International Journal of Electrical and Electronics Engineers
Volume 18, Issue No. 01, Jan-June 2026
ISSN (0) 2321-2055 ISSN (P) 2321-2045

Towards Fast and Accurate Violence Detection for Automated video Surveillance Applications

B. Mallikharjuna Reddy, A. Siva Karthik, B. Gangothri, and K. Adithya Sai
Department of ECE, Tirumala Engineering College, Narasaraopet, AP-522601

***Abstract*—The present work focuses on the problem of detecting real-time violence in large-scale surveillance environments, which are usually associated with limited coverage, delayed detection, and extensive use of human observers in traditional fixed camera-based systems. This paper introduces a novel drone-based, edge-enabled surveillance system utilizing a lightweight MobileNet-LSTM model along with optical flow motion analysis for spatio-temporal feature extraction. The drone-based system collects video streams from a bird's eye view and uses on-device processing with the help of Raspberry Pi. The proposed system, in contrast to typical convolutional neural network-based frame-based models, exploits temporal information and motion intensity for better detection of aggression. On the well-known dataset, the proposed system achieves the accuracy, precision, recall, and F1-scores of 88.1%, 82.05%, 96%, and 88.5% respectively. The high recall score reveals the effectiveness of the proposed system in detecting violent acts, which is a crucial factor for safety-focused applications. Moreover, by exploiting aerial mobility, the proposed approach minimizes surveillance gaps.**

***Index Terms*—Violence Detection, Edge AI, Drone Surveillance, MobileNet-LSTM, Optical Flow, Human Activity Recognition, Real-Time Video Analytics**

I. Introduction

Maintaining safety and security in large environments like universities and public places has become progressively hard owing to the increased density of people and interaction between them. The surveillance mechanisms, which are usually composed of CCTV cameras fixed in various places, depend excessively on manual observation by people, which is a process that can be inefficient because of fatigue and leads to late reactions in case of emergencies. What is more, the static nature of these cameras poses a risk of missing out on violent events like fighting.

The recent developments in computer vision and deep learning technology allow the use of automated systems that can analyze video recordings to detect human actions. The majority of methods that have been proposed up to now utilize frame-wise CNNs whose operation is based on the extraction of spatial information, failing to properly model the dynamics of temporal behavior. Thus, it becomes difficult for such models to differentiate between regular crowd behaviors and violence. While 3D CNN models and architectures can solve the problem, they bring considerable computational costs. To overcome these shortcomings, this paper suggests an intelligent drone-based approach for violence detection using edge computing. Our framework uses a light-weight MobileNet-LSTM model for capturing both static and dynamic attributes in addition to analyzing optical flow between two consecutive images for modeling the intensity of the motion. This approach allows us to deploy our model to a Raspberry Pi running on a drone for inference in real time without dependency on any cloud infrastructure, thus minimizing latencies and increasing scalability.

The design of the proposed architecture exploits the synergy between mobility based on drones and intelligent video analysis based on edge computing technologies to compensate for certain weaknesses of conventional surveillance systems. Leveraging lightweight deep learning models, spatio-temporal modeling, and motion analysis techniques, the proposed model is able to function effectively under limited computational resources but with consistent detection performance. On-device computation ensures independence from cloud services and, therefore, guarantees lower latency and real-time operation. In addition, the aerial view provides a greater field of vision and excludes potential blind spots.

II. Literature Survey

The recognition of human actions and detection of violence from surveillance videos has been intensively investigated via feature-based techniques and deep learning algorithms. First introduced by Ivan Laptev et al. [1], spatio-temporal interest points were used to obtain the motion and appearance features from video sequences. Though efficient for detecting basic human actions, such feature-based approaches are prone to noises and cannot adapt to various environmental conditions. Another study by Louis Kratz and Nishino [2] developed an anomaly detection technique using spatio-temporal patterns in surveillance videos. The proposed algorithm is able to detect abnormalities in crowded scenarios; however, it lacks the capability to discriminate between normal behaviors and violent actions. Due to improvements in deep learning techniques, convolutional neural network (CNN) is being used for automatic feature extraction. In their work titled "FlowNet: An End-to-End Architecture for Learning Optical Flow," Alexey Dosovitskiy et al. [3] have introduced a deep neural network based solution for computing optical flow, thereby providing a superior motion model than previous solutions. However, although FlowNet focuses on motion computation, it fails to directly classify the activities taking place. Subsequently, Ji Li et al. [4] introduced 3D Convolutional Neural Networks (3D-CNN) that jointly extract temporal and spatial information for violence detection tasks. This method is able to provide better results but requires high computational power making it impractical for implementation at the edge device. Recently,

