SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V11I9P104

Volume 11 Issue 9, 36-46, September 2024
© 2024 Seventh Sense Research Group®

Original Article

Enhanced Real-Time Surveillance and Suspect Identification Using CNN-LSTM Based Body Language Analysis

M. Archana[1], S. Kavitha[2], A.Vani Vathsala[3]

[1*]*Department of CSE, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India.*
[2]*Department of CTECH, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India.*
[3]*Department of CSE, CVR College of Engineering, Hyderabad, Telangana, India.*

[1]*Corresponding Author : am3323@srmist.edu.in*

Received: 24 June 2024 Revised: 05 August 2024 Accepted: 28 August 2024 Published: 30 September 2024

Abstract *- The exponential rise in criminal activities necessitates advanced methods for suspect identification and surveillance. This research aims to tackle this issue through the development of a sophisticated video analytics system leveraging computer vision and deep learning. The primary objective is to accurately identify suspects based on body language patterns extracted from video inputs. We propose a CNN-LSTM based Body Language Rule System (BLRS) that integrates Convolutional Neural Networks (CNNs) for spatial feature extraction and Long Short-Term Memory (LSTM) networks for temporal sequence learning. The system processes video frames to identify key body language indicators such as gestures, postures, and facial expressions. Extensive evaluations using the UCF-Crime Dataset demonstrate the model's high accuracy, with a precision of 95.5%, recall of 95.7%, and overall accuracy of 95.3%. The results indicate that the BLRS significantly outperforms traditional human action recognition models, providing robust and reliable identification of suspicious activities. This research concludes that integrating CNN and LSTM networks within a unified framework enhances real-time surveillance capabilities. The proposed system holds substantial potential for improving public safety and security by enabling more effective monitoring and identification of suspects through advanced body language analysis.*

Keywords - *Computer vision, Video analytics, Neural networks, Data analytics, Deep learning.*

1. Introduction

The rapid advancement of technology has brought about significant changes in how we approach security and surveillance. Traditional methods of monitoring human behaviour and activities have evolved from manual observation to highly sophisticated automated systems [1]. This evolution is particularly evident in the fields of computer vision and artificial intelligence, where the development of advanced algorithms has enabled more accurate and efficient analysis of visual data. Initially, surveillance systems relied heavily on manual monitoring, where human operators observed video feeds and identified suspicious activities[2]. This approach, while effective to some extent, was limited by human fatigue, subjective judgment, and the sheer volume of data that needed to be processed. As a result, semi-automated systems were introduced, which incorporated basic motion detection and alerting mechanisms to assist human operators. However, these systems still required significant human intervention and were prone to high false alarm rates[3].

The next level of development was the fully automated surveillance systems that were introduced into the market. These systems use image recognition technologies and artificial intelligence to process video feeds in real-time without the need for people's intervention. These systems are capable of analyzing a large number of records within a short time and coming up with consistent results; this is important in areas such as public safety, border control and other critical infrastructure protection. Computer vision is a branch of artificial intelligence that is concerned with giving computers the ability to use sight. It comprises a number of objectives, including object detection, image categorization, and scene generation. Deep learning, especially Convolutional Neural Networks (CNNs)[4] and Long Short Term Memory (LSTM) networks[5], has changed the face of computer vision and has given the field powerful tools for feature learning and classification.

CNNs are intended to learn spatial pyramids of features from input images automatically and flexibly. Some of the areas where they have displayed great competence include image recognition, which allows them to correctly categorize objects in an image. LSTMs, a form of Recurrent Neural Networks[6], are very effective at modeling the temporal characteristics of a sequence and, as such, are well suited for applications where there is a sequence of data, such as video analysis. Human Action Recognition (HAR)[7] is one of the important fields in computer vision that deals with the recognition of human activities from videos. The uses of HAR are in video monitoring, interaction between humans and computers, sports, and health checks. However, the following is a challenge that HAR systems are likely to encounter, at least to some extent: changes in the environment, occlusion, changes in the camera angle, and, lastly, the natural variability in human actions.

Body language analysis, a subset of HAR, specifically targets the non-verbal cues and movements of individuals to infer their intentions and actions. It involves studying gestures, postures, and facial expressions to understand a person's emotional state, intentions, and potential threat levels. Accurate body language analysis can significantly enhance the effectiveness of surveillance systems by providing additional context that is not captured through traditional motion detection methods. While deep learning models have greatly improved the accuracy of HAR, several challenges remain. One major challenge is the need for large, labeled datasets to train these models effectively. Collecting and annotating such datasets is both time-consuming and resource-intensive. Additionally, existing models may struggle with real-world scenarios where lighting conditions, background complexity, and subject occlusions vary significantly.

Another critical gap is the integration of multiple modalities, such as combining visual data with other sensory inputs like audio or thermal imaging, to enhance the robustness of HAR systems. Current research also highlights the need for models that can generalize well across different environments and populations, reducing biases that may arise from training on limited datasets. In security applications, the ability to accurately analyze body language can be a game-changer.

Traditional surveillance systems often rely on facial recognition or simple motion detection, which may not provide sufficient information to identify suspicious activities. Body language analysis can add a deeper layer of understanding, allowing systems to detect subtle cues that indicate potential threats. For example, unusual postures, nervous movements, or aggressive gestures can be early indicators of criminal intent.

Recent advancements in CNN and LSTM architectures have paved the way for more sophisticated body language analysis systems. CNNs can be used to extract spatial features from video frames, capturing details about body poses and movements. LSTMs can then analyze these features over time, identifying patterns that are indicative of specific actions or behaviors. The combination of CNNs and LSTMs in a unified framework, such as the proposed CNN-LSTM based Body Language Rule System (BLRS), offers a powerful approach to real-time suspect identification.

