Computer Vision Applications

COMP 388-002/488-002 Computer Science Topics

Daniel Moreira

Fall 2022

LOYOLA UNIVERSITY CHICAGO

Sensitive Video Analysis

COMP 388-002/488-002 Computer Science Topics

Daniel Moreira

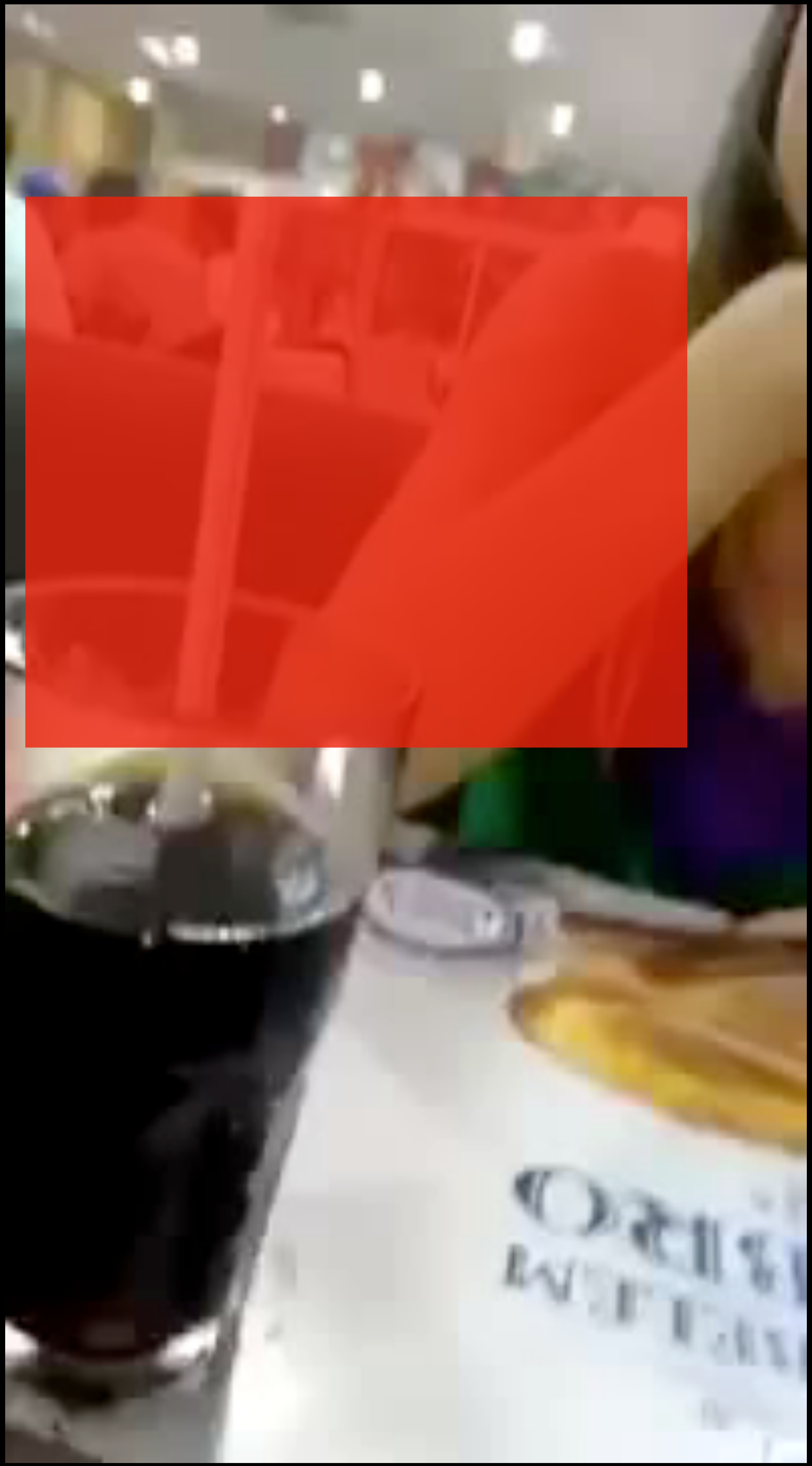
Fall 2022

LOYOLA
UNIVERSITY CHICAGO

Sensitive Video

"Motion pictures whose content may inflict harm (*e.g.*, trauma, shock, or fear) to particular audiences (*e.g.*, children or unwary spectators), due to the inappropriateness of content."


DailyMail.com
Justin Beiber getting his ass kicked!!




Why do we care?

Challenges

Big Data

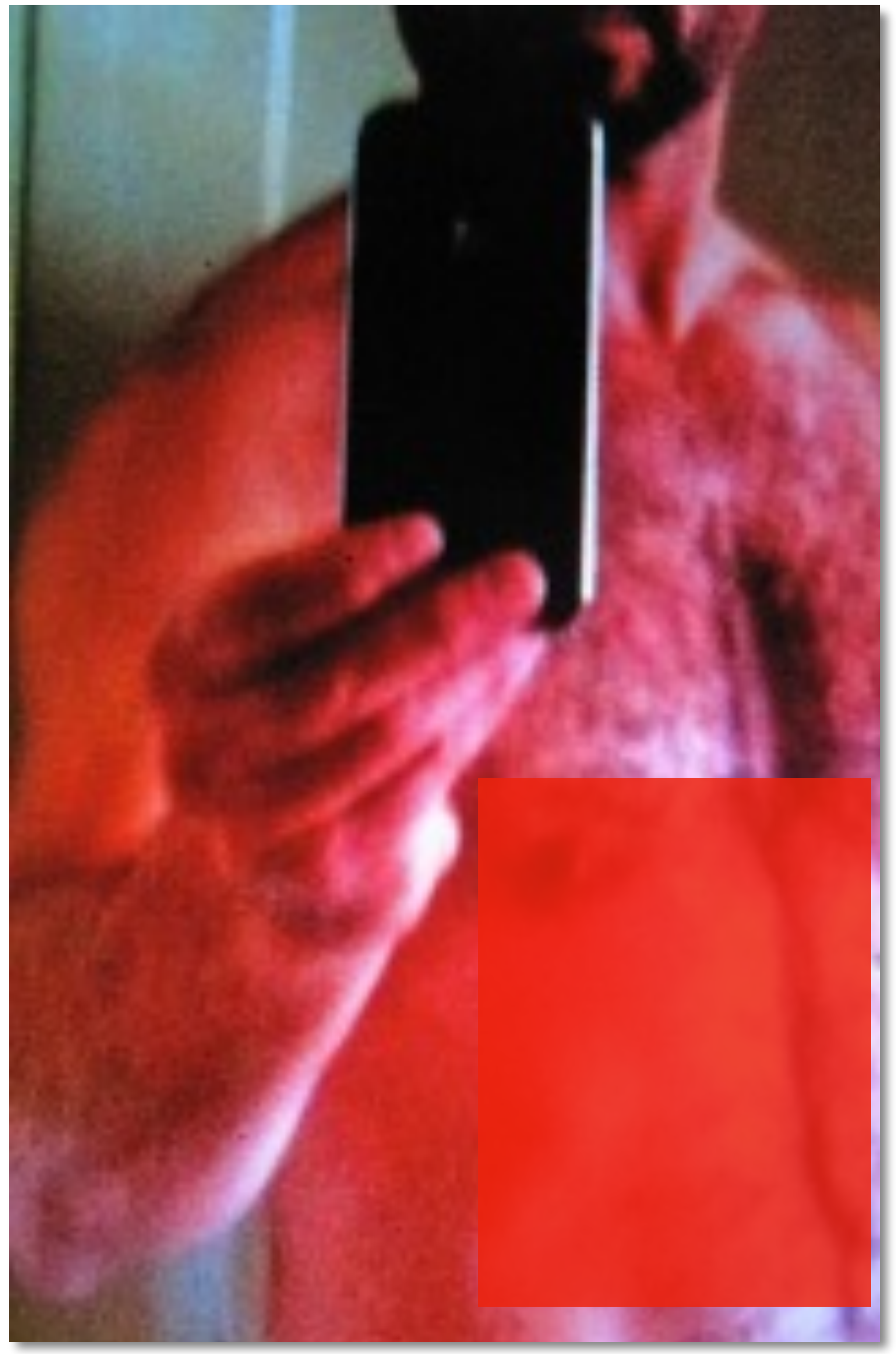
Challenges

Subjectivity

LOYOLA UNIVERSITY CHICAGO

Challenges

Pervasiveness



LOYOLA UNIVERSITY CHICAGO

Challenges

Urgency

LOYOLA UNIVERSITY CHICAGO

Tasks

Part I: Sensitive Video Classification

Part II: Sensitive Video Detection

Tasks

Part I: Sensitive Video Classification

Part II: Sensitive Video Detection

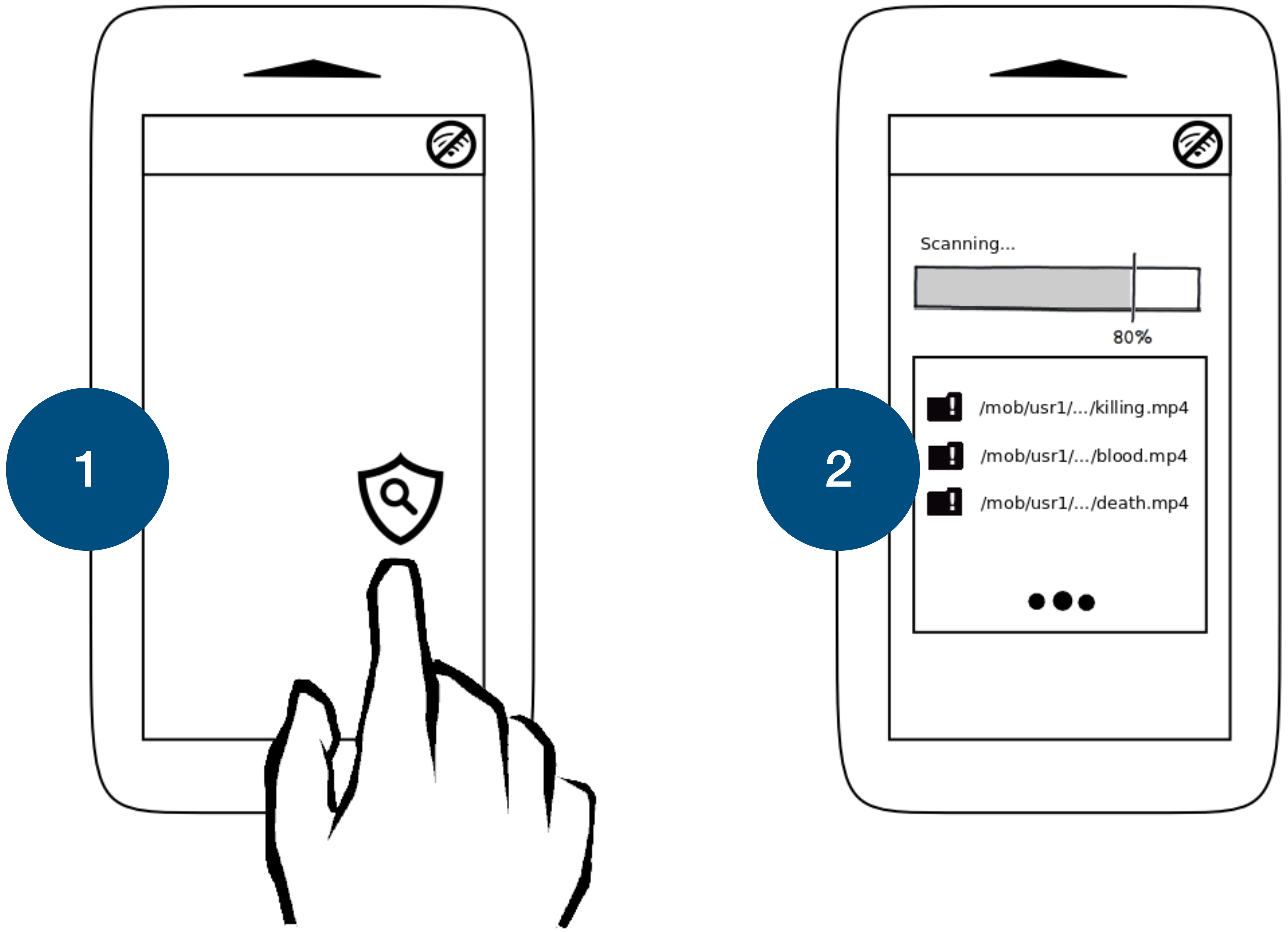
LOYOLA UNIVERSITY CHICAGO

Sensitive Video Classification

LOYOLA
UNIVERSITY CHICAGO

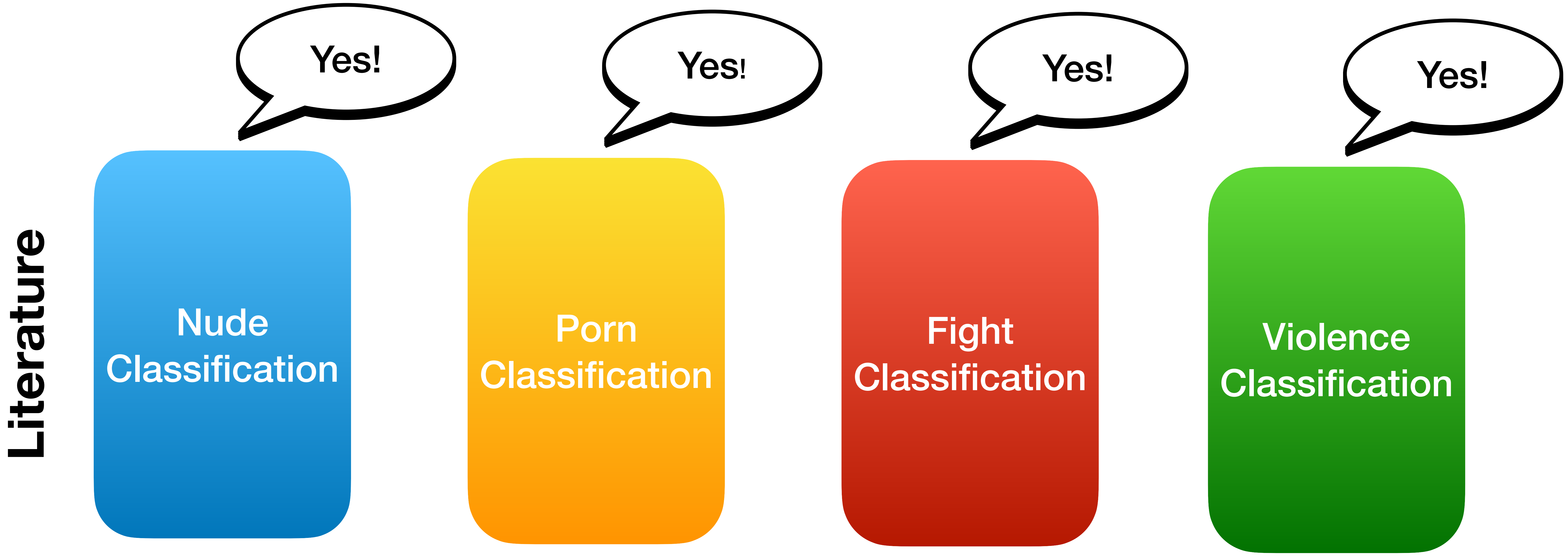
Task

Can a computer decide if a video is either sensitive or non-sensitive?



State of the Art

Can a computer decide if a video is either sensitive or non-sensitive?



LOYOLA UNIVERSITY CHICAGO

State of the Art

Literature

Nude Classification — Yes!

Porn Classification — Yes! — Motion is mandatory!

Fight Classification — Yes!

Violence Classification — Yes! — Motion is mandatory!

LOYOLA UNIVERSITY CHICAGO

Sponsor's Challenge

Will it run on mobile devices?

Motion is mandatory!

Motion is mandatory!

Yes!

Yes!

Yes!

Yes!

Literature

Nude Classification

Porn Classification

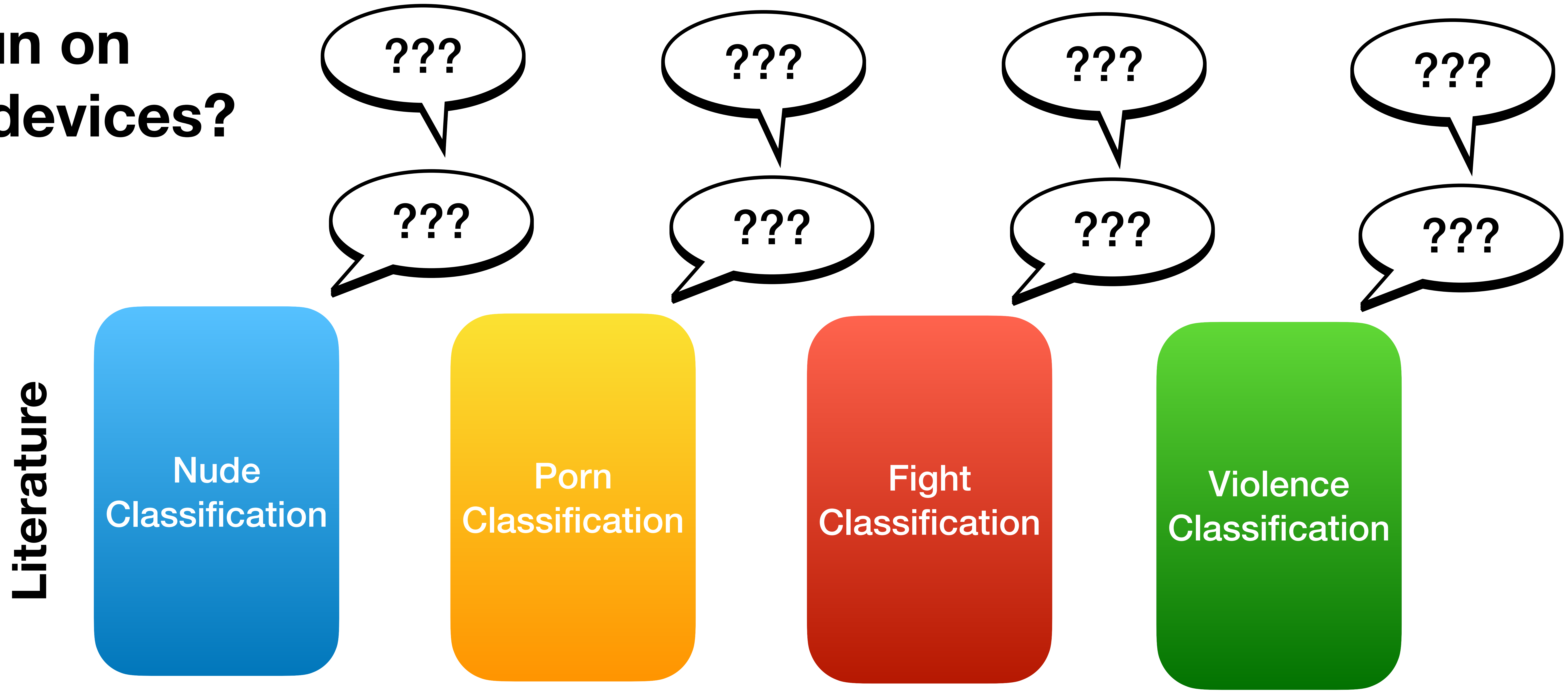
Fight Classification

Violence Classification

LOYOLA UNIVERSITY CHICAGO

Sponsor’s Challenge

Will it run on mobile devices?



LOYOLA UNIVERSITY CHICAGO

Sponsor's Challenge

Will it run on mobile devices?

Effectiveness

Motion is mandatory.
Spatiotemporal description takes time.

VS

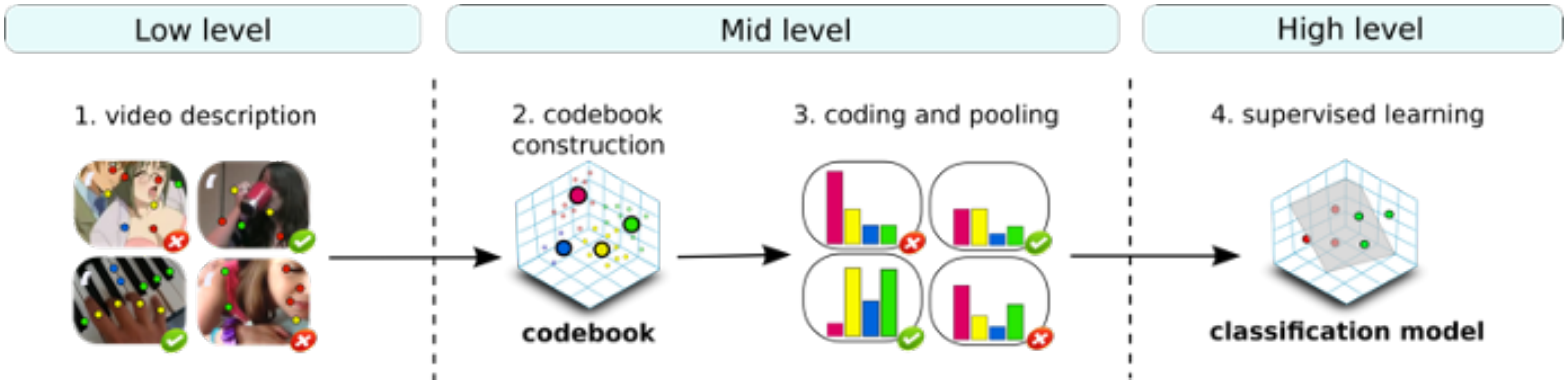
Efficiency

Small runtime.
Low-memory footprint.

