Journal of Current Research in Engineering and Science
JCRES
Bi-Annual Online Journal (ISSN: 2581 - 611X)

VOLUME:8, ISSUE:1, FEBRUARY 2025, PAPER NUMBER:1

Human Activity Detection for Surveillance

[1] Professor, [2,3,4,5] Final Year UG Students,

M.Subramanyam
Electronics and communication
P E S College of Engineering
Mandya, India
msubramanyam71@pesce.ac.in

Manoj M A
Electronics and communication
P E S College of Engineering
Mandya, India
manojmaarya@gmail.com

Rakesh M A
Electronics and communication
P E S College of Engineering
Mandya, India
rakeshma212@gmail.com

Madan Arya M A
Electronics and communication
P E S College of Engineering
Mandya, India
madanarya1720@gmail.com

Manish H D
Electronics and communicationP E S
College of Engineering
Mandya, India
hdmanish75@gmail.com

Human Activity Detection for Surveillance

Abstract - Surveillance systems are integral for public safety, with human activity detection serving as a pivotal component for threat identification and prevention. This paper proposes a novel approach integrating hybrid deep learning models, specifically a combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, for real-time human activity detection in surveillance footage. Leveraging Google Collab for model development and Raspberry Pi for hardware integration, the system utilizes dual cameras with 180-degree coverage to capture comprehensive visual data. Through this approach, the system aims to enhance surveillance and security measures by accurately detecting suspicious activities, thus contributing to the advancement of proactive threat mitigation strategies. The methodology encompasses model design, deployment, and performance analysis, demonstrating promising results in anomaly detection and severity assessment. This research underscores the significance of advanced machine learning techniques in bolstering the efficacy of surveillance systems, with practical implications across various domains including security, smart homes, and industrial monitoring.

Keywords—: Surveillance Systems, Machine Learning Model, Video data.

I. INTRODUCTION

Surveillance systems are integral to maintaining public safety, with human activity detection serving as a fundamental component for identifying and mitigating potential threats in real-time. Leveraging the advancements in deep learning, the analysis of surveillance footage has become increasingly automated and efficient. Traditional methods of monitoring public spaces have evolved significantly, particularly in the realm of Human Activity Detection for Surveillance (HAR).

This technology has garnered significant traction in recent years, spurred by the widespread adoption of wearable devices and a growing interest in context-aware applications. Its versatility extends beyond security applications, finding utility in health monitoring, sports performance analysis, and the creation of adaptive environments responsive to human behaviour. As such, the quest for enhancing surveillance efficacy through intelligent activity detection stands at the forefront of technological innovation, promising a safer and more responsive environment for all.

Surveillance systems stand as guardians of public safety, their efficacy reliant on the ability to swiftly detect and respond to potential threats. Central to this endeavor is the discipline of Human Activity Detection for Surveillance (HAR), a dynamic field propelled by advanced algorithms and machine learning. Traditionally, monitoring public spaces demanded extensive human oversight, but with the advent of deep learning, we've witnessed a transformative shift towards automated analysis of surveillance footage.

The structure of the remainder of the paper is organized as follows: Section 2 delves into the related work of Human activity detection. Section 3 explicates the methodology utilized in this study. Section 4 describes the machine learning models employed. Section 5 provides an explanation of the dataset used. Sections 6 and 7 respectively discuss the proposed architecture and present theresults and discussions. Lastly, Section 8 comprises the references

