A C3D-based Convolutional Neural Network for Frame Dropping Detection in a Single Video Shot


Chengjiang Long


Eric Smith


Arslan Basharat


Anthony Hoogs

Kitware Inc.
28 Corporate Drive, Clifton Park, NY 12065

CVPRW'2017 on Media Forensics

Motivation

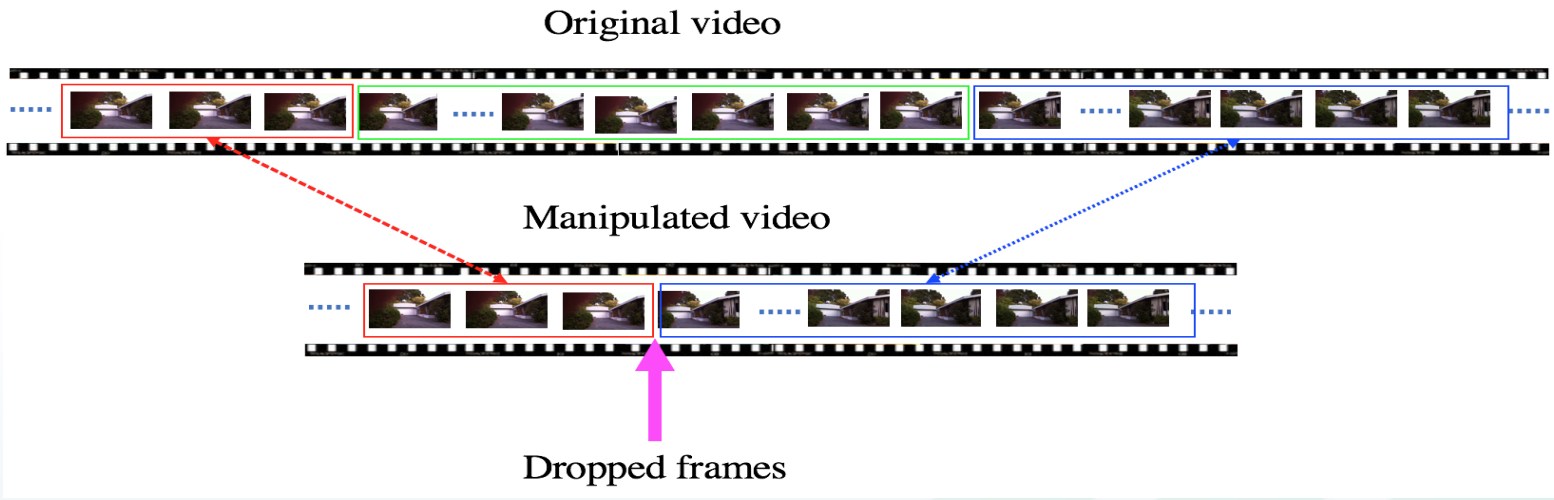
- Automatically determine if a video has temporal manipulations using the motion of the scene or camera without a reference video.
- Temporal manipulations include dropping frames, duplicating/looping frames, and speeding up/slowing down.
- In this paper, we present algorithms for detecting frame drops.

Original video

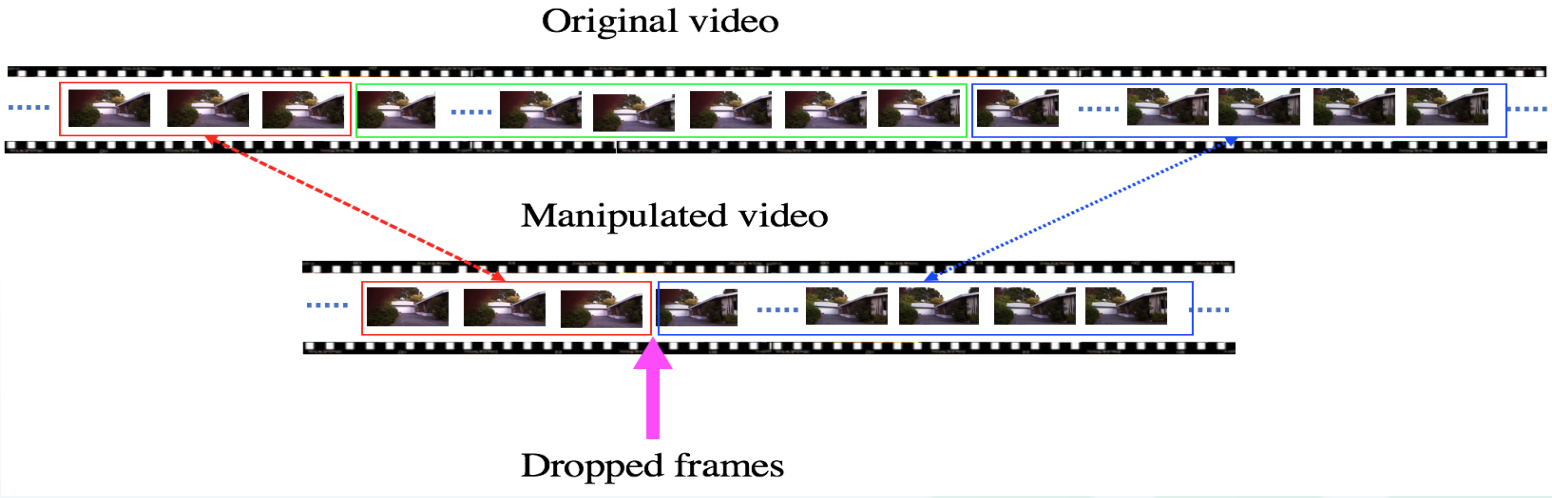
Manipulated video with frame drops, frame duplications and object removal.