1.1. Key Contributions

This research paper offers significant advancements in automated surveillance and human action recognition through the analysis of body language using cutting-edge deep learning techniques. The primary contributions are as follows:

1. Development of CNN-LSTM Based Body Language Rule System (BLRS): Introduction of a novel BLRS that combines CNNs for spatial feature extraction and LSTMs for temporal sequence learning, enabling the analysis and interpretation of body language from video inputs. This system identifies suspects based on body activity expressions, facial cues, and body pose estimations, providing a comprehensive tool for real-time surveillance.
2. Enhanced Data Pre-processing Method: Presentation of an innovative data pre-processing technique that converts video inputs into high-quality image sequences, optimizing the input for deep learning analysis. This method includes novel approaches for handling varying window sizes and the application of a skip-gram model to improve feature extraction.
3. Comprehensive Performance Evaluation: Extensive validation using multiple datasets, including the UCF-Crime dataset, demonstrating superior performance in identifying suspicious activities compared to traditional methods. Performance metrics such as precision, recall, F1-score, and accuracy highlight the system's effectiveness.
4. Implications for Future Research and Applications: The proposed BLRS framework can be integrated into existing security infrastructure, offering a scalable and efficient solution for enhancing public safety. The study lays the groundwork for future research on multimodal data integration and advanced deep learning applications in diverse surveillance contexts, paving the way for more robust and reliable automated surveillance systems.

In summary, this research significantly advances automated surveillance technology by introducing a robust system for body language analysis, enhancing real-time suspect identification, and improving overall security measures in various applications.

2. Literature Review

HAR has been an active research field in the computer vision and AI domain and has witnessed substantial growth in methods and techniques in the last decade. The use of HAR, which is one of the active and popular fields in computer vision as well as artificial intelligence, has been evolving in terms of the techniques and their application in recent years. Starting in 2018, the research has been mainly directed towards fine-tuning the HAR systems with the help of Deep Learning techniques.

Among the recent works, [8] proposed an effective activity recognition method employing a lightweight CNN and DS-GRU network for surveillance videos. It established that the proposed method produced notably higher accurate results owing to deep feature extraction and sequence learning. Also, [9] presented a transferable two-stream convolutional neural network for HAR to capture both spatial and temporal features and use transfer learning for feature extraction. These methodologies reveal that, in recent years, there has been a shift in the field to enhance the performance of deep learning models in overcoming challenges that come with the use of traditional feature-based approaches.

Despite these advancements, several gaps remain in the current literature. One of the primary challenges is the need for large, annotated datasets to train deep learning models effectively. Many existing studies rely on benchmark datasets such as UCF101 and Kinetics-400[10], which may not fully represent real-world scenarios with varying environmental conditions and occlusions. Furthermore, there is a need for more comprehensive evaluations that consider different lighting conditions, camera angles, and background complexities.

Recent advancements have significantly addressed some of the limitations identified in earlier studies. The integration of multimodal data sources, such as combining visual data with audio or thermal imaging, has been a key breakthrough. For instance, [11] developed a framework that uses depth-video sequences and weighted fusion of 2D and 3D autocorrelation gradient characteristics to improve classifier performance. This approach effectively enhances the robustness of HAR systems in varying environmental conditions.

Another significant breakthrough is the development of hybrid models that combine different deep learning architectures. [12] introduced a hybrid model that integrates CNNs with Grey Wolf Optimization (GWO) for action recognition, resulting in higher classification accuracy and reduced error rates. Additionally, the author proposed long-term temporal convolutions for action recognition, which significantly improved the ability to capture long-range dependencies in video sequences.

These advancements highlight the potential of deep learning to revolutionize HAR by providing more accurate and reliable systems. However, the complexity of these models often results in increased computational requirements, which can be a limitation for real-time applications.

The areas of practical usage of HAR research are numerous and cover many fields, including security, healthcare, sports, and human-computer interaction. In the area of security, HAR systems are employed in real-time monitoring and identification of threats, thus boosting situational analysis and control. Nasir, Mahmood, A. S. M, & Shafique, K. (2020) applied HAR for pedestrian identification in real-time visual surveillance, and this is evidence of how HAR can be used to enhance public safety.

In healthcare, HAR systems can monitor patient activities and detect abnormal behaviors, which is crucial for elderly care and rehabilitation. The study by [13] on action recognition using depth-video sequences underscores the importance of accurate activity monitoring in clinical settings. Furthermore, in sports analysis, HAR systems can provide detailed insights into athletes' performance, enabling coaches to devise better training strategies.

The implications of these applications are significant, as they can lead to improved safety, better healthcare outcomes, and enhanced athletic performance. However, the deployment of HAR systems in real-world settings also raises ethical and privacy concerns, which need to be addressed through appropriate regulatory frameworks.

Despite the progress made, there are ongoing debates and controversies in the field of HAR. One major debate revolves around the ethical implications of surveillance technologies. While HAR systems can enhance security, they also raise concerns about privacy and the potential for misuse. Researchers like [14] argue for the need to balance security benefits with privacy protections, advocating for transparent and accountable use of surveillance technologies.

Another problem is related to the bias of the HAR models. As many deep learning models are trained using datasets that may not include all persons of color, the outcomes are inherently bigoted. For example, [15] pointed out the prejudice of race and gender in action recognition systems and urged for larger and more diverse datasets.

There are also differing perspectives on the best methodologies for HAR. While some researchers advocate for the use of deep learning models due to their high accuracy, others point out the challenges related to computational complexity and the need for large, annotated datasets. This debate underscores the importance of continued research and innovation to develop more efficient and scalable HAR systems.

3. Development of CNN-LSTM Based Body Language Rule System (BLRS)

In this section, we delineate the methodology employed in developing the CNN-LSTM based Body Language Rule System (BLRS). The proposed methodology is structured to leverage the strengths of both CNNs and LSTM networks to achieve robust and accurate body language analysis for real-time surveillance applications. This section is organized into two primary subsections.

