SecureSpace: Automated Surveillance Alert System

Bahir K N, Tharani R, Jarsha Maria, Ridha Isha
Department of Information Technology
Rathinam College of Arts and Science
Tamil Nadu, India
{bahirofficial, tharaniraja2003, jerrylava100, ridha.isha20}@gmail.com

***Abstract*— A surveillance and Alert System is used for detecting the Violence Activity and fire spread that takes place in the surveillance environment and alerts the concerned authority. Today, the most fundamental impactful security feature that a premise can have is CCTV surveillance. Being the most well-known method of stopping and identifying undesirable activities, it is present in various premises. Like, A college campus with CCTV in a variety of structures, such as dormitories, classrooms, canteen, sports complex, auditorium, etc. Even if the incident had already occurred, manually searching for it in the recorded video is a time-consuming process. The aim of the project is to create a Long-term recurrent convolutional network (LRCN) system for the environment (for example the academic campus) to monitor CCTV surveillance and detect violent activity like — fighting, and fire incidents. The system can be used to create an alert message via mail with the video of the detected activity which will be sent to the respected authority.**

Keywords—Violence Detection, Alert Generation, LRCN, Fire detection, Non-violence

I. INTRODUCTION

Recognizing human actions in surveillance films has evolved into an active and forward-thinking research subject in computer vision and machine learning. The classification of visual content data is based on human activity, which exhibits general human behavior. Human behavior and activities are understood using several video features that categorize events as normal or abnormal. Walking, running on the ground, eating food, sitting down and rising from a chair, reclining in bed, choosing an object from a table or floor, and descending stairs are all examples of regular activities. Normal human actions are deviated from abnormal behaviors, often known as suspicious activities. The activities are inappropriate in one context but might be regarded as normal in another. Running in a playground, for example, is regarded as normal, whereas running in a bank or a marketplace is viewed as abnormal. The most significant unusual activities are those that physically show actions to inflict hurt or damage with aggressive behaviors. Fighting, murdering, and beating someone are the most typical forms of public violence Violent behaviors are manually monitored in a semi-automated method via the monitor screen of a security camera. This is not advantageous since continual monitoring is necessary, yet continuously watching screens to spot aggressive behaviors is tough. There is no room for laxity when it comes to monitoring such actions because they might occur at any time. Such semi-automated systems must be transformed into fully automated intelligent systems capable of detecting and recognizing violent behaviors without human supervision. Fully automated systems can identify human activity using computer vision and machine learning and are more successful and efficient than semi-automated systems in detecting object motions and recognizing human activity. Because of variables such as real-time categorization, the low video quality of security cameras, and unpredictable light intensity during monitoring, human activity detection is a tough process.

CCTV surveillance systems have developed into a forensic tool, which is used to gather evidence after an incident has occurred, as recording and storage technology and software like video analytics have become more effective. However, as CCTV surveillance systems become more easily integrated with monitoring devices, alarm systems, and access control devices, a third application of CCTV is growing rapidly i.e. assisting security personnel in detecting and interrupting security problems as they occur, or even before they occur.

II. RELATED WORKS

Using many different techniques, the challenge of violent action recognition at a distance has been achievable in the last few years. Although the initial attempts produced positive results, there are still a few limitations to these methods. For violence detection systems, both machine learning and deep learning methods have been implemented over the years to get efficient results. Some of these involve:
Abnormal activities (Violence and Fire) Detection through Machine learning techniques
[1] The proposed framework included a Bag-of-Words framework used for action recognition along with two action descriptors: STIP and MoSIFT on a new video database containing 1000 sequences divided into two groups: fights and non-fights. The experimental results were able to detect fights with 90.
[2] The approach was evaluated using two publically accessible datasets utilizing motion boundary SIFT (MoBSIFT) and movement filtering modules. To alleviate the temporal complexity posed by the MoBSIFT approach, the authors switched the optical flow estimation to dense optical flow estimation for the full frame once, removing the difference of Gaussian (DOG) pyramid-based flow estimation.
[9] Machine learning approach was proposed based on a Support Vector Machine (SVM), to detect if a human action, captured on a video is or not violent. Using a pose estimation

algorithm, it focuses mostly on feature engineering, to generate the SVM inputs. In particular, hand-engineered a set of input features based on key points (angles, velocity, and contact detection) and used them, under distinct combinations, to study their effect on violent behavior recognition from video. Overall, an excellent classification was achieved by the best-performing SVM model, which used key points, angles, and contact features computed over a 60-frame image input range.

