JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Surveillance Video Analysis using Event Recognition

[1]K.S Suresh Babu, [2]Ishaan Bipin Dwivedi, [3]Gaurav Sanjay Mishra, [4]Ayush Rajendra Dubey,

[5]Abhishek Satish Potare

[1]Professor, Department of Computer Science, [2345]Student, Department of Information Technology
Pillai College of Engineering (Autonomous), New Panvel, Maharashtra

Abstract: Video surveillance systems are becoming increasingly prevalent in various applications such as traffic monitoring, security, and law enforcement. However, the large amounts of video data generated by these systems pose significant challenges for effective management and analysis. Video summarization using Deep learning has emerged as a promising technique for condensing long videos into shorter and more meaningful summaries, thereby reducing the time and effort required for video analysis. In this study, we propose an enhanced video summarization approach that integrates both CNN and LRCN models. Firstly, the CNN model is employed to remove static frames from the input video, thereby reducing redundancy. The processed, shortened video is then sent to our LRCN model, trained on the extensive UCF Crime dataset for event recognition. The model processes the input video and tags frames based on the detected events. This results in an output video that is significantly shortened and tagged as per the detected events, giving insights about what is actually happening in the video. Our approach offers a significant reduction in the time and resources required to review surveillance footage, without overlooking critical events. It holds promise for enhancing the efficiency of security systems and has potential applications in various domains requiring video analysis.

Keywords: Video summarization, Convolutional Neural Network (CNN), Long-term Recurrent Convolution Network (LRCN), Deep learning, Event recognition, University of Central Florida (UCF).

1. INTRODUCTION

In the realm of security and surveillance, video surveillance systems play a pivotal role in monitoring and ensuring the safety of public and private spaces. The increasing use of surveillance cameras has led to an exponential increase in the volume of video data generated, which necessitates efficient processing techniques to extract meaningful information. Traditionally, this task has been performed by human operators; however, the sheer amount of data makes continuous human monitoring impractical. This is where automated video summarization becomes indispensable.

The advent of deep learning has revolutionized the field of computer vision, offering robust solutions to challenges that were previously insurmountable. Deep learning algorithms, particularly Convolutional Neural Networks (CNNs), have demonstrated remarkable success in image and video recognition tasks, outperforming traditional machine learning approaches. These algorithms can learn hierarchical representations from data, making them highly effective in understanding the content of surveillance footage.

Our research introduces a unique approach to video summarization for surveillance systems using Event Recognition. We employ Long-term Recurrent Convolutional Networks (LRCNs), which combine the spatial feature extraction capabilities of CNNs with the temporal sequencing prowess of Long Short-Term Memory (LSTM) networks. This synergy enables our model to recognize and interpret events over extended periods, a crucial requirement for surveillance applications.

To train our CNN model, we used a custom dataset which included videos from various CCTV sources. As for the LRCN model, we utilize the UCF Crime dataset, a comprehensive collection of surveillance videos featuring a wide array of anomalous events. This dataset provides a realistic benchmark for our system, ensuring that it can generalize well across various scenarios encountered in real-world surveillance. Unlike traditional methods that focus on shortening the video duration based on solely cutting out static frames, our system provides insights on what is actually occurring in the video, thus creating a summary that encapsulates the crux of the surveillance footage.

2. LITERATURE SURVEY

Recent literature has proposed various methodologies for video summarization and event recognition, aiming to enhance efficiency and accuracy in different applications. Ashvini Tonge and Sudeep D. Thepade [1]. introduced S-VSUM, a novel approach utilizing CNN-based models for static video content summarization. In this method, the model is trained on a dataset of annotated video frames using transfer learning. The authors extract features from the intermediate layers of the CNN model and use them to represent video frames. A clustering algorithm is used to group similar frames and select representative keyframes for video summary generation. Weiping Ding et al. [2]. presented an efficient summarization technique for surveillance systems, employing feature extraction and clustering algorithms. The method used local binary pattern (LBP) and histogram of oriented gradients (HOG) for feature extraction, and the combination of unsupervised and supervised learning techniques for video summarization. Sunil S. Harakannanava et al. [3] proposed a robust algorithm for video summarization in surveillance, utilizing supervised machine learning. They utilized various low-level features like color, motion, and texture to represent each video frame. The features are then processed through a Support Vector Machine (SVM) to classify them as important or unimportant. Yingxian Chen et al. [4] introduced MGFN for weakly-supervised video anomaly detection where the network first generates a heatmap highlighting the anomaly areas in each frame, which is then used to produce an anomaly score for the entire video sequence. Ghazaala Yasmin et al. [5] proposed a key moment extraction algorithm utilizing agglomerative clustering for video summarization. Mohammad Alijanpour et al. [6] presented a method based on Twin stream CNN for event recognition, one for spatial features from RGB frames and otherfor motion features. Muhammad Zeeshan Khan et al. [7] introduced a technique employing CNN and BLSTM for video summarization with scene boundary detection. The proposed approach consists of three major components: Scene Boundary Detection (SBD), Video Summarization using CNN, and Video Summarization using BLSTM. In the SBD component, a combination of color histogram and edge-based features are used to detect scene boundaries. In the CNN component, the model extracts spatial-temporal features from video frames. In the BLSTM component, a sequence of importance scores generated by the CNN component is fed to the bidirectional LSTM to capture the temporal dependencies between frames and generate the final summary. Omar Elharrouss et al. [8] proposed a method based on motion detection for surveillance system video summarization. The proposed method utilizes a combination of background subtraction, motion detection, and keyframe extraction techniques to identify significant events in the surveillance video. Tanveer Hussain et al. [9] suggested a cloud-assisted method for multi-view video summarization utilizing CNN and Bi-LSTM networks. Lixin Duan et al. [10] introduced a new method for event recognition using Aligned Spece-time Pyramid Matching (ASTPM), to measure distance between two video clips, and Adaptive Multiple Kernel Learning (A-KML) to fuse the information from multiple pyramids and features.

S. No	Name of paper with author	Technique	Conclusion
1	S-VSUM: Static Video Content Summarization using CNN (2023) Authors: Ashvini Tonge, Sudeep D. Thepade	CNN	The proposed methods creates a subset from the video this subset is known as these frames includes the relevant in formation of input video which is used later for summary.
			The limitation of this system is that the motion component of the original video is lost during this process.
2	Efficient Video Summarization for Smart Surveillance Systems (2023) Authors: Weiping Ding, Javier Del Ser, Amir H	CNN, Clustering	The proposed method uses Histogram and Intensity values to extract low level features, which are then used to cluster frames into different groups. These clusters are used to generate summaries.
			The limitation of this approach is the lack of temporal information.
3	Robust video summarization algorithm using supervised machine learning (2022) Authors: Sunil S Harakannanava, Shaik Roshan Sameer, Vikash Kumar, Sunil Kumar Behera	Support Vector Machine (SVM)	The proposed method uses supervised machine learning algorithm, SVM to train the lowlevel features which are extracted using histogram information of the image.
			The limitation of this implementation is its dependency on input data to train the SVM.

