Enhancing Violent Behavior Recognition in Schools Through YOLOv8 Optimization Using LSTM with Multi-Camera Model

Nghia. Phan Duc, Hoa. Doan Nguyen Thanh, Khang. Huynh Minh, Nha. Le Hoang Trang, Hai. To Thanh
Email: nghiapdse182439@fpt.edu.vn, hoadnt@fe.edu.vn,
khanghuynh0245@gmail.com, nhalehoangtrang2005@gmail.com,
HaiTTSE184086@fpt.edu.vn

***Abstract*—Violent behavior in schools poses a significant threat to the safety and well-being of students and staff. Effective recognition and prevention strategies are crucial for creating a secure educational environment. This paper presents an innovative approach to enhancing school violent behavior recognition by integrating YOLOv8 optimization with Long Short-Term Memory networks within a multi-camera surveillance model. YOLOv8, a state-of-the-art object detection framework, is optimized for real-time performance and accuracy in identifying violent actions. Including Long Short-Term Memory networks enables the system to capture temporal dependencies and improve the contextual understanding of sequential frames, thereby reducing false positives and enhancing recognition accuracy. By leveraging a multi-camera setup, the model ensures comprehensive coverage and minimizes blind spots, providing a holistic view of the monitored areas. The proposed system is evaluated on a publicly available dataset and a custom dataset collected from school environments. Experimental results demonstrate a significant improvement in the precision and recall of violent behavior recognition, underscoring the potential of this approach for real-world deployment. The findings highlight the effectiveness of combining advanced object detection algorithms with temporal analysis and multi-camera integration to create a robust and reliable violence detection system for educational institutions.**

***Index Terms*—Violent Recognition, Long Short-Term Memory, Multi-camera Integration, YOLO**

I. Introduction

Violence in schools is a growing concern globally, with incidents ranging from physical altercations to more severe forms of aggression that endanger the safety of students and staff. Traditional security measures, such as manual monitoring by school personnel and static surveillance cameras, have proven insufficient in effectively detecting and preventing these incidents in real time [1]. The increasing frequency and severity of school violence necessitate the development of advanced technological solutions that can provide timely and accurate recognition of violent behaviors, thereby enabling prompt intervention and ensuring a safe educational environment.

Effective violence detection systems in schools are crucial for several reasons. Firstly, they enhance overall security by providing early warnings and allowing for swift action to prevent escalation. Secondly, they contribute to a safer and more conducive learning atmosphere, which is essential for the academic and social development of students. Thirdly, advanced detection systems can assist in the collection of evidence and facilitate subsequent investigations, thereby holding perpetrators accountable and deterring future incidents [2]. Given these benefits, there is an urgent need for robust and reliable systems capable of real-time violent behavior recognition.

The primary objective of this study is to develop an enhanced violent behavior recognition system for school environments by integrating YOLOv8 optimization with Long Short-Term Memory (LSTM) networks in a multi-camera surveillance model. Specifically, this research aims to optimize YOLOv8 for Violence Detection: Fine-tune YOLOv8, a leading object detection algorithm, to accurately identify violent actions in a school setting, ensuring high precision and recall rates. Incorporate Temporal Analysis Using LSTM: Utilize LSTM networks to capture and analyze temporal sequences, thereby improving the system's ability to understand the context and reduce false positives. Implement a Multi-Camera Model: Design and deploy a multi-camera setup to provide comprehensive coverage of the monitored areas, ensuring no blind spots and enhancing the overall reliability of the detection system.

This paper follows a structured organization comprising five key sections. Session I, the Introduction sets the stage by discussing the increasing concern of violence in schools, the limitations of traditional security measures, and the necessity for advanced technological solutions. Session II, the related Work, reviews existing violence detection techniques. In session III, the Proposed Methodology shows the details of the system architecture, explaining the optimization of YOLOv8, the integration of LSTM networks to handle temporal dependencies, and the multi-camera model's design and data fusion techniques. Session IV, The Experimental Results describes the outline of the experimental setup, including hardware and software specifications, training and validation protocols, and evaluation metrics. Finally, session V, the Conclusion, summarizes the key contributions and findings, discussing the implications for enhancing school safety.

