









Real-Time Deep Learning-Driven Surveillance with Spatiotemporal Feature Extraction for Detection of Anomalous Human Behavior Across Dynamic Environments



Madhuri Pangavhane¹, Rahul Patil², Rajesh Bharati³, Deepak Gupta⁴, Prashant Ahire¹, Pramod Patil³,
Wasudeo Rahane³, Deepak Dharrao^{1*}

¹Department of Computer Science and Engineering, Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune 412115, India

²Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune 411044, India

³Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pune 411018, India

⁴Department of Computer Science & Engineering, Institute of Technology & Management, Gwalior 474001, India

Corresponding Author Email: deepak.dharrao@sitpune.edu.in

Copyright: ©2025 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ijss.150112>

ABSTRACT

Received: 24 October 2024

Revised: 26 November 2024

Accepted: 20 December 2024

Available online: 31 January 2025

Keywords:

Inception V3 network, suspicious activity detection, deep learning, video surveillance, automated surveillance systems, public safety, spatiotemporal features, attention mechanism

Nowadays continuous monitoring of public and private environments through Closed-Circuit Television (CCTV) is at peak attention. The identification and reporting of suspected human activity is crucial for the safety and security of the individuals and their belongings. Many researchers have provided numerous solutions for automated activity monitoring with CCTV and machine learning applications. But these models are struggling to provide high accuracy with real-time detection and triggering alert systems in a dynamic environment. The proposed model addresses these issues with a combined approach of convolutional and recurrent neural networks (Inception V3 and Bidirectional Long Short-Term Memory (BiLSTM)) with an attention mechanism to classify videos. This proposed model uses the strength of the Inception V3 network to extract spatial features from video frames, and the BiLSTM network processes these features in a time-dependent manner to the identification of suspicious human activities. Also, the attention mechanism added to proposed system to focuses on the most significant spatiotemporal variables for violence detection. This deep learning model is designed to extract spatiotemporal features and thus can extract complex patterns in human motion robustly. This model is trained with publicly available dataset to analyze the performance with accuracy in a dynamic environment. The proposed deep-learning model effectively identifies and categorizes numerous suspicious behaviors in real-time by scrutinizing video sequences. The performance analysis proves that the proposed model efficiently detects activities like loitering, aggressive behavior, and unauthorized access. This automated surveillance system strengthens security in homes as well as other public and private environments.

1. INTRODUCTION

The development of automated human activity monitoring through Closed-Circuit Television (CCTV) is in high demand for the safety and security of individuals and their belongings. The detection of anomalous human behavior is a challenging task due to varying environmental conditions in public and private places. The method for detecting suspicious human activities in residential and public areas addresses the need for increased safety and security of individuals [1]. With the increase in CCTV installation at residential and public places, there is a significant opportunity to use this infrastructure to implement proactive safety measures. However, the traditional security methods hampered the expected security by a lack of real-time monitoring, resulting in missed possibilities for timely intervention during suspicious activities like loitering, aggressive behavior, and unauthorized access [2]. Current methods of detecting suspicious activities and public safety

have relied mainly on manual monitoring, which is time-consuming, highly vulnerable to human mistake and delayed responses.

The existing automated techniques frequently failed to give real-time alerts and relied on computationally heavy and expensive frameworks [2]. Recent security measures in this regard have led to very rapid growth of surveillance systems to improve security. Most of the surveillance systems characterize human activities as the 'normal' or the 'abnormal'. The traditional modes of surveillance are based on manual observation and become complex and require significant human interaction hence they are time-consuming and likely to contain errors. Surveillance systems have evolved significantly with advancements in deep learning, enabling real-time detection of anomalous human behavior. However, existing approaches often falter in dynamic and unstructured environments where factors like occlusion, lighting variability, and complex human-object interactions challenge model

performance. To address such challenges, new system is proposed using deep learning techniques.