4	MGFN: Magnitude-Contrastive Glance-and-Focus Network for Weakly Supervised Video Anomaly Detection (2022) Authors: Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi1, Yik-Chung Wu	MGFN framework	The proposed method first generates a heatmap highlighting the anomaly areas in each frame, which is then used to produce an anomaly score for the entire video sequence. The limitation of this approach is the lack of temporal information.
5	Key moment extraction for designing an agglomerative clustering algorithm-based video summarization framework (2021) Authors: Ghazaala Yasmin, Sujit Chowdhury, Janmenjoy Nayak, Priyanka Das, Asit Kumar Das	Informative frame selection, Clustering	The proposed method uses an agglomerative clustering framework. The algorithm first detects key frames based on saliency detection and then clusters them to form key moments. Agglomerative clustering algorithm may not work well for certain types of videos with complex scenes or transitions, leading to suboptimal summaries.
6	Video Event Recognition using Two-Stream Convolutional Neural Networks (2021) Authors: Mohammad Alijanpour, Abolghasem Raie	CNN	The proposed method uses two stream CNN, one for spatial features from RGB frame and other for motion features. Optical flow estimation is very computationally expensive and could limit usage of system.
7	Video Summarization Based on ListNet Scoring Mechanism (2020) Authors: WU Guangli,GUO Zhenzhou, YAO Yanpeng, LI Leiting , WANG Chengxiang	CNN, ListNet Scoring Mechanism	The proposed methods first divided the video into shots, then features extracted using CNN, which is used to train ListNet model to give score for each frame. The limitations of this approach is that it is trained on the input data. Making it prone to overfitting.
8	Video Summarization using CNN and Bidirectional LSTM by Utilizing Scene Boundary Detection (2019) Authors: Muhammad Zeeshan Khan, Saleet ul Hassan, M.A. Hassan, Muhammad Usman Ghani Khan	CNN, LSTM	The proposed method combines visual and semantic features to generate video summary. The limitations of this implementation are the type of training datasets used and the assumption that the input video is pre-segmented into shots.
9	Cloud-Assisted MultiView Video Summarization using CNN and Bi-Directional LSTM (2019) Authors: Tanveer Hussain, Khan Muhammad, Amin Ullah, , Zehong Cao, Sung Wook Baik, Victor Hugo C.,	CNN and DB-LSTM based MVS framework	The proposed method extracts visual features from multiple views of a video and combines them using a CNN-Bi-LSTM architecture to generate a summary. The proposed approach heavily relies on cloud computing resources for video processing, which may not be feasible in all practical scenarios.
10	Visual Event Recognition in Videos by Learning from Web Data" (2017)	ASTPM, A-MKL	This first proposes a new method, called Aligned Space-Time Pyramid Matching (ASTPM), to measure the distance between any two video clips. Second, a new transfer learning method, referred to as Adaptive Multiple Kernel Learning (A-MKL), in

	Authors: Lixin Duan, Dong Xu, Ivor Wai-Hung Tsang, Jiebo Luo		order to fuse the information from multiple pyramid levels and features.
			The method relies heavily on predefined features. If these features do not capture necessary information about the event, the performance could be affected.

3. OBJECTIVE

Recognizing the increasing importance of video surveillance systems in various domains such as security, law enforcement, and traffic monitoring, this project aims to develop an advanced video summarization system. By integrating CNN and LRCN models, our objective is to enhance the efficiency and effectiveness of video analysis in surveillance footage. The primary goal is to automatically condense lengthy videos into concise summaries, emphasizing significant events while eliminating redundant static frames. This system seeks to streamline the process of video review, reducing the time and resources required for analysis. Ultimately, the objective is to empower security systems with improved decision-making capabilities and facilitate better management of surveillance data in real-world applications.

4. EXISTING SYSTEM

Despite the advancements in technology, the analysis of surveillance videos is still largely a manual process. The volume of video data from widely-deployed surveillance cameras has grown dramatically, overwhelming operators who cannot afford to view or analyze even a small fraction of their collections. Current automated methods have limited applicability, often operating over small areas and struggling with dense, crowded environments. They are primarily useful for after-the-fact investigation rather than real-time analysis. Furthermore, these methods often lack object-awareness, failing to distinguish disparate objects in processing.

5. PROPOSED SYSTEM

Our system first employs a Convolutional Neural Network (CNN) to filter out static frames from the input video, effectively reducing the amount of data that needs to be processed and eliminating unnecessary details. This results in a shortened video that contains only the frames likely to contain significant events. The shortened video is then passed to a Long-term Recurrent Convolutional Network (LRCN) for event detection. The LRCN model is capable of recognizing various ‘events’ or ‘crimes’ over time, providing a comprehensive overview of the activities in the video. This two-step process not only improves the efficiency of surveillance video analysis but also provides a more accurate representation of significant events.

