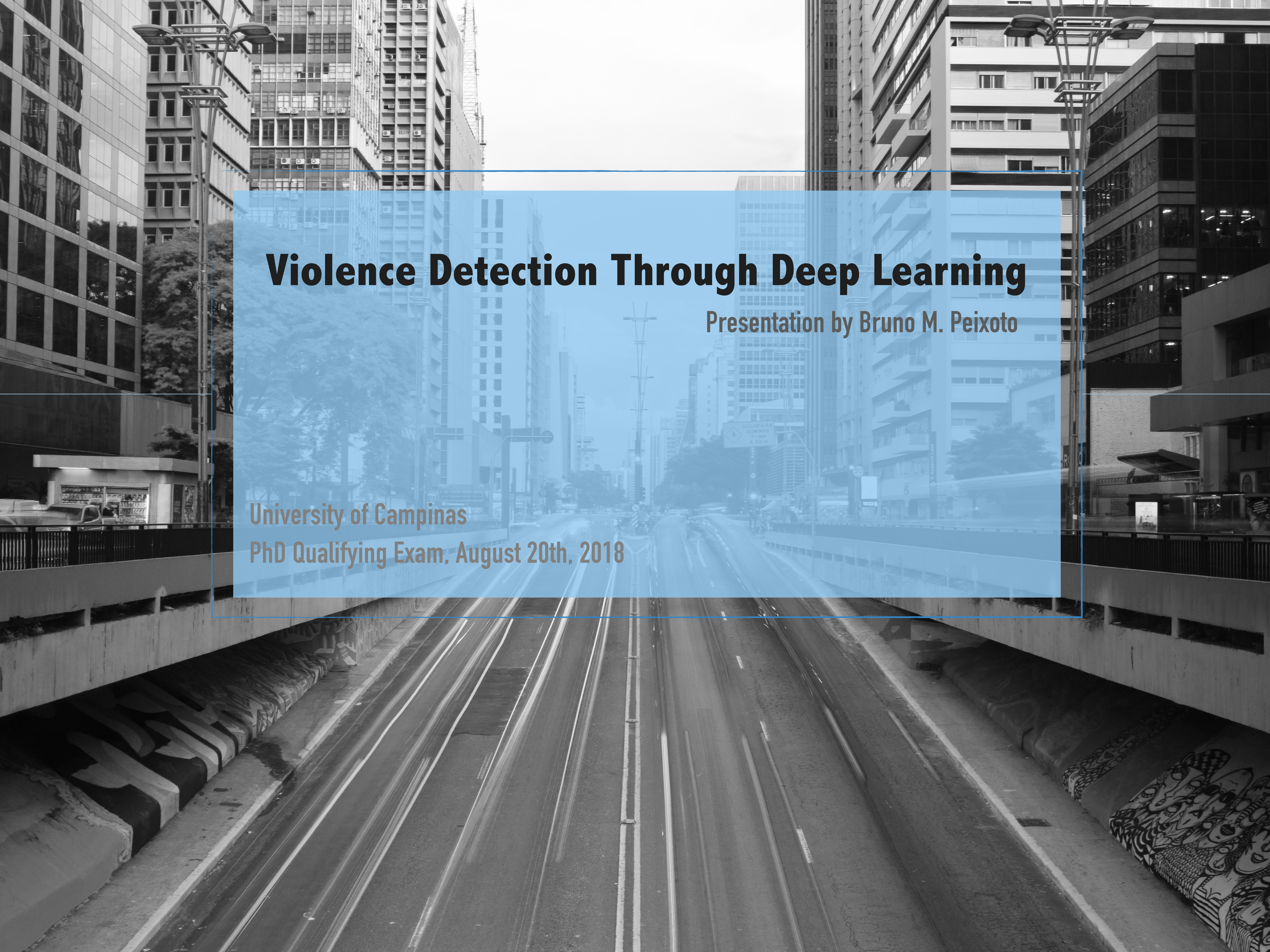


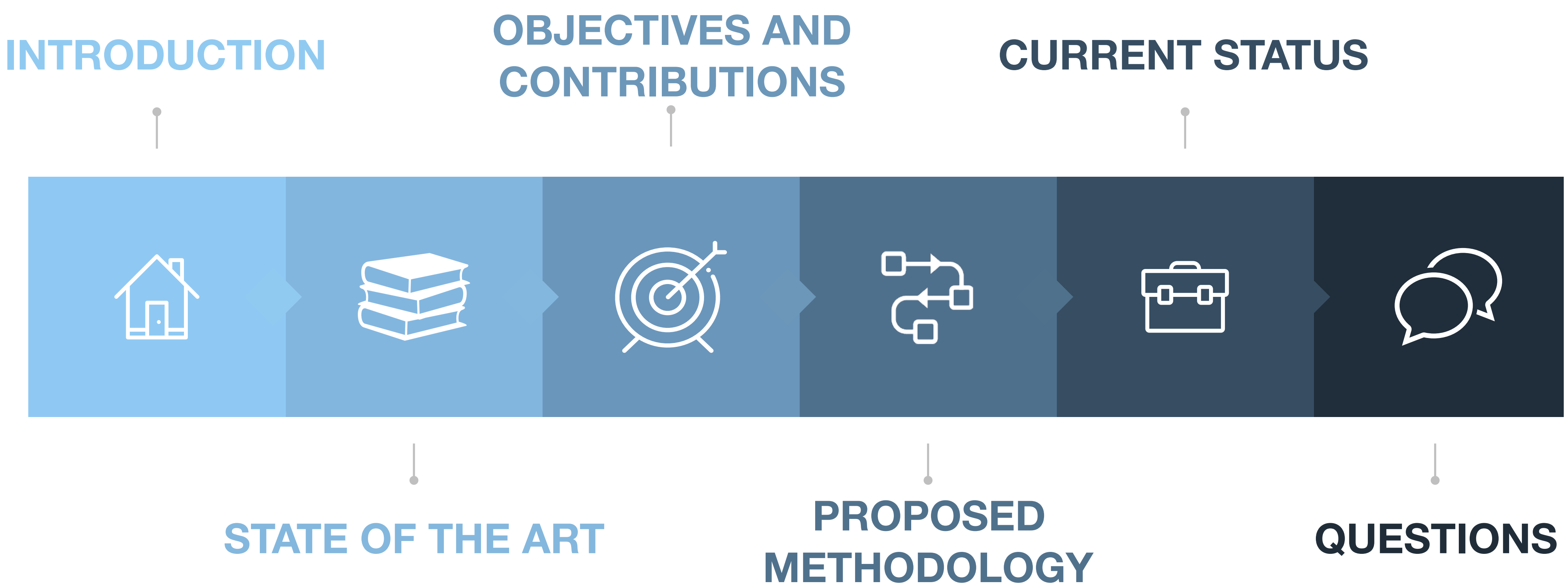
Violence Detection Through Deep Learning

