Detection of Real-world Fights in Surveillance Videos

Mauricio Perez, Anderson Rocha and Alex C. Kot

Corresponding author: mauricio001@e.ntu.edu.sg

1. Motivation

- Surveillance cameras (**CCTVs**) are commonly used in many places to enforce security, however their **efficiency is highly questionable**
- "Everyday, over **99% of surveillance videos** being recorded **are never watched** by anyone due to the limitations of traditional surveillance systems." (Vi Dimensions)
- Severe issue for some events, for example **missing fights** can lead to **impunity or serious injuries to those involved**

2. Related Work

- Previous research has **unrealistic or deficient characteristics**:
  - **Too broad definition of violence**, such as explosions, gunshots, etc.
  - Used **fights with artificial movements**, acted by movie actors or researchers
  - **Convenient video properties**: good viewpoint, high resolution, centralized, no occlusion, at close-range and etc
  - **Short clips** from trimmed videos

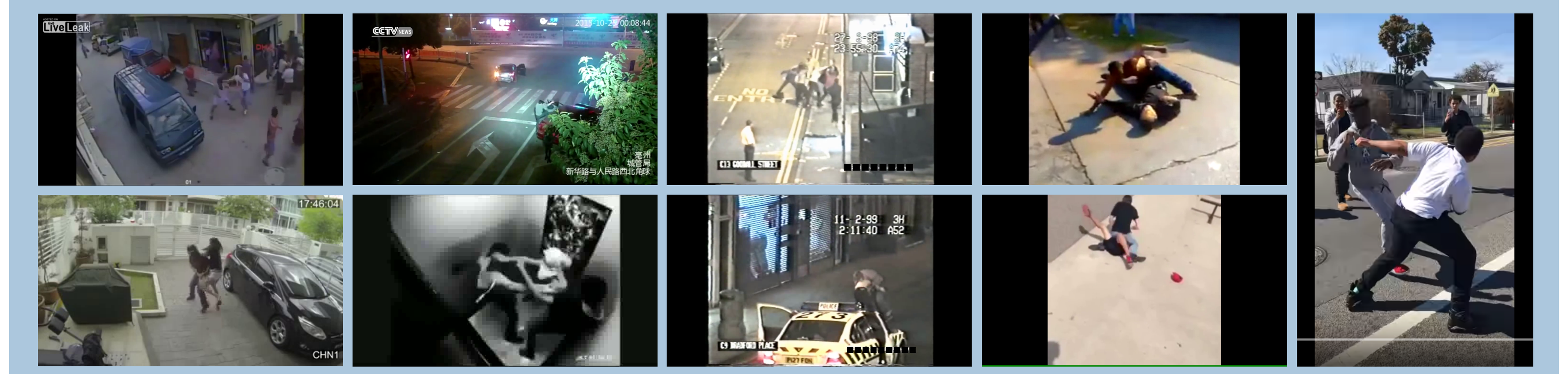
Related Datasets Details

	Name	Size	Characteristics
Trimmed	Hockey Fights	1,000 clips	Hockey players
	Movies	200 clips	Trimmed action movies
	Violent-Flows	246 clips	Crowd violence
Untrimmed	VSD	25 movies	Complete Hollywood movies
	RE-DiD	30 videos	Urban fights + Cars/Mobiles
	BEHAVE	4 videos	Acted fights + CCTVs
	CCTV-Fights	1,000 videos	Urban fights + CCTVs/Mobiles

3. Contributions

- **Creation of CCTV-Fights** dataset:
  - Challenging real-world fights
  - More than 8 hours of CCTV footage
  - Temporally annotated with begin and end of all fight instances
- **Foundational benchmark evaluation** of traditional methods
  - Feature extraction methods ranging from Deep Learning to Local Interest Points
  - Combined with different classifiers, including end-to-end CNN, LSTM and SVM

4. Novel Dataset: CCTV-Fights



- Diverse range of actions
- Multiple fights instances in the same video
- Fight segments temporal annotations
- Short and long videos (5 secs to 12 mins - 2 mins average)
- Videos from Non-CCTV sources as support data
  - Mainly mobile cameras, some very few car-cameras and drone/helicopter.

CCTV-Fights Statistics

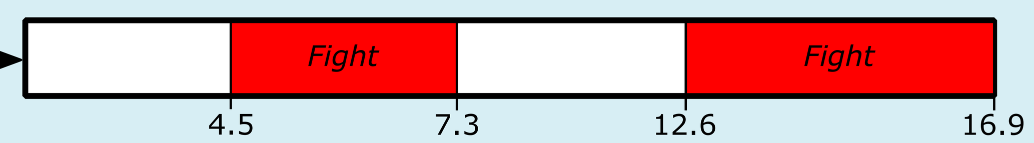
	Videos	Duration (hours)	Fight Instances
All	1,000	17.68	2,414 (2.41)
CCTV	280	8.54	747 (2.67)
Non-CCTV	720	9.13	1,667 (2.32)

5. Benchmark Methodology

Extract Features
- **Two-Stream** CNN: VGG16 [*Simonyan and Zisserman 2014*]
  - Spatial (Frames) + Temporal (Optical Flows)
- 3D-CNN: **C3D** [*Tran et al. 2015*]
- Local interest-points: **TRoF** [*Moreira et al. 2017*]

Classify Frames
- CNN Classification Layer
- LSTM
- SVM

Generate Segments
- Smooth scores with mean filter
- Merge continuous frames that satisfy a certain threshold



6. Experimental Results

Benchmark Results on CCTV-Fights

- mAP (Mean Average Precision):
  - Measurement of correct segments
- F-measure:
  - Precision and recall of fight frames

Features	Classifier	mAP	F-Measure
Two-Stream	CNN	**79.5%**	75.0%
	SVM	76.6%	72.8%
	LSTM	76.0%	**75.9%**
C3D	SVM	64.5%	58.6%
	LSTM	61.0%	58.1%
TRoF	SVM	69.2%	63.3%
	LSTM	63.8%	63.5%

Performance per Stream

Stream	mAP	F-Measure
Spatial	68.6%	61.0%
Temporal	**80.8%**	**75.3%**
Two-Stream	79.5%	75.0%

- Superior performance from Temporal Stream
- No advantage from fusing streams

Results by Videos Source

- Training strategies for **Temporal Stream**:
  - **All:** Both sources at the same time
  - **1-tiered:** Only with CCTV
  - **2-tiered:** First train with both, then fine-tune only with CCTV

Model	Source	mAP	F-Measure
All	All	80.8%	75.3%
	Non-CCTV	85.9%	79.6%
	CCTV	73.7%	66.7%
1-tiered	CCTV	72.1%	63.5%
2-tiered	**CCTV**	**75.6%**	**67.7%**

7. Conclusions

- Information from **explicit motion has a major positive impact** on performance
- Current **spatial features underperformed** greatly and could not positively complement the motion features
- **Sequential information could not be leveraged** by LSTM
- Information from **Non-CCTV sources benefit training** models that better generalizes for the CCTV videos, particularly through a 2-tiered training strategy

Acknowledgments

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, and the Infocomm Media Development Authority, Singapore.

We thank FAPESP DéjàVu grant #2017/12646-3, CAPES DeepEyes grant; and CNPq #304497/2018-5 for the financial support of this research.

Download Link

rose1.ntu.edu.sg/Datasets/cctvFights.asp