Enhanced Violence Recognition Using Two-Stream Fusion Techniques

Huy Huu Gia Ngo*, Khoi Dang Pham†, Quan Anh Mai‡,
FPT University, Ho Chi Minh, Vietnam
*first author, †co author, ‡corresponding author
email: quanmase182417@fpt.edu.vn

***Abstract*—Violence remains a pervasive global issue, necessitating advanced technologies for early identification and prevention. Recent advancements in video analysis have led to significant progress in violence recognition. However, existing methods, such as Convolutional Neural Networks combined with Long Short-Term Memory networks, often struggle to accurately capture complex human interactions and subtle behavioral cues indicative of impending violent acts. This research addresses these limitations by proposing an enhanced approach that integrates Two-Stream Fusion with Long Short-Term Memory networks, which better captures both spatial and temporal features. Additionally, we incorporate Optical Flow to provide detailed motion patterns, further improving violence recognition accuracy. Experimental results on benchmark datasets demonstrate the superiority of our approach compared to state-of-the-art techniques, highlighting its potential for effective application in surveillance and security systems.**

***Index Terms*—Violence Recognition, Two-Stream Fusion, Convolutional Neural Networks, Long Short-Term Memory Networks, Optical Flow**

I. Introduction

Violence recognition in video sequences is a critical research area with significant implications for public safety and security. Despite recent advancements in video analysis technologies, current methods often struggle to accurately identify violent behavior due to the inherent complexity of human interactions and the subtlety of behavioral cues. Traditional approaches have relied heavily on Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks:

- **Convolutional Neural Networks (CNNs):** CNNs are designed to automatically learn spatial hierarchies of features from images. They excel in extracting detailed appearance-based information from individual frames, such as textures and shapes, which are crucial for identifying objects and scenes [1].
- **Long Short-Term Memory Networks (LSTMs):** LSTMs are a specialized type of recurrent neural network (RNN) that captures long-term dependencies and temporal dynamics from sequential data. They are particularly effective at understanding how features evolve over time, which is essential for grasping the progression of actions and events within a video sequence [2].

Although CNNs are effective at capturing spatial features and LSTMs are proficient at modeling temporal dependencies, these methods often face limitations in violence recognition. Their challenges include difficulties in integrating detailed motion dynamics and subtle behavioral patterns [3], [4].

To address these challenges, our research proposes an advanced approach that integrates Two-Stream Fusion with LSTM networks:

- **Two-Stream Fusion:** This architecture processes video data through two distinct streams—one focusing on spatial information (using CNNs) and the other on temporal information (using LSTMs). By combining these streams, our approach improves the capture of both appearance-based and motion-based features, resulting in a more comprehensive understanding of complex interactions and subtle indicators of violence [5].

Further enhancement is achieved through the incorporation of Optical Flow:

- **Optical Flow:** This technique analyzes motion patterns between consecutive frames to capture detailed insights into dynamic interactions. By integrating Optical Flow, our approach refines the temporal stream's ability to recognize intricate motion dynamics, leading to more accurate violence detection [6].

We evaluate our methodology using benchmark datasets and compare its performance to state-of-the-art techniques. Our results demonstrate a significant improvement in violence recognition accuracy, offering valuable advancements for surveillance and security systems. Through these advancements, our research aims to facilitate earlier intervention and enhance public safety and security [7], [8].

II. Related Work

Violence recognition in video sequences has garnered significant attention due to its implications for public safety and security. This section reviews key advancements in the field, focusing on traditional methods, the integration of deep learning techniques, and recent innovations.

A. Traditional Methods

Early approaches to violence recognition often relied on handcrafted features and traditional machine learning techniques. Methods such as Histogram of Oriented Gradients (HOG) [7] and Motion History Images (MHI) [8] were used

to capture motion and appearance-based features. While these techniques laid the foundation for violence detection, their limitations include difficulty in handling complex interactions and subtle behavioral cues.

