Multistage Real-time Violence Detection using Convolutional Neural Network and Long Short-term Memory

Manh Dung Nguyen
Reseach Institute of Posts and Telecommunication
Posts and Telecommunications Institute of Technology
Hanoi, Vietnam
dungnm1@ptit.edu.vn

Soonghwan Ro
Smart Information Technology Engineering
Kongju National University
CheonAn, Korea
rosh@kongju.ac.kr

***Abstract*— Action recognition is a challenging research topic in the field of computer vision, with numerous practical applications, including violence detection. Early detection of violent behavior enables timely intervention, helping to prevent or minimize the damage caused by such incidents.**

In this paper, we present a multi-stage method for violence detection. In the first stage, groups at high risk of violence are detected using YOLO, the Deep SORT object tracking algorithm, and a 2D CNN image classification model. In the second stage, entropy-based features are employed to eliminate non-violent objects that visually resemble violent ones. In the final stage, 2D features of the remaining groups are extracted using a Convolutional Neural Network (CNN) and then fed into a Long Short-Term Memory (LSTM) network to determine whether the group is engaged in violent or normal activity.

Experimental results demonstrate that, compared to previous studies, the proposed method not only achieves effective detection of violent behavior but also reduces false positives, delivering strong performance suitable for real-world applications

Keywords—Convolutional Neural Network, Violence Detection, YOLO, Long Short-Term Memory

I. Introduction

Violence has long been a major concern in societies worldwide. It causes numerous negative consequences, harming physical health, mental well-being, property, and, in some cases, even human life. Despite the implementation of various preventive measures, violent incidents continue to occur frequently, showing no signs of decline.

If an effective monitoring mechanism can promptly detect and alert violent behavior, it would be possible to prevent such incidents or, at the very least, minimize the damage they cause.

In recent years, CCTV surveillance systems have rapidly expanded and are now widely deployed in hospitals, schools, streets, and other public areas. However, most existing systems remain limited to capturing and storing video data, while the actual monitoring process relies heavily on human operators. As a result, monitoring efficiency is low, and operational costs are high. To address these challenges, integrating artificial intelligence (AI) into surveillance systems has emerged as a prominent research trend, aiming to enhance monitoring capabilities and improve operational efficiency.

Machine learning refers to methods that automatically construct a mathematical model using sample data, also known as training data, enabling the system to learn from input data without the need for explicit programming. Although machine learning has been developed since the 1940s, its results were initially unimpressive, mainly due to difficulties in data collection and limitations in computational resources. By the end of the decade, advancements in computer hardware, along with the development of the internet, had made data collection much easier, thereby accelerating the growth of machine learning.

Recently, a branch of machine learning known as deep learning has emerged as one of the most powerful machine learning approaches. Deep learning encompasses a set of advanced techniques that utilize multi-layer neural networks, achieving remarkable results in various computer vision tasks. This significant advancement has also contributed to addressing challenging problems that remained unsolved, such as violence detection.

In this paper, we propose a deep learning-based multi-stage violence detection method using YOLO, Entropy-based classification and CNN–LSTM video classification. The remainder of the paper is organized as follows: Section II reviews related work. Section III describes the proposed method. Section IV reports the experimental results. Future research directions and discussions are presented in Section V

II. Related Work

In the literature, numerous methods have been proposed for the problem of violence detection [1, 2, 3]. These methods focus on the use of machine learning and image analysis.

A machine learning approach that has achieved good results is the *Fast Fight Detection* method [1]. This method assumes that, in a violent video, the regions of moving pixels have distinctive shapes and positions. First, the differences between consecutive frames are computed, and their absolute values are taken. Next, the resulting image is binarized to create motion regions. The K largest motion regions are selected for further processing. Finally, to classify these K motion regions, parameters such as centroid, perimeter, area, and the distances between them are calculated. Experimental results on the Movie, Hockey Fight, and UCF-101 datasets demonstrate that this method is effective and can be applied in real-world scenarios. However, it is not effective for classifying videos with continuous motion.