Spatiotemporal feature extraction has emerged as a promising direction, leveraging spatial and temporal dimensions for deeper context understanding. Despite its potential, its integration into real-time systems remains underexplored, particularly in dynamic, high-traffic scenarios. Convolutional, neural Networks or CNNs are among the most successful models, especially in the area of computer vision, particularly image and video. It provides the solution on major research gap of improving robustness and accuracy with ensemble-based of surveillance systems under varying environmental conditions. Firstly, extracted frames are passed through an Inception Network which is a convolutional neural network (CNN) model designed to extract important features from the input images (frames). It identifies relevant patterns in the frame like objects, textures, and other details. Secondly, it uses Bidirectional Long Short-Term Memory (BiLSTM) networks which are specialized for processing sequences, making them ideal for analyzing the temporal dynamics of the video, such as detecting actions or events that unfold over time. BiLSTM's superior ability to recognize patterns makes it suitable to the human activity classification problem. But if they are incorporated into CNNs one can generate highly effective automatic systems for the constant analysis of voluminous video and identification of prohibited activities. To improve the temporal feature extraction process, an attention mechanism to the BiLSTM network was added.

The proposed system aims to implement a deep leaning based technique (CNN+RNN) to address the task of detecting malicious behaviors of humans in surveillance videos. This proposed model provides more efficient anomalous activity detection and provides an efficient solution for improving the existing security systems in public spaces, transport structures, and other premises. This model focuses on spatiotemporal dependencies to identify human movements and activities.

The first section introduced the basic concept of human activities' classification and the methods developed earlier. Second section provides a comprehensive overview of the existing solutions and the challenges which motivated the development of the proposed system. Third section provides the materials and methodology used in the model. The next sections are validation, results discussion and conclusion.

2. LITERATURE REVIEW

There is a vast literature in the area of information analysis and human research. Some of the Human Detection and Recognition are classification, tracking, behavior analysis and object/motion detection. The following tools are typically used in the observation method: Thermal and CCD cameras and equipment for use at night.

This is especially so in cognitive games where we need to identify people (actors) for analysis, and person recognition suits this purpose well. Many articles are created with the target audience as the primary focus. The concept that employs walking lines and face characteristics. In addition, for analysis, they used the curvature-based matching approach (CBM) [3]. The authors wrote about the application of optical flows for object detection [4]. Further, he outlined the key processes involved in video surveillance. The application of the Hidden Markov Model (HMM) for human activity recognition from video [5]. To represent human behavior, they employed a

stochastic sequence of activities. The application of silhouette directionality for human activity recognition. It is presented in this work as a nonintrusive human activity recognition method [6, 7]. From the input movies, they reconstruct motion information and generate silhouettes (foreground) using the dynamic background-foreground separation method [8].

Many scholars have proposed a variety of techniques for recognizing human activities. A blob feature-based approach provides a method for detecting human activity in videos in their research study [9]. To extract the foreground elements in the video sequence, they employ background subtraction. Current surveillance systems do not recognize these terms because they are based on old technology. The ESMD Framework has been put forward to enhance current systems for detecting malicious conversations. It does this by expanding short-form words into full words and then categorizing the type of crime mentioned [10]. These aspects raise questions about the risks that may be associated with overt suspicious behavior. As crime rates have increased in urban and suburban regions, such activities should be detected to minimize such occurrences.

Surveillance in the past was done physically by people which was tiresome because suspicious activities were rarely or less than normal operational activities. However, with the coming of intelligent surveillance systems, different methods of surveillance have been established. Two scenarios that could pose a significant risk to human life if overlooked: recognizing possible gun-related offenses and detecting left luggage in videos [10]. The possibility of gun-related crimes and abandoned baggage in the video footage examined with the use of machine learning and deep learning methodologies [11, 12].

In video systems human activity detection is a technique of processing clips of a video and deciding which activities to incorporate in the movie. It is one of the growing fields in computer vision and artificial intelligence. The act of searching for undesirable human actions in specific environments and conditions is called suspicious activity recognition. To do this, the video is divided into frames and the subjects inside the processed frames are analyzed for their actions. Because human bodies are soft and can change their size and form at any given time, detecting humans has always been a problem. Face recognition and detection are difficult in indoor and outdoor environments due to various issues such as inadequate illumination and dynamic movements in the environment. Suspicious behaviors, such as breaking locks and stealing bags, can be detected by our system through YOLOv3. It provides efficient analysis and identification [13]. The convolutional layers which are capable of identifying and recognizing objects, especially in images, as demonstrated by recent applications of anomaly detection. Convolutional neural networks, on the other hand, are supervised and need to be labeled input for learning. The spatiotemporal architecture that may be used to detect suspicious activity in films, even in busy environments. Determine the characteristics or actions that are relevant to distinguishing between normal and questionable activity using space and spatial feature extraction. They have experimented performance of their model on the Avenue, Subway, and UCSD benchmark datasets show that at high speeds of up to 140 frames per second [14].

In another study the researchers divided the video into separate image frames to look for any odd behavior in the movie. It was also demonstrated that under sampling the k-space data using quick and multiclass dictionaries could lead

to significant improvements in the reconstructed image in magnetic resonance imaging. In magnetic resonance image reconstruction, a fast-orthogonal dictionary learning method is used for building sparse descriptions of images as described by previous study [7]. Video sensor systems that track human movement might be employed in biological and medical applications. Another research work presented a method to identify six abnormal behaviors from everyday activities of humans: They include; forward fall, backward fall, chest discomfort, fainting, vomiting and headache [10]. Those showing only the initiations of activity are employed in a statistical method termed “human activity anticipation”, aimed at determining further continuing activities. This is not to classify completed activities after the fact, but rather to detect when some operation has not been completed fully.

Methodologies for activity prediction are especially important for surveillance systems, which are needed to stop criminal activities and harmful behaviors [15, 16]. Additionally, video surveillance relies heavily on the early detection and prediction of human behavior. For example, by identifying illegal conduct beforehand to avoid unfavorable consequences. They used a Spatial-Temporal Implicit Shape Model to characterize the sparse local data gathered from a video in terms of spatiotemporal structure. Matching patterns using BiLSTM enables early detection of human activities [17]. Some previous studies often struggle with real-time processing, handling dynamic environments, and robust spatiotemporal feature extraction [18, 19]. Most approaches focus on static scenes and lack scalability. Our model addresses these gaps by leveraging deep learning for efficient real-time anomaly detection, integrating spatiotemporal features for higher adaptability to dynamic scenarios.

To spot possible firearm-related crimes and cases of unattended baggage in surveillance videos, another technique utilizes a deep neural network model to recognize handguns in images and a machine learning and computer vision system to identify abandoned luggage [20, 21].

3. PROPOSED WORK

The proposed architecture is based on convolutional and recurrent layers to classify suspicious human activities from surveillance videos for safety and security of individuals and their belongings.

3.1 Datasets

The dataset used in this study is a selected subset of the KTH Human Motion Dataset, a well-known standard for human activity recognition tasks [22]. The dataset used in our experiment is divided into two categories: violence and nonviolence, with a total of 2000 video clips in.avi format. Each video clip is very brief, lasting several seconds, and sampled at a frame rate of 25 frames per second (fps), making it ideal for evaluating spatiotemporal patterns in human movements.

The Violence folder contains 1000 video clips that primarily feature situations of physical aggressiveness, such as boxing and street fighting. These violent sequences feature a variety of fighting actions, including hand-to-hand combat (fists and other physical altercations) that resembles real-life street violence. These videos use various angles and surroundings to replicate real-life violent scenarios. The emphasis on fistfights,

particularly boxing, gives a clear and regulated set of aggressive movements, making this subset suitable for training models to recognize violent activity.

The Non-Violence folder contains 500 clips of common human activities like walking, jogging, and other non-aggressive actions. These clips were chosen to contrast with the violent activities by showing peaceful, regular behaviors. These activities, like as walking and jogging, are simple and consistent, allowing models to learn key traits that distinguish nonviolent from violent behavior. The sample data classes images are shown in Figure 1 and Figure 2, represent sample images of group activities.



Figure 1. Sample data classes images



Figure 2. Group of various activities

3.2 System architecture

The proposed system architecture is shown in Figure 3. This model is mainly divided in two phases. The first phase is designed with deep learning architecture that included Inception V3 network which eliminates the high-level features of the images and simplify the information process and used for spatial feature extraction. A Second phase is designed with Bidirectional Long Short-Term Memory (BiLSTM) networks which specialized for capturing the temporal dynamics of the videos. The model also includes an attention mechanism, which focuses on the most significant spatiotemporal variables for violence detection.

