Early Anomalus Action Detection in Surveillance Video Using MRCNN-LSTM Classification

D. Manju
Department of CSE-(CyS, DS) and AI&DS, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India
nuthana525@gmail.com (corresponding author)

Kishore K. Kumar
Department of ECE, ICFAI University, Raipur Chattisgarh, India
kishorekamarajugadda@gmail.com

Movva Pavani
Department of ECE, Nalla Malla Reddy Engineering College, Divyanagar, Hyderabad, India
drmovvapavani@gmail.com

Rajesh Kumar Verma
Department of CSE-(CyS,DS) and AI&DS, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India
rajeshverma.hyd10@gmail.com

Anand Kumar Saraswathi Rathod
Department of CSE-(CyS,DS) and AI&DS, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India
sr_ak@yahoo.com

Pavan N. V. S. Kumar
Department of Computer Science and Engineering, KLEF Education Foundation, India
nvspavankumar@kluniversity.in

V. S. N. Murthy
Department of Information Technology, Shri Vishnu Engineering College for Women, Bhimavaram, India
vsn.murthy87@gmail.com

Bh. Krishna Mohan
Department of Information Technology, RVR & JC College of Engineering, Guntur, India
bkm@rvrjc.ac.in

Received: 20 February 2025 | Revised: 18 April 2025, 9 May 2025, and 2 June 2025 | Accepted: 6 June 2025

 DOI: https://doi.org/10.48084/etasr.10656

ABSTRACT

Public space monitoring systems are critical for observing typical human behavior and detecting abnormal activities, especially in high-security environments. With the rise in public space thefts, there is a growing need for intelligent systems capable of detecting suspicious movements early enough to prevent criminal

acts. Although Convolutional Neural Networks (CNNs) are widely used in image classification, they are inadequate to differentiate between abnormal and normal behavior and identify criminal activity in its early stage. To overcome these limitations, this study proposes a new hybrid model that combines Mask R-CNN (MRCNN) with Long Short-Term Memory (LSTM) networks for accurate object detection, tracking, and sequential behavior analysis. The main contribution of this study is a multistage anomaly detection pipeline that involves frame conversion, contrast enhancement, background removal, object tracking, and feature extraction. The MRCNN-LSTM framework can extract both spatial and temporal characteristics to allow precise early-stage anomaly detection. Thorough testing on three benchmarking datasets, UCF Crime, Snatch1.0, and CUHK, exhibited excellent performance, with a 93.6% accuracy for the UCF Crime dataset. Performance metrics such as observation ratio and time duration were used to assess the responsiveness and effectiveness of the system in real-time surveillance scenarios. This research advances the field of intelligent surveillance by enabling proactive threat mitigation through the early and precise detection of anomalous behavior.

Keywords-anomaly detection; classification; LSTM; MRCNN; preprocessing; object detection

I. INTRODUCTION

Video surveillance systems have become indispensable tools for maintaining public safety, particularly in environments susceptible to crime. The increasing frequency of such incidents underscores the urgent need for automated real-time surveillance solutions capable of detecting abnormal human behavior without relying on manual monitoring. Although traditional surveillance systems simply record video, they lack the capacity for proactive threat detection and timely response before harm occurs. Recent advances in deep learning have enabled machines to autonomously detect anomalies in video streams. Techniques such as Convolutional Neural Networks (CNNs), Region-based CNNs, and Recurrent Neural Networks (RNNs) have shown promise in feature extraction and modeling temporal behaviors. However, these models face several limitations, such as high training times and computational overhead for large datasets, reduced performance in low-light or low-resolution environments, challenges in detecting rare or subtle anomalies due to class imbalance, and inadequate real-time alerting mechanisms for preemptive action. Moreover, many existing approaches focus primarily on either spatial object detection or temporal behavior modeling, but rarely integrate both effectively. Although methods such as Fast-AnoGAN, ADIHT, and DS-UTSSC explore anomaly detection from varied perspectives, and models such as YOLOv5 and Mask R-CNN serve object detection and segmentation tasks, few have been integrated into a unified system for early-stage anomaly detection in real-world surveillance videos.

To address these gaps, this study proposes a novel hybrid deep learning framework that combines Mask R-CNN for multiobject spatial detection with LSTM for temporal sequence analysis. This integrated pipeline aims to identify anomalies early, allowing for timely intervention. The key stages of the system include: frame conversion and contrast enhancement using FLAHE, background removal using GrabCut, real-time object tracking via YOLOv5, and sequential feature classification using an MRCNN-LSTM model. This architecture is capable of distinguishing between normal and abnormal behaviors early in activity sequences, facilitating prompt preventive action. The system was evaluated on three benchmark datasets, UCF-Crime, CUHK Avenue, and Snatch1.0, achieving accuracy scores of 93.6%, 94.9%, and 92.2%, respectively. The main contributions of this study are:

- A unified deep learning model (MRCNN+LSTM) for integrated spatio-temporal anomaly detection.
- A preprocessing pipeline optimized for enhancing performance in low-quality surveillance footage
- Empirical validation across multiple datasets, demonstrating notable improvement over existing state-of-the-art methods.
- Real-time applicability and alert mechanisms make the system suitable for deployment in high-security environments.

