International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 12 | Dec 2021 www.irjet.net p-ISSN: 2395-0072

Violence Detection in Surveillance Videos

Harsh Mandviya[1], Mubasheeruddin Siddiqui[2], Jayna Medtia[3], Prashant Kumar[4], Harshal Patil[5]

[1-5]*Dept. of Computer Engineering, Vishwakarma Institute of Technology, Pune, India*

Abstract— *Violence-related instances have recently surged dramatically in areas such remote roads, footpaths, shopping malls, elevators, sports stadiums, and liquor stores, which are tragically discovered only after it is too late. Our project's goal is to develop a complete system capable of real-time video analysis, which will aid in detecting the presence of any violent acts. We achieved an efficient solution that can be used for real-time analysis of video footage so that the concerned authority can monitor the situation. We have put forward a deep neural network for significant detection accuracy. A Convolutional Neural Network (CNN) is utilized to separate frame level highlights from a video which are then accumulated utilizing a variation of Long Short-Term Memory (LSTM) that utilizes convolutional entryways. CNN and LSTM are together utilized for the investigation of nearby movement in a video.*

Keywords— Deep learning, LSTM, CNN, Smart cities, Transfer learning, Violence detection

1. INTRODUCTION

Violence has increased dramatically in recent years, posing major hazards to people, systems, and structures. When violence occurs in public, the issue becomes even worse because most people are not held accountable and cannot be held accountable without proof. Because of their unusual nature, the majority of the horrible crimes occur in public. When we talk about violent activities, we're usually talking about a strange physical encounter between two or more people. Monitoring the surveillance has created a lot of difficulty for security personnel because they now have to go through the footage painstakingly to find the perpetrator and track his movements from one camera to the next, or view it in real-time to detect violent activities and behavior before or as they happen.

Because violence in a city might happen at any time, depending on humans to monitor and detect violent events is ineffective. Such behavior's frequently result in extremely unpleasant scenarios, making it critical for automatic identification of such events to occur using real-time video footage so that the appropriate, critical choice can be made by the appropriate authorities. As a result, the idea of putting in place methods and technology to detect such instances via video retrieval and real-time monitoring has been proposed. The primary goal is to eliminate the aforementioned real-world constraints and drastically and efficiently reduce crime rates.

Today, the measure of public violence has expanded significantly as much in high schools as in the street. This has assisted the specialists with recognizing these occasions and taking the fundamental measures. But almost all systems today require the human-inspection of these videos to identify such events, which is virtually inefficient. It is hence important to have such a down to earth framework that can naturally screen and distinguish the reconnaissance recordings. The advancement of different deep learning methods, on account of the accessibility of huge informational indexes and computational assets, has brought about a noteworthy change locally in computer vision. Different methods have been created to resolve issues, for example, object location, acknowledgment, following, activity acknowledgment, legend age, and so forth Notwithstanding, regardless of ongoing advancements in profound learning, not many methods dependent on profound learning have been proposed to resolve the issue of recognizing violence from recordings.

1.1 Literature Survey

Traditional methods for detecting violence centered on creating hand-crafted features that directly represented motion trajectory, limb position, local appearance, inter-frame variations, and so on. Nievas [1] proposed utilizing the Bag-of-Words framework by combining two such features: Motion Scale Invariant Feature Transform (MOSIFT) and Spatiotemporal Interest Points (STIP). They also introduced two well-known datasets for detecting violence.

Many efforts on violence detection have focused on developing end-to-end trainable neural networks that perform effectively with little to no pre-processing due to the popularity of deep learning approaches. Ding[2] utilized a 3D Convolutional Network to recognize violence directly from raw inputs. Following the success of two-stream networks on general activity recognition tasks, Dong[3] added acceleration streams with spatial and temporal ones for detecting person to person

violence. Dai et al proposed an LSTM that works over two streams to enhance the capture of temporal dynamics and a final SVM classifier for classification. The first CNN-LSTM models employed a fully linked normal LSTM layer that takes in 1-dimensional feature vectors as inputs and does not maintain the spatial aspects of the CNN-LSTM features. Using fully connected 2D LSTM layers, on the other hand, is not practicable due to the large number of parameters required. Sudhakaran[4] proposed utilizing ConvLSTM as the recurrent unit to aggregate frame-level features, which uses convolutions to conduct gate operations inside LSTM cells, greatly lowering parameter count. ConvLSTM can operate with 2D features without flattening them to 1D vector while preserving spatial information. They also discovered that practicing the difference between nearby frames improved performance. Hanson[5] later improved this work by utilizing BiConvLSTM, which uses long-range information in both temporal directions to provide bidirectional temporal encodings in feature vectors. Li[6] suggested a more efficient 3D CNN based on the Dense Net architecture, which requires less parameters. They used two deep neural networks to extract spatio-temporal features representing distinct concepts, which they then aggregated with a shallow network. Some studies combined visual and aural clues to identify violence in a multimodal manner. However, because audio signals are rarely available in surveillance footage, most studies focused on visual data. MobileNet, a lightweight 2D CNN that uses depth wise separable convolutions and intelligent design choices to build a quick and efficient model aimed toward mobile and embedded vision applications, was used in our research. We also used Separable Convolutional LSTM (SepConvLSTM), which is made up of depth wise separable convolutions that replace the convolution operations in the LSTM gates. Separable Convolutional LSTM was recently employed to speed up video segmentation tasks in a study. However, we were unable to locate any work applying SepConvLSTM in the field of activity recognition.

A. Dataset

Dataset is a collection of 1000 films culled from various ice hockey footage. There are 50 frames in each video. The backdrops of all the videos are the same. To avoid class imbalance, the mentioned dataset contains an equal amount of movies depicting violent and nonviolent action. For video violence recognition, we used the Hockey Fight dataset.

B. Network Architecture

Figure 2 depicts the planned network architecture. It has been demonstrated that, in addition to adding the LSTM after the CNN (which is designed to extract global temporal characteristics). The pre-trained CNN processes the two input frames. The last channel concatenates the 20 frames outputs from the bottom layer of the pre-trained model. To compare the high-level features of the 20 frames, the 20 frames outputs from the top layer of the pre-trained network are concatenated and fed into the other network model LSTM. To learn the global temporal features, the outputs from the CNNs are concatenated and transferred to a fully-connected layer and the LSTM cell. Finally, the LSTM cell's outputs are categorized by a fully-connected layer that contains two neurons representing the two categories (fight and non-fight).

2. Methodology

Our suggested method aims to create an end-to-end trainable neural network that can accurately recognize violent events while remaining computationally efficient. To that goal, we created a new and effective two-stream network for detecting violence. We have devised a simple approach that promotes the capturing of discriminative features by highlighting body motions in the frames and suppressing non-moving background information. This section begins with a description of the proposed network's design, followed by the Accuracy Evaluation.

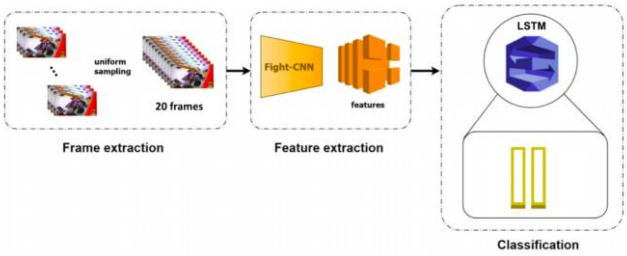


Fig.2: The proposed network architecture.

C. Convolutional Neural Networks

In comparison to regular neural networks, convolutional neural networks feature a distinct architecture. An input is transformed in a regular neural network by passing it through a series of hidden layers. Each layer is made up of a group of neurons that are all connected to the neurons in the layers preceding them. The output layer, which offers the predictions, is the final completely linked layer.

Convolutional neural networks, on the other hand, are a little different. Layers are divided into three categories: width, height, and depth. Furthermore, the neurons in one layer do not connect with all of the neurons in the next layer, but only a small portion of it. The final output will be reduced to a single vector of probability scores, which will be coordinated along the depth dimension.

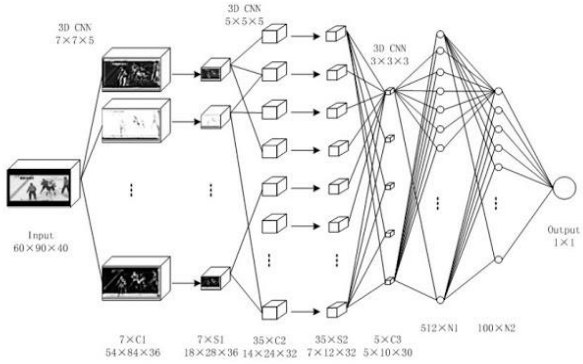
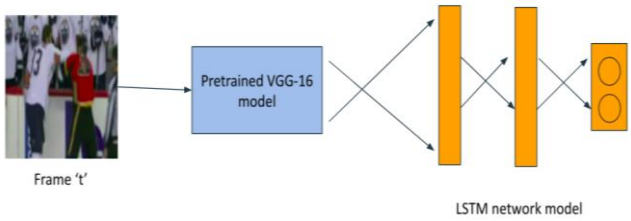


Fig.1: Violence detection by 3D convolutional networks

D. LSTM Architecture

The dimensions of the transfer values must be considered when defining the LSTM architecture. The VGG16 network generates a vector of 4096 transfer values from each frame. We process 20 frames from each video, resulting in a total of 20 x 4096 values per video. The classification must be done while keeping the video's 20 frames in mind. The video will be categorized as violent if any of them detects violence.

The temporal dimension is the first input dimension of LSTM neurons, which in our case is 20. The size of the characteristics vector is the second factor to consider (transfer values).



E. Recurrent Neural Network

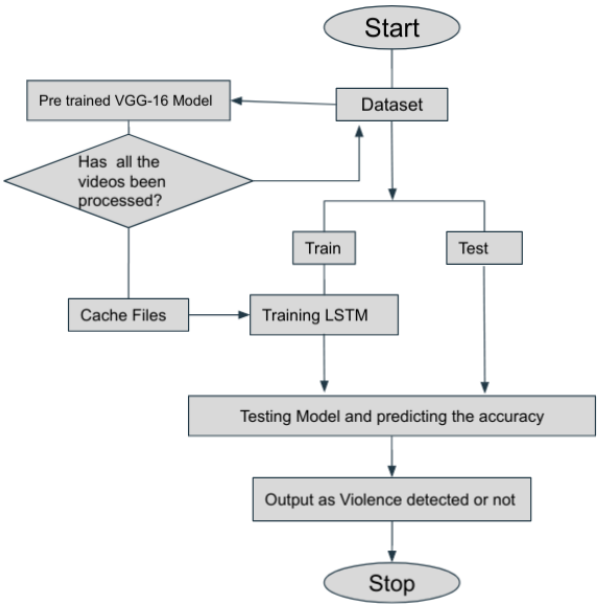
It is widely acknowledged that because like the long-term components, the gradient of the recurrent network can rapidly increase. The standard approach to dealing with an exploding gradient is to truncate it so that it remains within a tolerable range. While some research has found a solution to this problem by starting with a small number of unrolls and gradually increasing the size of unrolls as the loss plateaus. They discovered that clipping the gradients isn't essential in the second approach. They also claim that if the network isn't started from the little unrolls, it may never converge.

We discovered that even if the initial unroll is set to the length of the videos, the network may quickly converge. The lack of gradient clipping, on the other hand, causes the loss curve to oscillate during training, even if the training starts with a tiny unroll. As a result, the network's gradients are shortened in the range of -5.0 to 5.0. Clipping gradients into a smaller range (for example, from -1.0 to 1.0) has also been investigated. However, my experiment reveals that the network will scarcely converge to the lower minima as a result of this.

