# Violence Detection in Videos

Language and Image Processing Project
Mohamed Maher, and Tarun Khajuria

# Agenda

- Introduction and Motivation
- Related Work
- Dataset
- Methodology
- Results and Discussion
- Conclusion and Further Improvement

# Introduction and Related Work

- There is a large **increase in number of violence**, and terror incidents all over the world. [1]

+

- Wide usage of **surveillance cameras**.

---

- Automate the analysis of these videos and the detection of violence.

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

# Related Work

## Method 1

Observing the **variations among the video frames** where violence videos have large variations due to the fast motion during fights. [1][2]

## Method 2

Capture the **motion change patterns** in the analyzed videos along the frames of the video scene. [3][4]

## Approach 1

Extract ***Low Level Features*** and their derivatives **(Eg: HOG, SIFT, FHOG, MoSIFT, ViF)**. Then, use a machine learning ***classifier*** [5][6]

## Approach 2

Using **Convolutional** Neural Networks for Extracting **Spatial** Features from Frames + **Recurrent** Networks for capturing **temporal** features. [7]

# Dataset - (Hockey Fights)

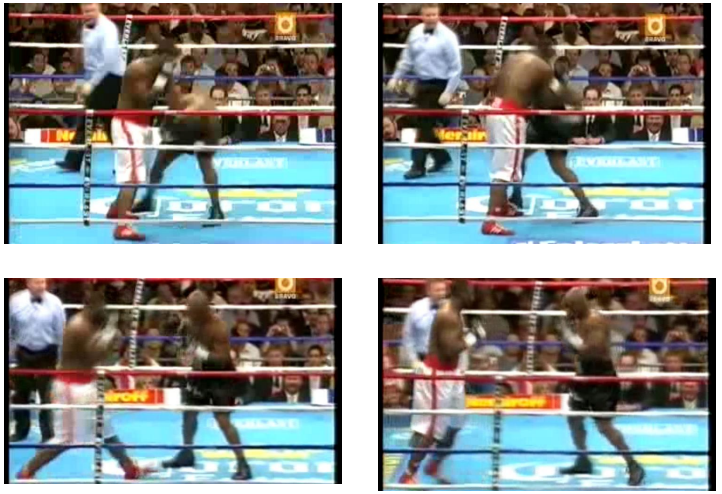
**1000 short video scenes** divided equally into violence hockey fights and non-violence scenes.

Hockey Dataset

# Dataset - (Movies Peliculas)

It has **200 different scenes** from various sports and movies divided also equally into violence and non-violence scenes.

Peliculas Dataset



# Dataset - (Technicolor Violent Scenes Dataset)

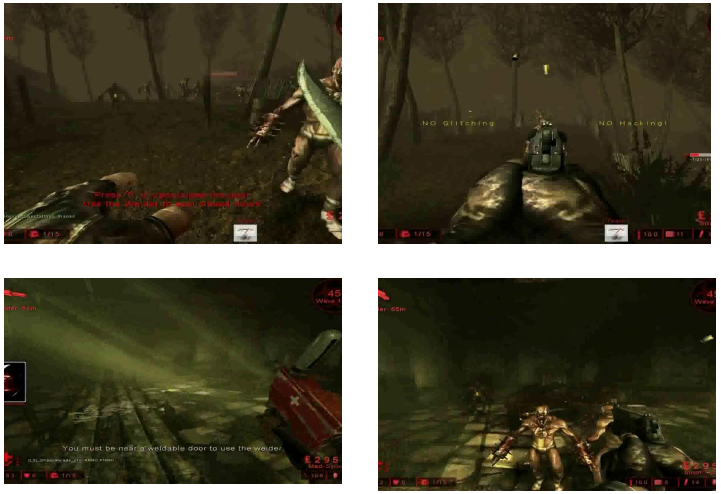
- Violent scene annotations in 32 movies
- Scene annotation for 86 Youtube short videos
- Only videos for youtube dataset provided due to copyright issues
- Audio and Visual features provided for all dataset

Image Source :
https://www.technicolor.com/dream/research-innovation/violent-scenes-dataset-description