6. METHODOLOGY

6.1. SYSTEM ARCHITECTURE

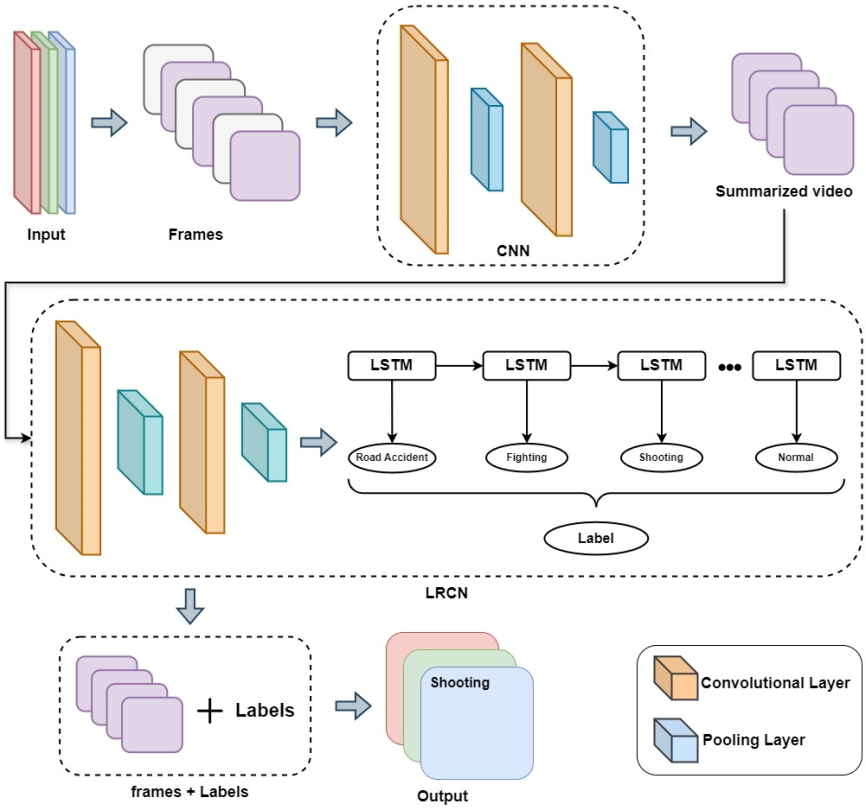

Fig 1. System Architecture

Our proposed system takes a surveillance video as input. The frames are extracted and resized to 224x224 before being passed to the ResNet50 model for feature extraction. These features are then passed to the CNN model for static frame removal. The CNN model is trained to score extracted features on a scale of 1-5. These scores are then normalized, after which frames not meeting a set threshold value are removed from the summary. This shortened summary is then passed to our LRCN model.

The LRCN model consists of a CNN which takes individual frames as input and extracts the spatial features, which are then passed to the LSTM model as input, which takes a sequence of these features as input, calculates the dependencies between features and provides respective tags to frames.

6.2. DATA ACQUISITION

The data acquisition for our project was a two-fold process, aimed at gathering diverse and representative datasets for training our Convolutional Neural Network (CNN) and Long-term Recurrent Convolutional Network (LRCN) models as per their roles in the project.

For the CNN model, we utilized CCTV footage from various sources. This footage provided a rich and diverse set of scenarios, lighting conditions, and activities. In contrast, the LRCN model was trained on the UCF 50 crime dataset. This dataset, specifically curated for crime detection, offered a comprehensive collection of video clips depicting different ‘events’ or ‘crimes’. However, the videos contained a lot of unnecessary frames which were not relevant to the event, therefore requiring intensive cleaning and trimming of videos. The model returned an accuracy of >15% on preliminary testing. The accuracy increased to 40% after the data was intensively trimmed and cleaned.

