Violent Interaction Detection in Video Source Using Deep Learning

Priyank Dave[a*] Nisha Khurana[b]

[a*] *PG Scholar, Computer Engineering Department, Gandhinagar Institute of Technology*

[b] *Professor,* Information Technology *Department,* Gandhinagar Institute of Technology

[a] *210120702001@git.org.in,* [b] *nisha.khurana@git.org.in*

Abstract

Violent activity detection is very important in some video analysis eventualities like railway and metro stations, prisons or alternative public places. With the rise of video data set to watch the demands of such a system, that acknowledges the violence and suspicious events has been increased automatically. Violence action detection has become a vigorous research area of computer vision and video processing to attract new researchers. During this work, we tend to discuss developing a method for the automated analysis of violence activity through video supply. Machine learning techniques and tools are used to detect sequences of violence interaction in videos.

Keywords: Violence Detection, Action and Activity Recognition, Anomaly Detection, Machine Learning for VD.

1. Introduction

The majority of the systems require manual human inspection, for identifying violent interaction scenarios in video. Which is both ineffective and practically impossible. Having such a practical system that can automatically monitor videos and identify the violent behavior of humans will be of immense help and assistance to the law-and-order establishment. In recent years, machine learning and computer vision techniques have made it possible to recognize human action from video. Automated video analysis techniques are utilized for a wide range of purposes, including: indexing and finding videos, Summary of a video [1], Recognizing action and activity [2], Classification of videos [3]

Due to its numerous applications in the fields of medicine, security, sports, and entertainment, among many others, action and activity detection has attracted a lot of study interest in the field of video categorization.

Action recognition is a technology that can identify human actions. Based on the complexity of the acts and the number of bodily parts involved, human activities are divided into four types.

The four types are gestures, actions, interactions, and group activities. Whereas violence detection is one of the broad classes that aim on identifying violent and injurious event patterns in an input video. Video classification can be used to identify specific actions from video. Sequences of images make up videos. Images taken from video frames can be used to extract features, and with the help of these features, predictions can be made.

As we can see in Fig. 1.1, Video classification can be used in applications like, activity recognition, violence detection, anomaly recognition.

*Priyank Dhirenkumar Dave

E-mail address: 210120702001@git.org.in

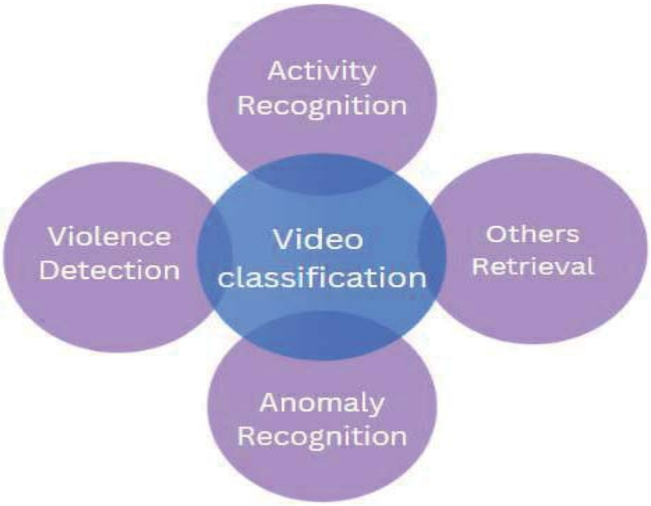


Fig 1.1 Video classification applications.

Fig. 1.2 shows clips extracted from video and makes predictions of activity in frame. As we can see in the image both are examples of running activity.

2. Background:

The community that studies action recognition has generally concentrated on relatively simple activities like clapping, walking, jogging, etc. Comparatively less research has been done on the recognition of specific events like fight behavior and violence detection. Such a feature might come in very handy for applications like video monitoring in jails, public spaces, psychiatric facilities, or even embedded in camera phones and social media.

Violence Detection is the selection or extraction of anomalous patterns from a typical surveillance video that may last for a very brief or an extended period of time. By starting countermeasures such as alerting the closest relevant departments for reaction activities, its early identification aids in either prevention or decreasing the residual harm in terms of human lives and their properties. Violence is any abnormally cruel human action, including hitting, harming, destroying, and wounding others.

Many real-world scenarios can be used with video data. When compared to the typical pattern of ongoing video surveillance, it is obvious that violent incidents are incredibly uncommon. As a result, using human resources to watch video streams might not be viable because it would take a lot of training to spot odd patterns. A video data system's primary objective is to identify and report any abnormal behavior that deviates from the typical pattern of activity.