Deep learning is a subset of machine learning that primarily focuses on the use of multi-layer neural networks. Compared to traditional machine learning, deep learning has achieved higher accuracy and efficiency in many computer vision tasks, particularly in image classification.

The Convolutional Neural Network (CNN) [4] is one of the most widely used architectures for image classification. CNNs were developed based on partially mimicking the way the human brain works—using spatial features to classify an image. They employ numerous learnable filters capable of automatically extracting features from images, enabling CNNs to "see" important features that manual feature extraction methods might fail to detect. During training, significant features are retained, while less relevant ones are eliminated from the system.

Although CNNs have achieved notable successe in image classification tasks, they are not as effective for action classification. The main reason is that an action is a sequence of consecutive images, and relying on a single image makes it difficult to produce an accurate prediction.

Unlike CNNs, the Long Short-Term Memory (LSTM) network [5] was designed with the idea of mimicking human thought processes, an aspect that traditional neural networks have not been able to achieve. Humans do not start their thinking from scratch.

At every moment, for example, when classifying an image from a movie, humans rely on previous images rather than only the current one, as CNNs do. An LSTM consists of a sequence of repeating neural network modules that mimic the way the human brain processes information. Each module contains four different neural network layers that interact in a special way. As a result, LSTMs can retain information over long periods, making them well-suited for learning from sequences of images in action classification tasks.

Many studies have shown that deep learning can be effectively applied to the problem of violence detection [6, 7, 8, 9]. Among these, the most accurate and efficient approach is the combination of CNN and LSTM [9]. First, consecutive frames are fed into a CNN to extract features. These features are then passed to a Bidirectional LSTM [10] to classify the sequence of frames as violent or non-violent. This method has been tested on the Hockey, Peliculas, and Collected Surveillance Camera datasets (the latter collected by the authors) and has achieved very good results, making it applicable for classifying videos with continuous motion. However, the method's accuracy decreases when classifying videos in which violent scenes occupy only a small portion of the frame. For example, in image number 2 from the PTIT dataset, the violent scene in the video accounts for only a small area of the frame, causing the algorithm to perform poorly.

The reason is that when using the entire image for feature extraction, the features representing violent behavior are not sufficiently prominent compared to other objects. To overcome this limitation, we propose a multi-stage violence detection method using CNN–LSTM. This model allows us to focus more closely on the locations where violent behavior occurs, thereby achieving higher accuracy.

III. Proposed Method

The proposed violence detection method is illustrated in Figure 1. This approach is divided into three main stages. In the first stage, groups of people with a high likelihood of violent behavior are detected and localized using YOLO [11], combined with the Deep SORT [12] object tracking algorithm. In the next stage, images of these high-risk groups from the video frames are passed to a CNN for feature extraction. Finally, in the last stage, the extracted features are fed into an LSTM to classify and determine whether the group is truly engaging in violent behavior or merely normal activity.

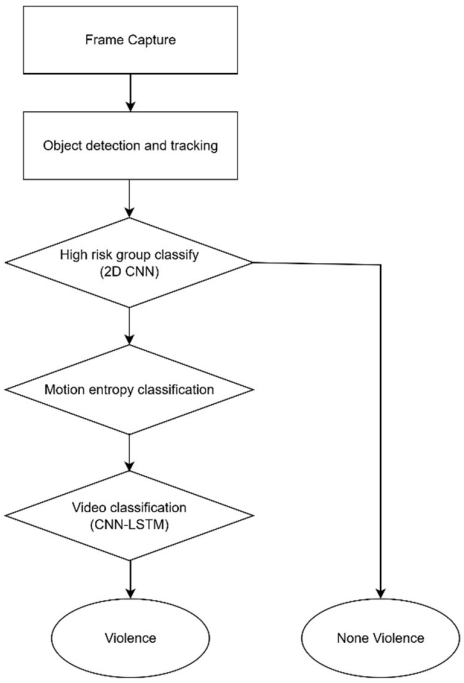

Fig. 1 Multi-stage Violence detection algorithm

A. High-risk Group Detection

YOLO [13] is employed for object detection and Deep SORT [12] for tracking to identify groups of individuals standing in close proximity. Segmented images of these groups are then classified by a 2D CNN to assess their potential risk of violent behavior. All groups classified as violent are added to a risk list for further processing, aimed at reducing false positives in violence detection. Fig

