technical Specifications and Implementation.

System Description	Details
Programming Language	Python version 3.7.4
Operating System	Windows 10
Processor	Intel (R) Core (TM) i7-9750H CPU-16 GB
Python libraries	Keras, Tensorflow, OpenCV, Sklearn, Xgboost, Numpy, Random, OS, and PIL

CONCLUSION AND FUTURE WORKS

In this work, a new real-time methodology for a smart video surveillance system is demonstrated. This methodology is designed to detect many abnormal scenarios with multi-surveillance cameras. For each camera, the time range for abnormality detection and the number of interest regions can be determined by the user. Moreover, the user can determine a specific abnormal object type and each region's anomaly motion direction. In the proposed method, the abnormal events depend on the event time, the location, the moving object type, and the motion direction. In addition, the human speed is included. The suggested scheme is composed of two parts: an SSD_MobileNet_V3 detector to detect the object type and a KCF tracker to track the abnormal object's motion. The human motion is classified by establishing a relationship between the real distance and the observed one in the video. The system is evaluated using the Authentically Distorted Surveillance Videos dataset along with the newly collected dataset.

We propose further investigation of human motion classification techniques, including analyzing separate classification for front-view versus side-view rather than using averaged values. This approach could potentially improve the algorithm's accuracy, and we have identified it as an important direction for future research. This work may be developed to support more abnormal scenarios like a falling person, fire detection, and fighting.

CONFLICT OF INTEREST

The authors have no financial interest to declare in relation to the content of this article.

REFERENCES

[1] Z. Zhang, Y. Song, C. Wang, and J. Yu, "Video surveillance taxonomy," ACM Computing Surveys, vol. 54, no. 9, pp. 1–36, 2021, doi:10.1145/3453161.

[2] K. Ullah, M. W. Ding, V. Palade, I. U. Haq, and S. W. Baik, "Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications," Applied Soft Computing, vol. 103, p. 107102, 2021. doi: 10.1016/j.asoc.2021.107102.

[3] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172–186, 2019. doi: 10.1109/TPAMI.2019.2929257.

[4] D. Kim, H. Kim, Y. Mok, and J. Paik, "Real-Time Surveillance System for Analyzing Abnormal Behavior of Pedestrians," Applied Sciences, vol. 11, no. 13, p. 6153, 2021. doi: 10.3390/app11136153.

[5] M. George, B. R. Jose, and J. Mathew, "Performance Evaluation of KCF based Trackers using VOT Dataset," Procedia Computer Science, vol. 125, pp. 560-567, 2018. doi: 10.1016/j.procs.2017.12.072.

[6] Z. Zhao, S. Lan, and S. Zhang, "Human Pose Estimation based Speed Detection System for Running on Treadmill," in Int. Conf. Culture-Oriented Science & Technology (ICCST), 2020, doi: 10.1109/iccst50977.2020.00108

[7] H. Li, S. Yang, J. Liu, Y. Yang, M. Kadoch, and T. Liu, "A Framework and Method for Surface Floating Object Detection Based on 6G Networks," Electronics, vol. 11, no. 18, p. 2939, 2022. doi: 10.3390/electronics11182939.

[8] C. Amrutha, C. Jyotsna, and J. Amudha, "Deep Learning Approach for Suspicious Activity Detection from Surveillance Video," in Int. Conf. Innovative Mechanisms for Industry Applications (ICIMIA), 2020. doi: 10.1109/icimia48430.2020.9074920.

[9] C. Sun, Y. Jia, H. Song, and Y. Wu, "Adversarial 3D Convolutional Auto-Encoder for Abnormal Event Detection in Videos," IEEE Transactions on Multimedia, vol. 23, pp. 3292–3305, 2021. doi: 10.1109/TMM.2020.3023303.

[10] S. Bouindour, R. Hu, and H. Snoussi, "Enhanced Convolutional Neural Network for Abnormal Event Detection in Video Streams," in IEEE Int. Conf. Artificial Intelligence and Knowledge Engineering (AIKE), 2019, pp. [page range]. doi: 10.1109/aike.2019.00039.

[11] J. S. Kim, M. G. Kim, and S. B. Pan, "A study on implementation of real-time intelligent video surveillance system based on embedded module," EURASIP Journal on Image and Video Processing, vol. 2021, no. 1, 2021. doi: 10.1186/s13640-021-00576-0.

[12] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in IEEE Int. Conf. Multimedia and Expo (ICME), 2017. doi: 10.1109/icme.2017.8019325.

[13] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning Temporal Regularity in Video Sequences," in IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016. doi: 10.1109/cvpr.2016.86.

[14] J. Yu, Y. Lee, K. C. Yow, M. Jeon, and W. Pedrycz, "Abnormal Event Detection and Localization via Adversarial Event Prediction," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 8, pp. 3572-3586, 2022. doi: 10.1109/TNNLS.2021.3053563.

[15] C. Liu, J. Ying, F. Han, and M. Ruan, "Abnormal Human Activity Recognition using Bayes Classifier and Convolutional Neural Network," in IEEE Int. Conf. Signal and Image Processing (ICSIP), 2018. doi: 10.1109/siprocess.2018.8600483.

[16] M. H. Kolekar, N. Bharti, and P. N. Patil, "Detection of fence climbing using activity recognition by Support Vector Machine classifier," in IEEE Region 10 Conf. (TENCON), 2016. doi: 10.1109/tencon.2016.7848029.