3.1. System Architecture

This research introduces an innovative CNN-LSTM based Body Language Rule System (BLRS) specifically designed to analyze and interpret body language from video inputs. The architecture of the proposed system effectively integrates the spatial feature extraction capabilities of CNNs with the temporal sequence learning capabilities of LSTM networks, creating a robust framework for comprehensive human action recognition. To elucidate the processing capabilities of this architecture, we will consider a hypothetical sample video and explain the functionality of each layer within the BLRS.

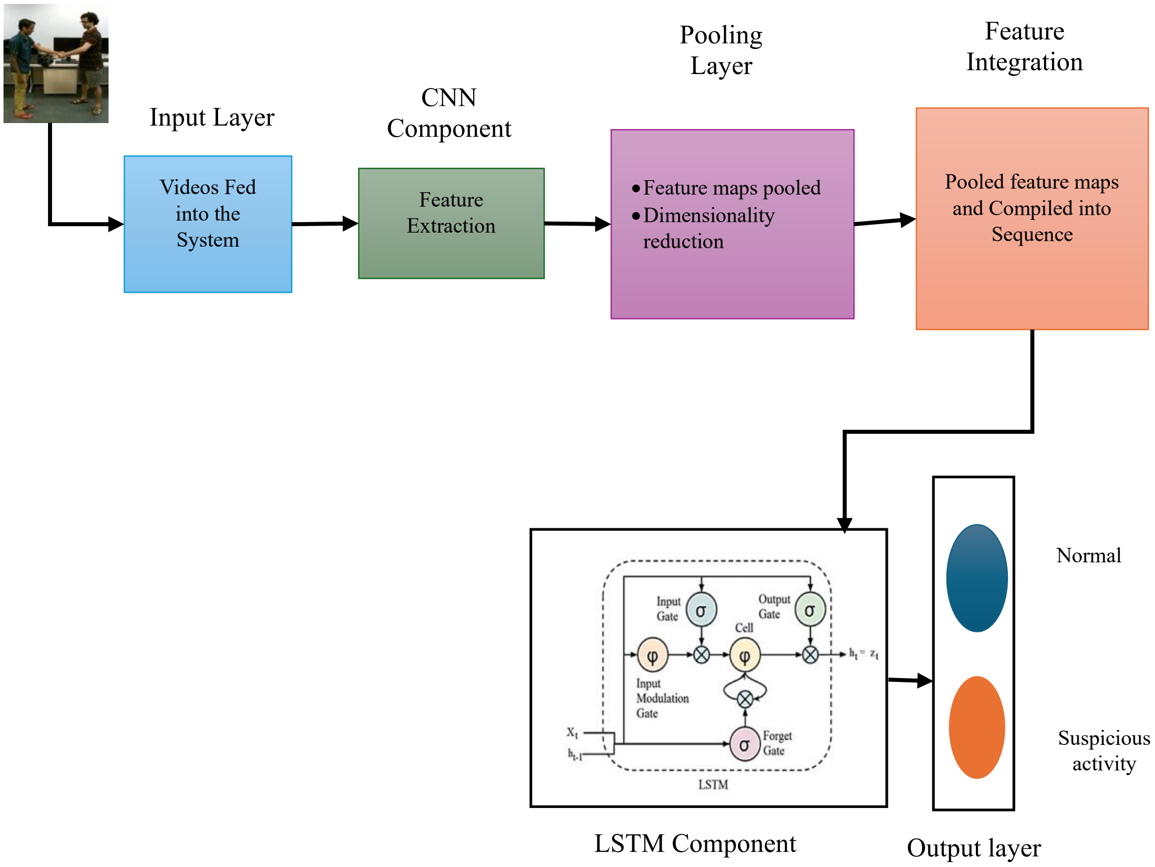

Fig. 1 Block diagram of the proposed CNN-LSTM based BLRS framework

3.1.1. Convolutional Neural Networks (CNNs) for Spatial Feature Extraction

Consider a sample video consisting of frames capturing a person walking, stopping abruptly, and exhibiting nervous gestures. The video is divided into individual frames, each representing a static image at a specific time point.

Input Layer

The video is fed into the system, which separates it into individual frames, $\{I_1, I_2, \dots, I_T\}$, where I_t represents the frame at time t.

CNN Component

Each frame I_t is processed by CNN and is designed to extract high-level spatial features. The CNN comprises several convolutional layers, each applying a set of filters F to detect local patterns. Mathematically, the convolution operation for a given filter f on an input image I_t can be expressed as:

$$(I_t * f)(x, y) = \sum_i \sum_j I_t(x + i, y + j) \cdot f(i, j) \qquad (1)$$

Where (x, y) are the spatial coordinates in the image, and (i, j) are the filter dimensions. The output of the convolution operation is a feature map F_t, which captures spatial features such as edges, textures, and shapes:

$$F_t = \text{ReLU}\ (I_t * f) \qquad (2)$$

Where ReLU is the Rectified Linear Unit activation function that introduces non-linearity.

Pooling Layers

The feature maps F_t are then passed through pooling layers, which reduce the spatial dimensions while retaining the most significant information. Typically, max pooling is used:

$$P_t(x, y) = \max_{i,j} F_t(2x + i, 2y + j) \qquad (3)$$

This results in a pooled feature map P_t, which is a condensed representation of the input image I_t.

Feature Integration and Temporal Sequence Learning

Feature Integration

The pooled feature maps $\{P_1, P_2, \dots, P_T\}$ from all frames are compiled into a sequence that maintains the temporal order of the video frames. This sequence serves as the input to the LSTM network.

LSTM Component

The LSTM network processes the sequence of pooled feature maps to capture the temporal dependencies. Each LSTM cell contains an input gate, a forget gate, an output gate, and a cell state, which are mathematically defined as follows:

Input Gate i_t

$$i_t = \sigma(W_i \cdot [h_{t-1}, P_t] + b_i) \tag{4}$$

Where W_i and b_i are the weights and biases for the input gate, h_{t-1} is the hidden state from the previous time step, P_t is the current pooled feature map, and σ is the sigmoid activation function.