From the UTMVT dataset for the benchmark of video forgery techniques

Dropped frame detection



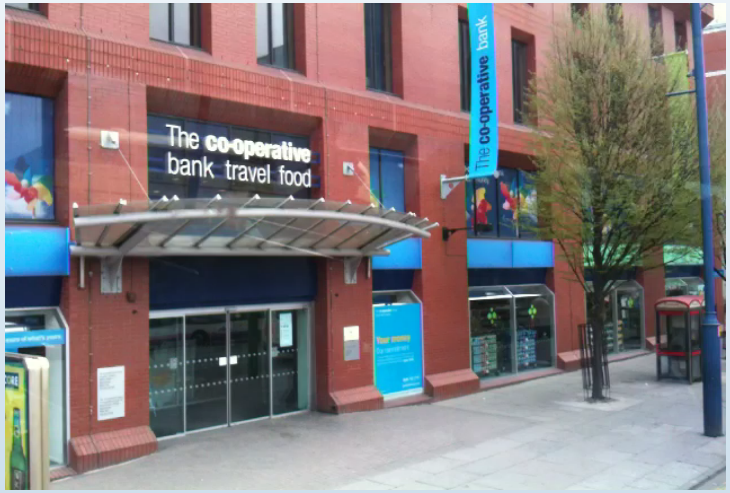
Dropped frame detection



- Frame drop is defined as removal of any number of consecutive frames within a shot.
- Currently focus on single shot videos
  - Surveillance, body camera, mobile phone, dash cam.

Frame Drop Detection Challenges

- The number of body camera, dash camera and surveillance videos are increasing every day.
- Distinguish between frame drops and fast/rapid motion of scene elements or camera motion. Must also be sensitive to rapid changes in static camera videos.
- To be effective, algorithms must be useful at an extremely low false positive operating point (1% FA is still 18 FP/min at 30fps).



Original video – moving camera

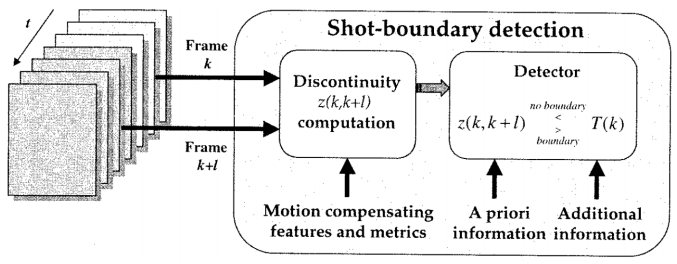
Original video – static camera

Our contributions

- ✓ Propose a 3D convolutional network for frame dropping detection + refine the confidence score with peak detection trick and a scale term.
- ✓ Develop a series of baselines including cue-based and learning-based methods.
- ✓ Evaluate the performance of frame dropping detection on YFCC100m dataset and Nimble Challenge 2017 dataset.

Related work

- **Detecting dropped frames in videos.**
  - Cue-based approaches. E.g. motion energy cue [S.Wolf 2009].
  - Learning-based approaches. E.g. SVM to classify tampered or non-tampered video, tampered or non-tampered frames [Thakur 2016].
- **Identifying Inter-frame Forgery in videos.**
  - Optical flow [Qi Wang 2014] [Juan Chao 2013]
  - Consistency of Correlation Coefficients of Gray Values[Qi Wang 2013]
- **Detecting shot boundary.**
  - [Alan F. Smeatona 2010] video shot boundary detection.
  - TRECVID: http://trecvid.nist.gov/



Shot-boundary detector framework [Hanjalic 2002]