Presentation by Bruno M. Peixoto

University of Campinas

PhD Qualifying Exam, August 20th, 2018

PRESENTATION OUTLINE

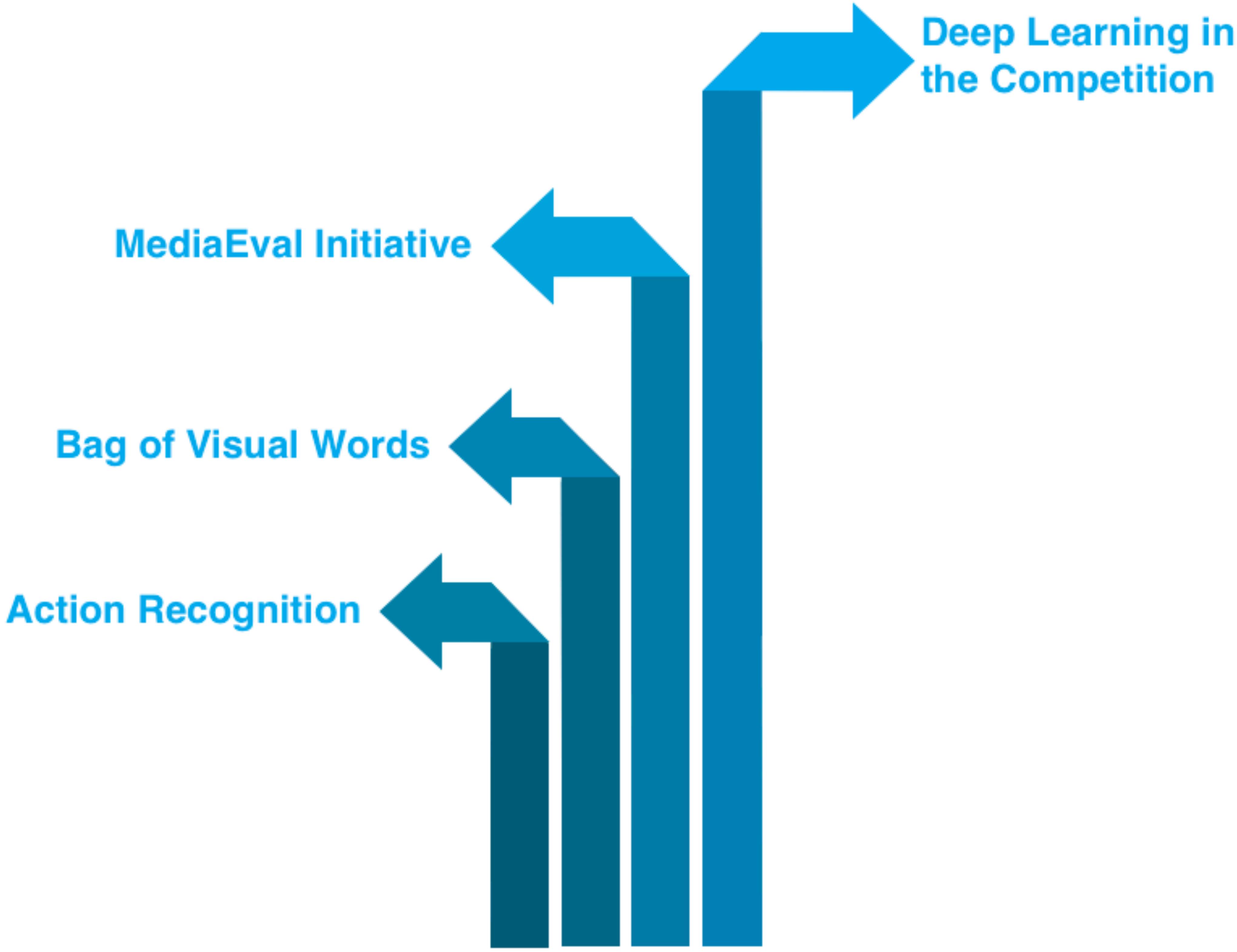


Why Detect Violence?

Reuters/Gonzalo Fuentes

Why Use Deep Learning?