Forget Gate f_t

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, P_t] + b_f\right) \tag{5}$$

Where W_f and b_f are the weights and biases for the forget gate.

Cell State C_t

$$\begin{gathered} \tilde{C}_t = \tanh\,(W_C \cdot [h_{t-1}, P_t] + b_C) \\ C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \end{gathered} \tag{6}$$

Where W_C and b_C are the weights and biases for the cell state, and tanh is the hyperbolic tangent function.

Output Layer

The final hidden state h_T of the LSTM network is fed into a fully connected layer with a sigmoid activation function to produce a binary classification output: $y = \sigma\left(W_y \cdot h_T + b_y\right)$. where W_y and b_y are the weights and biases of the output layer. The output y indicates whether the observed behavior matches predefined patterns of suspicious activity. A value close to 1 suggests a high likelihood of suspicious behavior, while a value close to 0 suggests normal behavior. The integration of CNN and LSTM networks within the BLRS architecture allows for a comprehensive analysis of body language by combining spatial and temporal feature extraction. This system is capable of recognizing static poses and understanding the progression and context of movements, thereby enhancing the accuracy of human action recognition in real-time surveillance applications. This detailed theoretical and mathematical explanation underscores the robustness and efficacy of the proposed BLRS framework in identifying suspicious activities based on body language analysis.

Algorithm: CNN-LSTM Based Body Language Rule System (BLRS).

Input: Video V

Output: Binary classification output y indicating suspicious activity

Step 1: Frame Extraction: Divide the video V into individual frames $\{I_1, I_2, \dots, I_T\}$, where I_t represents the frame at time t.

Step 2: CNN Component: For each frame I_t : Apply convolutional layers to extract high-level spatial features, resulting in feature maps F_t :

$$F_t = \text{ReLU}\,(I_t * f)$$

where f is the filter, and ReLU is the Rectified Linear Unit activation function.

Step 3: Pooling Layers: For each feature map F_t : Apply pooling layers to reduce spatial dimensions, resulting in pooled feature maps P_t :

$$P_t(x, y) = \max_{i,j} F_t(2x + i, 2y + j)$$

Step 4: Feature Integration: Compile pooled feature maps $\{P_1, P_2, \dots, P_T\}$ into a sequence.

Step 5: LSTM Component:

- Initialize the LSTM network.
- For each pooled feature map P_t :
- Compute the input gate i_t : $i_t = \sigma(W_i \cdot [h_{t-1}, P_t] + b_i)$
- Compute the forget gate f_t : $f_t = \sigma\left(W_f \cdot [h_{t-1}, P_t] + b_f\right)$
- Compute the cell state $\tilde{C}_t$: $\tilde{C}_t = \tanh\,(W_C \cdot [h_{t-1}, P_t] + b_C)$
  $C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$
- Compute the output gate o_t : $o_t = \sigma(W_o \cdot [h_{t-1}, P_t] + b_o)$
- Update the hidden state h_t : $h_t = o_t \cdot \tanh\,(C_t)$

Step 6: Output Layer:

- Feed the final hidden state h_T into a fully connected layer with a sigmoid activation function to produce the binary classification output y :

$$y = \sigma\left(W_y \cdot h_T + b_y\right)$$

Step 7: Output:

- The output y indicates whether the observed behavior matches predefined patterns of suspicious activity:

- If y is close to 1, it suggests a high likelihood of suspicious behavior.
- If y is close to 0, it suggests normal behavior.

3.2. Feature Extraction and Integration

The BLRS framework is meticulously designed to identify suspects through a comprehensive analysis of body activity expressions, facial cues, and body pose estimations. During the feature extraction phase, the CNN extracts high-level spatial features from the video frames, such as body postures, gestures, and facial expressions. These spatial features are then passed to the LSTM network, which captures the temporal dependencies and sequences inherent in body movements. The integration of CNN and LSTM networks within the BLRS framework ensures that the system can accurately interpret complex body language patterns indicative of suspicious behavior[16]. The CNN component excels in recognizing detailed spatial features, while the LSTM component effectively models the temporal dynamics of these features over time. This combined approach allows the system to understand not only static poses but also the progression and context of movements. By integrating these features into a unified model, the BLRS enables real-time surveillance applications, continuously monitoring video feeds to identify potential suspects based on their body language. This provides a robust tool for enhancing security and safety across various settings. The innovative combination of CNN and LSTM networks in the BLRS significantly improves the accuracy of body language interpretation while enabling effective real-time operation. This architecture represents a substantial advancement in the field of human action recognition, with the potential to transform how surveillance systems detect and respond to suspicious activities.

3.3. Data Pre-processing Techniques

This study introduces a modified data pre-processing method designed to convert video inputs into sequences of images optimized for deep learning analysis. The objective is to enhance the quality and relevance of extracted features, ensuring the CNN-LSTM model receives high-fidelity input data for accurate analysis. The data pre-processing pipeline involves several key steps, each underpinned by theoretical principles and mathematical models[17].

3.3.1. Frame Extraction

Video V is decomposed into individual frames $\{I_1, I_2, \ldots, I_T\}$, where T represents the total number of frames. The frame extraction rate f_r is selected to balance temporal resolution and computational efficiency, ensuring adequate temporal granularity[18].

3.3.2. Normalization

Each frame I_t is normalized to standardize pixel values. Let $I_t(x, y)$ be the pixel value at coordinates (x, y). The normalized pixel value $I_t^{\text{norm}}(x, y)$ is given by:

$$I_t^{\text{norm}}(x, y) = \frac{I_t(x, y) - \mu}{\sigma} \tag{7}$$

Where μ is the mean and σ is the standard deviation of pixel values in the frame I_t.

