ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar088

Smart Video Monitoring: Advanced Deep Learning for Activity and Object Recognition

Shashikumar D R[1]; Tejashwini N[2]; K N Pushpalatha[3]; Anurag Kumar[4]; Om Chavan[5]; Atharva Mishra[6]

[1]Computer Science and Engineering Sai Vidya Institute of Technology Bengaluru, Karnataka, India
[2]Computer Science and Engineering Sai Vidya Institute of Technology Bengaluru, Karnataka, India
[3]CSE (Data Science) Sai Vidya Institute of Technology Bengaluru, Karnataka, India
[4]Computer Science and Engineering Sai Vidya Institute of Technology Bengaluru, Karnataka, India
[5]Computer Science and Engineering Sai Vidya Institute of Technology Bengaluru, Karnataka, India
[6]Computer Science and Engineering Sai Vidya Institute of Technology Bengaluru, Karnataka, India

Publication Date: 2025/03/17

Abstract: This study explores the integration of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for the real-time recognition of human activities in video data. By harnessing the advantages of these two approaches, the system achieves high accuracy in detecting complex human actions. Specifically, CNNs address the spatial aspects of the task, while LSTMs handle the temporal sequences. A notable feature of the system is its categorization module, which enables users to select an action and identify similar actions, thereby enhancing productivity and usability.

Existing models often face challenges related to real-time inter- action capabilities and resilience to environmental disturbances. This study tackles these shortcomings by refining the CNN-LSTM framework to support real-time functionality and incorporating preprocessing techniques, such as frame extraction and normal- ization, to improve input data quality. The system's effectiveness is measured using indicators like accuracy, recall, and latency, demonstrating its advantages over traditional rule-based and basic deep learning approaches. The early findings are optimistic, demonstrating significant improvements in performance.

Nevertheless, challenges remain, particularly in tracking per- formance under occlusion or in cluttered environments. Future research should explore the integration of multi-modal data and advanced architectures, such as spatio-temporal graph con- volutional networks (STGCN), to further enhance recognition accuracy and system robustness.

In conclusion, the proposed CNN-LSTM hybrid architecture for activity recognition demonstrates potential for applications in video surveillance and beyond, including fields like healthcare and sports analytics. The system offers improved automated monitoring capabilities through enhanced accuracy, scalable human action detection, and user-friendly design.

How to Cite: Shashikumar D R; Tejashwini N; K N Pushpalatha; Anurag Kumar; Om Chavan; Atharva Mishra (2025). Smart Video Monitoring: Advanced Deep Learning for Activity and Object Recognition. *International Journal of Innovative Science and Research Technology*, 10(3), 168-172.
https://doi.org/10.38124/ijisrt/25mar088

I. INTRODUCTION

➤ *Context and Motivation*

Video surveillanceserves an essential function in ensuring public safety and protecting private property. These systems serve as a strong deterrent to crime and provide essential situational awareness. However, traditional surveillance ap- proaches rely heavily on human operators to monitor video streams, making them both labor-intensive and susceptible to errors due to operator fatigue. This challenge is exacer- bated by the growing volume of video data, underscoring the need for automated systems capable of efficiently and accurately recognizing human activities. Recent developments in deep learning offer powerful tools that could revolutionize surveillance practices, facilitating smarter and more efficient monitoring.

➤ *Problem Statement*

While progress has been achieved in activity recognition, existing surveillance systems still face significant obstacles. Challenges such as varying lighting conditions, occlusions, and dense environments impact the accuracy of human activity detection models. Furthermore, many current solutions are resource-intensive, which hinders their ability to achieving the real-time performance needed for rapid decision-making. These challenges hinder the adaptation of automated systems to dynamic environments. In addition, the lack of mechanisms to categorize and retrieve similar actions makes analyzing large volumes of footage cumbersome, resulting in delays in response times and resource allocation.

➢ *Objectives*

The main goals of this study are to:

- Design a hybrid CNN-LSTM model that combines CNN's ability to extract spatial features with LSTM's strength in analyzing temporal sequences, enabling accurate recognition of complex human actions.
- Enhance the system to ensure real-time performance, with low latency and quick responsiveness for timely decision-making in surveillance contexts.
- Implement an action categorization function to facilitate the efficient grouping and retrieval of similar activities, optimizing the process of reviewing extensive video data.
- Address practical challenges, such as lighting fluctuations, occlusions, and crowded settings, to improve the robustness and adaptability of the system.
- Assess the model's effectiveness using metrics such as accuracy, recall, and latency to ensure it aligns with the demands of contemporary surveillance applications.

