2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) | 978-1-6654-9400-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAECT57570.2023.10117775

# Violence Detection in Real Life Videos using Deep Learning

Bhaktram Jain
*Department of Networking and Communications*
*SRM Institute of Science and Technology*
Kattankulathur, India
bj5804@srmist.edu.in

Aniket Paul
*Department of Networking and Communications*
*SRM Institute of Science and Technology*
Kattankulathur, India
ap6919@srmist.edu.in

P Supraja
*Department of Networking and Communications*
*SRM Institute of Science and Technology*
Kattankulathur, India
suprajap@srmist.edu.in

***Abstract* — These days, it is essential to avoid or identify violence as soon as possible because it is spreading in an unpredictable way. Violence must be identified on real-time films taken by numerous surveillance cameras at all times and in all locations, which makes it difficult to perform. It might be time-consuming to continuously monitor CCTV surveillance cameras, thus it is essential to automatically spot any unusual activity. As soon as violent behaviours take place, it should be able to make a trustworthy real-time detection and notify the appropriate authorities. The dataset contains both violence and non-violence videos from real life situations. A methodology for detecting violence has been presented by us that uses a network similar to the U-NET with the encoder mobilenetv2 to extract spatial features before moving on to an LSTM block for the extraction of temporal features and binary classification. The results of the trial revealed that the precision is 95% and the accuracy is 94% utilising a dataset based on real life situations. The recommended model uses minimal computer resources while yet producing useful results.**

***Keywords — Violence Recognition, Intelligent Video Monitoring, U-NET, LSTM, Computer Vision.***

## I. INTRODUCTION

The prevalence of violence in daily life has long been considered one of the main problems. Any society's peace and harmony are quickly destroyed by it. However, there was a significant decrease in crime between 2014 and 2017. However, it began to climb once more in 2017. In 2017–18, there was a rise of 6.79%. Numerous variables contribute to violent behaviour in public places. The root causes of violence include societal and economic instability, together with personal greed, anger, and hatred. To address this issue, violence that is either anticipated or unanticipated has to be discovered early and stopped as soon as it is practical.

Human behaviour has lately been studied using deep learning and computer vision. Signal analysis allows one to quickly identify a pattern that results from illegal thought. It hasn't been done yet since it isn't technologically feasible. With the use of deep learning based computer vision, we can now quickly identify aggressive behaviour in public places. Currently, most companies and organizations use CCTV systems. Understanding violence and putting a stop to its harmful consequences, the government or policy-making authority may find it useful to use effective violence detection tools. All humans and members of society want safe streets, neighbourhoods, and workplaces. Explicit feature engineering is not a part of deep learning. Machine learning cannot compete with it.

## II. THE CHALLENGES OF VIOLENCE DETECTION IN VIDEOS

Violence detection has garnered interest from a wide range of people and has made significant advancements, but the detection algorithms are still in their infancy, and there aren't any now that work in every situation. Additionally, there are still a lot of significant issues in this subject that need to be resolved. The following factors primarily highlight the research's challenges:

### *A. Changes in the Background Environment*

The main challenge in a variety of computer vision applications might be attributed to the influence of variables like backdrop environment. The variety of viewpoints is primarily present. Different two-dimensional representations can be produced from the same movement when it is seen from various angles; The reciprocal occlusion between individuals and between individuals and the background makes feature

extraction challenging. Other impacting elements include dynamic environmental changes, crowded backdrops, adjustments to the lighting, and low-resolution photos.

### *B. Differences Between Classes and within Classes*

Even the same action can have distinct expressions for the majority of acts. Running, for instance, can be done in a variety of background settings, at various speeds ranging from slow to fast, and with various step lengths. Some non-periodic activities, such as running during a battle and escape, which is visibly different from the regular periodic running, have similar effects to other acts. It is evident that there are numerous action types and variations, which makes the study of behaviour recognition quite challenging. Additionally, various people's performance for the same action can vary substantially.

### *C. The Complexity of the Target Subject*

The behaviour can be classified as simple solitary behaviour, interactive behaviour, or group behaviour depending on how many subjects are involved. Simple action recognition is often a behaviour of a single person. such as sprinting, jumping, and waving. Hugging and touching have entailed the interaction of two different people, which are slightly more complex behaviours. Fighting and other violent behaviours are frequently created with two or more persons. The complexity of the interactions between several issues likewise gets more complicated.