Since decades, Violence Detection has been used to analyze video data utilizing soft computing methods that were initially based on conventional image processing methods. Early Violence Detection literature took into account a variety of factors for decision-making, including human motion acceleration, appearance, and motion flow, among many others Video data preparation, feature extraction, and segmentation into violent or nonviolent segments are all common processes in baseline research. Segmenting videos and cleaning up data are both part of the preprocessing stage. The term "features extraction" refers to the processing of individual-level (spatial) frames utilizing specific feature extractors including motion, speed, and optical flow.

3. Literature Survey

Violence detection methods can be categorized in two methods:

1. Violence detection using machine learning methods.
2. Violence detection using deep learning methods.

Machine learning methods have gotten more attention from recent decades, but in recent years deep learning approaches are popping up in the attention of researchers.

3.1 - Violence detection using machine learning methods

Penet et al. [6] investigated the different Bayesian network learning algorithms using temporal and multimodal information for a violent shot detection system.overall results gave a false alarm rate of 50% and 3% missed-detection. Nievas et al. [7] proposed to use the bag-of-words (BoW) approach and two motion descriptors, space–time interest points (STIP) and motion scale invariant feature transform (MoSIFT) for fight detection. Detect fights with 90% accuracy. A violent flow (VF) variation for violence detection based on the combination of SVM and Horn–Schunck optical flow algorithm was proposed by Arceda et al. [8]. Das et al. [9] classifier achieves 86%

accuracy.

3.2 - Violence Detection Based on Deep Learning Techniques

Ding et al. [10] used 3D convolution with back propagation strategy for violence detection. Xia et al. [11] used Bi-channels CNN with SVM are used for violence detection. Mu et al. [12] CNN are used to detect violent based on acoustic information. Meng et al. [13] Integrating frame of trajectory and Deep CNN to detecting human violent behavior in videos.

3.3 - Data set availability

Hence data is related to violence, so more dataset is not made publicly available but there are various datasets available like

1. RWF-2000 dataset for violence detection by Cheng et al. [4]
2. Movie Fights dataset by Nievas et al. [5]
3. Hockey Fights dataset by Nievas et al. [5]

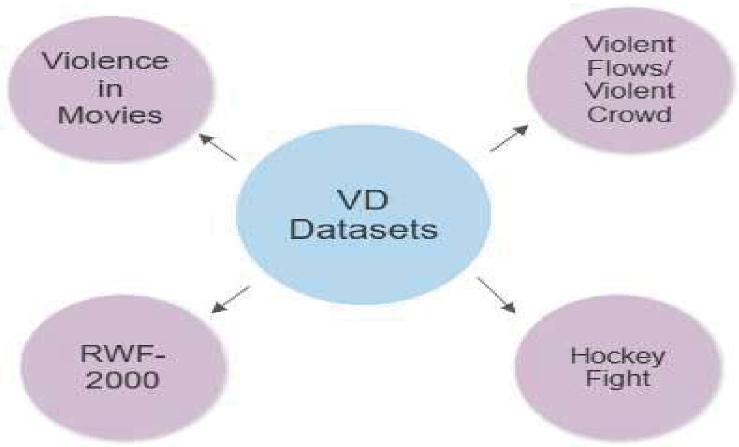


Fig. 3.1. Overview of VD Datasets.

4. Proposed Model

To classify violent or non-violent actions, our model must be able to anticipate sequences in consecutive frames, such as a pattern in the subjects' movement or a degree of their motion, and so on. This is not possible if only the spatial features (features that pertain to a specific frame) of the frames are considered. When detecting sequences in frames, temporal or time-related features must also be examined. The temporal features can be processed in either the forward or reverse sequence. Our model processes temporal characteristics in both directions in addition to spatial features, allowing the model to become more accurate while consuming less computational time.

