# Violence Detection In Videos

LTAT.01.005

**Name:** Tarun Khajuria
**ID:** B87775
**Name:** Mohamed Maher Abdelrahman
**ID:** B87782

**Supervisor:** Prof. Eduard Barbu, and Hasan Sait Arslan

**C**ourse Project Report

**Curriculum:** Master of Computer Science - 2018/19

**Semester:** Spring, 2019

**Presentation Date:** 29th May, 2019 - 12:00 to 02:00 PM

Institute of Computer Science
University of Tartu
Estonia

# Violence Detection In Videos

Mohamed Maher,Tarun Khajuria
University of Tartu, Estonia
mohamed.abdelrahman@ut.ee,tarunkhajuria42@gmail.com

## 1 INTRODUCTION

Recently, it has been reported that there is a large increase in number of violence, and terror incidents all over the world [1]. This has lead to wide usage of surveillance cameras to help authorities to take quick actions in case of violence incidents. Therefore, there is an irresistible urge to automate the analysis of these videos and automate the process of violence detection. Due to the development of several deep learning techniques, and availability of large datasets, it has achieved a great performance for different computer vision tasks. In this project, we are going to describe the related work in section 2. Then, 3 describes different attempts, and architectures used for solving this task. 4 shows the results achieved on benchmark datasets. Finally, 5 describes our final notes, and future ideas to tackle this problem.

## 2 RELATED WORK

Previous work has tackled the violence detection problem with two main different approaches. The first is through observing the variations among the video frames [3],[4] where violence videos have large variations due to the fast motion during fights. On the other hand, other approaches try to capture the motion change patterns in the analyzed videos [2],[6]. Both the previous classes of approaches are using either deep learning techniques of CNNs and RNNs [7] or low level features like histogram of oriented gradient (HOG) features that were proved to be good low level features for video analysis that achieved high performance in violence datasets [1] with some other derivatives from it. Then, we classify them into violence and non-violence scenes [9].

## 3 METHODOLOGY

### 3.1 Using CNN-ConvLSTM

Our main approach was based on [7] where we use a convolutional neural network for feature extraction of difference of frames in a video scene. Then, the output features are introduced to a Conv-LSTM network that can capture the temporal changes. In the Conv-LSTM gates, both the spatial and temporal features can be detected which helps in recognition of motion patterns in the video. 1 shows the used architecture where we tried two different pretrained CNNs on ImageNet for the CNN that extract high level features from difference of consecutive frames. In addition, we tried to concatenate the HOG features of each frame with the output from CNN in order to help the Conv-LSTM network in capturing better tempo-spatial features. It is worth mentioning that, the frames of video scene are pre-processed by scaling to a fixed image resolution, and normalization. In addition, during the training process and based on the literature, we augment some other modified frames like horizontal flipped ones, and cropped parts from the frames either from the center or from the corners. These augmented images help in regularization of the network, and avoid over-fitting over scenes used in the training process.

Detection.pdf

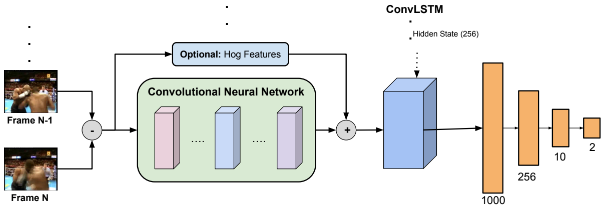

**Figure 1: CNN-ConvLSTM Architecture**

### 3.2 Detection from Caption Generators

Caption generation from videos is an emerging field and some good results have been reported in the field [5]. In [8] further improvements were reported using bidirectional proposal generation network, and then using the best proposal frame to start with the caption generation. The LSTM context from the proposal generation network is fed into the caption generation decoder to get the captions. We tried to use their implementation form Github [2] to generate captions but have not been able to replicate it to a performance that we can build upon. The plan was to generate the captions on the videos with violence and see if the captions can be used to classify a particular segment of the video. As the region proposal network itself can detect events within a particular video sequence, we adapted this network to provide proposal for violence related events in the sequence. The network used visual features like (HOG) from each frame and a sequence of 150 frames was used as input to the Bidirectional LSTM. The LSTM output from the forward part of the LSTM encodes the previous context and the reverse LSTM encodes the future context. A sigmoid activation on the output of each of the time frame was calcualted as the proposal score for the network. Finally we combine the forward and reverse proposal score by multiplying to get the final score. We trained the network by labeling a score of 1 for each frame containing violence and 0 for other frames.

### 3.3 Dataset

For the previous methodology that we have followed, we evaluated them based on three different benchmark datasets.

[1] https://www.vox.com/2018/12/10/18134232/gun-violence-schools-mass-shootings

[2] https://github.com/JaywongWang/DenseVideoCaptioning

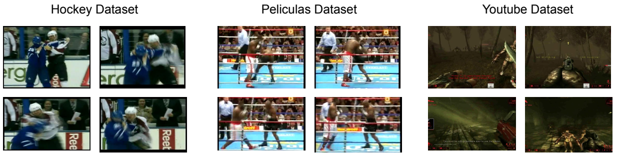

**Figure 2: Samples From Each Dataset**

- **Hockey Fights** [3]: it includes 1000 short video scenes divided equally into violence hockey fights and non-violence scenes.
- **Movies-Peliculus** [4]: it has 200 different scenes from various sports and movies divided also equally into violence and non-violence scenes.
- **Youtube Portion out of Violence Scenes Dataset** [5]: this portion is the only part where the videos are available to download. However, for other portions of the dataset only some low level features are open source. This dataset is different from the other ones where it contains scenes from first-person-shooter games where the camera itself isn't fixed in its place which adds more difficulty to the classification task. In addition, this dataset was smaller than other datasets with only 90 videos divided into violence and non-violence ones.