II. Related Knowledge

Violence detection in surveillance systems has been a critical area of research, particularly for ensuring safety in public

spaces like schools [3]. Traditional methods relied heavily on manual monitoring and basic motion detection algorithms, which often led to high false positive rates and missed incidents due to human error or limited system capabilities [3][4][5]. With the advent of advanced machine learning and computer vision techniques, more sophisticated approaches have been developed, offering improved accuracy and real-time detection capabilities [5].

A. Object Detection Methods (YOLO Series)

The YOLO (You Only Look Once) series has revolutionized object detection with its real-time processing capabilities and high accuracy [6]. YOLOv1 introduced a single-stage detection approach, where a single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation[7]. Subsequent versions, YOLOv2 and YOLOv3, improved upon this foundation by enhancing the network architecture, incorporating multi-scale predictions, and integrating better backbone networks. YOLOv4 and YOLOv5 continued this trend, optimizing for speed and accuracy, making them suitable for real-time applications. YOLOv8, the latest iteration, further refines these capabilities, offering enhanced performance through advanced optimization techniques and more efficient model architectures [8]. YOLOV8-pose is considered the most advanced model in real-time detection and pose recognition of multiple objects in one frame, and the model's limitations in detecting small objects have also been resolved. by integrating with machine learning algorithms that complement and support object recognition at a distance. With a compact version optimized for edge devices, YOLOV8n-pose also has mAPpose (50-95) up to 50.4 and the highest version for centralized servers or edge servers has mAPpose level (50- 95) up to 71.6

B. LSTM: Long short-term memory fix for sequential processing problems of RNNs

While object detection methods like YOLO excel at identifying static instances of objects or actions, understanding complex behaviors, especially violent actions, often requires analyzing temporal sequences.

Recurrent neural network (RNN) is an optimal neural network for processing sequential events that can carry frame information from the previous state to the following states, and then in the final state is the combination of all the frames. images to predict actions in a video[9]. However, recurrent neural networks still encounter some problems in remembering states from previous frames, leading to the fact that information can only be carried through a certain number of states, after which there will be a vanishing gradient, in other words, another way is that the model only learns from states near it.

To completely overcome this problem, the LSTM network, a type of RNN model was born to overcome the vanishing gradient problem on RNN by using special gates in its structure: forget gate, input gate, and output gate [9][10]. These gates help regulate the flow of information across time steps, effectively maintaining and updating hidden and memory cell states, helping to retain important information, and preventing gradients from being reduced too much during backpropagation. LSTM networks have been widely used in various domains for sequence prediction and analysis, from natural language processing to video analysis[11]. In the context of violence detection, LSTM enable the system to understand the progression and context of actions across multiple frames, thereby improving the accuracy of behavior recognition.

C. Multi-camera modeling: a behavioral decision method

Current approaches to violence detection vary widely in terms of techniques and effectiveness. Traditional methods, relying on heuristic-based motion detection, are prone to high false positive rates and limited in scope. More recent methods leveraging convolutional neural networks and deep learning have significantly improved detection accuracy but often require substantial computational resources [12]. Integrating object detection algorithms like YOLO with LSTM networks offers a promising solution, combining the strengths of both spatial and temporal analysis. Multi-camera models further enhance detection capabilities by providing comprehensive coverage and reducing blind spots. Compared to single-camera systems, multi-camera approaches are more robust but require sophisticated data fusion and synchronization methods [13][14].

The integration of YOLOv8 for real-time object detection, LSTM networks for temporal sequence analysis, and multi-camera systems represents a state-of-the-art approach to violence detection in schools. This combined methodology leverages the latest advancements in machine learning and computer vision to address the limitations of existing techniques, offering a robust solution for enhancing school safety.

III. A NEW MODEL FOR SCHOOL VIOLENT BEHAVIOR IDENTIFICATION SYSTEM

The proposed system integrates YOLOv8 for object detection, LSTM networks for temporal sequence analysis, and a multi-camera model for comprehensive surveillance coverage. The system is designed to operate in real-time, providing accurate and timely recognition of violent behaviors in school environments. The roles of the components are as follows:

YOLOv8 is selected for its superior performance in real-time object detection. The model is fine-tuned specifically for violence detection using a dataset containing various forms of violent actions such as fighting, bullying, and aggressive postures.

LSTM networks are integrated into the system to capture temporal dependencies and contextual information from sequential frames. The LSTM network is designed to process the sequence of bounding boxes and class probabilities generated by YOLOv8, learning the patterns associated with violent behaviors over time.