LOYOLA UNIVERSITY CHICAGO

Proposed Solution

Based on Bags of Visual Words that (BoVW)



Proposed Solution

Based on Bags of Visual Words that (BoVW)

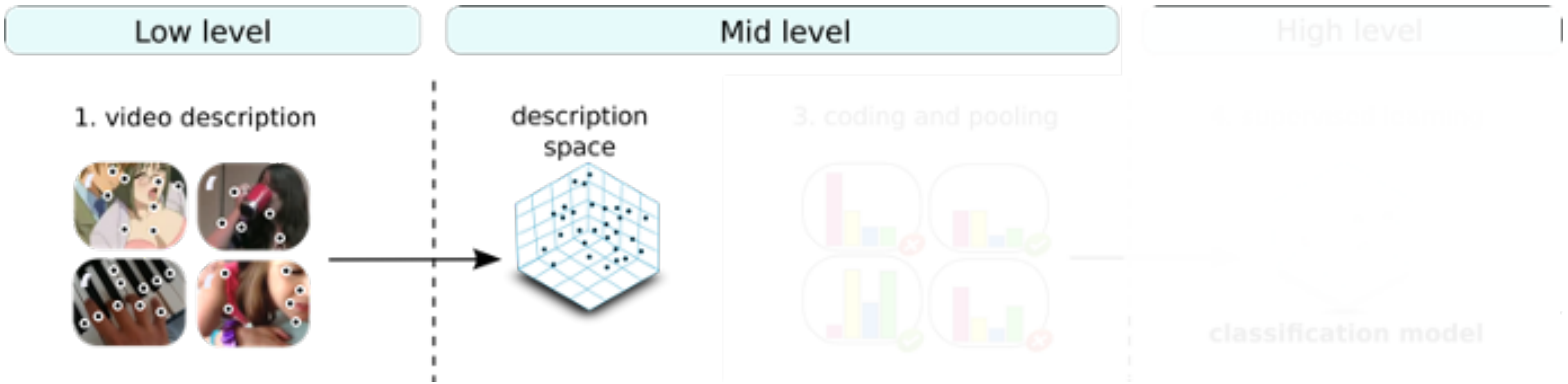
Proposed Solution

Based on Bags of Visual Words that (BoVW)

LOYOLA UNIVERSITY CHICAGO

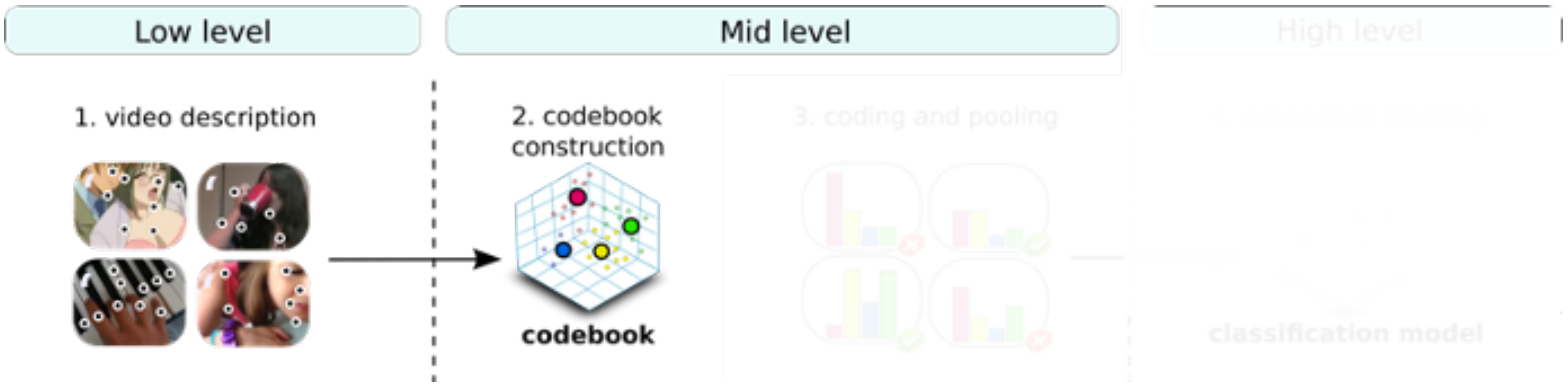
Proposed Solution

Based on Bags of Visual Words that (BoVW)



Proposed Solution

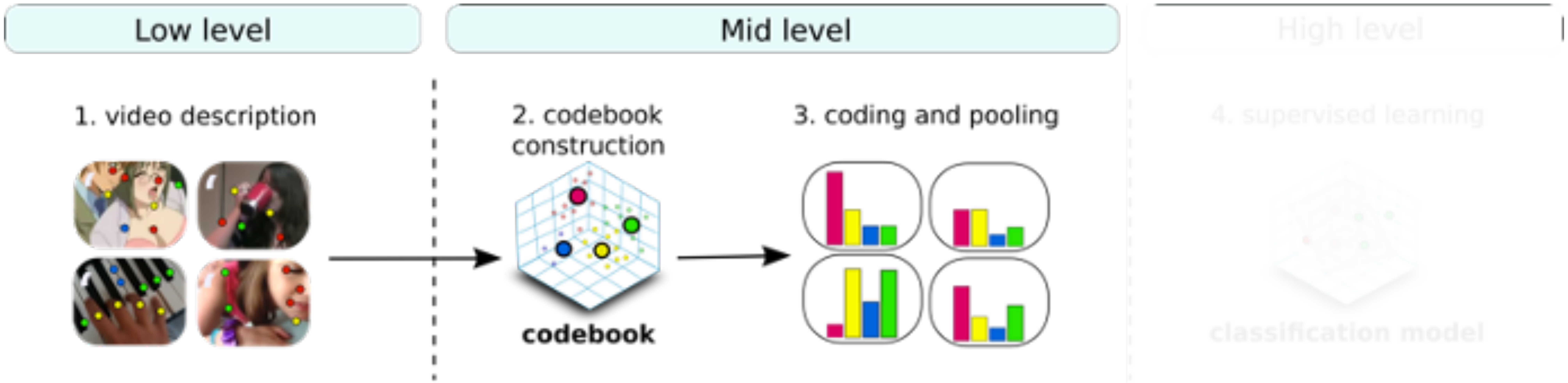
Based on Bags of Visual Words that (BoVW)



LOYOLA UNIVERSITY CHICAGO

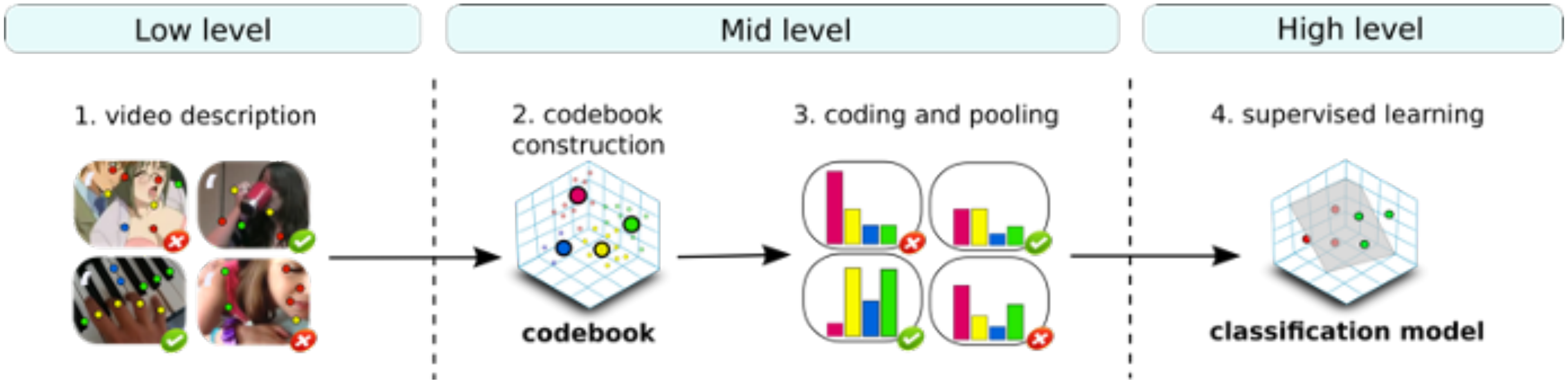
Proposed Solution

Based on Bags of Visual Words that (BoVW)



Proposed Solution

Based on Bags of Visual Words that (BoVW)



LOYOLA UNIVERSITY CHICAGO

Proposed Solution

Based on Bags of Visual Words that (BoVW)

Low level	Mid level	High level
Resolution reduction	GMM-based codebook	
Space-temporal local features	Coding: Fisher Vectors	Linear SVM
PCA Reduction	Pooling: single vector for entire footage	

LOYOLA UNIVERSITY CHICAGO

Proposed Solution

Based on Bags of Visual Words that (BoVW)

Low level	Mid level	High level
Resolution reduction	GMM-based codebook	
Space-temporal local features	Coding: Fisher Vectors	Linear SVM
PCA Reduction	Pooling: single vector for entire footage	

LOYOLA UNIVERSITY CHICAGO

Temporal Robust Features (TRoF)

Effectiveness

Motion is mandatory.
Spatiotemporal description takes time.

VS

Efficiency

Small runtime.
Low-memory footprint.

LOYOLA UNIVERSITY CHICAGO

Temporal Robust Features (TRoF)

RECAP

Inspiration on Speeded-Up Robust Features (SURF)
Hessian Matrix

Given an image pixel $I(x, y)$, a scale of interest σ,

and Gaussian second order derivative functions $\frac{\delta^2}{\delta x^2}G(\sigma)$, $\frac{\delta^2}{\delta y^2}G(\sigma)$, and $\frac{\delta^2}{\delta xy}g(\sigma)$,

the Hessian matrix H is given by:

$$H(x, y, \sigma) = \begin{bmatrix} \frac{\delta^2}{\delta x^2}g(\sigma) * I(x, y) & \frac{\delta^2}{\delta xy}g(\sigma) * I(x, y) \\ \frac{\delta^2}{\delta xy}g(\sigma) * I(x, y) & \frac{\delta^2}{\delta y^2}g(\sigma) * I(x, y) \end{bmatrix}$$

LOYOLA UNIVERSITY CHICAGO

Temporal Robust Features (TRoF)

RECAP

Inspiration on Speeded-Up Robust Features (SURF)
Hessian Matrix

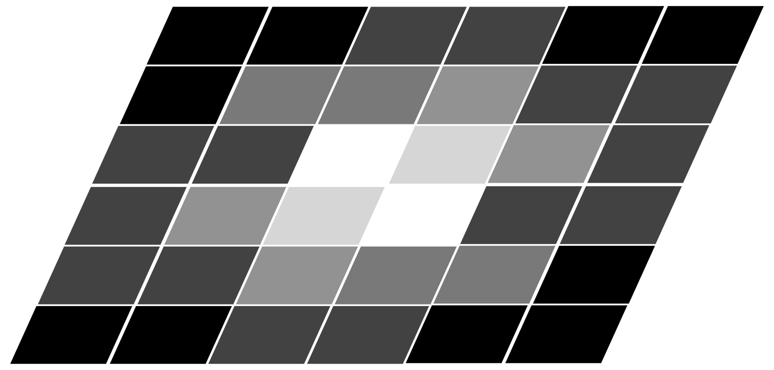
Given an image pixel $I(x, y)$, a scale of interest σ,

and Gaussian second order derivative functions $\frac{\delta^2}{\delta x^2} G(\sigma)$, $\frac{\delta^2}{\delta y^2} G(\sigma)$, and $\frac{\delta^2}{\delta xy} g(\sigma)$,

the Hessian matrix H is given by:

$$H(x, y, \sigma) = \begin{bmatrix} \frac{\delta^2}{\delta x^2} g(\sigma) * I(x, y) & \frac{\delta^2}{\delta xy} g(\sigma) * I(x, y) \\ \frac{\delta^2}{\delta xy} g(\sigma) * I(x, y) & \frac{\delta^2}{\delta y^2} g(\sigma) * I(x, y) \end{bmatrix}$$

Property: blobs with scale σ and centered at $I(x, y)$ will lead to a large $det(H)$.



Take the regions with large $det(H)$ as candidate keypoints.

LOYOLA UNIVERSITY CHICAGO

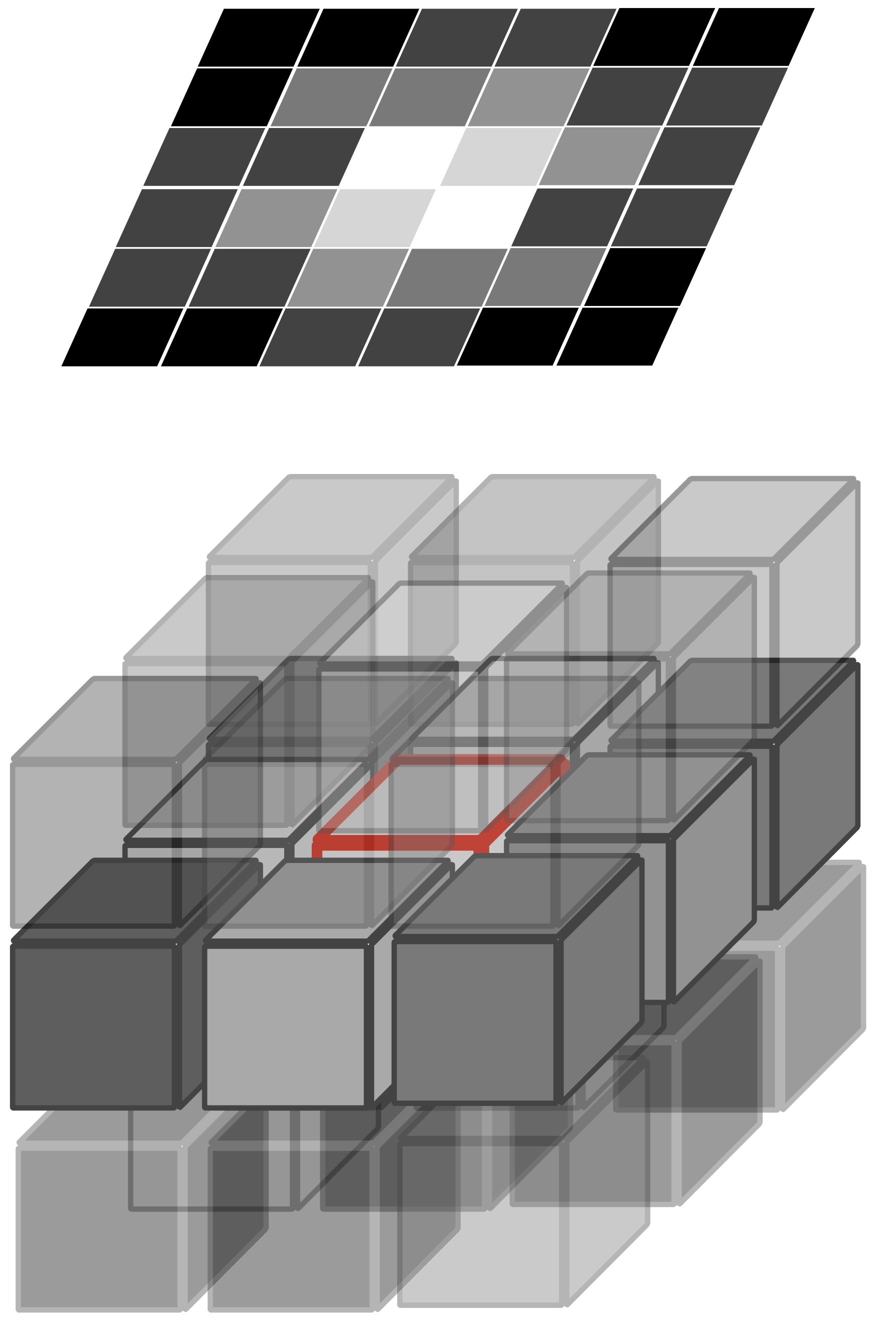
Temporal Robust Features (TRoF)

Inspiration on Speeded-Up Robust Features (SURF)
Statio-temporal Hessian Matrix

Given a **video voxel** $I(x, y, t)$, a scale of interest σ, and Gaussian second order derivative functions
$\frac{\delta^2}{\delta x^2}G(\sigma)$, $\frac{\delta^2}{\delta y^2}G(\sigma)$, $\frac{\delta^2}{\delta t^2}G(\sigma)$, $\frac{\delta^2}{\delta xy}g(\sigma)$, $\frac{\delta^2}{\delta xt}g(\sigma)$, and $\frac{\delta^2}{\delta yt}g(\sigma)$,

the Hessian matrix H is given by:

$$H(x, y, t, \sigma) = \begin{bmatrix} \frac{\delta^2}{\delta x^2}g(\sigma) * I(x, y, t) & \frac{\delta^2}{\delta xy}g(\sigma) * I(x, y, t) & \frac{\delta^2}{\delta xt}g(\sigma) * I(x, y, t) \\ \frac{\delta^2}{\delta xy}g(\sigma) * I(x, y, t) & \frac{\delta^2}{\delta y^2}g(\sigma) * I(x, y, t) & \frac{\delta^2}{\delta yt}g(\sigma) * I(x, y, t) \\ \frac{\delta^2}{\delta xt}g(\sigma) * I(x, y, t) & \frac{\delta^2}{\delta yt}g(\sigma) * I(x, y, t) & \frac{\delta^2}{\delta t^2}g(\sigma) * I(x, y, t) \end{bmatrix}$$



LOYOLA UNIVERSITY CHICAGO

Temporal Robust Features (TRoF)

RECAP

Inspiration on Speeded-Up Robust Features (SURF)
Integral Image

Data structure $I_{\sum}$ computed from a given image I that shares the same resolution (i.e., same number of rows and of columns).

Each "pixel" of $I_{\sum}$ has the following value:

$$I_{\sum}(x, y) = \sum_{i=0}^{x} \sum_{j=0}^{y} I(i, j)$$

i.e., it holds the sum of all the pixel values of I that spatially precede the position (x, y).

AD · MAJOREM · DEI · GLORIAM 1870 LOYOLA UNIVERSITY CHICAGO

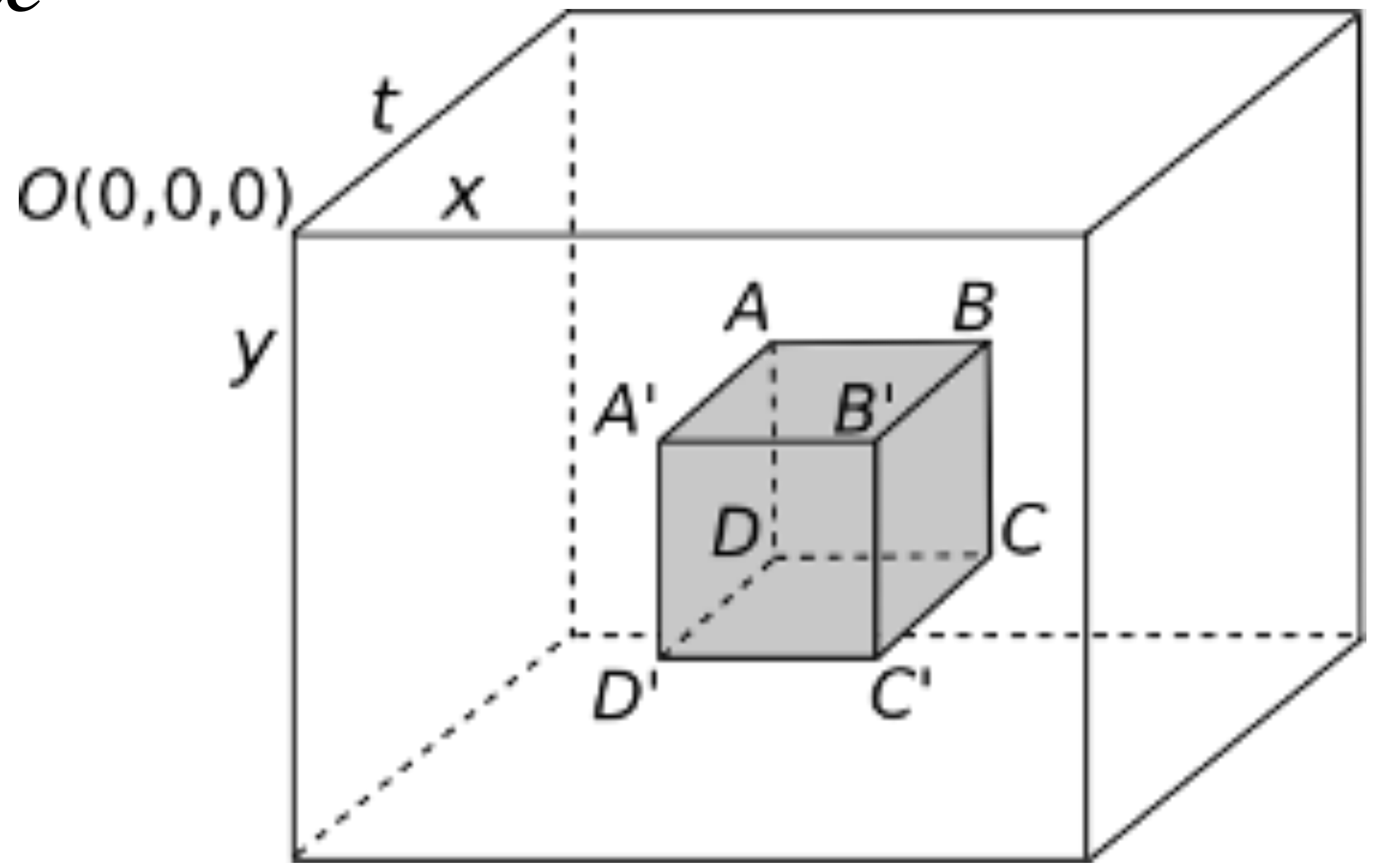
Temporal Robust Features (TRoF)

Inspiration on Speeded-Up Robust Features (SURF)
Integral **Video**

Convolutions supported by an integral video:

$$R = [(A + C) - (B + D) - (A' + C') + (B' + D')] \times filter\_value$$

Eight accesses for any filter size.



Temporal Robust Features (TRoF)

RECAP

Inspiration on Speeded-Up Robust Features (SURF)

Box Filters

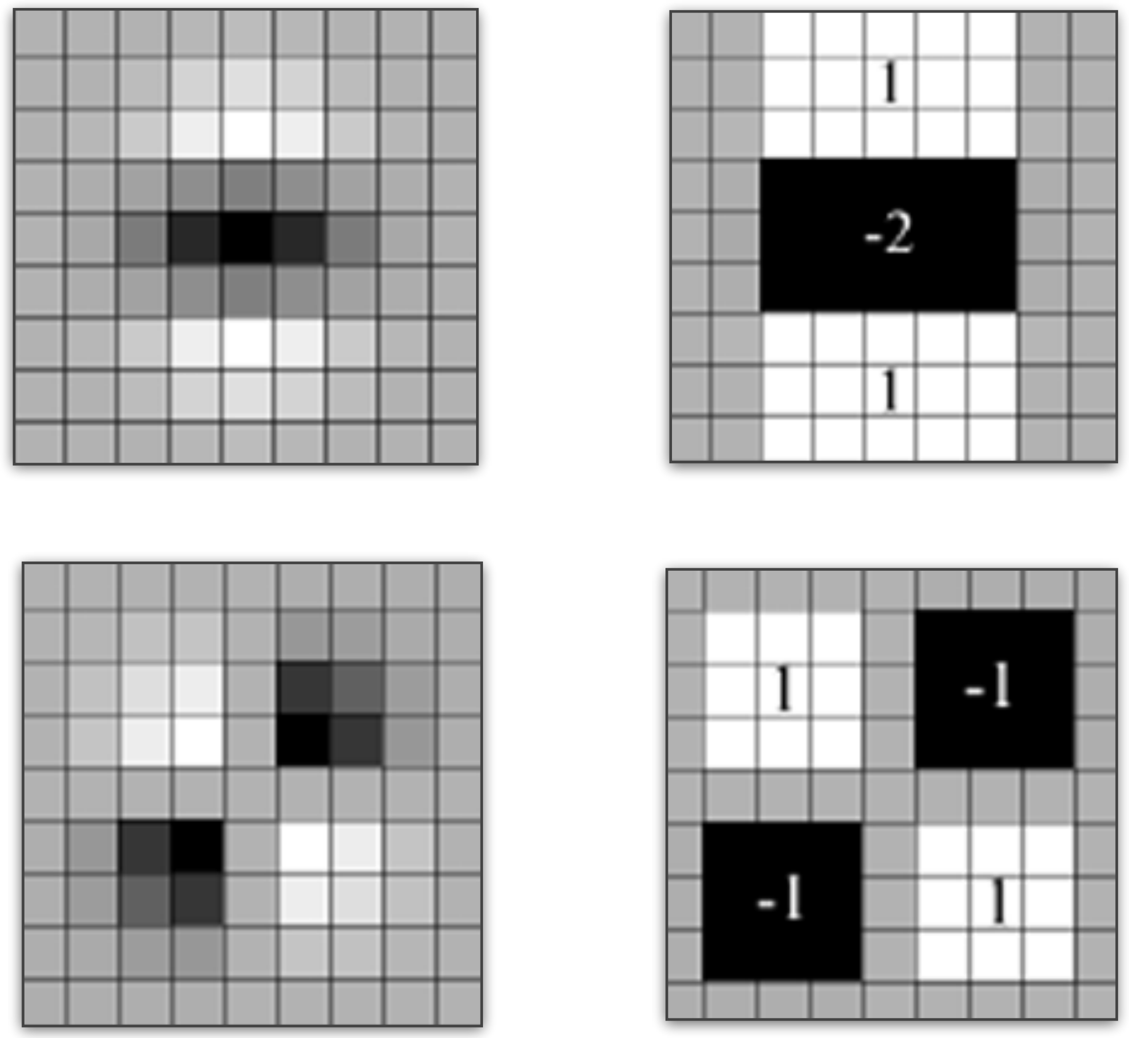
The Gaussian second order derivative functions $\frac{\delta^2}{\delta x^2}G(\sigma)$, $\frac{\delta^2}{\delta y^2}G(\sigma)$, and $\frac{\delta^2}{\delta xy}g(\sigma)$

can be approximated by box filters.