researches have been done to enhance efficiency and temporal modeling capabilities in video recognition. Christoph Feichtenhofer et al. [6], [8] developed SlowFast and X3D models that can detect slow semantics and fast motion components to deliver excellent results in video classification tasks. Nevertheless, these models need GPU-based computing power, which makes it impossible to deploy them on embedded edge devices. In addition, Kaidi Cao et al. [9] proposed temporal alignment methods in few-shot video classification for better generalization from few samples; however, the method needs significant computational power. Ming Cheng et al. [10] provided a massive dataset for violent content detection for better training of machine learning models, while Paolo Sernani et al. [11] performed benchmarking experiments using deep learning models, proving superior detection accuracy. However, all these solutions require centralized computation and cannot respond in real-time during deployment. Simultaneously, anomaly detection techniques such as the memory-guided system suggested by Hyunjong Park et al. [7] and explainable models created by Stanislaw et al. [12] emphasize the importance of learning about normal behavior patterns and anomalies. Although these techniques improve interpretability and anomaly detection performance, their design does not allow them to detect violence-related anomalies efficiently. In recent years, efforts made toward detecting violent activities have been directed at enhancing learning efficiency while minimizing reliance on large amounts of labeled data. Choqueluque-Roman and Camara-Chavez [13] suggested using a weakly-supervised learning scheme for violence detection in surveillance videos. Although such techniques can help in decreasing the amount of required annotation, they can be less accurate owing to insufficient supervision. Likewise, Aremu et al. [15] developed a model to detect weaponized violence by analyzing salient image features; nonetheless, the method does not address motion cues. Lastly, Vijeikis et al. [16] developed an efficient framework for violence detection, tailored for surveillance tasks, and offering superior computational speed, albeit lacking in adaptability for edge deployments. As can be seen from the literature reviewed above, all of these approaches are either highly complex in terms of computations or unable to accurately represent the dynamic characteristics of motion in a lightweight manner. In addition, most of these approaches are tailored toward offline or GPU-enabled applications, leaving out the issue of real-time implementation and limited processing power of the drones. Hence, in order to mitigate these shortcomings, a new lightweight approach that uses Optical Flow and MobileNet-LSTM architecture is introduced in this paper.

III. Objectives

The key focus of the research is to provide an algorithm and implementation of the real-time, edge-enabled framework for detection of violence using a drone. The main idea behind designing the system is to unite the capabilities of an airborne vehicle (i.e., its aerial mobility with embedded intelligence for acquiring and analyzing the stream of surveillance footage). The goals of the project include development of a computationally-efficient hybrid architecture for learning spatiotemporal patterns, based on the fusion of MobileNets for spatial representation and LSTM networks for modeling temporal data sequences. Another goal is to introduce motion estimation via optical flow, aimed at capturing the intensity and orientation of inter-frame motions for the purposes of improving sensitivity to aggressive patterns. At the same time, optimization of the architecture will be required to ensure that it runs on limited resources of the target device, thus being deployable in real-time without dependency on cloud computing services. Also, the project aims at developing efficient techniques for making decisions based on the outputs generated by the proposed model. They may include, for example, threshold-based classification and temporal averaging for mitigating instability and reducing false positives. In addition, the project requires implementation of event triggering and