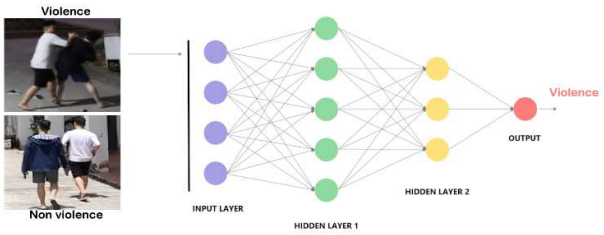

Fig. 2 High-risk of violence group classification

B. Motion Entropy Classification

Not all groups in the risk list are truly involved in violent behavior. In many cases, it is difficult to accurately determine the nature of the activity using a single image. As illustrated in Figure 3, certain scenes may appear visually similar to violent actions but are, in fact, normal activities. To address this, we introduce a

Fig. 3 Examples of non-violent activities that appear visually similar to violent actions

lightweight yet effective classifier, termed Motion Entropy Classification, to reduce false positives in violence detection. Through observation, we note that violent subjects tend to move more chaotically than normal ones, with greater variability in both velocity and direction. We use entropy to quantify the degree of motion chaos for each object, where a higher entropy value indicates a higher likelihood of violent behavior.

Based on this observation, we developed a method to measure the motion entropy of each individual within high-risk groups, identifying those with elevated entropy as potential violent actors. Groups containing at least one such individual are retained for further processing, whereas groups without are removed from the potential risk list.

The following describes the **Mathematical Definition of Motion Entropy**:

Let:

$$V_t = \frac{100}{H_t}\sqrt{|x_t - x_{t-1}|^2 + |y_t - y_{t-1}|^2} \quad (1)$$

$$\theta_t = \arctan\left(\frac{y_t - y_{t-1}}{x_t - x_{t-1}}\right) \quad (2)$$

Where:

- x_t, y_t are the object's coordinates at frame t
- H_t is object height, we normalize object's speed by its height to make it independent of the camera's viewing distance.
- V_t is the object's speed at frame t
- θ_t is the object's movement direction at frame t

The set of object speeds and directions over the last nnn frames, H(X)H(X)H(X), is used as a feature vector representing the object's motion entropy:

$$H(X) = [\theta t\text{-}n, Vt\text{-}n, \theta t\text{-}n+1, Vt\text{-}n+1 \ldots \theta t, Vt] \quad (3)$$

This vector captures the temporal variability of both velocity magnitude and direction, which is subsequently used as input to a Support Vector Machine (SVM) classifier trained to discriminate between individuals exhibiting high-risk violent motion patterns and those engaged in normal, non-violent activities

C. CNN-LSTM Violence activity classification

Although CNN-based image classification produces good results for detecting high-risk violent behavior, not all groups of individuals standing in close proximity are necessarily engaged in such actions. To reach a reliable final decision, an additional processing step is required to distinguish between genuine violent behavior and groups without violence. Motion entropy is an effective feature, as it captures the temporal characteristics of potentially violent subjects; however, it still has limitations, particularly in scenarios with dynamic backgrounds where accurate object tracking becomes challenging.

Violent behavior is a sequence of actions; therefore, it is necessary to observe a series of consecutive frames in order to make a final prediction. As previously mentioned, the most suitable model for classifying such sequential frame data is the CNN–LSTM architecture.