II. RELATED WORK

Many studies on human activity recognition have been conducted in the past few years. The existing research proposes various methods for identifying human behaviours captured in videos. Recognizing activities through cameras offers distinct advantages due to their ease of setup and accessibility. [1] focuses on providing assistance to elderly people by monitoring their activities in different indoor and outdoor environments using gyroscope and accelerometer data collected from a smart phone. [2] used nine MH algorithms as FS methods to boost the classification accuracy of the HAR and fall detection applications. They employed Residual Recurrent Neural Networks (Res RNN) for feature extraction, and evaluated their performance using metrics such as accuracy, precision, recall, and F1 score. [3] targets human detection in aerial video sequences captured by a moving camera mounted on an aerial platform. It addresses challenges like altitude variations, lighting changes, camera instability, and variations in viewpoints, object sizes, and colors. With low obtrusiveness, it utilizes the UCF-ARG dataset achieving an 80% accuracy rate.[4] system combines deep learning with a self-attention model and a wearable sensor-based human activity recognition framework. It exclusively utilizes a three-axis accelerometer, gyroscope, and linear acceleration for reliable performance, achieving a 90 percent accuracy rate and outperforming other sensors such as GPS or pressure sensors.[5] This paper mathematically characterizes hazardous situations using sensor data from mobile phones and context-aware technology. It explores accident prevention methods during mobile phone use.[6] discusses the challenges in human activity recognition and proposes a dynamic active learning-based method to address these challenges The method aims to reduce annotation costs and improve activity recognition performance by dynamically discovering new activities and patterns.

[7] suggests an efficient algorithm for solving this problem using an image descriptor capturing movement information and a classification approach. The novel abnormality indicator is generated from a hidden Markov model, which learns optical flow orientation histograms from the video frames.[8] proposed system utilizes CCTV footage to monitor human behavior on a campus, issuing gentle warnings when suspicious events occur. Key components include event detection and human behavior recognition, a challenging task. Various campus areas are under surveillance, with video footage serving as test data. The training process involves data preparation, model training, and inference, employing CNN and RNN neural networks. CNN extracts high-level features from images, while RNN handles classification, suitable for video processing. The system employs a pre-trained VGG-16 modelto predict behaviour and aid monitoring.[9] makes use of CCTV and webcams provide real-time video streaming, with webcams being cost-effective but potentially less secure. The system detects unauthorized persons using an AMD algorithm, tracks them upon user identification, and enhances moving object detection through background subtraction. AMD ensures thorough detection of moving objects, while a monitoring room camera generates alerts forsuspicious activity.[10] introduces a novel deep neural networkarchitecture , merging convolutional layers with Long Short-Term Memory (LSTM) units. This model efficiently extracts and classifiesactivity features with minimal parameters. LSTM, a variant ofrecurrent neural networks (RNNs), is adept at handling temporalsequences, making it ideal for processing raw data from mobilesensors. The architecture comprises two LSTM layers followed byconvolutional

layers, with a Global Average Pooling (GAP) layerreplacing the fully connected layer to reduce parameters.

Additionally, a Batch Normalization (BN) layer after GAP enhances convergence speed, yielding significant improvements. Evaluationon three public datasets (UCI, WISDM, and OPPORTUNITY) demonstrated outstanding performance: 95.78% accuracy on UCI-HAR, 95.85% on WISDM, and 92.63% on OPPORTUNITY. These results highlight the model's robustness and superior activity detection capability, surpassing previous findings while maintaining adaptability, parameter efficiency, and high accuracy.

III. METHODOLOGY

In fig.1 The human activity detection in surveillance, the process begins with the input of video data captured by cameras. This raw video data serves as the primary source of information for the subsequent analysis. The video data is then fed into a convolutional neural network (CNN), which specializes in extracting visual features from images or frames. The CNN processes each frame of the video, extracting relevant visual patterns, motion cues, and spatial relationships that characterize different human activities. he output of the CNN, which consists of the extracted visual features, is then passed on to a long short-term memory (LSTM) network. Unlike traditional neural networks, LSTM networks are capable of capturing temporal dependencies and sequential patterns in data. In this context, the LSTM analyses the sequence of visual features over time, learning the dynamic evolution of human activities within the video footage. Once the CNN and LSTM models have processed the video data and extracted relevant features, the combined model is trained using labelled data. During the training phase, the model learns to associate specific visual patterns and temporal sequences with corresponding human activities. Through iterative adjustments of model parameters, such as weights and biases, the model improves its ability to accurately classify different activities based on the learned features. After the training phase, the model is ready for inference. New, unseen video data can be input into the trained model, which then applies the learned classification rules to predict the human activities present in the footage. The model's output provides classifications or labels for each segment of the video, indicating the detected activities such as walking, running, standing, or interacting with objects.