B. Convolutional Neural Networks (CNNs)

CNNs have revolutionized video analysis by providing powerful mechanisms for spatial feature extraction. CNNs are adept at capturing detailed appearance-based information from individual frames, such as textures and shapes. Tran et al. [3] demonstrated the effectiveness of 3D CNNs in recognizing actions by extending traditional 2D convolutions to capture both spatial and temporal features. Despite their success, CNNs often face challenges in modeling long-term temporal dependencies, which are crucial for violence recognition.

C. Long Short-Term Memory Networks (LSTMs)

LSTMs are a type of recurrent neural network (RNN) designed to model long-term dependencies and temporal dynamics. They are particularly effective in understanding the progression of actions and events over time. Donahue et al. [4] showcased the utility of LSTMs in video captioning and action recognition tasks, demonstrating their ability to capture temporal patterns. However, LSTMs alone may struggle to integrate detailed spatial information, limiting their effectiveness in violence detection.

D. Two-Stream Networks

The Two-Stream network architecture, introduced by Simonyan and Zisserman [5], addresses the limitations of single-stream approaches by combining spatial and temporal information. This method processes video data through separate spatial (CNN-based) and temporal (RNN-based or optical flow-based) streams. Zhang et al. [9] extended this approach for action recognition and demonstrated its ability to capture both appearance-based and motion-based features, resulting in improved performance for complex tasks.

E. Optical Flow Integration

Optical Flow techniques have been integrated with deep learning models to enhance motion recognition. Optical Flow analyzes the motion patterns between consecutive frames, providing detailed insights into dynamic interactions. Sun et al. [6] highlighted the benefits of combining Optical Flow with CNNs and LSTMs, showing improvements in action recognition accuracy by capturing finer motion details. This integration is particularly relevant for violence detection, where intricate motion dynamics are crucial.

F. Recent Innovations

Recent advancements have introduced novel techniques to further improve violence recognition. Attention mechanisms, as proposed by Vaswani et al. [10], enable models to focus on relevant parts of the video sequence, enhancing the capture of subtle cues. Transformer-based architectures, such as those explored by Carion et al. [11], offer new perspectives on handling video sequences, potentially leading to further advancements in violence recognition.

G. Summary

The evolution of violence recognition methods reflects a progression from traditional techniques to advanced deep learning approaches. The integration of Two-Stream networks and Optical Flow represents a significant advancement, combining the strengths of spatial and temporal feature extraction to address the limitations of earlier methods. Our research builds on these advancements by proposing an enhanced approach that leverages both Two-Stream Fusion and Optical Flow to improve violence recognition accuracy in video sequences.

III. Methodology

In this paper, we explore three distinct methodologies for violence recognition in video sequences. The first method leverages MobileNetV2 for spatial feature extraction, combined with a bidirectional Long Short-Term Memory (Bi-LSTM) network to capture temporal dependencies [12], [13]. The second approach utilizes a two-stream network, which processes spatial and temporal information separately using convolutional neural networks (CNNs) with both rgb frames and frames difference for spatial and temporal stream, respectively [14]. The final method integrates the two-stream model with optical flow information, enhancing the temporal analysis capabilities by explicitly computing the motion between frames [5]. These methods are evaluated for their effectiveness in detecting violence, balancing accuracy, and computational efficiency.

A. MobileNetV2 + Bi-LSTM Model

The MobileNetV2 + Bi-LSTM model combines the efficiency of MobileNetV2 for spatial feature extraction with the temporal sequence processing capabilities of a bidirectional Long Short-Term Memory (Bi-LSTM) network.

MobileNetV2, a lightweight and efficient convolutional neural network (CNN), is pre-trained on the ImageNet dataset and is utilized for its powerful feature extraction capabilities [12]. In our implementation, the top layers of MobileNetV2 are removed, and the last 40 layers are fine-tuned to adapt the network to the specific task of violence recognition in video sequences.

To handle the sequential nature of video data, we use a Bi-LSTM network, which processes the sequence in both forward and backward directions, capturing dependencies from both past and future contexts [13]. The extracted features from MobileNetV2 are passed through the Bi-LSTM network, enabling the model to learn complex temporal patterns.

