Harnessing high-level concepts, visual, and auditory features for violence detection in videos

Bruno Malveira Peixoto

Advisors: Anderson Rocha
Zanoni Dias

University of Campinas
PhD Defense – July 14th, 2021

PRESENTATION OUTLINE

INTRODUCTION

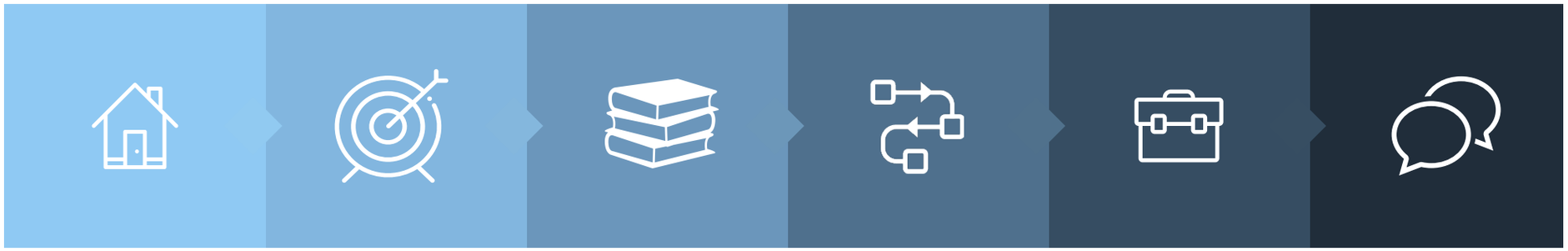
RESEARCH QUESTIONS

STATE OF THE ART

PROPOSED METHODOLOGY

RESULTS

CONCLUSION



Why Detect Violence?

Reuters/Gonzalo Fuentes

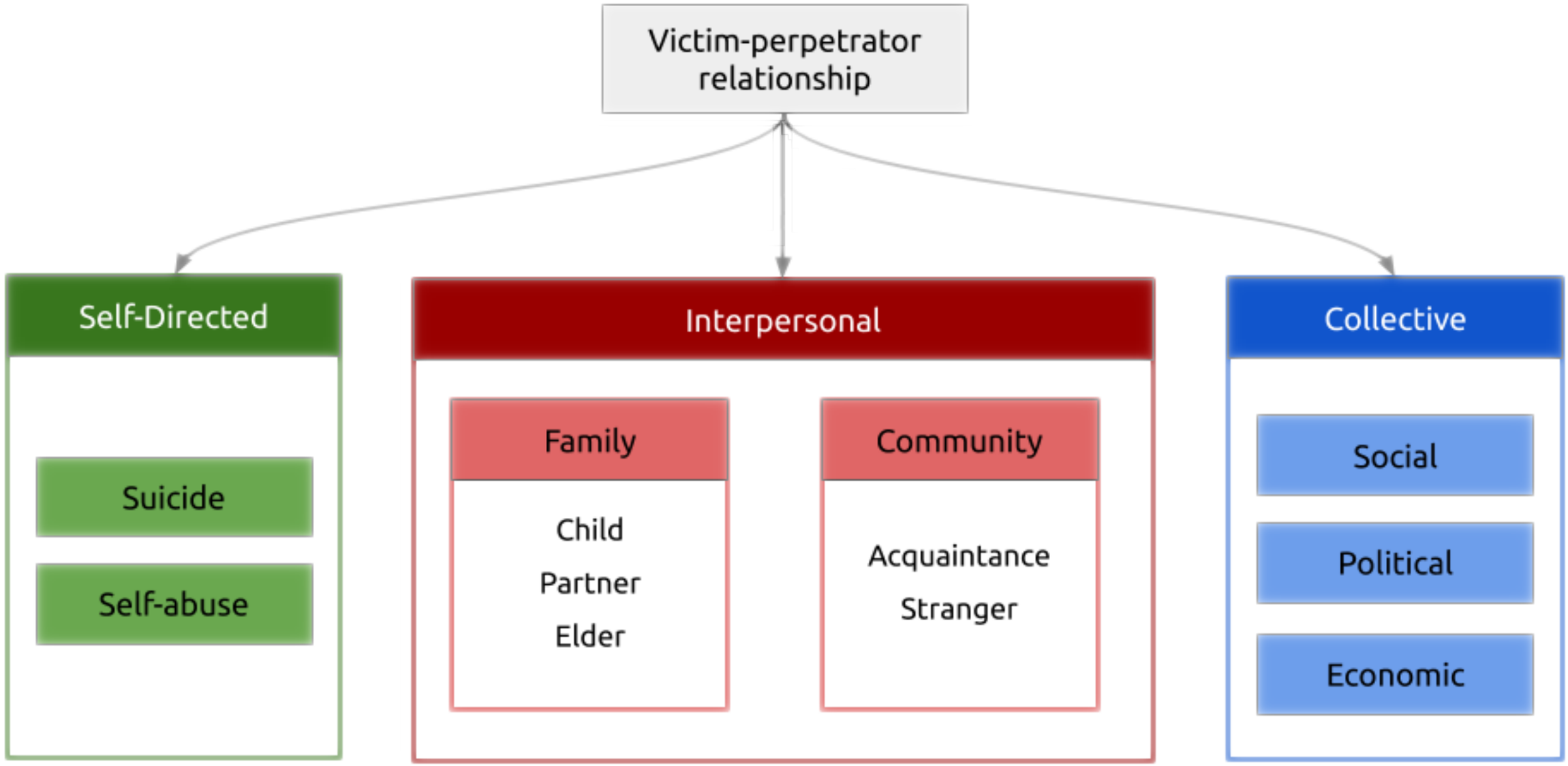
RESEARCH QUESTIONS

How to define violence in terms that a computer can understand?

How different kinds of violence interact to enable a system to understand the meaning of violence in general?

How to build a system that relies upon visual and sound signals characteristics to detect violence in the general sense?

DEFINITION OF VIOLENCE



Source: World Health Organization

DEFINITION OF VIOLENCE

Definitions from the MediaEval VSD task dataset

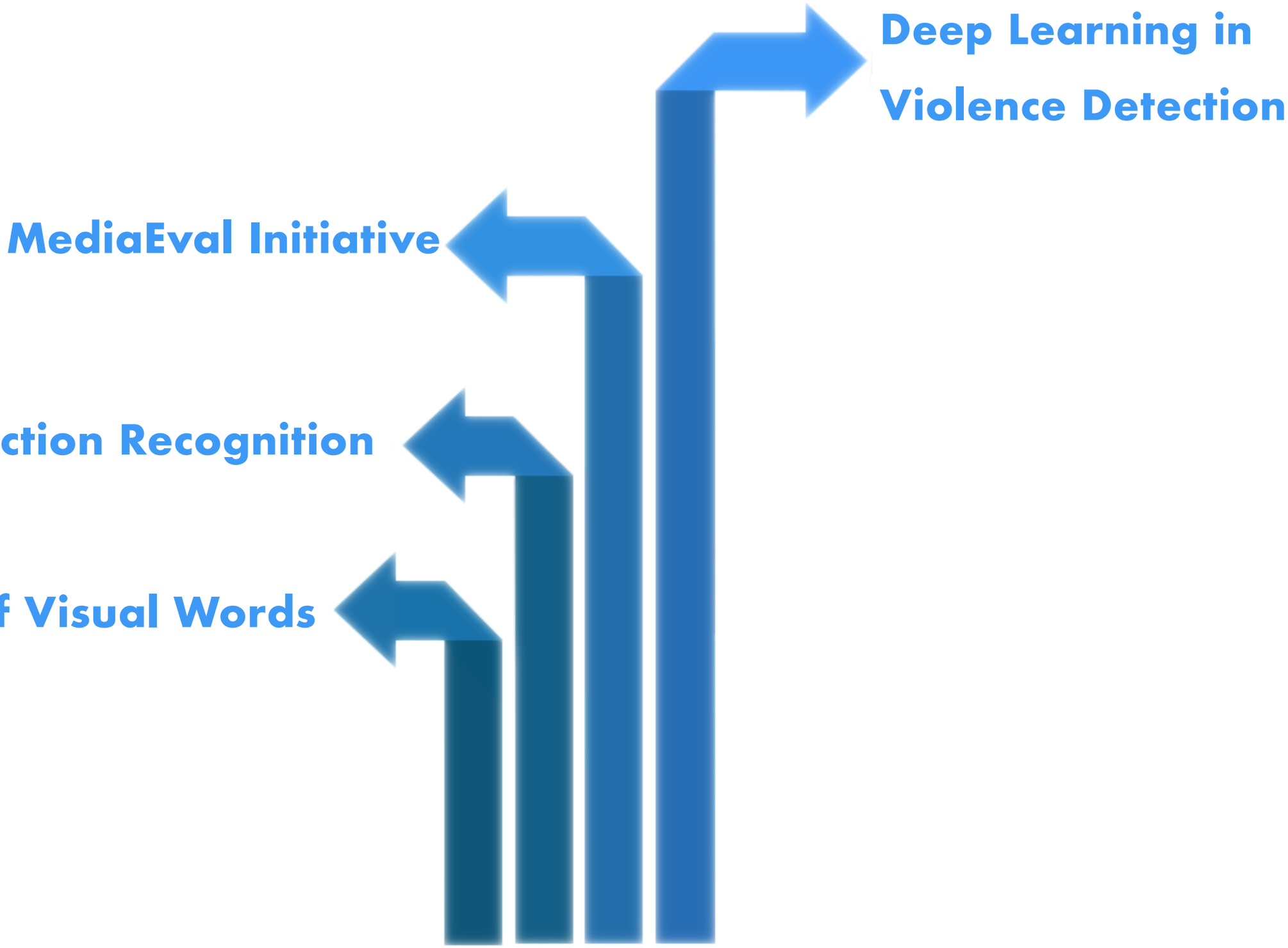
1. A scene is violent if it contains “physical violence or accident resulting in human injury or pain”.

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)

STATE OF ART


Deep Learning in
Violence Detection
MediaEval Initiative
Action Recognition
Bag of Visual Words

CLASSICAL DATASETS

Hockey Fights

Violence in Movies

Violent Flows

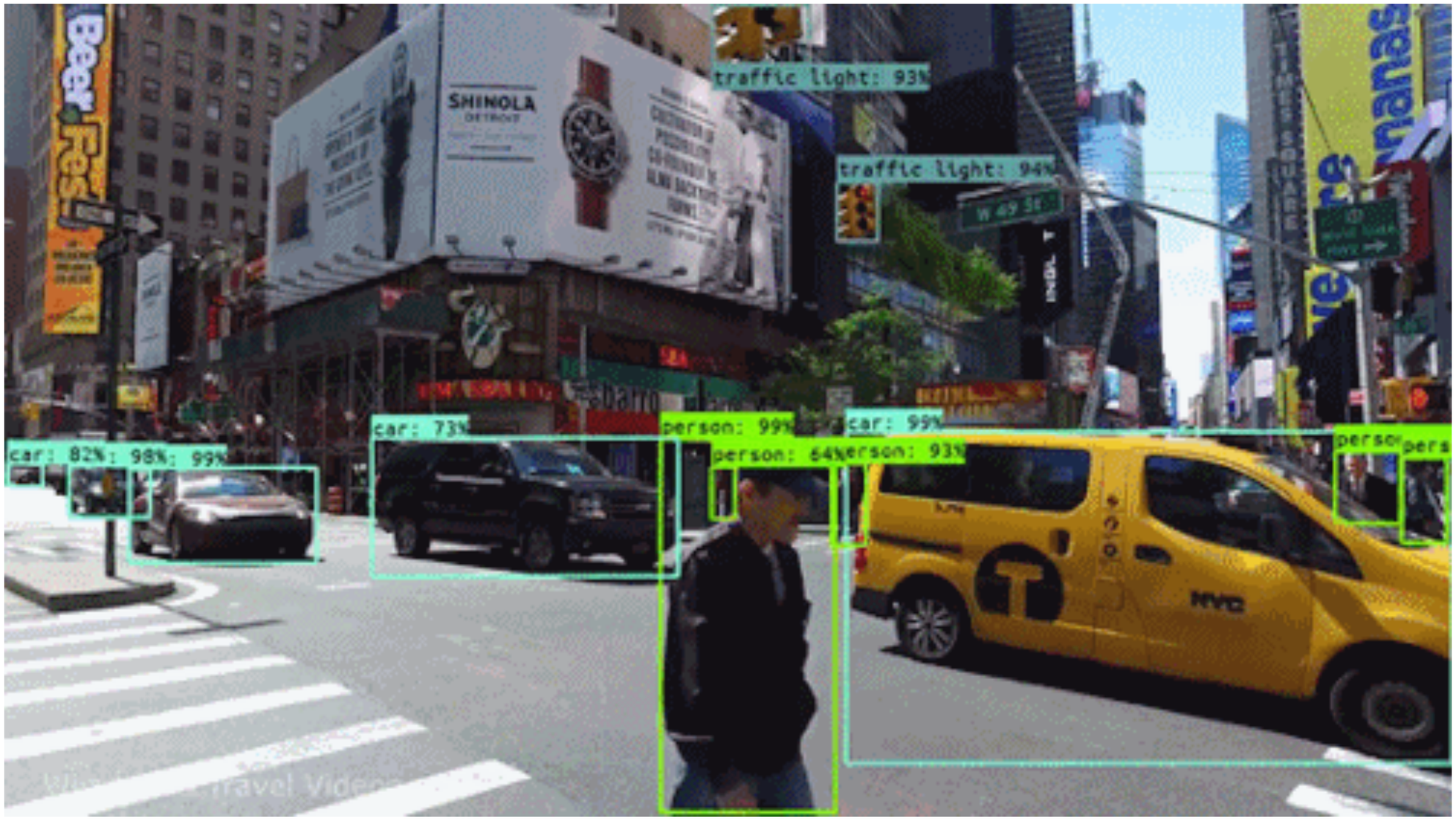
Source: A survey on video content rating: taxonomy, challenges and open issues, Pour et al. 2021

RECENT STUDIES IN VIOLENCE DETECTION

Year	Authors	Features / Classifier	Accuracy in Datasets (%)		
			Hockey	Movies	Flows
2011	Bermejo et al.	BoVW - MoSIFT + SVM(HIK)	90.90	89.50	-
2016	Bilinski et al.	FV / Sliding Window + SVM (χ^2)	93.70	99.50	96.40
2016	Dong et al.	Three-streams + LSTM	93.90	-	-
2017	Senst et al.	Lagrangian SIFT + SVM (χ^2)	94.42	94.95	93.12
2017	Meng et al.	CNN + Optical Flow + IDT	98.60	-	92.50
2017	Sudhakaran et al.	Convolutional LSTM	97.10	100.00	94.57
2018	Hanson et al.	Biconvolutional LSTM	96.96	100.00	92.18
2018	Mumtaz et al.	Transfer learning from Inception	99.28	99.97	-
2019	Ullah et al.	CNN + 3D CNN	96.00	99.99	98.00

Table 1 – Summary of studies in classical datasets.

