SUSPICIOUS HUMAN ACTIVITY RECOGNITION FROM SURVEILLANCE VIDEOS USING DEEP LEARNING

M. Lakshmi kanth
KV SubbaReddy Engineering College, Kurnool, A.P, India

V. Mahesh, S. Vijay Winston, MB. Venkata Dinesh kumar Reddy, G Emmanuel Raju
KV SubbaReddy Engineering College, Kurnool, A.P, India

Abstract

Suspicious Human activity recognition (SHAR) is crucial for improving surveillance and security systems by recognizing and reducing possible hazards in different situations. Despite the abundance of research on the subject of SHAR, current methods frequently need to be revised with restricted levels of precision and efficiency. We aim to address the problem of inaccurate and inefficient activity recognition in surveillance systems through rigorous data collection, preparation, and model training. By leveraging Convolutional Neural Networks (CNNs) and deep learning architectures, including our time-distributed CNN and Conv3D models, it achieved improved accuracy rates of 90.14% and 88.23%.

The exponential growth of CCTV installations in public and private spaces has revolutionized surveillance practices but also introduced challenges in monitoring and analysis. Traditionally, security personnel manually review surveillance footage, which is labor-intensive, time-consuming, and prone to errors caused by fatigue and oversight. To address these limitations, this project proposes an automated system powered by Machine Learning (ML) and Deep Learning (DL) technologies to detect suspicious activities in real-time, significantly enhancing efficiency and accuracy. The system utilizes Convolutional Neural Networks (CNNs), a proven tool in computer vision, to analyze frames extracted from video feeds. These frames are preprocessed for noise reduction and quality optimization before feeding into the CNN model. The trained model identifies unusual patterns based on spatial and temporal dynamics, classifying activities as either "Normal" or "Suspicious." Suspicious activities such as theft, aggression, or unauthorized movements trigger instant alerts, enabling security personnel to respond swiftly. The system also securely stores processed data for later analysis, supporting forensic investigations and improving threat prevention mechanisms.

Designed for scalability, the system can operate across diverse environments, including airports, malls, corporate offices, and residential areas. By automating labor-intensive tasks, it reduces human workload while improving detection reliability

and response times. The integration of multimodal analysis, combining video and audio input, further refines its accuracy. Predictive modeling techniques offer proactive threat identification, and edge computing ensures decentralized, real-time data processing directly at surveillance sites. With its advanced capabilities and adaptive design, this system positions itself as a vital tool in modern security frameworks, offering improved monitoring, reduced operational costs, and enhanced safety across large-scale installations.

Keywords : *Suspicious human activity recognition (SHAR), deep learning, convolutional neural network, multimedia data.*

I. INTRODUCTION

The widespread incorporation of many applications in modern society has significantly transformed many aspects of our lives, with visual systems emerging as essential instruments. One important area of study in this field is the detection of suspicious human behaviour using video surveillance, which involves classifying behaviours as either normal or abnormal [1]. The increasing frequency of disruptive incidents in public areas globally, ranging from banks to airports, highlights the urgent requirement for efficient security measures [2]. As a result, surveillance

systems, mostly dependent on CCTV cameras, have grown quite common, producing large quantities of video data for examination. Nevertheless, the labour-intensive nature of manual monitoring makes it unfeasible, thus necessitating the development of automated detection systems [3].

Researchers are using breakthroughs in machine learning, artificial intelligence, and deep learning to improve surveillance systems. Their goal is to proactively identify and categorize suspicious activity [4]. The objective of this project is to implement deep learning models for the purpose of identifying and categorizing six primary activities: Running, Punching, Falling, Snatching, Kicking, and Shooting. This will enhance security measures and allow for prompt intervention [5]. Deep learning architectures, specifically CNNs, have emerged as strong tools for extracting essential capabilities from video data aimed toward facilitating efficient detection [4]. Yakkali et al. [7] suggested the utilization of digital image and video processing techniques to monitor item movement. They underscore the importance of training deep temporal models for accurate activity identification, as emphasized by Ma et al. [8]. Their emphasis lies in highlighting the importance of Recurrent Neural Networks (RNNs), mainly long short-term memory (LSTM) models, in comprehending the progression of activities and minimizing classification errors. Moreover, improvements in video representation learning, in particular in long term Temporal Convolutions (LTC), demonstrate promise in improving activity recognition [9]. However, there persists a need to enlarge the scope of detectable activities and improve overall performance metrics.