IV. Proposed Methodology

This approach proposes a framework for detecting violent activities through the implementation of an architecture that utilizes the power of drone-based video capturing technology along with edge computing for deep learning. In particular, the proposed architecture is a pipeline-based approach that analyzes video streams obtained from a drone-based camera in real time, and conducts inference using a light model. With the help of drones, the system gains the capacity of surveillance in any environment where static cameras cannot work well because of possible blind spots in such cases. First, the video streams are segmented into consecutive frames and are processed by the means of various pre-processing techniques, including frame resizing to a particular size, normalization of frame pixels, and noise reduction. Such approaches increase the quality of features that may be used in the analysis process. The task of feature extraction is addressed by a lightweight CNN inspired by the MobileNet model that learns discriminating spatial representations at the level of single frames. MobileNet employs depthwise separable convolutions that help to drastically lower the computational complexity of the model and thus make it applicable for execution on low-end edge devices. The spatial features obtained with the use of convolutional operations are further used as input to an LSTM network that analyzes temporal correlations within frame sequences and describes the dynamic changes of human activities. Given the nature of violence, which implies quick and chaotic motions, the analysis of temporal relations is particularly important for classifying violent actions. In order to increase sensitivity to motion, the optical flow algorithm is applied to extract the displacements of individual pixels between neighboring frames and get explicit information regarding the speed and direction of motion. The final classification step involves using the extracted features as inputs and feeding them to fully connected layers and sigmoid activation functions to produce a probability score indicating whether there is any violent behavior. The decision process involved a threshold-based approach that categorizes the input video frames into violent and non-violent categories and uses temporal smoothing techniques to minimize

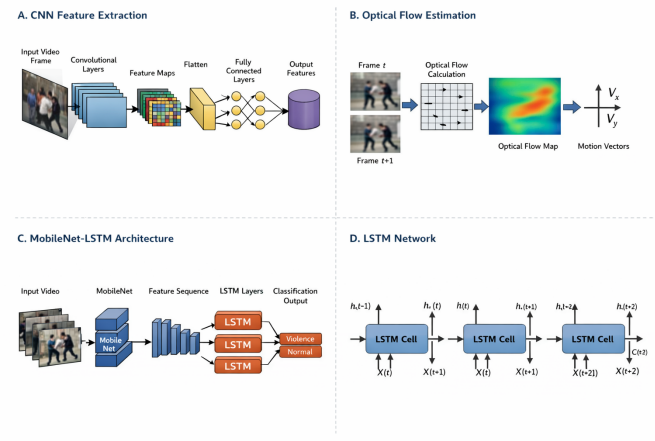

Fig. 1. System Architecure

variations in subsequent frames. The proposed methodology is implemented in a Raspberry Pi device embedded in the UAV, eliminating any need for cloud computing resources and minimizing communication overheads. If the input sequence frame is classified as having any violent activities, the system automatically triggers an alarm, captures the video clips and saves them in a storage device for future use while ensuring real-time video feeds are transmitted to a human observer at the base station.

A. Model Architecture

The entire design of the proposed solution, shown in Fig. 1, includes spatial feature extraction, motion estimation, and temporal modeling for successful violence detection. First of all, each single video frame is subjected to the process of convolutional neural network (CNN) operations that allow for the extraction of high-level spatial features related to human posture and contextual information from the scene. At the same time, optical flow is estimated between the subsequent frames that enables us to understand motion intensity and direction since it is important for aggressive behavior identification. Further on, the spatial feature vectors are arranged into a sequence and used as input for a MobileNet-LSTM neural network architecture, wherein MobileNet performs feature extraction in an efficient manner while LSTM layers are used to build the temporal model of the data. After passing through the LSTM cells, the model is able to understand behavior dynamics. Finally, the classification layer takes spatio-temporal and motion-aware features as its input to decide whether the behavior is violent or not.