STATE OF ART


Deep Learning in the Competition
MediaEval Initiative
Bag of Visual Words
Action Recognition

VIOLENCE DETECTION TASK IN MEDIAEVAL 2015

Team	CNN	Non-CNN Features	CNN + Others
Fudan-Huawei [1]	**23.5**	16.5	27.0
MIC-TJU [2]	17.4	**21.8**	**28.5**
RFA [3]	14.2	7.7	8.2
RUCMM [4]	11.8	10.6	21.6
KIT [5]	10.2	8.6	12.9
NII-UIT [6]	-	20.8	26.8
UMons [7]	9.67	9.56	-
TCS-ILAB [8]	-	6.4	-
ICL-TUM-PASSAU [9]	-	14.9	-
RECOD [10]	-	11.4	-

Table 1 - Results for the competition are measured in mean average precision (MAP), shown here in percentages.

OBJECTIVES AND CONTRIBUTIONS

REPRESENT VIOLENCE

Explore and find a robust representation of the concept of violence.

TEMPORAL INFORMATION

To reliably detect violence, we consider the action in relation to time.

LOCALIZE VIOLENCE

In some cases, only a specific interval of time is of interest. So we aim to localize them.

DEFINITION OF VIOLENCE

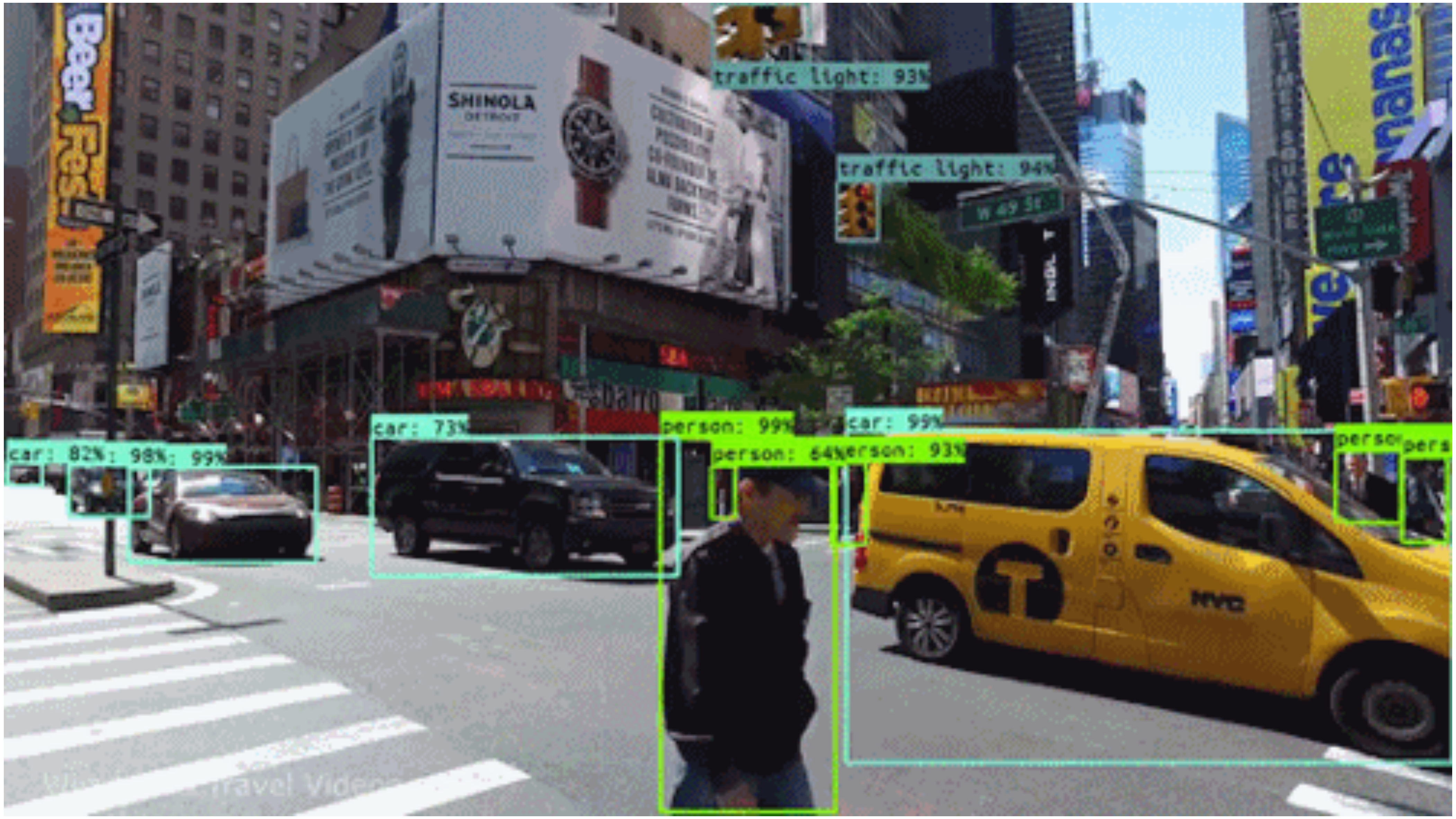
1. A scene is violent if it contains “physical violence or accident resulting in human injury or pain”.