Cue-based algorithms

Frame dropping

3 consecutive frames

Cue-based algorithms

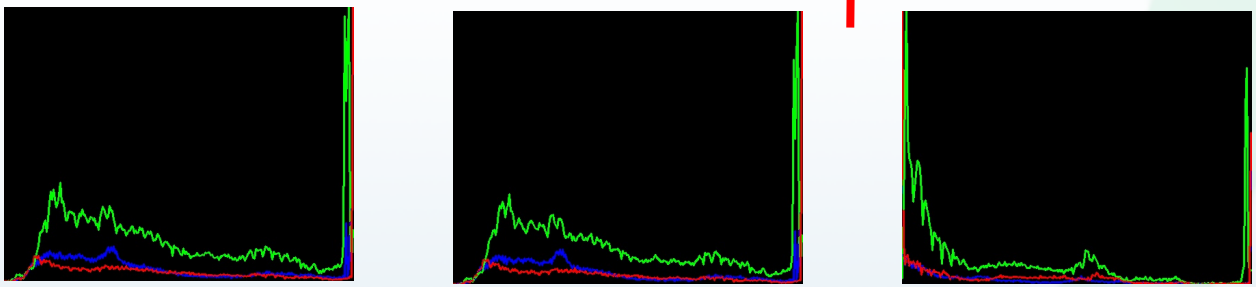
3 consecutive frames

Color histogram

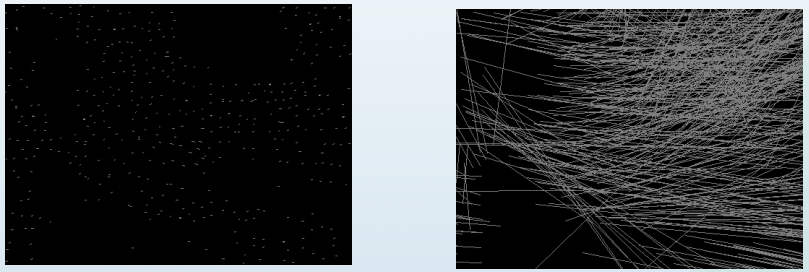
Cue-based algorithms

Frame dropping

3 consecutive frames