Unlike traditional neural networks, LSTM networks are capable of capturing temporal dependencies and sequential patterns in data. In this context, the LSTM z the sequence of visual features over time, learning the dynamic evolution of human activities within the video footage. Once the CNN and LSTM models have processed the video data and extracted relevant features, the combined model is trained using labelled data. During the training phase, the model learns to associate specific visual patterns and temporal sequences with corresponding human activities. Through iterative adjustments of model parameters, such as weights and biases, the model improves its ability to accurately classify different activities based on the learned features. After the training phase, the model is ready for inference. New, unseen video data can be input into the trained model, which then applies the learned classification rules to predict the human activities present in the footage. The model's output provides classifications or labels for each segment of the video, indicating the detected activities such as walking, running, standing, or interacting with objects.

In summary, the block diagram illustrates a pipeline for human activity detection in surveillance, leveraging a combination of CNN and LSTM models. By processing video data through these specialized architectures, the model can effectively capture both spatial and temporal characteristics of human activities, enabling accurate and robust detection and classification in real-world surveillance scenarios. In summary, the block diagram illustrates a pipeline for human activity detection in surveillance, leveraging a combination of CNN and LSTM models.

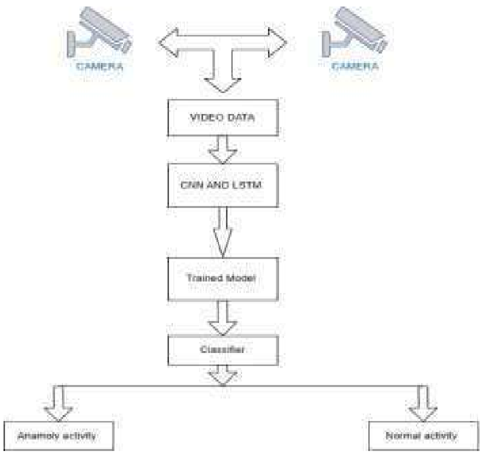

Figure 1 Proposed Methodology for Activity detection

By processing video data through these specialized architectures, the model can effectively capture both spatial and temporal characteristics of human activities, enabling accurate and robust detection and classification in real-world surveillance scenarios.

IV. MACHINE LEARNING MODELS

a) Convolution Neural Network (CNN)

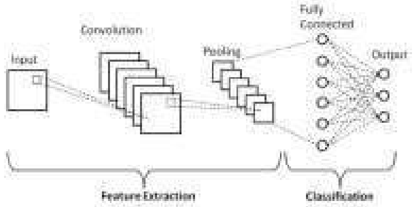

Figure 2 Convolution Neural Network

In human activity detection for surveillance, CNN (Convolutional Neural Network) models are utilized for their efficacy in analyzing visual data. Fig.2 shows a CNN model for this purpose typically consists of three main steps: input, feature extraction, and classification. At the input stage, raw image data or video frames are fed into the CNN model. These images serve as the primary source of information for the network to analyze.

The feature extraction phase, often referred to as the convolution and pooling layers, involves the application of filters to the input images. Convolutional layers detect various features such as edges, textures, and patterns within the input data. Pooling layers then down sample the feature maps produced by convolution, reducing their dimensionality while retaining essential information. Together, these layers effectively extract meaningful features from the input images, capturing relevant visual patterns that are crucial for human activity detection. Following feature extraction, the fully connected layers come into play for classification. These layers take the high-level features extracted by the convolution and pooling layers and map them to specific classes or categories of human activities. Through the process of training, the CNN learns to distinguish between different activities based on the extracted features, enabling accurate classification of observed behaviors in surveillance footage.

