Real-world Anomaly Detection in Surveillance Videos using YOLO-World and BiLSTM Framework

OPEN ACCESS

Volume: 13

Special Issue: 2

Month: January

Year: 2026

E-ISSN: 2582-0397

P-ISSN: 2321-788X

Citation:
Vadke, Janhavi Mandar. "Real-World Anomaly Detection in Surveillance Videos Using YOLO-World and BiLSTM Framework." *Shanlax International Journal of Arts, Science and Humanities*, vol. 13, no. 2, 2026, pp. 165–68.

DOI:
https://doi.org/10.34293/sijash.v13iS2-i2-Jan.10539

Janhavi Mandar Vadke
Department of Information Technology
Vidyalankar School of Information Technology, Mumbai, Maharashtra, India

Abstract
Detecting suspicious activities using Closed-Circuit Television (CCTV) is essential for improving security in public and semi-public spaces like schools, colleges, residential areas, parks, and workplaces. Conventional surveillance systems depend largely on constant human oversight from security staff, which can become ineffective and prone to mistakes when trying to monitor several camera feeds at once. This frequently leads to slow reactions or overlooking suspicious or unusual behaviors. This project suggests an automated suspicious activity detection system that uses deep learning techniques implemented using TensorFlow and Python in order to overcome these restrictions. A human-focused open-vocabulary detector called YOLO-World and BiLSTM Framework is employed. After suppressing the background noise, BiLSTM is employed to comprehend temporal behavior. In order to spot unusual or suspect human activity in real time, the system continuously examines live CCTV video streams. Instant notifications are created and transmitted to authorized personnel upon detection of such events, facilitating quicker and more informed decision-making. By automating the monitoring process and reducing reliance on humans, the suggested method greatly lessens the cognitive strain on security officers. It guarantees continuous observation without weariness, increases accuracy, and speeds up response times. The suggested method provides a more effective, scalable, and intelligent security mechanism than traditional surveillance systems. The technology helps create safer and more secure environments, which benefits society as a whole by facilitating prompt intervention and proactive threat identification.

Keywords: Surveillance Systems, Yolo-World, Bilstm, Suspicious, Closed-Circuit Television (Cctv), Unusual Behaviors

Introduction

In this system we present the object detection technique to determine the suspicious activities in open spaces. There will be an alert on the display in the monitoring room if there are any suspicious activities. The system will convert the videos into frames and using object detection algorithms the person will come to know. The algorithm will perform background subtraction and perform different object detection models. The system will scan through the real time video and detect the activity. The entire system is built using TensorFlow, object-detection, and OpenCV which are used for image processing and detection. This system is used on real time videos and can detect the activities.

When security guards are overburdened with the task of keeping an eye on numerous security monitors, it can become tiresome, and they begin to make mistakes and miss crucial moments when monitoring is essential. As there might be multiple screens, it becomes tough for the security guard to

keep a watch on all the screens; it might lead to any misconduct that the security might not be able to see. This may cause a big incident in the locality.

This system will detect the suspicious activity and will help the security guard to know if there are any activities happening which are not normal. This system will also assist organizations which do not have a security guard. This system will use the object detection technique to know whether the activity is suspicious.

The security has just to see at which location this activity has happened and take the necessary steps to prevent any further problems. The software will use real time videos and give the actual output at the same time to the authorized person.

Related Work / Literature Review

The results show that detection accuracy is not significantly affected by variations among pre-trained convolutional models. The study also identifies important obstacles and suggests promising avenues for further investigation.

This study talks about the growing problem of testing mobile apps by suggesting a vision-based robotic testing method that mimics the behaviors of expert testers. To solve the problem of identifying complicated hand movements from 2D images, the study uses an updated YOLOv5 algorithm for precise hand localization and an improved ResNet-152 model for action categorization.

This research focuses on human action recognition in movies, which entails comprehending both spatial and temporal information. Three deep learning techniques—Two-Stream CNN, 3D CNN and CNN+LSTM—are developed and assessed to address these problems. All three methods successfully identify human behaviors, according to experiments on the HMDB-51 dataset, with the best-performing algorithm determined by the outcomes of the experiments.

It suggests a deep learning method that blends recurrent neural networks (RNN) for temporal analysis with convolutional neural networks (CNN) for feature extraction. Five distinct sports categories are classified using this method. According to experimental data, performance is greatly enhanced by adopting suitable frame sequences, with classification accuracy reaching up to 96.66%.

This paper presents a hybrid framework. The suggested approach defines basic behavioral heuristics and incorporates them into physical equations. The resulting heuristic-based features outperform pre-trained convolutional networks, motion descriptors, and physics-based models on common benchmarks, achieving state-of-the-art performance.

The study presents a technique that uses Directional Motion History Images (DMHIs) to represent and translate human motion. Histograms of Oriented Gradients (HOG) features derived from DMHIs are used to recognize basic motions, and tests are carried out to find the best HOG parameters, such as bin numbers and cell sizes, for better recognition performance.

Advanced Research and Development Activity (ARDA) sponsored the VACE (Video Analysis and Content Extraction) program, containing a variety of video types, such as news, surveillance, and reconnaissance footage. The program's main goal is to improve video event detection, recognition, and comprehension, especially in surveillance footage.

Early approaches to surveillance relied on motion detection and background subtraction techniques. While these methods were computationally efficient, they were highly sensitive to environmental changes such as lighting variations and camera motion. Traditional machine learning techniques, including Support Vector Machines (SVMs) and Hidden Markov Models (HMMs), were later applied for activity recognition but required handcrafted features and lacked robustness.