Color histogram



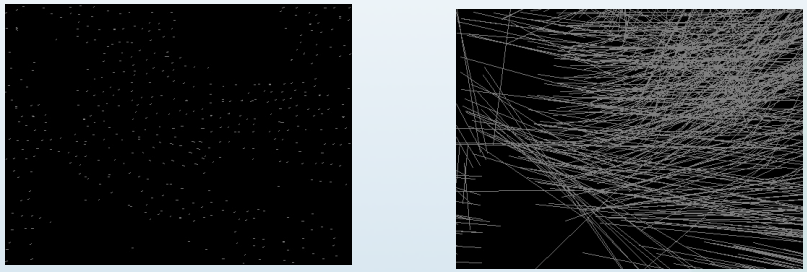
Optical flow

Cue-based algorithms

Frame dropping

3 consecutive frames

Color histogram



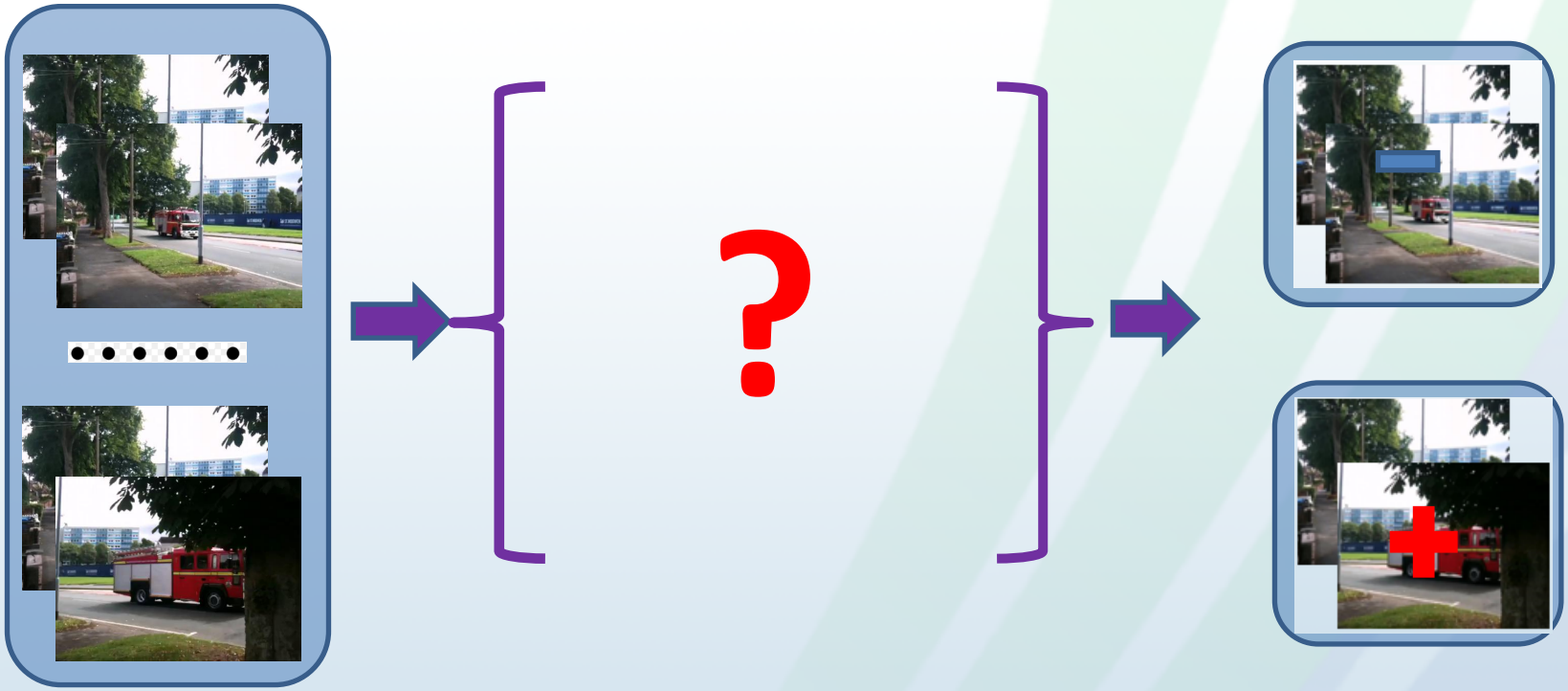
Optical flow

Motion energy

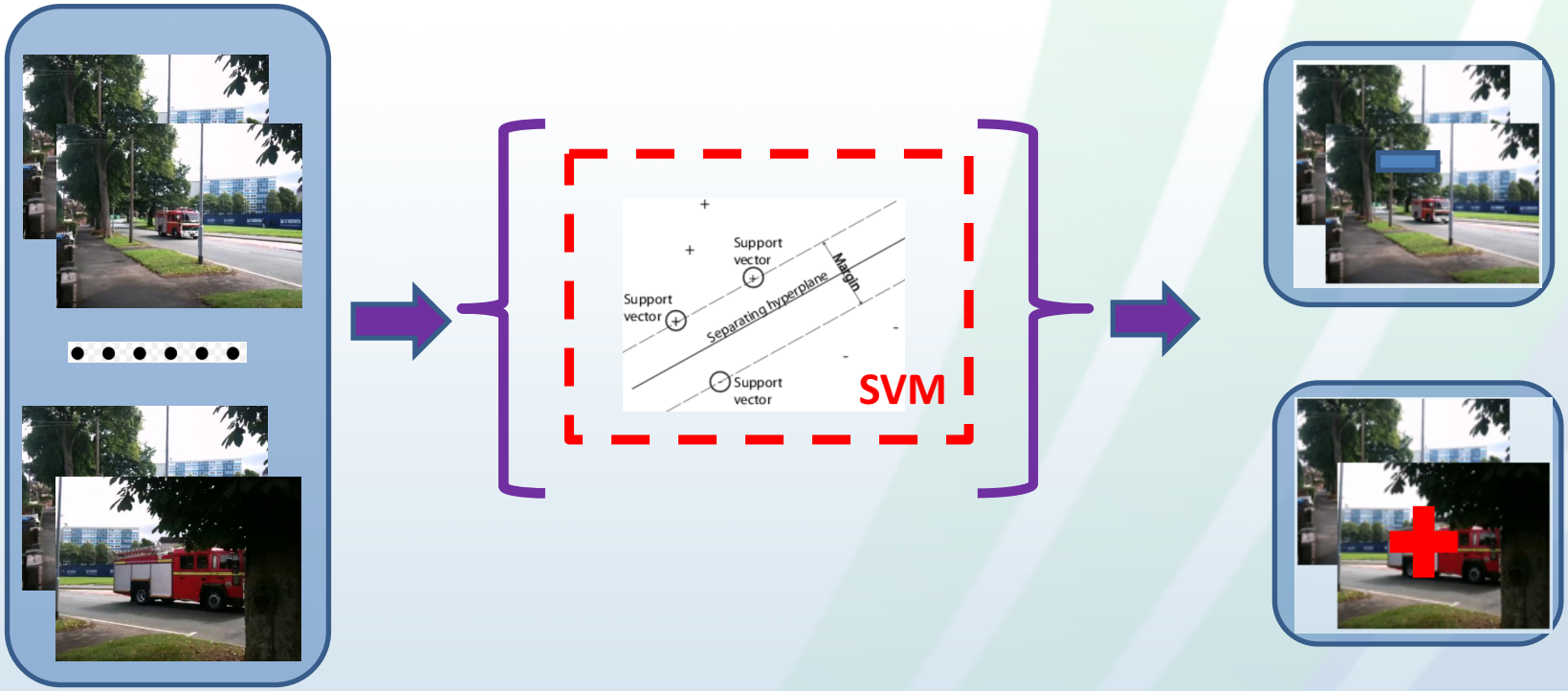
[S. Wolf, 2009]