B. Working Principle

The working mechanism of the suggested system depends on acquiring and analyzing video streams in real time to identify violent behavior by employing spatial and temporal deep learning and motion analysis. In terms of operation, this system functions through acquiring real-time video footage through the drone camera. It captures live aerial video streams of the surroundings, making its surveillance capabilities dynamic. The video stream is analyzed in real time by conducting frame-by-frame processing of the video streams through the embedded computing system installed in the drone. After the extraction of the video feed, its frames are segmented into individual frames and preprocessed using procedures like resizing, normalization, and de-noising to maintain consistency. The preprocessed frames undergo feature extraction through a convolutional neural network that is very lightweight, known as MobileNet, where spatial features related to human body posture, interaction behavior, and surroundings of the scene are extracted. The temporal relationships between consecutive frames are learned by passing these features into an LSTM neural network, allowing the detection of motion patterns related to acts of violence. In order to further enhance the accuracy of the detection process, the optical flow analysis technique has been incorporated into the system for the calculation of motion intensity and its direction from frame to frame. The additional information on movement serves as an important factor in identifying whether the action taking place is normal or violent. The calculated spatial and temporal characteristics, together with motion descriptors, have been sent for processing by the classifier module responsible for computing the probability of the violent act taking place. With the application of thresholding and temporal smoothing techniques, the system determines whether the recorded act was violent or non-violent. As a consequence of detecting a violent act, certain automated processes are initiated in order to record the incident and provide live video streaming to the ground station.

C. Hardware Implementation

The design framework is realized on an aerial surveillance platform consisting of a quadcopter. The drone frame acts as a robust mechanical housing with uniform load balancing that facilitates mounting electronic components. A flight controller called Pixhawk 2.4.8 will control flight dynamics, stabilization, and sensor fusion based on inputs from the sensors attached to the drone. The drone will use 2212 920KV brushless DC motors coupled with 10×4.5 propellers to facilitate sufficient lift and maneuverability. Motor speed will be controlled using Electronic Speed Controllers (ESCs) with commands from the flight controller.

TABLE I
HARDWARE COMPONENTS AND SPECIFICATIONS

Component	Description
Frame	Quadcopter Frame
Flight Controller	Pixhawk 2.4.8
Motors	2212 920KV Brushless Motors
ESC	Spedix ES30 HV 3–6S BLHeli_S
Propellers	10×4.5 HD Propellers
GPS Module	Neo-6M GPS
Transmitter	FlySky FS-i6
Receiver	FS-iA6B
Battery	2200mAh 3S Li-Po
Processing Unit	Raspberry Pi 3
Camera	USB Camera Module
Buzzer	Piezoelectric Buzzer

In regards to navigation and communication capabilities, the drone uses a Neo-6M GPS module for determining position

Fig. 2. Prototype of the Proposed Drone-Based Surveillance System

Fig. 3. Onboard Hardware Components Including Pixhawk Controller and GPS Module

and altitude data required in covering the surveillance area. An FliSky FS-i6 transmitter and FS-iA6B receiver will be used for remote manual control and communication respectively. Power is supplied to the system via a 3S 2200mAh Li-Po battery. Edge computing will be provided by a Raspberry Pi 3 device to carry out the proposed video stream analysis and implement the deep learning algorithm. Video streaming and analysis is conducted with the help of a USB camera module attached to the drone. Hardware implementation permits the smooth incorporation of air mobility and edge computing, making the surveillance process of any form of violence possible at all times. The ability to integrate further sensors and processing units makes the design scalable and flexible enough to be implemented in actual surveillance systems.

The architecture of the quadcopter used as a platform for surveillance is illustrated in Fig. 2. The drone body offers structural framework for incorporating the propulsion system, the landing gears, and the electronic components of the drone. BLDC motors along with propellers provide the necessary lifting power for the platform, while at the same time, the centralization of the components ensures uniform distribution of weight. The drone is capable of carrying extra weight, for instance, the camera system. The Fig. 3, shows the different components installed on board the drone. The Pixhawk 2.4.8 flight controller functions as the control unit responsible for controlling the drone's stability and flight process through the processing of sensor data and control signals that regulate motor performance via ESCs. The GPS module offers information related to navigation and positioning in real time, whereas the wiring installation indicates power distribution and communication integration on the drone.