2. A scene is violent if it contains physical violence which “one would not let an eight-year old child see”.

Source: Billy Elliot (2000)

Source: I Am Legend (2007)

REPRESENTING VIOLENCE

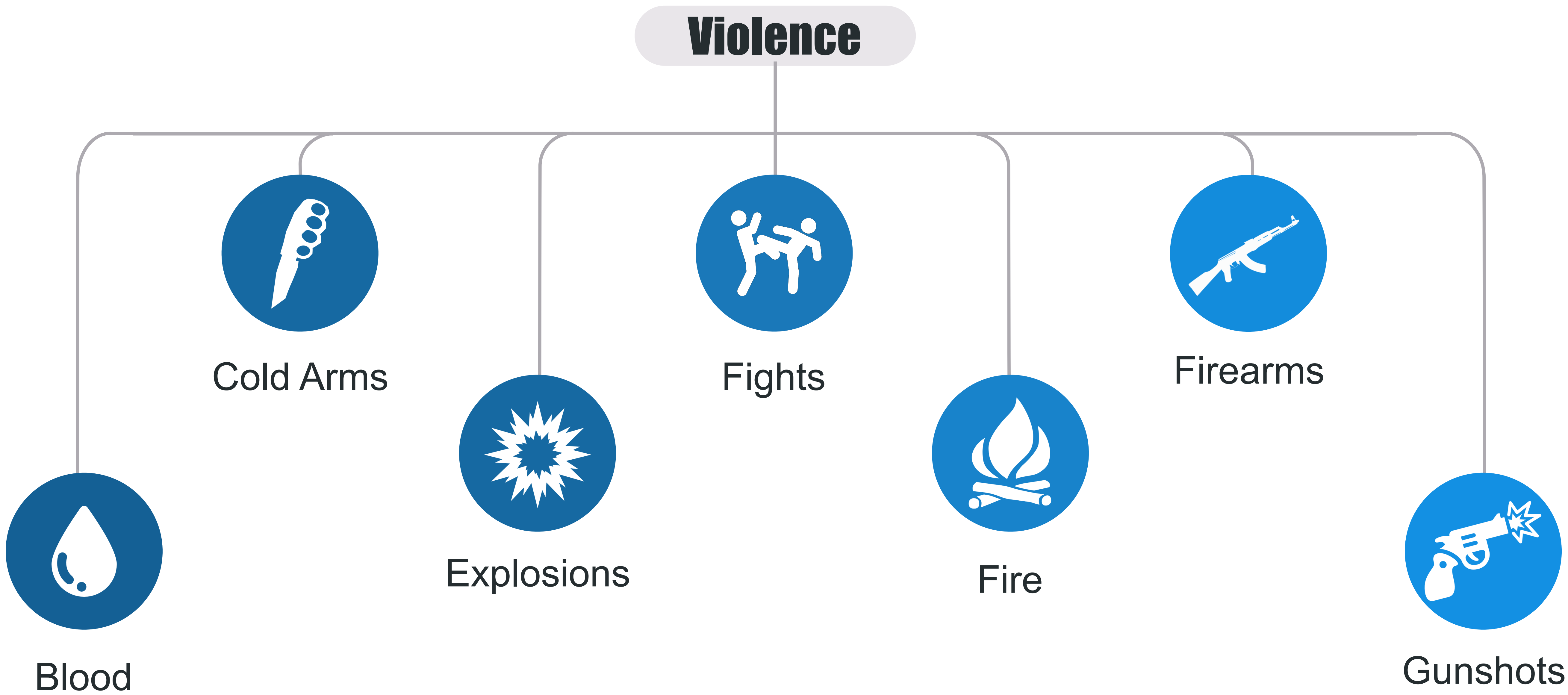


Source: Stefano Massa/Doctorcrowd

REPRESENTING VIOLENCE

Source: Billy Elliot (2000)

CONCEPTS OF VIOLENCE

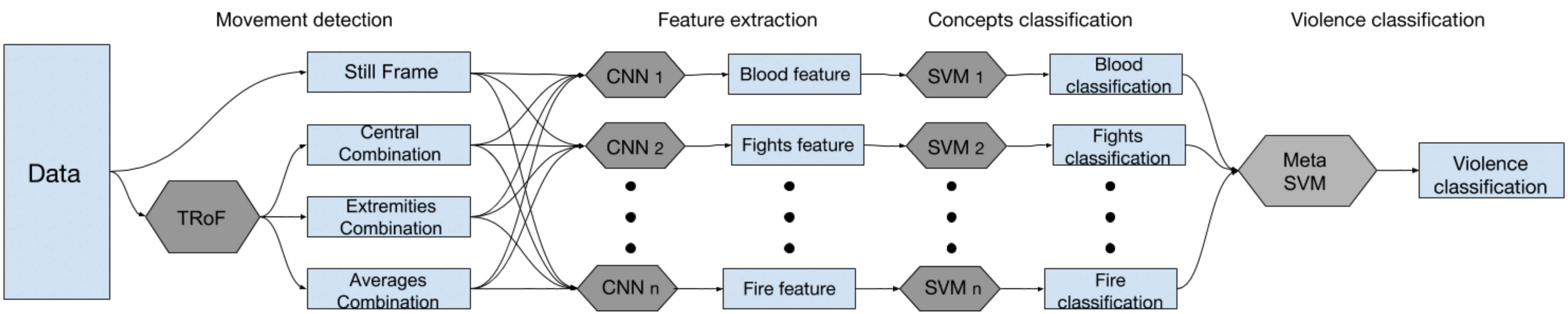


RELEVANCE OF INDIVIDUAL CONCEPTS

	(Percentage of Annotated Shots)	
Concept	**Non violent**	**Violent**
Blood	50.94	49.06
Cold Arms	**76.06**	23.94
Explosions	44.48	55.52
Fights	16.42	**83.58**
Fire	71.18	28.82
Firearms	66.63	33.37
Gunshots	44.57	55.43

Table 2 – Presence of concepts in violent scenes. Dataset for the MediaEval 2013 VSD Task.