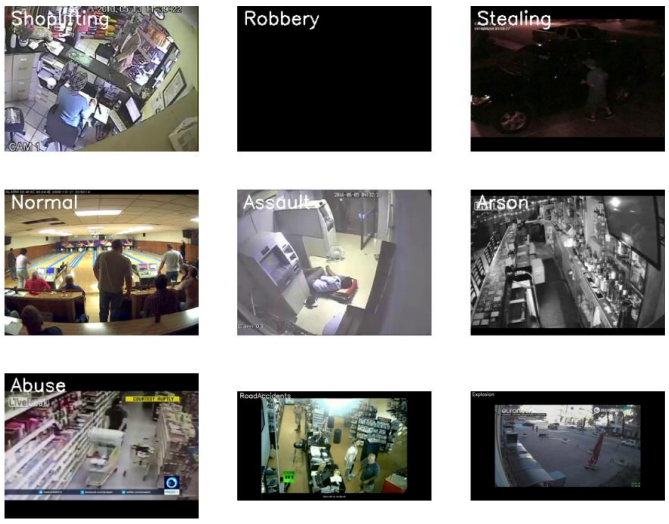

Fig 2. Various crimes/events in UCF50 Dataset

This data acquisition process ensured that our models were trained on relevant and representative data for their roles, thereby enhancing their performance and reliability.

6.3. CNN FOR STATIC FRAME REMOVAL

The structure of CNN is shown in the Fig 3. It has 3 convolution layers, 3 subsampling, 2 dropout and 2 dense layers.

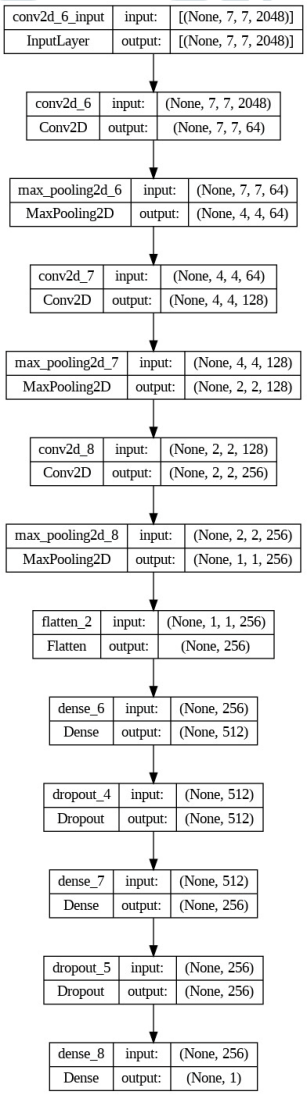

Fig 3. CNN model structure

6.4. LRCN FOR EVENT RECOGNITION

The structure of LRCN is shown Fig 4. It consists of 4 convolution layers, 4 subsampling layers, 3 dropout layers, 1 dense layer and 1 LSTM layer.