### *D. Complex Movement Patterns*

The speed and direction of the free movement are typically the only two factors that determine the movement style. Given that violent behaviour frequently takes place in a little amount of time and that the violent subject moves at a rapid pace, which also causes rapid changes in direction and confusion in the activities of various subjects, modelling violent behaviour's motion patterns is extremely challenging. In response to the aforementioned issues and challenges with violence detection, we review numerous efficient violence detection algorithms and research current approaches. Below, these techniques will be thoroughly explained.

## III. Related Work

Security camera footage is examined employing violence identification techniques derived from computer vision. Over the past several years, a number of sites have been equipped with these cameras and other surveillance equipment for the purpose of maintaining crime prevention to keep an eye on people's movements in areas like schools, hospitals, banks, marketplaces, and streets. Identifying whether a person's behaviour is acceptable or questionable is a component of behaviour analysis.

Over the past several years, a number of sites have been equipped with these cameras and other surveillance equipment for the purpose of maintaining crime prevention. The importance of violence detection in the video cannot be overstated because of the wide range of applications it has, including improving citizen security, preventing youngsters from acting violently, identifying threats, reducing first responders' response times, etc. Therefore, it is crucial to examine violent behaviours in people utilising surveillance footage. As a result, this section addresses the various approaches and strategies employed in earlier studies with a particular emphasis on how to spot violent behaviour in surveillance footage. The detection of violent actions from video collections has recently benefited significantly from the introduction of various artificial intelligence techniques. To further explain the advancement we divided this section into two pieces to better reflect our study on the diagnosis of violent behaviour.

In a noteworthy work by Deniz et al. [1], rapid movement styles were created by applying random transformations to the load spectrum of subsequent video frames, they were subsequently utilised as the major ingredient of the model to pinpoint violent behaviour. Studies show a 12% improvement in accuracy when compared to state-of-the-art techniques for detecting violent situations.

Using temporal and multi-modal data, Penet et al. [2] investigated the various Bayesian learning techniques for a violent detection.

Suarez et al. [3].'s investigation of the effectiveness of classifiers for machine learning to identify a brief combat based on three video datasets. In a separate work, Fu et al. [4] presented a minimal computation method for automatically detecting fights and the optical flow and BoW techniques were used as the foundation for the study's two feature extraction models.

Using spatial-temporal characteristics, a deep learning three stage violence identification system was recently suggested by Ullah et al. [5] in investigations. Three publicly accessible datasets were analysed, and experimental results indicate that the Hockey Fights dataset produced the best results in terms of accuracy.

For the purpose of identifying aggressiveness in video characteristics, a CNN-LSTM conjunction was innovated. The works use an LSTM derivative by Sumon et al. [6] and [7,8,9] to categorise gathered traits as violent or non-violent. Localization of spatiotemporal information included in the video enables local motion analysis by combining CNN and LSTM.

## IV. PROPOSED MODEL

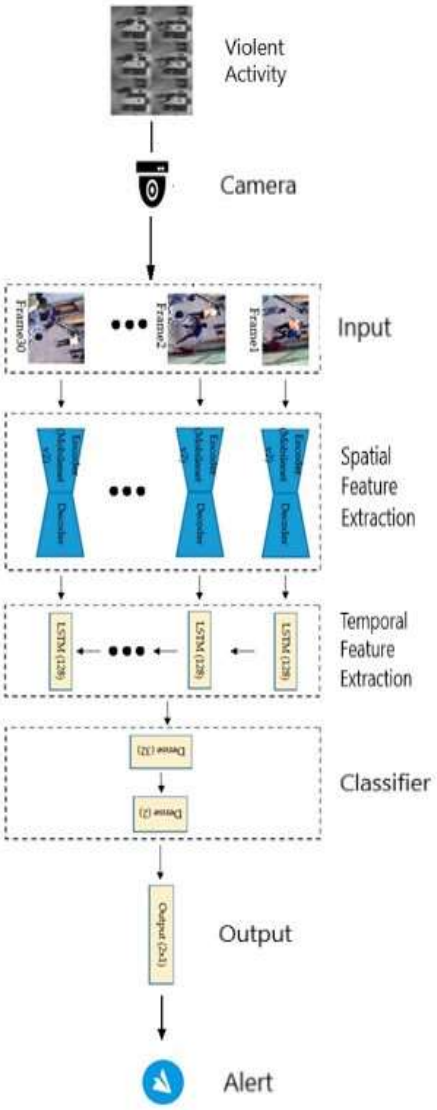



Figure 1. *Proposed model architecture diagram*

Reducing computing difficulty for implementation on low end devices is the suggested model's main goal while retaining performance comparable to cutting-edge violence detection techniques.