PIPELINE



Movement detection
Feature extraction
Concepts classification
Violence classification
Data
TRoF
Still Frame
Central Combination
Extremities Combination
Averages Combination
CNN 1
CNN 2
CNN n
Blood feature
Fights feature
Fire feature
SVM 1
SVM 2
SVM n
Blood classification
Fights classification
Fire classification
Meta SVM
Violence classification

RELATIVE VIOLENCE

Source: Billy Elliot (2000)

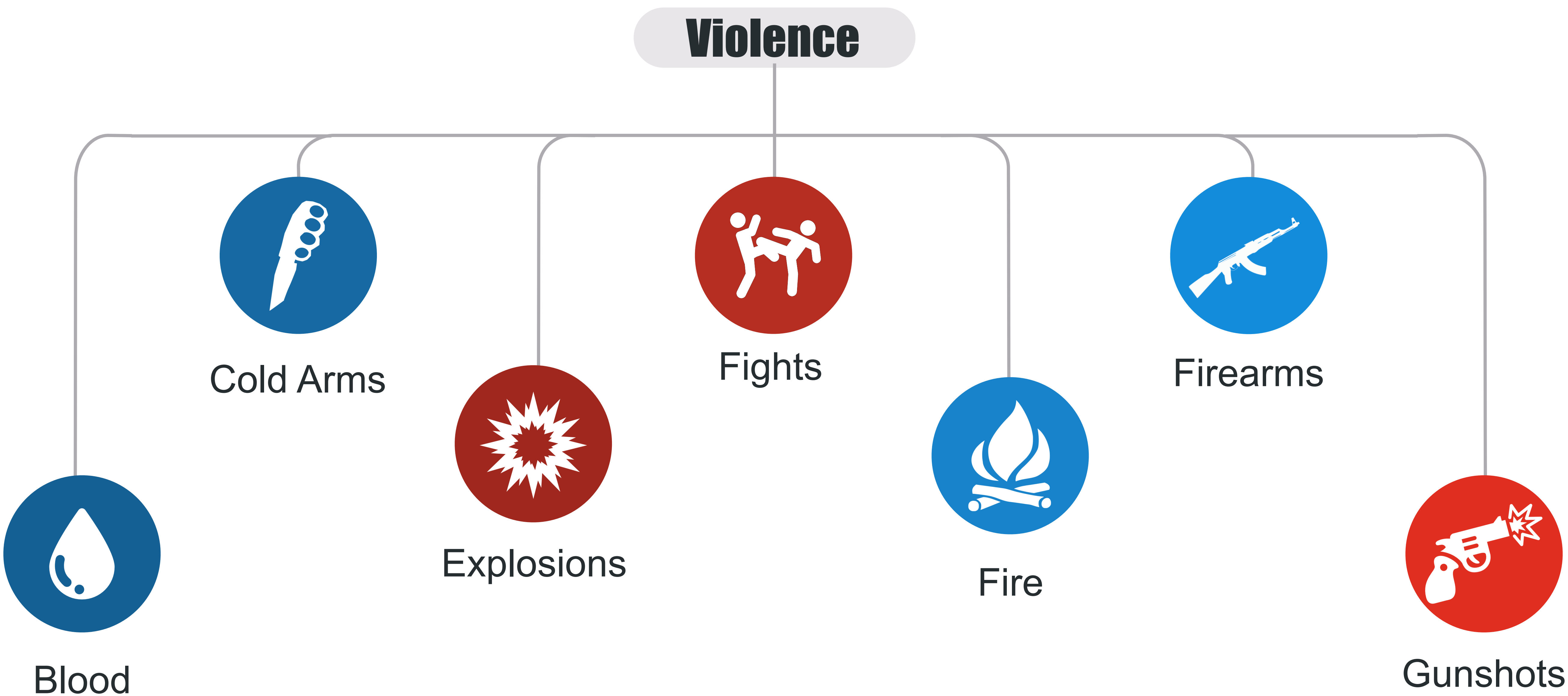
Classifying one scene is ambiguous.

“Humans agree more when they make relative statements.”[11]

Source: Billy Elliot (2000)

INCORPORATING TEMPORAL INFORMATION

Some concepts of violence convey passage of time.