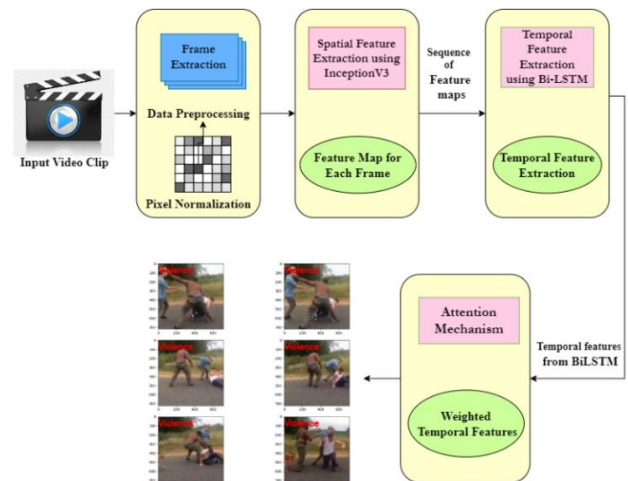


Figure 3. Video classification architecture

3.3 Data acquisition and preprocessing

The first step involves capturing video data from surveillance cameras that are fixed at various parts of the environment that have been deemed open space, transportation means, and other places. This video data becomes the input of the system, as further illustrated in the next videos on this page.

3.3.1 Preprocessing steps include:

Frame extraction: a process of partitioning the quantized footage into sections coming from the video frames.

Normalization: standardization, which involves bringing pixel values into a standard range of 0 to 1, is another commonly used feature pre-processing technique.

Data augmentation: augmentation involves flipping, rotation, scaling, cropping, and others to make the training data set more diverse and thereby minimize overfitting.

3.4 Spatial feature extraction using Inception V3

The extracted frames are passed through an Inception Network as illustrated in Figure 4, which is a convolutional neural network (CNN) model designed to extract important features from the input images (frames). Inception V3, a convolutional neural network pre-trained on the ImageNet dataset, to extract spatial features. Inception V3 is well-known for efficiently extracting detailed spatial patterns from images, thanks to its unique architecture, which includes inception modules that handle multi-scale feature extraction. This was crucial for comprehending the appearance and context of each frame.

For this task, ImageNet weights were loaded while excluding the top (completely linked) layers of the Inception V3 model. The remaining convolutional layers were frozen during training to avoid fine-tuning, allowing the model to concentrate on temporal dynamics without overfitting to spatial features.

It identifies relevant patterns in the frame like objects, textures, and other details.

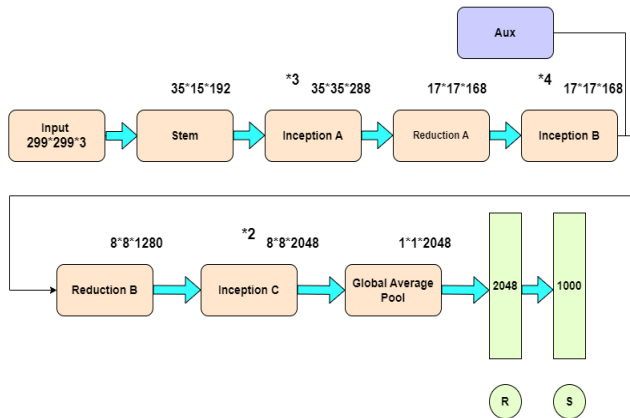


Figure 4. Inception V3 model

This network outputs Transfer Values which is greater than 15, which are essentially a set of learned features that describe the content of each frame. This process loops back and extract new features so that classification process will be more accurate. After getting transfer value as per our threshold, these features are grouped together into single pattern. The inception network organizes the features from multiple frames into a sequence that captures temporal information across the

video. Figure 4 illustrates the extraction of frames and identify the patterns and the convert them into the values which show the contents of frame.

3.5 Temporal feature extraction using BiLSTM

The grouped transfer values gathered from the output of Inception V3 network are fed into an BiLSTM network.

The temporal feature exactions important parameter in identifying anomalous human behavior across dynamic environments. To model the temporal evolution of actions across frames, a BiLSTM network, as illustrated in Figure 5, was employed. BiLSTM's are very useful in video analysis because they can process data in both forward and backward directions, capturing temporal dependencies from the past and future. This is critical for identifying behaviors such as punches, kicks, and movements that lead to violent confrontations.