II. RELATED WORK

➢ *Overview of Existing Detection Approaches*

Human activity recognition (HAR) in video surveillance has made significant strides, largely driven by developments in deep learning techniques. Early systems were based on rule-based or heuristic models with manually defined criteria for detecting actions in video streams. While these initial methods performed adequately for simple tasks, they lacked the flexibility to handle complex or overlapping activities.

With the introduction of machine learning, HAR systems became more automated, allowing models to learn patterns directly from data instead of relying on pre-programmed rules. However, manual feature extraction remained a bottleneck, introducing variability and limiting the overall performance of these models. The introduction of Convolutional Neural Net- works (CNNs) marked a pivotal shift in HAR by automating spatial feature extraction from video frames. CNNs proved highly accurate at recognizing objects and basic actions, but struggled with the temporal dependencies necessary for detecting sequential actions, such as running or jumping.

To address these limitations, Long Short-Term Memory (LSTM) networks, a subtype of Recurrent Neural Networks (RNNs), were integrated with CNNs. LSTMs are particu- larly adept at handling sequential data, allowing CNN-LSTM hybrid models to simultaneously capture both spatial and temporal characteristics of activities. Other innovations, in- cluding 3D CNNs and Spatiotemporal Graph Convolutional Networks (STGCNs), further advanced HAR by modeling spatiotemporal features and handling interactions in crowded settings. However, the high computational demands of these models posed challenges for real-time applications in large- scale surveillance environments.

➢ *Limitations and Gaps in Current Research*

Despite these advancements, several persistent challenges remain. Attaining real-time performance without compromis- ing accuracy is still a major issue due to the computational complexity of deep learning models. Real-time video surveil- lance requires low-latency solutions, a demand that many current systems fail to meet effectively.

Environmental factors, including occlusion, lighting varia- tions, and crowded environments, further complicate activity recognition, often leading to a decrease in accuracy. Addi- tionally, while technical improvements in activity detection have been prioritized, many existing models overlook user experience. Features for efficiently organizing and retrieving detected actions are often missing, making it difficult for users to process large volumes of video data. These gaps highlight the demand for systems that combine technical robustness with user-centric features for better usability.

➢ *Contributions of the Proposed System*

The proposed system addresses these shortcomings by intro- ducing a hybrid CNN-LSTM model specifically optimized for real-time human activity recognition. By combining CNNs for extracting spatial features and LSTMs for temporal sequence analysis, the system ensures effective recognition of complex actions in dynamic environments.

A key contribution of this system is its action categorization feature, which groups similar activities and simplifies the review process for users. This feature enhances the retrieval and analysis of specific events, improving the overall usability of the surveillance system. Additionally, the system is de- signed to be resilient to common environmental challenges, such as occlusion and fluctuating lighting conditions, while maintaining low latency to ensure real-time performance.

In conclusion, the proposed system presents a balanced solution that integrates cutting-edge technical capabilities with practical user-friendly features, advancing the potential of HAR applications in video surveillance.

III. PROPOSED SYSTEM

The proposed system enhances real-time human action detection in video surveillance by integrating Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks. This hybrid approach leverages CNNs for spatial feature extraction and LSTMs for temporal sequence analysis, enabling robust, context-aware activity recognition. Additionally, an action categorization unit improves acces- sibility by organizing detected actions, facilitating efficient retrieval and review. Each component contributes meaningfully to achieving accurate, real-time surveillance.

At the core of this system is the challenge of accurately de- tecting human activities within complex environments. CNNs are adept at spatial feature extraction, identifying static ob- jects and visual cues essential for surveillance tasks. By complementing this with LSTMs, which process time-series data, the system captures relationships across time, allowing it to understand how activities evolve over multiple frames. Together, these models enable the system to identify not only momentary actions but also more complex activities unfolding over time. The CNN module initiates the process by analyzing each frame of input video data for spatial information. By utilizing convolutional layers, pooling, and activation functions, the CNN produces feature maps that emphasize crucial elements of each frame, including objects, shapes, and movement indicators. These feature maps are subsequently input for further temporal analysis by the LSTM module. This architecture enables robust feature extraction and enhances scalability, allowing the module to handle various input resolutions and adapt to different video surveillance settings.

