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

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.

5. Benchmark Methodology

- Two-Stream CNN: VGG16 [Simonyan and Zisserman 2014]

- 3D-CNN: C3D [Tran et al. 2015]

- Local interest-points: TRoF [Moreira et al. 2017]

- Spatial (Frames) + Temporal (Optical Flows)

- CNN Classification Layer
  
  - LSTM
  
  - SVM

- Smooth scores with mean filter

- Merge continuous frames that satisfy a certain threshold

6. Experimental Results

- Superior performance from Temporal Stream

- No advantage from fusing streams

- 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

- Benchmark Results on CCTV-Fights

<table>
<thead>
<tr>
<th>Features</th>
<th>Classifiers</th>
<th>mAP</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td></td>
<td>79.5%</td>
<td>75.0%</td>
</tr>
<tr>
<td>LSTM</td>
<td></td>
<td>76.6%</td>
<td>72.8%</td>
</tr>
<tr>
<td>C3D</td>
<td>SVM</td>
<td>64.5%</td>
<td>58.6%</td>
</tr>
<tr>
<td>LSTM</td>
<td></td>
<td>61.0%</td>
<td>58.1%</td>
</tr>
<tr>
<td>TRoF</td>
<td>SVM</td>
<td>69.2%</td>
<td>63.3%</td>
</tr>
<tr>
<td>LSTM</td>
<td></td>
<td>63.8%</td>
<td>63.5%</td>
</tr>
</tbody>
</table>

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 #2017/13046-3, CAPES DeepEyes grant; and CNPq #304497/2018-5 for the financial support of this research.