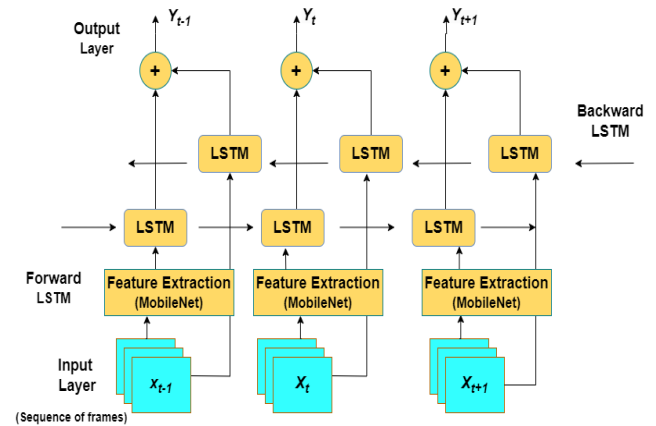


Figure 5. BiLSTM network

Forward LSTM: It processes the sequence in a forward direction, capturing dependencies from past to future. The forward LSTM can be denoted by the Eq. (1) where \vec{h}_t represents the hidden states at time step t .

$$\vec{h}_t = \text{LSTM}(x_t, \vec{h}_{t-1}) \quad (1)$$

Backward LSTM: It processes the sequence in a backward direction, capturing dependencies from future to past. The Eq. (2) denotes \overleftarrow{h} which represents the hidden states at time step t for backward LSTM.

$$\overleftarrow{h}_t = \text{LSTM}(x_t, \overleftarrow{h}_{t+1}) \quad (2)$$

Concatenation: The outputs of the forward and backward LSTMs at each time step is combined, forming a richer representation as shown in Eq. (3).

$$h_t = \begin{bmatrix} \vec{h}_t \\ \overleftarrow{h}_t \end{bmatrix} \quad (3)$$

By processing the sequence in both directions, BiLSTM can utilize knowledge from both past and future contexts, which is particularly effective for detecting temporal features as it requires context from both directions. The model with its parameters and values can be observed in the Table 1. The

model is built using a sequential API, indicating that layers are added sequentially one after another.

Table 1. Model specific configuration

| Model Parameters | Values |
|--------------------------|--------------------------|
| Model Architecture | Sequential |
| Input Shape | (16, 64, 64, 3) |
| Bidirectional LSTM Layer | 1 |
| Dropout Layers | 6 |
| Dense Layers | 4 |
| Dropout Rate | 0.25 |
| Loss Function | Categorical Crossentropy |
| Epochs | 50 |
| Batch Size | 16 |
| Validation Split | 0.2 |

The BiLSTM networks is recurring neural network model which is capable of remembering previous states, making them useful for tasks that require learning patterns over longer time intervals. These characteristic of BiLSTM make it ideal for analyzing the temporal dynamics of the video, such as detecting actions or events that unfold over time. The BiLSTM analyzed the feature maps created by the Inception V3 model for each frame in a sequential manner, learning movement patterns and changes across frames. This temporal study was critical in distinguishing between violent activities, which frequently feature quick, powerful motions, and nonviolent actions, which are more consistent and less abrupt.

3.6 Attention mechanism

To improve the temporal feature extraction process, an attention mechanism to the BiLSTM network was added [23]. The attention layer was designed to direct the model's attention to the most relevant frames in the sequence, giving higher weight to frames that provided more critical information for recognizing violent behaviors. For example, in a boxing battle, frames with punches are more essential than frames with combatants moving into position. This selective weighting approach helps the model prioritize crucial moments, resulting in better performance.

3.7 Compatibility and benefits of ensemble modeling

The proposed system is designed and developed to capture the abnormal or anomalies human behaviour from CCTV video clips. So, there is a need of capturing video sequences and compare to recently captured video clips to make differentiate between normal or abnormal activities. This can be achieved with capturing spatial features from video frames and compare through temporal pattern to classify human activities. The Inception network superiors in extracting spatial features (what the person is doing in each frame), while the BiLSTM network efficient to handle temporal understanding (how the person's behaviour changes over time). Combining these two strengths allows the system to capture both spatial and temporal aspects of suspicious behaviour [23, 24].

By the ensemble of two networks with attention mechanism, the proposed model can generalize better across different behaviours and scenarios. The Inception V3 Network (CNN) ensures that visual features are captured accurately, while the BiLSTM support to capture generalize behaviour patterns over

time, making it better at detecting anomalies in human behaviour.

Finally, the model included a fully connected (Dense) layer with a sigmoid activation function that produced a probability score for binary classification: 1 for violent and 0 for nonviolent video clips. This score showed the model's confidence in its prediction, allowing for a clear separation of the two classes [25, 26].

4. RESULTS AND DISCUSSIONS

This section presents the performance tests done on the proposed approach. The model was trained on the KTH dataset using the Adam optimizer, with a learning rate of 0.0001 to allow for slow convergence. The binary cross-entropy loss function was used due to the binary character of the classification job.