[11] This paper describes the movement sensors and a microphone collect motion and speech data, which is used to extract a set of motion and audio attributes that may be used to identify bullying instances from everyday life events. Time-domain and frequency-domain motion characteristics are retrieved, while audio features are calculated using traditional MFCCs. The wrapper technique is used to implement feature selection. These motion and audio characteristics are then combined to generate combined feature vectors for classification, and LDA is used to reduce dimension further. A BPNN is taught to detect bullying and separate it from regular daily life activities. In addition, the authors offer an action transition detection approach to decrease computing complexity for practical use. As a result, the bullying detection system will only be activated when an action transition event is observed. The combined motion-audio feature vector surpasses isolated motion and acoustic features in simulation, obtaining an accuracy of 82.4 and a precision of 92.2. Furthermore, the action transition approach can cut computing costs in half.

[12] The authors provide a cascaded technique of detecting violence based on motion boundary SIFT (MoBSIFT) and movement filtering. The surveillance movies are examined using a movement filtering algorithm based on temporal derivative in this technique, which prevents most peaceful acts from passing through feature extraction. Only filtered frames may be used for feature extraction. Motion boundary histogram is retrieved and coupled with scale-invariant feature transform (SIFT) and histogram of optical flow feature to generate MoBSIFT descriptor. Because of its great tolerance to camera motions, the experimental findings reveal that the proposed MoBSIFT beats the existing approaches in accuracy. The use of movement filtering in conjunction with MoBSIFT has also proven to minimize time complexity.

Abnormal activities (Violence and Fire) Detection through Deep learning techniques

[3] The suggested model is a U-Net-like network that extracts spatial features and then utilizes LSTM for temporal feature extraction and classification. Five folds of cross-validation were used, with three different datasets: Hockey Fights, Movie Fights, and RWF-2000. Experiments using a dataset based on RWF-2000 revealed an average accuracy of 0.82 2 and an average precision of 0.81 3.

[4] The solution uses an architecture in the Fast-RCNN style that has been temporally extended. First, using motion appearance (dynamic pictures), tracking algorithms, and pre-trained person detectors, a spatiotemporal suggestion (action tubes) is constructed. Then, using deep neural networks, spatiotemporal information from an input video and the action suggestions is being extracted. The classification of each action as violent or not violent is done by training a classifier based on multiple-instance learning. With the help of three open datasets, Hockey Fight, RLVSD, and RWF-2000, an accuracy of 97.3, 92.88, and 88.7, is achieved respectively.

[5] Another proposed framework uses a lightweight CNN architecture for detecting anomalies in the actions found in videos. The three categories of abnormal behaviors are falling suspicious action, and violence. To achieve high classification performance, the proposed framework adopts SG3Is (stacked grayscale 3-channel images) to train a lightweight CNN. SG3Is provide a potent alternative to classical methods of motion representation, such as optical flow and dynamic images. The experiments on UR Fall, Avenue, Mini-Drone Video, and Hockey Fights datasets show that the proposed framework can efficiently detect various anomalies found in these datasets with accuracies of 98.86, 95.28, 95.81, and 99.74, respectively.

[6] The proposed model is a combination of ResNet and ConvLSTM used for anomaly detection from surveillance cameras. Convolutional Neural Networks (CNN)and Recurrent Neural Networks (RNN) structure. CNN to extract essential features from each input video data frame. Whereas (RNN) helps investigate a series of frames to find any abnormal events. Implement the Residual Networks (ResNets) part of CNN. Then used Convolutional LSTM (ConvLSTM) as a recurrent network (RNN) to find the anomalies in our video dataset. The video file is divided into n frames, the difference between each frame is the input to the CNN (i.e., ResNet50). The output of the ResNet50 will then go to the RNN (i.e. ConvLSTM). After this process is done for all n frames, the output goes to a max-pooling layer followed by several fully connected layers to get final results. The proposed model was implemented on the UCF-Crime dataset to compare experimental results with other methods applied to evaluate how well our model works. AUC and Accuracy metrics were used for evaluation.