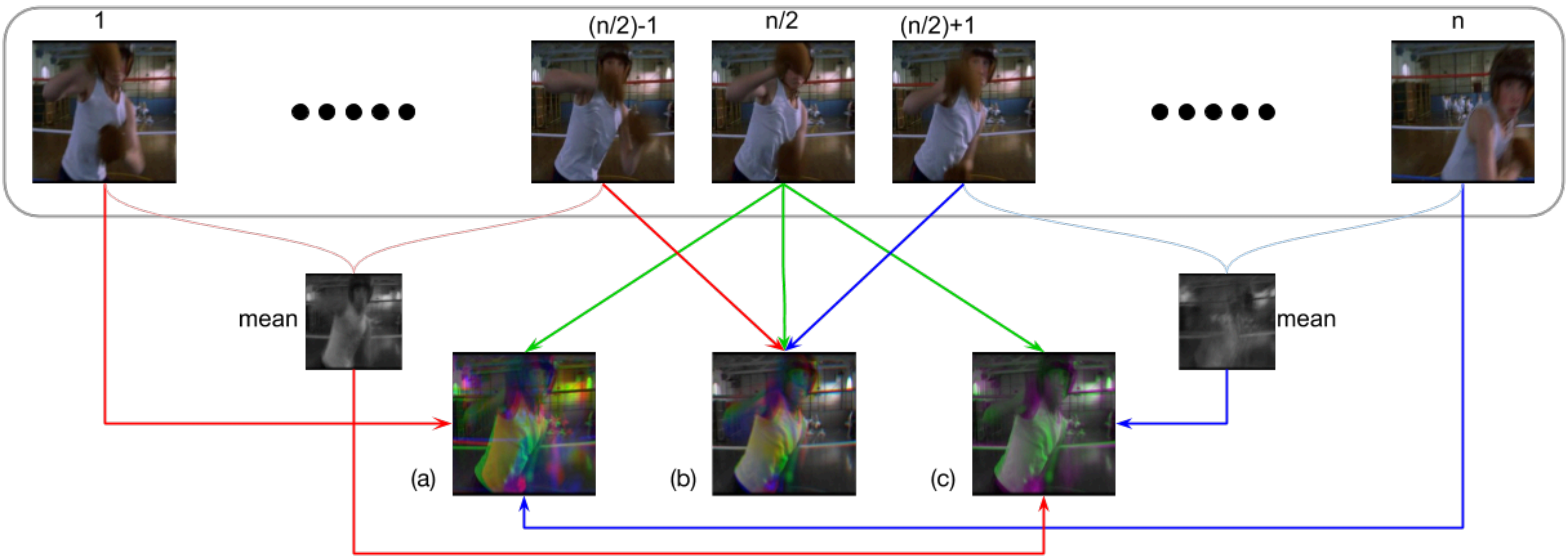


INCORPORATING TEMPORAL INFORMATION

Temporal Robust Features - TRoF

Identify which frames belong to a specific movement

Combine these frames into a single image input



COMBINATIONS

a) Original F

Combination

c) Average Combination

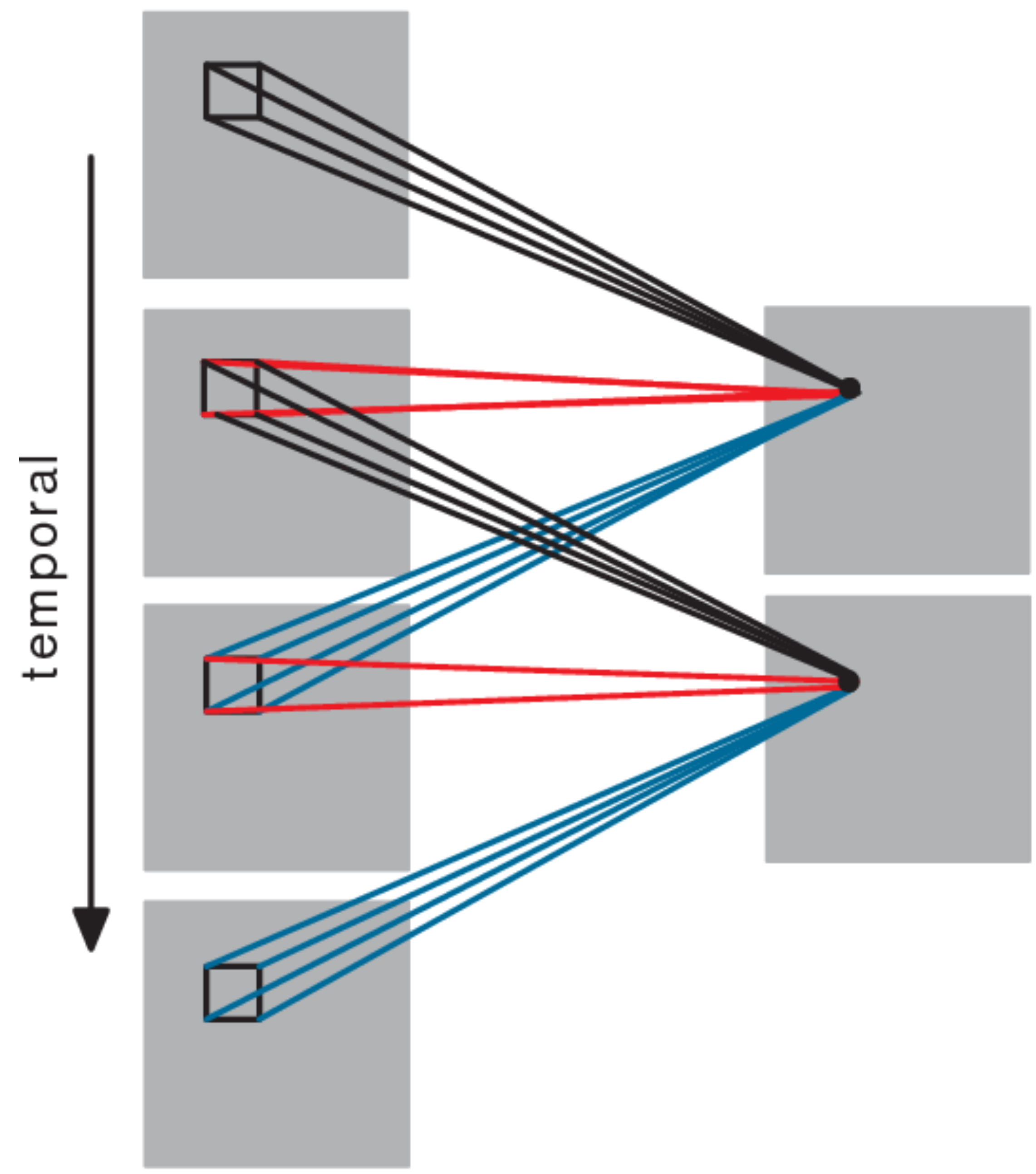
d) Extremities Combination

Source: Billy Elliot (2000)