Kitware

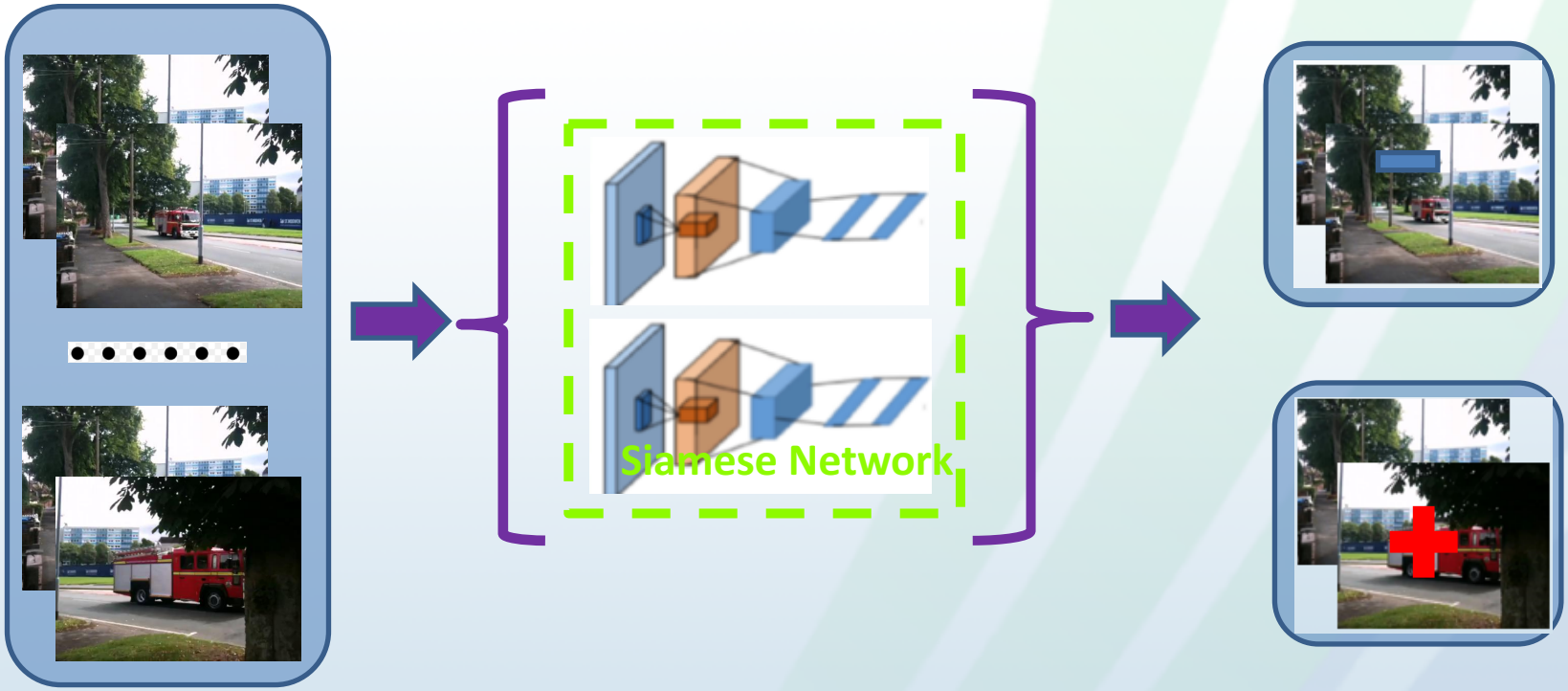
Learning-based algorithms



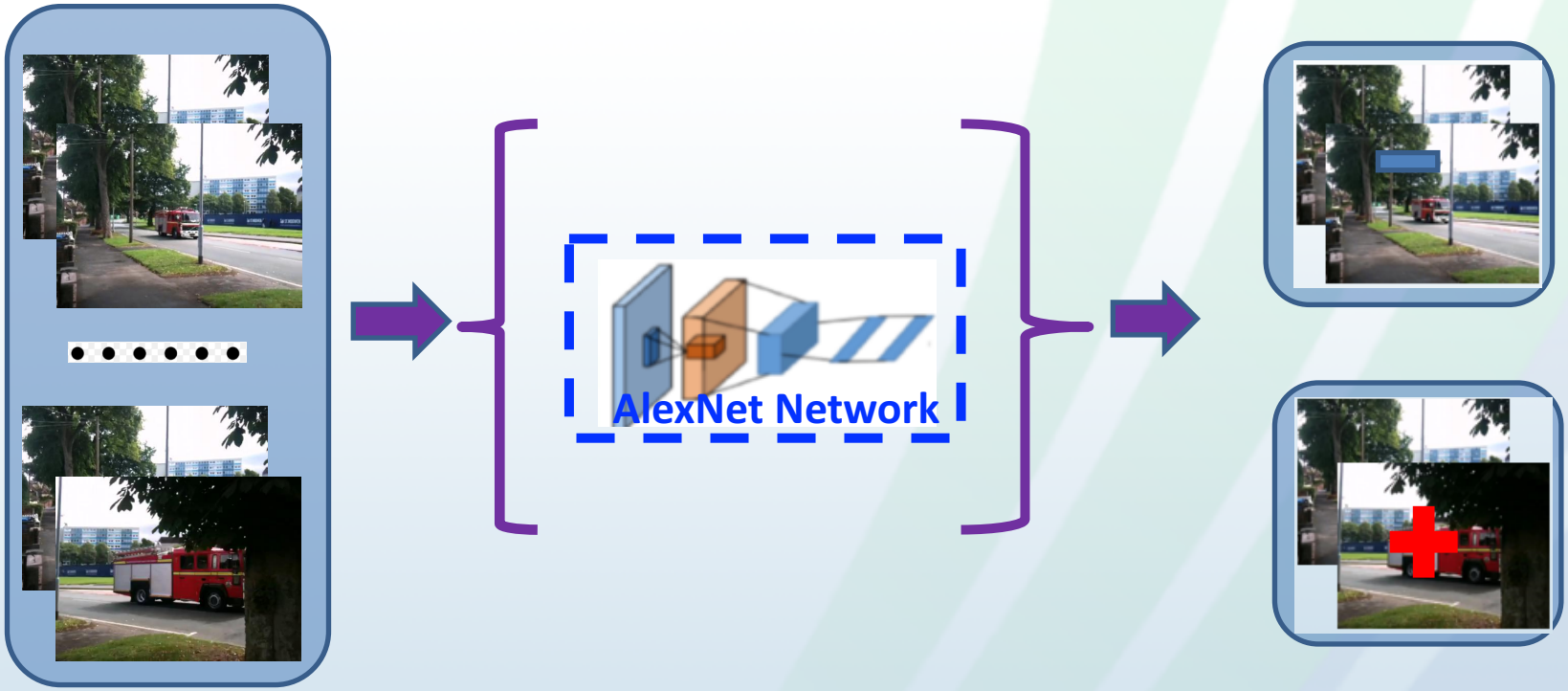
Learning-based algorithms (1)



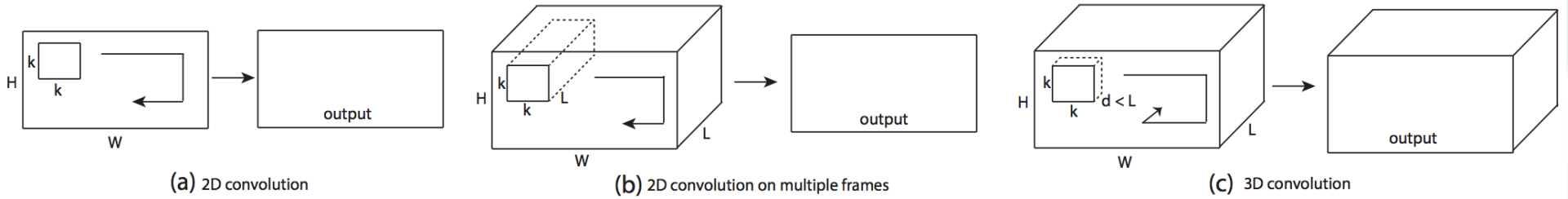
Learning-based algorithms (2)

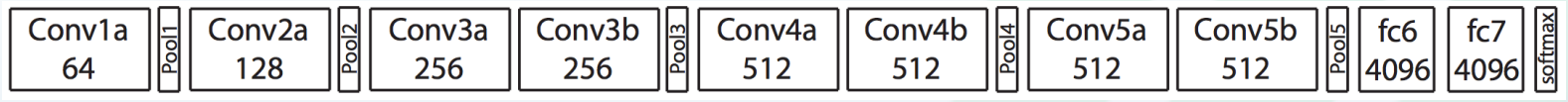


Learning-based algorithms (3)