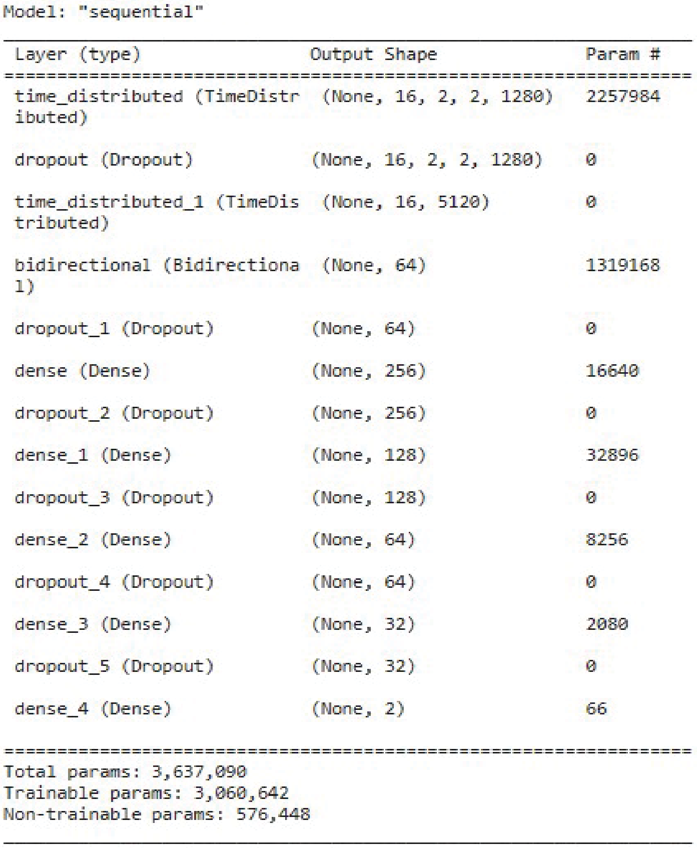
The proposed model aims to keep performance similar to cutting-edge violence detection models while reducing the computational complexity.

Three steps primarily make up the proposed algorithm:

1. Spatial feature extractor
2. Temporal feature extractor
3. Classifier

The videos' frames have been extracted. The captured frames are reshaped to 64 x 64 pixels (denoted as x y). The training data is a Numpy array, with each row indicating a sequence or pattern in video. A sequence could include a degree of movement and movements, such as whether an arm movement is a punch or a handshake, and so on. A series can be extracted with as few as two frames. However, we extracted temporal features (time-related features) using 16 sequential frames.

When the model receives a video frame, the first step is spatial feature extraction using a network. This network employs MobileNet V2 as an encoder used to perform sequential-time-distributed static single frame spatial feature extraction. As a classifier for the model, we selected MobileNet V2, a simple state-of-the-art classifier for spatial feature extraction. MobileNet V2 gets great accuracy while using a much smaller network size. Then it is passed to BiLSTM for temporal feature extraction. Following fig shows the structure of the model.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 time_distributed (TimeDistr  (None, 16, 2, 2, 1280)   2257984
 ibuted)

 dropout (Dropout)           (None, 16, 2, 2, 1280)    0

 time_distributed_1 (TimeDis  (None, 16, 5120)         0
 tributed)

 bidirectional (Bidirectiona  (None, 64)               1319168
 l)