The CNN–LSTM architecture employs a Convolutional Neural Network (CNN) to extract the 2D spatial features of the input images, which are then processed by a Long Short-Term Memory (LSTM) network to capture the temporal dependencies before producing the final prediction.

To select an appropriate CNN model, we experimented with several widely used architectures trained on the ImageNet dataset. The results, shown in Table 1, indicate that ResNet18 [14] provides the best balance between accuracy and model complexity. Therefore, ResNet18 was chosen for 2D feature extraction. We modified its final layer to produce a 256-dimensional feature vector instead of the 1000-dimensional output in the original version.

This 256-dimensional feature vector is then used as the input to the LSTM network for violent behavior classification. The features of high-risk regions from consecutive frames are fed into the LSTM to extract spatiotemporal features before being passed to a Softmax classifier to produce the final decision.

We employ a Bidirectional LSTM (Bi-LSTM) instead of a standard LSTM to enhance temporal context modeling. The architecture of a Bi-LSTM is illustrated in *Figure 4*.

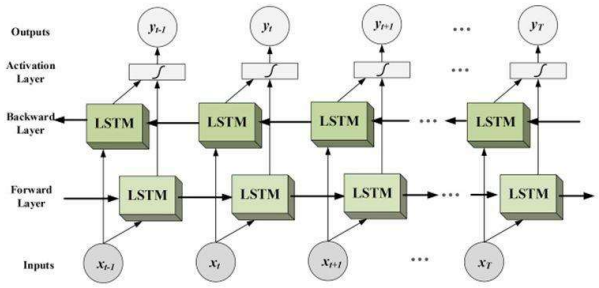

Fig. 4 BI-LSTM architecture

Unlike a unidirectional LSTM, a Bi-LSTM stores information from both the past and the future, allowing the model to make more informed predictions when processing sequences of violent behavior that require temporal information from both directions. Furthermore, we adopt a **two-layer LSTM architecture** because experimental results showed that it performs better than a single-layer LSTM. Increasing the number of layers beyond two yields only marginal accuracy improvements, while significantly increasing processing time

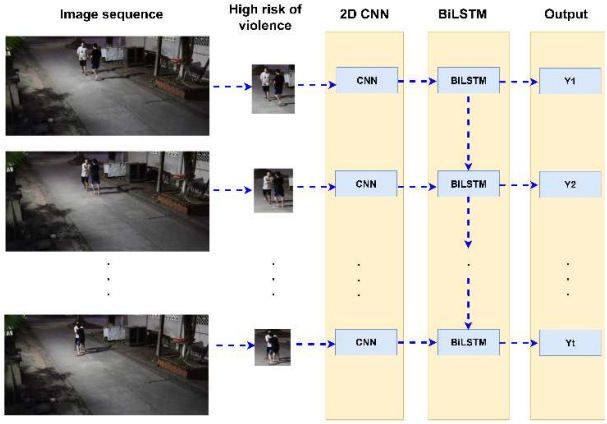

Fig. 5 CNN-BiLSTM Architecture for video classification

Figure 5, show the proposed CNN–BiLSTM fusion architecture for classifying image sequences to identify human violent behavior.

D. Experiment And Evaluation

1) Dataset

To evaluate the accuracy and effectiveness of the proposed algorithm, we conducted experiments on three datasets: Hockey Fight, Peliculas, and PTIT, as summarized in Table 1

No	Table 1. Experiment dataset		
	Dataset	***#violence***	***#non-violence***
1	Hockey Fight	500	500
2	Peliculas	100	100
3	PTIT VIOLENCE	120	90

Hockey Fight Dataset – This dataset contains both violent and non-violent scenes from ice hockey games. It consists of a total of 1,000 video clips, with 500 violent samples and 500 non-violent samples. Each video has a duration of 2 seconds, and all videos share the same frame resolution. In the violent samples, the violent action occupies most of the frame. The videos share a common background and contain background motion.