Compute the $det(H)$ quickly by using the box filters and the integral image!

$$H(x, y, \sigma) = \begin{bmatrix} \frac{\delta^2}{\delta x^2}g(\sigma) * I(x, y) & \frac{\delta^2}{\delta xy}g(\sigma) * I(x, y) \\ \frac{\delta^2}{\delta xy}g(\sigma) * I(x, y) & \frac{\delta^2}{\delta y^2}g(\sigma) * I(x, y) \end{bmatrix}$$

Bay's



$\frac{\delta^2}{\delta y^2}G(\sigma)$

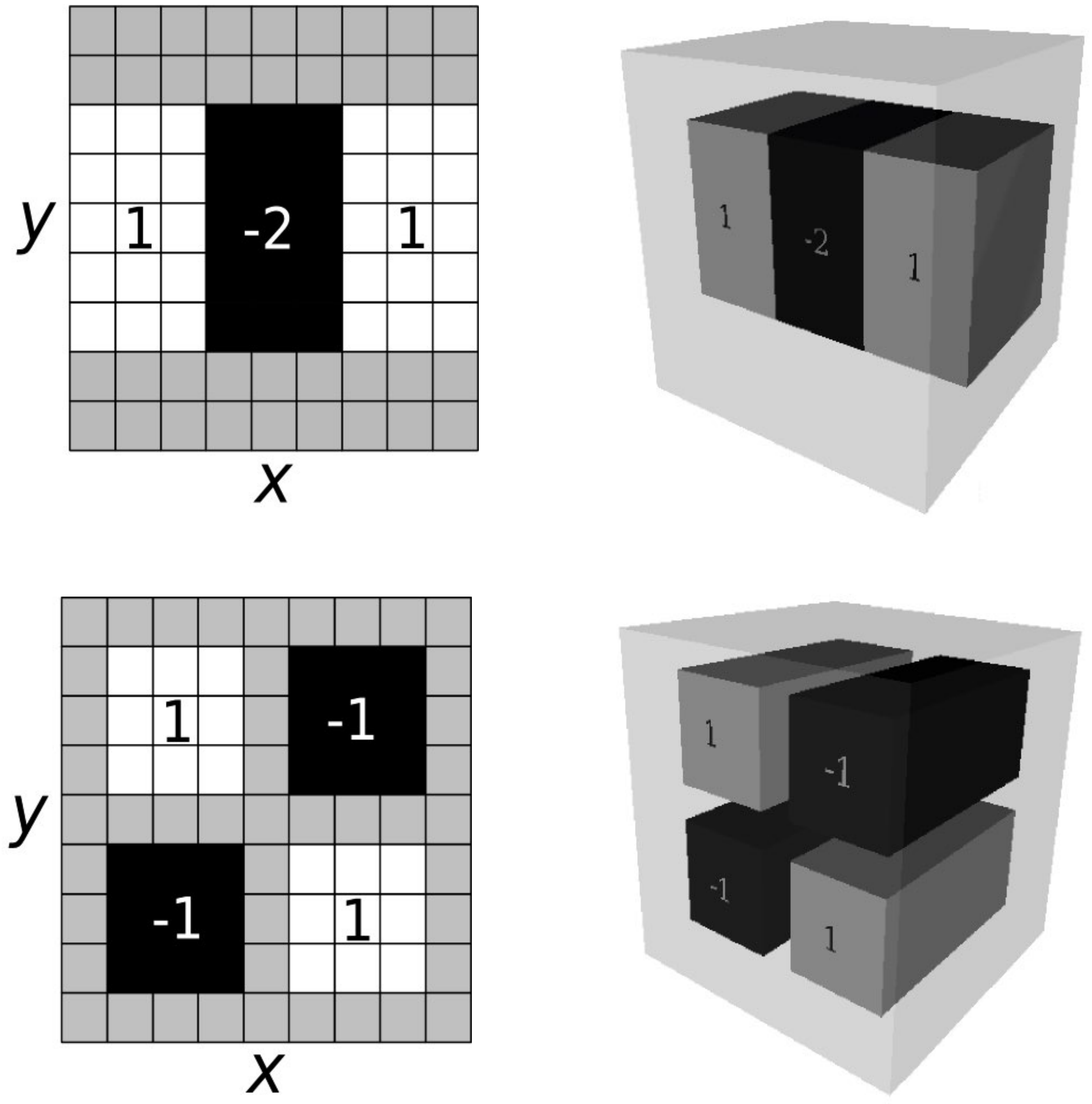
$\frac{\delta^2}{\delta xy}g(\sigma)$

LOYOLA UNIVERSITY CHICAGO

Temporal Robust Features (TRoF)

Inspiration on Speeded-Up Robust Features (SURF)

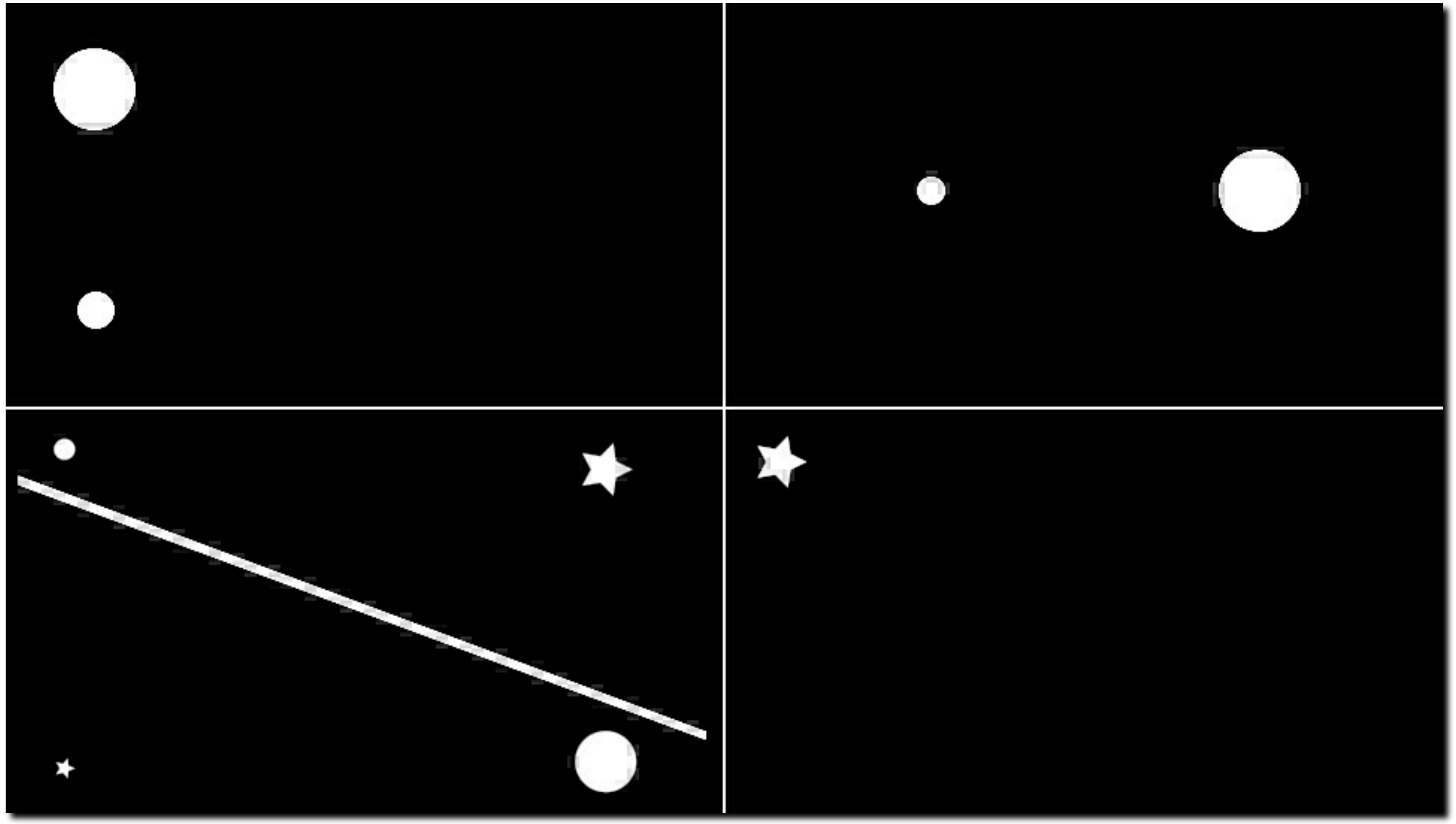
3D Box Filters



SURF TRoF

Temporal Robust Features (TRoF)

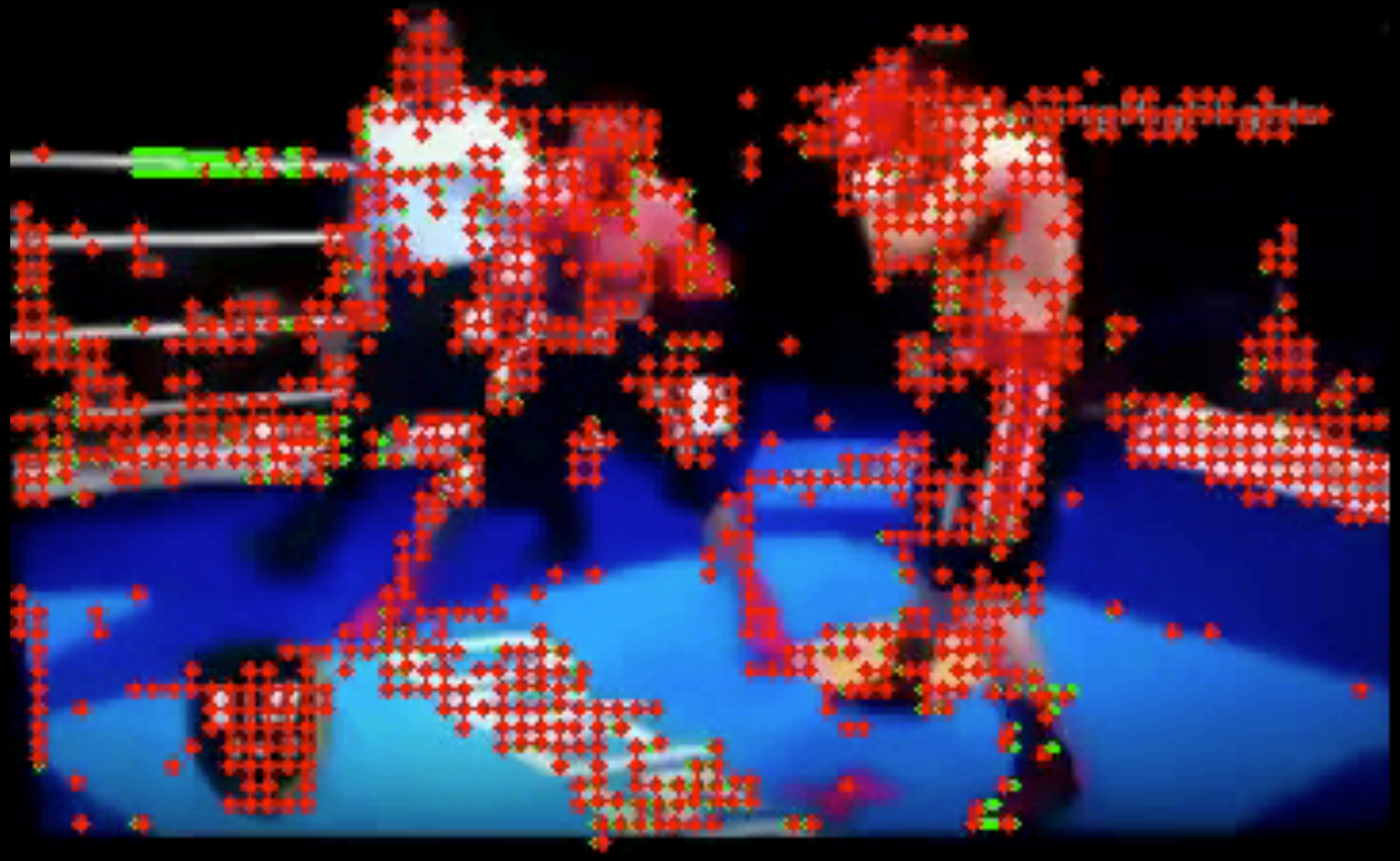
TRoF Detector



Original

TRoF

Dense Trajectories
(Wang et al., 2013)



STIP
(Laptev et al., 2008)

Temporal Robust Features (TRoF)

RECAP

Inspiration on Speeded-Up Robust Features (SURF)

Keypoint Description

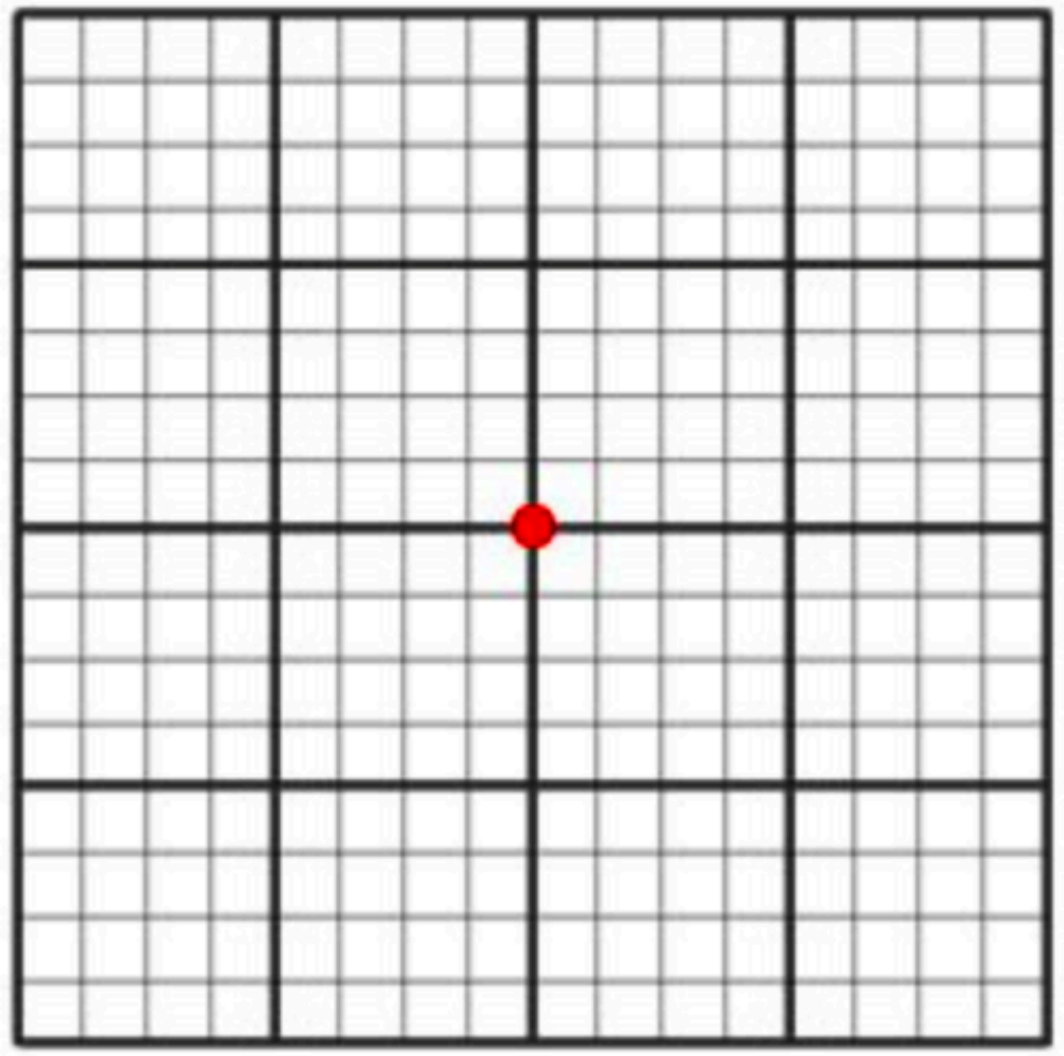
For each rotated keypoint, sample a 4 x 4 window on its neighborhood, according to the keypoint scale.

For each one of the 4 x 4 cells, compute 4 sums:
(1) $\sum d_x$, (2) $\sum |d_x|$, (3) $\sum d_y$, and (4) $\sum |d_y|$.

Fill out a feature vector with the 4 x 4 x 4 = 64 values.

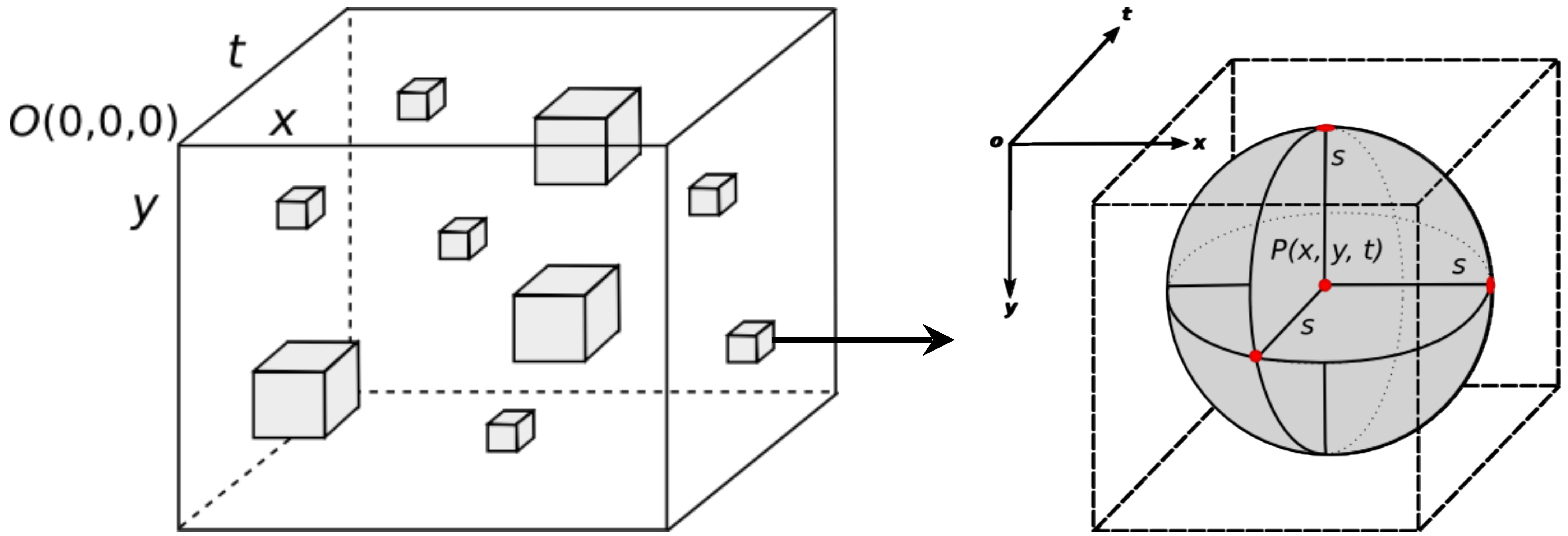
d_x

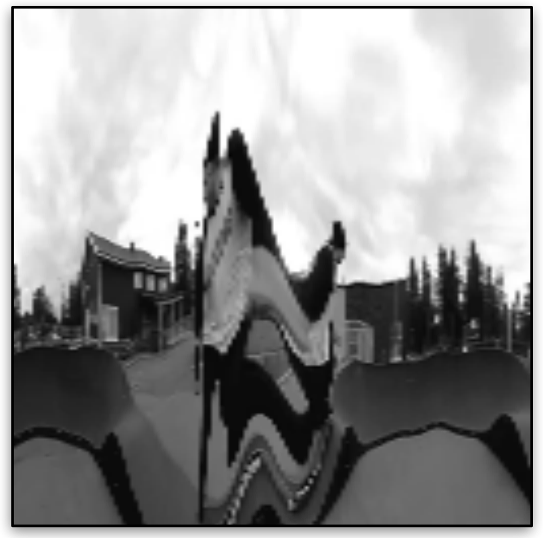
d_y



Temporal Robust Features (TRoF)

TRoF Descriptor
How to describe each detected blob?





SURF 64D	SURF 64D	SURF 64D
[x,y]	*[x,t]*	*[y,t]*

192 D

LOYOLA UNIVERSITY CHICAGO

Proposed Solution

Based on Bags of Visual Words that (BoVW)

Low level	Mid level	High level
Resolution reduction	GMM-based codebook	
Space-temporal local features	**Coding: Fisher Vectors***	Linear SVM
PCA Reduction	Pooling: single vector for entire footage	

*Perronin et al., 2010

Perronnin, F., Sanchez, J., and Mensink, T.
Improving the fisher kernel for large-scale image classification
European Conference on Computer Vision (ECCV), 2010

LOYOLA UNIVERSITY CHICAGO

Proposed Solution

Based on Bags of Visual Words that (BoVW)

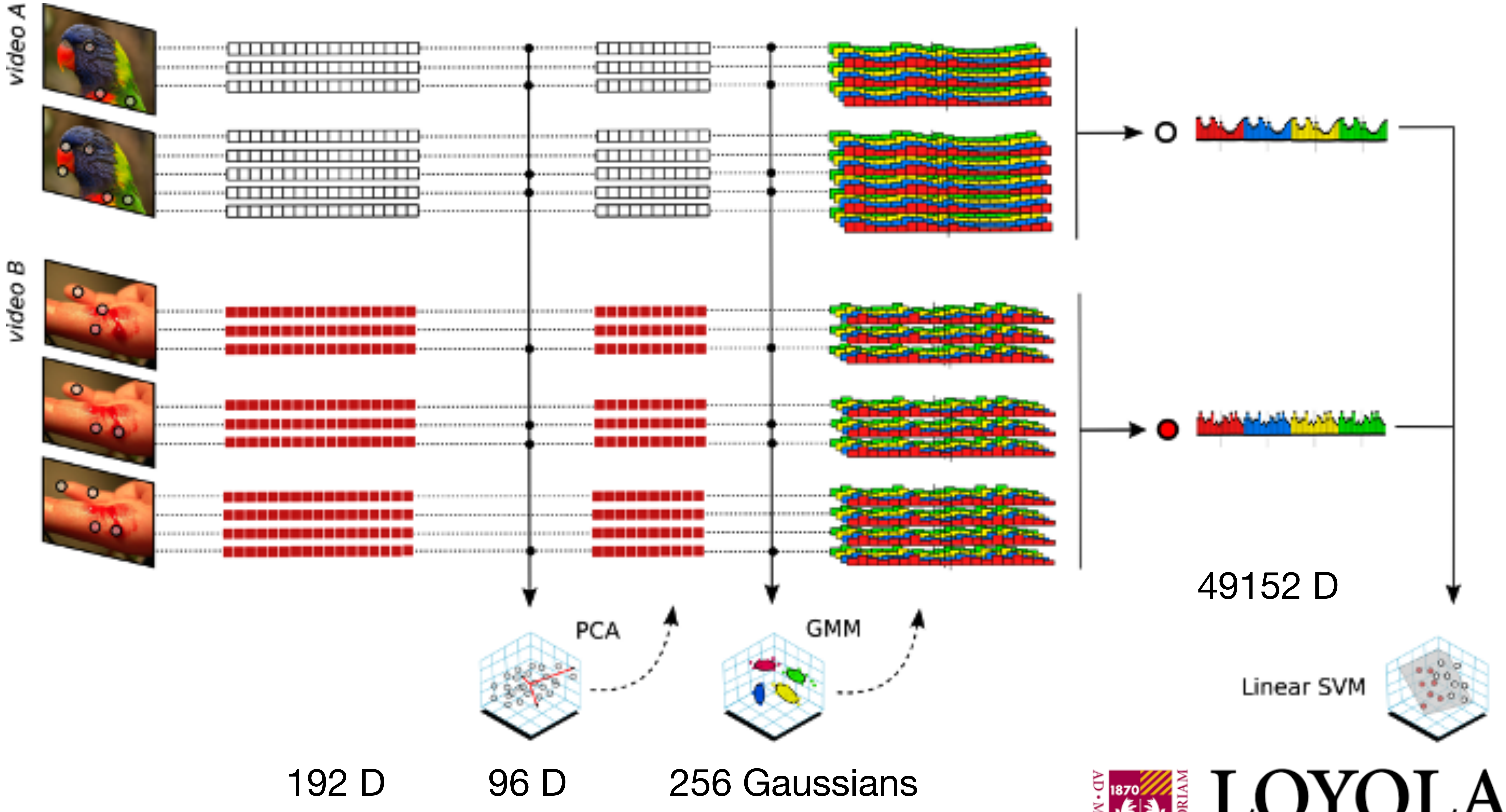
Low level	Mid level	High level
Resolution reduction	GMM-based codebook	
Space-temporal local features	Coding: Fisher Vectors	Linear SVM
PCA Reduction	**Pooling: single vector for entire footage***	

*Average Pooling

LOYOLA UNIVERSITY CHICAGO

Proposed Solution

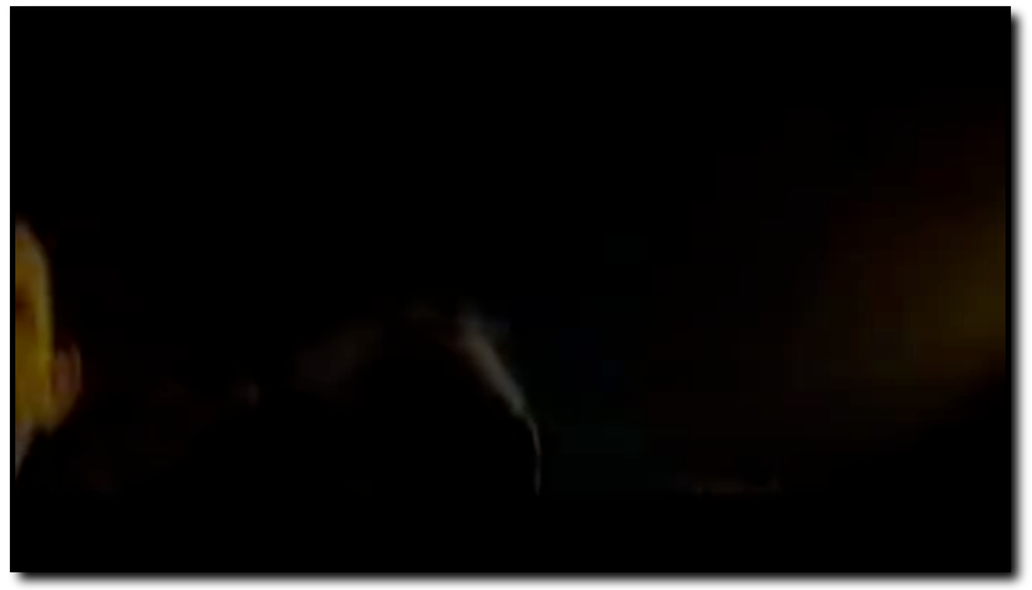
Inference Time



Violence Results

Dataset
MediaEval 2013



"Content one would not let a child see." [2]