[7] Here it presents an intelligent fire/smoke detection approach based on the YOLOv2 network, which aims to achieve a high detection rate, low false alarm rate, and high speed. Then detector is deployed on a single-board embedded system, NVIDIA Jetson Nano, as a standalone application for real-time video processing. CNN-based fire detection based on a pre-trained VGG16 and Resnet50 as baseline architecture. YOLOv2 algorithm helped to identify and locate fire and smoke objects using a video camera. The proposed model was a Deep Neural Designer tool in MATLAB to build YOLOv2 neural network layers. The model is constructed using CNN with 21 layers which were compared to the other object detectors such as R-CNN and Fast R-CNN. MATLAB was used with our bench-test dataset of fire and smoke videos. These 3 were run simultaneously resulting in YOLOv2 25 times faster than R-CNN and 23 times than the Fast R-CNN object detectors.

[8] The Convolutional Neural Network (CNN) models have been evaluated with the proposed MobileNet model. The MobileNet model has been contrasted with AlexNet, VGG-16,

and GoogleNet models. The simulations have been executed using Python from which the accuracy of AlexNet is 88.99 and the loss is 2.480. The accuracy of VGG-16 is
96.49 and the loss is 0.1669, The accuracy of GoogleNet is 94.99 and the loss is 2.92416. The proposed MobileNet model accuracy is 96.66 and the loss is 0.1329.

[10] It presents three deep learning-based models for violence detection and tests them on the AIRTLab dataset. The author proposed transfer learning-based models which are C3D combined with an SVM classifier and C3D combined with new fully connected layers. It helped get stable accuracy results on all three tested datasets, being better, in many cases, than the related works tested on the Hockey Fight and Crowd Violence. Thereby suggesting going with transfer learning-based models for the task of violence detection. The models based on 3D CNNs perform better than well-known 2D CNNs pre-trained on ImageNet and combined with a recurrent module to extract the spatiotemporal features of the videos in the datasets, suggesting continuing the research about 3D architectures for violence detection. Moreover, all the proposed models demonstrated more capability of identifying violent videos than non-violent, given that most of the errors are false positives. Whilst this behavior is partially affected by the fact that the samples from the two classes are unbalanced, it also validates the design of the AIRTLab dataset in checking the robustness against false positives.

III. PROPOSED MODEL

The workflow of our activity recognition model is depicted in Figure 1. First, we gathered video data from Kaggle. Before partitioning the video dataset into training and testing data, various preprocessing techniques such as resizing, extracting frames, and normalizing data are applied. On the processed data, we trained the LRCN model. For activity recognition, our model may upload testing video data from YouTube and can be live video to detect Violence, Fire, and Non-violence

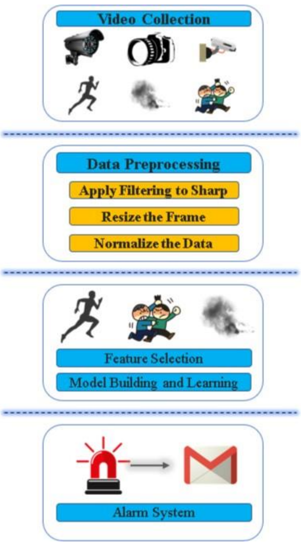


Figure 3.1: Use case Diagram

A. Dataset Collection

We have used the Real Life Violence Situations Dataset to train our model for Violence and Non-violence data. The dataset contains 1000 Violence and 1000 non-violence videos containing many real street fights situations in several environments and conditions. The average duration of the videos is 5 seconds and the majority of these videos are from CCTV footage. To train the model with the fire, the FIRESENSE dataset was used.

B. Model Training

A Long-Term Recurrent Convolutional Network (LRCN) is used. This approach combines the CNN and LSTM layers into a single model for image and video description. The gathered spatial information is sent to an LSTM layer for the Convolutional layers-based temporal sequence modeling at each time step. The network learns spatiotemporal features in this manner during an end-to-end training session, leading to a robust model. We have an LRCN network with four input nodes as shown in Figure 3.1.