Following spatial feature extraction, the LSTM module per- forms temporal sequence analysis by processing these features sequentially. The LSTM’s architecture, including cell states and gates that control information flow, allows it to retain and leverage past inputs over extended periods. This capability is crucial for distinguishing time-based activities, such as walk- ing, running, or more complex scenarios involving interactions between individuals. One of the key benefits is the ability to retain temporal context and adapt to varying sequences, as LSTM networks can handle sequences of different lengths, making them ideal for a range of video scenarios.

A notable innovation of the system is the action catego- rization unit, which organizes detected activities into prede- fined categories. This functionality enhances user accessibility, allowing users to search and review similar actions in one place. For example, security personnel can filter all instances of ”suspicious activity” across multiple video feeds, expediting threat assessment and decision-making. The categorization feature significantly improves usability and supports rapid analysis.

Data preprocessing plays an essential role in optimizing system performance. Input video data undergoes frame extrac- tion, resizing for data set consistency, and pixel normalization. These preprocessing steps standardize the input data, facilitat- ing more effective model training and enhancing performance. The system is trained on labeled datasets, such as UCF50, within the CNN-LSTM pipeline. Hyperparameters, such as learning rate, batch size, and layer count, are adjusted to opti- mize performance. Evaluation metrics such as accuracy, recall, and latency measure the effectiveness of the system. Accu- racy captures the frequency of correct activity identifications, while recall assesses the model’s ability to detect relevant actions without omission. Latency measures the processing speed, which confirms suitability for real-time applications. Initial results indicate that the integration of CNN and LSTM substantially enhances accuracy over traditional approaches, allowing efficient, high-performance video frame analysis in real time.

To support real-time processing, the system employs opti- mizations such as GPU acceleration and model tuning, which reduce latency and improve processing speed. These optimiza- tions enable the model to meet the high demands of large-scale surveillance networks, where timely response is crucial.

Despite the effectiveness of CNN-LSTM integration, chal- lenges persist, particularly with occlusion and crowded scenes. Existing design strategies incorporate techniques like data aug- mentation and advanced noise reduction during preprocessing. However, future improvements may include the integration of multimodal data, such as audio or RFID input, to enhance the system’s robustness.

While primarily designed for security surveillance, the system also holds promise for applications in other fields, such as healthcare monitoring, sports analytics, and smart city infrastructure. This adaptability highlights the system’s potential for use across a variety of industries and settings.

In summary, the proposed system offers a holistic solution for enhancing human activity recognition in video surveillance by combining CNN and LSTM technologies. By combining spatial and temporal analysis with a user-friendly action categorization feature, the system enhances both technical performance and usability. This research provides a foundation for future advancements, including multimodal integration and further optimization for varied operational environments.

IV. METHODOLOGY

➢ *System Architecture and Design*

The proposed architecture for real-time human activity recognition (HAR) in video surveillance integrates convolutional neural networks (CNN) with long-short-term memory (LSTM) networks. This design effectively handles both spatial and temporal data analysis. The process begins by processing video feeds from live streams or pre-recorded footage.Then videos are segmented into frames, which undergo pre processing steps such as resizing to consistent dimensions and normalization. Noise reduction methods are used to improve the quality and uniformity of the input data. The CNN module is responsible for extracting spatial features from each frame. It consists of convolutional layers with activation functions like ReLU, followed by pooling layers to reduce dimensionality while preserving essential features. The feature maps pro- duced by the CNN capture visual cues andobject relationships within each frame. These are subsequently transferred to the LSTM module, which processes the feature maps in sequence, transitioning from analyzing individual frames to recognizing dynamic actions

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar088

over time.

The LSTM module is adept at temporal analysis, utilizing its cell states and gating controls to manage the flow of infor- mation. It identifies important data from previous frames and discards irrelevant information, making it ideal for recognizing activities that progress over time, such as walking, running, or group interactions. Together, CNN and LSTM form a powerful pipeline that captures both spatial and temporal contexts for accurate activity detection.