Training: 18 movies
Test: 7 movies

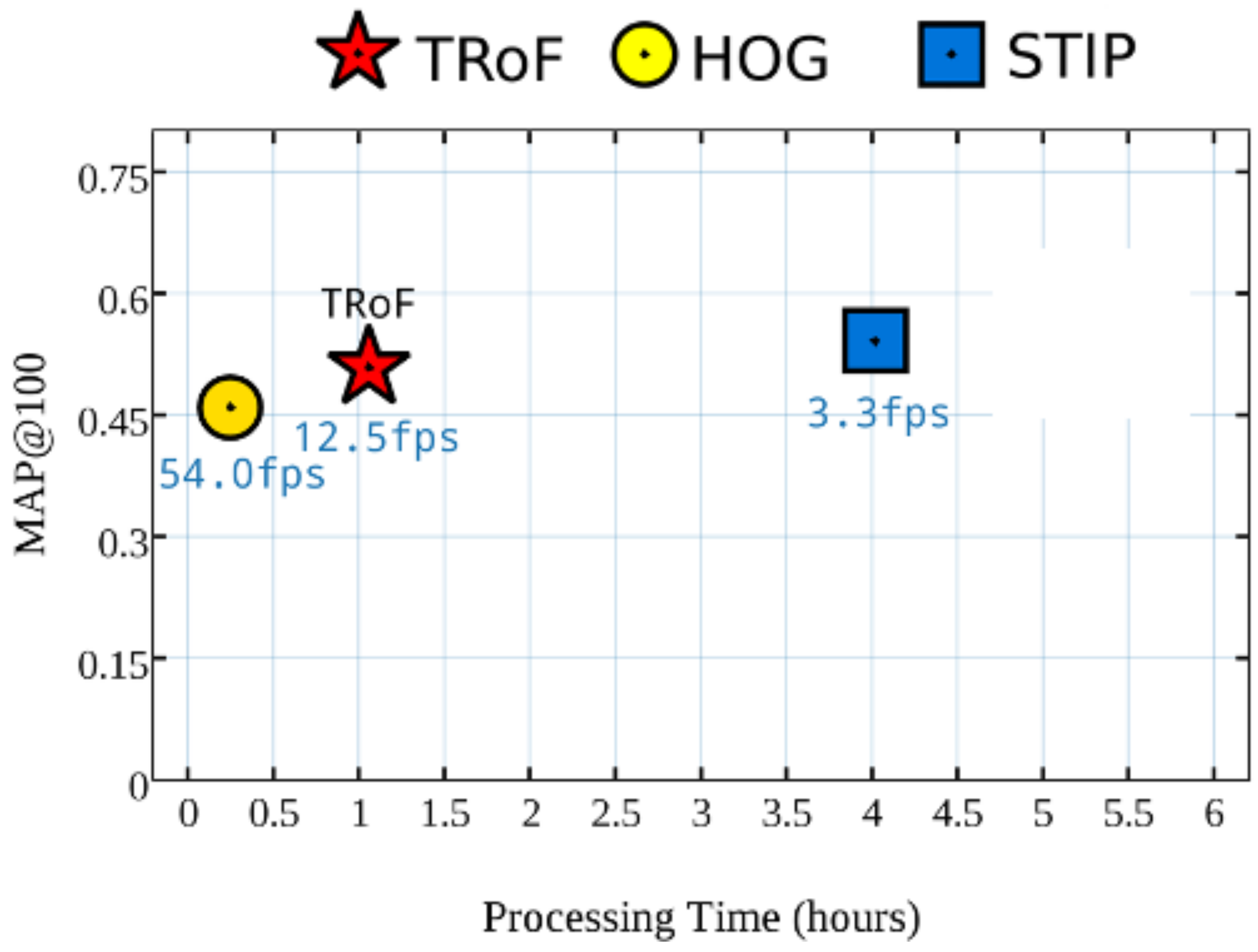
Shot-based segmentation and classification.

Metric: Mean Average Precision (MAP)

[2] Demarty et al., *Benchmarking Violent Scenes Detection in Movies*. In IEEE CBMI, 2014

Violence Results

MAP vs. Runtime



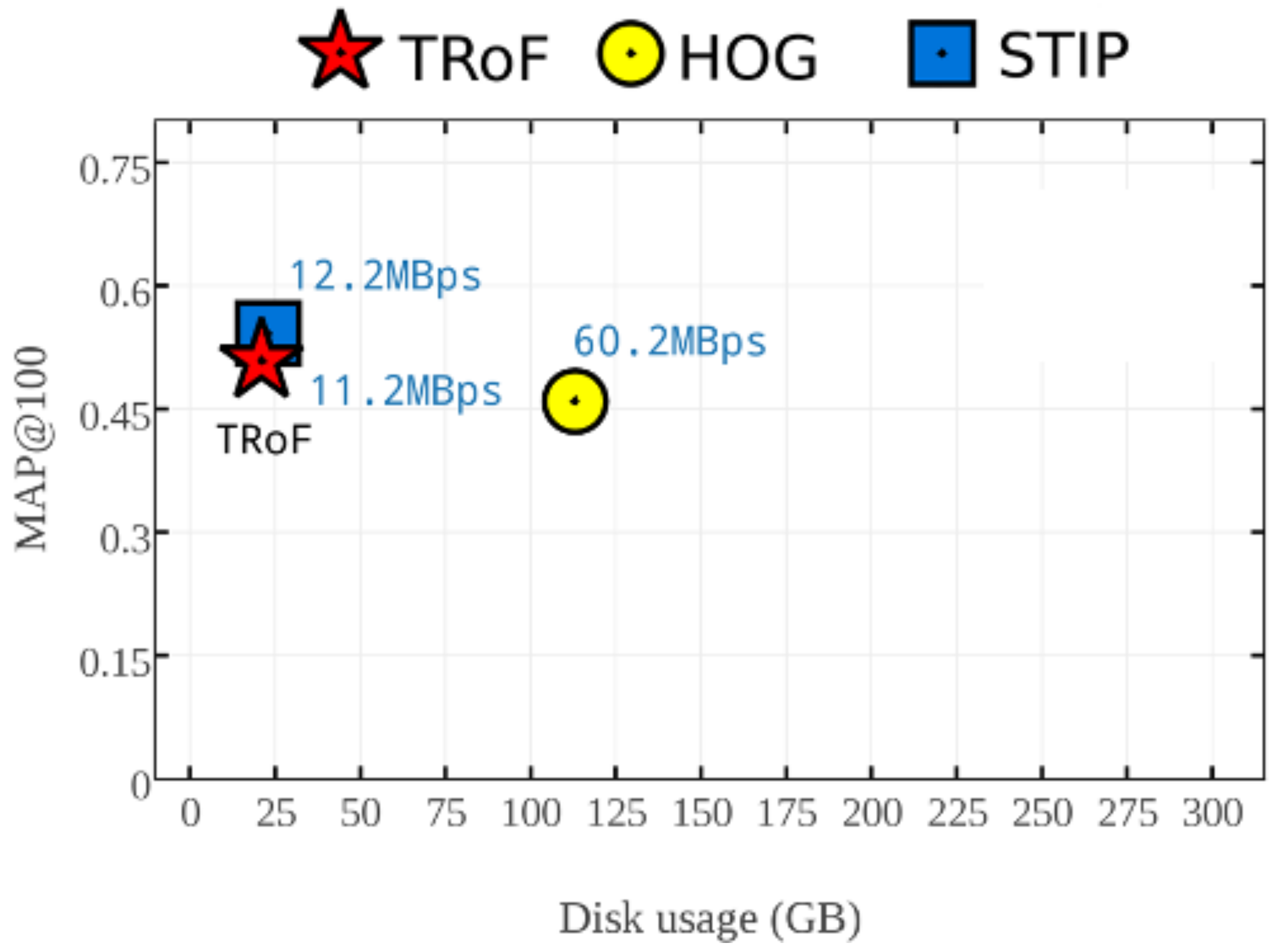
2013 MediaEval Dataset

LOYOLA UNIVERSITY CHICAGO

Violence Results

MAP vs. Memory Footprint

2013 MediaEval Dataset



LOYOLA UNIVERSITY CHICAGO

Violence Results

True Positive Sample

LOYOLA UNIVERSITY CHICAGO

Violence Results

False Negative Sample

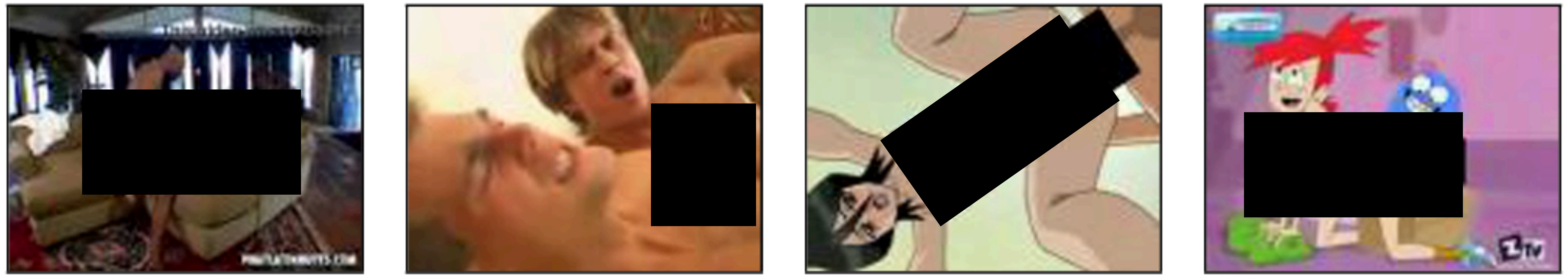
LOYOLA UNIVERSITY CHICAGO

Violence Results

False Negative Sample

Pornography Results

Dataset
Porn-2k



(a) (b) (c) (d)

(e) (f) (g) (h)

"Any explicit sexual matter with the purpose of eliciting arousal." [1]

140h of video
1000 porn clips
1000 non-porn clips

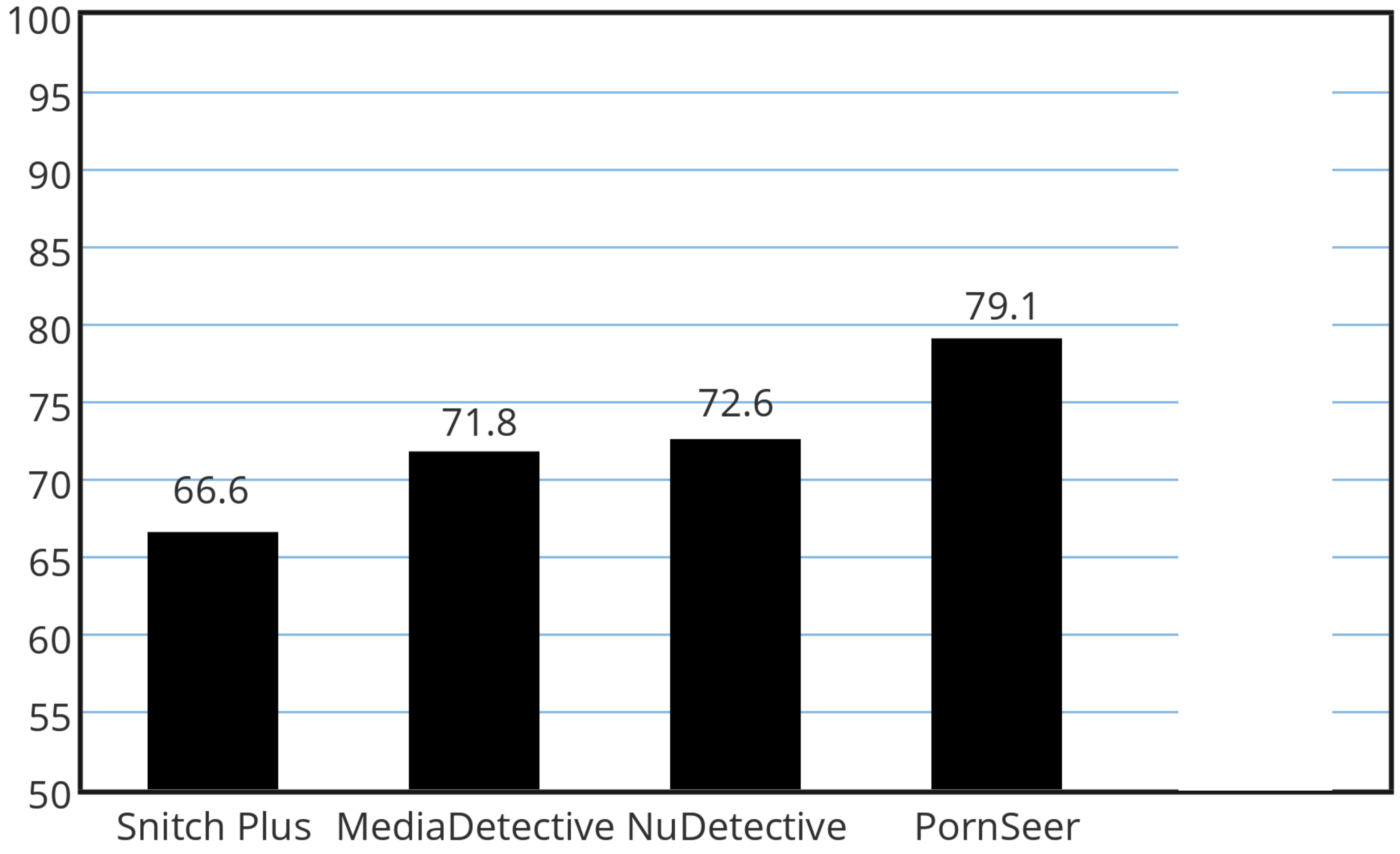
Metric: Classification Accuracy

YouTube Vine
vimeo Porn sites

[1] Short et al., *A review of internet pornography use research: Methodology and content from the past 10 years.* Cyberpsychology, Behavior, and Social Networking 15, 2012