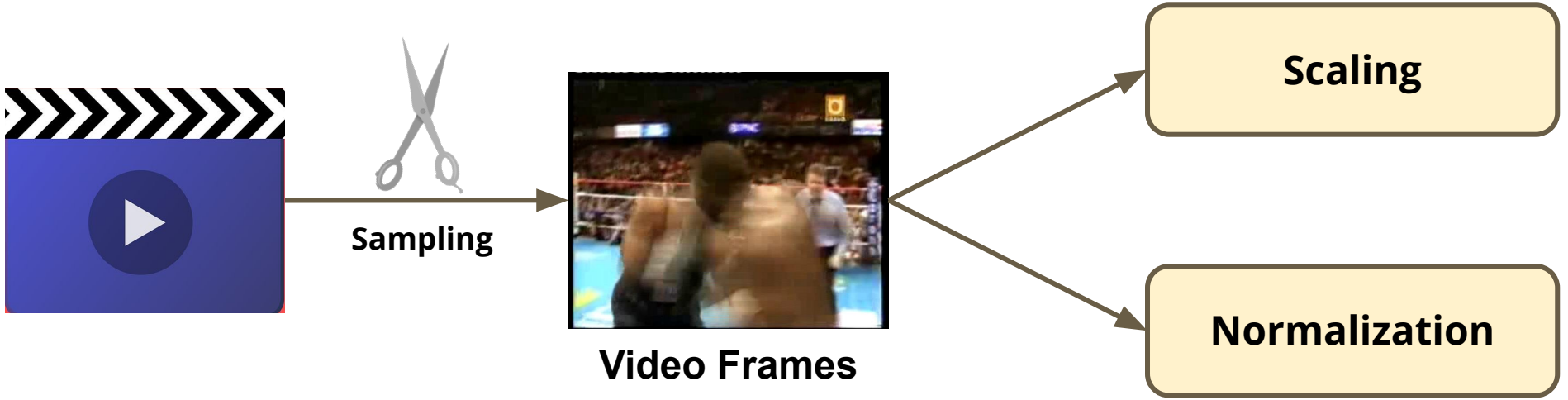
# Dataset - (Youtube Section of Violence Scenes Dataset Peliculas)

- This portion is the only part where the **videos are available**.
- This dataset is different from the other ones where the **camera itself is moving** in some scenes which adds more difficulty to the classification task.
- This dataset was smaller than other datasets with only **90 videos** divided into violence and non-violence ones.

