Temporal Robust Features for Violence Detection

Daniel Moreira[1,2], Sandra Avila[1], Mauricio Perez[1], Daniel Moraes[1], Vanessa Testoni[3], Eduardo Valle[1], Siome Goldenstein[1], Anderson Rocha[1]
[1]RECOD Lab., University of Campinas, Brazil — [2]University of Notre Dame, USA — [3]Samsung Research Institute, Brazil

Highlights

- Solution for the automatic detection of violent video.
- Temporal Robust Features (TRoF) are proposed for violent motion description.
- Classification quality is similar to the literature, while being more efficient in terms of runtime and memory footprint.

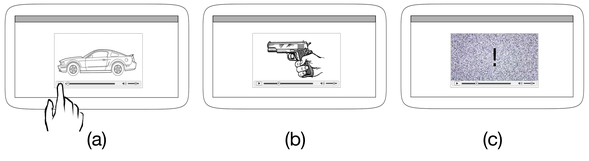
The Problem

What? — Computer-aided violence detection in camera footage and violent movie filtering.

Why? — Solutions from the literature are effective, but not efficient.
We aim at **fast detection** with **low memory footprint**.

Fast detection
For saving lives and catching red-handed criminals.

Low memory footprint
For running on smartphones and tablets.

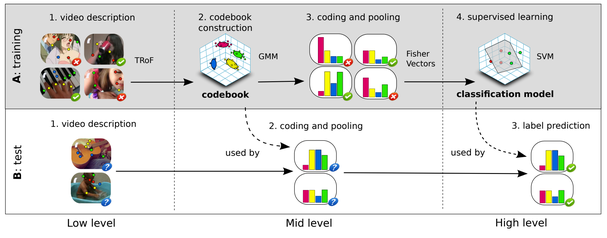


Contact

Daniel Moreira (daniel.moreira.comp@gmail.com)

The Solution

Framework
Three-level BoVW-based machine learning solution, with training and test pipelines.



TRoF Blob Detector

SURF inspired

- Four-variable Hessian maximization.
- Non-maximal suppression of four spatiotemporal scale octaves.
- Use of **3D Gaussian box approximative filters** allied with **integral video** for fast computation.

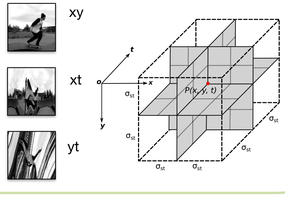
$$H(x,y,t,\sigma_{st}) = \begin{bmatrix} L_{xx}(x,y,t,\sigma_{st}) & L_{xy}(x,y,t,\sigma_{st}) & L_{xt}(x,y,t,\sigma_{st}) \\ L_{xy}(x,y,t,\sigma_{st}) & L_{yy}(x,y,t,\sigma_{st}) & L_{yt}(x,y,t,\sigma_{st}) \\ L_{xt}(x,y,t,\sigma_{st}) & L_{yt}(x,y,t,\sigma_{st}) & L_{tt}(x,y,t,\sigma_{st}) \end{bmatrix}$$

3D box filters
Gaussian approximations quickly convolved through integral video.

TRoF Blob Descriptor

HOG based

- Sampling of *xy*, *xt*, and *yt* blob central planes.
- *64D* HOG features.
- 192D feature vectors.



Watch TRoF!

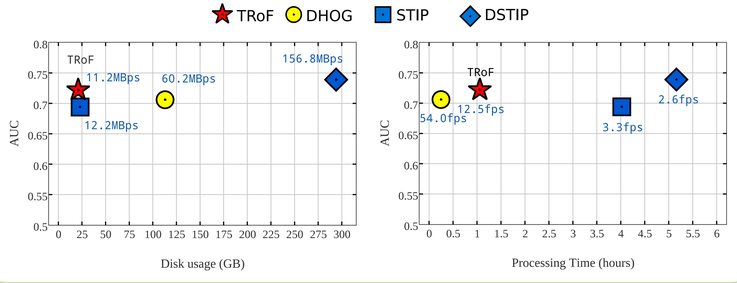
Results

Dataset: **MediaEval** 2013 Violent Scenes Detection task

- Training: 18 Hollywood titles — Test: 7 Hollywood titles.
- Violence: “content which one would not let an 8-year old child see”.

Solution	Media	MAP
Derbas et al.	audio & video	0.690
Tan and Ngo	audio & video	0.689
Dai et al.	audio & video	0.682
Lam et al.	audio & video	0.596
TRoF	video only	0.508

Solution	MAP	AUC
DHOG	0.459	0.706
STIP	0.541	0.694
DSTIP	0.588	0.739
TRoF	0.508	0.722



Conclusions

- Motion description is key for violent detection.
- Dense descriptors are more effective but less efficient.
- TRoF is a non-dense motion-aware video description solution that deals well with the effectiveness vs. efficiency tradeoff.
- TRoF is suitable for generalization tasks (e.g., violence detection, pornography detection, etc.).
- Further investigation is needed regarding action recognition.

Acknowledgements: