Violence Detection Model

Neeraj Upadhyay [1]
IT Department
IPEC, Ghaziabad, India
neerajofficialnu@gmail.com

Mohit Karakoti [2]
IT Department
IPEC, Ghaziabad, India
mohitkarakoti7777@gmail.com

Ms. Tanya Sharma [3]
Ass. Prof. (IT Department)
IPEC, Ghaziabad, India
tanya.sharma@ipec.org.in

ABSTRACT

A vital field of machine intelligence with uses in surveillance and public safety is violence detection. The accuracy and limitations of several models, such as SVM, 3-D CNN,oriented violent flow , Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid techniques, are examined in this paper. Attention methods and transformer-based topologies have demonstrated the potential to enhance detection performance. Despite progress, challenges including dataset bias, environmental changes, and high processing costs still exist. Real-time processing and attaining robustness in a variety of settings are challenges for many models. This study compares the pros and cons of the current methodology through an organized evaluation. We provide a explanatory evaluation of their accuracy using established datasets available on the internet. We present extensive research of their accuracy on popular datasets. This report also highlights drawbacks in the current body of knowledge and suggests future direction. Reliability, scalability, and efficiency must be improved for systems that detect anomalies.

Keywords: action recognition, fight detection, video surveillance ,CNN , Deep Learning .

I. INTRODUCTION

Violence has been sharply rising in this day and age, endangering people, infrastructure, and systems. When violence occurs in public, most people are not held accountable and cannot be held responsible without evidence, which makes the issue worse. Because they are so unusual, the majority of the horrible crimes occur in public. The term "violent activities" typically refers to an atypical physical encounter between two or more individuals [11]. Security personnel now face many challenges as a result of surveillance monitoring, as they must meticulously go through the footage to identify the offender in particular, follow his movements from camera to camera, or watch it in real-time to identify violent behaviors and activities before or as they happen. The caliber of the surveillance footage they are given is another significant barrier to this. Since the defendant will contest their existence in the picture, the majority of surveillance film snapshots are not admissible in court.

Relying on a human to keep an eye out for and identify violent incidents is ineffective in urban settings where violence can happen at any moment. Such acts typically result in really uncomfortable situations, which is why it is so important that they be automatically detected by real-time video footage so that the relevant authorities can make the necessary, important choice. Consequently, the concept of putting in place tools and systems to identify such occurrences through real-time monitoring and video retrieval has been presented. Eliminating the afore mentioned practical limitations and effectively reducing crime rates are the main goals.

Technology has transformed our world and daily lives. We may now combine technology to detect objects and movements, creating a system for detecting possible threats. A vast quantity of research on human action recognition has been established in recent years detection research: hockey fight dataset and movies fight [14].Recent research has established a significant body of knowledge on recognizing human actions [12]. Deep learning outperforms traditional handcrafted approaches for identifying video sequences [20,13]. Research on violence detection often uses two prominent datasets: movies and hockey [14].

Here violence detection is using supervised learning [16] in which these models are trained on labelled datasets in which there are clips of violence and non-violence or fight and no fight . Some of widely used datasets are hockey fight dataset ,Real life violence dataset , Movie fight dataset etc.

These datasets have a bunch of videos either they are taken from a recording or from a CCTV footage and then later they are clipped at very short lengths typically 5 seconds in which action is happening[15] .

These models typically use neural networks for capturing spatial features like movement, fight,non-fight, aggression, etc and these frames are processed to identify the pattern in action[19]. Also RNN can be used for capturing temporal dependencies which helps models to understand past frames and future frames [17] .In this paper, we would be doing a study on these models and creating a literature survey which would be used in further implementation paper.

II. LITERATURE SURVEY

Violence detection is a core concept to detect violence in society. With advancements in technology especially in artificial intelligence we have found that violence detection has changed rapidly over the last decade. In artificial intelligence, there are various models and architectures that we can train to detect anomalies**.** In [1] we have studied a sturdy model that uses 3DCNN+LSTM which gives us an accuracy of 87.20%. This model uses DANIS-16 Dataset. But this model has faced problems in handling movie camera scenarios [1].

Another paper “A review on state of the art violence detection techniques” uses Hough Forest methodology for violence recognition [2]. That model showed 84%-96% accuracy on multiple datasets like UCF, hockey fight, Media eval, etc [2]. It worked completely fine in less crowded frames but face problems in detection in the crowded frame.

3D Conv LSTM-CNN was used on HIS dataset which shows 97-99% accuracy but in this model, OA decreases with a reduction in no of training [3]. [4]use CNN and RNN for anomaly detection and it gives a whooping 92-98% accuracy. This model usesd UCF crime dataset but it gave too many false positives . In [5]Vif+oVif, a hybrid model using 2 technologies gave an accuracy of 92-98%on hockey fight, violent flow dataset but the major problem was that this model was not working fine in crowded places.