Our C3D-based approach: motivation

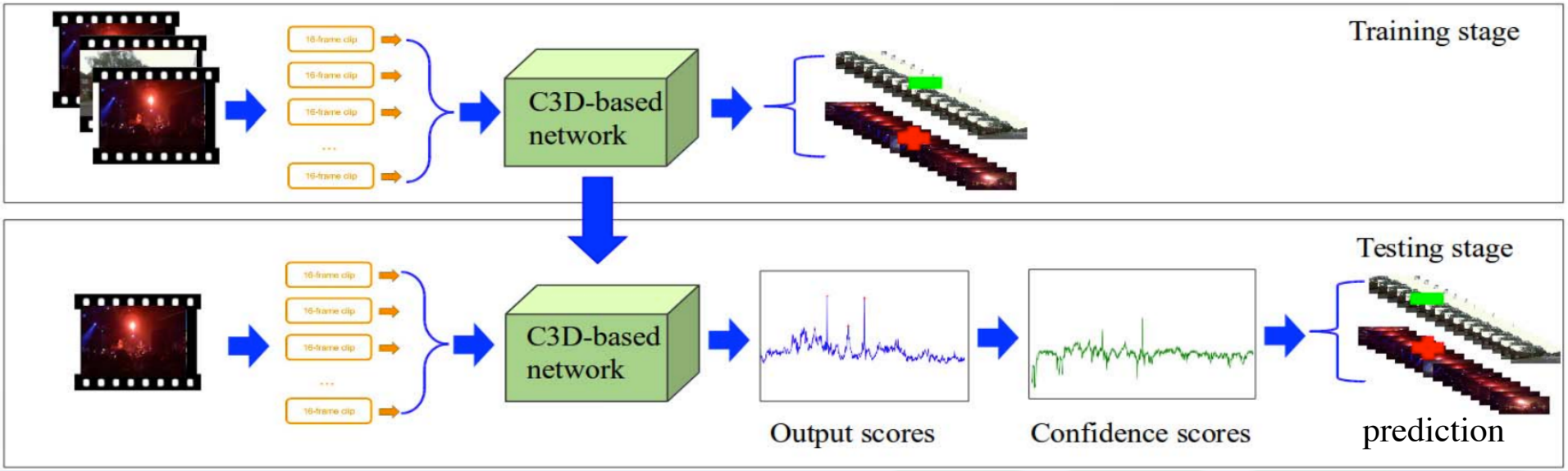


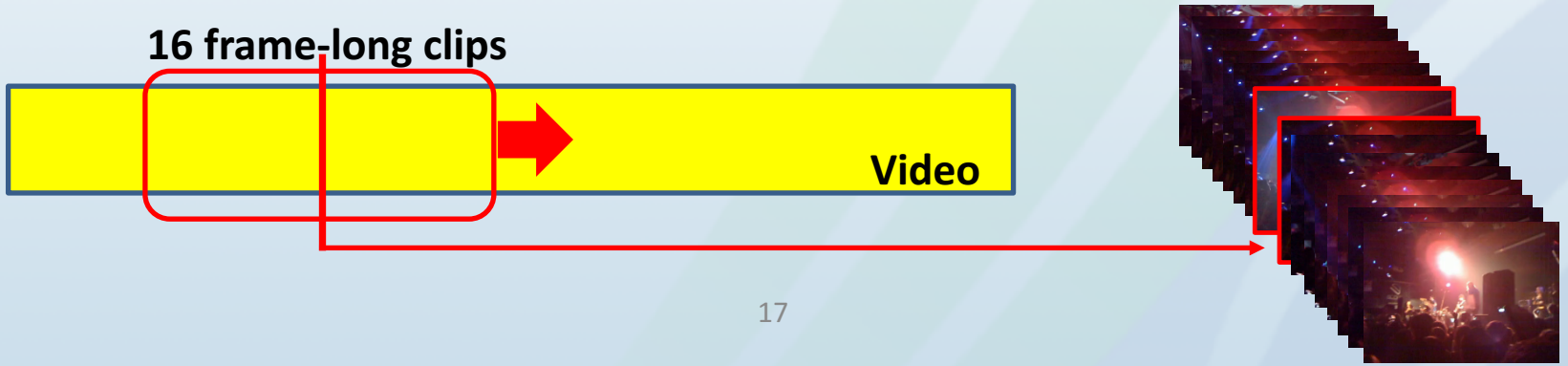


[Du Tran et al. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proc. ICCV 2015.]

- The C3D network is a 3D convolutional network using 3x3x3 convolution kernels and 66 million parameters originally designed for action recognition.
  - This architecture allows us to make use of motion cues due to convolution across the temporal dimension.

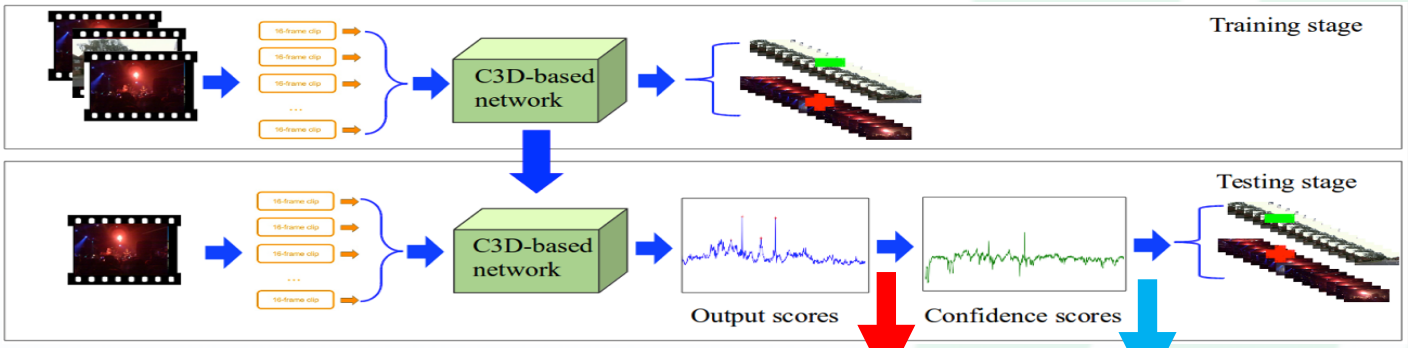
Our C3D-based approach: pipeline





Kitware

Our C3D-based approach: confidence score



$$f_{conf}(i) = \begin{cases} f(i) - \lambda\Delta(i) & \text{when } n_p < 2 \\ \frac{f(i)}{n_p} - \lambda\Delta(i) & \text{otherwise} \end{cases} \quad (1)$$

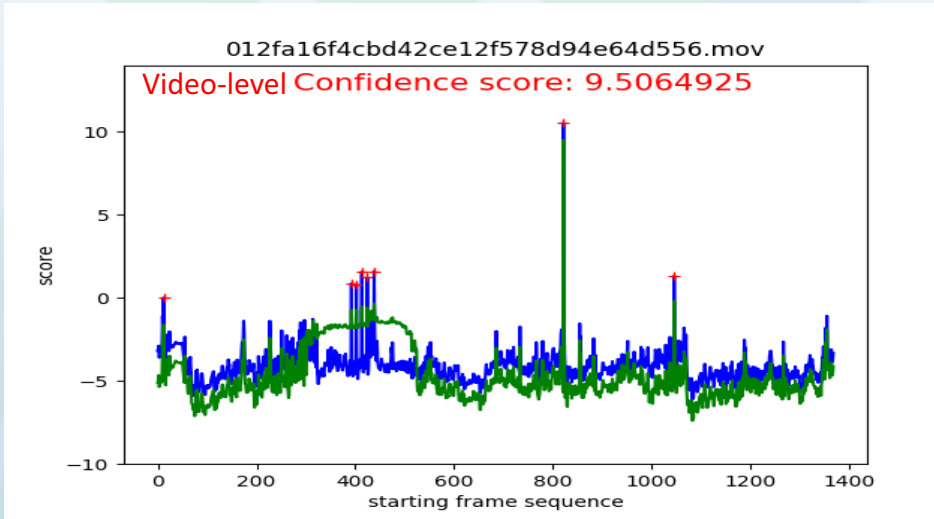
where

$$\Delta(i) = \underset{k\in W(i)}{\text{median}}\, f(k) - \min_{k\in W(i)} f(k) \quad (2)$$