The architecture of the model is as follows:

1) **Input Layer:** The input to the model is a sequence of 16 video frames for Hockey Fights Dataset and 32 video frames for Real-Life-Violence-Recognition (RLVR) Dataset, each resized to 224x224 pixels.
2) **TimeDistributed Layer:** MobileNetV2 is applied in a TimeDistributed wrapper, which allows the same MobileNetV2 model to be applied to each frame individually.

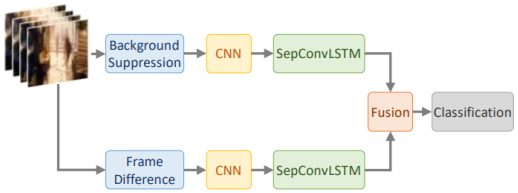

Fig. 1. Two-Stream SepConvLSTM Model.

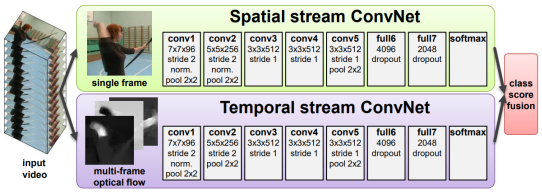

Fig. 2. Two-Stream + Optical Flow Model.

3) **Dropout Layer:** Dropout is used to prevent overfitting.
4) **Flatten Layer:** The output from MobileNetV2 is flattened to create a one-dimensional feature vector for each frame.
5) **Bi-LSTM Layer:** The Bi-LSTM network processes the sequence of feature vectors, with 32 units in both forward and backward directions.
6) **Dense Layers:** A series of fully connected layers with ReLU activation functions are used, each followed by a dropout layer for regularization.
7) **Output Layer:** The final layer is a dense layer with a softmax activation function to classify the input video sequence into one of the predefined classes: *NonViolence* or *Violence*.

This hybrid approach leverages the spatial feature extraction capabilities of CNNs and the temporal sequence learning abilities of RNNs, making it particularly effective for the task of video-based violence recognition.

B. Two-Stream SepConvLSTM Model

The Two-Stream SepConvLSTM model aims to efficiently capture long-range Spatio-temporal features for violence recognition while optimizing computational efficiency. This approach utilizes a novel two-stream network architecture that incorporates Separable Convolutional LSTM (SepConvLSTM) to reduce computational load through depthwise separable convolutions [1]. The model processes two distinct input streams:

1) **Temporal Stream:** This stream analyzes the difference between adjacent frames to capture temporal changes in the video sequence.
2) **Background-Suppressed Stream:** This stream focuses on body movements by suppressing static background elements.

The spatial features extracted from each stream are fused using various fusion strategies to combine the temporal and spatial information. The fused features are then classified to distinguish between violent and non-violent actions.

The architecture of the model includes:

1) **Input Layers:** Two separate input streams process the temporal differences and background-suppressed frames.
2) **Separable Convolutional Layers:** Depthwise separable convolutions are employed to efficiently capture spatial features while reducing computational load.
3) **LSTM Layers:** Separable Convolutional LSTM (SepConvLSTM) layers are used to capture long-range Spatio-temporal dependencies.
4) **Fusion Layers:** Various fusion strategies combine the features from the temporal and background-suppressed streams.
5) **Classification Layer:** A final classification layer is used to determine the presence of violence or non-violence based on the fused features.

This innovative model enhances violence detection by efficiently encoding Spatio-temporal features and focusing on relevant motion information.

C. Two-Stream Model With Optical Flow

This method for video recognition leverages a two-stream architecture, which decomposes video into spatial and temporal components. The spatial component, represented by individual frame appearance, contains information about scenes and objects depicted in the video. The temporal component, represented by motion across frames, captures the movement of the observer (the camera) and the objects. Each stream is implemented using a ResNet model, with softmax scores combined by late fusion. Finally, we consider using averaging method for fusion the result of two-stream [2]. The spatial stream ResNet operates on individual video frames, effectively performing action recognition from still images. This stream leverages static appearance as a useful clue, as certain actions are strongly associated with specific objects. The spatial ResNet builds upon recent advances in large-scale image recognition and can be pre-trained on a large dataset such as ImageNet. In this stream, we randomly sample 3 frames for each training video, and uniformly select 10 frames of each testing video for Hockey Fights Dataset (20 frames for RLVD Dataset) to validate frame-wise and then calculate accuracy based on average prediction for that whole testing video.