 dropout_1 (Dropout)         (None, 64)                0

 dense (Dense)               (None, 256)               16640

 dropout_2 (Dropout)         (None, 256)               0

 dense_1 (Dense)             (None, 128)               32896

 dropout_3 (Dropout)         (None, 128)               0

 dense_2 (Dense)             (None, 64)                8256

 dropout_4 (Dropout)         (None, 64)                0

 dense_3 (Dense)             (None, 32)                2080

 dropout_5 (Dropout)         (None, 32)                0

 dense_4 (Dense)             (None, 2)                 66

=================================================================
Total params: 3,637,090
Trainable params: 3,060,642
Non-trainable params: 576,448
_________________________________________________________________
```

Fig. 4.1. Structure of the model.

5. Results

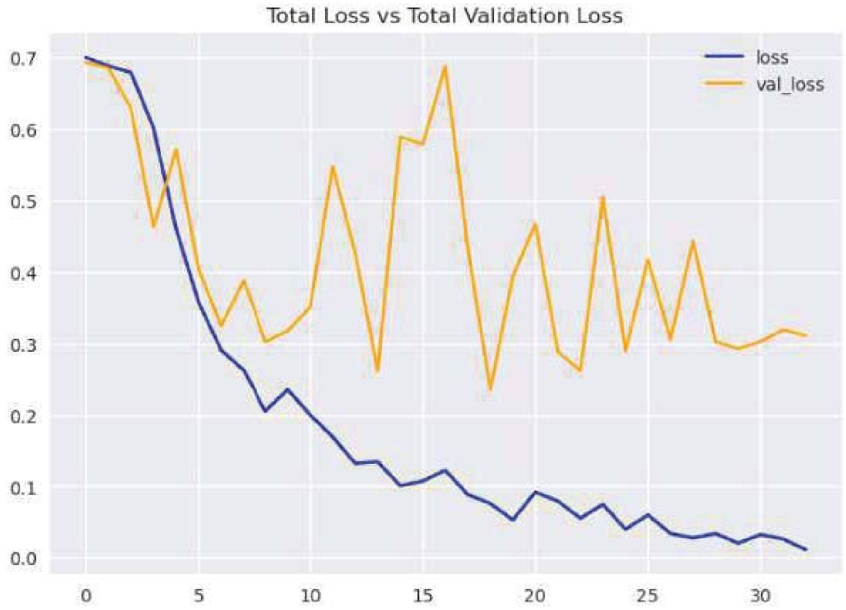


Fig. 5.1. Total loss vs Total Validation Loss.

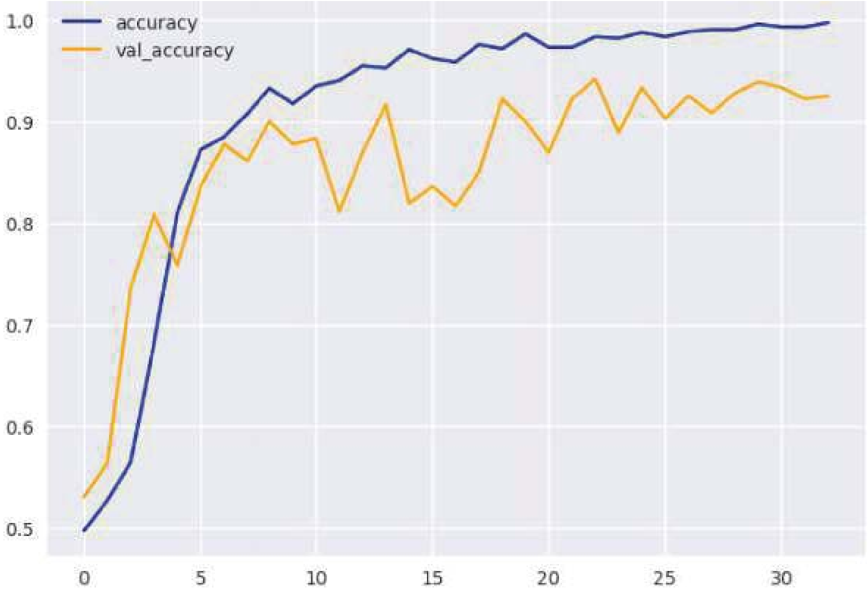


Fig. 5.2. Total Accuracy vs Total Validation Accuracy.

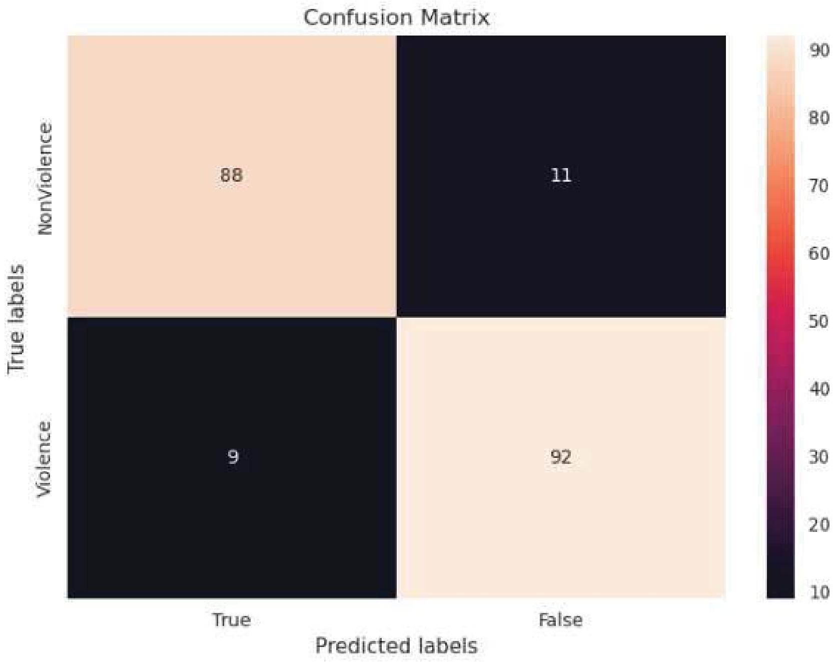
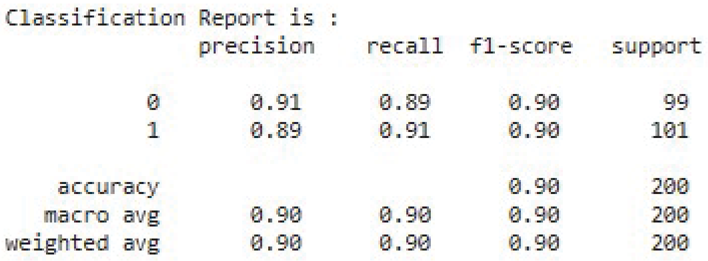


Fig. 5.3. Confusion Matrix.