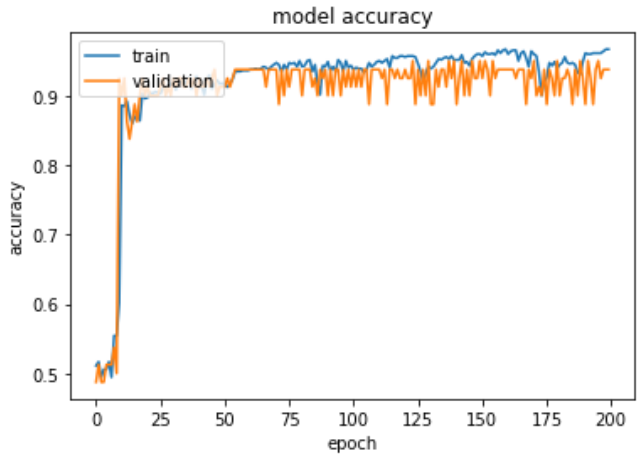
F. LSTM

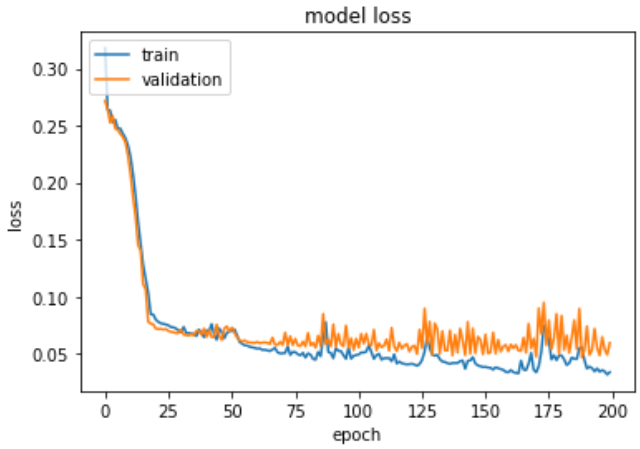
Long Short-term Memory is a technique that is utilized in sequence learning undertakings. The memory use capacity of LSTM contrasts from the normal recurrent neural organizations (RNN). Its memory doors in the modules make it conceivable to keep the fundamental data and disregard immaterial data. The gates decide to pass or toss a few pieces of the information as indicated by its relevance by thinking about the past information. All in all, the doors in LSTM figure out how much the new data relies upon the past data. Hence, the connection between the components of a grouping can be learned. In this case, the data consists of a sequence of images and the network can connect the information in frames which are taken at different times from the videos. During this interaction, the system recollects the past frames while analyzing the current frame. The system learns the transient changes happening during the video handling and those progressions give huge data to perceive the activities. During the LSTM experiments, an LSTM model with one LSTM layer, three dense (1024, 50, 2) and three activation layers (ReLU, sigmoid, Softmax) are used. At the end of the architecture, the Softmax layer is used with two classes instead of binary classification by sigmoid. Therefore, the prediction confidences in the output can be observed. So that mean squared error is used as the loss function which gives better results than the cross entropy loss function.

3. Flow Diagram



4. Results





Hockey fight dataset includes 1000 clips which are divided into 500 fight clips and 500 non-fighting footage from hockey games. Our Datasets is split into 80% for training and 20% for testing. We trained our LSTM model with an epoch of 200 and having batch size as 500. We have got the accuracy and loss as shown in the above two graphs. The accuracy of our model came out to be approximately 93% with a loss of 4.96% only which tells us our model is quite efficient.

5. Conclusion

The main objective of this study is detecting fight scenes from surveillance cameras in a fast and accurate way. Nowadays, the rate of violence in our environment is escalating, posing a hazard to people, structures, and systems. There has always been a need for a better system to assist the police in monitoring the violence, which is often difficult to handle because it is a group activity and the elimination process to discover the perpetrator is time-consuming. The results of our experiment show that CNN+LSTM architecture may be used effectively to train a model over Hockey's dataset. Our work can assist law enforcement in keeping their area under control.

By consolidating CNN with LSTM, the accuracy increases to a specific edge when contrasted with transfer learning models alone.

6. Limitations

Aside from the high accuracy in identifying violence, the real-time processing speed, the capacity to detect violence events frame by frame, and the capacity to handle variable length detection are all features that the system offers proves to be a limitation for such and which should be overcome as this can affect the usability of such models in real life scenarios.

7. Future Outcomes

Our proposed system can be altered to increase its performance in a variety of ways. By specifying the severity of the detected violence, the system can be improved.. To boost performance, the suggested Deep Learning architecture can be tweaked by changing the hyper parameters. In addition, the system can be expanded to identify other forms of crimes, such as fires and burglaries. Also we aim to build an integrated system (application) for violence detection which can detect violence in real time and alert the concerned authority with a simple but effective user interface.

This surveillance camera dataset can be extended by adding new samples from security camera footage on streets or underground stations.

8. Acknowledgment

The project was supported by the **Vishwakarma Institute of Technology, Pune**. We are thankful to **Dr. Prof. Kulkarni Milind** for guiding us throughout the project with his precise suggestions.

References

[1] E. B. Nievas, O. D. Suarez, G. B. Garc´ıa, and R. Sukthankar, "Violence detection in video using computer vision techniques," in the International conference on Computer analysis of images and patterns. Springer, 2011, pp. 332–339.

[2] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, "Violence detection in video by using 3d convolutional neural networks," in International Symposium on Visual Computing. Springer, 2014, pp. 551–558.

[3] Z. Dong, J. Qin, and Y. Wang, "Multi-stream deep networks for person to person violence detection in videos," in Chinese Conference on Pattern Recognition. Springer, 2016, pp. 517–531.

.

[4] Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," published in 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2017, pp. 1–6.

[5] A. Hanson, K. Pnvr, S. Krishnagopal, and L. Davis, "Bidirectional Convolutional LSTM for the detection of violence in videos," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 0–0. Dai.

[6] J. Li, X. Jiang, T. Sun, and K. Xu, "Efficient violence detection using 3d convolutional neural networks," in 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2019, pp. 1–8.