$$W(i) = \{i - \frac{w}{2}, \ldots, i + \frac{w}{2}\}. \quad (3)$$

$$\max_i f_{conf}(i)$$

Video-level score

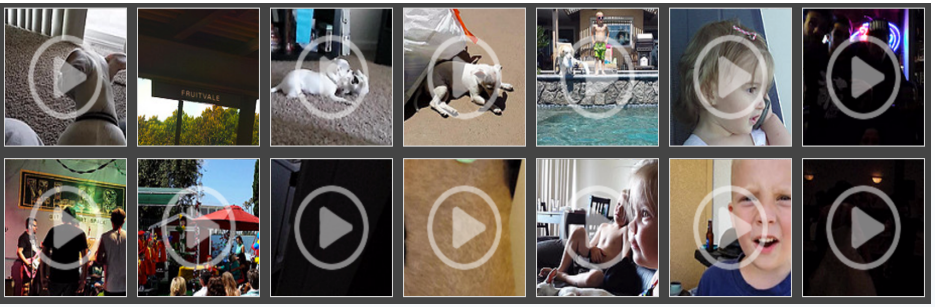


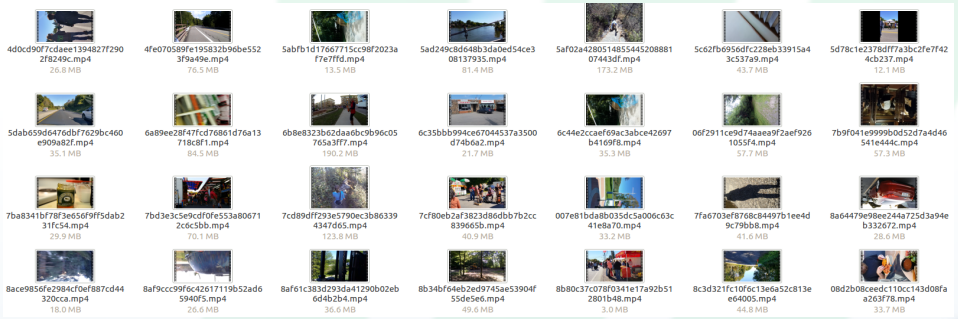
Dataset-training and validation

- 2,394 iPhone 4 videos downloaded from Medifor RankOne Browser.
- We removed videos with length smaller than 1 min or longer than 3 mins.

Dataset	# of original videos	# of Manipulated videos	Source
iPhone 4 video Dataset	314 (264 for training, 50 for validation)	15,700 = 314 x5x10 (drop duration: 0.5, 1, 2, 5, 10 seconds)	MediFor Wrold Dataset

Test datasets





Dataset	Raw video #	Manipulated video#	Source
YFCC100cm Dataset	53	53 x5x10 = 2650 (drop durationa:0.5, 1, 2, 5, 10 s.). Frame-level ground-truth information is available.	Yahoo Flickr Creative Commons 100 Million (YFCC100m) dataset (http://www.yfcc100m.org)
NC17-Dev2-Beta1	209	No information available for drop durations for the videos, and has only video-level ground-truth information	Nimble Challenge 2017 Evaluation sub-dataset (https://www.nist.gov/itl/iad/mig/nimble-challenge-2017-evaluation)

Baselines

Method	Brief description	Learning?
Color histogram	RGB 3 channel histograms + L2 distance.	No
Optical flow	The optic flow [11, 1] with Lucas-Kanade method + L2 distance.	No
Motion energy	Based on temporal information difference [13] sequence.	No
SVM	770-D feature vector (3x256-D RGB histogram + 2-D optic flow).	Yes
Pairwise Siamese Network	Siamese network architecture (2 conv layers + 3 fc layers + contrastive loss).	Yes
Triplet Siamese Network	Siamese network architecture (Alexnet-variant + Euclidean&contrastive loss).	Yes
Alexnet [3] Network	Alexnet-variant network architecture.	Yes

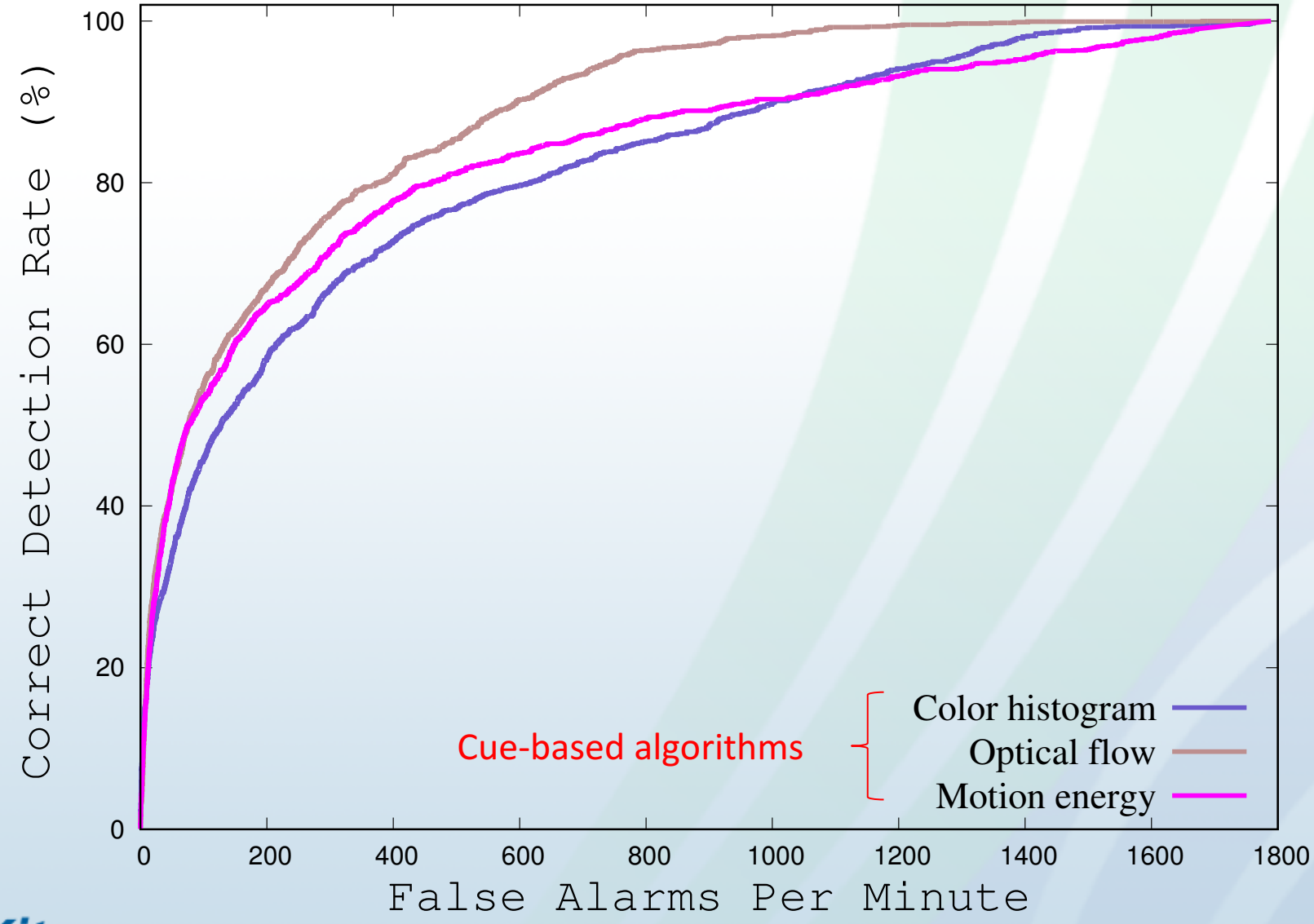
[1] J. Chao, et al. A novel video inter-frame forgery model detection scheme based on optical flow consistency. In *International Workshop on Digital Watermarking*, 2012.