V. Results and Discussion

The proposed algorithm for violence detection has been evaluated using the data set that contains both violent and non-violent video clips. The data set has been separated into training and validation parts. For the model training, the MobileNet-LSTM based deep learning algorithm with Optical Flow analysis has been used. To evaluate the performance of the proposed system, the metrics such as accuracy, precision, recall, and F1 score have been calculated by analyzing the confusion matrix. According to the confusion matrix results, the model is capable of differentiating between violent and non-violent acts with high precision. From the quantitative evaluation, it can be seen that the accuracy attained by the proposed method is 88.1%. This means that the system has been able to classify the vast majority of videos accurately. The precision attained by the system is 82.05%. This is because the system sometimes incorrectly classifies some non-violent activities as violent activities. This results in some false positives. The recall value attained by the system is 96%. This means that the system has been able to identify most of the violent activities that occur. Therefore, very few violent activities have gone undetected. The F1 score attained by the system is 88.5%. This suggests that the model is efficient for use in situations that require the detection of violent events, such as security and law enforcement. From the system point of view, the performance of the model remains stable in real-time settings when using it on an edge device. Using MobileNet greatly decreases computation complexity; as a result, it allows us to get good features for the model while not losing the capability of detecting aggressive actions. Applying temporal smoothing to predictions allows us to further stabilize the system since it eliminates unnecessary fluctuation during consecutive frame processing. However, the presence of false positives shows that sometimes the model incorrectly identifies fast, non-aggressive actions as an aggressive one. Despite the limitation, the system provides a nice balance between the efficiency of aggressive action detection and computation costs. As shown from the confusion matrix, the algorithm accurately

TABLE II
Confusion Matrix

	Predicted Non-Violence	Predicted Violence
Actual Non-Violence	80	21
Actual Violence	4	96

categorizes 96 samples as violent events (true positives) and 80 as non-violent events (true negatives). However, there are

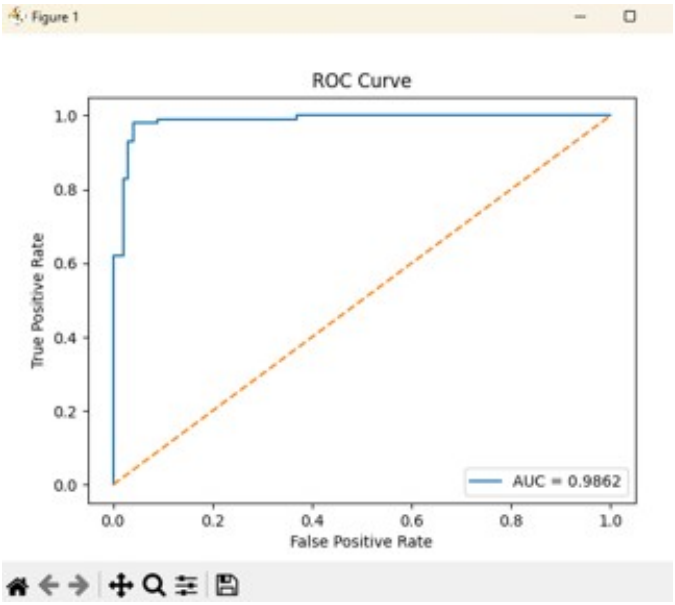

Fig. 4. ROC-AUC Curve

TABLE III
COMPARISON WITH EXISTING METHODS

Method	Accuracy (%)	Precision (%)	Recall (%)	Edge Deployable
CNN-Based Methods	85	80	78	Yes
3D CNN	92	88	90	No
SlowFast/X3D	93	89	91	No
Proposed Method	**88.1**	**82.05**	**96**	**Yes**

21 samples that should have been categorized as non-violent but were instead categorized as violent events (false positives), and 4 samples of violent events that were not detected (false negatives). The fewer false negatives show that the model is efficient in identifying violent events.

Receiver Operating Characteristics (ROC) curve will be used for evaluating the classification abilities of the suggested algorithm through different levels of decision thresholds. It depicts the relationship between the True Positive Rate (recall) and the False Positive Rate. This particular model has a great capacity for separating violent acts from nonviolent ones since it gives good results regarding the high True Positive Rate while the False Positive Rate remains in control. Such an approach proves the validity of integrating spatio-temporal learning along with motions to identify aggressiveness from regular behavior.