For all the experiments, we have used 80-20% training and testing fixed splits from each dataset and we have extracted a sample of frames from each video scene. 2 shows an example of type of videos in each dataset . All our implementations for the previous techniques can be found in Github[6].

## 4 RESULTS AND DISCUSSION

### 4.1 Using CNN-ConvLSTM

Using the CNN-ConvLSTM approach, we have tried to use Alex-Net for the CNN part in one trial, and ResNet-18 for another trial. In addition, we tried to extract HOG features from each two consecutive frames, subtract them from each other and augment with output from AlexNet / ResNet. The training process was done for 50 epochs in each trial using a batch size of 16 images, learning rate of 1e-4, memory size of 256 of the Conv-LSTM layer, and RMSProp optimizer with 0.5 decay rate. The testing accuracy is reported for each dataset in 1.

### 4.2 Detection from Caption Generators

The re-purposed proposal generation network is trained with the 150 frame segments from Technicolor violence dataset. As large part of the dataset was provided in form of visual and audio features. For this task the visual features for each frame were used in form of a sequence of 150 frames to get the final event score for each frame using the proposal generation network. We train the network by creating a binary vector ground truth label for the frame sequence and minimizing the binary cross entropy loss over the sequence.

**Table 1: Results of different CNN-ConvLSTM architectures used**

| | Hockey Fights | Peliculas Movies | Youtube |
|---|---|---|---|
| **AlexNet ConvLSTM** | 93.5% | 97.56% | 80.95% |
| **ResNet18 ConvLSTM** | 94.0% | 100% | 85.71% |
| **ResNet18 ConvLSTM +HOG** | 96.0% | 100% | 80.95% |

In this way each frame 's output score represents how likely the frame is having a violent scene. The network was trained over Youtube video features and Youtube+ Movie scene video features from the Technicolor Violent scene description data-set. We extracted 576 and 1273 sequences from these respective video sets. As this task of predicting the exact scene form a sequence containing violence is complex the accuracy of prediction over the test sets was not very high. We obtained 60.1 and 63.6 percent accuracy over the two sequence sets which is very close to baseline accuracy for a binary classification. As the networks do improve the test set accuracy when training over more data, we feel that larger amout of data is required to train the network for this specific task.

## 5 CONCLUSION AND FUTURE WORK

Automatic violence detection has become a crucial thing that needs to be achieved nowadays. We have implemented tried several approaches to classify video scenes into violent or non-violent including CNN-LSTM architecture that tries to capture different motion patterns and differences among video frames. This approach performed very well on three different benchmark datasets. Adding some low features like HOG features to the LSTM cells improved the performance for one dataset. As a future work, we recommend to study the effect of using different kinds of these features like MoSIFT, and VIF by adding them to the LSTM layer. In addition, there should be a study on the effect of different sampling rates of frames out of videos on the performance of classification.

On the other hand, we tried to generate captions for the scenes and use the words generated in classification of videos but this approach failed to perform well due to the lack of large training data. In addition, to the best of our knowledge, we couldn't find either an ontology for violence words or a dataset for videos with captions describing the violence actions. Thus, we don't expect that this approach can act as a standalone method for violence detection. However, it can be integrated with other approaches to provide some new features that can help in the classification process like generation of some words like running, crowd, fire, etc. As a future work in this approach, we recommend to train this network with more data and use the generated captions with word embedding layer to feed the vectors of these generated words to the LSTM layer of the first approach which can act as new features that may enhance the classification accuracy.

One of the challenges in this task, is the lack of large and general open source datasets. We need to test the

[3] https://bit.ly/2Wb0fhD
[4] https://bit.ly/2wo11Zm
[5] https://www.technicolor.com/dream/research-innovation/violent-scenes-dataset
[6] https://github.com/mmaher22/ViolenceDetector

approaches on more general datasets with more real life situations as the benchmark ones used are always biased to certain type of scenes like Hockey dataset. In addition, other good datasets like Technicolor VSD doesn't provide videos themselves but only some acoustic and visual features extracted from the frames.

## REFERENCES

[1] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. 2011. Violence Detection in Video Using Computer Vision Techniques. In *Computer Analysis of Images and Patterns*, Pedro Real, Daniel Diaz-Pernil, Helena Molina-Abril, Ainhoa Berciano, and Walter Kropatsch (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 332–339.

[2] Piotr Bilinski and Francois Bremond. 2016. Human violence recognition and detection in surveillance videos. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 30–36.

[3] Liang-Hua Chen, Hsi-Wen Hsu, Li-Yun Wang, and Chih-Wen Su. 2011. Violence detection in movies. In *2011 Eighth International Conference Computer Graphics, Imaging and Visualization*. IEEE, 119–124.

[4] Oscar Deniz, Ismael Serrano, Gloria Bueno, and Tae-Kyun Kim. 2014. Fast violence detection in video. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, Vol. 2. IEEE, 478–485.

[5] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-Captioning Events in Videos. In *International Conference on Computer Vision (ICCV)*.

[6] Paolo Rota, Nicola Conci, Nicu Sebe, and James M Rehg. 2015. Real-life violent social interaction detection. In *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3456–3460.

[7] Swathikiran Sudhakaran and Oswald Lanz. 2017. Learning to detect violent videos using convolutional long short-term memory. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–6.

[8] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. 2018. Bidirectional Attentive Fusion with Context Gating for Dense Video Captioning. In *CVPR*.

[9] Peipei Zhou, Qinghai Ding, Haibo Luo, and Xinglin Hou. 2018. Violence detection in surveillance video using low-level features. *PLoS one* 13, 10 (2018), e0203668.