REPRESENTING VIOLENCE



Source: Stefano Massa/Doctorcrowd

REPRESENTING VIOLENCE

Source: Billy Elliot (2000)

CONCEPTS OF VIOLENCE

Blood

Cold Arms

Fights

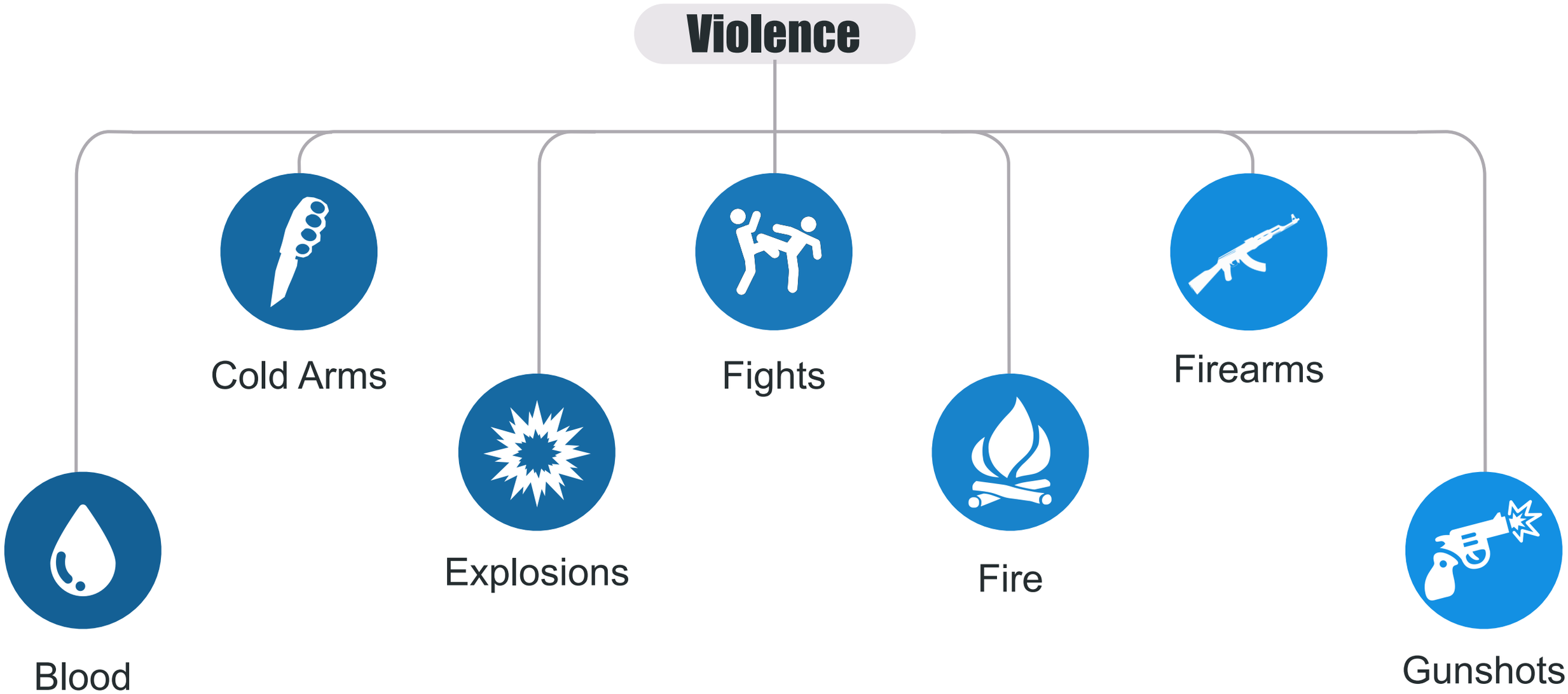
Explosions

Fire

Firearms

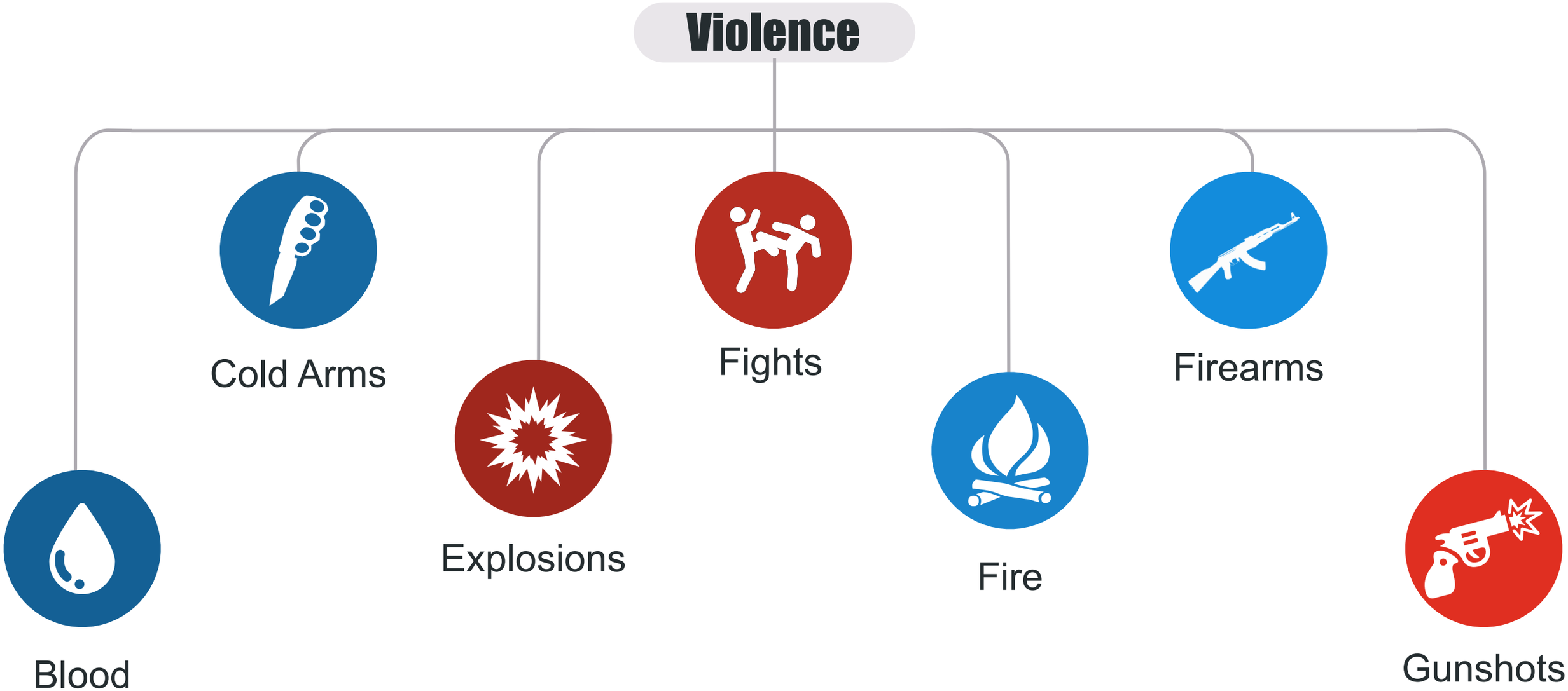
Gunshots

CONCEPTS OF VIOLENCE



INCORPORATING TEMPORAL INFORMATION

Some concepts of violence convey passage of time.

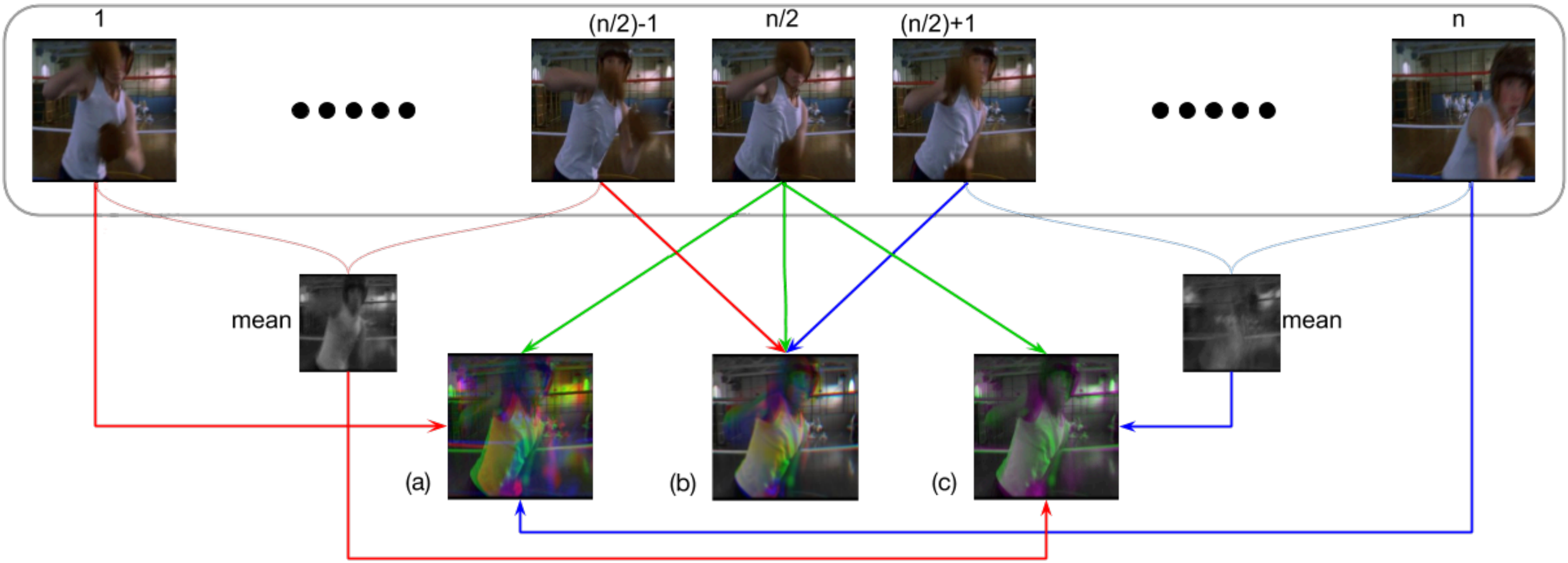


INCORPORATING TEMPORAL INFORMATION

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

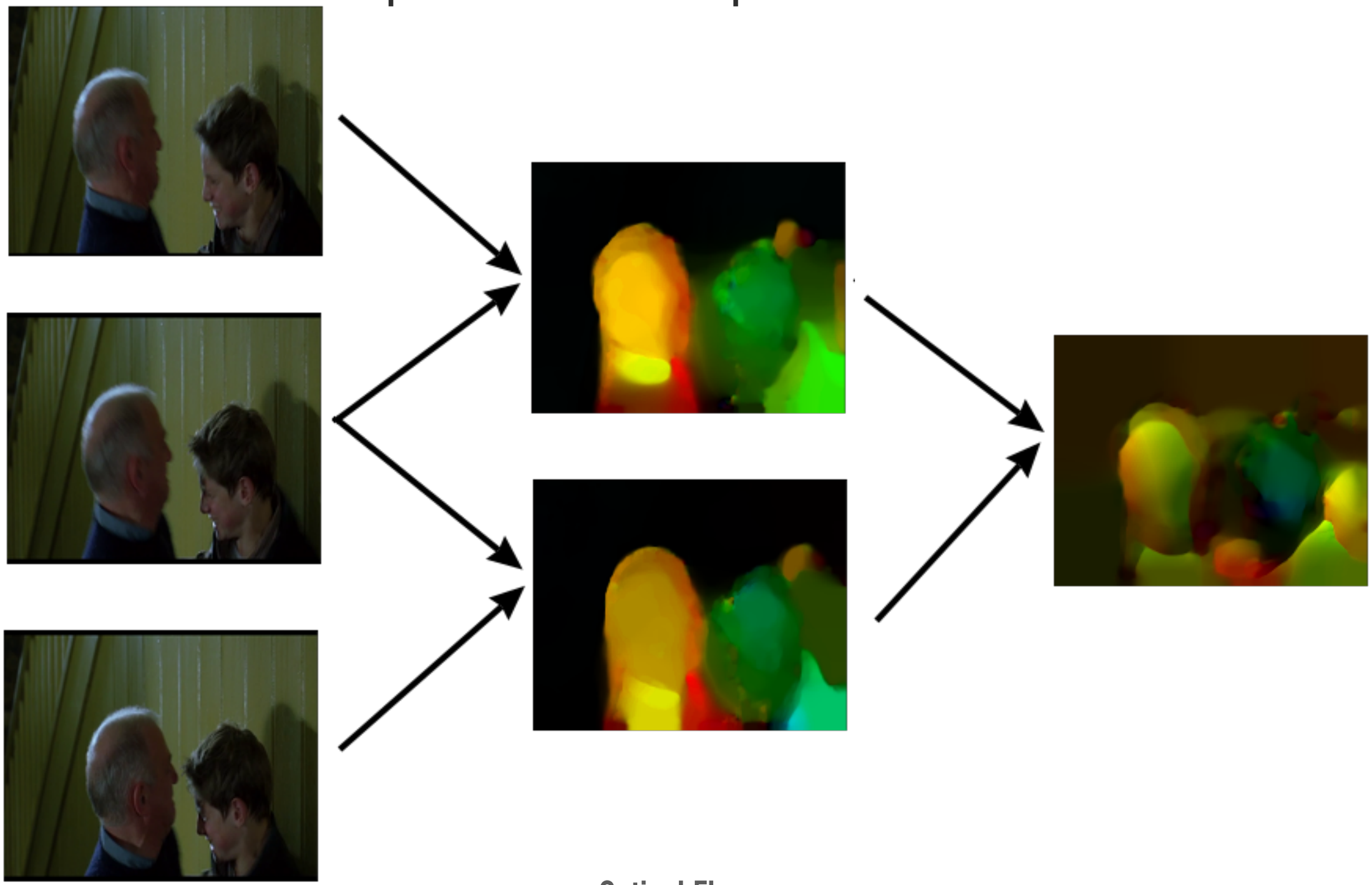
c) Average Combination

d) Extremities Combination

Source: Billy Elliot (2000)