TEMPORAL INFORMATION IN THE NETWORK

3D Convolutional Neural Network

(a) 2D convolution



(b) 3D convolution

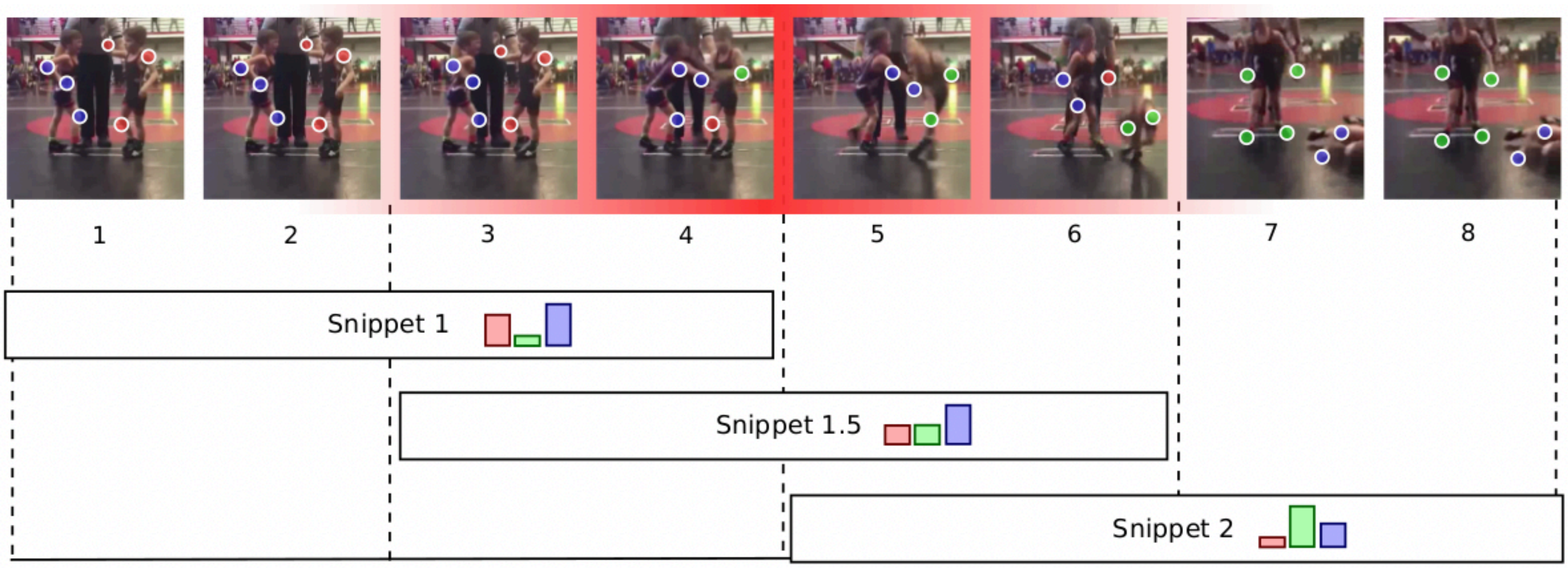
Source: Ji, Shuiwang et al. "3D Convolutional Neural Networks for Human Action Recognition."(2010)

LOCALIZE VIOLENCE

Overlapping Snippets

Varying Lengths

Key frame detection



Source: Moreira, Daniel Henriques "Sensitive-video analysis." (2016)

CURRENT STATUS

A paper will be presented in the 2018 ARES/WSDF conference, in Hamburg, Germany.

CURRENT STATUS

Breaking the concept of violence into seven sub-concepts.

Meta-Classification for Violence.

Combination of frames detected by TRoF.

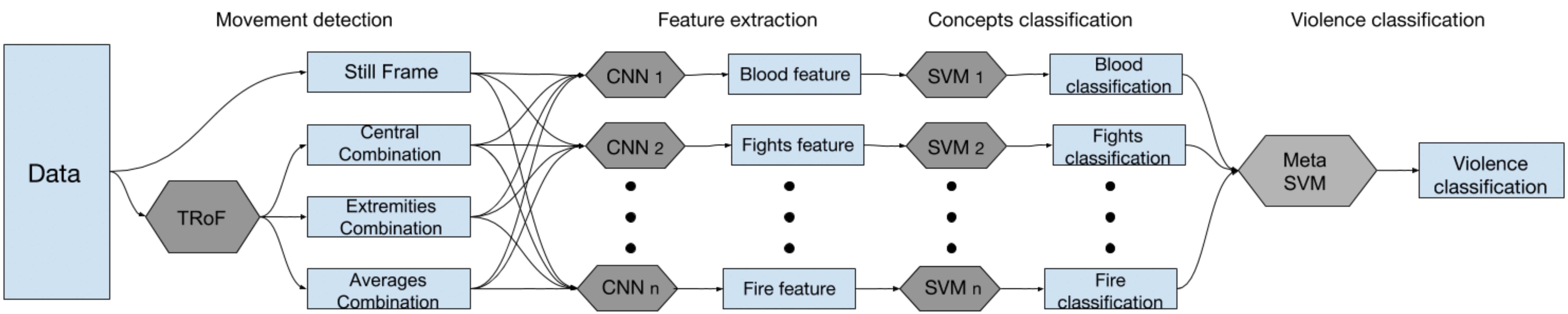
DATASET - MEDIAEVAL 2013 VSD TASK

Hollywood movies

Training set: 17 movies, 2088 min.