The proposed model first employs an extractor of spatial and temporal features, then categorization. The U-NET design, seen in Figure 2, comprises 23 convolutional layers, works with very few training samples, and offers improved performance for segmentation tasks. It permits the simultaneous use of global location and context.

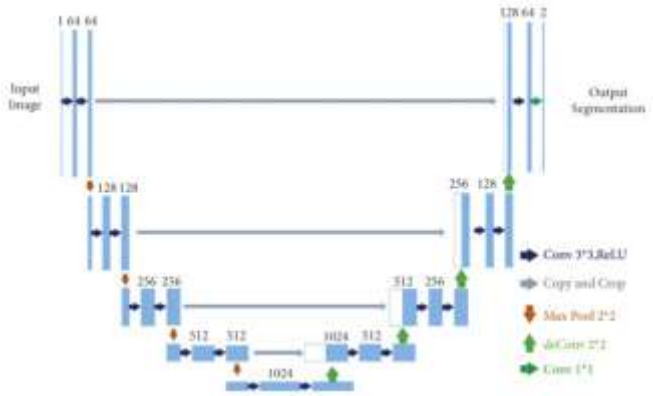


Figure 2. *U-Net and MobileNet V2 network model*

As a result, the second stage receives a fresh queue of frame characteristics to extract temporal information from. The sequential information between successive video slices is obtained using LSTM. Using this information, violent or non-violent incidents are classified using a 2 layer classifier with dense layer. Figure 1 depicts the intended model's architecture. The predictor for the method we employed was MobileNet V2, a straightforward modern classifier for extracting spatial features.

Table 1. TOTAL PARAMETERS

| Layer | Output Shape | No. of Parameters # |
|---|---|---|
| Time distribution (U-Net features extractor) | (30, 64, 64, 1) | 1,907,041 |
| LSTM | (128) | 2,163,200 |
| Dense | (32) | 4128 |
| Dense | (2) | 66 |
| Total parameters: 4,074,435<br>Trainable parameters: 3,457,219<br>Nontrainable parameters: 617,216 | | |

The binary cross-entropy loss function was chosen as the error function in this study because the successfully performs the task of binary classification for the dataset containing videos. The equation is shown below:

$$BCE = -\frac{1}{output\ size} \sum_{i=1}^{output\ size} y_i \cdot \log\log \dot{y_1} + (1 - y_i) \cdot \log\log(1 - \dot{y_1})$$

In this instance, $y_i$ stands for the label or class.
$\hat{y}$ stands for the projected probability of the data.

In summary, MobileNet V2 has improved accuracy while using fewer computations and learning parameters. As seen in Figure 2, we added MobileNet V2 to the feature extractor that resembles U-Net [10,11]. The encoder used by the model was previously trained using the Imagenet dataset. As a result of the frames' unlabeled geographical data, training efficacy is increased. Most information about violence is transitory and visible in action rather than still images. The locations where the scenes in the security camera movie take place also vary greatly. The difficulty of connecting the development of these traits into aggressive behaviour over time is decreased by providing a feature extractor that is effective and efficient before training.

## V. Experiments and Results

### A. Dataset

Real-Life Violence Situations Dataset, which has 2000 video clips, is one of several datasets that are accessible; nonetheless, in our study, we use it: Security cameras in diverse real-world scenarios gathered thousand violent and thousand non-violent films. Frequent clips and equal amounts of violence were chosen as the training approach. The models' input comes from the video data that is frame-by-frame data obtained from the videos.

### B. Results

In this part, we look at how well the suggested model performs when identifying and categorising violent and non-violent films. The trials showed that the suggested model performs well and is relatively simple and quick.