INCORPORATING TEMPORAL INFORMATION

Optical Flow and Optical Acceleration



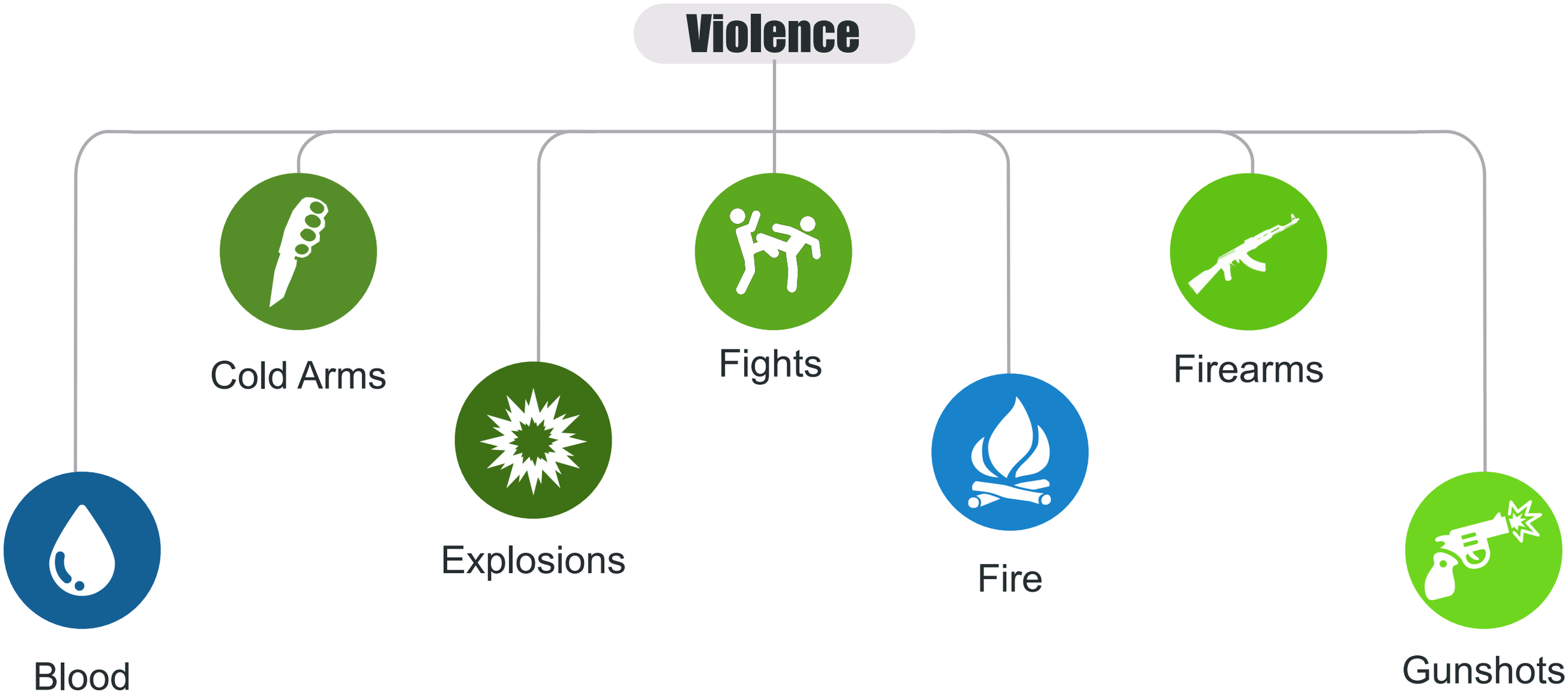
Original Frames

Optical Flow

Optical Acceleration

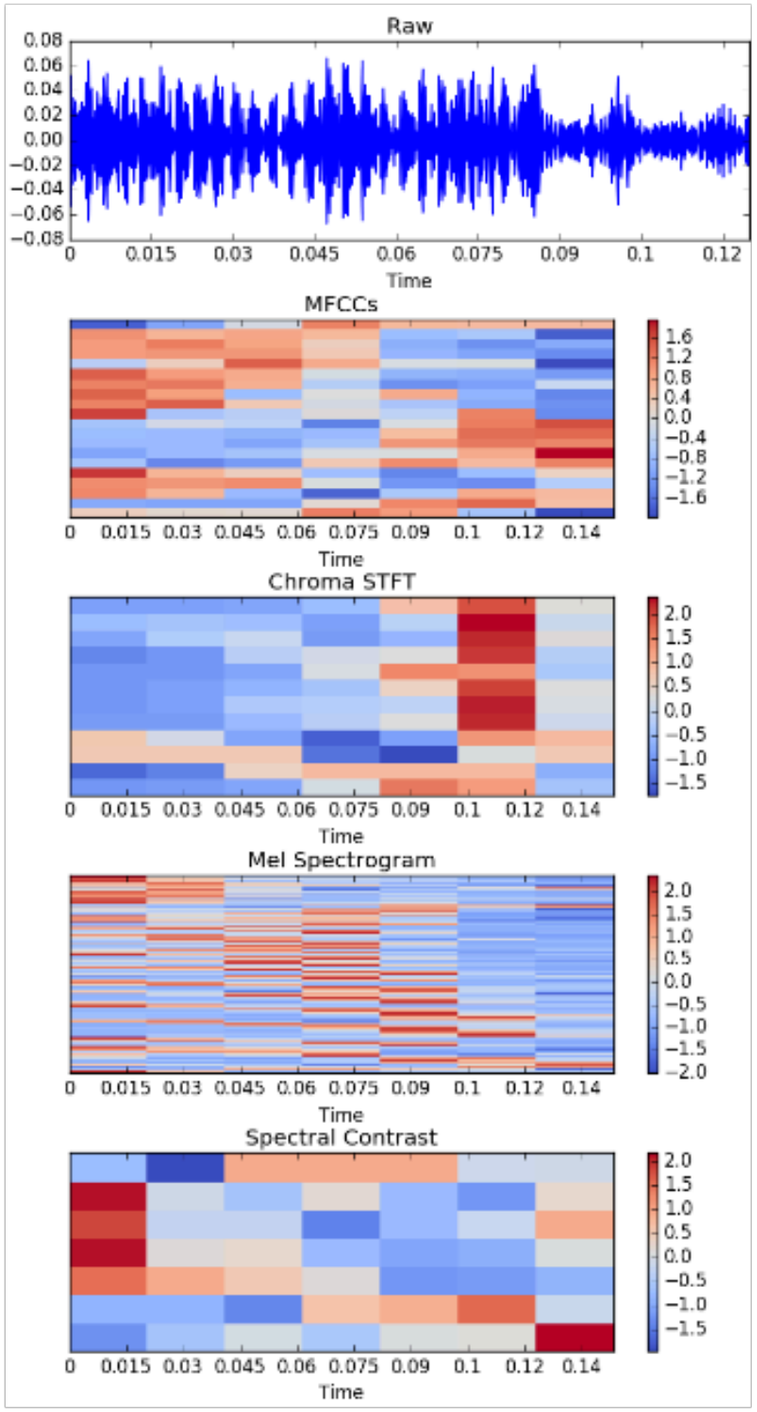
INCORPORATING AUDIO INFORMATION

Different concepts of violence have different sound signals.