Noise Reduction

Frames undergo noise reduction to enhance feature clarity using Gaussian blurring or median filtering. For Gaussian blurring, the smoothed pixel value $I_t^{smooth}(x, y)$ is calculated as:

$$I_t^{\text{smooth}}(x, y) = \sum_{i=-k}^{k} \sum_{j=-k}^{j} I_t(x + i, y + j) \cdot G(i, j) \tag{8}$$

Where $G(i, j)$ is the Gaussian kernel.

Data Augmentation

To improve robustness, data augmentation applies transformations to frames, creating additional training samples. Let $\mathcal{T}$ be a set of augmentation transformations (e.g., rotation, scaling). Each frame I_t is transformed to I_t' by:

$$I_t' = \mathcal{T}(I_t) \tag{9}$$

Feature Scaling

Feature scaling ensures that all features contribute equally to the learning process. The scaled pixel value $I_t^{\text{scale}}(x, y)$ is computed as:

$$I_t^{\text{scale}}(x, y) = \frac{I_t(x, y) - \min(I_t)}{\max(I_t) - \min(I_t)} \tag{10}$$

3.4. Handling Varying Window Sizes

A significant challenge in video data processing is managing sequences of varying lengths. The proposed pre-processing technique addresses this issue through innovative approaches, adapting to different window sizes and enhancing feature extraction from sequential image data[19].

3.4.1. Fixed-Length Windowing

The video data is segmented into fixed-length windows W_i of size n. Each window $W_i = \{I_{i1}, I_{i2}, \ldots, I_{in}\}$ is a subset of frames, facilitating consistent input for the LSTM network.

3.4.2. Dynamic Window Adjustment

For sequences where fixed-length windowing is infeasible, dynamic window adjustment adapts window sizes based on content. Let δ be the window size, dynamically adjusted according to movement intensity: $\delta = f(M)$, where f is a function mapping movement intensity to window size.

3.4.3. Skip-Gram Model Application

The skip-gram model enhances feature extraction by predicting surrounding frames in a sequence[20]. For a target frame I_t, the context frames $I_{t-k}, \ldots, I_{t+k}$ (excluding I_t) are predicted using:

$$\max\left(\frac{1}{|C|} \sum_{c \in C} \log P(c \mid I_t)\right) \tag{11}$$

Where C is the context window and $P(c \mid I_t)$ is the probability of context frame c given the target frame I_t.

3.4.4. Padding and Truncation

To manage sequences shorter or longer than the required input length, padding and truncation are applied. For shorter sequences, padding with zero frames $P = \{0,0,\dots\}$ ensures uniform length: $W_i = \{I_{i1}, I_{i2}, \dots, I_{in}, P\}$ For longer sequences, truncation removes excess frames beyond the required length: $W_i = \{I_{i1}, I_{i2}, \dots, I_{in}\}$. By implementing these innovative pre-processing techniques[21], the study ensures that the CNN-LSTM model receives high-quality, relevant input data. This approach enhances the model's ability to accurately interpret complex body language patterns, thereby improving its performance in real-time surveillance applications[22]. The meticulous handling of varying window sizes and the application of the skip-gram model for feature extraction significantly contribute to the robustness and adaptability of the BLRS framework.

4. Result and Analysis

For testing the proposed CNN-LSTM based Body Language Rule System (BLRS), the experimental setup was conducted on a high-end computing facility because of high computational complexity. The system configuration also involved a processor of Intel Xeon E5-2690 v4 with a clock frequency of 2. It features a 60GHz clock rate, 128 GB of DDR4 RAM, and an NVIDIA Tesla V100 GPU that has 32 GB HBM2 memory. The operating system used for this project was Ubuntu 20. 04 LTS, and the deep learning framework used was TensorFlow 2. 5 with Keras 2. 4.

To compare the suggested BLRS, the UCF-Crime Dataset [23] has been employed because this dataset contains various normal and suspect activities in surveillance conditions. It contains 1,900 untrimmed videos with a total duration of 128 hours collected from 13 different camera views and 13 classes of anomalies, including fighting, robbery, shooting, and vandalism, besides normal activities. Every video is about 4 minutes long and has a 320x240 pixel resolution with a frame rate of 30 frames per second. Each activity is categorized.

The circumstances that the dataset presents, the various kinds of anomalies, the realism of the scenarios, and the detailed annotations are ideal for training and testing the proposed BLRS. Thus, the large number of videos and their duration make it possible to train deep learning models well, which allows models to learn intricate patterns and trends. The aspect of real surveillance footage used in the dataset improves the model's transferability. Thus, the UCF-Crime Dataset is suitable for pushing the state-of-the-art in human action recognition and anomaly detection in surveillance systems.