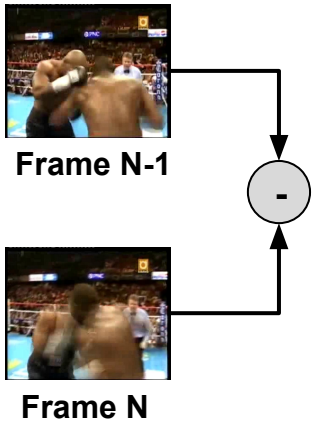
Youtube Dataset



# Methodology - *CNN + Conv-LSTM*

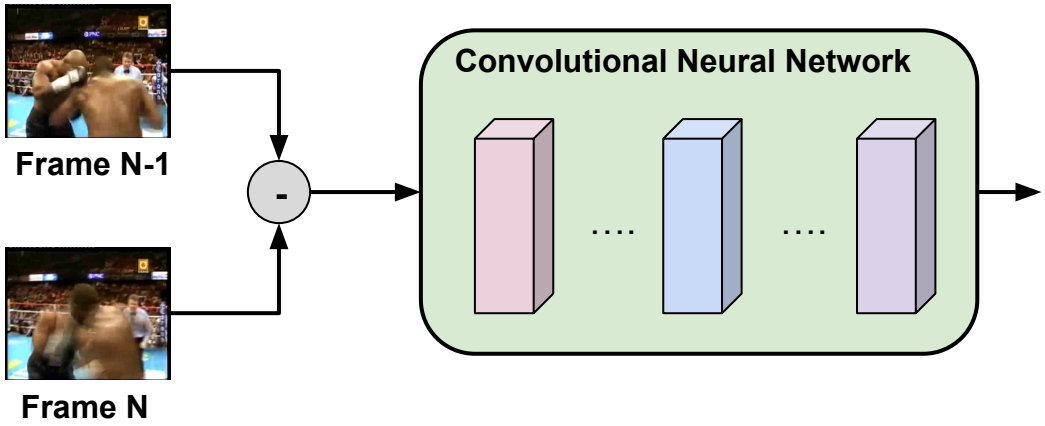


# Methodology - *CNN + Conv-LSTM*

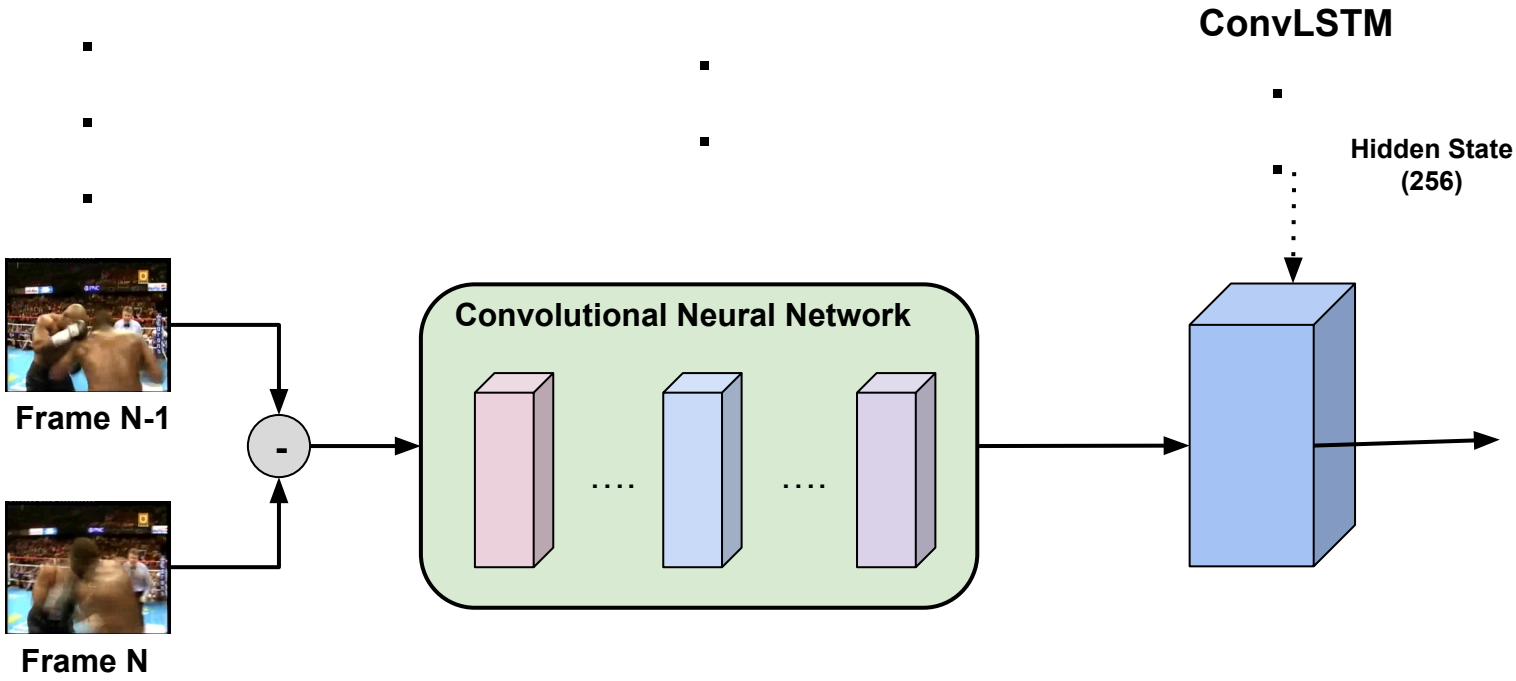



- We randomly augment some *horizontally flipped*, *center cropped* and *corner cropped* images to have better generalization.