Peliculas Dataset – This dataset consists of both violent and non-violent scenes extracted from Hollywood movies, football matches, and various other events. It contains a total of 200 video clips, with 100 violent samples and 100 non-violent samples. Each video has a duration of 2 seconds; however, the frame resolutions vary across the dataset. In the violent samples, violent actions occupy most of the frame. The videos feature diverse environments and human subjects, and background motion is also present.

PTIT Dataset – This dataset was collected by us for research purposes at the Posts and Telecommunications Institute of Technology (PTIT). It consists of a total of 210 video clips, including 110 violent samples and 90 non-violent samples. All videos share the same frame resolution but vary in duration. They were recorded in different contexts and at varying distances from the camera, ranging from close-up to long-range shots.

2) Experiment result

The experiments were implemented in Python using the PyTorch deep learning framework, with the following hardware configuration:

- OS: Windows 10
- CPU: Intel Core i9-10900K
- RAM: 32 GB
- GPU: NVIDIA GeForce RTX 4060

The evaluation was conducted on all three datasets—Hockey Fight, Peliculas, and PTIT—using two CNN models (ResNet18 and VGG16 [15]) and two LSTM variants including standard LSTM and Bidirectional LSTM. To investigate the effect of the number of timesteps (i.e., the number of frames fed into the LSTM for prediction) on classification accuracy, we tested configurations with 10 and 15 timesteps. The datasets were split into 80% for training and 20% for testing. Experimental results, expressed in terms of classification accuracy, are presented in Table 2.

The processing times for ResNet18 with 10 and 15 timesteps were **60 ms** and **75 ms**, respectively, demonstrating suitability for real-time applications.

The results indicate that our proposed method—**Fight Region Candidate + ResNet18 + 2-layer Bi-LSTM**—achieves the best accuracy. This approach outperforms baseline CNN–LSTM models that do not include a preprocessing step for suspicious region detection. Furthermore, replacing ResNet18 with VGG16 yields negligible improvement in accuracy while significantly increasing the CNN's computational complexity.

The experiments also show that the **15-timestep Bi-LSTM** configuration provides better accuracy compared to the 10-timestep setting, while still meeting real-time processing requirements

A. Conclusion and future work

Automatic detection of violent behavior is essential for timely intervention, prevention, and early warnings, thereby reducing potential harm to human health, property, and psychological well-being.

This paper has presented an effective approach for violent behavior detection by integrating 2D CNN classification, motion entropy analysis, and a hybrid CNN–BiLSTM video classification framework. Experimental results demonstrate that the proposed method achieves superior performance compared to existing approaches, while maintaining low computational cost, making it fully suitable for real-time applications.

For future work, we plan to extend the proposed framework to handle more complex real-world scenarios, such as crowded environments and scenes with severe occlusions. We also aim to incorporate more advanced object tracking and background modeling techniques to further improve motion entropy estimation, as well as explore transformer-based architectures for enhanced spatiotemporal feature learning

In addition, we plan to collect larger and more diverse datasets to further improve the accuracy and robustness of the model

No	Table 2. Experiment Result									
	Model	*Hockey Fight dataset*			*Peliculas dataset*			*PTIT dataset*		
		Precision	*Recall*	*F1 Score*	*Precision*	*Recall*	*F1 Score*	*Precision*	*Recall*	*F1 Score*
1	Resnet18+2 LSTM 10 steps	0.93	0.94	0.94	0.86	0.89	0.87	0.84	0.84	0.83
2	Resnet18+2 LSTM 15 steps	0.95	0.95	0.95	0.87	0.91	0.9	0.82	0.86	0.85
3	Resnet18+2Bi-LSTM 10 steps	0.95	0.97	0.96	0.89	0.91	0.9	0.85	0.87	0.85
4	Resnet18+2Bi-LSTM 15 steps	0.96	0.97	0.97	0.9	0.92	0.87	0.87	0.88	0.87
5	FightRegion Candidate + Resnet18+2 LSTM 10 steps	0.96	0.97	0.95	0.9	0.91	0.9	0.9	0.92	0.92
6	FightRegion Candidate + Resnet18+2 LSTM 15 steps	0.97	0.97	0.97	0.93	0.93	0.9	0.93	0.93	0.93
7	**FightRegion Candidate + Resnet18+2 BiLSTM 10 steps**	**0.99**	**0.98**	**0.99**	**0.92**	**0.94**	**0.92**	**0.93**	**0.94**	**0.94**
8	**FightRegion Candidate + Resnet18+2 BiLSTM 15 steps**	**0. 99**	**0.99**	**0.99**	**0.99**	**0.97**	**0.97**	**0.97**	**0.99**	**0.98**