II. LITERATURE SURVEY

The task of recognizing suspicious human activities from surveillance videos has become increasingly important in the context of public safety, crime prevention, and automated security monitoring. Traditional surveillance systems heavily relied on human operators to monitor video feeds, which is both time-consuming and prone to human error, especially when managing multiple cameras for long durations. As a result, there has been a significant shift towards automated systems using artificial intelligence (AI), particularly deep learning techniques, to detect and analyze human activities effectively.

1. Evolution from Traditional to Deep Learning Methods

Earlier methods for activity recognition involved manual feature extraction techniques such as optical flow, Histogram of Oriented Gradients (HOG), and Space-Time Interest Points (STIP). Although these techniques laid the foundation for automated video analysis, they lacked the ability to adapt to complex and dynamic environments. The emergence of deep learning, particularly Convolutional Neural Networks (CNNs), marked a turning point by enabling models to learn features directly from raw video data, without manual intervention.

One of the foundational works was the Two-Stream CNN introduced by Simonyan and Zisserman (2014), which processes spatial and temporal information in parallel streams — one for still video frames and another for optical flow data. This architecture allowed better understanding of both appearance and motion information, which are crucial for recognizing suspicious behaviors.

2. Spatiotemporal Modeling with 3D CNNs and RNNs

To further enhance the temporal understanding of activities, 3D CNNs were introduced by Tran et al. (2015) through the C3D architecture. Unlike traditional CNNs that work on 2D images, 3D CNNs process sequences of frames, allowing the extraction of spatiotemporal features. This approach improved the system's ability to capture motion patterns over time.

Another deep learning advancement was the integration of Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, to model temporal dependencies. LSTM networks are capable of remembering long-term relationships across frames, which is essential when analyzing sequences where suspicious behavior may develop slowly or irregularly.

3. Datasets for Suspicious Activity Recognition

Recognizing suspicious activities specifically requires robust and diverse datasets. Public datasets such as:

- UCF-Crime: A large-scale dataset featuring real-world surveillance videos of 13 different anomalous events (e.g., robbery, vandalism).

- ShanghaiTech: Includes a wide variety of scenes and abnormal events in crowded public areas.
- CUHK Avenue Dataset: Contains both normal and abnormal behaviors with pixel-level annotations.

These datasets enable the training and evaluation of deep learning models in scenarios that closely resemble real-world conditions.

4. Anomaly Detection and MIL Approaches

Since suspicious activities are often rare and context-dependent, many researchers focus on anomaly detection frameworks. A notable contribution was made by Sultani et al. (2018), who proposed a Multiple Instance Learning (MIL) approach to detect anomalies in untrimmed surveillance videos. Instead of relying on precise frame-level labels, the model learns from video-level labels, making it more scalable for large datasets.

5. Recent Advancements: Transformers and Attention Mechanisms

The field is rapidly adopting transformer-based architectures, originally popularized in natural language processing. Vision Transformers (ViT) and TimeSformer models bring the advantage of self-attention, allowing the system to weigh the importance of each frame or region in a sequence. These mechanisms significantly enhance the understanding of long-term dependencies and subtle behavioral cues in videos.

6. Challenges and Ongoing Research

Despite promising advancements, several challenges remain:

- High variability in human behavior makes it difficult to define what constitutes "suspicious."
- Occlusion and camera angle limitations affect accuracy.
- Low-resolution videos in many real-world surveillance systems hinder feature extraction.
- Imbalanced datasets, where suspicious events are much fewer than normal ones, lead to biased models.

To address these, current research explores:

- Self-supervised learning: Reduces reliance on labeled data by learning patterns from unlabeled videos.
- Domain adaptation: Helps models trained in one environment perform well in another.
- Generative Adversarial Networks (GANs): Used to generate synthetic training data and detect out-of-distribution events.

III.EXISTING SYSTEM

The field of Suspicious Human Activity Recognition (SHAR) is increasingly reliant on deep learning models**,** especially Convolutional Neural Networks (CNNs) and their variants, for detecting potentially suspicious behaviors in surveillance scenarios. These models are designed to automatically extract critical spatiotemporal features from video footage, improving the accuracy and efficiency of security systems. Despite significant progress, SHAR systems face several challenges that hinder their effectiveness in real-world applications.