Pornography Results

Classification Accuracy

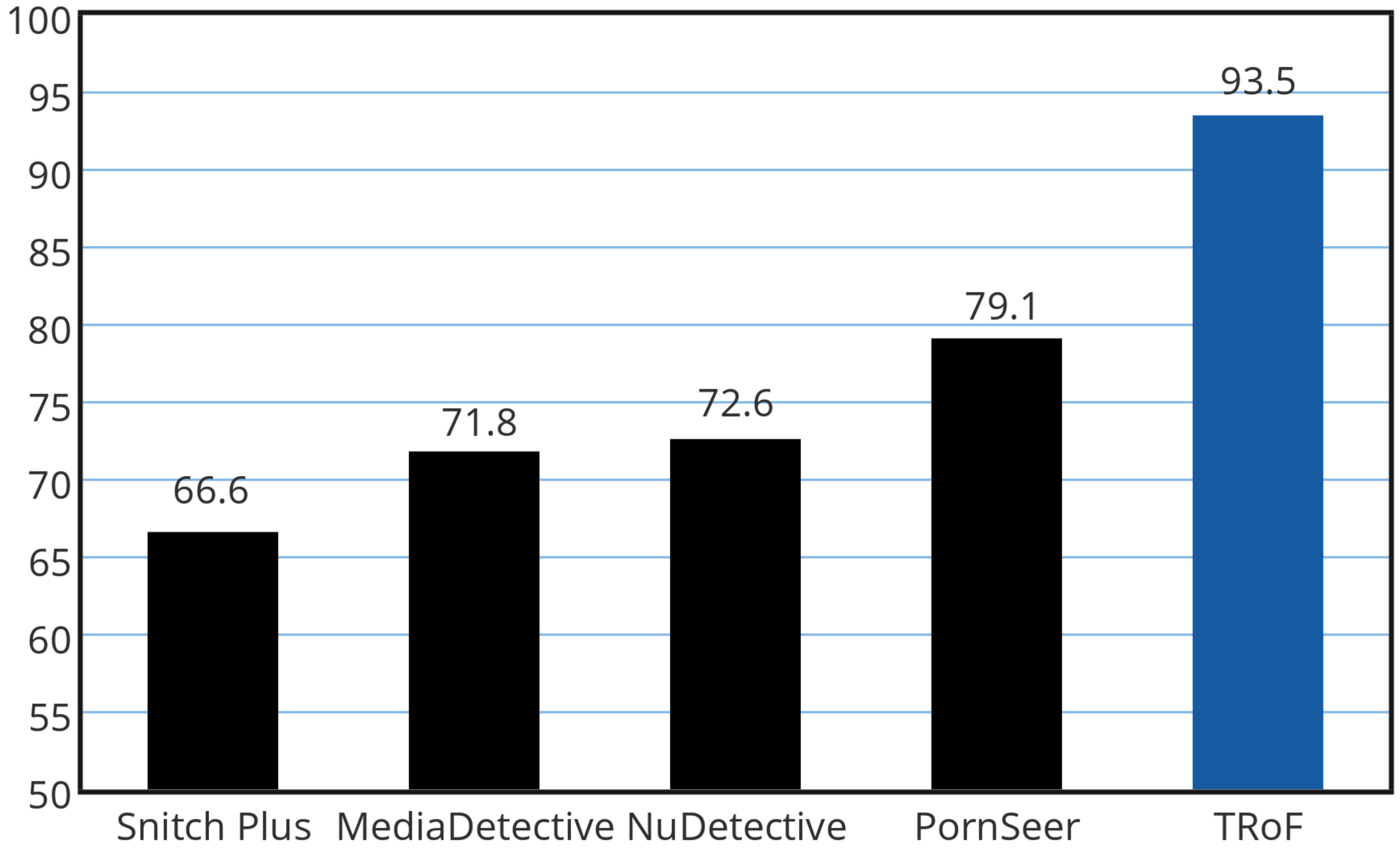


Porn-2k Dataset

LOYOLA UNIVERSITY CHICAGO

Pornography Results

Classification Accuracy

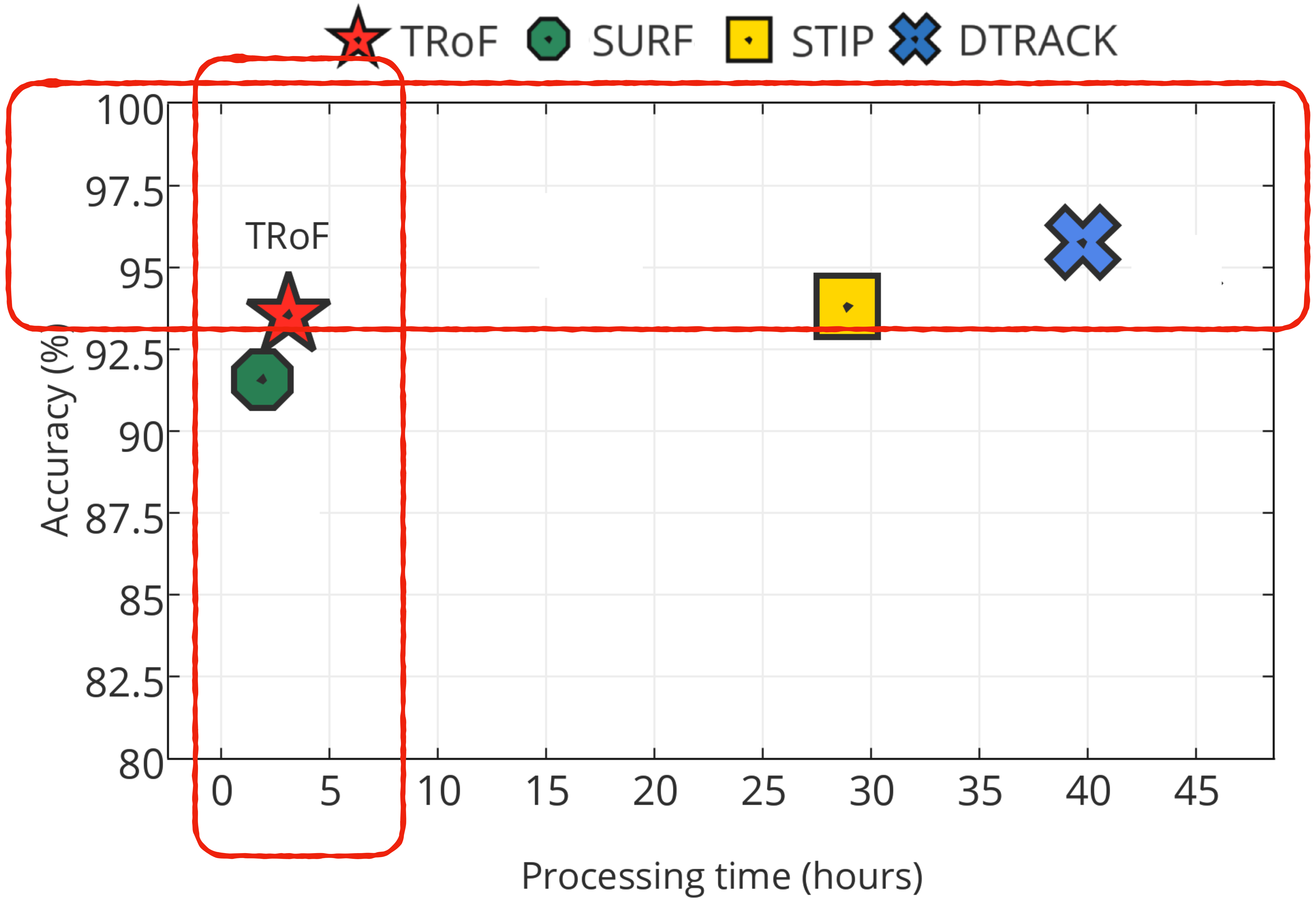


Porn-2k Dataset

LOYOLA UNIVERSITY CHICAGO

Pornography Results

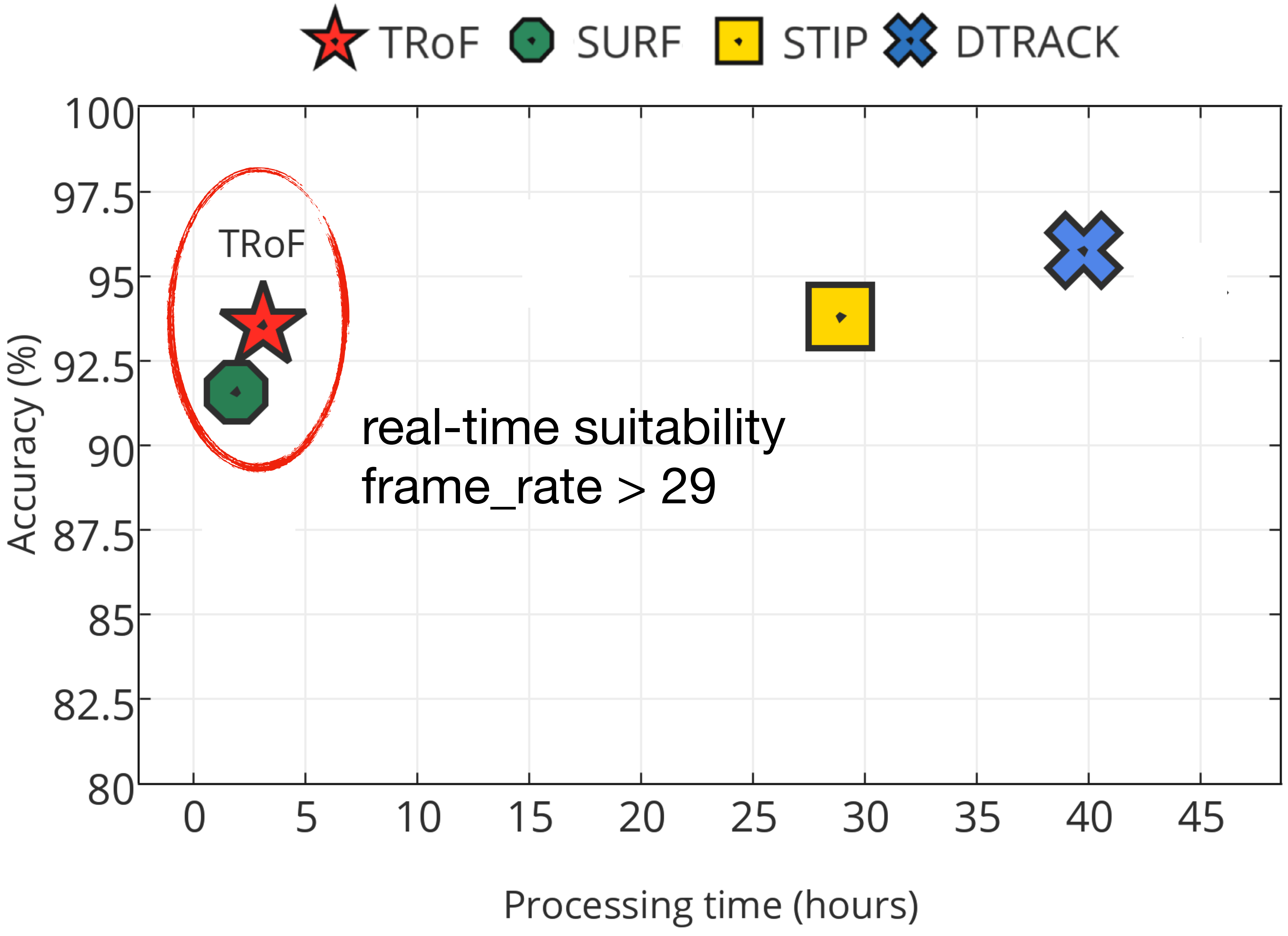
Accuracy vs. Runtime



Porn-2k Dataset

Pornography Results

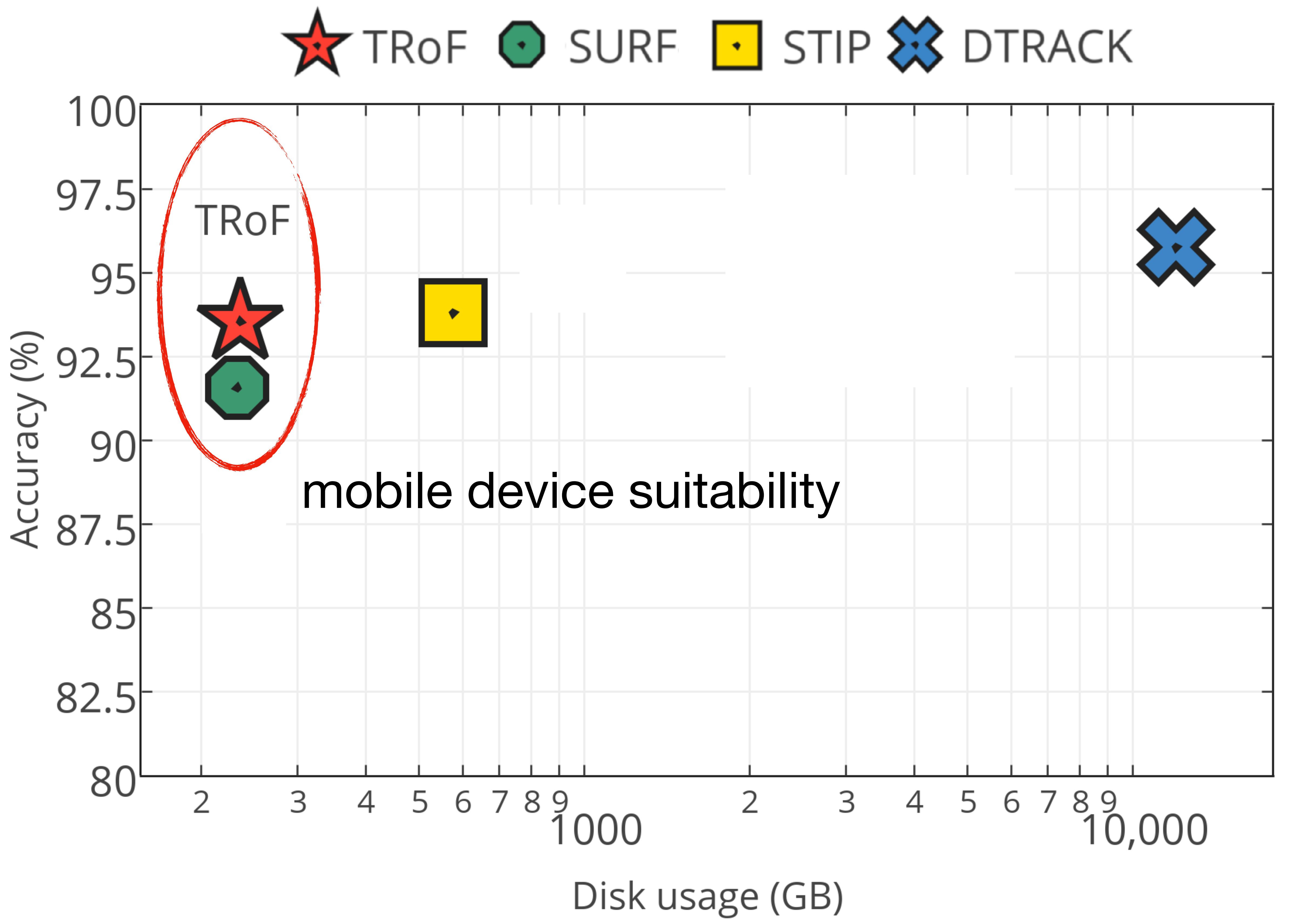
Accuracy vs. Runtime



Porn-2k Dataset

Pornography Results

Accuracy vs. Memory Footprint



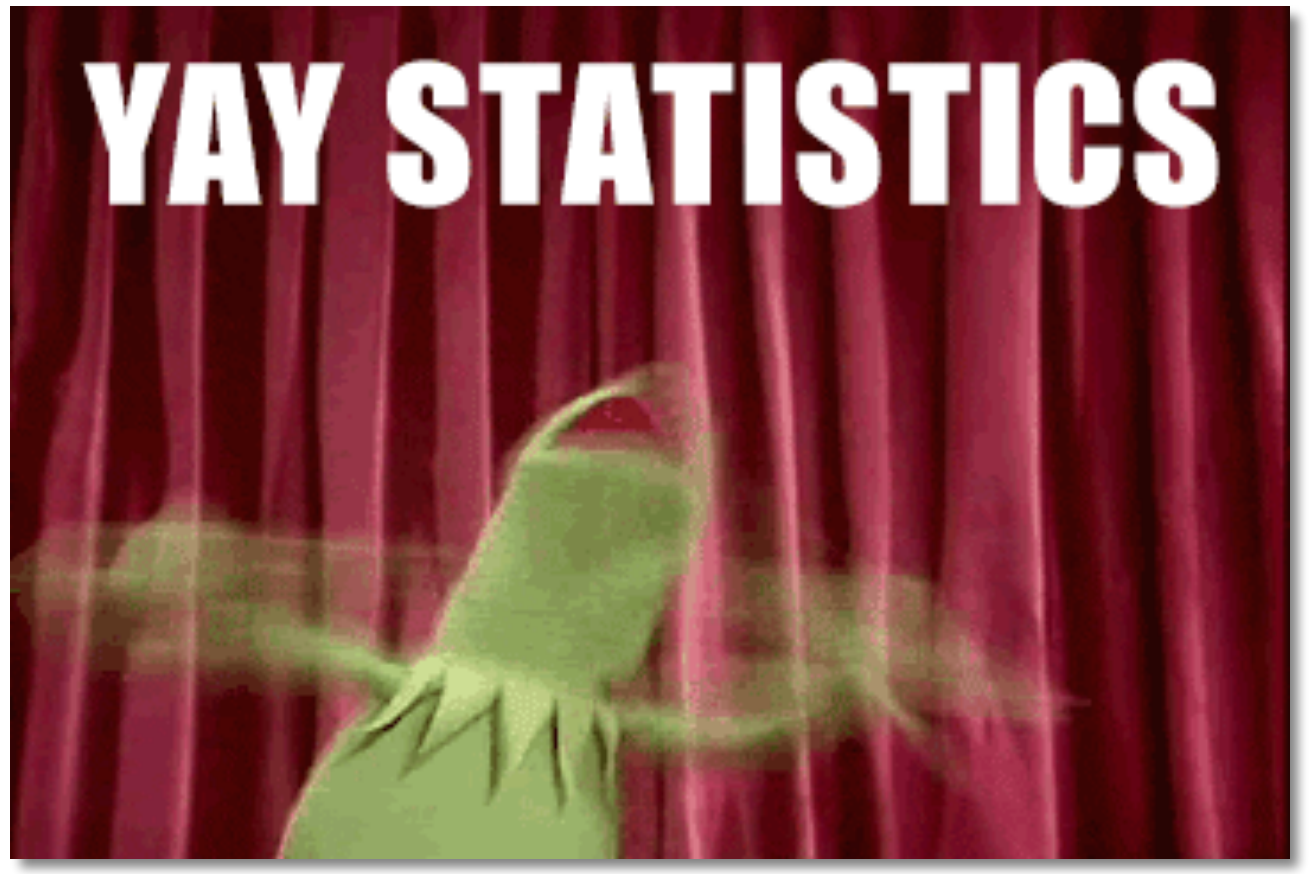
Porn-2k Dataset

LOYOLA UNIVERSITY CHICAGO

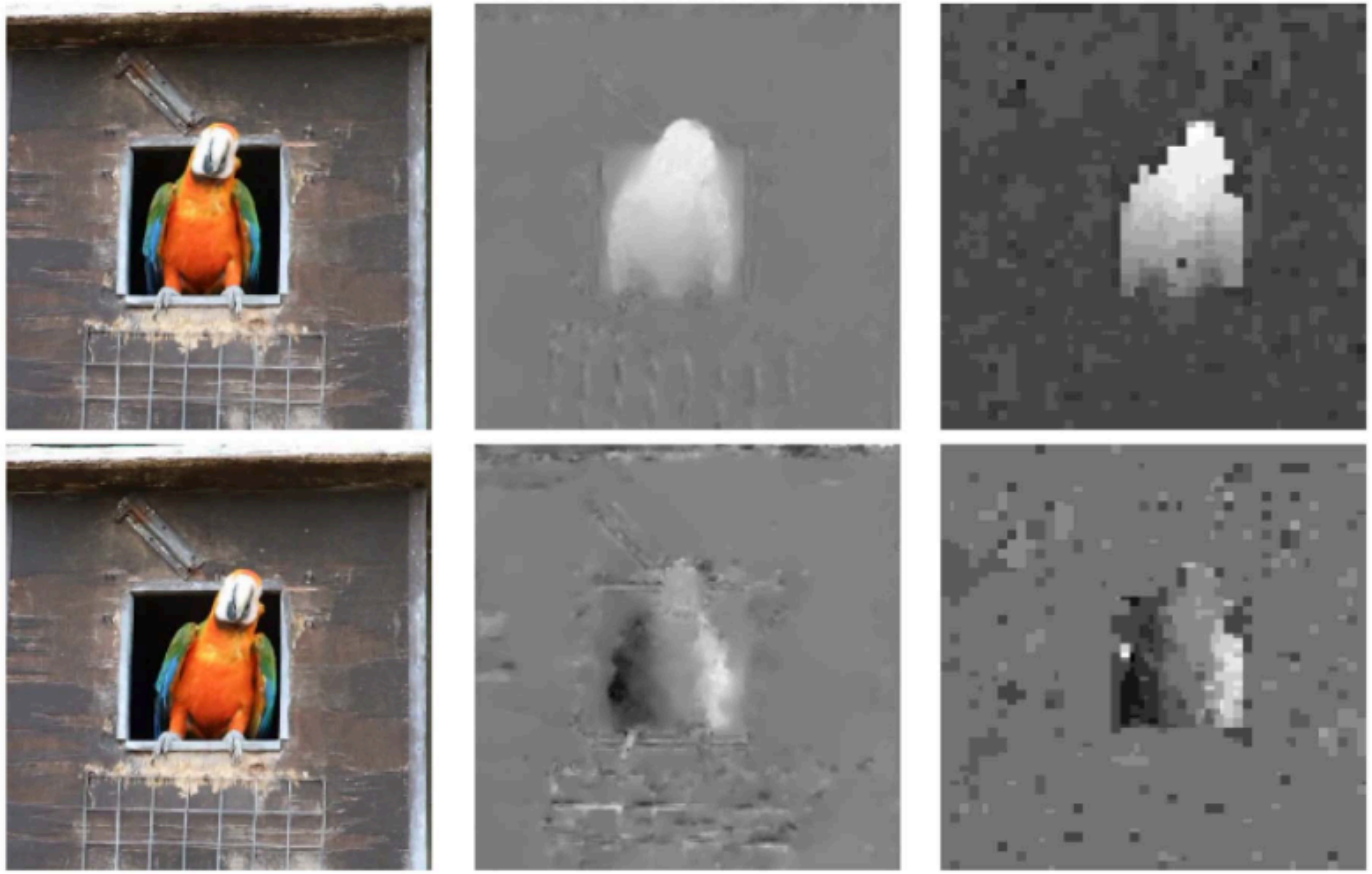
Training Protocol

Folding Blurb
5x2-fold cross validation
Non-parametric pairwise Wilcoxon signed-rank test,
with Bonferroni's *p*-correction

Reference
Demšar, J.
Statistical comparisons of classifiers over multiple data sets
ACM Journal of Machine Learning Research (JMLR) 7 (1), 2006



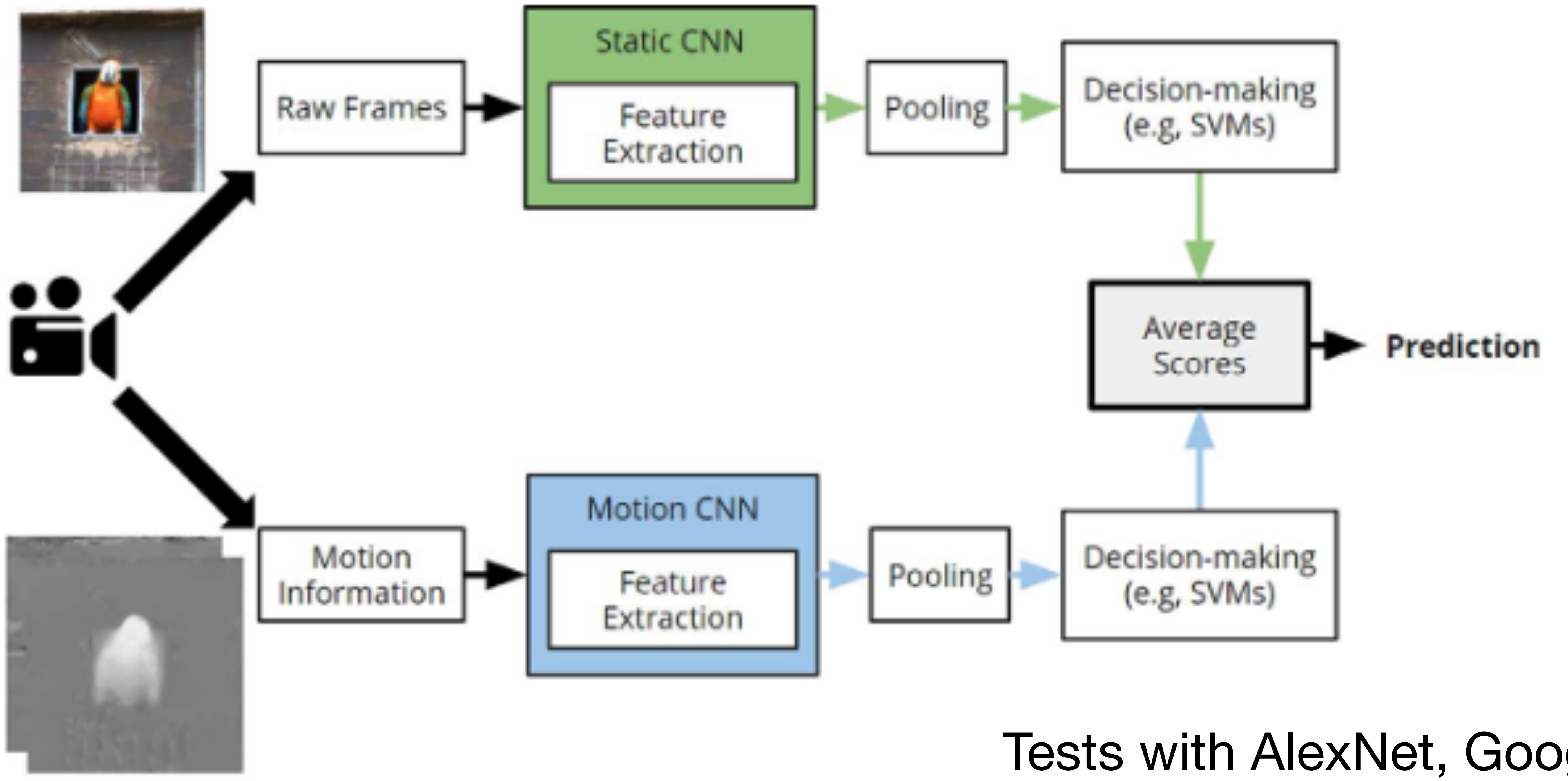
Deep Learning?



(a) Sequential Raw frames (b) Optical Flow (c) Motion Vectors

Perez, M., at al.
Video pornography detection through deep learning techniques and motion information
Elsevier Neurocomputing 230, 2017

Deep Learning?



Tests with AlexNet, Googlenet, and VGG.
Best results so far.
Portable to mobile devices?

Tasks

Part I: Sensitive Video Classification

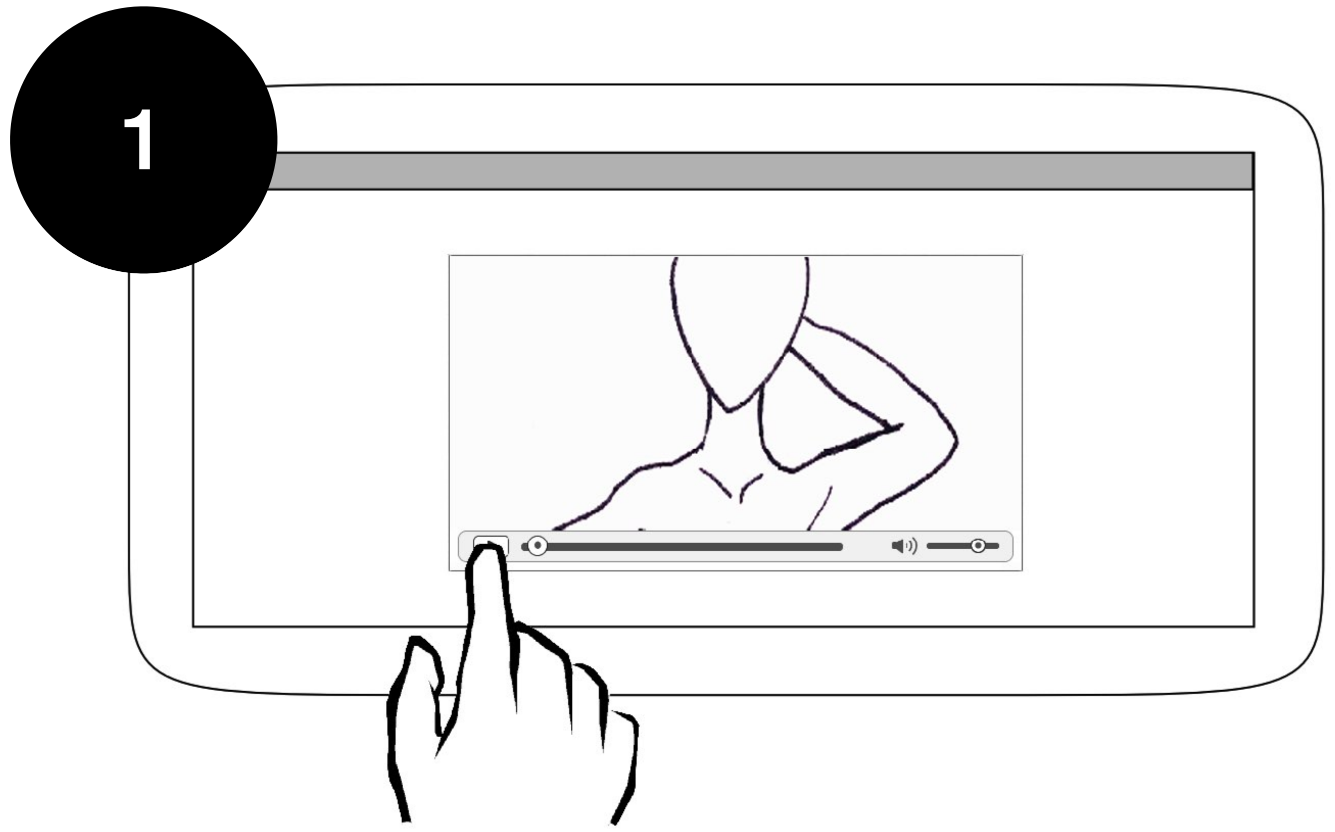
Part II: Sensitive Video Detection

Sensitive Video Detection

LOYOLA
UNIVERSITY CHICAGO

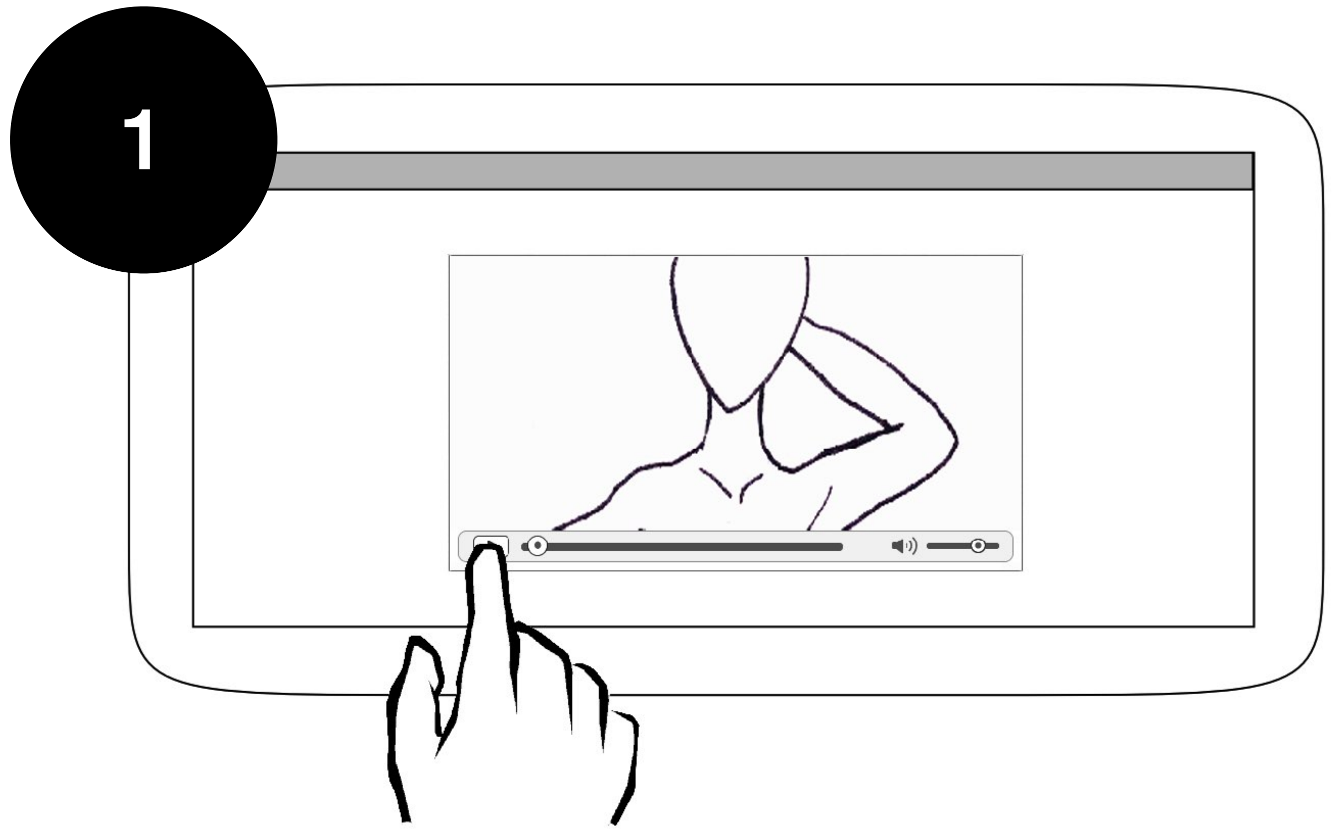
Task

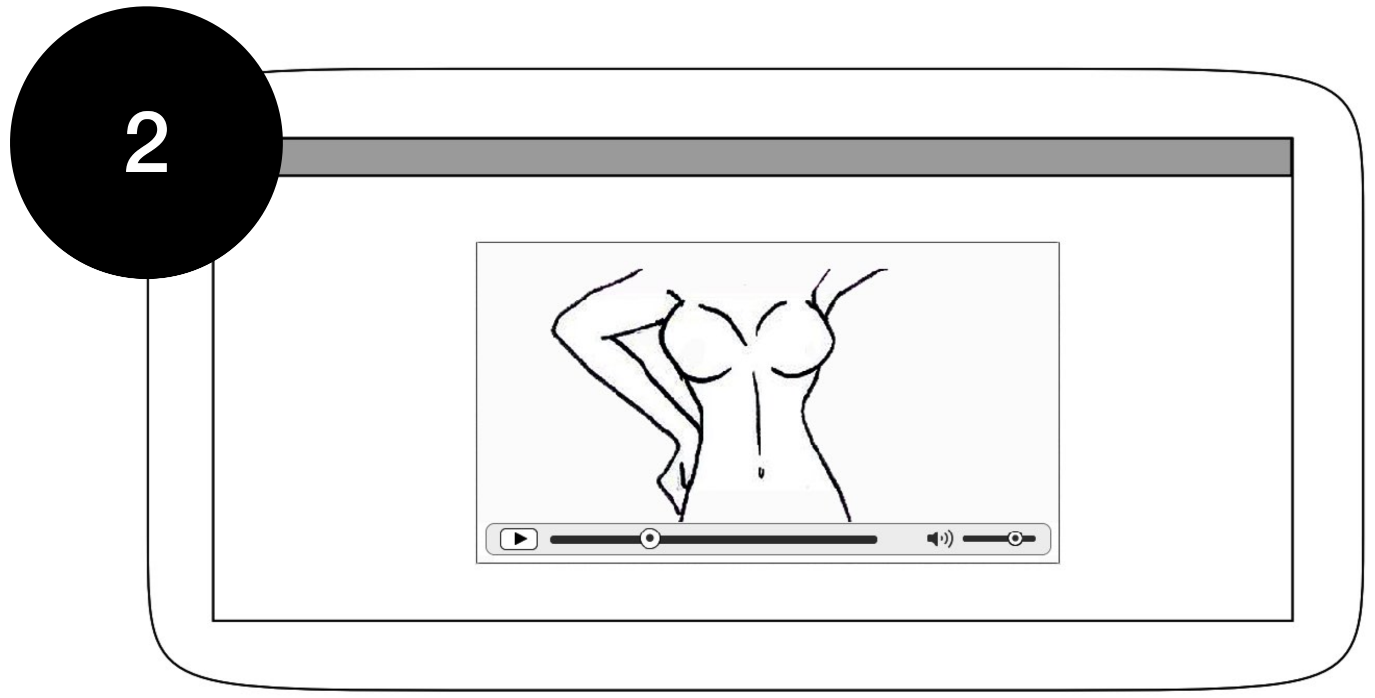
Can a computer detect (or localize) sensitive scenes within the video timeline?