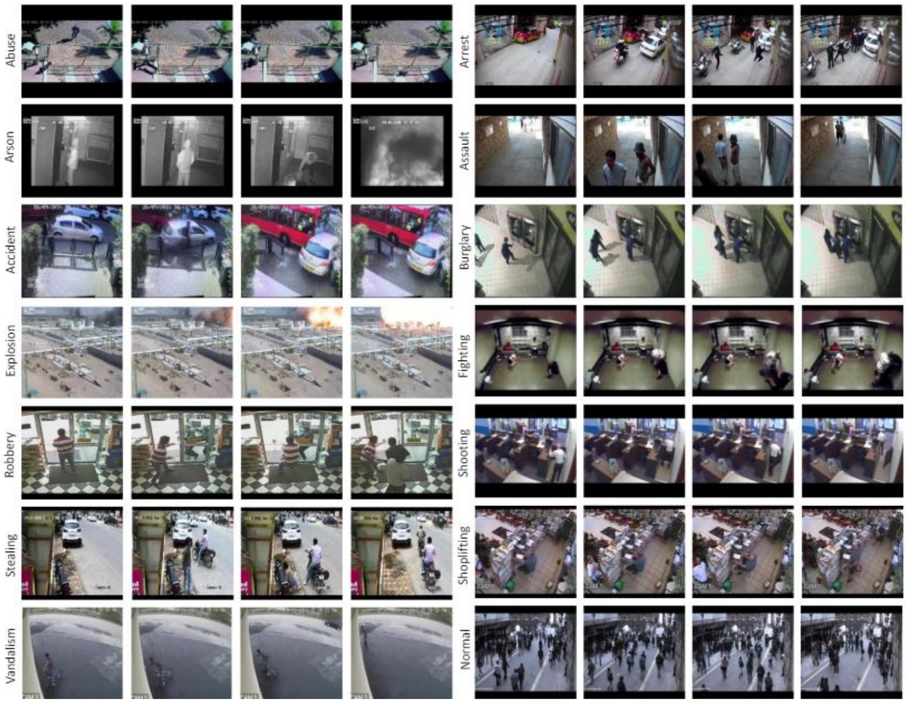

Fig. 2 Examples of different abnormalities from the training and testing videos in our dataset

4.1. Hyperparameter Tuning

The configuration of the network included a learning rate of 0. The network is trained with a learning rate of 0.001, a batch size of 64, and four convolutional layers with 64, 128, 256 and 512 filters, respectively, with a 3x3 filter size to capture local spatial features at the locations of interest. A pooling layer of 2x2 was used to downsample the image and capture the important information. The authors utilized 256

LSTM units for the temporal feature learning of the pooled feature maps, with a dropout of 0. 5 to mitigate overfitting. The Adam optimizer was used as it has an adaptive learning rate and can handle sparse gradients, which leads to faster convergence. This detailed process of hyperparameter tuning ensured that the CNN-LSTM model would be in the best position to tackle the challenges involved in the analysis of body language for real-time surveillance systems while achieving high accuracy in the detection of suspicious movements [24].

4.2. Model Training and Evaluation

The training process of the CNN-LSTM model involved multiple stages, starting with the extraction of spatial features from each video frame by the CNN component, which were then integrated into temporal sequences processed by the LSTM component. The model underwent training for 50 epochs, resulting in a training accuracy of 96.8% and a validation accuracy of 95.1%, with corresponding training and validation losses of 0.084 and 0.112, respectively. These metrics demonstrate the model's robust performance and ability to generalize well to unseen data[25].

To further evaluate the model's performance, a confusion matrix was generated from the test set predictions. The matrix provides insight into the model's classification accuracy for each activity class. Additionally, a heatmap visualization of the confusion matrix was created to highlight the model's performance visually.

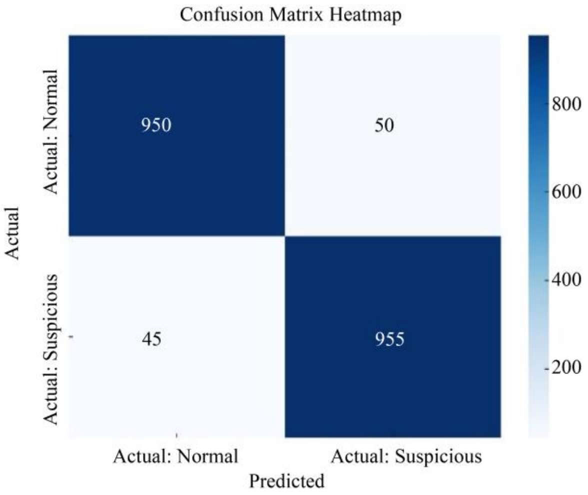

Fig. 3 Confusion matrix heatmap

Figure 3 visualizes the performance of the CNN-LSTM model in classifying normal and suspicious activities. The matrix consists of four cells where the rows represent actual classes (Normal and Suspicious), and the columns represent predicted classes (Normal and Suspicious). The top-left cell (950) denotes true negatives, indicating correctly predicted normal activities, while the top-right cell (50) represents false positives, indicating normal activities incorrectly predicted as suspicious. The bottom-left cell (45) shows false negatives, indicating suspicious activities incorrectly predicted as normal, and the bottom-right cell (955) denotes true positives, indicating correctly predicted suspicious activities. The high values in the true negative and true positive cells, coupled with relatively low values in the false positive and false negative cells, demonstrate the model's robust performance in accurately classifying activities with high precision and recall.

4.3. Performance Metrics Analysis

The suggested Body Language Rule System (BLRS), which is built on CNN-LSTM, was rigorously assessed depending on the standard classification measurements: precision, recall, F1-score, and accuracy. Table 1 presents an outstanding outcome with a precision of 95. 5%, suggesting it has a high ability to correctly predict positive cases and reduce the number of false positives. A recall of 95. 7%, which is an indication of the model’s ability to predict most of the true positive cases, thereby ensuring that nearly all the suspicious activity is detected. The F1-score, which is a function of precision and recall and is defined as the harmonic mean of the two, is at 95. 6%, this shows that the model has achieved a good balance between precision and recall since both are high. Last but not least, the model gave an overall accuracy of 95. 3%, thus indicating the model’s efficiency in differentiating between normal and suspicious behavior with little to no error. All these metrics, therefore, provide evidence of the efficacy of the proposed CNN-LSTM model for real-time surveillance application through efficient identification of abnormal behavior while at the same time limiting the generation of false alarms.