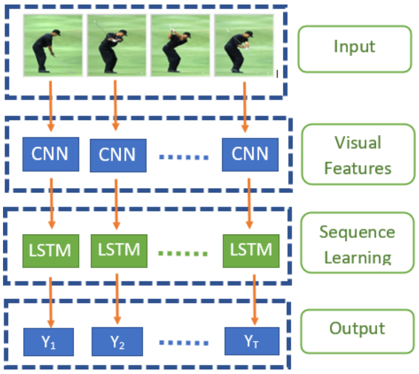


Figure 3.2: Long-Term Recurrent Convolutional Network

C. Architecture

Footage from surveillance is broken down into frames. These frames are given as input to the LRCN model to detect Violence, Fire, and Non-violence activities in the given sequence of input frames. If not violent the frame is discarded. If fire or violence is detected, the input footage is sent along with a mail alert to concerned authorities.

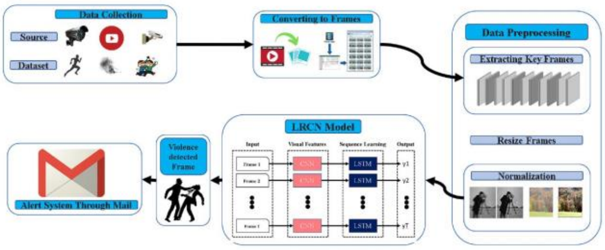


Figure 3.3: High-level architecture diagram

IV. RESULTS AND DISCUSSION

We provide a model for surveillance and alert that works efficiently. An intriguing objective for future research would be to identify action sequences that lead to the initiation of the violence. It is critical to detect violence in video data streams.

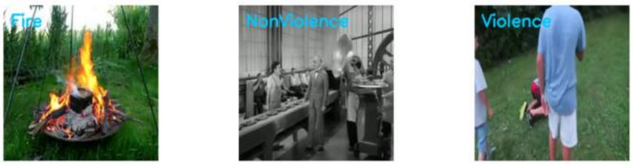
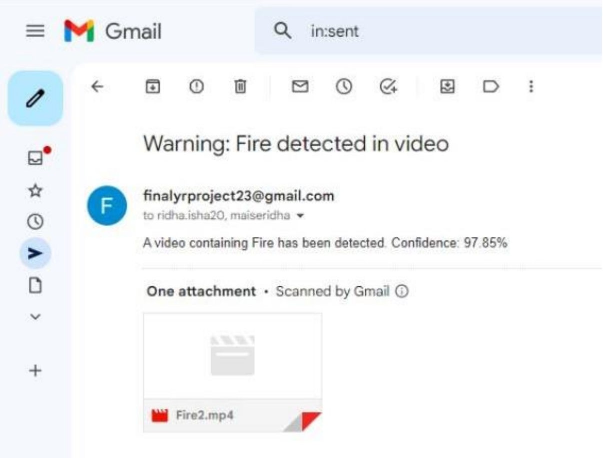


Figure 4.1: Action Recognition

To detect suspicious activity, The model architecture is defined. A basic LRCN model with four CNN layers followed by an LSTM layer is created. For model construction, a sequential model is used. Then the model summary is displayed. The constructed LRCN model is returned. In Figure 4.1 it can be seen that the model can identify Non-violence, Violence, and Fire

Further, the accuracy of the model is also evaluated for the better results. The model achieved an accuracy of about 83 percent, which is not bad.

For the alert system, we are using the library like smtplib and Email Message. By defining the user and password the message can be sent through the defined email to the targeted email address, the SMTP server is used to define the email. After using the if clause which contains the targeted email and the message that is needed to be sent. Then, the mail is sent along with the video that is tagged by the detection as seen in Figure 4.3



Warning: Fire detected in video

finalyrproject23@gmail.com
to ridha.isha20, maiseridha

A video containing Fire has been detected. Confidence: 97.85%

One attachment • Scanned by Gmail

Fire2.mp4

Figure 4.3: Alert through Gmail

V. FUTURE SCOPE

The future of violence and smoke detection with the LRCN model is broad and promising, and as technology advances, the potential applications are likely to continue expanding, providing a safer environment for individuals and communities.

1. Improving accuracy: The accuracy of CNNs and LRCNs in violence and fire detection can be further improved by incorporating more diverse training data and fine-tuning the models for specific use cases.

2. Integrating additional information: Combining other sources of information such as audio or text data with video data can enhance the performance of violence detection models.

3. Handling real-world scenarios: The models need to be tested and improved in real-world scenarios to handle various lighting conditions, camera angles, and background distractions.