```
Classification Report is :
              precision    recall  f1-score   support

           0       0.91      0.89      0.90        99
           1       0.89      0.91      0.90       101

    accuracy                           0.90       200
   macro avg       0.90      0.90      0.90       200
weighted avg       0.90      0.90      0.90       200
```

Fig. 5.4. Classification Report.

6. Conclusion

The suggested model extracts spatial features using MobileNet V2 as an encoder, then extracts temporal features using Bi-LSTM followed by a classifier. There are 3,637,090 parameters in the model. The model's architecture makes it computationally quick and light. Experiments using a complicated dataset of real security camera footage based on RWF2000 revealed average accuracy of 0.90.

Although our suggested model performed satisfactorily, it still has to be further verified using more common datasets where it is difficult to identify one to many or many to many violent acts, including the use of weapons. Again, by analyzing the past sequence of events for an individual or a group of people, the detection model can be extended to a prevention model. We will expand this work in the future to address the aforementioned challenges in detecting violent and non-violent activities.

References

1. Wu, J., Zhong, S.h., Liu, Y., 2020. Dynamic graph convolutional network for multi-video summarization. Pattern Recognition 107, 107382.
2. Dang, L.M., Min, K., Wang, H., Piran, M.J., Lee, C.H., Moon, H., 2020. Sensor-based and vision-based human activity recognition: A comprehensive survey. Pattern Recognition 108, 107561.
3. Sasithradevi, A., Roomi, S.M.M., 2020. Video classification and retrieval through spatio-temporal radon features. Pattern Recognition 99, 107099.
4. Cheng, M.; Cai, K.; Li, M. RWF-2000, 2021. An open large scale video database for violence detection. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15, p. 4183–4190.
5. Bermejo Nievas, E.; Deniz Suarez, O.; Bueno García, G.; Sukthankar, R., 2011. Violence detection in video using computer vision techniques. In International Conference on Computer analysis of Images and Patterns; Springer: Berlin/Heidelberg, Germany, p. 332–339
6. Penet, C.; Demarty, C.H.; Gravier, G.; Gros, P., 2012. Multimodal information fusion and temporal integration for violence detection in movies. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, p. 2393–2396.
7. Bermejo Nievas, E.; Deniz Suarez, O.; Bueno García, G.; Sukthankar, R., 2011. Violence detection in video using computer vision techniques. In International Conference on Computer analysis of Images and Patterns; Springer: Berlin/Heidelberg, Germany, p. 332–339.
8. Arceda, V.M.; Fabián, K.F.; Gutíerrez, J.C., 2016/ Real Time Violence Detection in Video; IET: Talca, Chile.
9. Das, S.; Sarker, A.; Mahmud, T., 2019. Violence detection from videos using hog features. In Proceedings of the 2019 4th International Conference on Electrical Information and Communication Technology (EICT), Khulna, Bangladesh, p. 1–5.
10. Ding, C., Fan, S., Zhu, M., Feng, W., Jia, B., 2014. Violence detection in video by using 3d convolutional neural networks. In: International Symposium on Visual Computing, p. 551–558. Springer.
11. Xia, Q., Zhang, P., Wang, J., Tian, M., Fei, C.,2018. Real time violence detection based on deep spatiotemporal features. In: Chinese Conference on Biometric Recognition, p. 157–165. Springer.
12. Mu, G., Cao, H., Jin, Q. 2016. Violent scene detection using convolutional neural networks and deep audio features. In: Chinese Conference on Pattern Recognition, p. 451–463. Springer.
13. Meng, Z., Yuan, J., Li, Z., 2017. Trajectory-pooled deep convolutional networks for violence detection in videos. In: International Conference on Computer Vision Systems, p. 437–447. Springer.
14. https://www.nytimes.com/2019/05/14/technology/facebook-live-violent-content.html
15. Serrano Gracia I, Deniz Suarez O, Bueno Garcia G, Kim TK., 2015. Fast fight detection. PLOS ONE 10(4):e0120448.
16. Zhou P, Ding Q, Luo H, Hou X., 2018. Violence detection in surveillance video using low-level features. PLOS ONE 13(10):e0203668.

17. Ribeiro PC, Audigier R, Pham QC. RIMOC, 2016. a feature to discriminate unstructured motions: application to violence detection for video-surveillance, 144, p. 121–143.
18. Yao C, Su X, Wang X, Kang X, Zhang J, Ren J., 2021. Motion direction inconsistency based fight detection for multiview surveillance videos, p. 1–11.
19. Febin IP, Jayasree K, Joy PT., 2020. Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm. 23(2):p. 611–623.