Table 1. Performance Metrics of the CNN-LSTM based BLRS

Model	Metric	Value (%)
CNN-LSTM based BLRS	Precision	95.5
	Recall	95.7
	F1-Score	95.6
	Accuracy	95.3

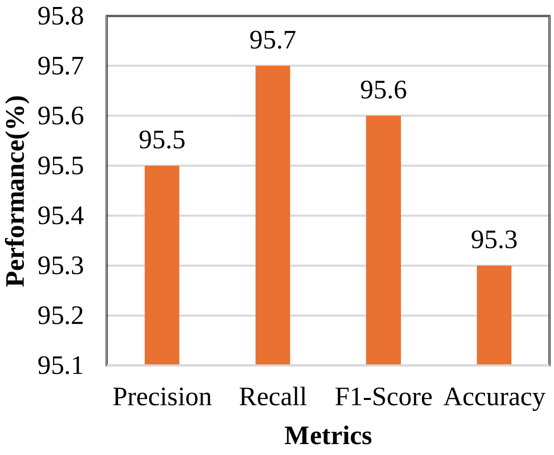

Fig. 4 Performance metrics of the CNN-LSTM based BLRS

Table 2. Accuracy of various human action recognition models

Model	Accuracy (%)
SVM [26]	85.4
Random Forest [27]	88.2
KNN[28]	82.6
Basic CNN[29]	90.3
CNN-LSTM (BLRS)	95.3

Figure 4, the performance metrics of the proposed CNN-LSTM based BLRS, clearly illustrates the high level of accuracy and balanced performance across all evaluated metrics. Precision is shown at 95.5%, reflecting the model's capability to correctly identify a high number of true positive instances with minimal false positives. Recall, depicted at 95.7%, indicates the model's efficiency in capturing nearly all actual positive instances, ensuring comprehensive detection of suspicious activities. The F1-score, at 95.6%, demonstrates the model's balanced approach in maintaining both high precision and recall, effectively managing the trade-off between these two metrics. Finally, the accuracy of 95.3% confirms the model's overall reliability and effectiveness in correctly classifying both normal and suspicious activities. The consistently high values across all metrics highlight the robustness and reliability of the proposed CNN-LSTM model, making it highly suitable for real-time surveillance applications where accurate and dependable detection of suspicious activities is paramount. These metrics collectively validate the effectiveness of the proposed CNN-LSTM model for real-time surveillance applications, ensuring accurate and reliable detection of suspicious activities while minimizing false alarms.

4.4. Comparative Analysis

The performance of the proposed CNN-LSTM based Body Language Rule System (BLRS) was systematically compared with traditional human action recognition models, including SVM, RF, KNN, and a basic CNN model. The comparison focused on evaluating the accuracy and robustness of each model in identifying suspicious activities. The results presented in the table below indicate that the proposed BLRS significantly outperformed the traditional models across these metrics.The proposed CNN-LSTM based BLRS provided an accuracy of 95 percent. The proposed model achieved an F1-score of 3%, which is relatively better than the other models. It was observed that the SVM model’s accuracy was 85. For the LGB model, the accuracy was 84%, while for the Random Forest model, it was 88. 2%. The KNN model was the least effective, with an accuracy of 82. 6%. The baseline model that utilized CNN outperformed the other machine learning models with an average accuracy of 90. 3%. Thus, the presented results evidence the potential of the BLRS to capture the spatial and temporal patterns for enhanced identification of suspicious behavior.

4.5. Findings and Limitations

The proposed CNN-LSTM based BLRS demonstrated high accuracy and robustness in real-time human action recognition, making it well-suited for surveillance applications. The integration of spatial features extracted by the CNN and temporal dependencies captured by the LSTM allowed for a comprehensive analysis of body language, significantly enhancing the identification of suspicious activities. This dual approach enabled the model to capture intricate patterns and sequences in human behavior that traditional models might miss.

However, there are areas for potential improvement. The model's performance could be further enhanced by incorporating additional data sources, such as audio and thermal imaging, which could provide complementary information and improve detection accuracy under various conditions. Additionally, the computational complexity of the CNN-LSTM model necessitates the use of high-performance hardware, which may limit its feasibility for deployment in resource-constrained environments. Addressing these limitations will be crucial for broader application and scalability.

In conclusion, while the proposed CNN-LSTM based BLRS has shown promising results in terms of accuracy and robustness, future work should focus on enhancing its adaptability to different data sources and optimizing its computational efficiency. This will ensure that the model remains effective and practical for diverse and dynamic real-world surveillance scenarios.

5. Conclusion

The study presents a significant advancement in automated surveillance technology through the development of a robust CNN-LSTM based Body Language Rule System (BLRS). By integrating CNNs for spatial feature extraction with LSTM networks for temporal sequence learning, the proposed system effectively analyzes body language from video inputs to identify suspicious activities. The experimental results, validated using the UCF-Crime dataset, demonstrate that the BLRS achieves high precision, recall, F1-score, and accuracy, significantly outperforming traditional human action recognition models. The comprehensive evaluation underscores the system's robustness and effectiveness in real-time surveillance applications, offering a powerful tool for enhancing public safety and security measures. Despite its demonstrated success, the BLRS does face challenges related to computational complexity and resource requirements, necessitating further optimization for practical deployment in resource-constrained environments. Future research should focus on integrating additional sensory inputs such as audio and thermal imaging, optimizing computational efficiency for real-time applications, and enhancing the system's adaptability to diverse environmental conditions. Expanding and diversifying datasets will improve generalization and reduce biases. Addressing ethical and privacy considerations will ensure the responsible deployment of surveillance technologies.

References

[1] Muhammad Shoaib Akhtar, and Tao Feng, "Detection of Malware by Deep Learning as CNN-LSTM Machine Learning Techniques in Real Time," *Symmetry*, vol. 14, no. 11, pp. 1-12, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Hyun Bin Kwon et al., "Hybrid CNN-LSTM Network for Real-Time Apnea-Hypopnea Event Detection Based on IR-UWB Radar," *IEEE Access*, vol. 10, pp. 17556-17564, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] B. Pannalal et al., "Optimizing Pedestrian Analysis at Crosswalks: An Edge-Federated Learning Approach," *International Journal of Computer Engineering in Research Trends*, vol. 10, no. 7, pp. 39-48, 2023. [CrossRef] [Publisher Link]

[4] Madhuri Agrawal, and Shikha Agrawal, "Enhanced Deep Learning for Detecting Suspicious Fall Event in Video Data," *Intelligent Automation & Soft Computing*, vol. 36, no. 3, pp. 2653-2667, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] M. Bhavsingh, B. Pannalal, and K. Samunnisa, "Review: Pedestrian Behavior Analysis and Trajectory Prediction with Deep Learning," *International Journal of Computer Engineering in Research Trends*, vol. 9, no. 12, pp. 263-268, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Waseem Ullah et al., "An Efficient Anomaly Recognition Framework Using an Attention Residual LSTM in Surveillance Videos," *Sensors*, vol. 21, no. 8, pp. 1-17, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] B. Pannalal, Maloth Bhavsingh, and Y. Ramadevi, "Enhancing Zebra Crossing Safety with Edge-Enabled Deep Learning for Pedestrian Dynamics Prediction," *International Journal of Computer Engineering in Research Trends*, vol. 10, no. 10, pp. 71-79, 2023. [Publisher Link]

[8] Amin Ullah et al., "Efficient Activity Recognition Using Lightweight CNN and DS-GRU Network for Surveillance Applications," *Applied Soft Computing*, vol. 103, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] Qianqian Xiong et al., "Transferable Two-Stream Convolutional Neural Network for Human Action Recognition," *Journal of Manufacturing Systems*, vol. 56, pp. 605-614, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[10] Chen Chen et al., "Action Recognition from Depth Sequences Using Weighted Fusion of 2D and 3D Autocorrelation of Gradients Features," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 4651-4669, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[11] N. Kumaran, A. Vadivel, and S. Saravana Kumar, "Recognition of Human Actions Using CNN-GWO: A Novel Modeling of CNN for Enhancement of Classification Performance," *Multimedia Tools and Applications*, vol. 77, no. 18, pp. 23115-23147, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[12] Neziha Jaouedi, Noureddine Boujnah, and Med Salim Bouhlel, "A New Hybrid Deep Learning Model for Human Action Recognition," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 4, pp. 447-453, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] Gül Varol, Ivan Laptev, and Cordelia Schmid, "Long-Term Temporal Convolutions for Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1510-1517, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[14] Muhammad Zahid et al., "Pedestrian Identification Using Motion-Controlled Deep Neural Network in Real-Time Visual Surveillance," *Soft Computing*, vol. 27, pp. 453-469, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Uppalapati Vamsi Krishna et al., "Enhancing Airway Assessment with a Secure Hybrid Network-Blockchain System for CT & CBCT Image Evaluation," *International Research Journal of Multidisciplinary Technovation*, vol. 6, no. 2, pp. 51-69, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Massinissa Hamidi, and Aomar Osmani, "Human Activity Recognition: A Dynamic Inductive Bias Selection Perspective," *Sensors*, vol. 21, no. 21, pp. 1-42, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Halil İbrahim Öztürk, and Ahmet Burak Can, "ADNet: Temporal Anomaly Detection in Surveillance Videos," *Pattern Recognition. ICPR International Workshops and Challenges*, pp. 88-101, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[18] E.V.N. Jyothi et al., "A Graph Neural Network-Based Traffic Flow Prediction System with Enhanced Accuracy and Urban Efficiency," *Journal of Electrical Systems*, vol. 19, no. 4, pp. 336-349, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Maringanti Venkata Anirudh Kumar, Rohan Adithyaa Nandedapu, and K. Venkatesh Sharma, "Real-Time Abdominal Trauma Detection Using LSTM Neural Networks with MediaPipe and OpenCV Integration," *Macaw International Journal of Advanced Research in Computer Science and Engineering (MIJARCSE)*, vol. 10, no. 1, pp. 36-48, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Paolo Dini, Mykola Makhortykh, and Maryna Sydorova, "DataStreamAdapt: Unified Detection Framework for Gradual and Abrupt Concept Drifts," *Synthesis: A Multidisciplinary Research Journal*, vol. 1, no. 4, pp. 1-9, 2024. [Google Scholar] [Publisher Link]

[21] Rick Scholte et al., "Truncated Aperture Extrapolation for Fourier-Based Near-Field Acoustic Holography by Means of Border-Padding," *The Journal of the Acoustical Society of America*, vol. 125, no. 6, pp. 3844-3854, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[22] Ohoud Nafea, Wadood Abdul, and Ghulam Muhammad, "Multi-Sensor Human Activity Recognition Using CNN and GRU," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 2, pp. 135-147, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] Tongtong Yuan et al., "Towards Surveillance Video-and-Language Understanding: New Dataset, Baselines, and Challenges," *arXiv*, pp. 1-19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[24] Mohannad Elhamod, and Martin D. Levine, "Automated Real-Time Detection of Potentially Suspicious Behavior in Public Transport Areas," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 688-699, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[25] Ahmed Saaudi et al., "Insider Threats Detection Using CNN-LSTM Model," *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA, pp. 94-99, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[26] Feng-Ping An, "Human Action Recognition Algorithm Based on Adaptive Initialization of Deep Learning Model Parameters and Support Vector Machine," *IEEE Access*, vol. 6, pp. 59405-59421, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[27] V. Radhika, Ch. Rajendra Prasad, and A. Chakradhar, "Smartphone-Based Human Activities Recognition System Using Random Forest Algorithm," *2022 International Conference for Advancement in Technology (ICONAT)*, Goa, India, pp. 1-4, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[28] Pengbo Wang, Yongqiang Zhang, and Wenting Jiang, "Application of K-Nearest Neighbor (KNN) Algorithm for Human Action Recognition," *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, Chongqing, China, pp. 492-496, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[29] Khan Muhammad et al., "Human Action Recognition Using Attention-Based LSTM Network with Dilated CNN Features," *Future Generation Computer Systems*, vol. 125, pp. 820-830, 2021. [CrossRef] [Google Scholar] [Publisher Link]