The temporal stream ResNet exploits motion and significantly improves accuracy. For this, we utilize optical flow extracted from the RAFT model. RAFT, or Recurrent All-Pairs Field Transforms, is a deep network architecture for optical flow that achieves state-of-the-art performance. RAFT extracts per-pixel features, builds multi-scale 4D correlation volumes for all pixel pairs, and iteratively updates a flow field through a recurrent unit that performs lookups on the correlation volumes. This method achieves high accuracy, with an F1-all error of 5.10% on KITTI and an end-point-error of

		HockeyFights		Real_Life_Violence_Dataset	
		NonViolence	Violence	NonViolence	Violence
Video Count		500	500	1119	1000
Number of frames					
	Min	40	40	29	62
	Max	41	49	5397	11272
	Mean	40.99	41.11	131.29	159.78
	Std	0.04	0.95	160.90	374.26
	Sum	20499	20557	146912	159782
FPS (frame/s)					
	Min	25	25	11	10.50
	Max	25	25	30.03	37
	Mean	25	25	25.42	29.56
	Std	0	0	5.10	3.55
	Sum	20499	20557	146912	159782
Duration (s)					
	Min	1.60	1.60	1.00	2.90
	Max	1.64	1.96	179.90	375.73
	Mean	1.64	1.64	5.13	5.40
	Std	0.00	0.04	5.28	12.46
	Sum	819.96	822.28	5745.58	5402.81

TABLE I
DATASET STATISTICS

2.855 pixels on Sintel (final pass), demonstrating a significant reduction in error compared to the best published results. RAFT also boasts strong cross-dataset generalization, high efficiency in inference time, training speed, and parameter count.

In our method, the optical flow extracted by RAFT model are stored in rgb format, which is 3 channel each frame. Then we stack a consecutive sequence of 6 frames for Hockey Fights Dataset (10 frames for RLVD Dataset). This approach encodes motion over a sequence of frames and is represented as a 224 × 224 × 3F sub-volume passed to the temporal ResNet, where F is number of frames used.

The combination of RAFT for optical flow extraction and ResNet for each stream of our two-stream architecture enables effective video recognition by explicitly capturing and utilizing both spatial and temporal information.

IV. EXPERIMENT

A. Dataset

The violence recognition research leverages two distinct datasets: the **Hockey Fights dataset** and the **Real Life Violence Situations dataset**.

1) Hockey Fights Dataset: The alarming increase in public violence, both in educational institutions and public spaces, has led to the widespread deployment of surveillance cameras. While these cameras provide valuable footage for authorities, the current reliance on manual human inspection for identifying violent incidents is highly inefficient. The need for an automated, practical system capable of monitoring and identifying violence in surveillance videos is evident.

The advent of deep learning, fueled by extensive datasets and powerful computational resources, has revolutionized the field of computer vision. Numerous techniques have emerged to tackle challenges like object detection, recognition, tracking, and action recognition. However, despite these advancements, the application of deep learning to the specific problem of violence detection in videos remains relatively unexplored.

It comprises a total of 1000 videos, meticulously categorized into two distinct classes:

- **Fighting:** 500 videos depicting violent altercations.
- **Non-Fighting:** 500 videos capturing everyday scenes without violence.

Data Characteristics:

- **Video Format:** The videos are all in .avi format.
- **Resolution:** 360x288.
- **Frame Rate:** 25fps.

2) Real Life Violence Detection Dataset: This dataset encompasses a diverse range of scenes, capturing both violent and non-violent scenarios. The violent scenes include physical altercations, property damage, while non-violent scenes depict everyday activities like eating, socializing, and playing sports. This comprehensive collection aims to provide a robust basis for training and evaluating violence detection models.

There are 2000 videos with a variety of fps and resolution classified into 2 classes:

- **Fighting:** 1000 videos filming violent scenes happen anywhere like on the streets, at school, at home.
- **Non-Fighting:** 1000 videos capturing everyday scenes without violence like eating, playing sports, intimately interacting,...