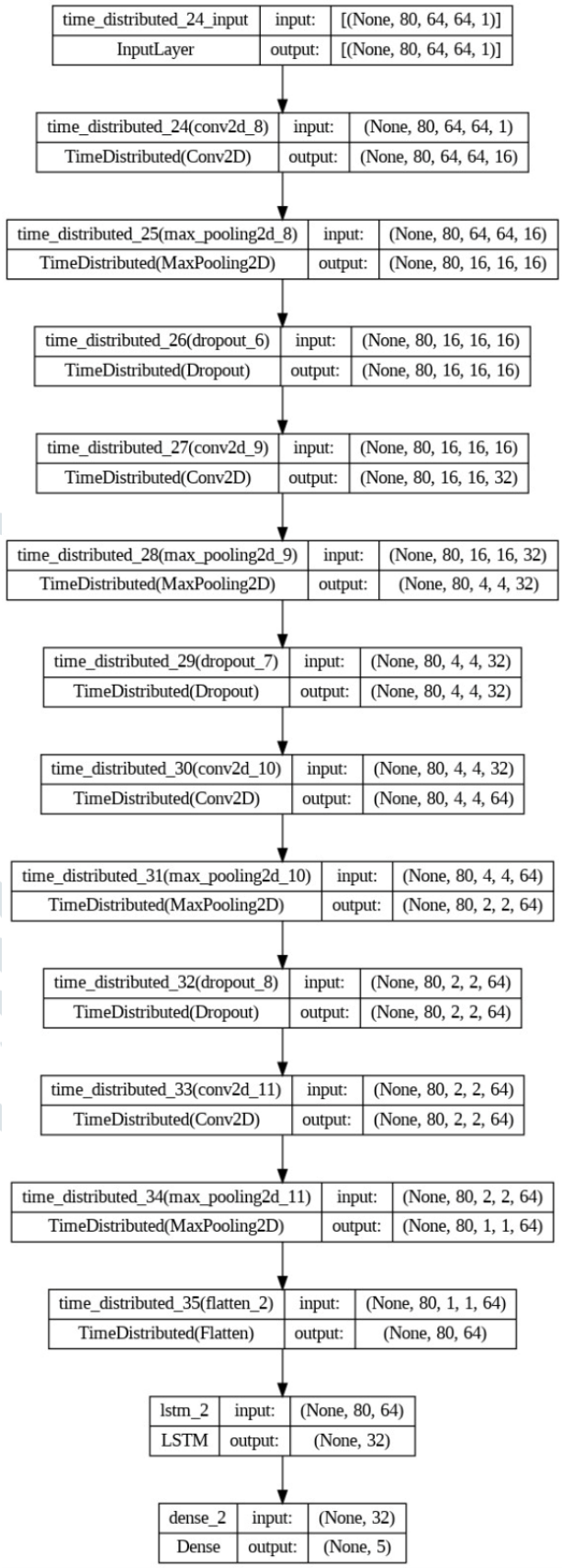

Fig 4. LRCN model structure

7. RESULTS

For our CNN model we tried various combination of Epochs (Number of samples processed an instance) and Batch size (Number of samples processed an instance). Out of these, the model with 50 Epochs and 32 Batch size, and dataset split of 80-20 [Train-Test split] performed the best with an accuracy rate of 77.5%.

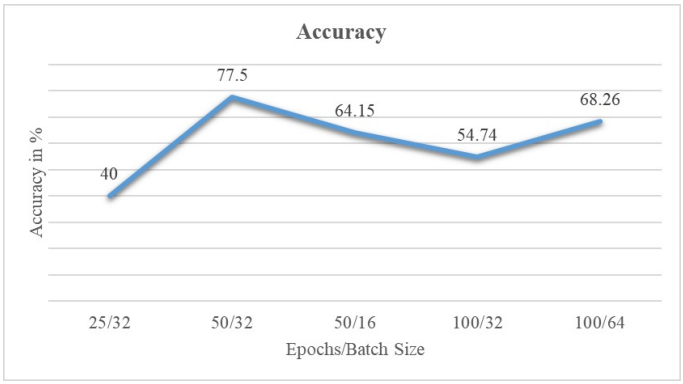

Fig 5. CNN model's accuracy over Epochs & Batch size

For our LRCN model we tried various Sequence Length (Number of frames used from each video) increasing this value can significantly increase the accuracy at the cost of increased inference time (time taken to process a single frame) while decreasing this value can decrease the inference time as the cost of accuracy. In our testing we found out that Sequence Length of 40 gave the best result with an accuracy of 40%.

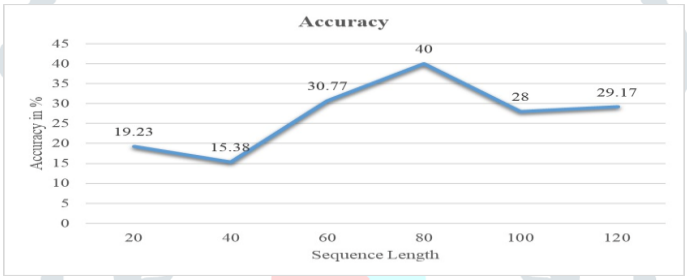

Fig 6. LRCN model's accuracy over Epochs & Batch size

SCRENNSHOTS

Fig 7.1 Input video

Fig 7.2 Output video

The input video was 41 seconds long and the summarized output was 6 seconds long.

Fig 8.1 Input video

Fig 8.2 Output video

The input video was 14 minutes 35 seconds long and the output summary was 2 minutes and 24 seconds long.