In [1], a comprehensive overview of deep and shallow anomaly detection techniques examined model explainability. In [2], a skip-connected dilated RNN was proposed to identify temporal patterns for video-based snatch theft detection. In [3], patch-based processing limitations were addressed using a hierarchical temporal one-class network to enhance time series anomaly detection. In [4], anomaly detection was improved by leveraging expert feedback within the detection pipeline. Logsy [5] is a self-attentive classification-based system to generalize anomaly detection in unstructured logs. HDAD [6] is a hyperdimensional computing-based method for detecting anomalies in vehicle sensor data. In [7], model performance was measured with softmax metrics on geometrically transformed images for deep anomaly detection. In [8], deep reinforcement learning was used to optimize anomaly detection on residual maps. ADIHT [9] uses an Inverted Hash Table (IHTable) for the detection of privacy-enforced time series anomalies. The LRCRD [10] uses a low-rank collaborative representation to detect hyperspectral anomalies. In [11], scalable bitmap and Bayesian approaches were investigated for anomaly detection in time series data for industrial systems. Fast-AnoGAN [12] is a semi-supervised approach that computes anomaly scores using image difference analysis and two-stage background prediction. In [13], pure background pixels were employed to predict background covariance and enhance hyperspectral detection. DS-UTSSC [14] is an unsupervised hyperspectral target detection algorithm that uses data sphering. In [15], a self-supervised anomaly detection framework was proposed, assuming that normal samples are more reconstructible than anomalies. In [16], Gaussian Process Regression (GPR) was used for video-based anomaly classification in surveillance systems. In [17], a deep model blended spatial-temporal RNNs and region-based CNNs for

efficient pedestrian detection. In [18], it was observed that normal frame reconstruction errors may emphasize anomalies in surveillance videos.

In [19], a Siamese CNN was employed to compare motion and appearance differences for efficient anomaly detection. In [20], a spatio-temporal autoencoder was proposed, employing reconstruction errors for anomaly detection. In [21, 22], special loss functions were designed to encode normal data and detect anomalies as outliers. In [23], low-contrast satellite images were improved with a slantlet transform and gamma correction. In [24], contrast enhancement was optimized with a new white balancing parameter. In [25, 26], different contrast enhancement methods, including DWT-SVD, were compared on different images. In [27], a multiscale dark-pass filter was introduced to suppress pixel-edge artifacts for improved image contrast. In [28], a YOLOv5-based model was used to improve real-time small object detection in UAV imagery. In [29, 30], YOLO-Z variants were developed with spatial optimization for different scales, illustrating the versatility of YOLOv5 in different tracking applications. In [31], a Gaussian Mixture Fully Convolutional Variational Autoencoder (GMFC-VAE) model was proposed for video anomaly detection and localization. This model can encode spatial and temporal relationships with fully convolutional layers and provide fine-grained anomaly localization without using optical flow or hand-crafted features. This study used various public datasets. The UCF-Crime dataset [32, 33] is a large surveillance dataset that captures a broad set of real-world anomalous activities, including theft, assault, and accidents. Another significant dataset comes from CUHK [34], which includes realistic abnormal behavior scenarios, further facilitating strong model testing and benchmarking. In [35], an attribute-based method was introduced for anomaly detection, which does not resort to reconstruction losses or adversarial training. The model employed high-level semantic attributes to distinguish normal from abnormal events and was competitive with other approaches while being more interpretable. In [36], a Multi-Branch GAN-based framework integrated context learning for surveillance videos. Having multiple branches enables the model to learn varied contextual information, enhancing its performance to detect faint anomalies that single-branch models are prone to overlook. In [37], early action prediction was investigated by adopting a hybrid deep learning architecture that incorporates 3D CNNs with Bidirectional LSTM (Bi-LSTM). This approach anticipated future actions at the beginning of the video sequence, which is very helpful for pre-emptive anomaly detection. In [38], a deep context-sensitive feature extraction technique was developed for anomaly detection. Using context-aware features, this approach yielded a better detection accuracy in surveillance datasets. This work underlined the need for contextual knowledge to separate normal and abnormal behavior. In [39], the focus was on panic detection for crowded scenes. This method was tailored for scenarios involving mass gatherings, where early detection of panic behavior can prevent disasters. The model integrated crowd-motion features with behavior recognition, demonstrating effectiveness in complex, high-density settings.

This study aimed to develop an anomaly detection framework specifically designed for real-time video surveillance systems. Although conventional benchmark datasets such as UCF Crime and CUHK Avenue are used extensively in the literature, they tend to deal mainly with a general set of anomalies in fairly controlled or structured scenes. Such data are very important for measuring broad performance, but commonly do not sample localized, short-duration, risky criminal acts such as snatch burglaries that are now on the rise in open and semi-public places. To overcome this limitation, the Snatch1.0 dataset includes a set of real and simulated surveillance videos, recording incidents such as chain snatching and mobile theft. The primary objective of using this dataset was to test the performance of the proposed anomaly detection system on video surveillance data that closely resembles the target real-world applications. Snatch1.0 offers scenes with varied lighting, depth complexity, and crowd density conditions, commonly faced in real surveillance videos but not widely covered in available public datasets. Testing the proposed method on Snatch1.0, as well as on UCF and CUHK, aimed to show how the model is generalizable and usable in practice for detecting common and niche anomalies in a variety of surveillance scenarios. This added support enhances the overall objective of creating a system that is accurate in research datasets but also deployable in real-world public safety and crime reduction applications.