In summary, the block diagram of a CNN model for human activity detection comprises input, feature extraction (consisting of convolution and pooling layers), and classification (implemented through fully connected layers). This architecture effectively leverages the power of machine learning to analyze visual data and identify human activities in surveillance scenarios.

b) Long Short Term Memory(LSTM)

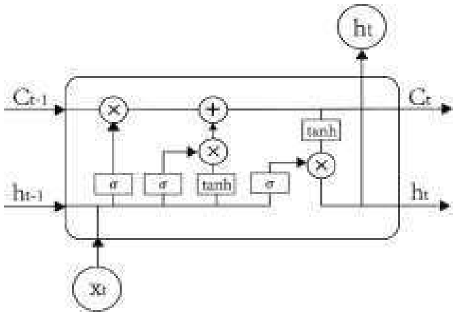

Figure 3 Single Long short term memory cell

In the LSTM architecture, each LSTM cell incorporates three crucial gates: forget, input, and output gates. These gates regulate the flow of information within the cell state, which serves as a memory unit, enabling the LSTM to store and propagate information across time steps. Figure 3 illustrates the interconnections between these gates and the q cell state. The forget gate, which receives input from both the current time step Xt and the previous output $ht-1$, utilizes a sigmoid activation function to determine which information to retain or discard. If the output of the sigmoid function is 1, the corresponding information is preserved, while an output of 0 signifies complete removal. Equation (1) demonstrates the computation of the forget gate. Subsequently, the input gate decides what new information from the current input ($Xt,ht-1$) should be added to the cell state. This information is combined with a new candidate vector $C\sim t$, generated by the tanh activation function, to produce an updated cell state Ct. Equations (2)-(4) detail the calculations involved in determining the input gate, new candidate values, and cell state, respectively. Following this, the output gate is determined based on filtered information, employing two different activation functions, and specifies the next hidden state. The previous hidden state $ht-1$ and the current input xt xt are first passed through a sigmoid activation function, while the updated cell state is fed into a tanh activation function. The outputs of these functions are then multiplied to generate the next hidden state. Equations (5)and h, outline the computation of the output gate and hidden state, respectively. In essence, the LSTM cell operates as a memory unit, selectively erasing, reading, and writing information based on the decisions made by the forget, input, and output gates. This mechanism enables LSTMs to effectively capture long-term dependencies and relationships in sequential data.

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \qquad f_t \text{ represents forget gate} \tag{1}$$

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \qquad i_t \text{ represents input gate} \tag{2}$$

$$\tilde{C}_t = \tanh(W_C.[h_{t-1}, x_t] + b_c) \qquad \tilde{C}_t \text{ represents candidate values} \tag{3}$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \qquad C_t \text{ represents Cell state} \tag{4}$$

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \qquad o_t \text{ represents output gate} \tag{5}$$

$$h_t = o_t \times \tanh C_t$$

h_t represents hidden state

where x is the input data, σ is the sigmoid activation function, tanh is the hyperbolic tangent activation function, W is the weight matrix.

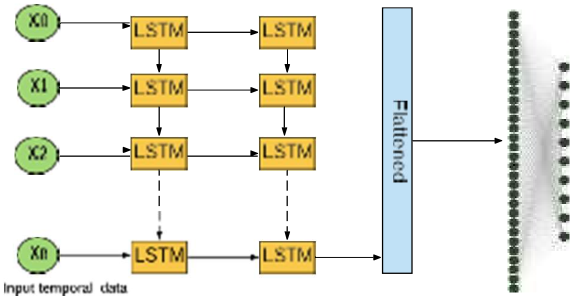

Figure 4 Architecture of the LSTM model for human activity recognition.

V. DATASET DESCRIPTION

The UCF101 dataset has emerged as a fundamental resource for researchers delving into human activity detection, particularly within surveillance contexts. Leveraging Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) models, this dataset serves as a cornerstone for training and evaluating such algorithms. CNNs excel at extracting spatial features from images or frames, making them adept at recognizing patterns within video sequences. Meanwhile, LSTM models are proficient at capturing temporal dependencies, crucial for understanding how activities unfold over time. By combining these architectures, researchers can develop robust systems capable of not only identifying individual actions but also comprehending their sequential nature. The UCF101 dataset's diverse range of activities, spanning from everyday actions to sports, ensures that models trained on it gain a comprehensive understanding of human behaviour, thereby enhancing the efficacy of surveillance systems in monitoring and analysing complex scenarios.