Task

Can a computer detect (or localize) sensitive scenes within the video timeline?

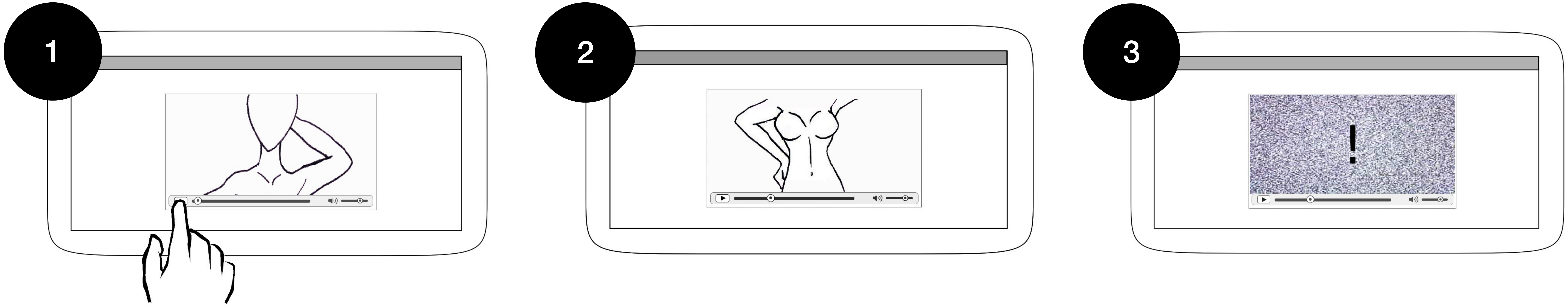




LOYOLA UNIVERSITY CHICAGO

Task

Can a computer detect (or localize) sensitive scenes within the video timeline?



Why do we care?

The Washington Post
The Intersect
A 12-year-old girl live-streamed her suicide. It took two weeks for Facebook to take the

The New York Times
Teenager Is Accused of Live-Streaming a Friend's Rape

SOUTH FLORIDA
Miami Herald
Another girl hangs herself while streaming it live — this time in M

Man shot, killed while live-streaming
CNN

CNN BUSINESS
Markets Tech Media Success Perspectives Video
U.S. Edition +
Seven weeks later, videos of New Zealand attack still circulating on Facebook and Instagram

AD · MAJOREM · DEI · GLORIAM 1870
LOYOLA
UNIVERSITY CHICAGO

State of the Art

Can a computer detect (or localize) sensitive scenes within the video timeline?

Literature

?

?

Yes!

Surveillance Violence Detection

No, you can't.

Hollywood Violence Detection

LOYOLA UNIVERSITY CHICAGO

State of the Art

Can a computer detect (or localize) sensitive scenes within the video timeline?

No, you can't.

Yes!

Literature

?

?

Surveillance Violence Detection

Hollywood Violence Detection

LOYOLA UNIVERSITY CHICAGO

Sponsor's Challenge

Can a computer detect sensitive content other than violence?

Literature

?

?

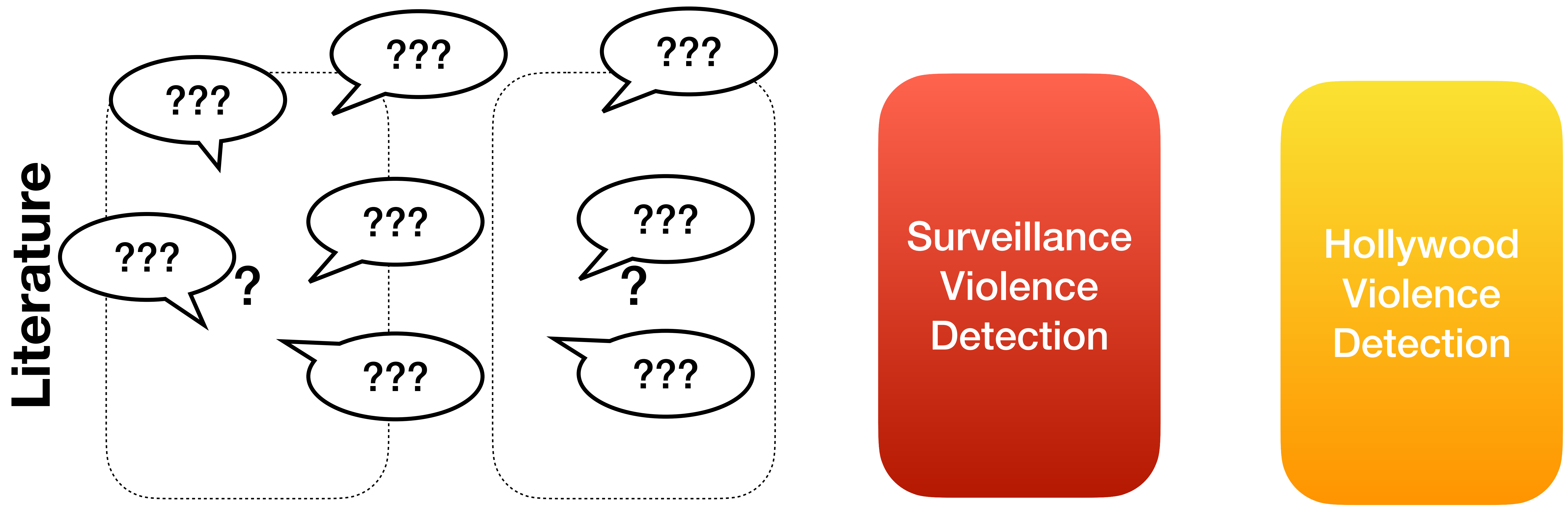
Surveillance Violence Detection

Hollywood Violence Detection

LOYOLA UNIVERSITY CHICAGO

Sponsor's Challenge

Can a computer detect sensitive content other than violence?