A strategic placement of multiple cameras is essential to ensure comprehensive coverage of the monitored areas. Cameras are positioned to minimize blind spots and provide overlapping

fields of view, allowing the system to capture incidents from multiple angles.

Our envisioned system architecture consists of two key components, denoted as Part A and Part B. Part A will include recognizing the subject's posture and behavior to collect data. at a camera angle for Part B, while Part B undergoes processing in the central cloud server to make the final decision. A comprehensive description of the system architecture is illustrated in Figure 1.

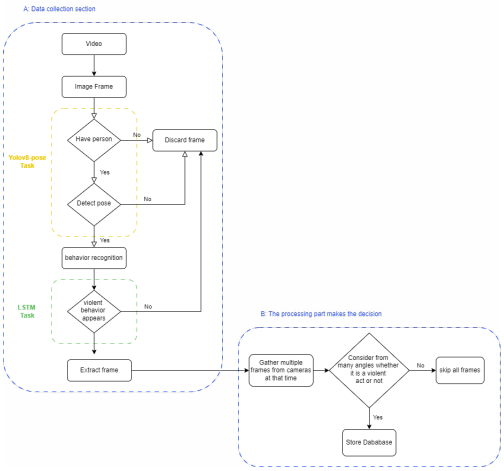

Fig. 1. The proposed architecture

When data from camera angles reaches the server. The processing of these images on the server (Part B) will start with focusing on calculating the percentage of action probability based on different image angles. The behavior then records this information in the database. More detailed information about the architecture and behavioral extension process will be shown in the following Figure 2 model.

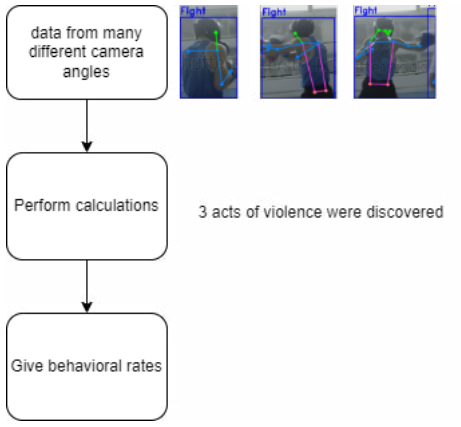

Fig. 2. Package segment

IV. Experimental results

In the experiment below to demonstrate the ability to handle a series of behaviors, we used videos about 1 minute in length that are close to the actual duration of a real-life fight or bullying with the standard frame rate of modern cameras supporting 60 FPS to test the recognition ability of LSTM and RNN models in each case of using one camera and multiple cameras respectively. First, let's take a look below at the parameters of the yolov8-pose model used in this study to identify key points in the human body.

Model	mAPpose(50-95)	mAPpose(50)	Speed A100 (ms)	params (M)	FLOPs (B)
YOLOv8m-pose	65.0	88.8	2.00	26.4	81.0

TABLE I
DETAILS OF THE PARAMETERS OF THE YOLOV8-POSE MODEL USED.

Below is the accuracy of 2 models LSTM and RNN on 5 actions of running, walking, hugging, talking, and fighting shown in Table 2 and Table 3.

Behavior	Accuarncy
Fighting	95,67%
Running	84,16%
Walking	80,5%
Talking	78,91%
Hugging	78,56%

TABLE II
THE DETAILED ACCURACY OF LSTM ON EACH BEHAVIOR IS TESTED ON APPROXIMATELY 1-MINUTE-LONG VIDEOS WITH A FRAME RATE OF 30 FPS.

Behavior	Accuarncy
Fighting	83,5%
Running	78,0%
Walking	78,21%
Talking	78,16%
Hugging	75,86%

TABLE III
THE DETAILED ACCURACY OF RNN ON EACH BEHAVIOR IS TESTED ON APPROXIMATELY 1-MINUTE-LONG VIDEOS WITH A FRAME RATE OF 30 FPS.

Through the results, we can see the reduced ability of RNN compared to LSTM in identifying frame sequences containing long behaviors, especially in videos with objects changing behavior in the last frames. Next, we will review the results of experiments on single and multiple-camera deployments.