The suggested violence detection model's performance was validated using averages of the following:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

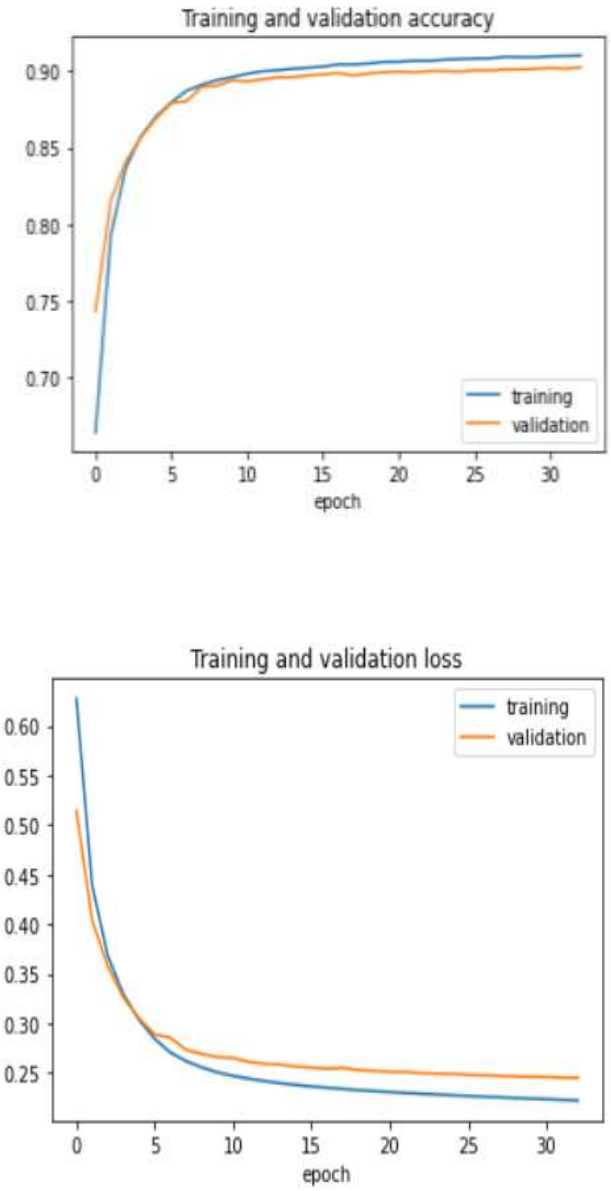
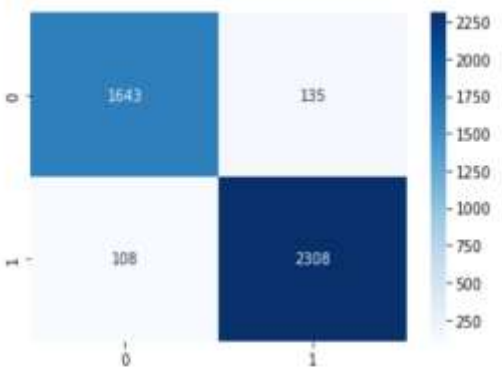



Figure 3. *Accuracy of MobileNetV2 during training and validation in real world violent situations*

The indicated design is operationally lightweight and nevertheless produces good results, as evidenced by trials utilising a complex real-life violent scenarios dataset that exhibited an accuracy of 95.69% and precision of 94%. Figure 4 shows an analysis and evaluation of the performance of our suggested model over the dataset using the evaluation measures.




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| NonViolence | 0.94 | 0.92 | 0.93 | 1778 |
| Violence | 0.94 | 0.96 | 0.95 | 2416 |
| accuracy | | | 0.94 | 4194 |
| macro avg | 0.94 | 0.94 | 0.94 | 4194 |
| weighted avg | 0.94 | 0.94 | 0.94 | 4194 |

Figure 4. *Confusion Matrix and Classification report*

## VI. Conclusion

This study presents a brand-new and useful technique for detecting abusive tendencies in actual surveillance film. The suggested model uses a network similar to the U-NET with the encoder mobilenetv2 to extract spatial features before moving on to an LSTM block for the extraction of temporal features and binary classification**.** The design of the model makes it algorithmically light and speedy. Using a dataset of real-world violent incidents, five folds of cross-validation were carried out. By utilising the usefulness of our proposed system may potentially be improved using technologies of other smart grid installations, which will speed up the reaction time of larger systems operated by people inside of dwellings and other intricate structures. (such as factories, mines, parking lots, and shopping malls), as suggested by Wei et al. [12,13].

The results of the experiments revealed an average precision of 94% and accuracy of 95.69%. The proposed model, despite being small and computationally cheap, obtained high accuracy. Our concept is useful for edge devices or time-sensitive applications. Such technologies can be used in CCTV surveillance of public locations to safeguard citizens [14]. The presence of weapons and other violent items could be analysed to evaluate the level of violence. Even though low-light film might be intellectually stimulating to classify, a variety of techniques for a approach to identify violent incidents in dim or dark conditions can be explored using image recognition.

## Acknowledgments

We would like to express our profound gratitude to **Dr. Supraja P.** an Associate Professor in the SRM Institute of Science and Technology's Department of Networking and Communication, for providing us with the opportunity to work on our project under her direction. She encouraged us to go into the academic fields that piqued our interest while giving us the freedom to do so,

## References

[1] O. Deniz, I. Serrano, G. Bueno and T. -K. Kim, "Fast violence detection in video," 2014 International Conference on Computer Vision Theory and Applications (VISAPP), 2014, pp. 478-485.

[2] C. Penet, C. -H. Demarty, G. Gravier and P. Gros, "Multimodal information fusion and temporal integration for violence detection in movies," 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 2393-2396, doi: 10.1109/ICASSP.2012.6288397.

[3] Gracia, I. S., Suarez, O. D., Garcia, G. B., & Kim, K. (2015). Fast Fight Detection. *PLOS ONE*, *10*(4), e0120448. https://doi.org/10.1371/journal.pone.0120448

[4] Eugene Yujun Fu, Hong Va Leong, Grace Ngai, and Stephen Chan. 2016. Automatic Fight Detection in Surveillance Videos. In Proceedings of the 14th International Conference on Advances in Mobile Computing and Multi Media (MoMM '16). Association for Computing Machinery, New York, NY, USA, 225–234. https://doi.org/10.1145/3007120.3007129

[5] 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, doi: 10.3390/s19112472.

[6] Sumon, Shakil Ahmed, Raihan Goni, Niyaz Bin Hashem, Tanzil Shahria, and Rashedur M. Rahman. "Violence detection by pretrained modules with different deep learning approaches." *Vietnam Journal of Computer Science* 7, no. 01 (2020): 19-40.

[7] Soliman, Mohamed Mostafa, Mohamed Hussein Kamal, Mina Abd El-Massih Nashed, Youssef Mohamed Mostafa, Bassel Safwat Chawky, and Dina Khattab. "Violence recognition from videos using deep learning techniques." In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pp. 80-85. IEEE, 2019.

[8] Butt, Umair Muneer, Sukumar Letchmunan, Fadratul Hafinaz Hassan, Sultan Zia, and Anees Baqir. "Detecting video surveillance using VGG19 convolutional neural networks." *International Journal of Advanced Computer Science and Applications* 11, no. 2 (2020).

[9] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017, pp. 1-6, doi: 10.1109/AVSS.2017.8078468.

[10] Vijeikis, R., Raudonis, V., & Dervinis, G. (2022). Efficient Violence Detection in Surveillance. *Sensors (Basel, Switzerland)*, *22*(6). https://doi.org/10.3390/s22062216

[11] Sandler, Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. "Mobilenetv2: Inverted residuals and linear bottlenecks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510-4520. 2018.

[12] Wei, W., Xia, X., Wozniak, M., Fan, X., Damaševičius, R. and Li, Y., 2019. Multi-sink distributed power control algorithm for cyber-physical-systems in coal mine tunnels. *Computer Networks*, *161*, pp.210-219.

[13] Wei, W., Song, H., Li, W., Shen, P. and Vasilakos, A., 2017. Gradient-driven parking navigation using a continuous information potential based on wireless sensor network. *Information Sciences*, *408*, pp.100-114.

[14] Patel, Mann. "Real-Time Violence Detection Using CNN-LSTM." *arXiv preprint arXiv:2107.07*