The model was trained for 20 epochs, but early halting guaranteed that it did not always complete the maximum number of epochs. To ensure a balance of violent and non-violent video clips in both training and validation sets, the data was stratified into 80% training and 20% validation.

Furthermore, the results of several analyses are quantified, addressed, and compared to previous research in stationary action recognition. This proposed pipeline architecture uses a combination of convolutional and recurrent neural networks (Inception V3 and BiLSTM) with attention mechanism to classify videos. The Inception network extracts spatial features from video frames, and the BiLSTM network processes these features in a time-dependent manner to produce a classification result.

Figure 6 depicts the result of the loss for the convolutional neural network model, while Figure 7 displays the accuracy result.

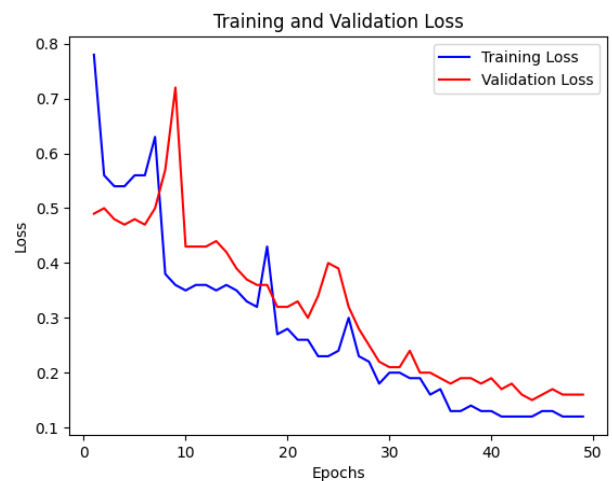


Figure 6. Loss graph of proposed model

The above graph of loss validation in Figure 6 illustrates that the training loss as shown in blue line decreases consistently, which indicates that the learning rate of model is good on the training data. However, the red line indicated that the validation loss shows some fluctuations at initial phase, but then it displays significant improvement in comparison to training result analysis. The validation loss equivalent to the training loss, showcases the efficient performance on unseen data also. The spikes in the validation loss suggest moments

where the model has difficulty predicting certain behaviours or sequences, indicating room for improvement, possibly through regularization techniques or better hyperparameter tuning.

The training accuracy (blue line) improves steadily, indicating that the model becomes more proficient at predicting human behaviour during training as shown in Figure 7. The validation accuracy (red line), although fluctuating at the beginning, stabilizes later but remains consistently lower than the training accuracy. This further supports the hypothesis of overfitting, as the model performs significantly better on the training data than on the validation data.