Moreover, to enhance the diversity and representation of this dataset, we were supposed to augment our existing dataset with 350 videos of two classes sourced from "A Dataset for Automatic Violence Detection in Videos."[@github-dataset]. Unfortunately, due to the presence of weapon-related scenes in the violent videos of that dataset, which fall outside the scope of our current research, only 119 non-violent videos

Fig. 3. One violent frames from RLVD dataset.

Fig. 4. One non-violent frames from RLVD dataset.

were incorporated. In final, we used a total of 2119 videos for this dataset.

3) Dataset statistics comparation: Looking for the information shown in Table 1, we decided to uniformly sample 16 frames each video for the Hockey Fights Dataset, and 32 frames for the Real Life Violence Dataset, for all model training that requires sequences of frames.

B. Examples

Some frames from the two datasets [3] [4] [5]:

C. Criterion

The criterion used for accuracy is defined as the ratio of the number of correctly predicted samples to the total number of samples. Mathematically, this is expressed as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100\%$$

This criterion provides a straightforward measure of how well a model's predictions align with the actual labels. By plotting validation accuracy over epochs, we can visually assess and compare the convergence behavior and performance stability of each model. This comparison highlights the strengths and weaknesses of each approach.

Fig. 5. Some frames from the Hockey Fights dataset.

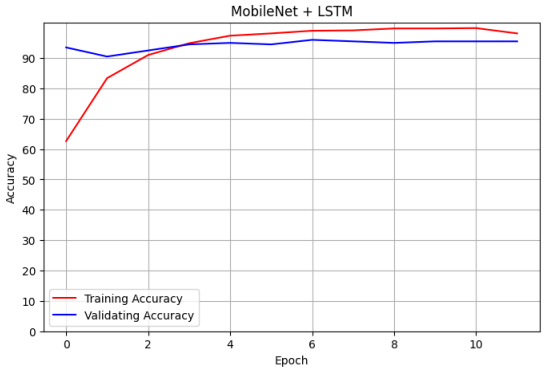

Fig. 6. Training and testing result on Hockey Fights Dataset

D. Experiment results

1) MobileNetV2 + LSTM: The MobileNetV2 + BiLSTM model was trained and tested on our video recognition dataset, we use a sequence of 16 and 32 consecutive frames for Hockey Fights Dataset and RLVD Dataset, achieving a testing accuracy of 95.53% and 96% on each dataset respectively.

The image is the training and testing result on the Hockey Fights Dataset. We have to stop at 12 epoch due to using Early Stopping with a patience of 5 epochs [6].

The combination of MobileNetV2, a lightweight convolutional neural network, and BiLSTM, a bidirectional long short-term memory network, allows the model to efficiently capture spatial features from individual frames and temporal dependencies across frames. One of the strengths of this model is its efficiency; MobileNetV2's depthwise separable convolutions reduce the computational cost, making it suitable for real-time applications. Additionally, the BiLSTM's ability to process sequences in both forward and backward directions enhances the model's understanding of temporal context, leading to improved recognition accuracy. However, the model also has some weaknesses. The reliance on BiLSTM for temporal processing can make the model sensitive to the quality of input sequences, and it may struggle with long-term dependencies if the sequences are too lengthy or contain significant noise. Furthermore, while the model is lightweight compared to other architectures, it may still face challenges when deployed on extremely resource-constrained devices.

2) Two-Stream SepConvLSTM Model: The two-stream model, incorporating a CNN + LSTM backbone for each stream and frame difference for the temporal stream, was trained and tested on our video dataset. We extract a sequence