The introduction of deep learning transformed video surveillance research. Convolutional Neural Networks (CNNs) enabled automatic feature extraction for object detection and classification. Region-based detectors such as R-CNN, Fast R-CNN, and Faster R-CNN improved detection accuracy but suffered from high computational complexity. YOLO addressed this limitation by introducing a single-stage detection framework capable of real-time performance.

For temporal activity recognition, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks demonstrated strong capability in modeling sequential data. Bidirectional LSTM (BiLSTM) networks further improved performance by learning contextual information from both past and future frames. Recent studies indicate that combining CNN-based spatial feature extraction with LSTM-based temporal modeling yields superior performance in abnormal activity detection.

Research Gap

Most models perform well on specific datasets, but fail in new or unseen real-world scenes due to domain shifts (lighting, camera angles, crowd density). Most anomaly detection systems rely on closed-set object detectors trained on limited categories and fixed environments. These detectors fail when deployed in new surveillance scenes with unseen object types or contextual variations.

CNN and transformer-based approaches often analyze short clips or frame-level features, failing to capture long-term temporal dependencies critical for recognizing suspicious behavior (e.g., loitering, stalking).

Research Methodology and Analysis

Dataset: Datasets were collected from publicly available sources such as Kaggle and online surveillance repositories.

Labelling: For object detection, suspicious activities like knife point, gun point, and criminal activities are clearly labelled and defined.

TensorFlow: The loss function sparse_categorical_crossentropy is used. 16 strides are used, which tells how far the filter moves in every step along one direction.

OpenCV: The cv2 library is used in which methods like VideoCapture are used for getting live stream from CCTV cameras.

Frame Extraction Module: In this module the real time CCTV footage will be converted into frames which will be further used by the next module.

Background Subtraction Module: In this module the background of the frame obtained by the frame extraction module will be subtracted and the person will be only visible to the system.

Activity Detection Module: The activity the person is doing will be detected in this module and this model will be using the machine learning algorithm for this.

YOLO-Based Object Detection: YOLO is employed to detect humans and potential threat objects in each frame. The model performs object detection in a single forward pass, enabling real-time performance. Detected bounding boxes and class probabilities are used as spatial features for further analysis.

I. Segregating Normal and Suspicious Activity Module: When activity is detected, then this module will determine in which category the activity goes. There are two categories: Normal and Suspicious.

J. Temporal Feature Modeling Using BiLSTM: Sequential features extracted from YOLO detections are fed into a BiLSTM network. The BiLSTM captures temporal dependencies and behavioral patterns across consecutive frames, allowing the system to distinguish between normal and suspicious activities.

K. Alert Module: In this module if there is any suspicious activity detected it will provide an alert to the security or the user.

Implemented Methodologies

The methodology adopted in this research includes dataset preparation, model training, and system integration.

A. Dataset Collection

Video datasets were collected from publicly available sources such as Kaggle and online surveillance repositories. The datasets include various scenarios with normal activities and suspicious behaviors.

B. Data Annotation and Augmentation

Frames were annotated with bounding boxes and activity labels. Data augmentation techniques such as rotation, flipping, and scaling were applied to improve model generalization.

C. Model Training

YOLO was trained to detect humans and relevant objects, while the BiLSTM network was trained on sequential features to recognize temporal behavior patterns. Cross-entropy loss and Adam optimizer were used during training.

Results and Performance Evaluation

TensorFlow and OpenCV were used for model development and deployment. The system was evaluated using metrics such as accuracy, precision, recall, and F1-score. The proposed YOLO–BiLSTM framework achieved an overall accuracy of approximately 92% in detecting suspicious activities. The system demonstrated real-time performance with low latency, making it suitable for live surveillance applications. Comparative analysis showed that the proposed approach outperformed traditional CNN-only models. The integration of spatial and temporal modeling significantly improves the system's ability to recognize complex human behaviors. YOLO provides fast and accurate object detection, while BiLSTM effectively captures temporal dependencies. Challenges remain in crowded environments and low-light conditions, which will be addressed in future work.

Conclusion and Future Work

This paper presented an intelligent suspicious activity detection framework using YOLO and BiLSTM for automated CCTV surveillance. The proposed system reduces human monitoring effort, improves detection accuracy, and enhances public safety. Future work includes deployment on edge devices, multi-camera integration, and incorporation of privacy-preserving mechanisms.

References

1. Md. Haidar Sharif, Lei Jiao, Christian Omlin, "Deep crowd anomaly detection: state-of-the-art, challenges, and future research directions," 20 February 2025.
2. Tao Zhang, Zhengqi Su, Jing Cheng, Feng Xue, Shengyu Liu, "Machine vision-based testing action recognition method for robotic testing of mobile application," International Journal of Distributed Sensor Networks, August 4, 2022.
3. Zeqi Yu, Wei Qi Yan, "Human Action Recognition Using Deep Learning Methods," Published in: 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ). Publisher: IEEE.
4. Mohammad Ashraf Russo, Alexander Filonenko, Kang-Hyun Jo, "Sports Classification in Sequential Frames Using CNN and RNN," 2018 International Conference on Information and Communication Technology Robotics (ICT-ROBOT), September 2018.
5. Sadegh Mohammadi, Alessandro Perina, Hamed Kiani, Vittorio Murino, "Detecting Violent Events in Videos," European Conference on Computer Vision, October 2016.
6. Makoto Murakami, Joo Kooi Tan, Hyoungseop Kim, Kyushu Institute of Technology, Seiji Ishikawa, "Human motion recognition using directional motion history images," January 2010.
7. J.D. Prange, "Detecting, recognizing and understanding video events in surveillance video," Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, 2003.