In [6] we saw a different hybrid model using HOG+SVM again, the Hockey fight dataset with an accuracy of 89% but the only problem was that it had a problem detecting the complex motion. [7]uses HOG+Random Forest in which random forest is used as a classifier for anomaly detecting. Again it uses the Hockey fight dataset with an accuracy of 86%. Here the only concern was accuracy because it show less accuracy compared to models that leverage deep learning.

3D CNN along with SVM was used in [8]on variety of datasets like the Hockey fight, crowd violence, and Movie violence dataset. This model shows an accuracy of 98%. But the only shortcoming was that it is computationally very expensive.

[9] uncovers CNN along deep audio features which was tested on a different dataset Media eval
2015 . In this paper, we didn’t get any accuracy but it showed very high false positives. Finally in [10] we used vit along with structured neural learning ang gave accuracy of of 99%. This model is tested on various dataset like RWF,UCF, XD Violence, UBI Fight etc that’s why accuracy fluctuate and change very much with different dataset.

III. SUMMARY OF LITERATURE SURVEY

Ref.	Title	Model Used	Accuracy	Dataset	Shortcoming
[1]	A 3D CNN-LSTM-based image-to-image foreground segmentation	3D CNN+LSTM	87.20%	DANIS-2016	Lack capability of handling moving camera scenario
[2]	A Review on State-ofthe-art Violence Detection Techniques	Hough forest methodology for recognition	84%-96%	UCF , hockey fight,Media eval etc	Works fine in less crowded frames but do not work fine in crowded scenarios
[3]	Hyperspectral Image Classification via a Novel Spectral–Spatial 3D ConvLSTM-CNN	3DConvLSTM-CNN	97%-99%	HSI Dataset	OA decreases with reduction in no of training

[4]	KIANNET: an attentionbased CNN-RNN Model	CNN-RNN	92.98%	UCF Crime	Overfitting, too many false positive
[5]	Violence detection using Oriented VIolent Flow	Vif+oVif	87%-88%	Hockey fight, violent flow dataset	Appropriate for only non-crowd scenario
[6]	Suspicious and Violent Activity Detection of Humans	HOG +SVM	89%	Hockey fight	Have a problem detecting the complex motion
[7]	Violence Detection from Videos Using HOG	HOG+ Randomforest	86%	Hockey fight	Less accuracy compared to deep models
[8]	Violence Detection in Videos by Combining 3D Convolutional Neural Networks and Support Vector Machines	3D CNN+SVM	98%	Hockey+crowd violence+ Movie violence dataset	Computationally it is very expensive
[9]	Violent Scene Detection Using Convolutional Neural Networks and Deep Audio Features	CNN+Deep audio features	Not mentioned	Media Eval 2015	False positive score is very high
[10]	Transformer and Adaptive Threshold Sliding Window for Improving Violence Detection in Videos	Vit+Structured Neural learning	99%	RWF,UCF,X-D Violence,UBI Fight etc	Accuracy fluctuate with different datasets drastically

IV. RESULT

After the above summary of the literature survey now we have found that the model's accuracy is determined by 2 major factors, first the dataset and another model technology. As we have seen models that use HOG along with SVM or Random Forest classifier get an accuracy of 86-89% but they were having problems in detecting complex motion. As advanced AI models are used like 3D CNN, LSTM, and 3DConv-LSTM we found that accuracy increases to 98% But some models have problems of overfitting and some have problems of being computationally expensive. So we found that if we want to make a robust model then we have to use AI models that are computationally inexpensive. Along with that, we have to use an improved and precise dataset because the dataset can change the accuracy of the model drastically. We would focus on building a model on pre trained model and try to achive maximum accuracy.

V. CONCLUSION

In this review, we finally have examined a number of precise and updated methods for anomaly detection, such as CNN with deep features, 3D CNN with SCN, CNN-LSTM architecture, oriented violent flow, and SCM-based model that were promising. Along with their advantages and sturdiness, there are still issues in those models, which include limitations in real-time processing, dependencies, on countable data present on the internet, high computational complexity, and sensitivity to environmental factors. Changes can be made after we have done the examination of methods along with their drawbacks. In order to improve our review, emphasize a hybrid model along with fine-tuning. We state that accuracy, reliability, efficiency, and adaptability can still be enhanced. Further studies and advancement may result in more reliable and expendable violence detection models that can handle complex real-world scenarios. Also we would be using a bunch of performance metrics[18] in our implementation paper to support our claims for our model .