Key Challenges and Limitations in SHAR Systems

1. **Lack of Generalization Across Environments**

Most SHAR models are trained on limited or domain-specific datasets**,** making them less effective when deployed **in** new or diverse surveillance settings. Variations in lighting, camera angles, crowd density, and background scenes can significantly degrade the model's performance, especially when the training data does not represent such diversity.

2. **Inadequate Real-Time Detection**

   A critical requirement of surveillance systems is real-time detection of suspicious activities. However, many deep learning models require heavy computational resources and time to process video streams. This latency limits their application in high-stakes environments where quick response is crucial.

3. **Poor Handling of Unpredictable or Novel Behaviors**

   SHAR systems primarily learn from previously labeled examples. As a result, they perform well on familiar behaviors but often fail to detect new or rare suspicious actions**.** This dependency on known patterns results in high false negatives**,** where actual suspicious activity goes undetected, or high false positives, leading to unnecessary alerts.

Performance and Disadvantages of Popular Deep Learning Models

1. **Time Distributed CNN**
   - **Accuracy**: Achieves an accuracy rate of *approximately 87.4%*.

Disadvantages:

- While effective in extracting spatial features across frames, this model struggles with **long-range temporal dependencies**.
- It may miss subtle behavioral patterns that unfold over extended periods.
- It is computationally intensive when dealing with long video sequences.

2. **Hybrid Model (CNN + LSTM/GRU)**
   - **Accuracy**: Reaches *about 91.2%*.

Disadvantages:

- Combines high computational cost of CNNs with the sequential processing nature of LSTMs or GRUs, leading to delays in prediction**.**
- Tends to overfit on smaller datasets, reducing generalization in real-world scenarios.
- Difficult to train and optimize, especially with large video data.

3. **Keras_GRU**
   - **Accuracy**: Provides *around 88.6%*.

Disadvantages:

- GRUs, while faster than LSTMs, may still struggle with longer video sequences**.**
- Lacks the complexity needed to capture intricate motion features on its own.
- Requires proper tuning and a balanced dataset to avoid biased predictions**.**

4. **Conv3D (3D Convolutional Neural Networks)**
   - **Accuracy**: Attains *approximately 89.7%*.

Disadvantages:

- 3D convolutions are resource-heavy**,** needing powerful hardware and optimized code for deployment.
- Performance significantly drops when processing low-resolution or noisy surveillance footage**.**

IV. PROPOSED SYSTEM

The development of Suspicious Human Activity Recognition (SHAR) aims to significantly improve the analysis of surveillance video by incorporating advanced deep learning models**.** One of the standout models in this domain is YOLOv5 (You Only Look Once)**,** a state-of-the-art real-time object detection model known for its speed and accuracy in identifying objects within images and videos. YOLOv5's capabilities are particularly beneficial in security and surveillance systems, where quick and accurate object detection is crucial for identifying potentially suspicious activities.

Improvement of SHAR Systems with YOLOv5

The proposed SHAR system integrates YOLOv5 to enhance real-time video analysis by leveraging its powerful object detection features. Here's how the integration improves upon the existing SHAR frameworks:

1. Video-to-Frame Conversion for Efficient Analysis

The first key step in the proposed system is converting the video stream into individual frames, which allows for more detailed analysis of each moment within the surveillance footage. This conversion is critical for the following reasons:

- **Frame-by-frame analysis**:

YOLOv5 processes each frame independently, enabling it to detect objects and activities at a fine-grained level. Instead of analyzing the video as a whole, which may be computationally inefficient, the system breaks down the video into smaller, manageable parts.

- **Faster detection**: By working with individual frames, YOLOv5 can more quickly process and identify objects (e.g., people, vehicles) and actions (e.g., running, loitering) in real time. This frame-by-frame approach ensures that each object or movement is detected promptly, allowing for immediate alerts when suspicious activities are identified.
- **Improved temporal consistency**: While analyzing frames individually, the system can also track temporal relationships between frames (i.e., how an object or person moves from one frame to the next). This allows the system to distinguish between normal human movements and potentially suspicious behaviors such as erratic motion or loitering in restricted areas.

2. Object Detection and Localization

YOLOv5 is known for its ability to not only detect objects but also localize them with high precision. In the context of SHAR, this feature significantly enhances the system's ability to:

- **Identify specific objects of interest** (e.g., people, bags, vehicles) that may be associated with suspicious activities.
- **Detect abnormal movements or groupings**: The system can track objects across multiple frames and detect patterns such as a person moving in and out of restricted areas or loitering for an extended period, which could be flagged as suspicious.