Test set: 7 movies, 923 min.

PIPELINE



Movement detection
Feature extraction
Concepts classification
Violence classification
Data
TRoF
Still Frame
Central Combination
Extremities Combination
Averages Combination
CNN 1
CNN 2
CNN n
Blood feature
Fights feature
Fire feature
SVM 1
SVM 2
SVM n
Blood classification
Fights classification
Fire classification
Meta SVM
Violence classification

META-CLASSIFICATION RESULTS

CNN: LeNet architecture fine-tuned for each concept.

SVM: Linear, power-mean and rbf kernels.

Solution	MAP@100	AUC
Original Frames	0.677	0.764
Central Combination	0.682	0.772
Extremities Combination	**0.701**	**0.783**
Averages Combination	0.696	0.779
TRoF [12]	0.508	0.722
LIG-Multimodal [13]	**0.690**	-
Fudan-Multimodal [14]	0.682	-
NII-UIT-Multimodal [15]	0.596	-

Table 3 – Results on the MediaEval 2013 dataset. All multimodal competitors' solutions employed five or more description modalities. Competitors did not report AUC.


How good is each specific
concept in classifying violence?
Scott Strazzante, The Chronicle

RELEVANCE OF INDIVIDUAL CONCEPTS

Training set: 15 movies, 1839 min.

Test set: 2 movies, 249 min.

	Concept x Concept	Concept x Violence
Blood	0.724	0.513
Cold Arms	0.740	0.504
Explosions	0.748	0.634
Fights	0.778	**0.686**
Fire	0.631	0.522
Firearms	0.736	0.501
Gunshots	**0.809**	0.617

Table 4 - Results for the AUC of each concept when classifying shots by its presence and classifying violence by itself.

RELEVANCE OF INDIVIDUAL CONCEPTS

	(Percentage of Annotated Shots)	
Concept	**Non violent**	**Violent**
Blood	50.94	49.06
Cold Arms	**76.06**	23.94
Explosions	44.48	55.52
Fights	16.42	**83.58**
Fire	71.18	28.82
Firearms	66.63	33.37
Gunshots	44.57	55.43

Table 2 – Presence of concepts in violent scenes. Dataset for the MediaEval 2013 VSD Task.

REFERENCES

[1] Qi Dai, Rui-Wei Zhao, Zuxuan Wu, Xi Wang, Zichen Gu, Wenhai Wu, Yu-Gang Jiang:Fudan-Huawei at MediaEval 2015: Detecting Violent Scenes and Affective Impact in Movies with Deep Learning.
[2] Yun Yi, Hanli Wang, Bowen Zhang, Jian Yu: MIC-TJU in MediaEval 2015 Affective Impact of Movies Task.
[3] I. Mironica et al., RFA at MediaEval 2015 Affective Impact of Movies Task: A Multimodal Approach.
[4] Qin Jin, Xirong Li, Haibing Cao, Yujia Huo, Shuai Liao, Gang Yang, Jieping Xu: RUCMM at MediaEval 2015 Affective Impact of Movies Task: Fusion of Audio and Visual Cues.
[5] M. Vlastelica P. et al, KIT at MediaEval 2015 - Evaluating Visual Cues for Affective Impact of Movies Task.
[6] V Lam et al., NII-UIT at MediaEval 2015 Affective Impact of Movies Task.
[7] Omar Seddati, Emre Kulah, Gueorgui Pironkov, Stéphane Dupont, Saïd Mahmoudi, Thierry Dutoit: UMons at MediaEval 2015 Affective Impact of Movies Task including Violent Scenes Detection.
[8] Rupayan Chakraborty, Avinash Kumar Maurya, Meghna Pandharipande, Ehtesham Hassan, Hiranmay Ghosh, Sunil Kumar Kopparapu: TCS-ILAB - MediaEval 2015: Affective Impact of Movies and Violent Scene Detection.
[9] George Trigeorgis, Eduardo Coutinho, Fabien Ringeval, Erik Marchi, Stefanos Zafeiriou, Björn W. Schuller: The ICL-TUM-PASSAU Approach for the MediaEval 2015 "Affective Impact of Movies" Task.
[10] Daniel Moreira, Sandra Eliza Fontes de Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, Anderson Rocha: RECOD at MediaEval 2015: Affective Impact of Movies Task.
[11] A. Kovashka et al. WhittleSearch: Interactive Image Search with Relative Attribute Feedback. 2012
[12] Daniel Moreira, Sandra Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. Pornography classification: The hidden clues in video space–time. Forensic Science International, 268:46–61, 2016.
[13] Nadia Derbas, Bahjat Safadi, and Georges Quénot. LIG at MediaEval 2013 Affect Task: Use of a Generic Method and Joint Audio-Visual Words.
[14] Qi Dai, Jian Tu, Ziqiang Shi, Yu-Gang Jiang, and Xiangyang Xue. Fudan at MediaEval 2013: Violent Scenes Detection Using Motion Features and Part-Level Attributes.
[15] Vu Lam, Duy-Dinh Le, Sang Phan, Shinichi Satoh, and Duc Anh Duong. NII-UIT at MediaEval 2013 Violent Scenes Detection Affect Task. In MediaEval, 2013.

Questions?

Thank You!

Suggestions?