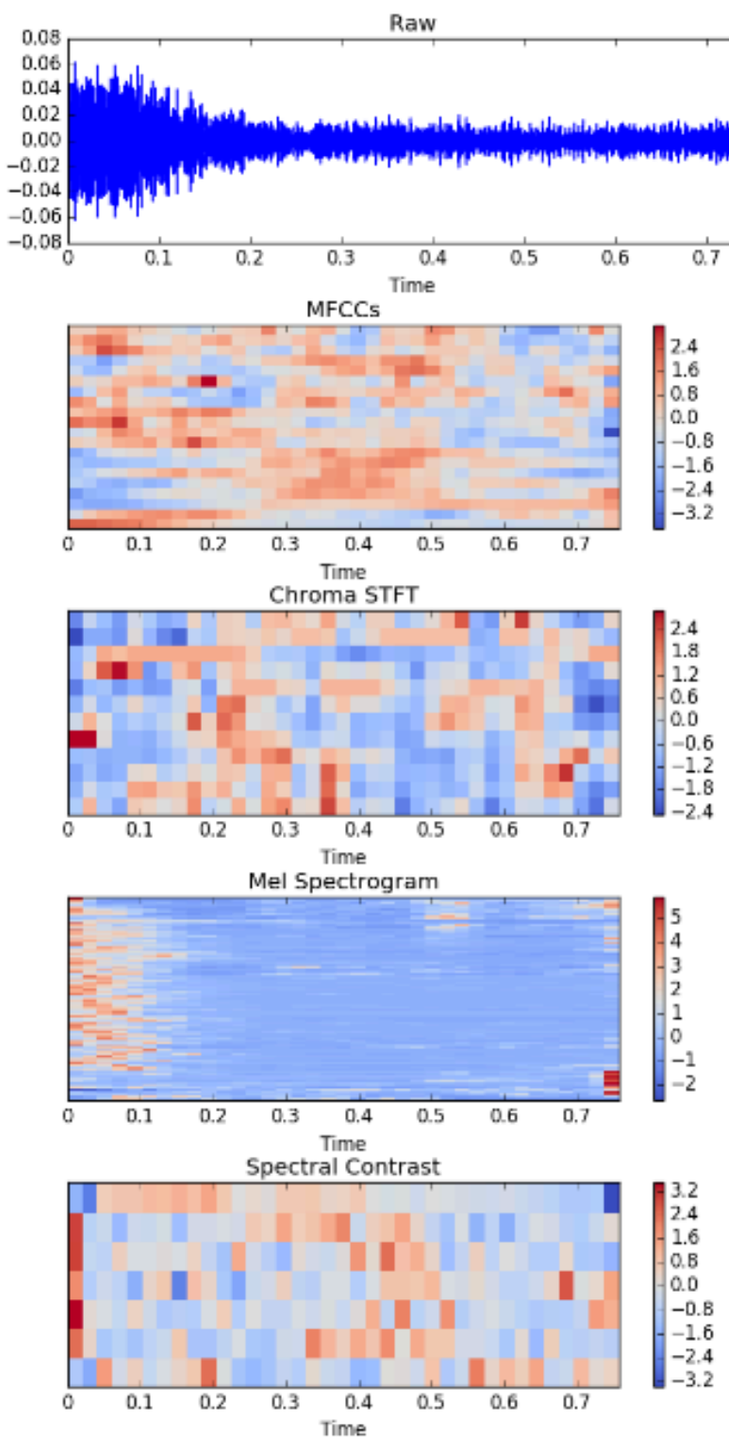


INCORPORATING AUDIO INFORMATION

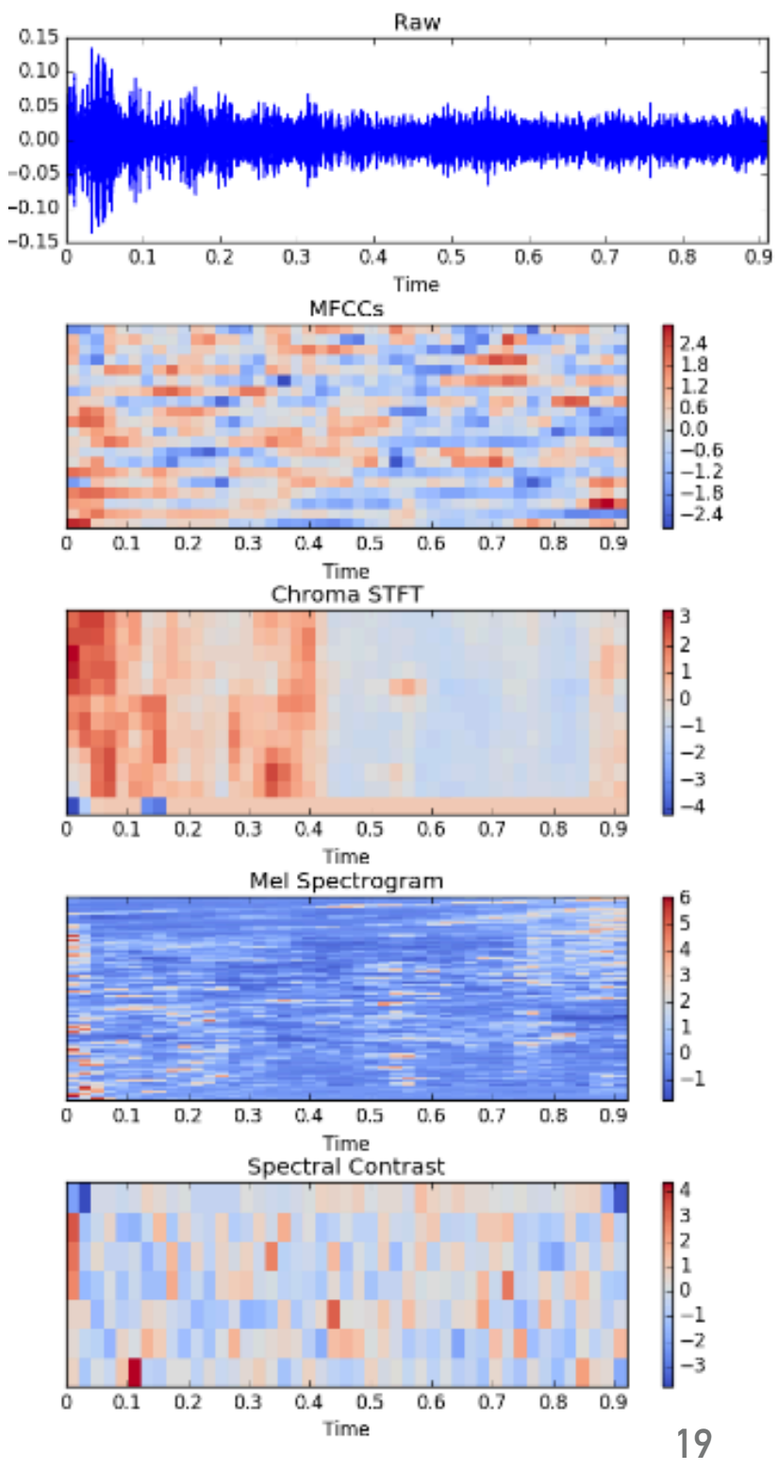
Gunshot



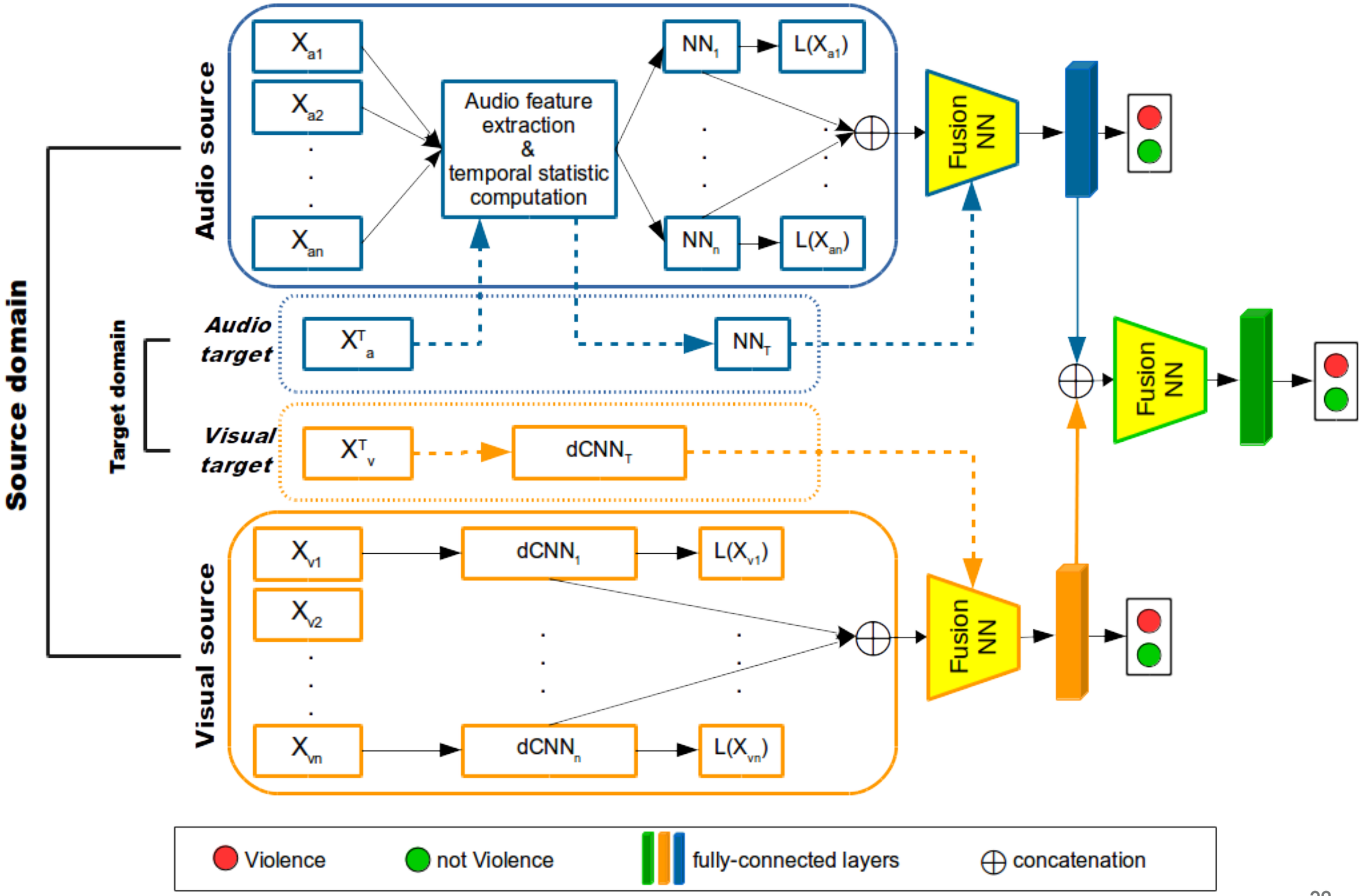
Explosion



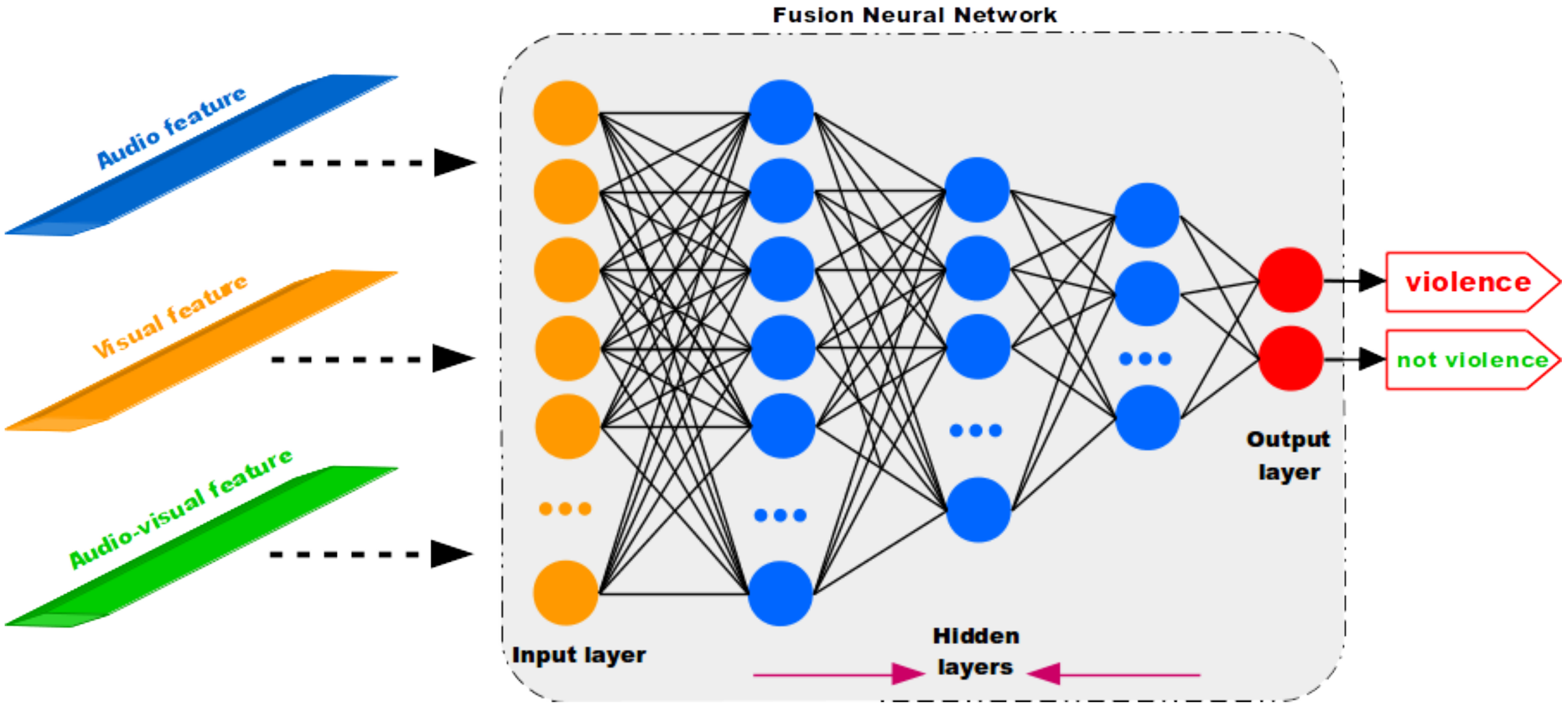
Fight



PIPELINE


Source domain
Target domain
Audio source
Audio target
Visual target
Visual source
Audio feature extraction & temporal statistic computation
Fusion NN
Violence
not Violence
fully-connected layers
concatenation

FUSION NETWORK



DATASET – MEDIAEVAL 2013 VSD TASK

Hollywood movies

Training set: 17 movies, 2013 min.

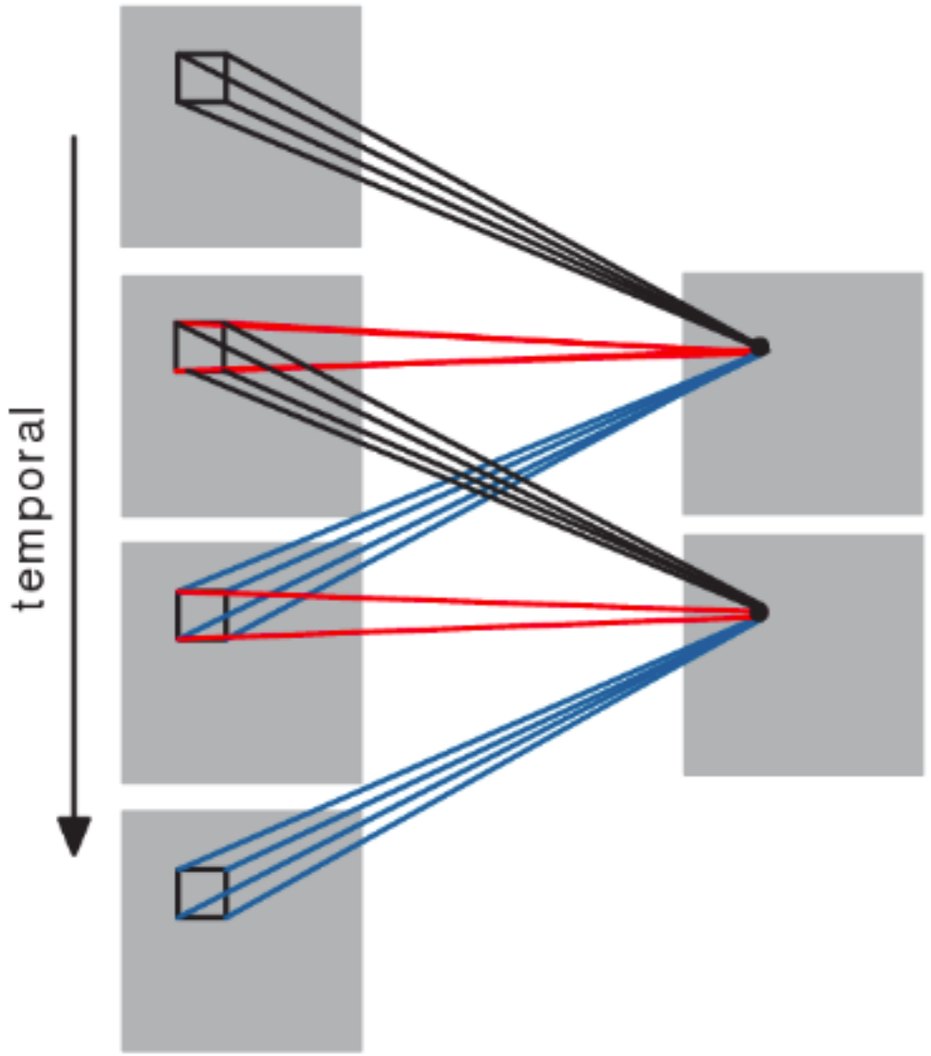
Test set: 7 movies, 885 min.

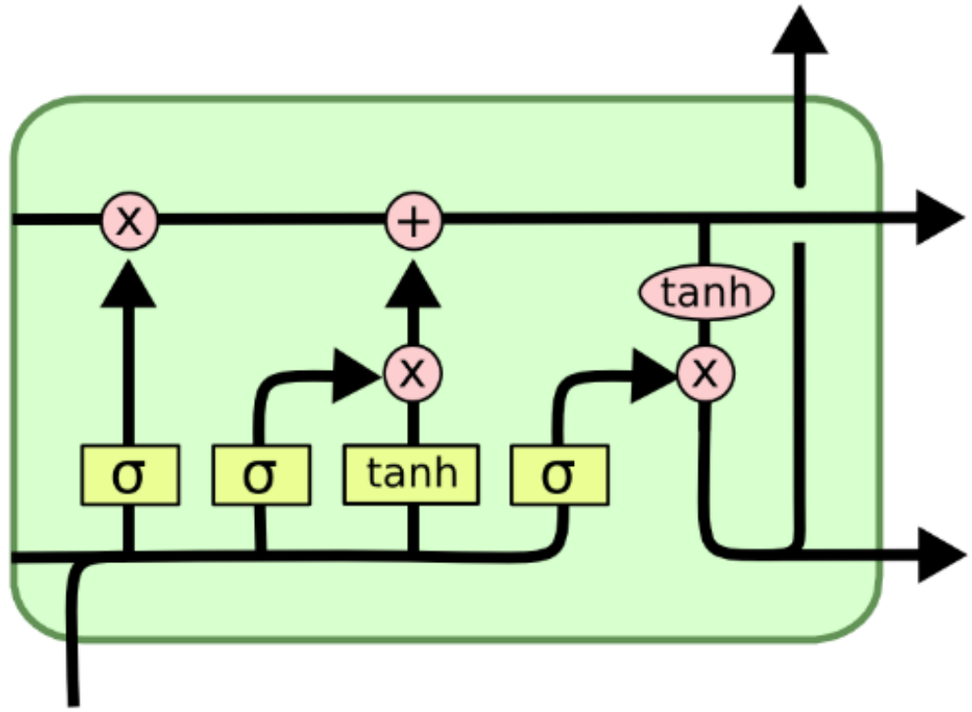
RESULTS – VISUAL INPUT TRANSFORMATIONS

Concept	Raw Frames	TRoF Combinations			Optical Flow Inputs	
		Central	Extremities	Average	Flow	Acceleration
Blood	**74.2**	73.8	70.6	69.9	68.3	58.2
Cold Arms	**81.6**	64.4	71.5	70.8	61.9	76.5
Explosions	**79.4**	71.3	75.9	74.2	77.8	70.6
Fights	73.1	70.4	73.2	71.5	**76.8**	74.3
Fire	70.1	70.7	70.6	69.9	68.1	**71.2**
Firearms	60.8	58.4	58.5	59.1	62.3	**66.8**
Gunshots	69.3	66.8	66.4	64.0	63.6	**73.1**
Violence	66.7	**68.4**	62.8	65.3	65.0	58.7
Concatenation	72.4	70.5	73.3	**73.6**	67.9	72.1
Fusion Network	**74.4**	74.1	73.8	74.2	68.2	72.8

Table 3 – Accuracy percentages for each input transformation on every sub-concept.

TIME-BASED ARCHITECTURES





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

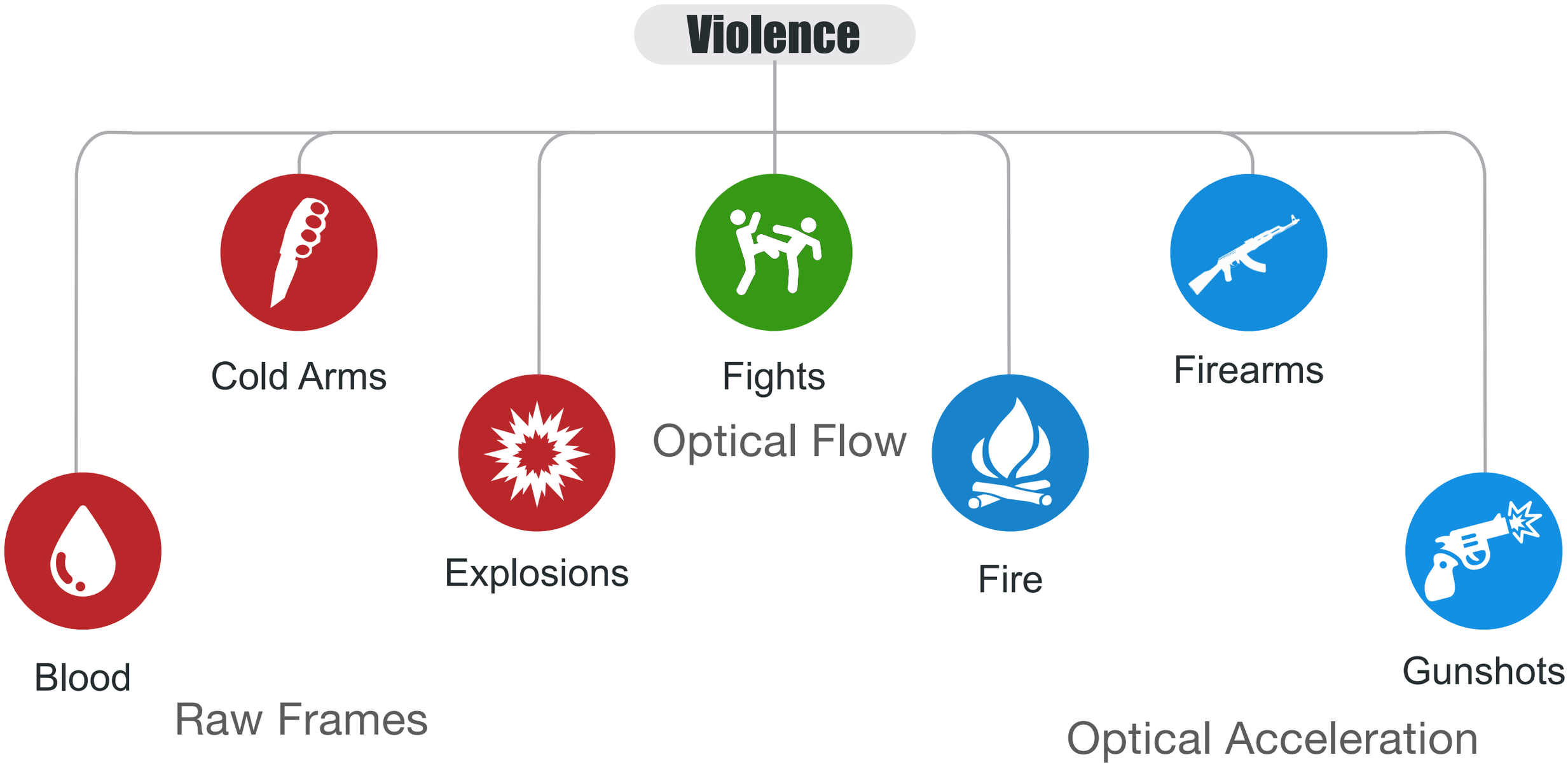
Source: Colah’s Blog. “Understanding LSTM Networks” (2015)

RESULTS – TIME-BASED ARCHITECTURES

Concept	Raw Frames		Optical Flow		Optical Acceleration		Central TRoF	
	C3D	CNN-LSTM	C3D	CNN-LSTM	C3D	CNN-LSTM	C3D	CNN-LSTM
Blood	58.0	57.2	59.2	60.2	**60.2**	58.2	56.8	57.4
Cold Arms	58.3	54.2	66.5	66.2	**75.3**	69.0	63.5	64.0
Explosions	**77.1**	61.4	66.4	64.	73.0	68.1	69.4	69.0
Fights	**70.5**	53.7	68.0	66.9	65.4	61.7	68.2	66.3
Fire	60.2	55.6	60.3	61.3	**64.9**	61.9	62.7	63.6
Firearms	61.0	60.3	63.2	65.0	**66.5**	62.3	62.0	63.4
Gunshots	65.3	56.8	62.6	64.1	**68.6**	66.8	63.8	64.5
Violence	62.3	55.9	58.1	58.6	**68.3**	63.6	62.3	60.1
Fusion	67.3	63.3	67.2	64.8	**69.2**	64.2	66.8	65.2

Table 4 – C3D and CNN-LSTM accuracy percentages with different input transformations.

RESULTS – BEST INPUTS FOR EACH CONCEPT



RESULTS – COMBINED VISUAL AND AUDIO FEATURES

Concept	Best Visual Features	Audio Features	Visual + Audio Features
Blood	**74.2**	61.0	66.5
Cold Arms	81.6	66.9	**83.2**
Explosions	**79.4**	65.3	77.3
Fights	76.8	61.9	**77.2**
Fire	**71.2**	67.8	64.5
Firearms	66.8	62.4	**73.3**
Gunshots	73.1	70.7	**74.5**
Violence	68.4	**72.8**	72.1
Fusion	75.3	63.0	**78.5**

Table 5 – Accuracy percentages of the best visual features combined with audio features.

SPECIALIZED DATASET – NTU-CCTV FIGHTS

1000 Videos

18 hours

CCTV and mobile cameras

Miguel Oliveira

RESULTS – SPECIALIZED DATASET

	mAP
Two-Stream	**0.795**
C3D	0.645
TRoF	0.692
Fights Detector	0.623
Fusion	0.652

Table 6 – Pre-trained network tested in specialized dataset

RESULTS – SPECIALIZED DATASET

	Original Training		Specialized Training	
	Acc. (%)	mAP	Acc. (%)	mAP
Fights	77.2	-	**78.8**	-
Fusion Network	78.5	0.656	**79.6**	**0.661**

Table 7 – Training only the fights detector network with the specialized dataset and tested on MediaEval 2013

PUBLISHED WORKS

Breaking down violence: A deep-learning strategy to model and classify violence in videos.

B. Peixoto, S. Avila, Z. Dias, and A. Rocha. 2018

In Proceedings of the 13th International Conference on Availability, Reliability and Security (ARES 2018).

Concepts and input manipulation results

– Ivan Alvarado/Reuters

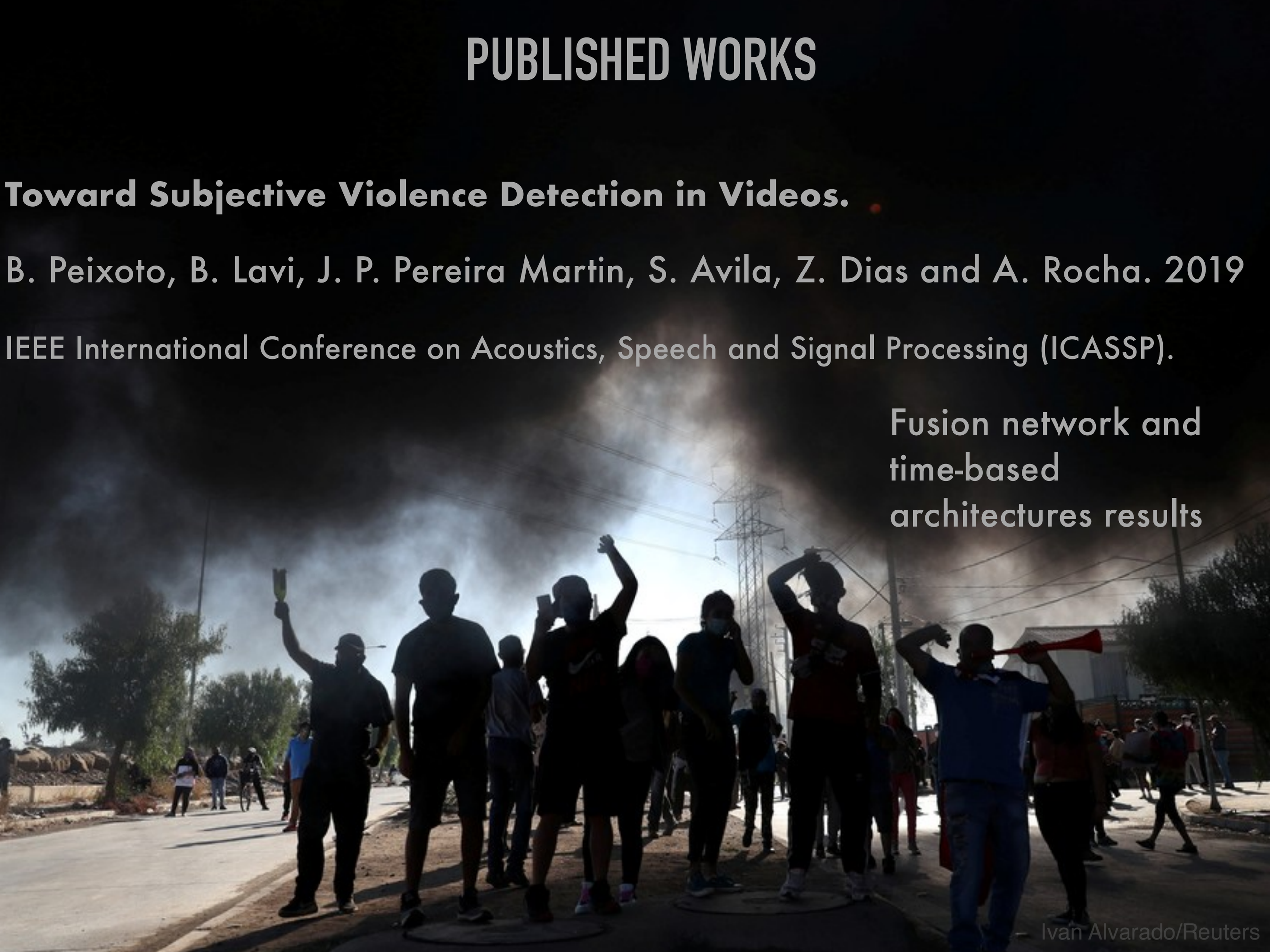
PUBLISHED WORKS

Toward Subjective Violence Detection in Videos.

B. Peixoto, B. Lavi, J. P. Pereira Martin, S. Avila, Z. Dias and A. Rocha. 2019

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Fusion network and time-based architectures results



– Ivan Alvarado/Reuters

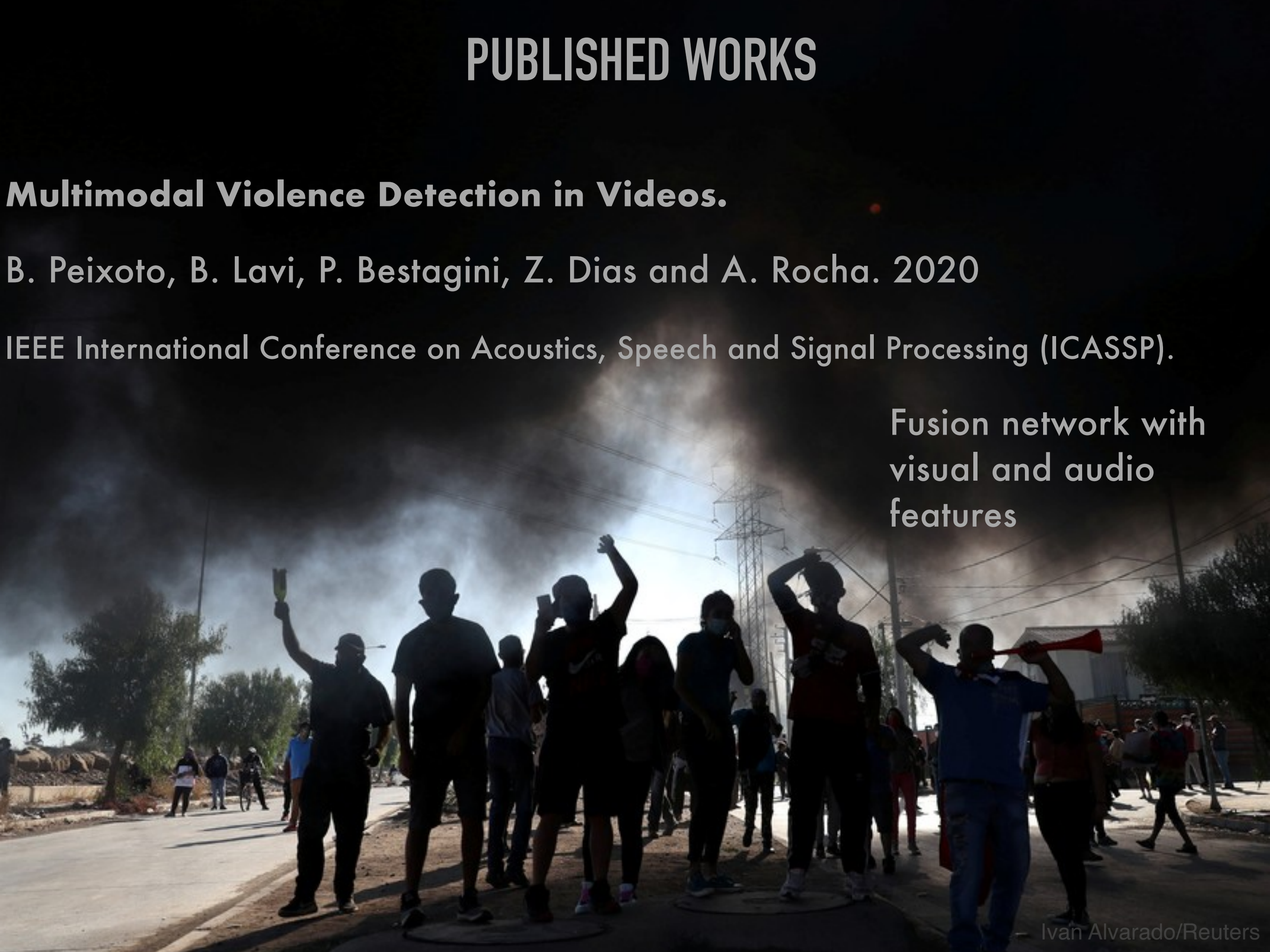
PUBLISHED WORKS

Multimodal Violence Detection in Videos.

B. Peixoto, B. Lavi, P. Bestagini, Z. Dias and A. Rocha. 2020

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Fusion network with visual and audio features



– Ivan Alvarado/Reuters

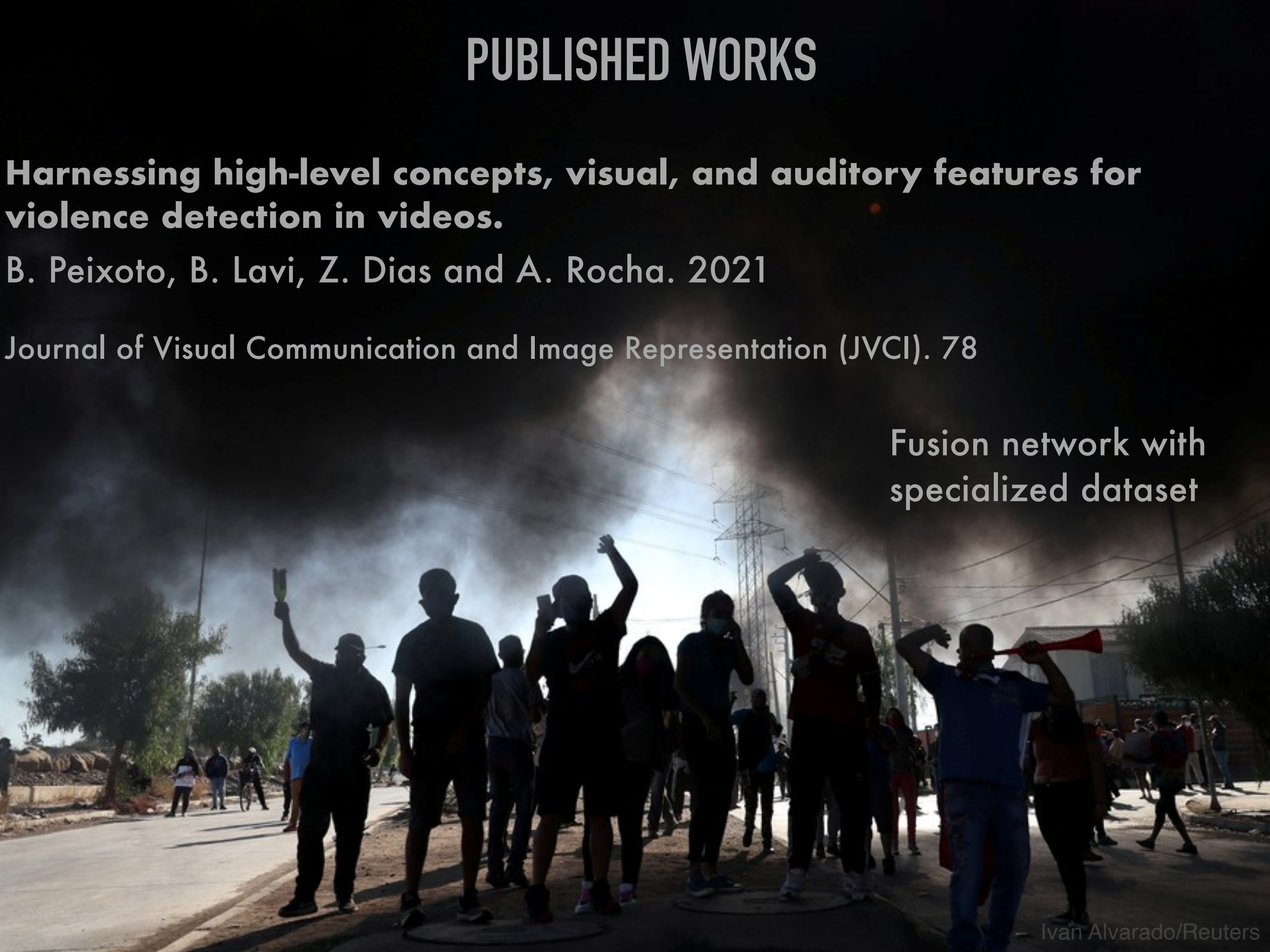
PUBLISHED WORKS

Harnessing high-level concepts, visual, and auditory features for violence detection in videos.

B. Peixoto, B. Lavi, Z. Dias and A. Rocha. 2021

Journal of Visual Communication and Image Representation (JVCI). 78

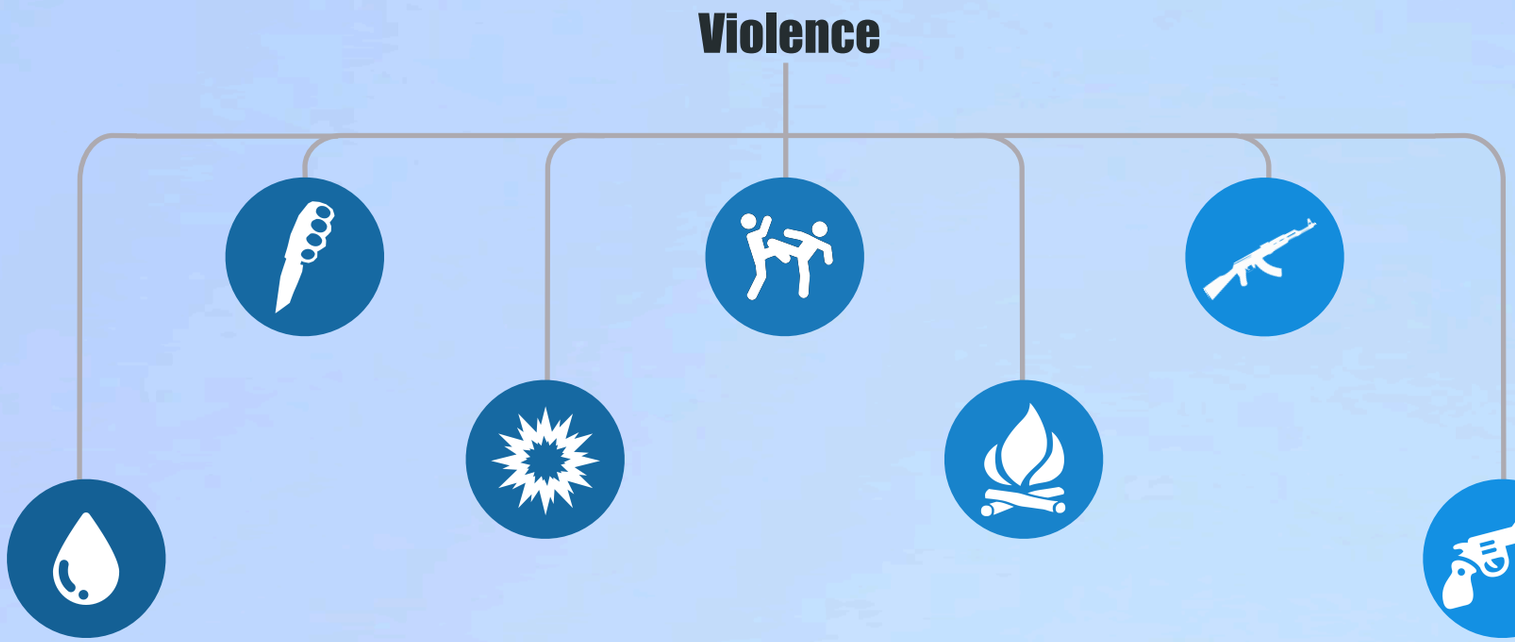
Fusion network with specialized dataset



– Ivan Alvarado/Reuters

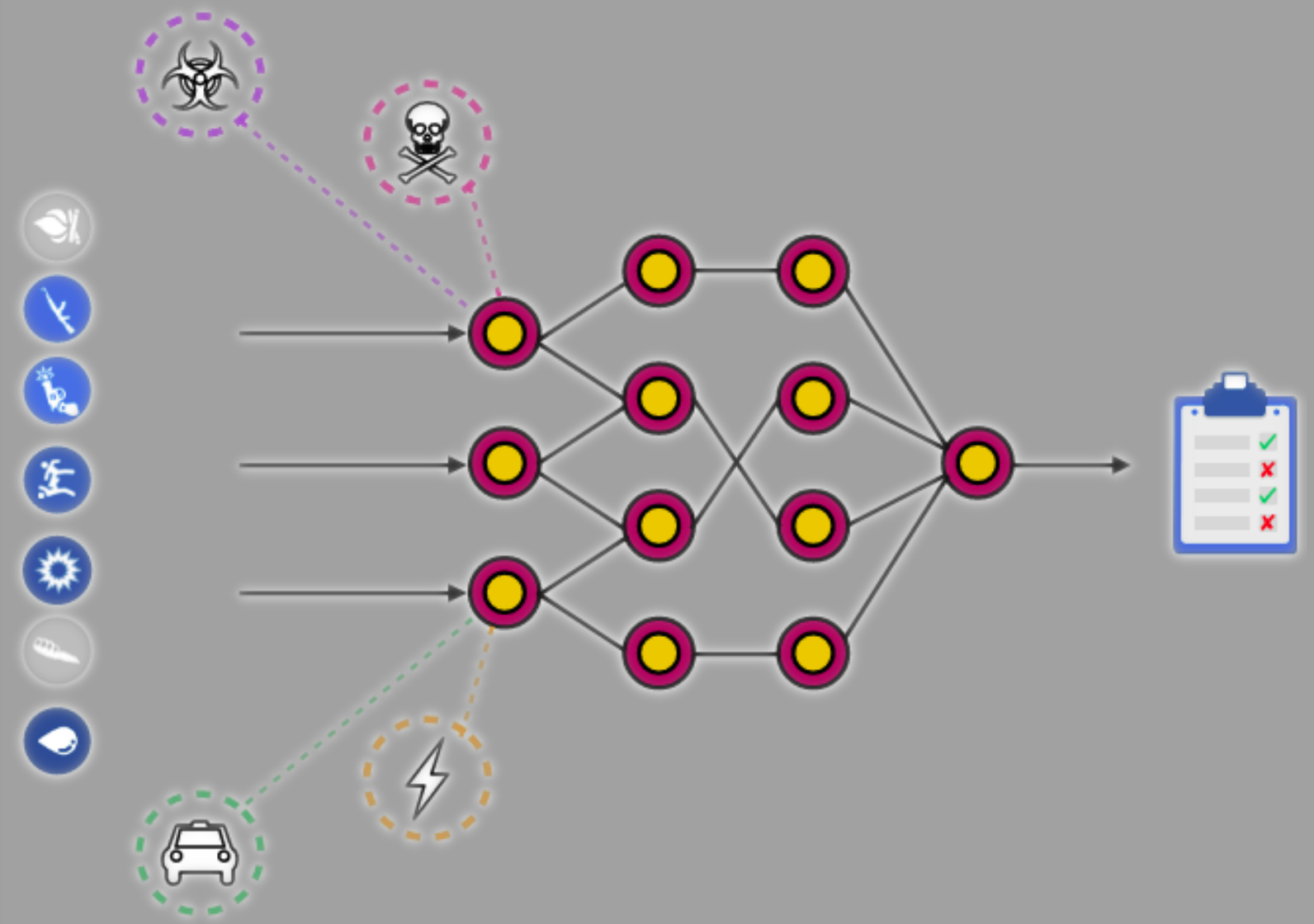
CONCLUSION

How to define violence in terms that a computer can understand?



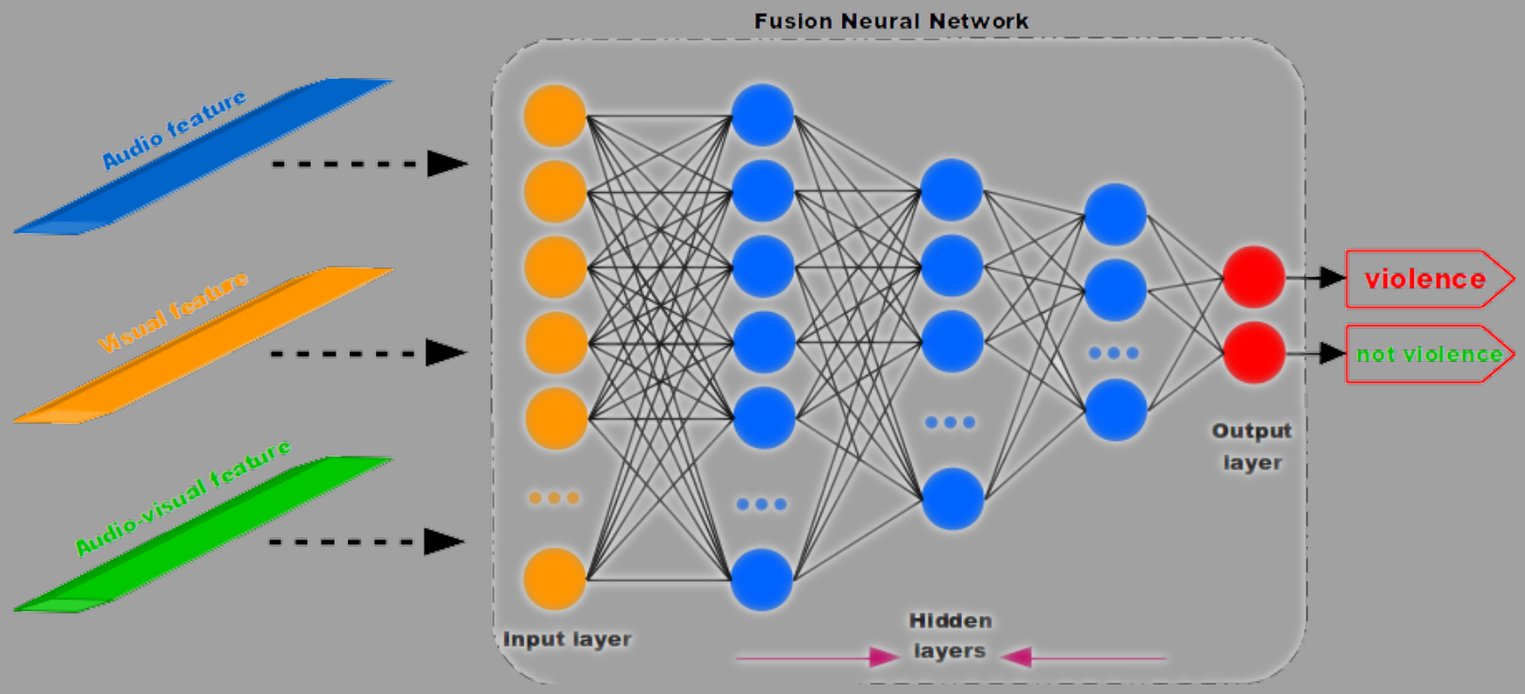
CONCLUSION

How different kinds of violence interact to enable a system to understand the meaning of violence in general?



CONCLUSION

How to build a system that relies upon visual and sound signals characteristics to detect violence in the general sense?



FUTURE WORK

Specialized datasets for each desired sub-concept.

Self-supervised learning to deal with non-labeled data.

UNICAMP

CAPES

CNPq

Instituto de Computação

UNIVERSIDADE ESTADUAL DE CAMPINAS

recod.ai

Tiago Viana

Thank You!

RESULTS – EXAMPLES

Firearms

(a)

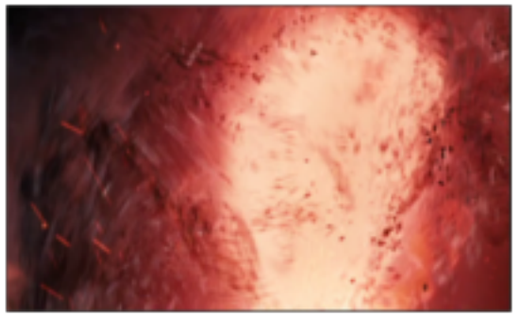
(b)

(c)

Explosions

(d)

(e)



(f)

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

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

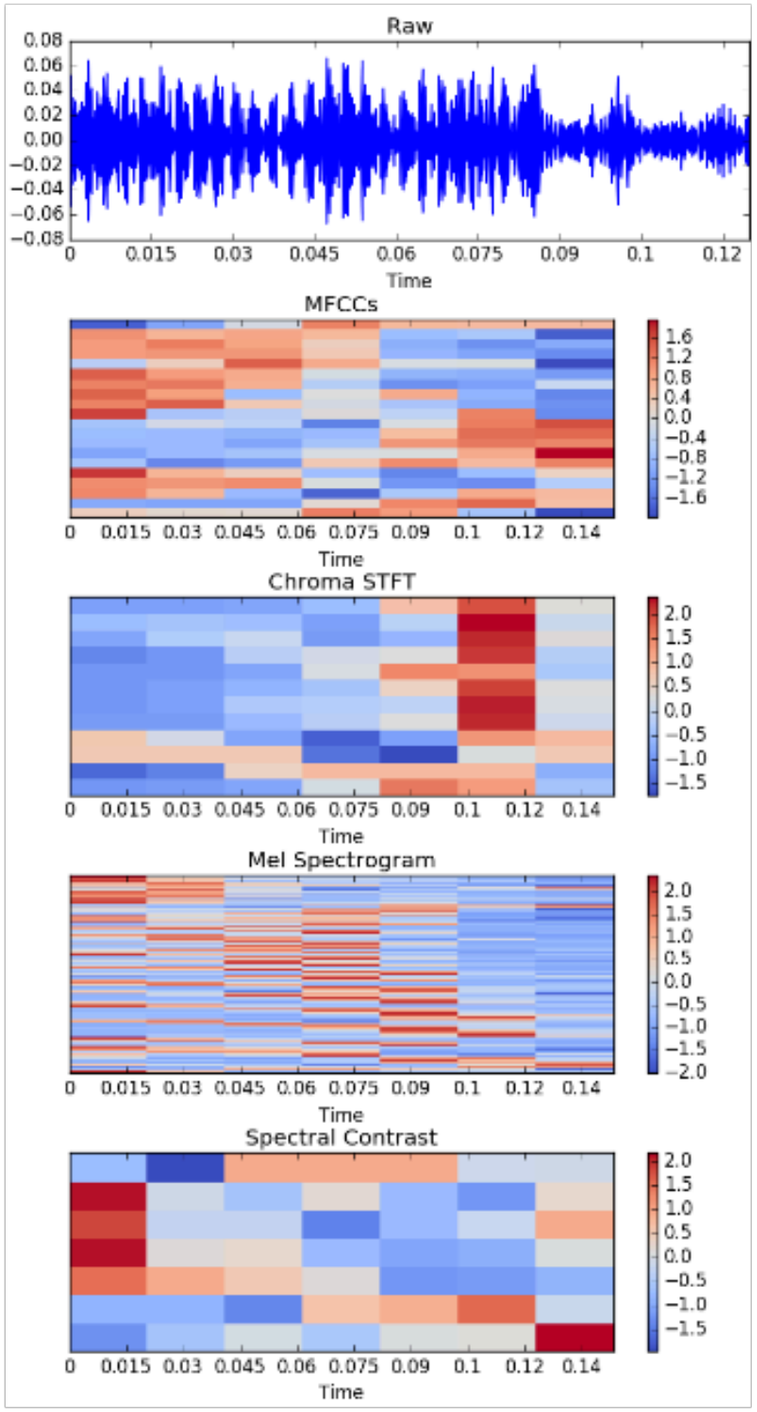
PRESENCE OF INDIVIDUAL CONCEPTS IN VSD 2013 DATASET

	Blood	Cold Arms	Explosions	Fights	Fire	Firearms	Gunshots
Armageddon	0.86%	0.04%	1.61%	3.07%	9.66%	4.02%	0.09%
Billy Elliot	0.24%	1.88%	0.00%	2.00%	1.08%	0.00%	0.00%
Dead Poets Society	0.36%	0.84%	0.00%	0.31%	3.39%	0.51%	0.00%
Eragon	5.27%	13.94%	0.45%	10.91%	22.00%	0.00%	0.00%
Fight Club	8.20%	0.20%	0.26%	4.59%	2.71%	5.37%	0.08%
Harry Potter 5	4.97%	2.80%	2.14%	5.73%	16.96%	0.00%	0.00%
I Am Legend	6.43%	2.45%	0.35%	4.08%	1.45%	9.47%	0.53%
Independence Day	0.52%	0.89%	4.13%	1.67%	12.75%	8.87%	2.34%
Leon	8.17%	1.14%	0.15%	1.16%	0.51%	13.71%	0.92%
Midnight Express	2.08%	0.45%	0.00%	5.41%	3.82%	7.10%	0.25%
Pirates of the Caribbean	0.94%	0.01%	1.16%	13.67%	26.15%	29.19%	3.01%
Reservoir Dogs	37.23%	1.89%	0.00%	4.12%	0.22%	19.31%	0.78%
Saving Private Ryan	27.31%	23.75%	16.45%	13.78%	14.85%	68.65%	33.23%
The Bourne Identity	3.66%	2.52%	0.09%	2.90%	0.48%	6.86%	0.47%
The Sixth Sense	1.08%	4.85%	0.00%	0.14%	2.08%	0.92%	0.04%
The Wicker Man	0.47%	1.26%	0.14%	0.30%	3.30%	4.32%	0.18%
The Wizard of Oz	0.00%	32.88%	1.08%	1.20%	6.65%	7.34%	0.00%

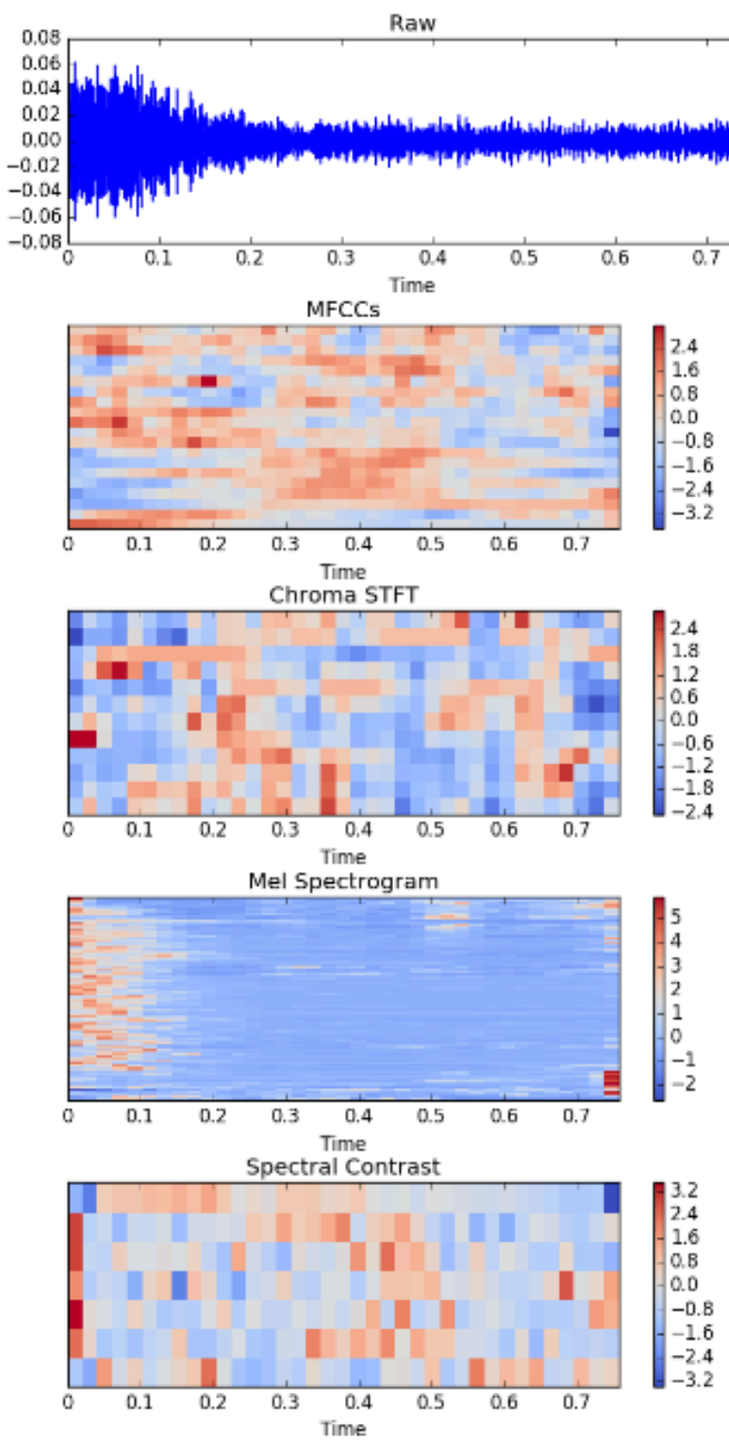
Percentage of positive samples for each sub-concept in each movie