II. PROPOSED METHOD

Video surveillance systems are intended to identify anomalies, including chain snatching and other unforeseen events, which are important for public safety and security. This study offers an integrated method that combines contrast enhancement technology with object tracking and deep learning-based classification for better anomaly detection in actual video frames. Such an approach can improve detection accuracy along with response capabilities by examining and classifying abnormal activities in real-time. Video footage enters the process by converting into frames, and then the system applies Fuzzy Logic-based Adaptive Histogram Equalization (FLAHE) preprocessing to eliminate noise while improving picture quality. The video surveillance system performs background extraction using GrabCut and YOLOv5, which operate as object detectors. The detection system generates information for low- and high-level features of the identified objects. The LRCNN-LSTM classification unit accepts feature inputs, which allows precise identification of anomalous as well as non-anomalous actions to minimize performance errors. The system detects suspicious activity early to provide rapid alert and response functions and improve public security. Performance tests utilizing precision, recall, accuracy, observation ratio, and time taken were used to measure detection accuracy and operational efficiency, as well as the speed of the system. The system is developed in Python with image processing functionalities and deep learning libraries.

This method ensures a strong and efficient answer for detecting anomalies in video surveillance systems that apply to real-world security situations. Frames are extracted from the video recording during the first step. FLAHE improves frame quality by reducing noise. Grabcut is an operation that removes background components from frames that were initially

processed. The software performs background removal with YOLOv5, which identifies particular objects in the frames. Two kinds of features are obtained from the identified objects: low-level and high-level features. The MRCNN-LSTM classifier uses extracted features for precise anomalous and non-anomalous human action prediction, without generating misclassification or misprediction errors. Figure 1 illustrates the visual representation of the proposed framework. MRCNN-LSTM classification, combined with contrast enhancement and object tracking and background removal, enables the proposed system to identify anomalous human behavior in surveillance videos at an early stage.

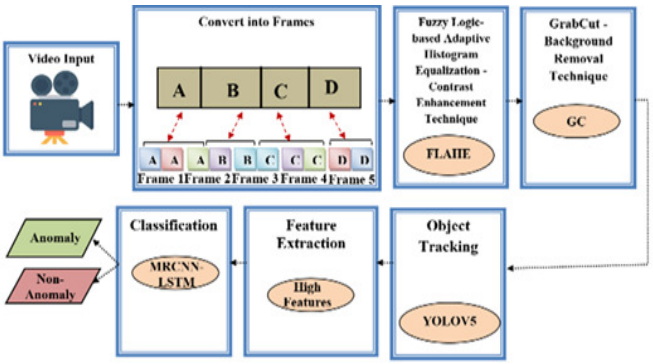

Fig. 1. The framework of the proposed method.

The system provides important services for public security through quick alerts and swift responses to suspicious activities, and its evaluation assesses its performance with respect to its efficiency and speed.

A. *Fuzzy Logic-based Adaptive Histogram Equalization (FLAHE)*

The Cumulative Distribution Function (CDF) is given by:

$$CDF(i) = \sum p(j), \text{ for } j = 0 \text{ to } I \quad (1)$$

where $CDF(i)$ calculates the CDF after obtaining $p(j)$, which stands for histogram probability values at intensity level j. $CDF(i)$ refers to the CDF at intensity level i.

$$F(\mathrm{i}) = \sum p(j) * w(j), \text{ for } j = 0 \text{ to } I \quad (2)$$

B. *Grabcut*

The fuzzy membership function determines weight values that guide the gain function to perform enhancement adjustments. Such adaptive processing improves image contrast without damaging the distinct image features.

$$E(u) = \sum_i (D(i,u_i) + \lambda * V(i,u_i)) \quad (3)$$

where $E(u)$ represents the energy of the segmentation, i iterates over the pixels in the image, u_i represents the label (foreground or background) assigned to pixel i, $d(i,u_i)$ is the data term measuring color similarity, $v(i,u_i)$ is the smoothness term measuring spatial coherence, and λ is a parameter controlling the trade-off between data and smoothness terms.

C. *YoloV5*

The formula to express predicted bounding boxes is:

$$A_x = \sigma(M_x) + C_x \quad (4)$$

$$A_y = \sigma(M_y) + C_y \quad (5)$$

$$A_{wd} = S_{wd}.e^{T_{wd}} \quad (6)$$

$$A_{ht} = S_{ht}.e^{T_{ht}} \quad (7)$$

This equation contains anchor box dimensions (S_{wd}, S_{ht}) and the coordinate value of the image top-left corner along with predicted bounding box dimensions $(A_x, A_y, S_{wd}, S_{ht})$. Additional information about anchor box dimension sizes is available in Figure 2, along with details about bounding box prediction locations.

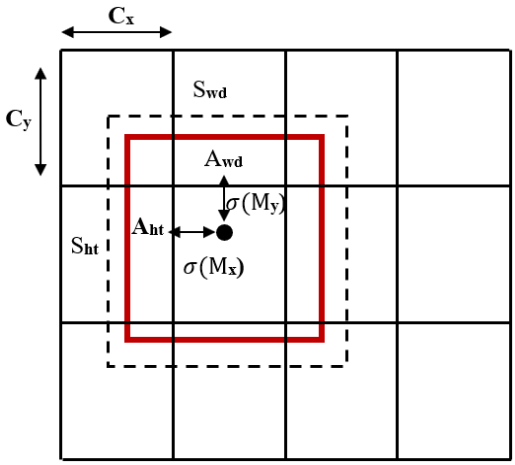

Fig. 2. YOLOv5 bounding box prediction.

D. *Mask R-CNN (MRCNN)*

MRCNN was chosen for its superior spatial feature extraction and instance segmentation capabilities, which are essential for accurate anomaly detection in complex video scenes. While it is computationally heavier, its integration with LSTM ensures a robust and reliable hybrid model for detecting both spatial and temporal anomalies. Therefore, the MRCNN model is a popular instance segmentation framework that extends the Faster R-CNN object detection architecture by incorporating a pixel-level segmentation branch.

Once the objects get tracked, the most important and required features are extracted. This method extracts two kinds of features: Low-level and high-level features.

$$B_i^{Low} = \{B_{HO}, B_{HG}\} \quad (8)$$

Low-level features resemble features such as Histogram of optical flow B_{HO} and Histogram of gradient B_{HG}. High-level features resemble features such as Shape B_S, Texture B_T, Surf B_{Sf}, SIFT B_{SI}, Boundary box B_{Bb}, and human skeleton B_{Hs}.

$$B_i^{High} = \{B_S, B_T, B_{Sf}, B_{SI}, B_{Bb}, B_{Hs}\} \quad (9)$$

The loss functions measure the difference between the predicted values and the ground truth annotations at various levels.

E. Proposed Hybrid Algorithm

Algorithm 1: MRCNN-LSTM
Input: Video input, Threshold 35%,
Batch size = 1
Output: Predicted frame

1: For each input video:
Load video frames sequentially.
Store them in a list or array for processing.
2: Apply Preprocessing
For each frame in the video:
Convert the frame (resize, normalize, grayscale)
Apply FLAHE to enhance contrast.
Use the GrabCut algorithm to segment the foreground (rough ROI isolation).
Use YOLOv5 to detect and localize objects in the frame.
3: Feature Extraction using MRCNN
For each detected object in the frame:
Extract the bounding box from YOLOv5.
Crop the region of interest (ROI) from the original image.
Pass the cropped ROI to the Mask R-CNN model.
Extract the feature vector from MRCNN output
4: Dataset Splitting and Custom Feature Representation
(i) A custom feature representation R is created, $R \in R^{k \times d}$, where k is the number of detected objects (e.g., people, bags) and d is the number of frames per video
(ii) For each frame i in the video (total N frames):
Compute the feature vector $x_i^k = R(x_i)$
(iii) Then, binary classification is performed
The system determines whether a frame is normal or abnormal by comparing the current feature vector x with a reference feature x^*.
if $d(x^*, x) <= Threshold$ then
Normal
else
Abnormal
(iv) Create Dataset:
Combine features and labels.
Use train_test_split:
- 70% data for training
- 30% data for testing
5: Adding LSTM Layer
(i) Feed the sequence of feature vectors $[x_1, x_2, \ldots, x_n]$ to LSTM.
LSTM captures how object states change over time.
This helps to identify temporal anomalies (i.e., patterns that deviate over time).
(ii) Train the LSTM model with labelled sequences:
Input: sequences of feature vectors (temporal sequences).
Output: binary classification (normal or abnormal).
(iii) At inference:
Predict the status (normal/abnormal) of each frame or sequence. LSTM analyses how features change frame-to-frame (i.e., over time).
Calculate metrics.

The proposed technique enables real-time video anomaly identification by combining LSTM networks for temporal analysis and Mask R-CNN for spatial feature extraction. The output is a classification of frames as either normal or abnormal. First, video input is processed, and frames are retrieved one after the other. Frame quality is improved by pre-processing methods, such as YOLOv5 for initial object detection, GrabCut for foreground extraction, and adaptive histogram equalization. Mask R-CNN then uses labeled normal and abnormal frames for training before segmenting objects and extracting features. A custom classification function uses a threshold to identify abnormalities in the split dataset. Finally, LSTM refines classification by analyzing sequential dependencies and flags anomalies if their estimated probability is more than 35%. Frame weight vectors are fine-tuned based on errors generated from positive samples.

This study initially selected hyperparameters using a manual tuning approach, based on previous literature and empirical observations. To refine these choices and ensure optimal performance, a random search strategy was employed over a predefined range of values for key hyperparameters: Learning rate (range:0.001), Optimizer: Adam (β_1 = 0.9, β_2 = 0.999, ε = 1e-8), Loss function: cross-entropy loss (for classification), Batch size (16), Number of LSTM units (64), Dropout rate (0.3 for LSTM layers, L2 regularization), Number of training epochs (100). The selection was guided by validation performance (accuracy, precision, recall) on a held-out validation set drawn from each dataset. Random search was preferred over grid search due to its efficiency in exploring larger spaces with fewer iterations. The final configuration used in the experiments yielded the best balance between detection accuracy and inference time [38, 39].

III. EXPERIMENTAL RESULTS AND DISCUSSION

The experiments were carried out on a PC equipped with an Intel Core i7 latest generation processor, 16 GB RAM, and a dedicated NVIDIA Tesla P100. To evaluate the real-time applicability of the model, the average inference time per frame

was measured during the testing, including frame processing time at approximately 38-45 ms per frame on average (22-26 fps), which is suitable for near-real-time surveillance scenarios. The system was optimized using efficient frame conversion techniques (e.g., FLAHE, GrabCut) and light-weight tracking (YOLOv5), together allowing low-latency detection despite the integration of deep models such as Mask R-CNN and LSTM.

The performance of the proposed anomaly detection framework was evaluated using three datasets:

- Snatch1.0 contains low-resolution videos from Hyderabad, India [2], focusing on rare chain-snatching incidents. Out of 37,485 normal events, only 35 snatch thefts (each lasting 4-5 seconds) were identified over 4.5 hours. The dataset was divided into 816 10-second clips for analysis.
- UCF Crime consists of 1,900 real-world surveillance videos (128 hours) with 13 natural anomalies such as fighting, robberies, and accidents. It supports both general anomaly detection and specific anomaly classification, with a 70-30 train-test split [33].
- The CUHK [34] dataset has 16 training videos and 21 testing videos. The dataset has some problems, such as a slight camera shaking in testing video 2 (frames 1051-1100), and the training data includes a few outliers.

To ensure the protection of individual privacy, all datasets used in this project are publicly available with existing consent provisions. No Personally Identifiable Information (PII) was extracted, stored, or utilized from any video content included in this study. All datasets were handled with strict adherence to privacy-preserving protocols. Each dataset was used in full compliance with its respective licensing terms, which explicitly include ethical use provisions and protections for human subjects.

Performance was evaluated using metrics such as precision (correctly detected anomalies), recall (detected anomalies out of all actual anomalies), accuracy (overall detection correctness), observation ratio (normal vs. anomalous event distribution), and time duration (processing speed and efficiency) [33, 37].

A. Result Images

The proposed method successfully detected abnormal behaviors in videos. Through this capability, emergency responders can take early action against unusual activities, resulting in improved protection of public safety while reducing security risks and stopping potential incidents from happening.

B. Result Tables

Table I demonstrates the performance analysis of the proposed MRCNN-LSTM, which achieved 94.9% accuracy on the CUHK dataset, 89.2% on Snatch1.0, and 93.6% on UCF Crime. MRCNN-LSTM achieved higher accuracy on CUHK due to the relatively lower number of actions in a sequence, even compared with results from previous works [36, 37].

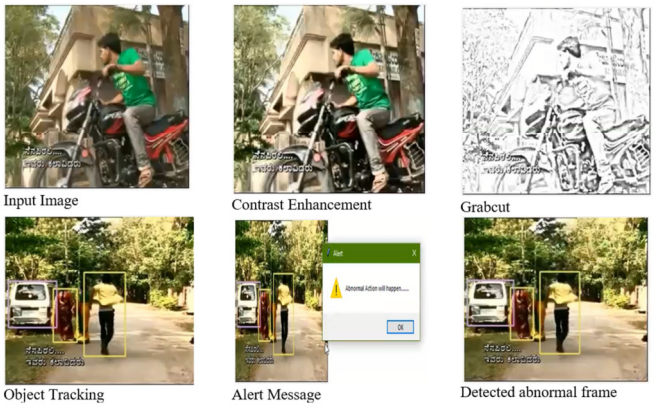

Fig. 3. Overall analysis and identification of anomalous activities in the video footage of the Snatch1.0 dataset.

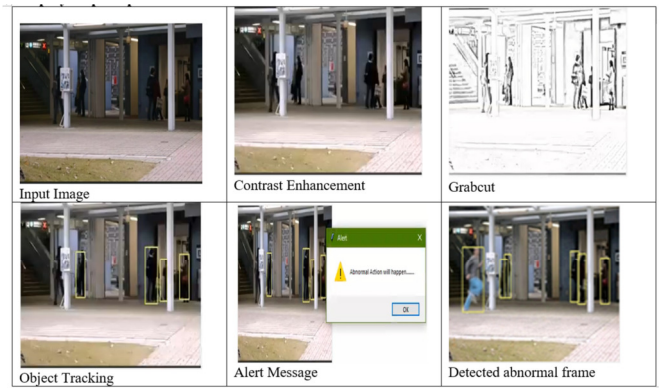

Fig. 4. Anomalous event identification in the CUHK dataset.

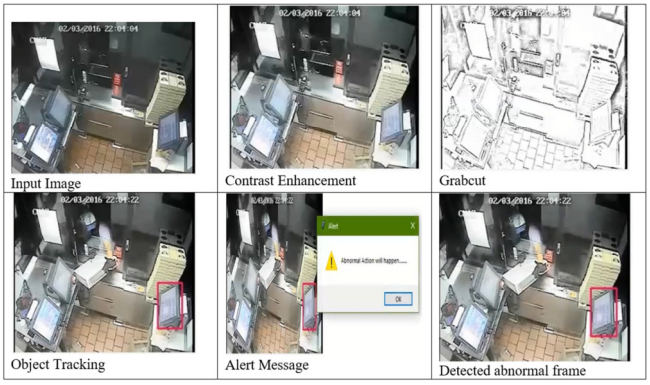

Fig. 5. Identification of anomalous events in the UCF Crime dataset.