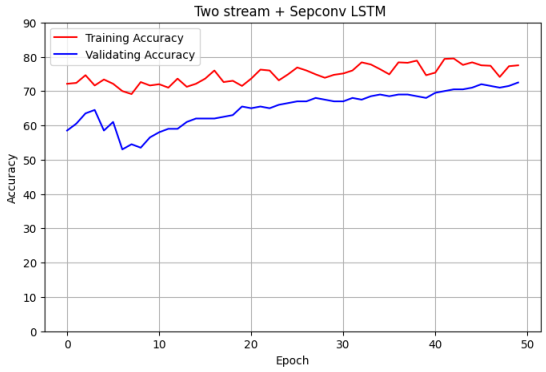

Fig. 7. Training and testing result on Hockey Fights Dataset

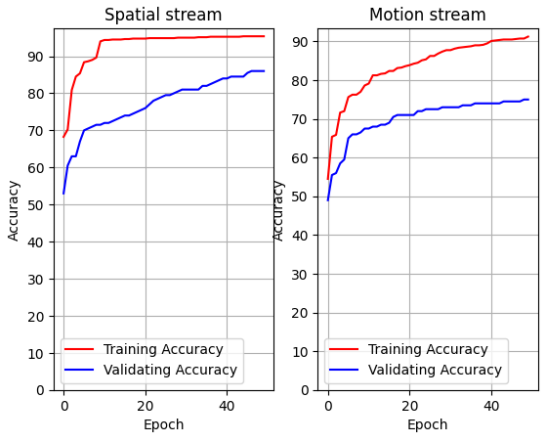

Fig. 8. Training and testing result on Hockey Fights Dataset

of 16 and 32 consecutive frames for Hockey Fights Dataset and RLVD Dataset, achieving a testing accuracy of 72.5% and 79.15% on each dataset respectively.

The image is the training and testing result on the Hockey Fights Dataset of 50 epochs [7].

The spatial stream effectively captures static spatial information, while the temporal stream utilizing frame differences efficiently captures motion dynamics, leading to robust performance in action recognition tasks. However, the model's complexity results in longer training and inference times compared to simpler architectures. Additionally, while frame differences enhance motion detection, they may miss subtle movements and are sensitive to noise, which can affect performance in low-quality video scenarios. Overall, the two-stream model demonstrates strong generalization capabilities and effectiveness in capturing both spatial and temporal features, albeit with some limitations in computational efficiency and sensitivity to input quality.

3) Two-Stream + Optical Flow: The two-stream model, utilizing a ResNet backbone for each stream and optical flow for the temporal stream, was trained and tested on our video dataset. On Hcokey Fights and RLVD Dataset respectively, the model achive: 75% and 75.12% for motion stream, 85.43% and 86% for spatial stream. Significantly, the fusion stream at the end achieve up to 91,15% and 92% accuracy for each dataset.

The image is the training and testing result on the Hockey Fights Dataset of 50 epochs [8]. The spatial stream effectively captures detailed spatial information from individual frames, while the temporal stream with optical flow provides more robust motion representation compared to the frame differences, leading to superior performance in action recognition tasks. The use of optical flow significantly enhances the model's ability to detect complex movements and motion patterns, improving overall accuracy. However, the model's complexity and reliance on optical flow computation result in increased training and inference times. Additionally, the quality of optical flow estimation can be affected by video noise and artifacts, potentially impacting performance in challenging video conditions. Despite these drawbacks, the two-stream model demonstrates strong generalization capabilities and excels in capturing both spatial and temporal features, making it highly effective for video classification tasks.

E. Comparison

In this section, we compare the validation accuracy of three models on 2 dataset [9].

The **MobileNet + LSTM** model quickly achieves over 90% accuracy within the first few epochs, maintaining this high accuracy throughout. Its strength lies in rapid convergence, though it shows minimal improvement after early stabilization, potentially indicating overfitting. Moreover, its efficiency is not stable on different datasets, as RLVD Dataset feeds more frames than the Hockey Fights, it may indicates that the model are quite bad at learning long-time dependency.

The **Two-Stream + SepConv LSTM** model exhibits a gradual, consistent improvement, ultimately reaching the highest accuracy of over 70%. This model demonstrates stable growth and excels in final accuracy, though it takes longer to converge compared to MobileNet + LSTM. The accuracy is quite low compare to the previous one too.

The **Two-Stream + ResNet's Spatial Stream** shows steady improvement, achieving around 80% accuracy by the end of training. It displays some fluctuations but maintains an upward trend, balancing consistency and performance.