Based on the analysis, although there exist deep learning techniques that have higher levels of accuracy in the detection process, the methods employed are not efficient enough because they use highly complex networks that cannot be used in real-time. On the other hand, the proposed technique performs best in terms of recall value; hence, there will always be minimal errors or failures in detecting any violent activity. The technique, therefore, would be most applicable in scenarios that require a high level of security such as safety-related surveillance systems. The proposed classification ability of the violence detection algorithm is estimated in terms of classification accuracy, precision, recall, and F1-scores that

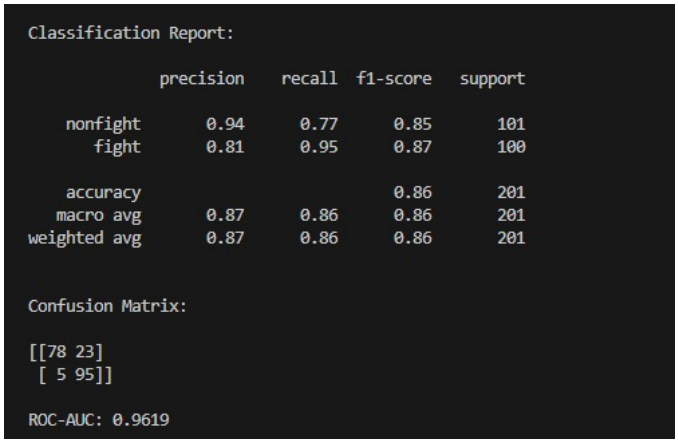

Fig. 5. Terminal Output Showing Performance Metrics and Confusion Matrix

are calculated based on the provided classification report and confusion matrix, as presented in Fig. **??**. Overall accuracy of the classifier is 86%, which means satisfactory classification results for both considered classes. As far as non-violence is concerned, the classifier demonstrates excellent precision (0.94) and rather low recall (0.77). It means that, while the classifier shows a high rate of correctness when classifying non-violent cases, there are still mistakes and some data belonging to non-violence category is classified as being violent. Violent videos show even higher precision (0.81) and extremely high recall rate (0.95), indicating that it does not miss any actual violent case while being quite precise in its predictions. Thus, for safety purposes, the obtained results are highly suitable and demonstrate that the proposed model is capable to detect actual violent behavior. Confusion matrix supports the findings, presenting 95 correctly classified videos and only 5 false-negative cases within the violent category and 78 classified non-violent videos and 23 false positives among them. The model's high ROC-AUC value is 0.9619.

Output of the proposed violence detection system in real-time form is shown in Fig. 2. In the figure, we can see that the proposed algorithm detects the violence among the individuals and the occurrence is classified as "VIOLENCE" with a score of 0.83. Along with the classification result, the system calculates the motion value of the event, which is 2.05 based on the optical flow technique. This helps in determining the presence of high-intensity motions that can be related to aggression. Using both spatial classification and motion analysis results in better accuracy in detection of violent events as the appearance and motion aspects are analyzed for the same. These findings prove that the suggested model focuses on maximizing the recall rate for violent incidents while ensuring acceptable accuracy levels, making it appropriate for live surveillance applications where false negatives should be minimized.

VI. CONCLUSION

This study provides a new framework based on edge computing and drones for the detection of violent activities in real-time. The method includes the use of MobileNet spatial feature