TABLE I. PERFORMANCE ANALYSIS OF MRCNN+LSTM

Dataset	Accuracy	Precision	Recall
CUHK	94.9%	90.2	92.1
Snatch 1.0	92.2%	91.2	92.4
UCF Crime	93.6%	90.3	91.9

Figure 6 shows a normalized confusion matrix for the UCF Crime dataset, which contains 14 classes (13 crime categories plus a Normal class). The diagonal line indicates correct predictions, with the color intensity reflecting the model's accuracy for each class. Off-diagonal elements, especially the

lighter blue and red dots, show misclassifications. A red dot in the lower right indicates a notable misclassification, suggesting the model may have confused severe crimes such as Assault or Robbery with a different or less severe category. The matrix is normalized, meaning that values are shown as proportions rather than raw counts, which is important for interpreting performance in imbalanced datasets such as UCF Crime. The matrix exhibits strong diagonal dominance, showing that most predictions correctly match the ground truth labels. However, despite this overall strong performance, the presence of light off-diagonal values indicates some degree of confusion between visually or temporally similar activities (e.g., running vs. escaping), which may be attributed to overlapping visual features between activities, limited training samples for rare anomalies, and class imbalance.

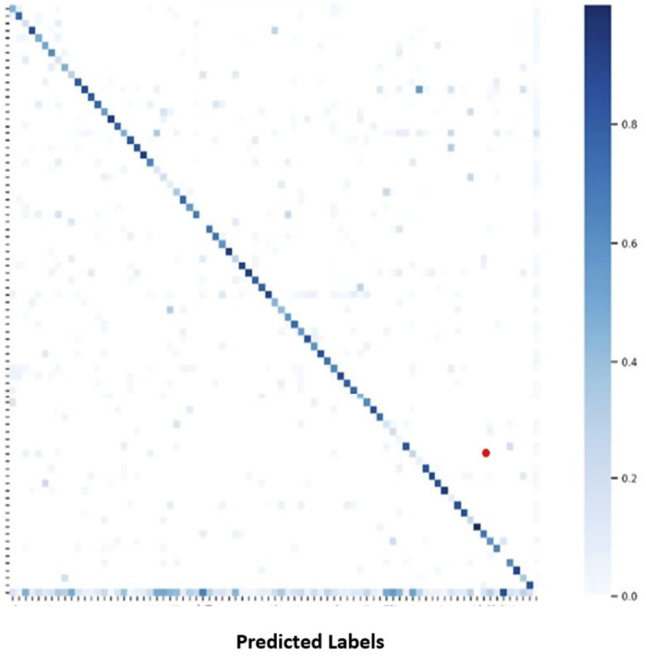

Fig. 6. Confusion Matrix Normalized for UCF Crime dataset.

Table II summarizes the performance of the MRCNN-LSTM model on a subset of classes from UCF-Crime. The proposed MRCNN-LSTM framework shows consistently high scores across both normal and abnormal classes. For high-priority security events, such as fighting and stealing, the model achieves F1-scores of 0.89 and 0.87, respectively, reflecting a balanced trade-off between false positives and false negatives. AUC, ranging from 0.89 to 0.936, further confirms the model's ability to discriminate between different classes, even in the presence of overlapping behavioral patterns.

Table III demonstrates the observation ratio evaluation metrics of the proposed MRCNN-LSTM. Observation ratio refers to the proportion of the video sequence or frame sequence that is processed or observed by the model before deciding on the class or anomaly type. An observation ratio of 0.1 means that only 10% of the total video frames from the beginning are used to make predictions. An observation ratio of 1.0 means that the entire video sequence is processed before inference. This setup simulates early event detection, which is critical in real-world surveillance scenarios, where detecting suspicious behavior as early as possible can improve response time. Therefore, the proposed model predicts suspicious behaviors under various uncertain circumstances.

TABLE II. MRCNN-LSTM PERFORMANCE ON UCF-CRIME

Class	Precision	Recall	F1-score	AUC
Normal	0.92	0.94	0.93	0.936
Fighting	0.90	0.89	0.88	0.94
Shoplifting, Stealing, and Shooting	0.88	0.86	0.88	0.93
Running	0.86	0.82	0.86	0.91
Loitering	0.87	0.81	0.83	0.90
Vandalism	0.88	0.84	0.86	0.92
Abuse, Assault, and Arrest	0.85	0.80	0.82	0.89

TABLE III. PERFORMANCE OF MRCNN-LSTM BASED ON OBSERVATION RATIO

Dataset	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
CUHK	84.67	85.89	86.09	88.90	89.00	89.99	90.23	92.19	93.69	96.99
Snatch1.0	82.03	83.09	84.24	84.90	85.56	86.12	86.99	87.56	89.81	92.99
UCF Crime	85.81	86.10	87.81	88.90	90.94	91.60	93.08	93.65	94.59	94.89

IV. CONCLUSION

The proposed anomaly detection method involves a sequence of steps: preprocessing, classification, and prediction. Initially, video footage is split into frames, and FLAHE is applied for contrast enhancement to improve frame quality. The GrabCut technique segments the frames to isolate foreground objects, followed by object detection using YOLOv5, which identifies specific objects for accurate tracking. Low- and high-level features are extracted from these objects to capture appearance and behavior characteristics. These features are then fed into the MRCNN-LSTM model, which classifies frames as anomalous or non-anomalous. This hybrid approach offers accurate object detection by ensuring precise segmentation of objects, improved feature extraction using advanced preprocessing to enhance frame quality, and temporal analysis using LSTM to capture time-dependent anomalies in video sequences. The method was tested on three datasets, CUHK, Snatch1.0, and UCF Crime, achieving accuracies of 94.9%, 92.2% and 93.6%, respectively. The high accuracy on CUHK is attributed to its simpler sequences. By enabling early detection of anomalies, this approach allows for proactive measures to be taken, enhancing public safety and security by preventing unusual events before they occur.