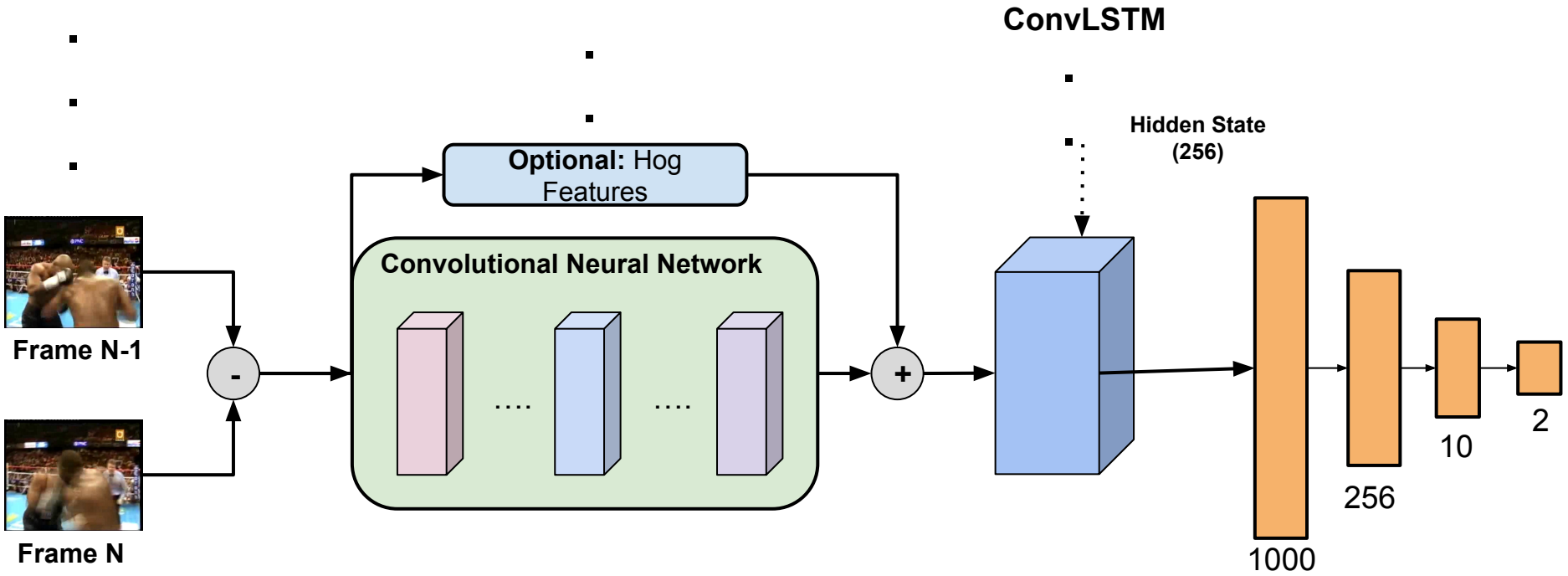
# Methodology - *CNN + Conv-LSTM*



# Methodology - *CNN + Conv-LSTM*



# Methodology - *CNN + Conv-LSTM*



# Results (CNN + ConvLSTM)

- Using 80-20% Training and Testing Splits of each Dataset.
- Batch Size = 16
- Number of Epochs = 50
- Memory Size = 256
- Learning Rate = 1e-4
- RMSProp Optimizer with 0.5 Decay Rate

# Results (CNN + ConvLSTM)

| | 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% |

Compared to state-of-the-art:

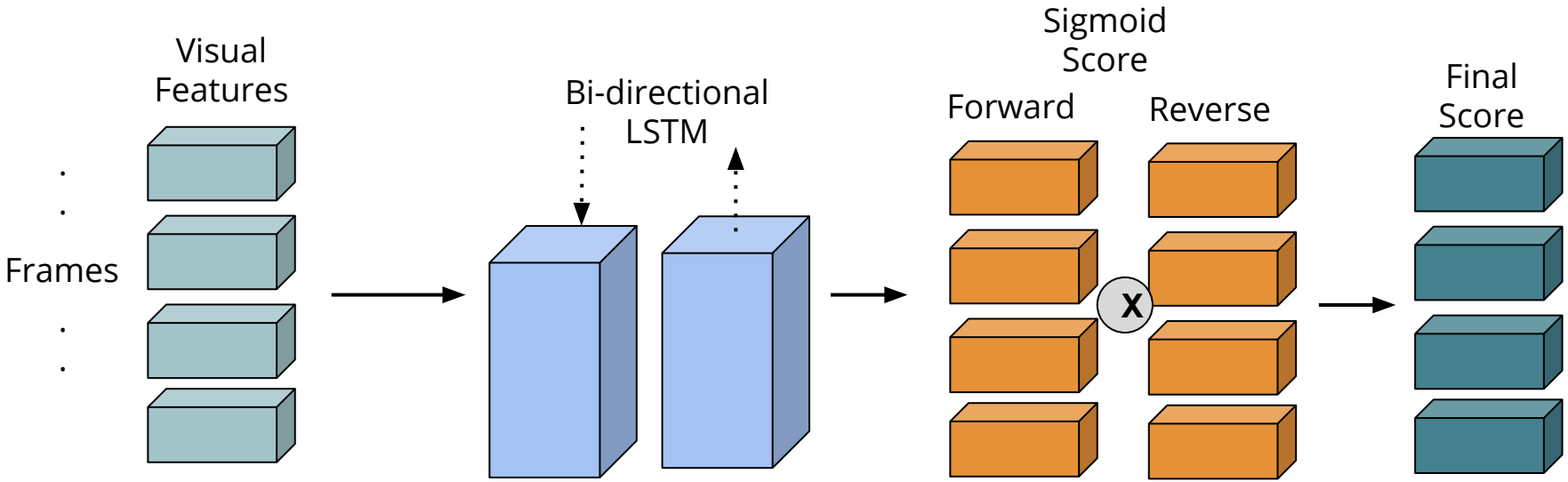
- 97.1% on Hockey Fights → Using 5 fold CV.
- 100% on Peliculas Movies → Using 5 fold CV.
- No reporting for accuracy of the Youtube portion only from Violence Scenes Dataset.

# Caption Generators

- Dense Captioning Images
- Dense Video Captioning
- Proposal Generation Networks

"a person in a black and white photo","man is looking at the camera""a wall behind the man","the leg of a person"

# Proposal Generation Networks



# Results (Proposal Generation Network)

- Using 70-30% Training and Testing Splits of Technicolor Movie Dataset.
- Batch Size = 16
- Number of Epochs = 30
- No of frames = 150
- Learning Rate = 1e-4

| Dataset | Train Instances | Accuracy |
|---|---|---|
| Youtube Videos | 576 | 60.1% |
| Youtube + Movie | 1273 | 63.6% |

# Conclusions

- **CNN-ConvLSTM** architecture **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.

# Future Work

- Trying to feed other kinds of low level features **(ViF and MoSIFT)** to LSTM layer and study their effect.

# Conclusions

- Generating captions for the scenes and use the words generated in classification of videos but this approach failed to perform well due to **lack of large training data**.

- There is ***no ontologies*** for violence words.

# Future Work

- 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 Important Note / Challenge

- The **lack of large and general open source datasets.**
- We need to test the approaches on more general datasets with more real life situations.
- The benchmark ones used are always biased to certain type of scenes like **Hockey dataset**.
- Good Datasets like Technicolor VSD doesn't provide videos themselves but only some ***acoustic and visual features*** extracted from the frames.

# Github Repository

https://github.com/mmaher22/ViolenceDetector

# References


1. 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.
2. 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.
3. 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
4. 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
5. Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno Gar-cía, 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.
6. Swathikiran Sudhakaran and Oswald Lanz. 2017. Learning to detect violent videos using convolutional long short-term memory. In2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE,1–6.
7. Peipei Zhou, Qinghai Ding, Haibo Luo, and Xinglin Hou. 2018.Violence detection in surveillance video using low-level features.PLoS one13, 10 (2018), e0203668.


Thank
You
QUESTIONS