VI. REFERENCES

[1] T. Akilan, Q. J. Wu, A. Safaei, J. Huo, and Y. Yang, “A 3D CNN-LSTM-Based Image-to-Image Foreground Segmentation,” IEEE Transactions on Intelligent Transportation Systems, vol. XX, no. X, pp. XX–XX, Feb. 2019, doi: 10.1109/TITS.2019.2900426

[2] M. Ramzan, A. Abid, H. U. Khan, S. M. Awan, A. Ismail, M. Ahmed, M. Ilyas, and A. Mahmood, "A Review on State-of-the-Art Violence Detection Techniques," IEEE Access, vol. XX, pp. 1-1, 2017, doi: 10.1109/ACCESS.2019.2932114

[3] G. Farooque, L. Xiao, J. Yang, and A. B. Sargano, “Hyperspectral Image Classification via a Novel Spectral–Spatial 3D ConvLSTM-CNN,” Remote Sensing, vol. 13, no. 21, pp. 4348, Oct. 2021, doi: 10.3390/rs13214348

[4] S. Ahmadi Vosta Kolaei, "KianNet: An attention-based CNN-RNN model for violence detection," Ph.D. dissertation, Faculty of Graduate Studies and Research, University of Regina, Regina, SK, Canada, Apr. 2024. [Online]. Available: https://ppl-ai-file-upload.s3.amazonaws.com/web/directfiles/46387881/e872912b-3af0-4ced-856a-dfeb6f281afe/KIANNET-AN-ATTENTION-BASED-CNNRNN-MODEL.pdf.

[5] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using Oriented VIolent Flows," *Image and Vision Computing*, vol. 46, pp. 1-10, 2016, doi: 10.1016/j.imavis.2016.01.006.

[6] P. K. Roy and H. Om, "Suspicious and Violent Activity Detection of Humans Using HOG Features and SVM Classifier in Surveillance Videos," Advances in Soft Computing and Machine Learning in Image Processing, Studies in Computational Intelligence, vol. 730, pp. 277–294, Springer, 2018, doi: 10.1007/978-3-319-63754-9_13.

[7] S. Das, A. Sarker, and T. Mahmud, "Violence Detection from Videos using HOG Features," in *Proceedings of the 4th International Conference on Electrical Information and Communication Technology (EICT)*, Khulna, Bangladesh, Dec. 20-22, 2019, pp. 1-6. doi: 10.1109/EICT48899.2019.9036102.

[8] S. Accattoli, P. Sernani, N. Falcionelli, D. N. Mekuria, and A. F. Dragoni, "Violence Detection in Videos by Combining 3D Convolutional Neural Networks and Support Vector Machines," *Applied Artificial Intelligence*, vol. 34, no. 4, pp. 329-344, 2020. doi: 10.1080/08839514.2020.1723876.

[9] G. Mu, H. Cao, and Q. Jin, “Violent Scene Detection Using Convolutional Neural Networks and Deep Audio Features,” in Proc. CCPR 2016, Part II, CCIS 663, Singapore: Springer, 2016, pp. 451–463. DOI: 10.1007/978-981-10-3005-5_37

[10] F. J. Rendón-Segador, J. A. Álvarez-García, and L. M. Soria-Morillo, " Transformer and Adaptive Threshold Sliding Window for Improving Violence Detection in Videos" Sensors, vol. 24, no. 5429, 2024. doi: 10.3390/s24165429

[11] Marszalek, I. Laptev, and C. Schmid, "Actions in Context," in 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 2009, pp. 2929–2936

[12] S.-R. Ke et al., "A review on video-based human activity recognition," Computers, vol. 2, no. 2, pp. 88–131, Jun. 2013, doi: 10.3390/computers2020088.

[13] A. B. Sargano, P. Angelov, and Z. Habib, "A Comprehensive Review on Handcrafted and LearningBased Action Representation Approaches for Human Activity Recognition," Appl. Sci., vol. 7, no. 110, pp. 1–37, Jan. 2017

[14] E. Bermejo, O. Deniz, G. Bueno, and R. Sukthankar, "Violence Detection in Video Using Computer Vision Techniques," *E.T.S.I.Industriales, Universidad de Castilla-La Mancha and Intel Labs Pittsburgh*, 2013.

[15] M. Perez, A. C. Kot, and A. Rocha, "Detection of real-world fights in surveillance videos," in 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2662– 2666.

[16] Batta Mahesh, "Machine Learning Algorithms - A Review," International Journal of Science and Research (IJSR), vol. 9, no. 1, Jan. 2020. doi: 10.21275/ART20203995.

[17] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential Recurrent Neural Network for Action Recognition," International Journal of Computer Vision, vol. 128, no. 3, pp. 804-820, 2020.

[18] E. Beauxis-Aussalet and L. Hardman, "Visualization of confusion matrix for non-expert users," in Proceedings of the IEEE Symposium on Information Visualization (IEEE InfoVis), 2014.

[19] J. Yang and J. Li, "Application of deep convolution neural network," International Centre for Wavelet Analysis and Its Applications, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, 2025.

[20] D. Wu, N. Sharma, and M. Blumenstein, “Recent Advances in VideoBased Human Action Recognition using DeepLearning: A Review,” International Joint Conference on Neural Networks (IJCNN), 2017, pp. 2865-2872