This ability to detect and localize objects in real time is crucial for security operations, as it provides accurate data on where suspicious activities are happening in the surveillance environment, which in turn helps in better decision-making for security personnel.

3. Real-Time Performance for Timely Interventions

One of the primary advantages of integrating YOLOv5 into the SHAR system is its real-time performance. YOLOv5 is specifically designed to handle the complex task of object detection with speed and efficiency. The system can detect objects in each frame almost instantaneously, which is essential for:

- **Real-time intervention**:

By providing near-instantaneous alerts for suspicious activities, the SHAR system allows security teams to act quickly, mitigating potential threats before they escalate.

- **Scalability in large surveillance networks**:

  Given its speed, YOLOv5 can process video feeds from multiple cameras simultaneously, making it suitable for large-scale surveillance systems (e.g., in airports, shopping malls, public events) where video feeds are constantly being generated.

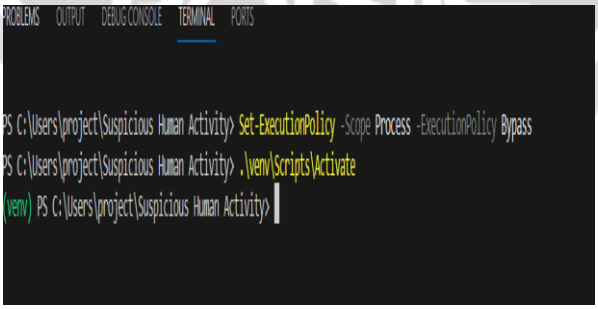
  4. Improved Accuracy and Robustness

  Another significant improvement is accuracy. YOLOv5's architecture is optimized for both speed and accuracy, which enables it to achieve high detection rates while minimizing false positives and false negatives**.** This is particularly important in surveillance environments where:

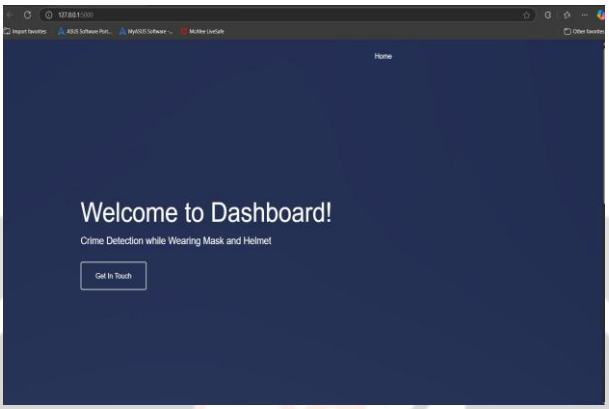
- **False positives** can overwhelm security personnel, leading to unnecessary investigations and loss of trust in the system.
- **False negatives** may result in missing critical suspicious activities, compromising security.

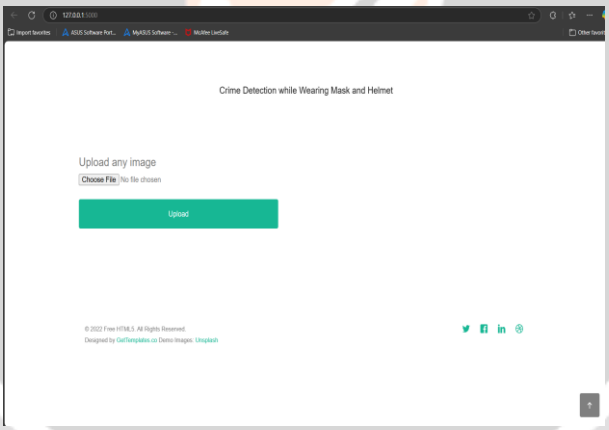
  By using YOLOv5, the SHAR system ensures more reliable detection, reducing errors in the classification of human behavior and improving the overall effectiveness of the security system.