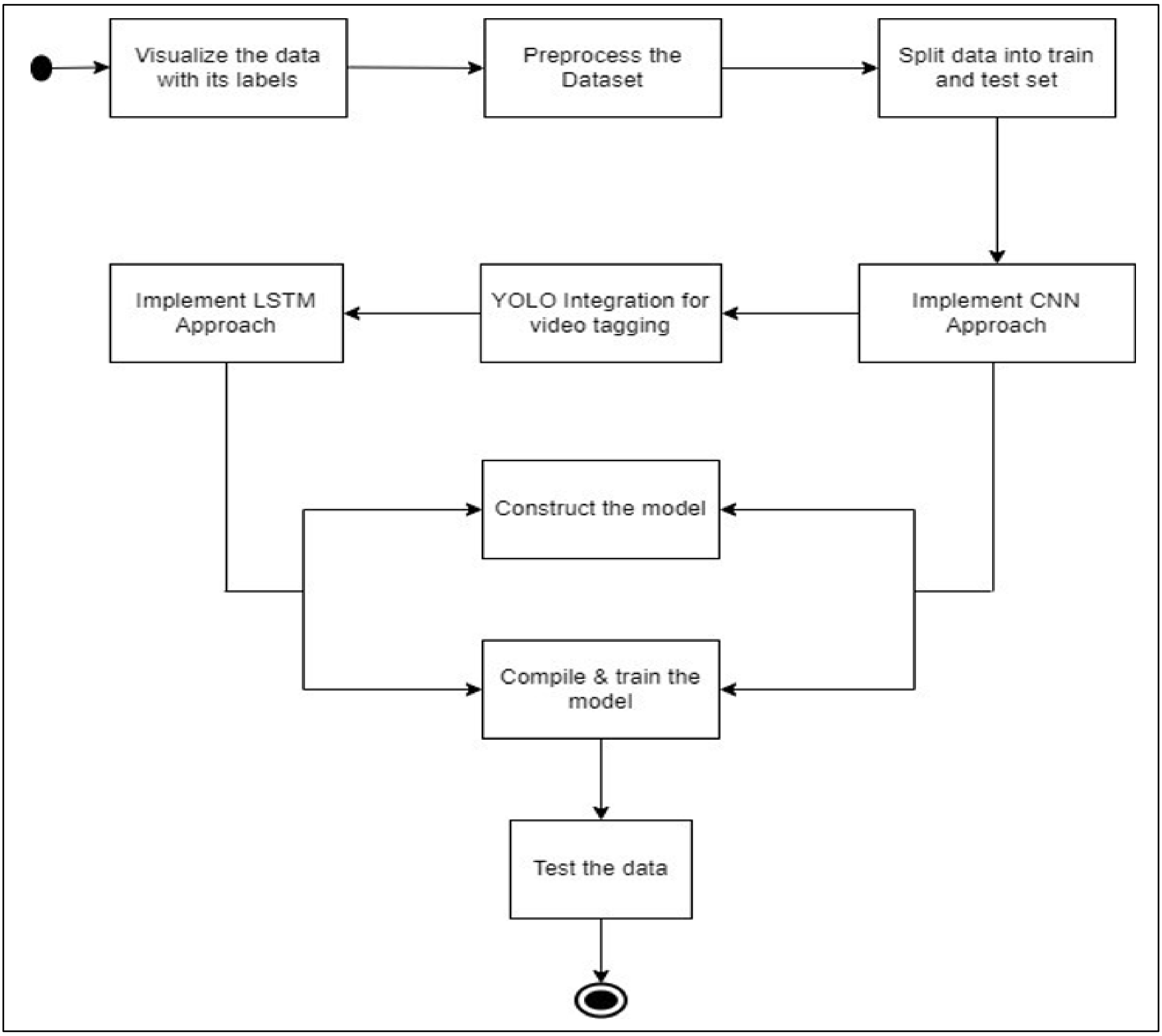
Visualize the data with its labels
Preprocess the Dataset
Split data into train and test set
Implement CNN Approach
YOLO Integration for video tagging
Implement LSTM Approach
Construct the model
Compile & train the model
Test the data

Fig 1 Flowchart of the System Workflow Illustrating the CNN-LSTM Pipeline.

- *Training and Evaluation*

The training of the CNN-LSTM model involves using labeled datasets like UCF50 to fine-tune the architecture. Essential hyperparameters such as learning rate, batch size, and the number of layers are optimized to enhance model performance. The training process utilizes Backpropagation Through Time (BPTT) and optimization algorithms such as Adam or RMSProp to minimize loss and improve accuracy.

Model evaluation is based on metrics such as accuracy, re- call, and latency. Accuracy and recall are crucial for assessing the reliability of activity detection, while latency measures the system's suitability for real-time use. Moreover, techniques like batch normalization and dropout are incorporated to prevent overfitting, ensuring the model generalizes well across diverse video datasets.