8. CONCLUSION

In conclusion, the development and implementation of the dual-model approach for surveillance video analysis represent a significant advancement in video summarization and activity detection. The first model, based on Convolutional Neural Networks (CNNs), effectively identifies and removes static frames from long surveillance videos, thereby generating concise summaries containing only frames with notable actions, achieving an impressive accuracy level of 77.5%. This capability not only reduces the storage requirements for video archives but also streamlines the process of reviewing video content by eliminating the need to manually sift through uneventful frames.

Furthermore, the integration of a second model, which combines CNNs with Long Short-Term Memory (LSTM) networks, enhances the system's ability to detect specific activities within the selected frames. Despite a moderate accuracy rate of 40%, this model successfully categorizes actions such as shooting, road accidents, fighting, and normal activities, providing valuable insights into the nature of events captured in the summarized videos.

Overall, the combined utilization of these models offers a practical solution for optimizing video surveillance operations by significantly reducing the time and effort required for video analysis while maintaining a reasonable level of accuracy in identifying critical activities. Future enhancements could focus on refining the activity detection model to improve accuracy and expanding the range of identifiable actions, ultimately enhancing the utility and effectiveness of the surveillance video analysis system.

9. FUTURE SCOPE

The current project demonstrates promising advancements in surveillance video analysis; however, there are several areas that could be explored to enhance the system's performance and applicability:

- Explore Multi-Modal Fusion - Investigate integrating audio with video frames for a comprehensive understanding of surveillance footage using multi-modal fusion techniques.
- Enhance Activity Detection Accuracy - Implement strategies to mitigate overlapping class predictions from the CNN + LSTM model (LRCN) by refining the model architecture or using post-processing techniques such as class-specific thresholds or ensemble methods.
- Optimize for Real-Time Processing - Implement optimizations like GPU acceleration and lightweight architectures to enable real-time video analysis.
- Expand Activity Classes with Dataset Expansion - Increase dataset diversity to include more annotated examples beyond basic classes (e.g., normal, road accidents, fighting, shooting, explosion) to enable the CNN + LSTM model (LRCN) to learn and recognize a broader range of activities for improved surveillance video analysis.

10. REFERENCES

[1] Ashvini Tonge, Sudeep D. Thepade, “S-VSUM: Static Video Content Summarization using CNN” (2023)
[2] Weiping Ding, Javier Del Ser, Amir H. Gandomi, Victor Hugo C. De Albuquerque, Tanveer Hussain, Kham Muhammad, “Efficient Video Summarization for Smart Surveillance Systems” (2023)
[3] Sunil S Harakannanava, Shaik Roshan Sameer, Vikash Kumar, Sunil Kumar Behera, “Robust video summarization algorithm using supervised machine learning” (2022)
[4] WU Guangli,GUO Zhenzhou, YAO Yanpeng, LI Leiting, WANG Chengxiang, “Video Summarization Based on ListNet Scoring Mechanism” (2020)
[5] Ghazaala Yasmin, Sujit Chowdhury, Janmenjoy Nayak, Priyanka Das, Asit Kumar Das, “Key moment extraction for designing an agglomerative clustering algorithm-based video summarization framework” (2021)
[6] Mohammad Alijanpour, Abolghasem Raie, “Video Event Recognition using Two-Stream Convolutional Neural Networks” (2021)
[7] WU Guangli, GUO Zhenzhou, YAO Yanpeng, LI Leiting, WANG Chengxiang, “Video Summarization Based on ListNet Scoring Mechanism” (2020)
[8] Muhammad Zeeshan Khan, Saleet ul Hassan, M.A. Hassan, Muhammad Usman Ghani Khan, “Video Summarization using CNN and Bidirectional LSTM by Utilizing Scene Boundary Detection” (2019)
[9] Tanveer Hussain, Khan Muhammad, Amin Ullah, , Zehong Cao, Sung Wook Baik, Victor Hugo C, “Cloud-Assisted Multi-View Video Summarization using CNN and Bi-Directional LSTM” (2019)
[10] Lixin Duan, Dong Xu, Ivor Wai-Hung Tsang, Jiebo Luo, “Visual Event Recognition in Videos by Learning from Web Data” (2017)
[11] https://bleedaiacademy.com/human-activity-recognition-using-tensorflow-cnn-lstm/