VI. PROPOSED ARCHITECTURE

In the methodology of human activity detection for surveillance using CNN + LSTM models, the approach integrates the strengths of both convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to effectively analyze video data from surveillance cameras. Initially, raw video data captured from two cameras is fed into the CNN model. The CNN processes this visual data, extracting meaningful visual features that represent different aspects of human activities such as motion patterns, spatial relationships, and object interactions.

These extracted visual features serve as a rich representation of the input video frames. Subsequently, the output of the CNN, consisting of visual features, is passed on to the LSTM network. The LSTM, known for its capability in capturing temporal dependencies and sequential patterns, performs sequence learning on the visual features obtained from the CNN. By analysing the temporal evolution of visual features over time, the LSTM model can discern complex activity patterns and dynamics inherent in the surveillance footage. As the LSTM processes the sequential data, it learns to recognize patterns indicative of various human activities, such as walking, running, gesturing, or interacting with objects. The LSTM's ability to retain information over long sequences enables it to capture the temporal context crucial for accurate activity detection. Finally, the output of the LSTM model represents the detected human activities based on the learned patterns and sequences from the input video data. These detected activities can include a wide range of behaviors and actions observed in the surveillance footage, providing valuable insights for security monitoring, anomaly detection, or behavioural analysis purposes.

We employ widely known performance evaluation criteria, namely, accuracy, precision, and recall, to measure the recognition performance of the proposed system. Accuracy measures the total percentage of the accurate recognition rate of the classifier.

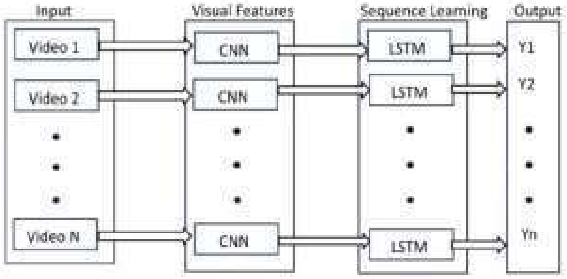

Figure 5 CNN and LSTM Model

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \times 100$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100$$

Here, TP classifies the positive class as positive, FP classifies the positive class as negative, TN classifies the negative class as negative, and FN classifies the negative class as positive.

Overall, the combined CNN + LSTM approach leverages the complementary strengths of both architectures, enabling robust and accurate human activity detection in surveillance scenarios by effectively capturing both spatial and temporal dynamics inherent in video data.

VII. RESULTS AND DISCUSSION

Fig.6 illustrates the performance results of the CNN-LSTM model for Human Activity Recognition (HAR) using the UCF101 Video dataset. The model demonstrates superior classification performance, achieving an average accuracy of 99.87% and a precision of 99.82%. The recall for all human activities is 99.81%. Analysis of the table reveals that across ten folds, the 7th fold attains a maximum accuracy of 100%, while the remaining folds achieve over 99% accuracy consistently. This consistency underscores the Exceptional classification performance of the CNN-LSTM model for HAR tasks.

Metrices	1st	2nd	3rd	Avg
Accuracy	99.86	99.86	99.93	99.87
Precision	99.66	99.81	99.99	99.82
Recall	99.79	99.78	99.88	99.81

Figure 6 Performance result of the proposed CNN-LSTM Model on the UCF dataset

The confusion matrix of the CNN-LSTM model on the UCF101 dataset, depicted in Figure 7, showcases the individual recognition accuracy of human activities. Notably, activities such as Run, Sit down, Standup, and Walk achieve 100% accuracy, indicating the model's capability to capture both spatial and temporal features from video data. However, misclassifications occur between Lie down and Fall activities. This misclassification may be attributed to the similarity in patterns between sudden falls, characteristic of Fall activities, and the steady posture of Lie down activities.