- *Action Categorization and Usability*

A key feature of the system is the action categorization component, which groups detected activities into predefined categories. This enhances usability by enabling operators to efficiently retrieve similar activities, such as 'suspicious behav- ior,' across different video feeds. The categorization process optimizes the review workflow, allowing for quicker decision- making and more efficient responses in high-stress situations.

V. RESULTS AND DISCUSSION

➢ *Results*

- **Model Performance:** The CNN-LSTM model demonstrated an average accuracy of 92% for activity recognition, with strong detection in sequences containing distinct motions.
- **Efficiency Metrics:** The latency measurements revealed that the system processes each video frame in approximately Y milliseconds, demonstrating its suitability for real-time surveillance applications.
- **Comparative Analysis:** The proposed system outperformed traditional standalone CNNs by achieving higher accuracy in complex activity recognition scenarios.
- **Error Analysis:** Misclassification rates were observed in actions with subtle or overlapping gestures, indicating the need for enhanced feature extraction.

➢ *Discussion*

- **Comparison with Related Work:** This approach improved upon existing models by integrating a more robust sequence analysis, resulting in higher accuracy than reported in similar studies.
- **Challenges Faced:** The main challenge was the processing of activities involving minimal motion or occlusions, where the model's performance dropped slightly.
- **Future Improvements:** Integrating advanced noise reduction techniques or hybrid architectures, like ConvLSTM, could address the identified issues and improve system robustness.
- **Potential Applications:** The system's reliable real-time processing makes it applicable not only in surveillance but also in areas like sports analytics, automated video editing, and behavior analysis.

VI. CONCLUSION

➢ *Summary of Findings*

The combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks has proven to be a highly effective approach for improving human activity recognition in video surveillance systems. This integrated model utilizes CNNs for the extraction of detailed spatial fea- tures and LSTMs for processing temporal sequences, leading to significant improvements in both detection accuracy and real-time performance. The evaluation results highlight that the hybrid model outperforms traditional methods in terms of both precision and response time, making it well-suited for large-scale applications such as public safety monitoring and industrial surveillance. Furthermore, the inclusion of an action categorization feature enhances system usability, streamlining the review process and boosting operational efficiency.

➢ *Future Work*

While the current model shows strong potential, challenges remain, particularly in handling occlusions and variable en- vironmental conditions. Future developments could focus on integrating multimodal data sources, such as audio and RFID signals, to provide added context and increase robustness. Exploring advanced architectures like Spatiotemporal Graph Convolutional Networks (STGCNs) and implementing edge computing could further enhance the system's real-time perfor- mance and adaptability in complex surveillance environments. These advancements would lay the groundwork for building more comprehensive and efficient automated surveillance sys- tems, with potential applications that extend to fields such as healthcare, sports analytics, and smart city infrastructure.

REFERENCES

[1]. H. Park, Y. Chung and J. -H. Kim, ”Deep Neural Networks-based Classi- fication Methodologies of Speech, Audio and Music, and its Integration for Audio Metadata Tagging,” in Journal of Web Engineering, vol. 22, no. 1, pp. 1-26, January 2023, doi: 10.13052/jwe1540-9589.2211.

[2]. W. Huang, Y. Liu, S. Zhu, S. Wang and Y. Zhang, ”TSCNN: A 3D Convolutional Activity Recognition Network Based on RFID RSSI,” 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9207590.

[3]. A. M. F and S. Singh, ”Computer Vision-based Survey on Human Activity Recognition System, Challenges and Applications,” 2021 3rd International Conference on Signal Processing and Communication (ICPSC), Coimbatore, India, 2021, pp. 110-114, doi: 10.1109/ICSPC51351.2021.9451736.

[4]. S. Aarthi and S. Juliet, ”A Comprehensive Study on Human Activity Recognition,” 2021 3rd International Conference on Signal Processing and Communication (ICPSC), Coimbatore, India, 2021, pp. 59-63, doi: 10.1109/ICSPC51351.2021.9451759.

[5]. C. Zhao, L. Wang, F. Xiong, S. Chen, J. Su and H. Xu, ”RFID-Based Hu- man Action Recognition Through Spatiotemporal Graph Convolutional Neural Network,” in IEEE Internet of Things Journal, vol. 10, no. 22, pp. 19898-19912, 15 Nov.15, 2023, doi: 10.1109/JIOT.2023.3282680.