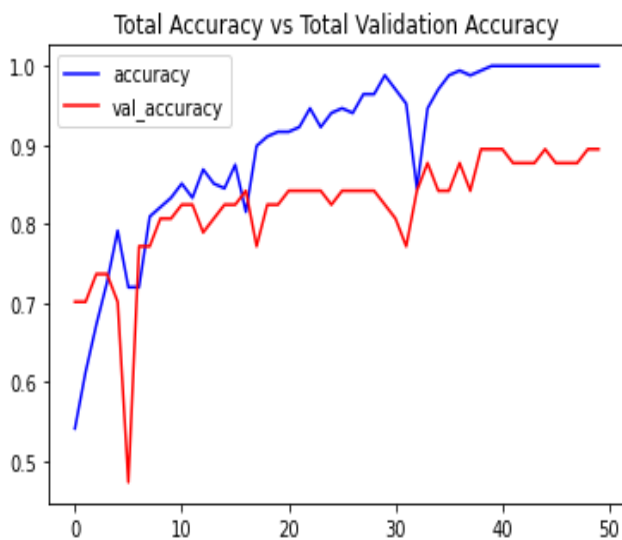


Figure 7. Accuracy graph of proposed model

A graphical representation of proposed ensemble model shows that combining an Inception V3 network and BiLSTM is a powerful approach for capturing human behavior, especially in multimodal data like video, where spatial (Inception) and temporal (BiLSTM) patterns are crucial. The validation results prove that the ensemble model of inception Network (CNN) with BiLSTM benefits from the CNN's capability to capture complex details in every frame, like body posture or object detection (e.g., a person holding a suspicious item), and the BiLSTM's ability to analyse the sequence of actions (e.g., detecting the event when someone following a suspicious path or lingering in restricted areas). The model demonstrates a consistent improvement in accuracy and a decrease in loss across training and validation datasets. The validation accuracy stabilizes after an initial fluctuation, indicating that the model generalizes well to unseen data. The alignment between training and validation loss curves suggests that overfitting is effectively minimized. The improved performance can be attributed to the effective use of spatiotemporal feature extraction and appropriate parameter tuning, addressing the research objective of achieving accurate anomaly detection in dynamic environments. The result analysis shows that training accuracy consistently improved and reached stability after ~30 epochs. As this combined approach allows the proposed system to efficiently handle both short-term and long-term behaviour patterns. For instance, a person walking normally (short-term) in comparison to the somebody pacing back and forth suspiciously over a longer period (long-term).

5. CONCLUSIONS

The proposed ensemble model of Inception V3 network and BiLSTM architecture with attention mechanism proven the significant improvement in real-time suspicious human activity detection through CCTV cameras. The significant advancements in the field of suspicious activity detection enable its versatile application across various domains. Additionally, research in related areas, such as activity tracking, holds promise for enhancing its utilization in multiple fields. Continuous monitoring against predefined conditions aids in detecting prescribed activities of interest efficiently. This approach facilitates real-time performance and eliminates the need for extensive training required by machine learning-based methods. Human behaviors exhibit complexity and diversity in natural environments. Thus, in this paper, we formulate suspicious action detection for security systems. We achieved an accuracy of approximately 92%. However, the present technique for extracting features produces precise outcomes only in regulated settings. Integrating more advanced techniques for feature extraction could improve performance. Also broadening the training dataset to encompass suspicious videos of diverse actions and resolutions is vital for enhancement.

REFERENCES

- [1] Gorodnichy, D., Mungham, T. (2008). Automated video surveillance: Challenges and solutions. ACE Surveillance (Annotated Critical Evidence) Case Study. In NATO SET-125 Symposium "Sensor and Technology for Defence against Terrorism", Mainheim. <https://www.researchgate.net/publication/229040125>.
- [2] Ren, J., Xia, F., Liu, Y., Lee, I. (2021). Deep video anomaly detection: Opportunities and challenges. In 2021 International Conference on Data Mining Workshops (ICDMW), Auckland, New Zealand, pp. 959-966. <https://doi.org/10.1109/ICDMW53433.2021.00125>
- [3] Ko, T. (2008). A survey on behavior analysis in video surveillance for homeland security applications. In 2008 37th IEEE Applied Imagery Pattern Recognition Workshop, Washington, DC, USA, pp. 1-8. <https://doi.org/10.1109/AIPR.2008.4906450>
- [4] Robertson, N., Reid, I. (2006). A general method for human activity recognition in video. Computer Vision and Image Understanding, 104(2-3): 232-248. <https://doi.org/10.1016/j.cviu.2006.07.006>
- [5] Yang, Y., Sun, J., Li, H., Xu, Z. (2016). Deep ADMM-Net for compressive sensing MRI. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, pp. 10-18. <https://dl.acm.org/doi/abs/10.5555/3157096.3157098>.
- [6] Zhan, Z., Cai, J.F., Guo, D., Liu, Y., Chen, Z., Qu, X. (2015). Fast multiclass dictionaries learning with geometrical directions in MRI reconstruction. IEEE Transactions on Biomedical Engineering, 63(9): 1850-1861. <https://doi.org/10.1109/TBME.2015.2503756>
- [7] Piyathilaka, L., Kodagoda, S. (2013). Gaussian mixture based HMM for human daily activity recognition using 3D skeleton features. In 2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA), Melbourne, VIC, Australia, pp. 567-572.