REFERENCES

[1] L. Ruff *et al.*, "A Unifying Review of Deep and Shallow Anomaly Detection," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 756–795, Feb. 2021, https://doi.org/10.1109/JPROC.2021.3052449.

[2] D. Roy and K. M. C., "Snatch theft detection in unconstrained surveillance videos using action attribute modelling," *Pattern Recognition Letters*, vol. 108, pp. 56–61, Jun. 2018, https://doi.org/10.1016/j.patrec.2018.03.004.

[3] L. Shen, Z. Li, and J. Kwok, "Timeseries Anomaly Detection using Temporal Hierarchical One-Class Network," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13016–13026, 2020.

[4] S. Das, W. K. Wong, T. Dietterich, A. Fern, and A. Emmott, "Discovering Anomalies by Incorporating Feedback from an Expert,"

ACM Transactions on Knowledge Discovery from Data, vol. 14, no. 4, Mar. 2020, https://doi.org/10.1145/3396608.

[5] S. Nedelkoski, J. Bogatinovski, A. Acker, J. Cardoso, and O. Kao, "Self-Attentive Classification-Based Anomaly Detection in Unstructured Logs," in *2020 IEEE International Conference on Data Mining (ICDM)*, Sorrento, Italy, Nov. 2020, pp. 1196–1201, https://doi.org/10.1109/ICDM50108.2020.00148.

[6] R. Wang, F. Kong, H. Sudler, and X. Jiao, "Brief Industry Paper: HDAD: Hyperdimensional Computing-based Anomaly Detection for Automotive Sensor Attacks," in *2021 IEEE 27th Real-Time and Embedded Technology and Applications Symposium (RTAS)*, Nashville, TN, USA, May 2021, pp. 461–464, https://doi.org/10.1109/RTAS52030.2021.00052.

[7] I. Golan and R. El-Yaniv, "Deep Anomaly Detection Using Geometric Transformations," in *Advances in Neural Information Processing Systems*, 2018, vol. 31.

[8] C. Lee, J. Kim, and S. Kang, "Semi-supervised Anomaly Detection with Reinforcement Learning," in *2022 37th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, Phuket, Thailand, Jul. 2022, pp. 933–936, https://doi.org/10.1109/ITC-CSCC55581.2022.9895028.

[9] C. Zhang, W. Zuo, P. Yang, Y. Li, and X. Wang, "Outsourced privacy-preserving anomaly detection in time series of multi-party," *China Communications*, vol. 19, no. 2, pp. 201–213, Feb. 2022, https://doi.org/10.23919/JCC.2022.02.016.

[10] Z. Wu, H. Su, and Q. Du, "Low-Rank and Collaborative Representation for Hyperspectral Anomaly Detection," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Yokohama, Japan, Jul. 2019, pp. 1394–1397, https://doi.org/10.1109/IGARSS.2019.8900381.

[11] S. M. A. Karim, N. Ranjan, and D. Shah, "A Scalable Approach to Time Series Anomaly Detection & Failure Analysis for Industrial Systems," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA, Jan. 2020, pp. 0678–0683, https://doi.org/10.1109/CCWC47524.2020.9031262.

[12] Y. Jiang, W. Wang, and C. Zhao, "A Machine Vision-based Realtime Anomaly Detection Method for Industrial Products Using Deep Learning," in *2019 Chinese Automation Congress (CAC)*, Hangzhou, China, Nov. 2019, pp. 4842–4847, https://doi.org/10.1109/CAC48633.2019.8997079.

[13] Y. Zhang, M. Xu, Y. Fan, Y. Zhang, and Y. Dong, "A Kernel Background Purification Based Anomaly Target Detection Algorithm for Hyperspectral Imagery," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Yokohama, Japan, Jul. 2019, pp. 441–444, https://doi.org/10.1109/IGARSS.2019.8900606.

[14] S. Chen, X. Li, and L. Zhao, "Hyperspectral Anomaly Detection with Data Sphering and Unsupervised Target Detection," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, Kuala Lumpur, Malaysia, Jul. 2022, pp. 1975–1978, https://doi.org/10.1109/IGARSS46834.2022.9884083.

[15] J. Luo, J. Lin, Z. Yang, and H. Liu, "SMD Anomaly Detection: A Self-Supervised Texture–Structure Anomaly Detection Framework," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022, https://doi.org/10.1109/TIM.2022.3194920.

[16] K. W. Cheng, Y. T. Chen, and W. H. Fang, "Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 2909–2917, https://doi.org/10.1109/CVPR.2015.7298909.

[17] D. Varga and T. Szirányi, "Robust real-time pedestrian detection in surveillance videos," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 1, pp. 79–85, Feb. 2017, https://doi.org/10.1007/s12652-016-0369-0.