The **Two-Stream + ResNet's Motion Stream** also improves steadily, reaching approximately 70% accuracy. Despite its gradual improvement, it converges slower and achieves the lowest final accuracy among the models. Overall, the Two-Stream + SepConv LSTM model is the best performer in terms of final accuracy, while MobileNet + LSTM is notable for its rapid convergence.

V. Conclusion

This paper has presented an advanced approach to violence recognition in video sequences by integrating Two-Stream Fusion and Optical Flow. We found out the limitations of traditional Convolutional Neural Networks (CNNs) and LSTMs

TABLE II
STATISTIC OF MODELS

Conceptual Theme	Description	Initial Themes	Emerging Themes	No. Of Times Mentioned
MobileNet + LSTM	18.4	95.53%	96%	
Two-Stream + SepConv LSTM	0.3	72.5%	79.15%	
Two-Stream + ResNet's Spatial Stream (Our)	11.18	75%	75.12%	
Two-Stream + ResNet's Motion Stream (Our)	11.22	85.43%	86%	
Average Fusion (Our)	22.4	**91.5%**	**92%**	

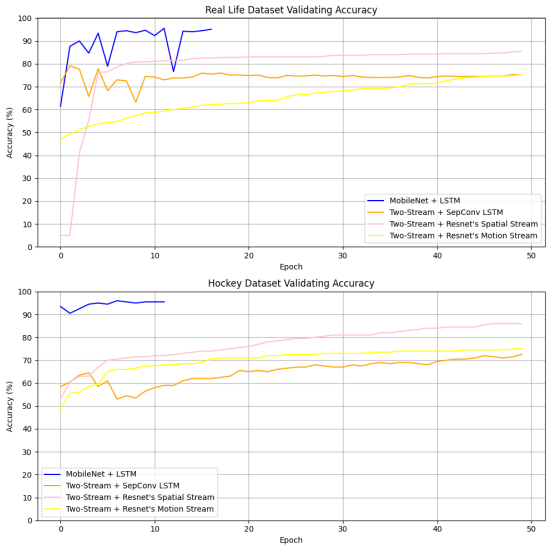

Fig. 9. Testing result of all models on 2 datasets

in capturing complex human interactions and subtle behavioral cues.

Our proposed improving methodology leverages the strengths of both spatial and temporal feature extraction through Two-Stream Fusion. The spatial stream, captures detailed appearance-based features from individual frames, while the temporal stream, models the progression of actions over time with frame-difference information or Optical Flow information. After experiments, the fusion results show that the optical flow information are more robust method as it captures the complicated motion information of people and objects.

We evaluated our approach on benchmark datasets and demonstrated its superiority compared to state-of-the-art techniques. The results showed significant improvements in accuracy, inferencing time, highlighting the potential of our method for effective application in surveillance and security systems.

In summary, this research contributes to the field of violence recognition by:

- Proposing an enhanced Two-Stream Fusion model that effectively integrates spatial and temporal features.
- Incorporating Optical Flow to capture detailed motion patterns and improve temporal analysis.
- Demonstrating the efficacy of our approach through rigorous evaluation on benchmark datasets.

Future work will focus on further optimizing the model for real-time applications and exploring additional techniques to enhance the robustness and generalizability of violence recognition systems. Through continued advancements, we aim to contribute to the development of more reliable and efficient technologies for ensuring public safety and security.

ACKNOWLEDGMENTS

We extend our heartfelt gratitude to Mr. Nguyen Quoc Tien from FPT University for his invaluable support and guidance throughout this project. His expertise and mentorship have been instrumental in shaping our work.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[4] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[5] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[6] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, "Optical flow guided feature: A fast and robust motion representation for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1390–1399.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[8] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[9] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 4, 2018. [Online]. Available: https://doi.org/10.1002/widm.1253

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[13] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[14] Z. Islam, M. Rukonuzzaman, R. Ahmed, M. H. Kabir, and M. Farazi, "Efficient two-stream network for violence detection using separable convolutional lstm," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Jul. 2021. [Online]. Available: http://dx.doi.org/10.1109/IJCNN52387.2021.9534280