REFERENCES

[1] I. S. Gracia, O. D. Suarez, G. B. Garcia, and T.-K. Kim, "Fast fight detection," PLoS ONE, vol. 10, no. 4, Apr. 2015, Art. no. e0120448

[2] P. C. Ribeiro, R. Audigier, and Q. C. Pham, "RIMOC, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance," Comput. Vis. Image Understand., vol. 144, pp. 121–143, Mar. 2016.

[3] E. Y. Fu, H. Va Leong, G. Ngai, and S. Chan, "Automatic fight detection in surveillance videos," in Proc. 14th Int. Conf. Adv. Mobile Comput. Multi Media, Nov. 2016, pp. 225–234.

[4] [S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," in 2017 International Confe-rence on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEng-Technol.2017.8308186.

[5] Ralf C. Staudemeyer and Eric Rothstein Morris, "Understanding LSTM - a tutorial into Long Short-Term Memory Recurrent Neural Networks". arXiv, 2019.

[6] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, ''Violence detection in video by using 3D convolutional neural networks,'' in Proc. Int. Symp. Visual Comput., 2014, pp. 551–558.

[7] S. Sudhakaran and O. Lanz, ''Learning to detect violent videos using convolutional long short-term memory,'' in Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS), Aug./Sep. 2017, pp. 1–6.

[8] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, ''Violence detection using spatiotemporal features with 3D convolutional neural network,'' Sensors, vol. 19, no. 11, p. 2472, May 2019.

[9] Seymanur Akti, Gozde Ayse Tataroglu and Hazim Kemal Ekenel, "Vision-based Fight Detection from Surveillance Cameras". IEEE, 2019.

[10] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," in IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, Nov. 1997, doi: 10.1109/78.650093.

[11] Joseph Redmon, Santosh Divvala, Ross Girshick and Ali Farhadi, "You Only Look Once: Unified, Real-Time Object Detection". arXiv, 2016.

[12] Nicolai Wojke, Alex Bewley and Dietrich Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric". arXiv, 2017

[13] Alexey Bochkovskiy, Chien-Yao Wang and Hong-Yuan Mark Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection". arXiv, 2020.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, "Deep Residual Learning for Image Recognition". arXiv, 2015.

Karen Simonyan and Andrew Zisser-man, "Very Deep Convolutional Networks for Large-Scale Image Recognition". arXiv, 2015.

Nguyen Manh Dung received his Bachelor's degree in Electronics and Telecommunications from Hanoi University of Science and Technology in 2005. He obtained his Master's degree in Information Technology from Kongju National University in 2009, and his Ph.D. in Information Technology from Kongju National University in 2019. He is currently a lecturer and researcher at the Faculty of Electronics Engineering 1, Posts and Telecommunications Institute of Technology (PTIT). His research interests include image processing, computer vision, algorithms, and artificial intelligence

Soonghwan Ro received B.S., M.S., and Ph.D. degrees from the Department of Electronics Engineering at Korea University in 1987, 1989, and 1993, respectively. He was a research engineer of Electronics and Telecommunications Research Institute and University of Birmingham in 1997 and 2003, respectively. Since March 1994, he has been a professor at Kongju National University, Korea. His research interests include 5G communications, mobile networks, and embedded systems.