[18] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, Jun. 2018, pp. 6479–6488, https://doi.org/10.1109/CVPR.2018.00678.

[19] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Computer Vision and Image Understanding*, vol. 156, pp. 117–127, Mar. 2017, https://doi.org/10.1016/j.cviu.2016.10.010.

[20] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X. S. Hua, "Spatio-Temporal AutoEncoder for Video Anomaly Detection," in *Proceedings of the 25th ACM International Conference on Multimedia*, Jul. 2017, pp. 1933–1941, https://doi.org/10.1145/3123266.3123451.

[21] L. Ruff *et al.*, "Deep One-Class Classification," in *Proceedings of the 35th International Conference on Machine Learning*, Jul. 2018, pp. 4393–4402.

[22] A. Gao and J. Liu, "STEAD: Spatio-Temporal Efficient Anomaly Detection for Time and Compute Sensitive Applications." arXiv, Mar. 11, 2025, https://doi.org/10.48550/arXiv.2503.07942.

[23] H. Singh, A. Kumar,. L. K. Balyan, and G. K. Singh, "Slantlet filter-bank-based satellite image enhancement using gamma-corrected knee transformation," *International Journal of Electronics*, vol. 105, no. 10, pp. 1695–1715, Oct. 2018, https://doi.org/10.1080/00207217.2018.1477199.

[24] M. Kumar and A. K. Bhandari, "Contrast Enhancement Using Novel White Balancing Parameter Optimization for Perceptually Invisible Images," *IEEE Transactions on Image Processing*, vol. 29, pp. 7525–7536, 2020, https://doi.org/10.1109/TIP.2020.3004036.

[25] D. Raj and P. Mamoria, "Comparative analysis of contrast enhancement techniques on different images," in *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, Greater Noida, Delhi, India, Oct. 2015, pp. 27–31, https://doi.org/10.1109/ICGCIoT.2015.7380422.

[26] S. P. Brintha, S. P. Premnath, and S. E. Lawrence, "Contrast enhancement using discrete wavelet transform and adaptive tranfer function," in *International Conference on Information Communication and Embedded Systems (ICICES2014)*, Chennai, India, Feb. 2014, pp. 1–4, https://doi.org/10.1109/ICICES.2014.7034054.

[27] D. Vijayalakshmi and M. K. Nath, "A Novel Contrast Enhancement Technique using Gradient-Based Joint Histogram Equalization," *Circuits, Systems, and Signal Processing*, vol. 40, no. 8, pp. 3929–3967, Aug. 2021, https://doi.org/10.1007/s00034-021-01655-3.

[28] W. Zhan *et al.*, "An improved Yolov5 real-time detection method for small objects captured by UAV," *Soft Computing*, vol. 26, no. 1, pp. 361–373, Jan. 2022, https://doi.org/10.1007/s00500-021-06407-8.

[29] A. Benjumea, I. Teeti, F. Cuzzolin, and A. Bradley, "YOLO-Z: Improving small object detection in YOLOv5 for autonomous vehicles." arXiv, Jan. 03, 2023, https://doi.org/10.48550/arXiv.2112.11798.

[30] M. Anwer, S. M. Khan, M. U. Farooq, and Waseemullah, "Attack Detection in IoT using Machine Learning," *Engineering, Technology & Applied Science Research*, vol. 11, no. 3, pp. 7273–7278, Jun. 2021, https://doi.org/10.48084/etasr.4202.

[31] Y. Fan, G. Wen, D. Li, S. Qiu, M. D. Levine, and F. Xiao, "Video anomaly detection and localization via Gaussian Mixture Fully Convolutional Variational Autoencoder," *Computer Vision and Image Understanding*, vol. 195, Jun. 2020, Art. no. 102920, https://doi.org/10.1016/j.cviu.2020.102920.

[32] "Real-world Anomaly Detection in Surveillance Videos." [Online]. Available: https://www.crcv.ucf.edu/projects/real-world/.

[33] "UCF Crime Dataset." [Online]. Available: https://www.kaggle.com/datasets/odins0n/ucf-crime-dataset.

[34] "Avenue Dataset." [Online]. Available: https://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html .

[35] T. Reiss and Y. Hoshen, "An Attribute-based Method for Video Anomaly Detection," *Transactions on Machine Learning Research*, Jul. 2024.

[36] D. Li, X. Nie, R. Gong, X. Lin, and H. Yu, "Multi-Branch GAN-Based Abnormal Events Detection via Context Learning in Surveillance Videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 3439–3450, Feb. 2024, https://doi.org/10.1109/TCSVT.2023.3325451.

[37] D. Manju, M. Seetha, and P. Sammulal, "Early Action Prediction Using 3DCNN With LSTM and Bidirectional LSTM," *SSRN Electronic Journal*, 2021, https://doi.org/10.2139/ssrn.3815963.

[38] A. Phapale and S. Bhingarkar, "Deep Context-Aware Feature Extraction for Anomaly Detection in Surveillance Videos," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 21633–21638, Apr. 2025, https://doi.org/10.48084/etasr.9810.

[39] A. B. Altamimi and H. Ullah, "Panic Detection in Crowded Scenes," *Engineering, Technology & Applied Science Research*, vol. 10, no. 2, pp. 5412–5418, Apr. 2020, https://doi.org/10.48084/etasr.3347.