[17] S. Lao, D. Wang, F. Li, and H. Zhang, "Human running detection: Benchmark and baseline," Computer Vision and Image Understanding, vol. 153, pp. 143-150, 2016. doi: 10.1016/j.cviu.2016.03.005.

[18] V. D. R. Seethi and P. Bharti, "CNN-based Speed Detection Algorithm for Walking and Running using Wrist-worn Wearable Sensors," in IEEE Int. Conf. Smart Computing (SMARTCOMP), 2020. doi: 10.1109/smartcomp50058.2020.00064.

[19] N. D. Mane, "Real-Time Anomaly Detection in Video Surveillance: A Mathematical Modeling and Nonlinear Analysis Perspective with MobileNet and Bi-LSTM," Deleted Journal, vol. 31, no. 2s, pp. 306-319, 2024. doi: 10.52783/cana.v31.651.

[20] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in IEEE Int. Conf. Image Processing (ICIP), 2017. doi: 10.1109/icip.2017.8296547.

[21] T. Ganokratanaa, S. Aramvith, and N. Sebe, "Unsupervised Anomaly Detection and Localization Based on Deep Spatiotemporal Translation Network," IEEE Access, vol. 8, pp. 50312-50329, 2020. doi: 10.1109/ACCESS.2020.2979869.

[22] N. S. Ghedia and C. H. Vithalani, "Outdoor object detection for surveillance based on modified GMM and Adaptive Thresholding," International Journal of Information Technology, vol. 13, no. 1, pp. 185–193, 2020. doi: 10.1007/s41870-020-00522-9.

[23] K. K. Verma, B. M. Singh, and A. Dixit, "A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system," International Journal of Information Technology, vol. 14, no. 1, pp. 397-410, 2019. doi: 10.1007/s41870-019-00364-0.

[24] S. Mishra and S. Jabin, "Anomaly detection in surveillance videos using deep autoencoder," International Journal of Information Technology, vol. 16, no. 2, pp. 1111-1122, 2023. doi: 10.1007/s41870-023-01659-z.

[25] Y. I. M. Reyna, C. J. G. del Ángel, and J. H. D. Acuña, "MEDAVET: Traffic vehicle anomaly detection mechanism based on spatial and temporal structures in vehicle traffic," 2024. [Online]. Available: https://www.researchgate.net/publication/380096029.

[26] M. M. Saif, H. S. Tantawy, and A. El-Marakeby, "INTELLIGENT TRAFFIC SIGNAL CONTROL USING SPATIO-TEMPORAL DATA AND REINFORCEMENT LEARNING," Journal of Al-Azhar University Engineering Sector, vol. 20, no. 75, pp. 511-526, 2025. doi: 10.21608/auej.2025.329865.1723.

[27] T. M. Soliman, A. Elshenawy, and H. Tantawy, "ADAPTIVE CRUISE CONTROL WITH LANE KEEPING ASSIST USING REINFORCEMENT LEARNING," Journal of Al-Azhar University Engineering Sector, vol. 19, no. 73, pp. 1349-1368, 2024. doi: 10.21608/auej.2024.298711.1676.

[28] M. Mustafa, R. Abobeah, and M. Elkholy, "A COMPREHENSIVE APPROACH TO AUTONOMOUS VEHICLE NAVIGATION," Journal of Al-Azhar University Engineering Sector, vol. 0, no. 0, pp. 167-182, 2024. doi: 10.21608/auej.2024.275435.1634.

[29] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," in IEEE Conf. Computer Vision and Pattern Recognition, 2018, pp. 89-95.

[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," Computer Vision – ECCV 2016, pp. 21-37, 2016. doi: 10.1007/978-3-319-46448-0_2.

[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 2017. doi: 10.1109/TPAMI.2016.2577031.

[32] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, and M. Tan, "Searching for MobileNetV3," arXiv:1905.02244v5 [cs.CV], 2019.

[33] J. Panero and M. Zelnik, Human Dimension and Interior Space [Online]. New York: Watson-Guptill, 1979. Available:http://books.google.ie/books?id=pQ1QAAAAMAAJ&q=)Human+Dimension+and+Interior+Space&dq=)Human+Dimension+and+Interior+Space&hl=&cd=2&source=gbs_api.

[34] R. C. Browning, E. A. Baker, J. A. Herron, and R. Kram, "Effects of obesity and sex on the energetic cost and preferred speed of walking," Journal of Applied Physiology, vol. 100, no. 2, pp. 390-398, 2006.
[35] C. A. Franco, Authentically Distorted Surveillance Videos Dataset [Online]. IEEE DataPort, 2022. Available: https://ieee-dataport.org/open-access/authentically-distorted-surveillance-videos-dataset
[36] M. A. A. Al-Qaness, A. A. Abbasi, H. Fan, R. A. Ibrahim, S. H. Alsamhi, and A. Hawbani, "An improved YOLO-based road traffic monitoring system," Computing, vol. 103, no. 2, pp. 211-230, 2021. doi: 10.1007/s00607-020-00869-8.
[37] P. R and M. P, "An implementation of intelligent YOLOv3-based anomaly detection model from crowded video scenarios with optimized ensemble pattern extraction," The Imaging Science Journal, pp. 1-22, 2023. doi: 10.1080/13682199.2023.2255335.