Model	One camera	Multi camera
LSTM	79,52%	93.65%
RNN	73.15%	84.33%

TABLE IV
DISPLAY DETAILED TEST RESULTS ON VIDEO UNDER 1 CAMERA ANGLE AND MULTIPLE CAMERA ANGLES OF 2 MODELS LONG SHORT-TERM MEMORY AND RECURRENT NEURAL NETWORKS

At a special camera angle with camera angles only from behind the subject, both models show that confusion between the behaviors of talking, hugging and fighting leads to reduced results, while with the 3 camera angles in Experimental results have been much improved. Next, to calculate the accuracy in the general multi-camera model I used the following formula

$$\text{behavioral rate} = \left(\frac{\text{the highest number of detected}}{\text{total number of detected}} \right) \times 100\% \quad (1)$$

V. CONCLUSION

This research article provides a thorough examination of efficiently identifying violent behavior in schools. By fine-tuning Yolov8 recognition models with LSTM, the study has successfully addressed the crucial issue of restraining and preventing violent incidents in modern educational settings, leading to enhanced overall performance and safety in schools. Our suggested model, which utilizes Yolov8's human skeleton recognition and LSTM algorithm, has demonstrated promising outcomes in practical assessments, surpassing existing methods in terms of accuracy and efficiency. These results underscore the effectiveness of violent behavior detection systems in enhancing safety within school premises.

REFERENCES

[1] Himanshu Gupta, and Syed Taqi Ali, "Violence Detection using Deep Learning Techniques", in *2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*, Aug. 2022.

[2] Gul e Fatima Kiani, and Taheena Kayani, "Real-time Violence Detection using Deep Learning Techniques", in *2022 3rd International Conference on Innovations in Computer Science & Software Engineering (ICONICS)*, Dec. 2022.

[3] N. Murali Krishna, Ramidi Yashwanth Reddy, Mallu Sai Chandra Reddy, Kasibhatla Phani Madhav, and Gaikwad Sudham, "Object Detection and Tracking Using Yolo", in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, Sep. 2021.

[4] Reza Ghanbari, and Keivan Borna, "Multivariate Time-Series Prediction Using LSTM Neural Networks", in *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, Mar. 2021.

[5] Bakhita Salman, Mohammed I. Thanoon, Saleh Zein-Sabatto, and Fenghui Yao, "Multi-camera Smart Surveillance System", in *2017 International Conference on Computational Science and Computational Intelligence (CSCI)*, Dec. 2017.

[6] M. Alruwaili, M. N. Atta, M. H. Siddiqi, A. Khan, A. Khan, Y. Alhwaiti, and S. Alanazi. "Deep learning-based YOLO models for the detection of people with disabilities". *IEEE Access*, Dec. 2023, pp. 2543-2566.

[7] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.

[8] Utsavi Patel, Rohan Vaghela, Yashvi Popat, Hirva Patel, Jigar Sarda, and AkashKumar Bhoi, "Multi-Class Event Classification using YOLOv8", in *2024 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, Jan. 2024.

[9] Svetoslav Zhelev, and Dimiter R. Avresky, "Using LSTM Neural Network for Time Series Predictions in Financial Markets", in *2019 IEEE 18th International Symposium on Network Computing and Applications (NCA)*, Sep. 2019.

[10] Linkai Wang, Jing Chen, Wei Wang, Ruofan Wang, Lina Yang, and Mai An, "A Time Series Prediction Model Based on Long Short-Term Memory Networks", in *2021 7th International Conference on Systems and Informatics (ICSAI)*, Nov. 2021.

[11] Jiayi Sun, and Wenming Guo, "Time Series Prediction Based on Time Attention Mechanism and LSTM Neural Network", in *2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS)*, Feb. 2023.

[12] A. Mumtaz, A. B. Sargano, and Z. Habib, "Violence Detection in Surveillance Videos with Deep Network Using Transfer Learning", in *2018 2nd European Conference on Electrical Engineering and Computer Science (EECS)*, Dec. 2018.

[13] Reena Kumari Behera, Pallavi Kharade, Suresh Yerva, Pranali Dhane, Ankita Jain, and Krishnan Kutty, "Multi-camera based surveillance system", in *2012 World Congress on Information and Communication Technologies*, Nov. 2012.

[14] Maria Gadelkarim, Mazen Khodier, and Walid Gomaa, "Violence Detection and Recognition from Diverse Video Sources", in *2022 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2022.