- <https://doi.org/10.1109/ICIEA.2013.6566433>
- [8] Khan, Z.A., Sohn, W. (2011). Abnormal human activity recognition system based on R-transform and kernel discriminant technique for elderly home care. *IEEE Transactions on Consumer Electronics*, 57(4): 1843-1850. <https://doi.org/10.1109/TCE.2011.6131162>
- [9] Zhou, X., Bhanu, B. (2007). Integrating face and gait for human recognition at a distance in video. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(5): 1119-1137. <https://doi.org/10.1109/TSMCB.2006.889612>
- [10] Bashir, F., Usher, D., Casaverde, P., Friedman, M. (2008). Video surveillance for biometrics: Long-range multi-biometric system. In *2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance*, Santa Fe, NM, USA pp. 175-182. <https://doi.org/10.1109/AVSS.2008.28>
- [11] Alahi, A., Ortiz, R., Vandergheynst, P. (2012). Freak: Fast retina keypoint. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, pp. 510-517. <https://doi.org/10.1109/CVPR.2012.6247715>
- [12] Yu, G., Yuan, J., Liu, Z. (2012). Predicting human activities using spatiotemporal structure of interest points. In *Proceedings of the 20th ACM International Conference on Multimedia*, Nara, Japan, pp. 1049-1052. <https://dl.acm.org/doi/abs/10.1145/2393347.2396380>.
- [13] Gowsikhaa D., Manjunath, Abirami S. (2012). Suspicious human activity detection from surveillance videos. *International Journal on Internet & Distributed Computing Systems (IJIDCS)*, 2(2): 141-148. <http://www.ijidcs.org/issues/v2n2/ijidcs-17-141-148.pdf>.
- [14] Ryoo, M.S. (2011). Human activity prediction: Early recognition of ongoing activities from streaming videos. In *2011 International Conference on Computer Vision*, Barcelona, Spain, pp. 1036-1043. <https://doi.org/10.1109/ICCV.2011.6126349>
- [15] Bordoloi, N., Talukdar, A.K., Sarma, K.K. (2020). Suspicious activity detection from videos using YOLOV3. In *2020 IEEE 17th India Council International Conference (INDICON)*, New Delhi, India, pp. 1-5. <https://doi.org/10.1109/INDICON49873.2020.9342230>
- [16] Wan, B., Jiang, W., Fang, Y., Luo, Z., Ding, G. (2021). Anomaly detection in video sequences: A benchmark and computational model. *IET Image Processing*, 15(14): 3454-3465. <https://doi.org/10.1049/ipr2.12258>
- [17] Ali, M.M., Qaseem, M.S., Rasheed, M.A., Khan, M.K.A. (2020). ESMD: Enhanced suspicious message detection framework in instant messaging applications. In *2020 Fourth International Conference on Inventive Systems and Control (ICISC)*, Coimbatore, India, pp. 777-784. <https://doi.org/10.1109/ICISC47916.2020.9171122>
- [18] Loganathan, S., Kariyawasam, G., Sumathipala, P. (2019). Suspicious activity detection in surveillance footage. In *2019 International Conference on Electrical and Computing technologies and applications (ICECTA)*, Ras Al Khaimah, United Arab Emirates, pp. 1-4. <https://doi.org/10.1109/ICECTA48151.2019.8959600>
- [19] Sultani, W., Chen, C., Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 6479-6488. <https://doi.org/10.1109/CVPR.2018.00678>
- [20] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S. (2016). Learning temporal regularity in video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 733-742. <https://doi.org/10.1109/CVPR.2016.86>
- [21] Dharrao, D.S., Uke, N.J. (2019). Fractional Krill-Lion algorithm based actor critic neural network for face recognition in real time surveillance videos. *International Journal of Computational Intelligence and Applications*, 18(2): 1950011. <https://doi.org/10.1142/S1469026819500111>
- [22] Hassan, N., Miah, A.S.M., Shin, J. (2024). A deep bidirectional LSTM model enhanced by transfer-learning-based feature extraction for dynamic human activity recognition. *Applied Sciences*, 14(2): 603. <https://doi.org/10.3390/app14020603>
- [23] KTH Human Motion. <https://www.kaggle.com/datasets/beosup/kth-human-motion/data>.
- [24] Kale, S., Patil, K., Satghare, P., Dharrao, D. (2018). Real time object tracking system with automatic pan tilt zoom features for detecting various objects. *International Journal of Recent Technology and Engineering*, 6(6). <https://www.ijrte.org/portfolio-item/E1713116517/>.
- [25] Ahire, P., Lokhande, M., Patil, R., Shirsath, T., Zaware, S., Zalki, T. (2023). Intrusion detection using camera and alert management. In *2023 International Conference on Inventive Computation Technologies (ICICT)*, Lalitpur, Nepal, pp. 1117-1121. <https://doi.org/10.1109/ICICT57646.2023.10134400>
- [26] Naik, D., Jaidhar, C.D. (2022). A novel multi-layer attention framework for visual description prediction using bidirectional LSTM. *Journal of Big Data*, 9(1): 104. <https://doi.org/10.1186/s40537-022-00664-6>