4. Real-time implementation: The implementation of these models in real-time systems for surveillance and public safety can greatly benefit from advancements in edge computing and low-latency processing.

5. Explaining predictions: The interpretability of violence and fire detection models needs to be improved to gain more confidence in the predictions made by the systems. This can help in reducing false alarms and improving overall performance.

6. The LRCN model can also be used in conjunction with other sensors such as carbon monoxide detectors and temperature sensors to provide a complete fire warning system. This can enable rapid evacuation and intervention and reduce the damage caused by fires.

7. The LRCN model can be used with unmanned aerial vehicles (UAVs) to detect smoke and other fires in remote areas, allowing for early detection of forest fires and more efficient monitoring of large areas.

Continued research and development in these areas will further advance the capabilities of CNNs and LRCNs in violence and fire detection and enhance their potential to promote public safety.

VI. CONCLUSION

With the aim of expanding the range of our Smart Surveillance Systems solutions, we are developing a real-time violence and fire detection technology that can be integrated with a surveillance system. Its primary function is to ensure public safety through visual crowd surveillance, so any violent activity generates automatically an alert is generated. The violence detection module can be incorporated into the surveillance systems installed in airports, schools, parking lots, prisons, shopping malls, and other indoor and outdoor public access areas.

An efficient surveillance system is a large-scale computer vision, data analysis, and decision-making challenges. Hence the 'smart' approach to extracting information from the surveillance footage employs the synergy of several innovative technologies that power the next key components of the Smart Surveillance Systems.

ACKNOWLEDGMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible. Primarily, we thank "THE ALMIGHTY" for the strength and encouragement have given to us throughout all the challenging moments of this work. Profound gratitude and unwavering respect to the following captioned persons for the help and support extended to making this project possible.

We thank Dr. P.K.A. Chitra, M.E., Ph.D., Project Coordinator and all faculty members of Department of Information Technology, Rathinam Campus, iNurture Education Solution pvt ltd for their constructive suggestions. We convey special thanks to our supervisor Dr. P.K.A. Chitra, M.E., Ph.D, for her inestimable support, guidance, and motivation.

We dedicate sincere respect to our parents for their moral motivation in completing the project.

REFERENCES

[1] Bermejo Nievas, E., Deniz Suarez, O., Bueno Garc´ıa, G., Sukthankar, R. (2011).

[2] Violence Detection in Video Using Computer Vision Techniques. In: Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., Kropatsch, W. (eds) Computer Analysis of Images and Patterns. CAIP 2011. Lecture Notes in Computer Science, vol 6855. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23678- 5

[3] Zhang, T.; Yang, Z.; Jia, W.; Yang, B.; Yang, J.; He, X. A new method for violence detection in surveillance scenes. Multimedia. Tools Appl. 2016, 75, 7327–7349.

[4] A Machine Learning Approach to Detect Violent Behaviour from Video.

[5] A Combined Motion-Audio School Bullying Detection Algorithm Liang Ye, Peng Wang, Le Wang, Hany Ferdinando, Tapio Sepp¨anen and Esko Alasaarela.

[6] Febin, I.P., Jayasree, K. Joy, P.T. Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm. Pattern Anal Applic 23, 611–623. https://doi.org/10.1007/s10044-019-00821-3

[7] Vijeikis, R.; Raudonis, V.; Dervinis, G. Efficient Violence Detection in Surveil-

[8] lance. Sensors 2022, 22, 2216. https://doi.org/10.3390/s22062216

[9] Choqueluque-Roman, D.; Camara-Chavez, G. Weakly Supervised Violence Detec- tion in Surveillance Video. Sensors 2022, 22, 4502. https://doi.org/10.3390/s22124502

[10] Mehmood, A. LightAnomalyNet: A Lightweight Framework for Efficient Abnor- mal Behavior Detection. Sensors 2021, 21, 8501. /doi.org/10.3390/s21248501

[11] Vosta, Soheil, and Kin-Choong Yow. 2022. "A CNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras" Applied Sciences 12, no. 3: 1021. https://doi.org/10.3390/app12031021

[12] Saponara, S., Elhanashi, A. and Gagliardi, A. Real-time video fire/smoke detection based on CNN in antifire surveillance systems. J Real-Time Image Proc 18, 889–900(2021). https://doi.org/10.1007/s11554-020-01044-0