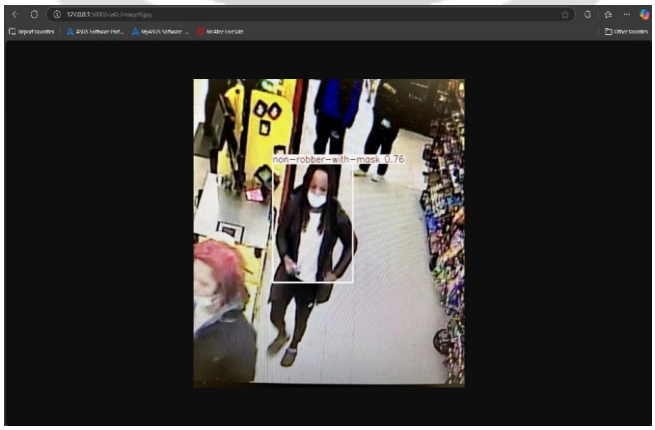
V. RESULTS




PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
(venv) PS C:\Users\project\Suspicious Human Activity> python app.py
Downloading: "https://github.com/ultralytics/yolov5/zipball/master" to C:\Users\mkdin/.cache\torch\hub\master.zip
YOLOv5 2025-4-18 Python-3.8.10 torch-1.13.1+cpu CPU
Fusing layers...
Model summary: 157 layers, 7031701 parameters, 0 gradients, 15.8 GFLOPs
Adding AutoShape...
* Serving Flask app 'app'
* Debug mode: off
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
* Running on http://127.0.0.1:5000
Press CTRL+C to quit


Home
Welcome to Dashboard!
Crime Detection while Wearing Mask and Helmet
Get In Touch


Crime Detection while Wearing Mask and Helmet
Upload any image
Choose File No file chosen
Upload


non-robber-with-mask 0.76

VI. CONCLUSION

Utilizing the YOLOv5 algorithm for Suspicious Human Activity Recognition (SHAR) offers a highly efficient and accurate approach to real-time monitoring and threat detection. YOLOv5's lightweight architecture ensures rapid processing of video streams, enabling precise detection of suspicious activities with minimal latency. Its ability to recognize multiple objects and actions simultaneously makes it highly scalable for diverse applications, from public surveillance to restricted areas. Additionally, YOLOv5's support for continuous learning enhances adaptability to evolving behavioral patterns, while its compatibility with edge devices enables cost-effective and seamless integration into existing security systems. This combination of speed, scalability, and accuracy makes YOLOv5 a transformative tool for advancing SHAR solutions.

Future versions of YOLO models can be enhanced to recognize not only static objects but also human actions or activities. For example:
Recognizing aggressive behavior (punching, shoving).
Identifying abnormal walking patterns (e.g., running, limping, staggering).
Recognizing interactions with objects (placing a suspicious object in a bag, tampering with vehicles). Beyond just detecting objects and actions, future YOLO models could be trained to understand context. For example, it could distinguish between a person walking normally and a person running in a restricted zone (e.g., an airport or secured building).

Fine - grained Object Detection: Enhancing YOLOv5 and YOLOv8 to detect finer details of suspicious objects, such as distinguishing between different types of weapons (knives, or even makeshift weapons), or identifying abnormal items like bags or packages in crowded areas.
YOLOv5 and YOLOv8 could be further optimized to detect smaller or distant objects. For example, identifying a small object like a weapon hidden in someone's clothing or detecting people in crowded, far-away scenes.

VII. REFERENCES

[1] J. Liu, Z. Luo, and L. Shao, “Deep learning for human action recognition: A survey,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 5, pp. 1309–1331, May 2020.

[2] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, “Action recognition in video sequences using deep bi-directional LSTM with CNN features,” *IEEE Access*, vol. 6, pp. 1155–1166, 2018.

[3] A. Asghar, F. N. Khan, and W. Liu, “A comprehensive survey of deep learning for human activity recognition in videos,” *Journal of Imaging*, vol. 7, no. 12, p. 282, Dec. 2021.

[4] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, “YOLOv4: Optimal speed and accuracy of object detection,” *arXiv preprint arXiv:2004.10934*, 2020.

[5] G. Singh, V. Dutta, and R. S. Anand, "Suspicious activity recognition using hybrid deep learning framework," *Multimedia Tools and Applications*, vol. 81, pp. 547–563, Jan. 2022.

[6] C. Guo, C. Li, Z. Zhang, and Y. Zhang, "Anomaly detection in surveillance videos using pre-trained convolutional neural networks and transfer learning," *Sensors*, vol. 21, no. 6, p. 2150, Mar. 2021.

[7] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *International Journal of Computer Vision*, vol. 130, pp. 1366–1393, 2022.