INCORPORATING AUDIO INFORMATION

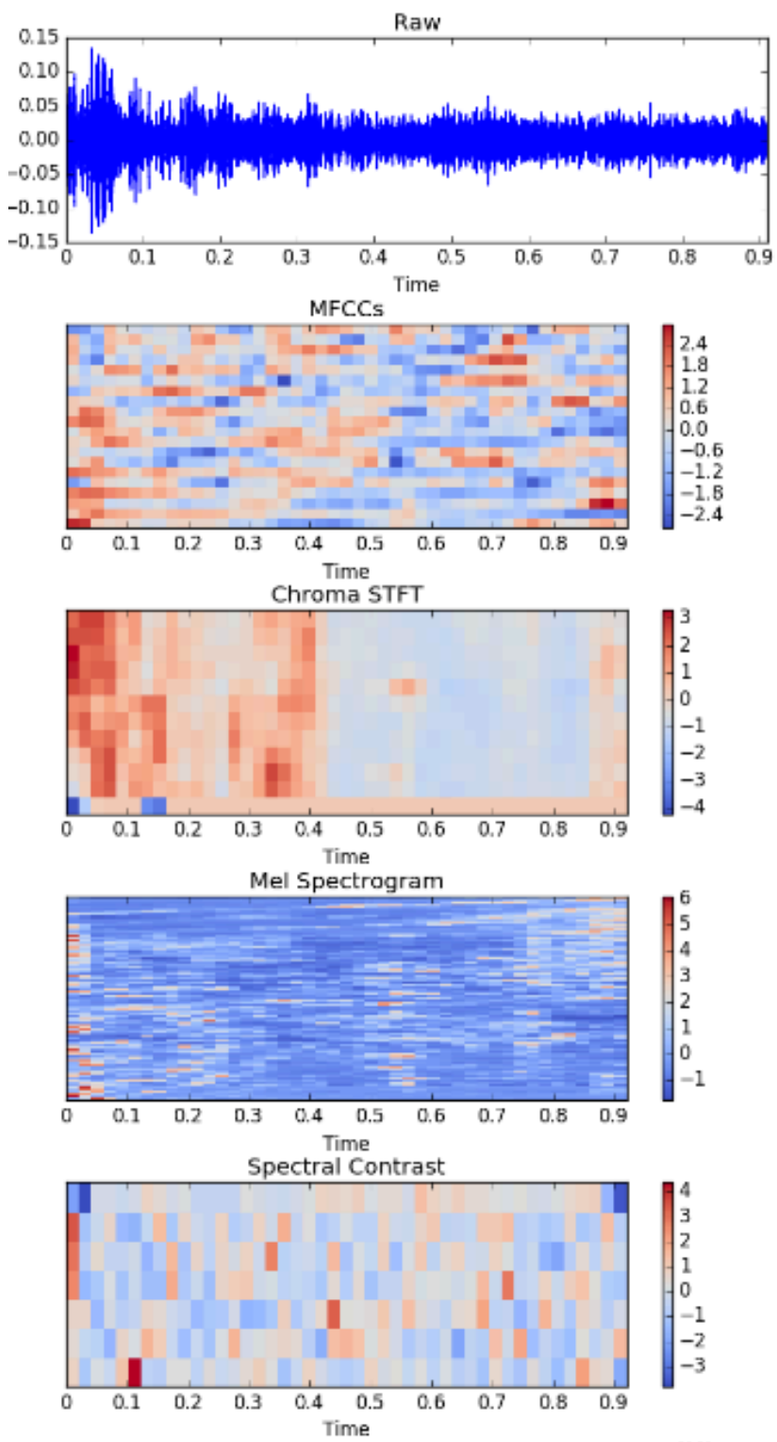
Gunshot



Explosion

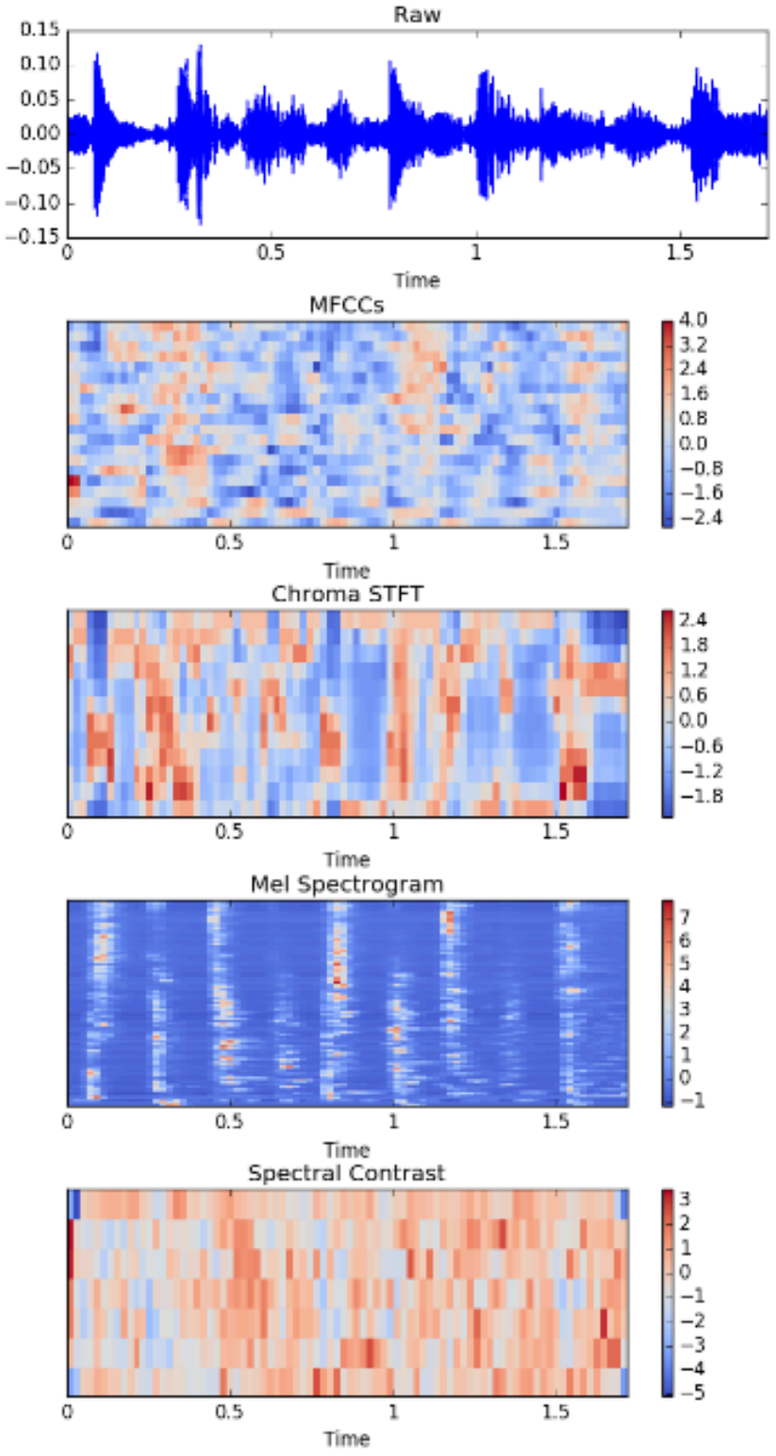


Fight

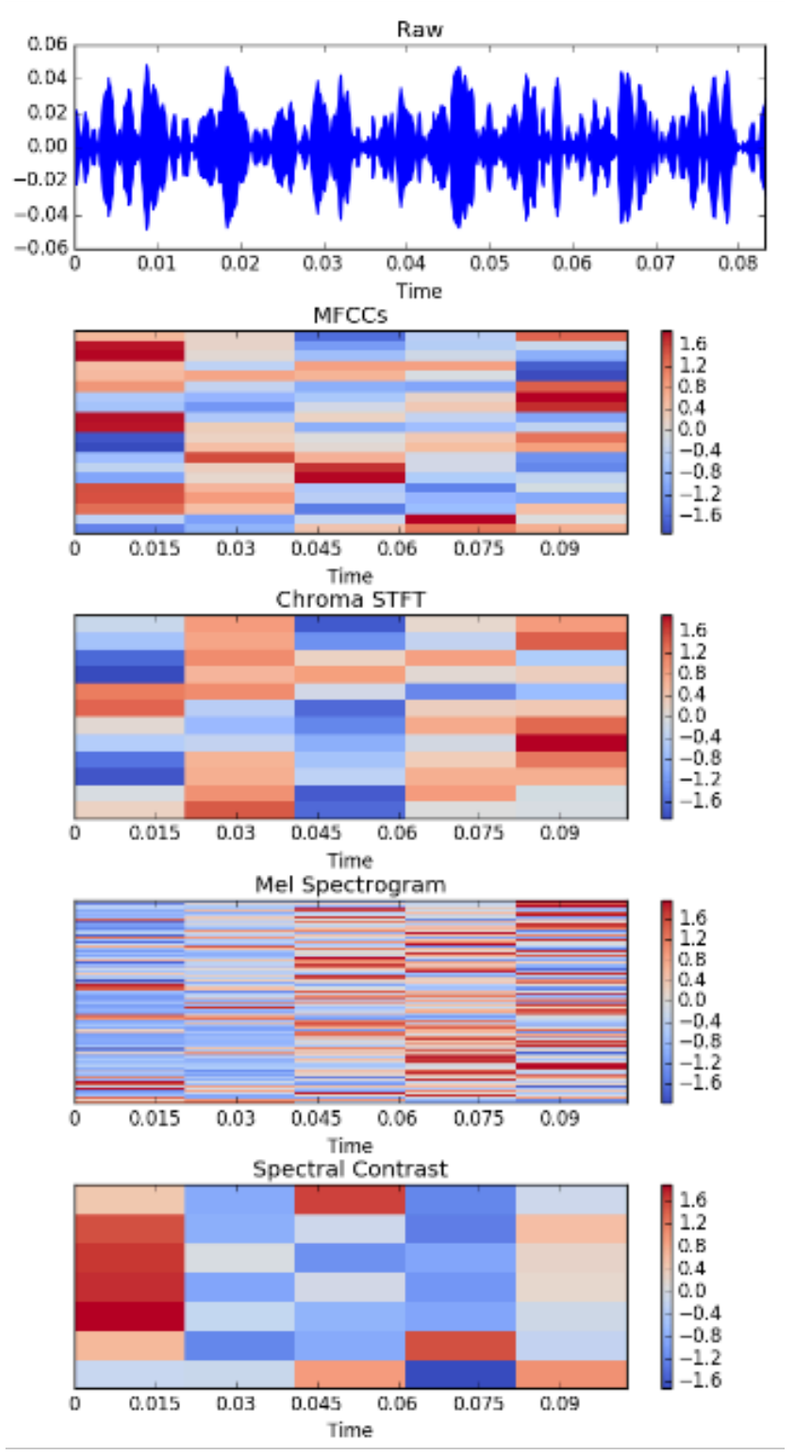


INCORPORATING AUDIO INFORMATION

Cold arms

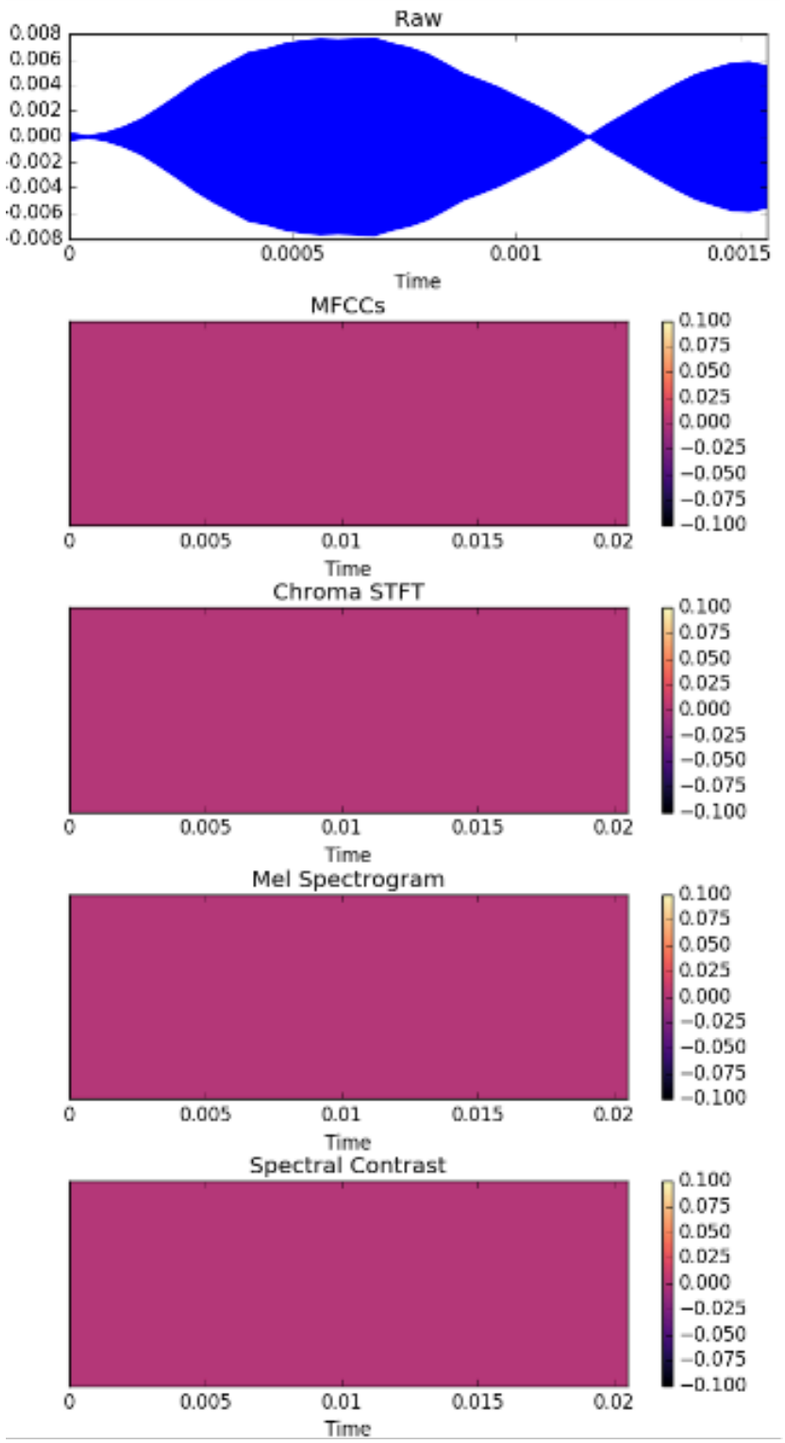


Firearm

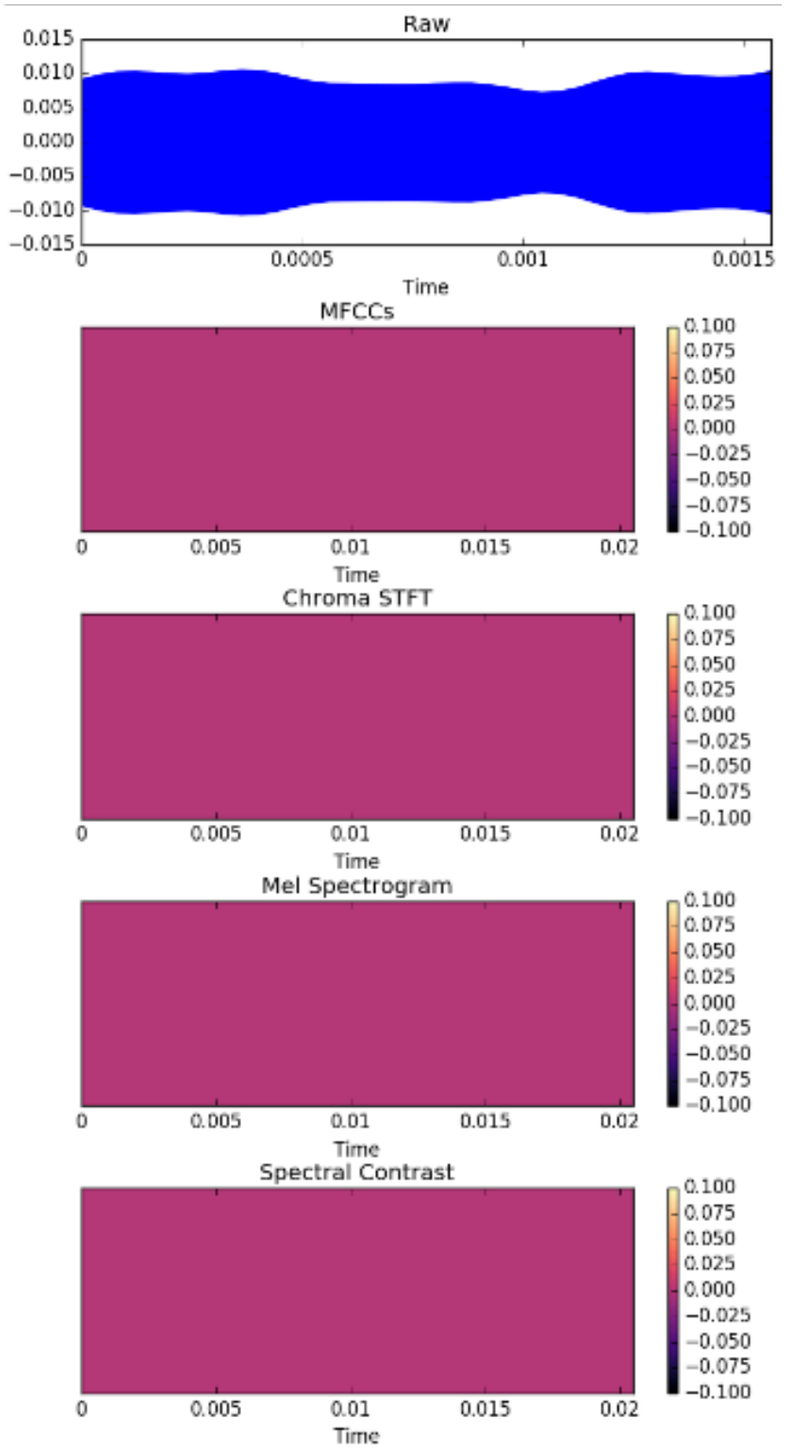


INCORPORATING AUDIO INFORMATION

Blood



Fire



Why Use Deep Learning?