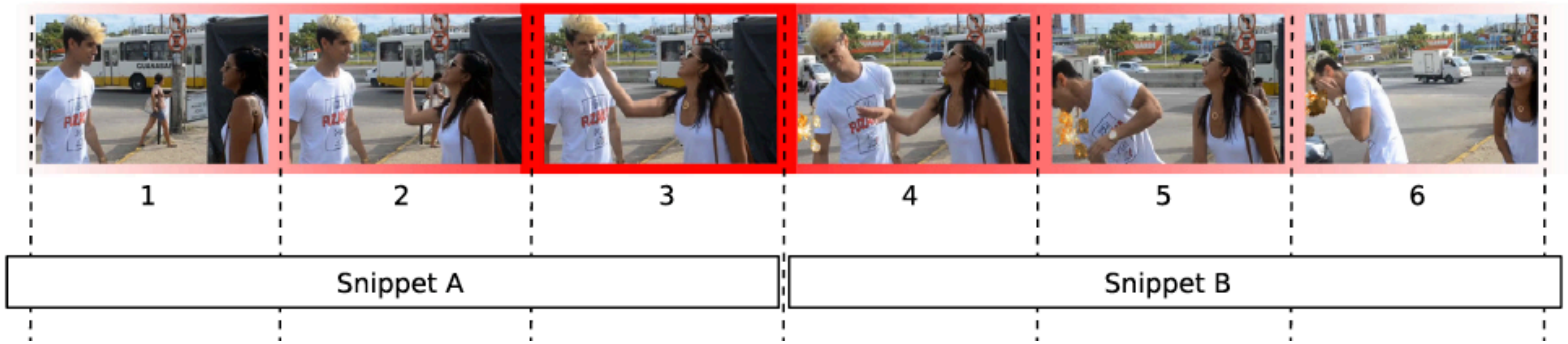
LOYOLA UNIVERSITY CHICAGO

Proposed Solution

Video Snippet Segmentation

Proposed Solution

Video Snippet Segmentation



Non-overlapping Snippets

Proposed Solution

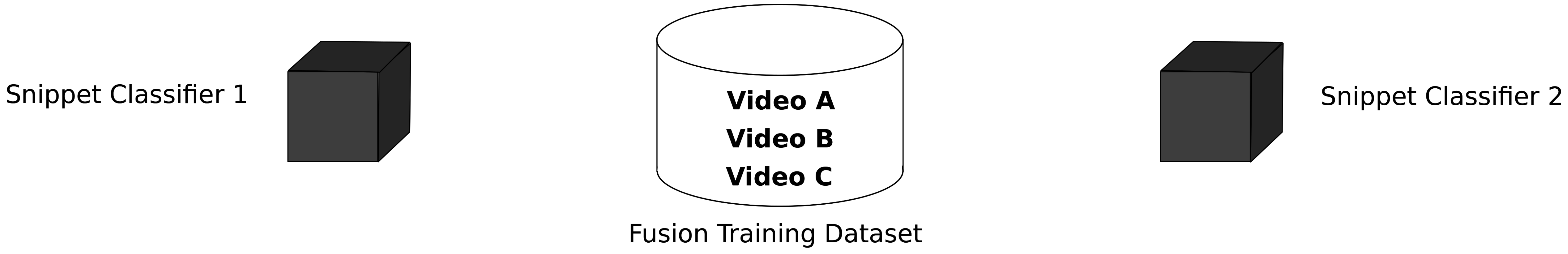
Video Snippet Segmentation

Overlapping Snippets

LOYOLA UNIVERSITY CHICAGO

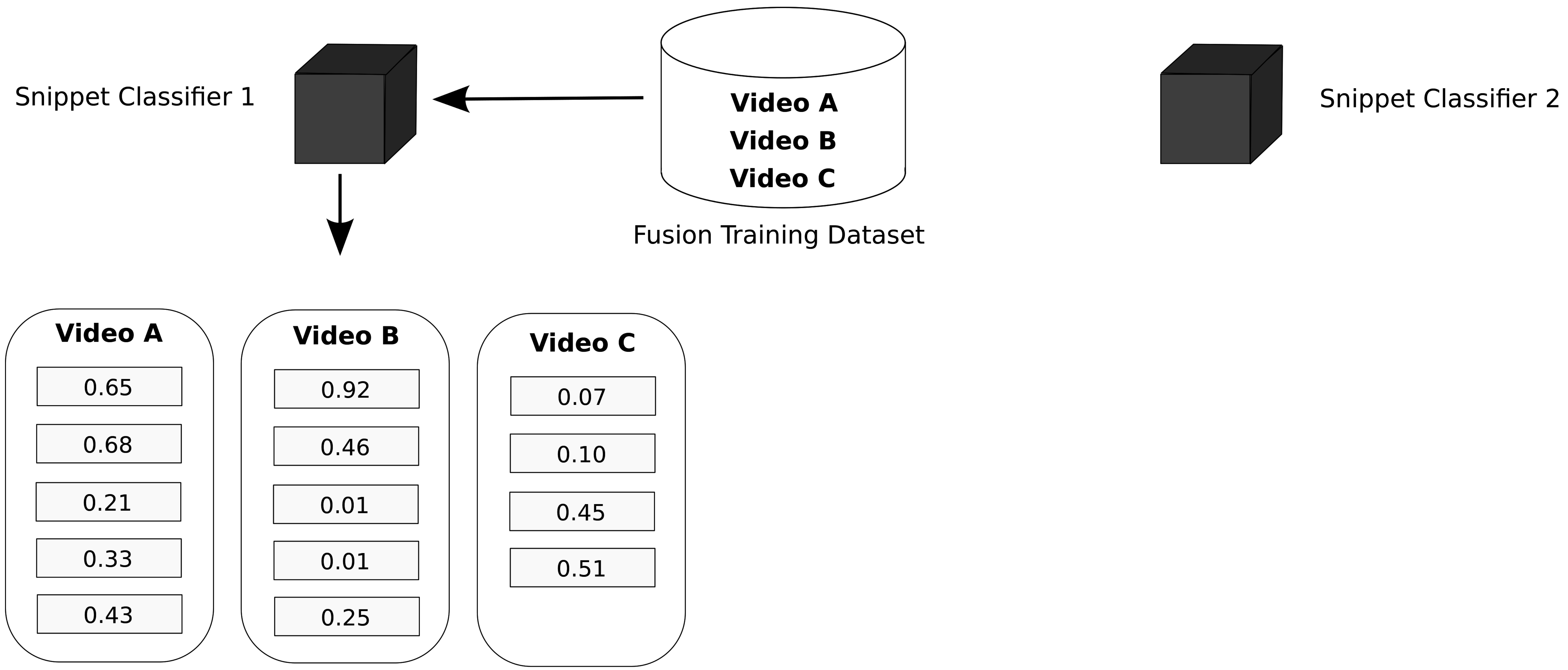
Proposed Solution

Snippet Classification



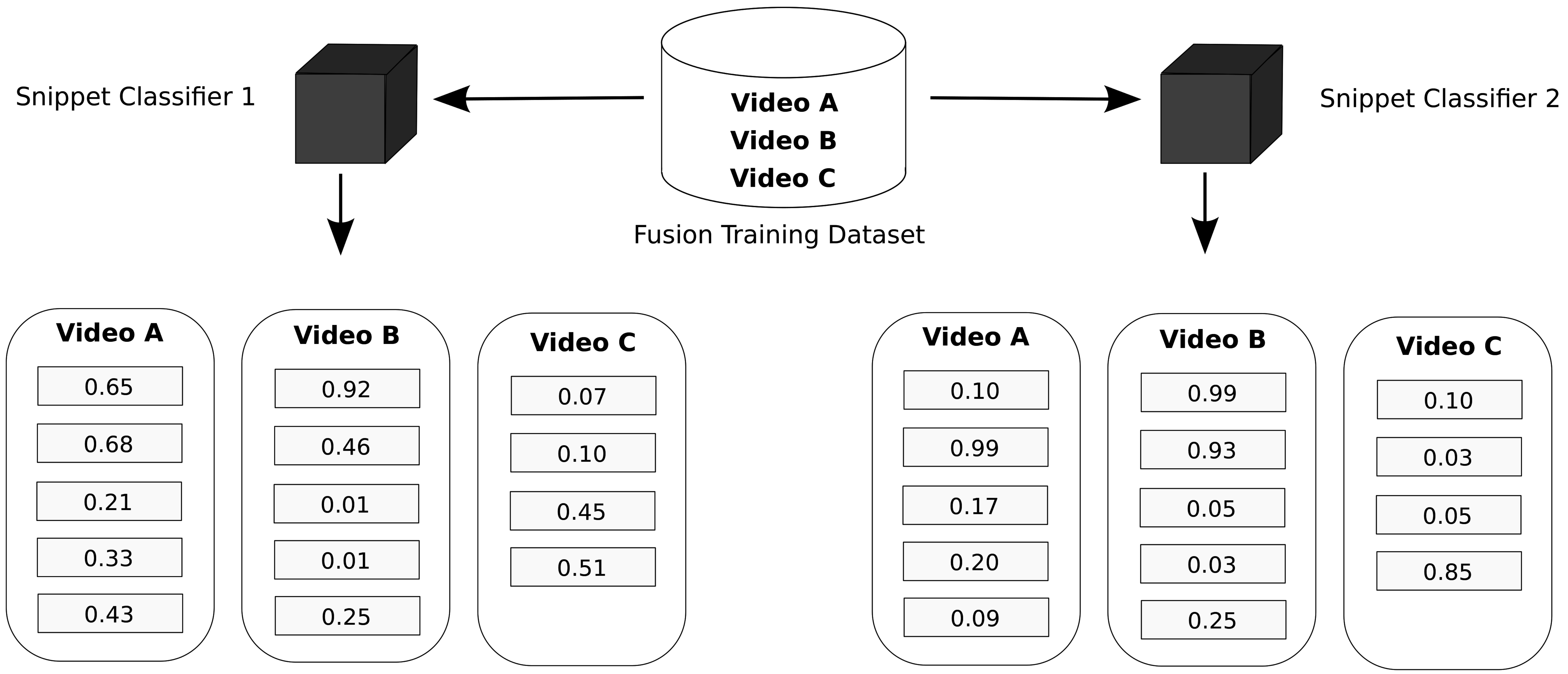
Proposed Solution

Snippet Classification



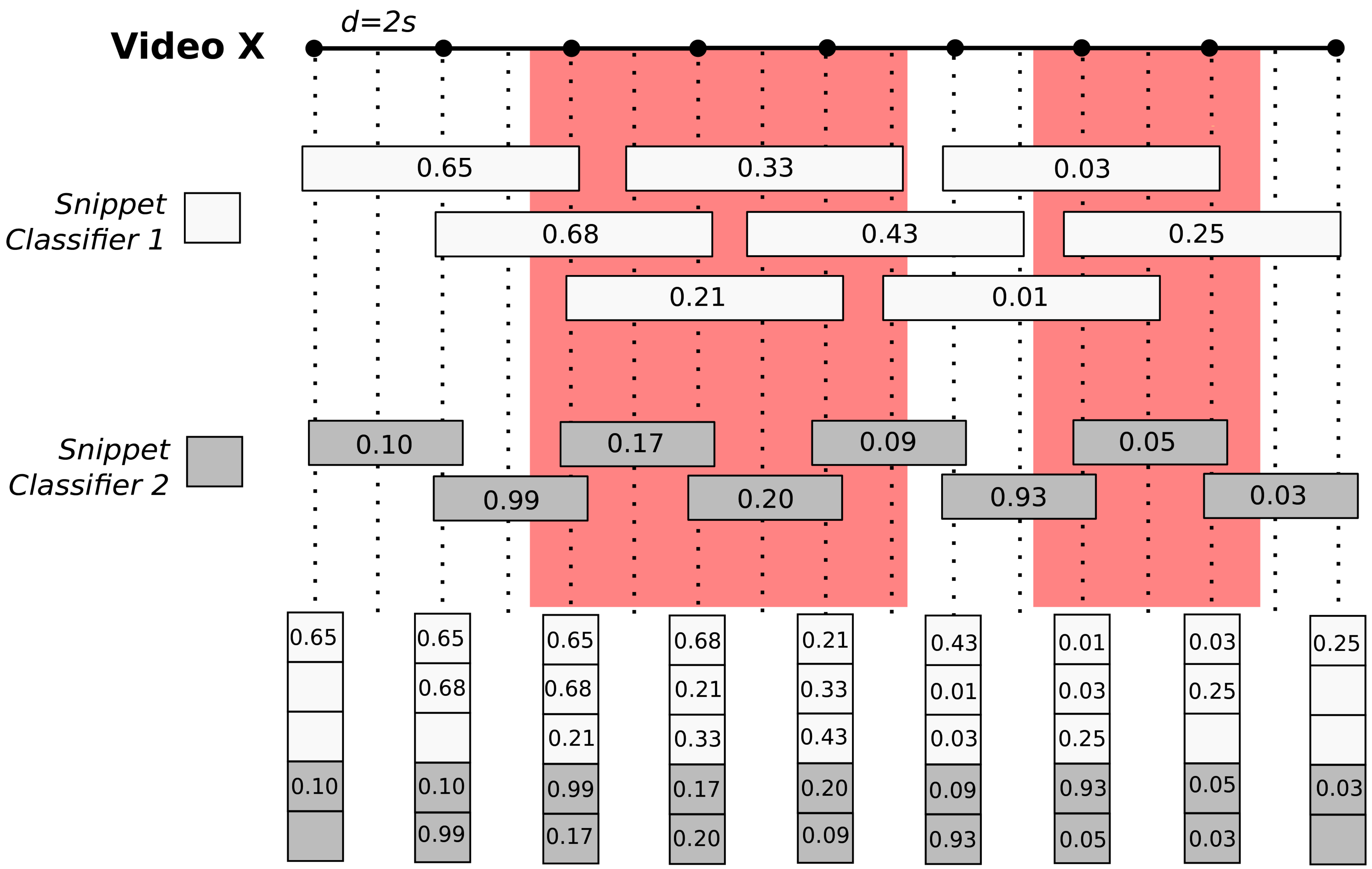
Proposed Solution

Snippet Classification



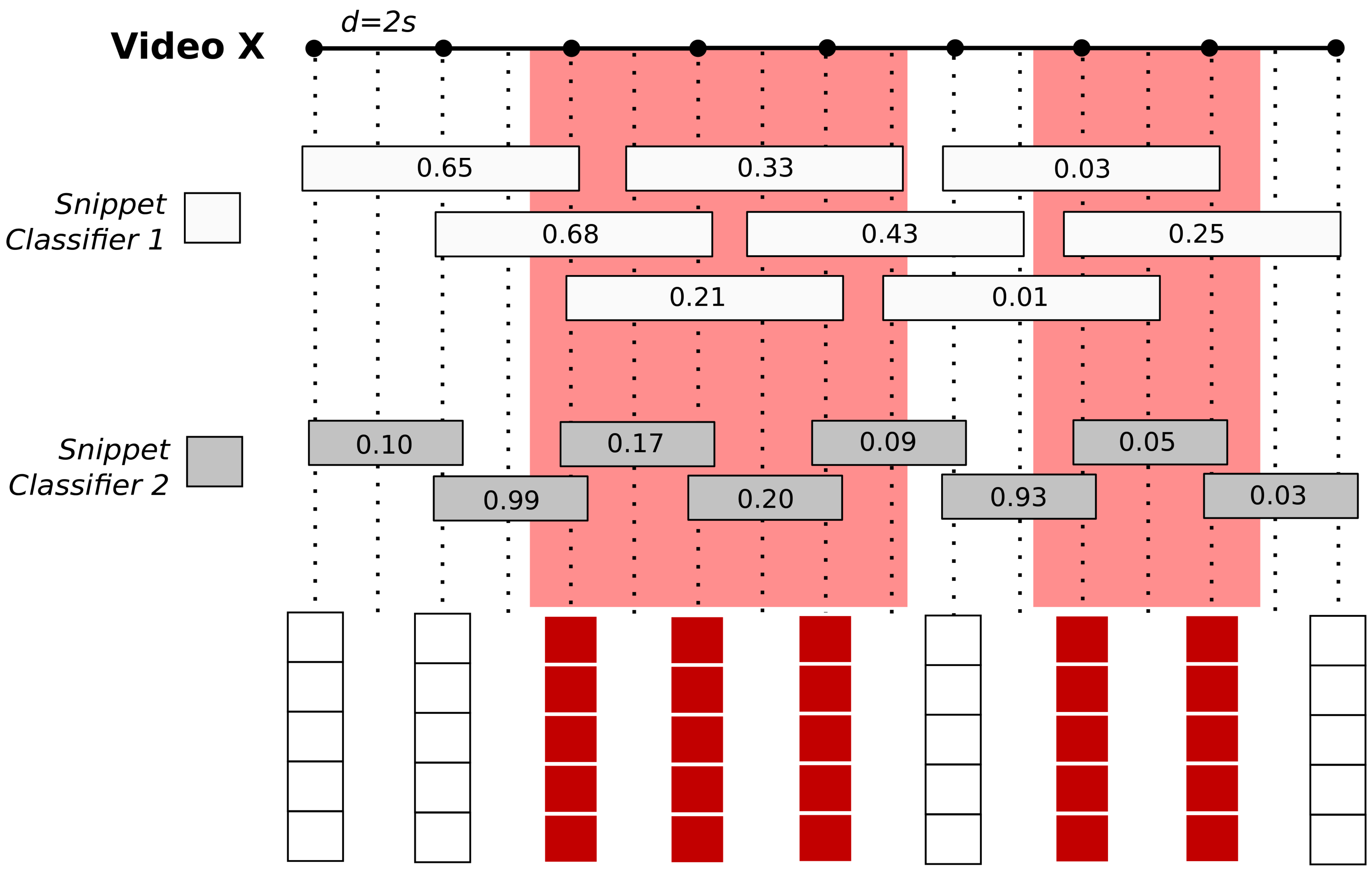
Proposed Solution

Late Fusion of Snippet Classifiers



Proposed Solution

Late Fusion of Snippet Classifiers

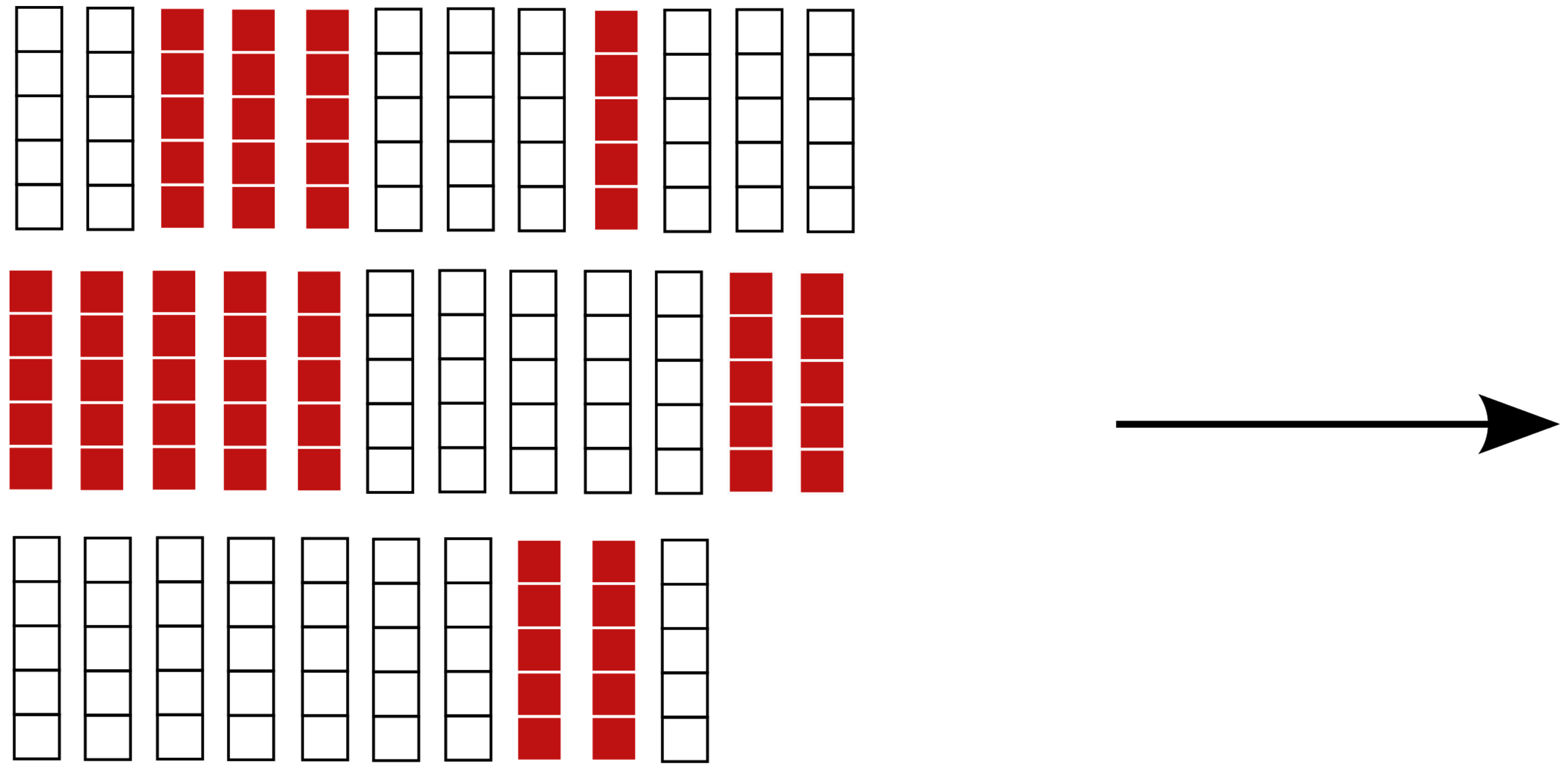


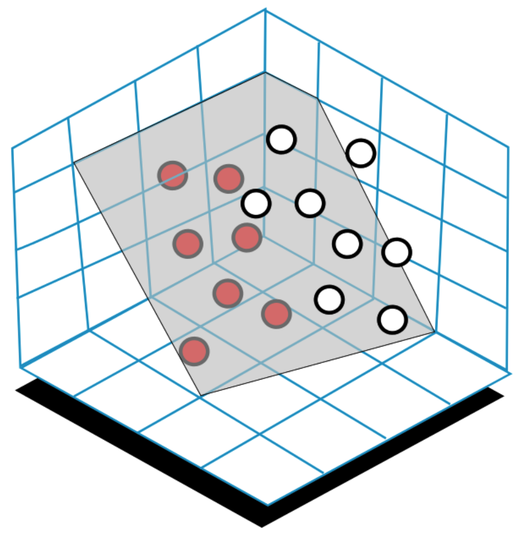
LOYOLA UNIVERSITY CHICAGO

Proposed Solution

Classification of Fusion Vectors

Training Time



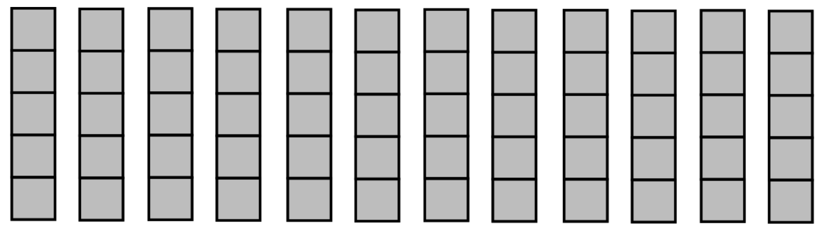


fusion classification model

Proposed Solution

Classification of Fusion Vectors

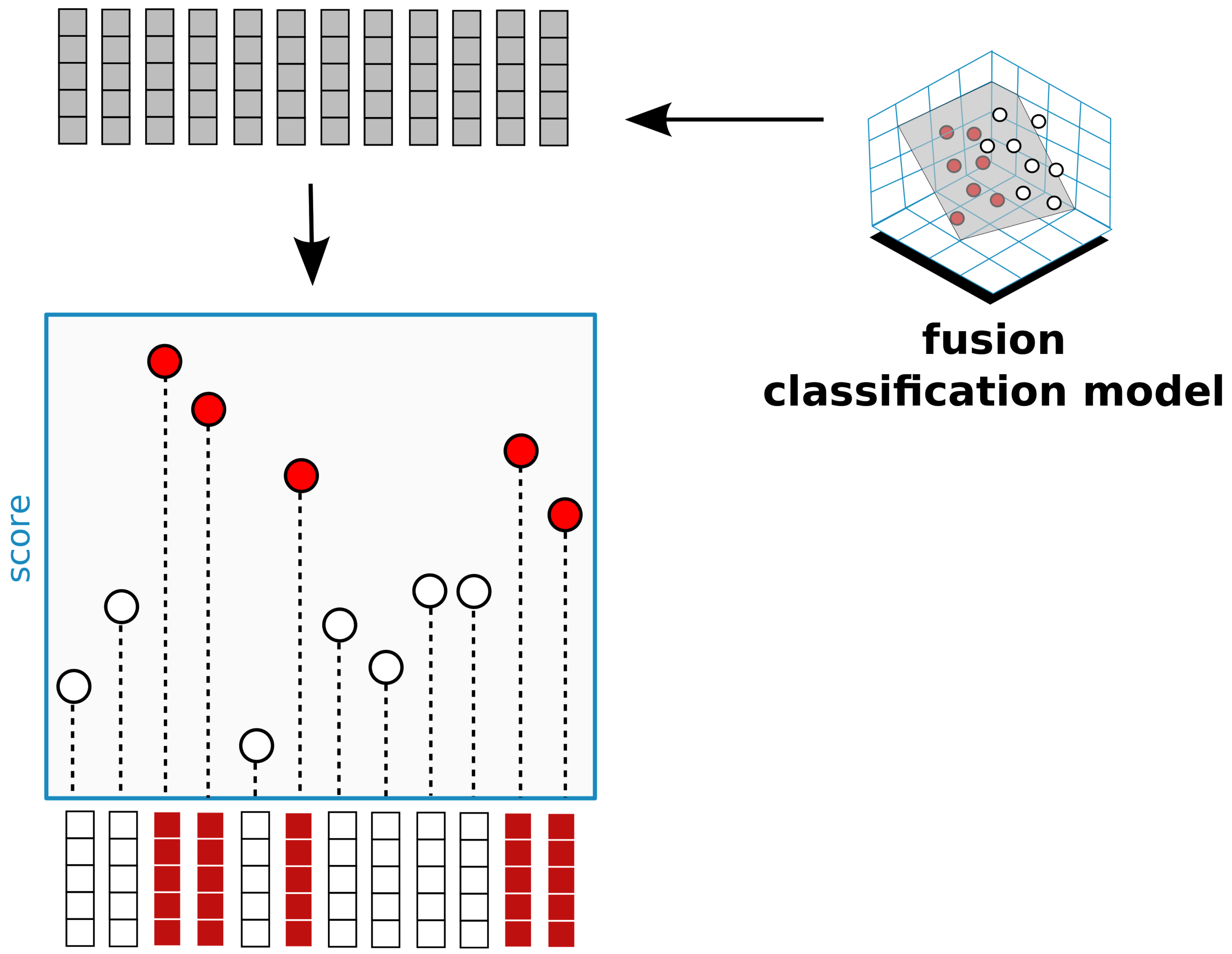
Inference Time



Proposed Solution

Classification of Fusion Vectors

Inference Time

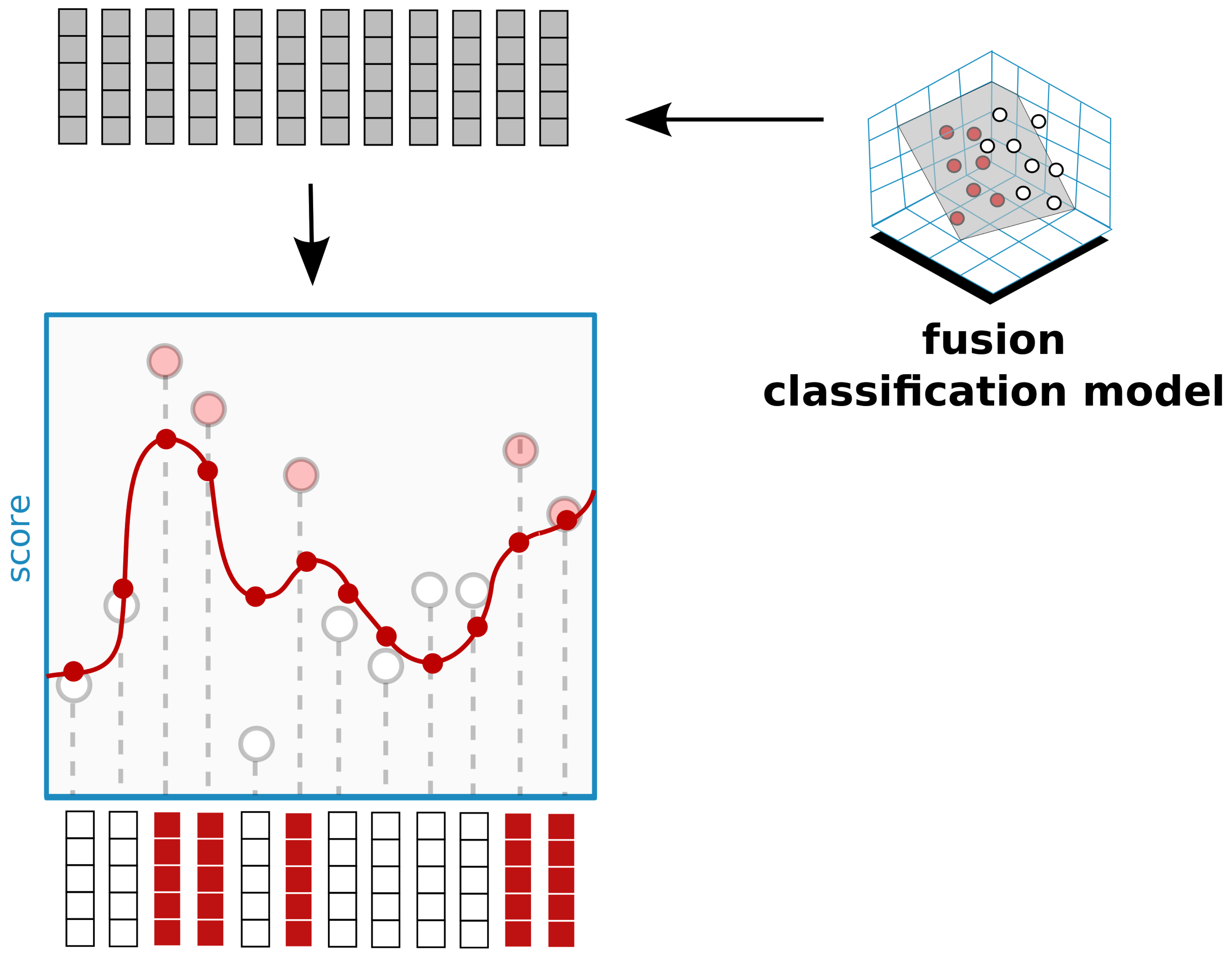


LOYOLA UNIVERSITY CHICAGO

Proposed Solution

Classification Score Smoothing

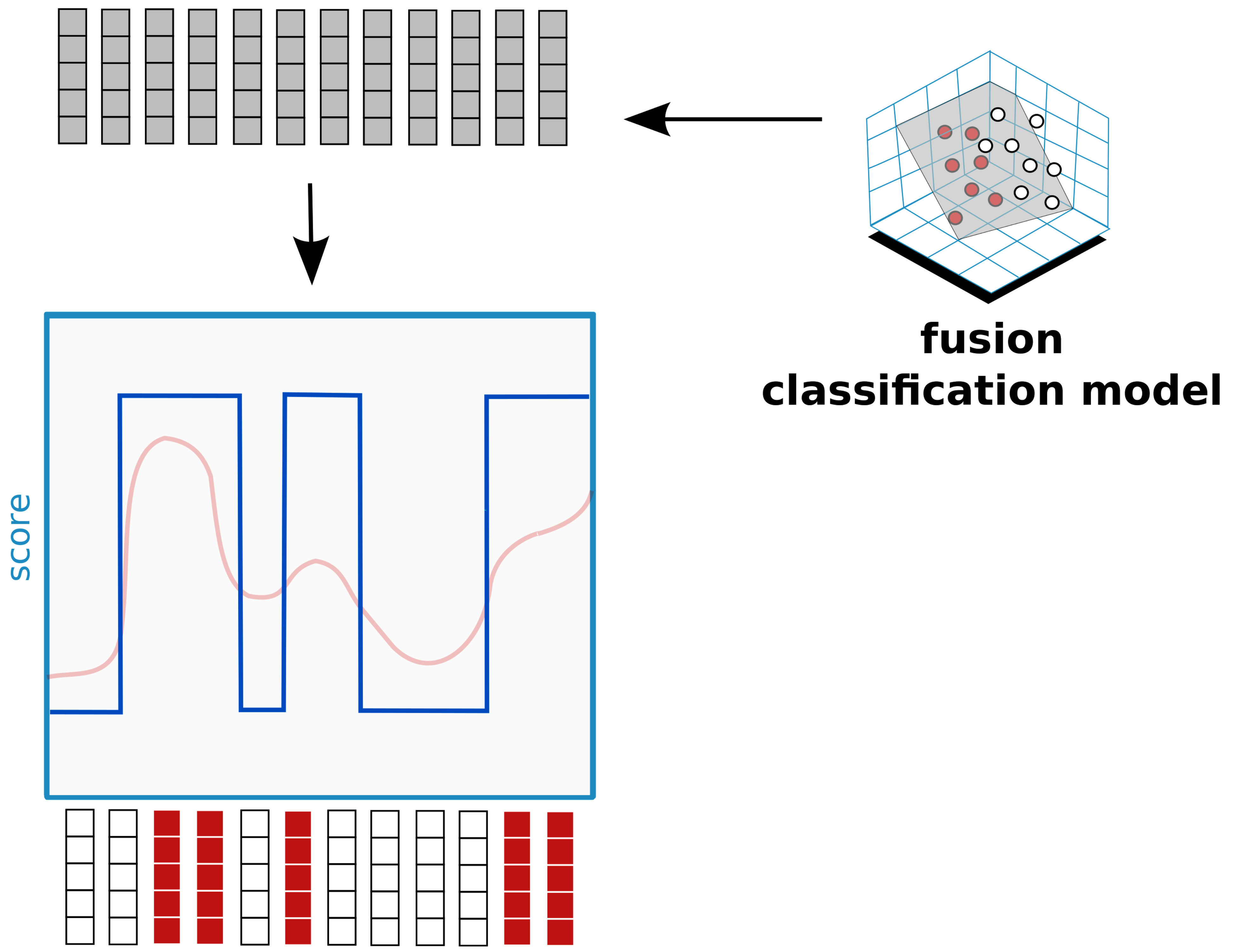
Inference Time



Proposed Solution

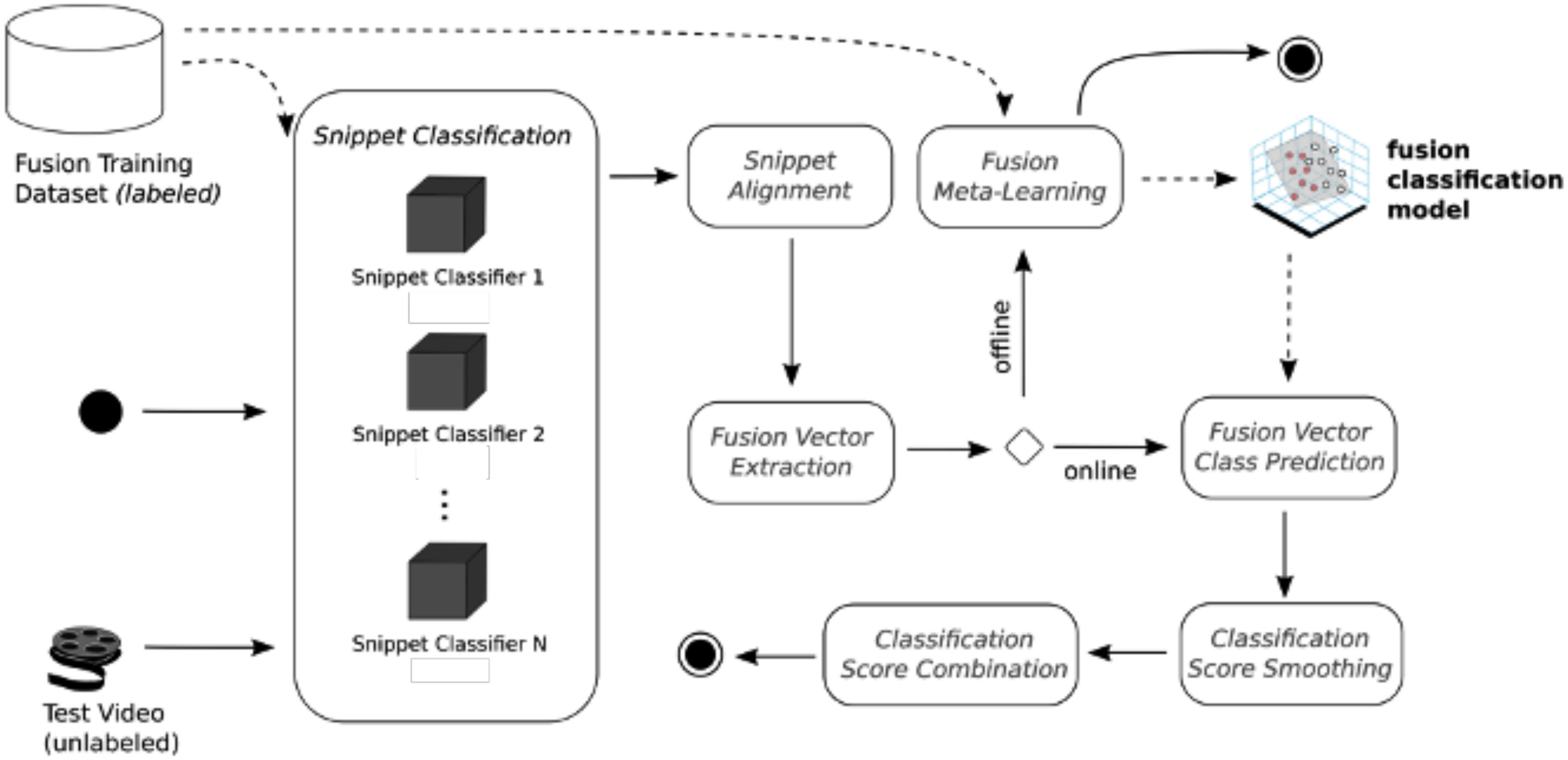
Classification Score Combination

Inference Time



Proposed Solution

Summary

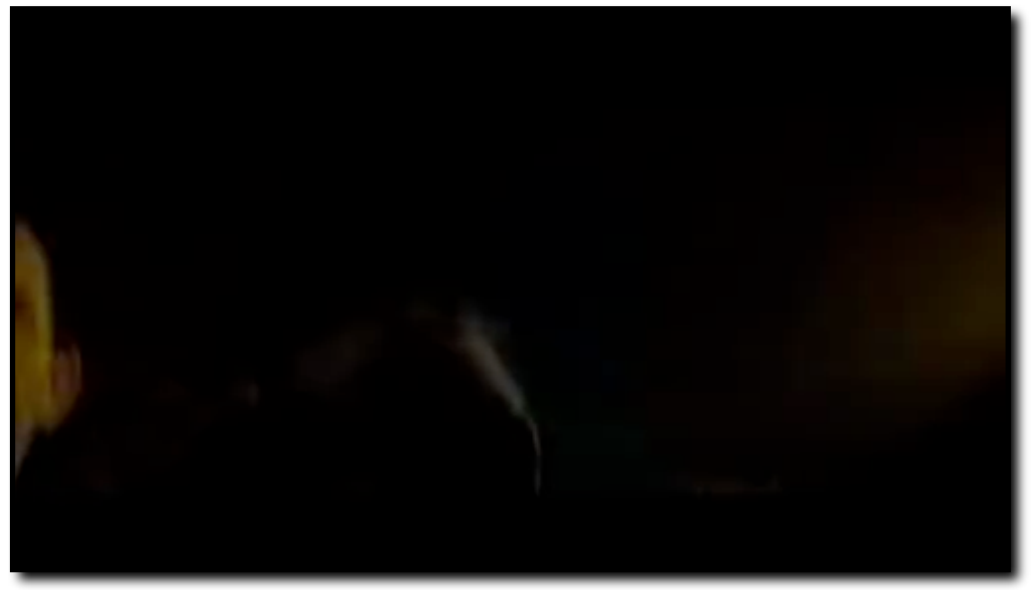


LOYOLA UNIVERSITY CHICAGO

Violence Results

Dataset
MediaEval 2014



"Content one would not let a child see." [2]

Training: 24 movies
Test: 7 movies

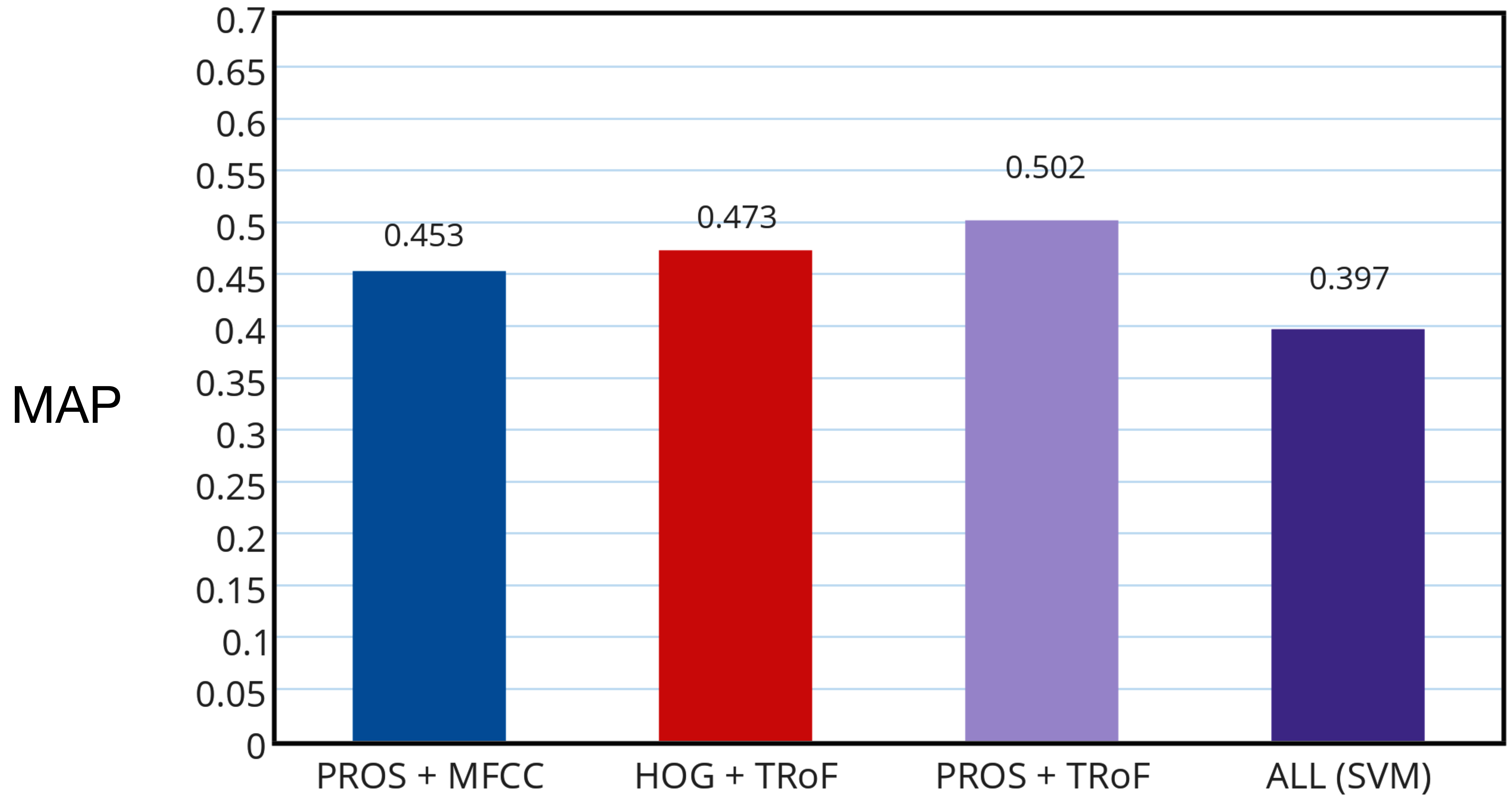
Frame-level annotation.