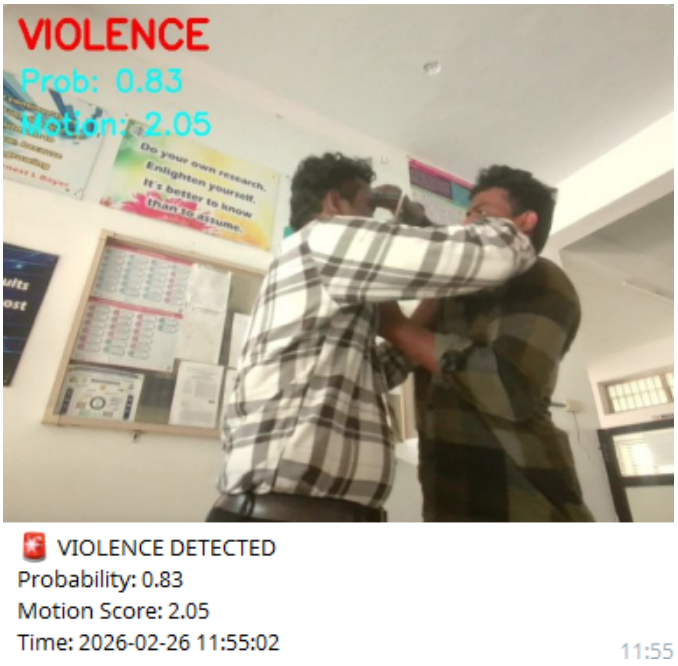

Fig. 6. Real-Time Violence Detection Output

extraction combined with LSTM for temporal modeling along with optical flow motion estimation to accurately extract visual information about human movements. Through the deployment of the trained model on a Raspberry Pi embedded in the drone, the process can be carried out through decentralized processing with low latency, thereby minimizing any possible communication overhead.

As evident from the experimental results, it can be seen that the system performs reliably in terms of detection, especially when it comes to recall, which means that most of the violence incidents are detected properly by the system. Such an outcome is important, as missing out on detection may lead to serious problems in the case of applications related to security purposes. It must be noted that although the precision is not at its best because of some false detections, this suggests that the architecture of the proposed approach strikes a balance between detection sensitivity and computational effectiveness.

The suggested approach is an efficient solution for intelligent surveillance because it deals with several issues concerning coverage, latency, and automation. The direction for future research involves improving the accuracy of detection with context-based modeling, incorporating object detection into the system, and boosting its performance with the use of hardware acceleration methods like FPGA implementation.

VII. Future Work

- **Model Enhancement:** Improve detection precision by incorporating attention mechanisms and context-aware deep learning models to reduce false positives in complex environments.
- **Hardware Optimization:** Deploy the system on hardware accelerators such as FPGA or edge GPUs to achieve lower latency and improved energy efficiency for real-time operation.
- **System Extension:** Integrate object detection and multi-modal inputs (e.g., audio) to enhance scene understanding and improve robustness in diverse surveillance scenarios.

References

[1] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.
[2] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1446–1453.
[3] A. Dosovitskiy, P. Fischer, E. Ilg, et al., "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2015, pp. 2758–2766.
[4] J. Li, X. Liu, and H. Zhang, "Efficient violence detection using 3D convolutional neural networks," in *Proc. IEEE Int. Conf. Advanced Video and Signal-Based Surveillance (AVSS)*, 2019.
[5] M. M. Soliman, M. Kamal, M. Nashed, et al., "Violence recognition from videos using deep learning techniques," in *Proc. Int. Conf. Intelligent Computing and Information Systems (ICICIS)*, 2019.
[6] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019.
[7] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020.
[8] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020.
[9] K. Cao, J. Ji, Z. Cao, et al., "Few-shot video classification via temporal alignment," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020.
[10] M. Cheng, K. Cai, and M. Chen, "An open large-scale video database for violence detection," in *Proc. Int. Conf. Pattern Recognition (ICPR)*, 2021.
[11] P. Sernani, N. Falcionelli, F. Dragoni, et al., "Deep learning for automatic violence detection: Tests on the AIRTLab dataset," *IEEE Access*, vol. 10, pp. 1–12, 2022.
[12] S. Stanislaw, et al., "Discrete neural representation for explainable anomaly detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022.
[13] D. Choqueluque-Roman and G. Camara-Chavez, "Weakly supervised violence detection in surveillance video," *Sensors*, vol. 22, no. 12, p. 4502, 2022.
[14] FFmpeg Developers, "FFmpeg tool," Version be1d324, 2016. [Online]. Available: http://ffmpeg.org/
[15] T. Aremu, Z. Li, and R. Alameeri, "Any object is a potential weapon: Weaponized violence detection using salient image," *arXiv preprint arXiv:2207.12850*, 2022.
[16] R. Vijeikis, V. Raudonis, and G. Dervinis, "Efficient violence detection in surveillance," *Sensors*, vol. 22, no. 6, p. 2216, 2022.
[17] ASUS, "Xtion 2 depth sensor," 2017. [Online]. Available: https://www.asus.com/
[18] Intel, "RealSense Depth Camera D435," 2022. [Online]. Available: https://www.intelrealsense.com/