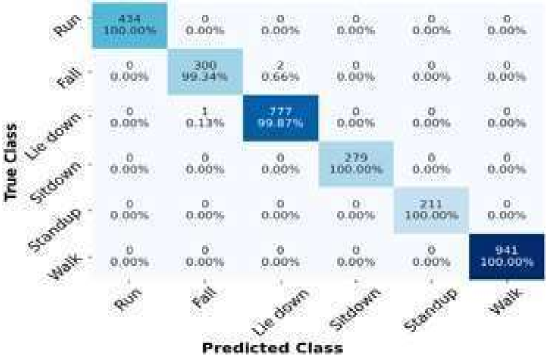

Figure 7 Confusion Matrix of the Proposed UCF101 Dataset

VIII. CONCLUSION

In conclusion, our CNN-LSTM approach for human activity recognition, utilizing the UCF101-Human Activity Recognition dataset, has demonstrated promising results in the domain of surveillance. By effectively leveraging both CNN and LSTM models, we were able to robustly extract spatial and temporal features from video data, respectively. The ConvLSTM and LRCN architectures exhibited superior performance compared to other deep learning approaches, achieving notable accuracies of 80% and 92%, respectively. Notably, the LRCN model showcased higher accuracy while requiring less training time, making it a compelling choice for real-time surveillance applications. Our evaluation metrics, including total loss, validation loss, total accuracy, and validation accuracy, underscored the effectiveness of our proposed approach in accurately detecting human activities in surveillance videos.

For future work, we aim to enhance the model's capabilities to recognize actions involving multiple individuals performing different activities simultaneously within the frame. This will require annotated datasets containing information about each person's activity along with their bounding box coordinates. Alternatively, we could explore the feasibility of performing activity recognition on each individual separately, albeit at the cost of increased computational complexity. By addressing these challenges, we can further improve the robustness and versatility of our CNN-LSTM model for human activity detectionin surveillance scenarios, thereby advancing the field ofintelligent video surveillance.

IX. REFERENCES

[1] Hayat, Ahatsham, et al. "Human activity recognition for elderly people using machine and deep learning approaches." Information 13.6 (2022): 275.

[2] Al-Qaness, Mohammed AA, et al. "The applications ofmetaheuristics for human activity recognition and fall detection using wearable sensors: A comprehensive analysis." Biosensors 12.10 (2022): 821.

[3] Aldahoul, Nouar, et al. "A comparison between various human detectors and CNN-based feature extractors for human activity recognition via aerial captured video sequences." IEEE Access 10 (2022): 63532-63553.

[4] Khatun, Mst Alema, et al. "Deep CNN-LSTM with self-attention model for human activity recognition using wearable sensor." IEEE Journal of Translational Engineering in Health and Medicine 10 (2022): 1-16.

[5] Sun, Bowen, et al. "Context awareness-based accident prevention during mobile phone use." IEEE Access 8 (2020): 27232-27246.

[6] Bi, Haixia, et al. "Human activity recognition based ondynamic active learning." IEEE Journal of Biomedical andHealth Informatics 25.4 (2020): 922-934.

[7] Wang, Tian, et al. "Abnormal event detection based onanalysis of movement information of video sequence." Optik 152 (2018): 50-60.

[8] Amrutha, C. V., C. Jyotsna, and J. Amudha. "Deep learning approach for suspicious activity detection from surveillance video." 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE, 2020.

[9] Divya, P. Bhagya, et al. "Inspection of suspicious human activity in crowd sourced areas captured in surveillance cameras." International Research Journal of Engineering and Technology (IRJET) 4.12 (2017).

[10] Xia, Kun, Jianguang Huang, and Hanyu Wang. "LSTM-CNN architecture for human activity recognition."IEEE Access 8 (2020): 56855-56866.

[11] Subramanyam, M., and S. S. Parthasarathy. "Dynamic User Activity Prediction using Contextual Service Matching Mechanism." International Journal of Advanced Computer Science and Applications 13.3 (2022).

[12] M Subramanyam, SS Parthasarathy“Ambient intelligent framework for modelling critical medical events based on context awareness”International Journal of Electrical and Computer Engineering (IJECE) 14 No 3(2024) 3106-3115