Metric: Mean Average Precision (MAP)

[2] Demarty et al., *Benchmarking Violent Scenes Detection in Movies*. In IEEE CBMI, 2014

Violence Results

Multimodal Fusion
(Audio + Video)



Audio
Prosodic features (PROS)
MFCC

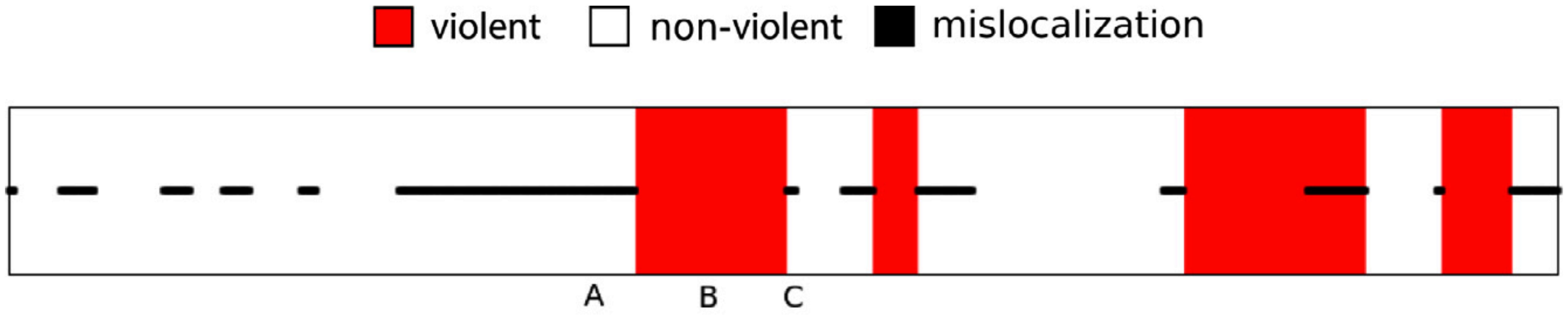
Video
HOG
TRoF

Fusion
SVM

LOYOLA UNIVERSITY CHICAGO

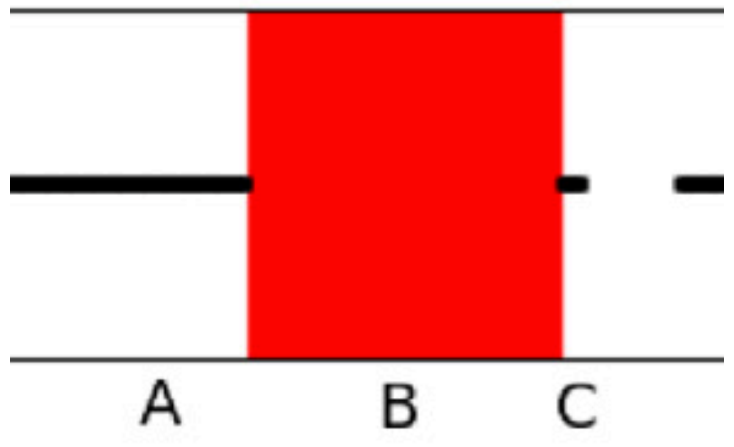
Violence Results

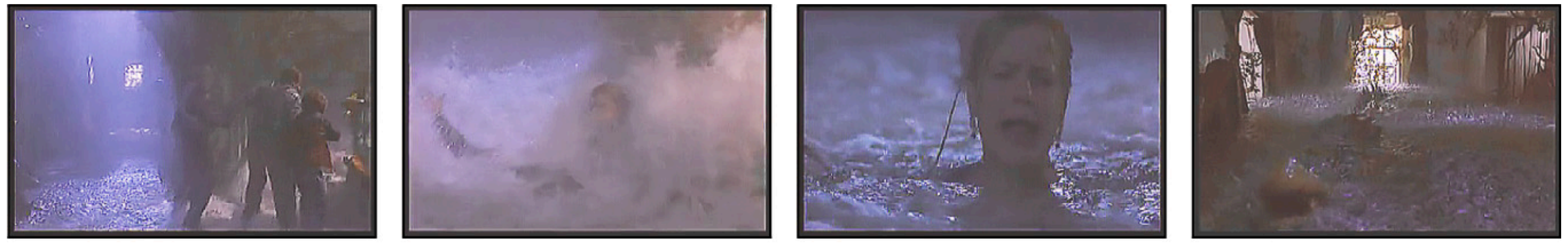
Qualitative Results



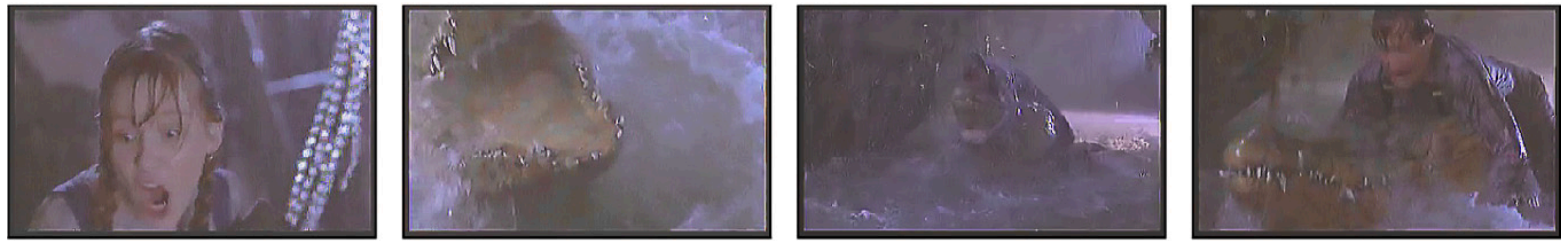
Violence Results

Qualitative Results





(a) A_1: non-violent (b) A_2: non-violent (c) A_3: non-violent (d) A_4: non-violent



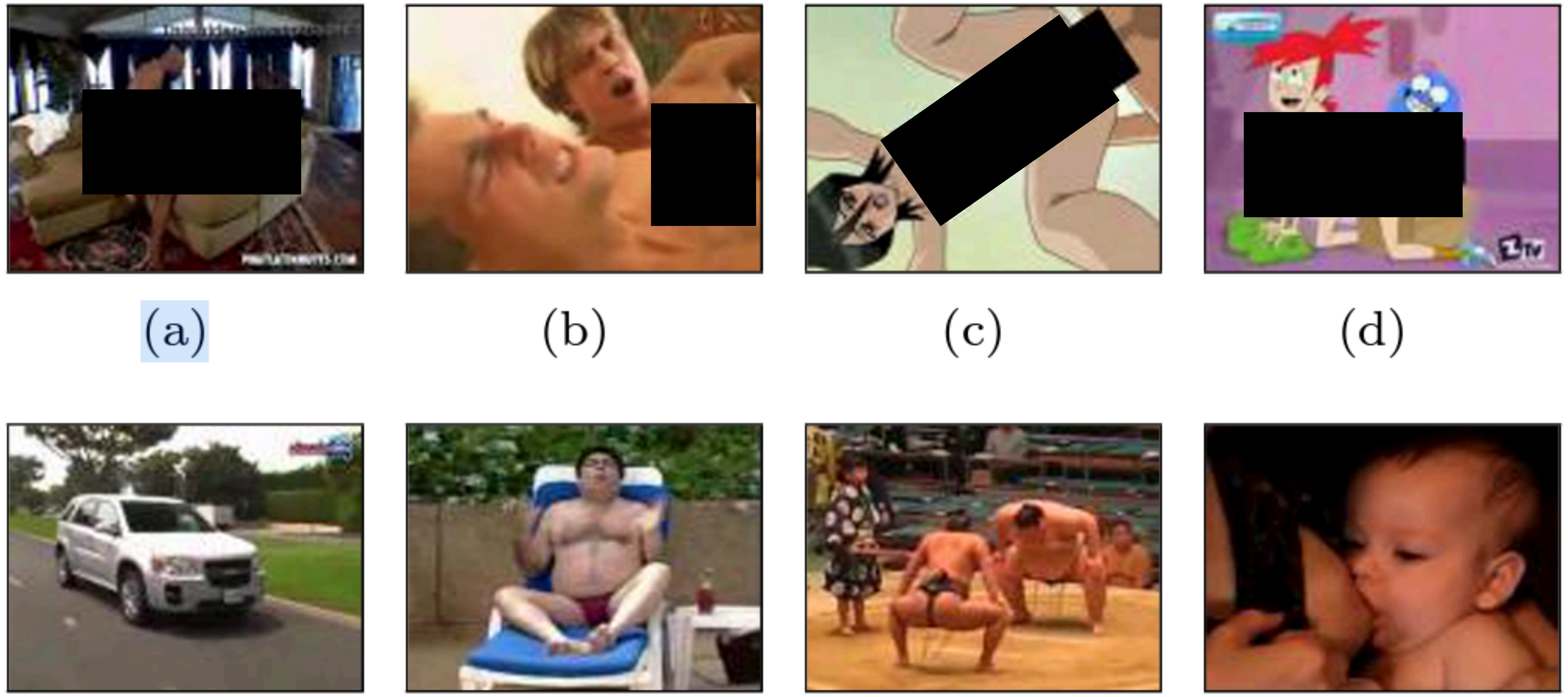
(e) B_1: violent (f) B_2: violent (g) B_3: violent (h) B_4: violent

(i) C_1: non-violent (j) C_2: non-violent (k) C_3: non-violent (l) C_4: non-violent

Pornography Results

Dataset

Porn-2k



(a) (b) (c) (d)

(e) (f) (g) (h)

"Any explicit sexual matter with the purpose of eliciting arousal." [1]

140h of video

Frame-level annotation

Metric: frame-level classification accuracy.

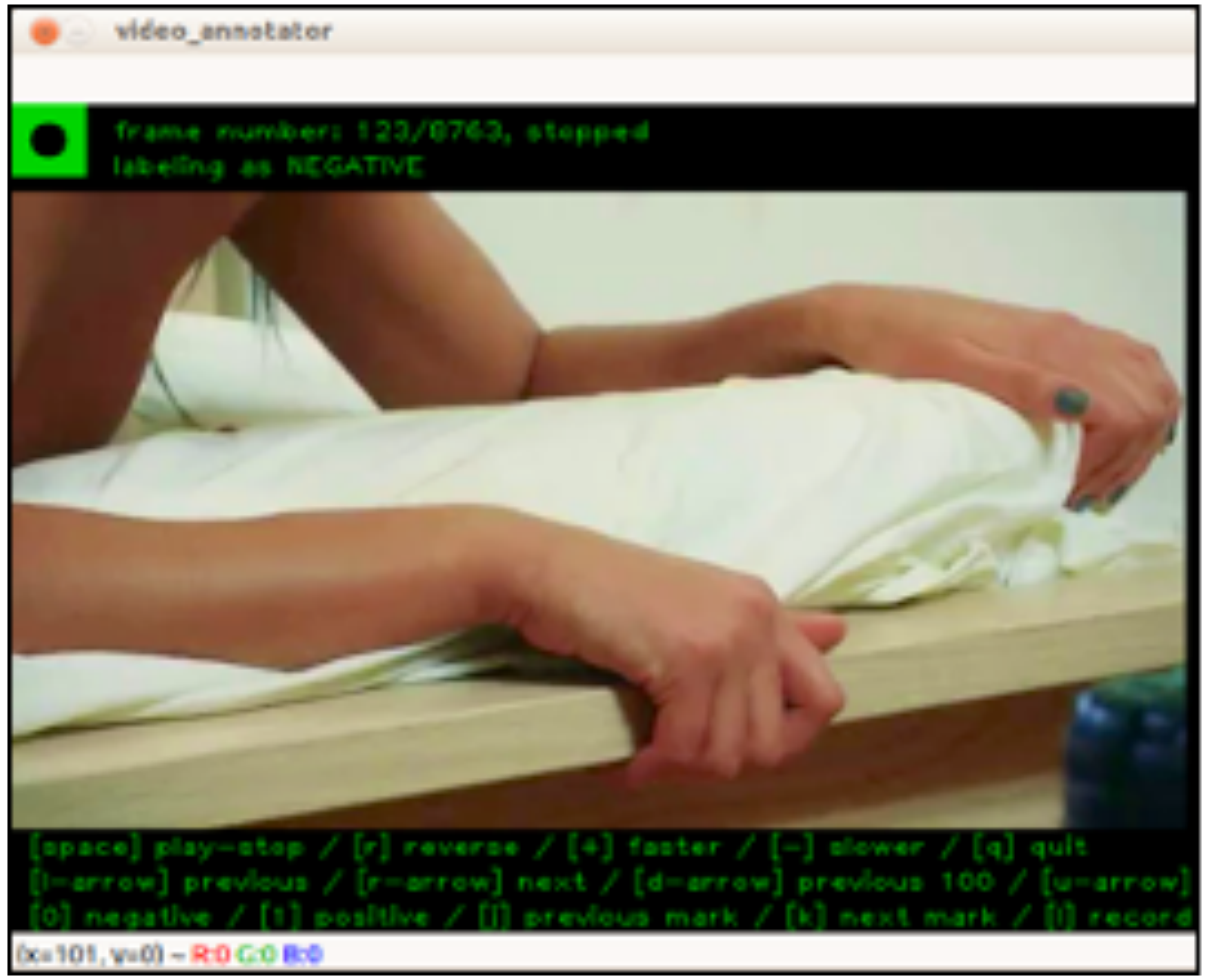
YouTube Vine

vimeo Porn sites

[1] Short et al., *A review of internet pornography use research: Methodology and content from the past 10 years.* Cyberpsychology, Behavior, and Social Networking 15, 2012

Pornography Results

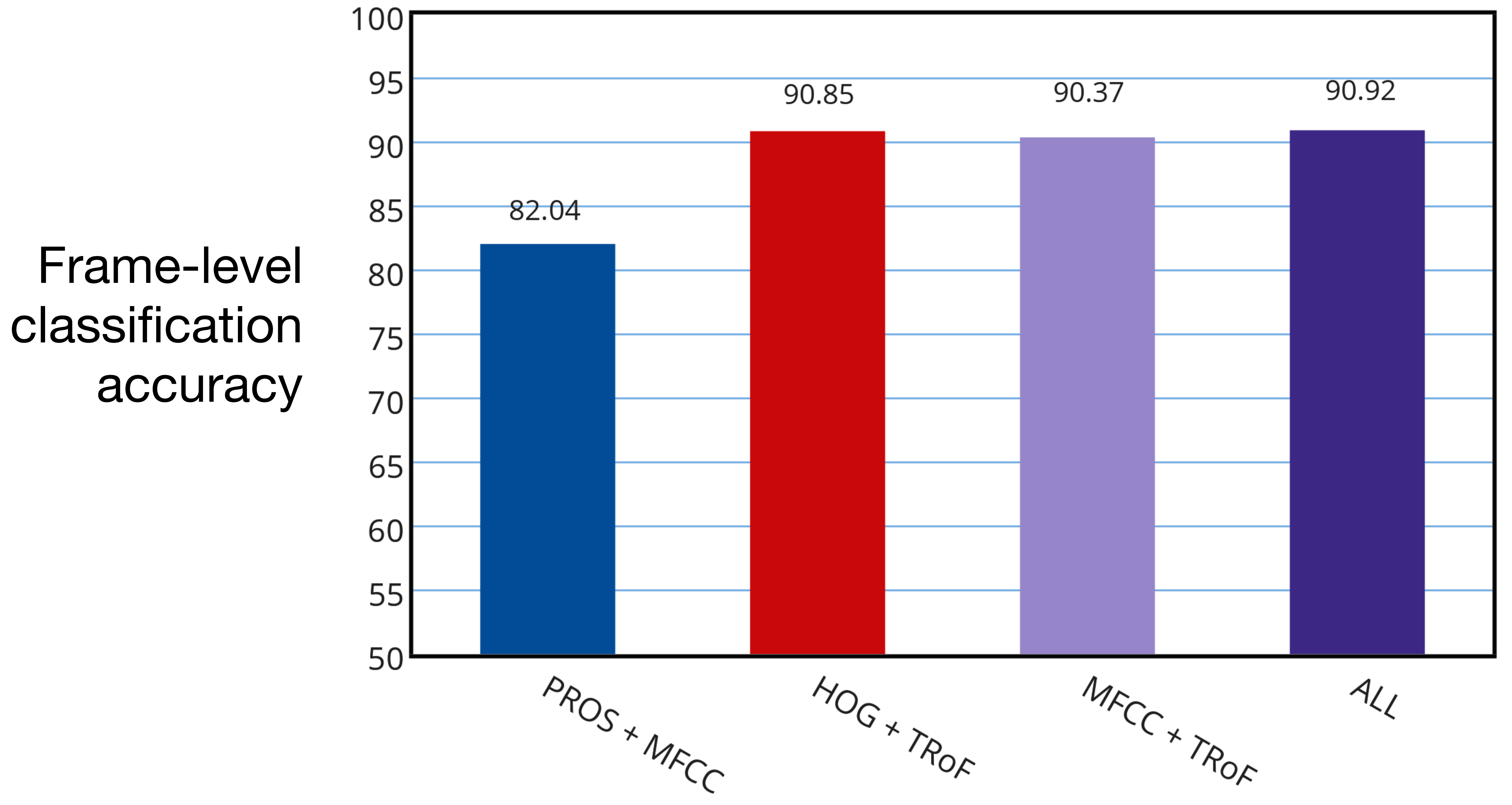
Dataset
Porn-2k



Frame-level annotation tool.

Pornography Results

Multimodal Fusion
(Audio + Video)



Audio
Prosodic features (PROS)
MFCC

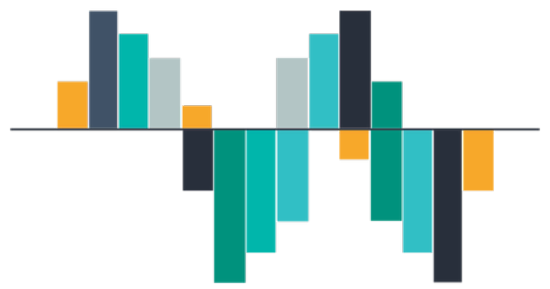
Video
HOG
TRoF

Fusion
SVM

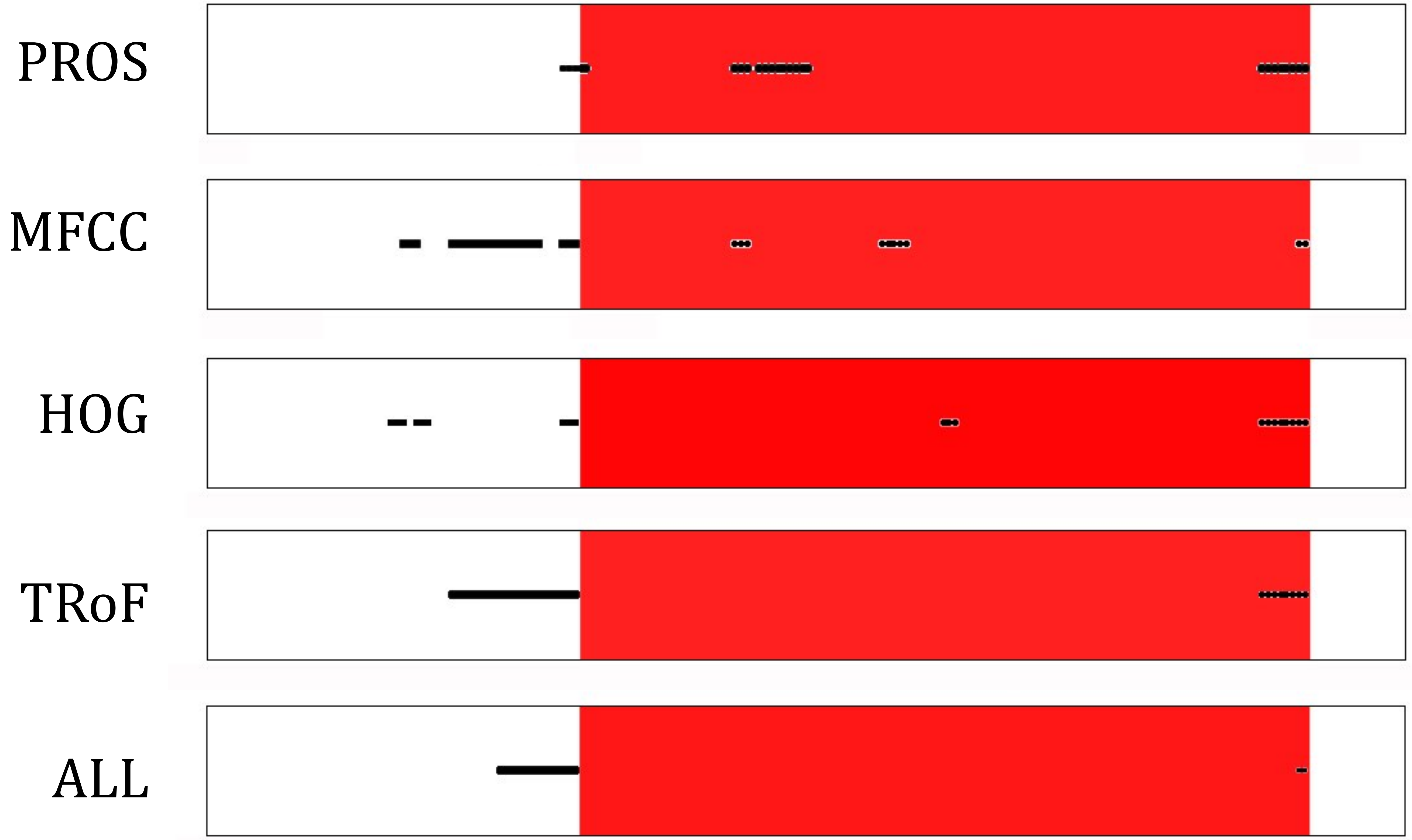
LOYOLA UNIVERSITY CHICAGO

Pornography Results

Qualitative Results



The solution misses 5 minutes
in every hour of pornographic content



Accomplishments


3 Journals

WACV 2017

3 Conference Papers

INTELLECTUAL PROPERTY PATENTED INTELLECTUAL PROPERTY

1 patent

Frame-level annotated porn video dataset

Violent Scenes Detection Competition

LOYOLA UNIVERSITY CHICAGO

SAMSUNG

RECOD
reasoning for complex data

UNICAMP

Future Work

Cryptography and Machine Learning
Can Machine Learning techniques be trained over sensitive encrypted data?

Advantages
Human intelligibility is destroyed by encryption.

Applications
Child pornography detection
and other sensitive data.

Hint
https://bit.ly/2YGEOmD

Encrypted Deep Learning Training and Predictions with TF Encrypted Keras

LOYOLA UNIVERSITY CHICAGO