[3] A. Krizhevsky, et al. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[11] Q. Wang, et al. Video inter-frame forgery identification based on optical flow consistency. *Sensors & Transducers*, 2014.

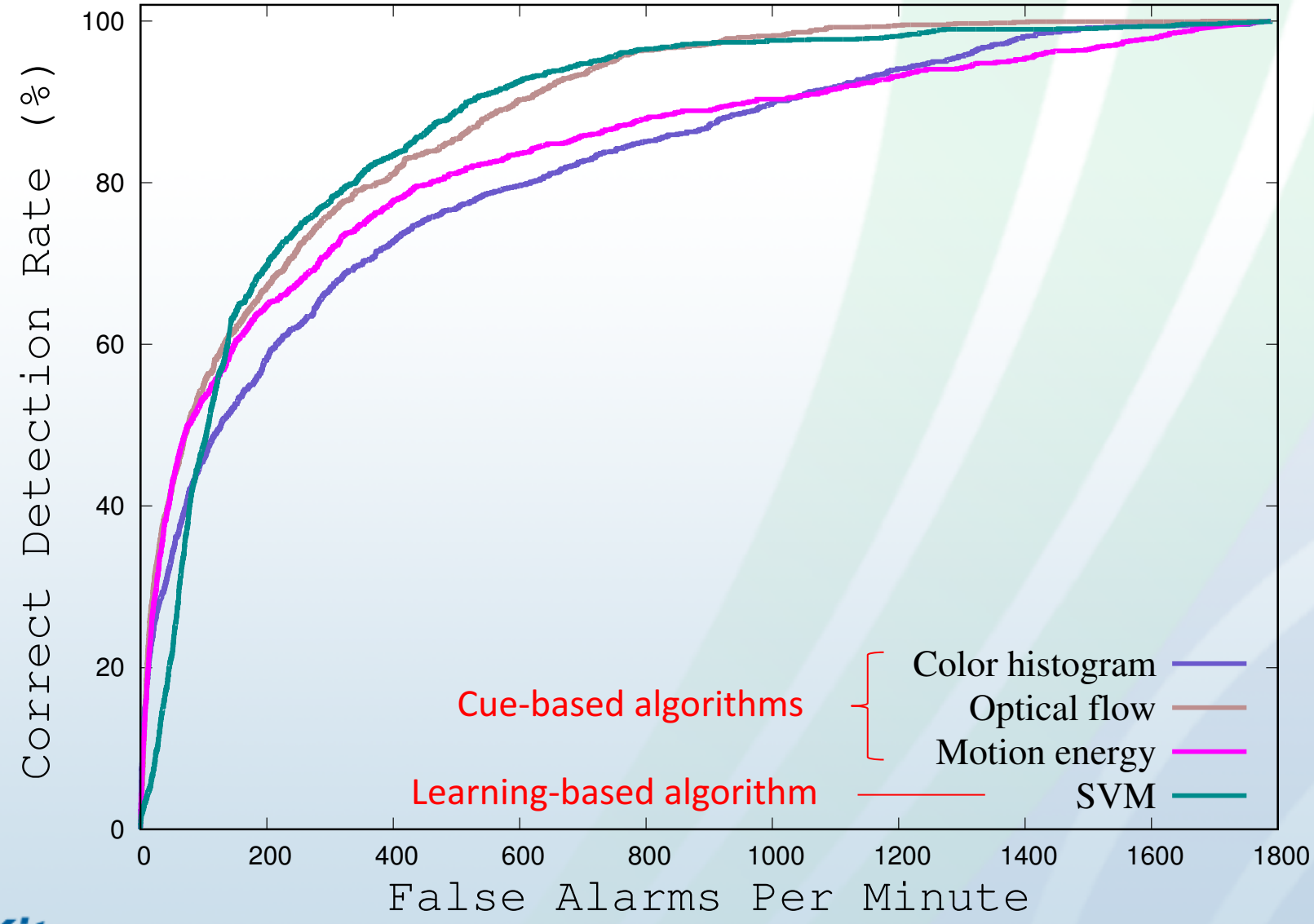
[13] S. Wolf. A no reference (nr) and reduced reference (rr) metric for detecting dropped video frames. In *National Telecommunications and Information Administration (NTIA),* 2009.

Performance – 0.5s drop duration

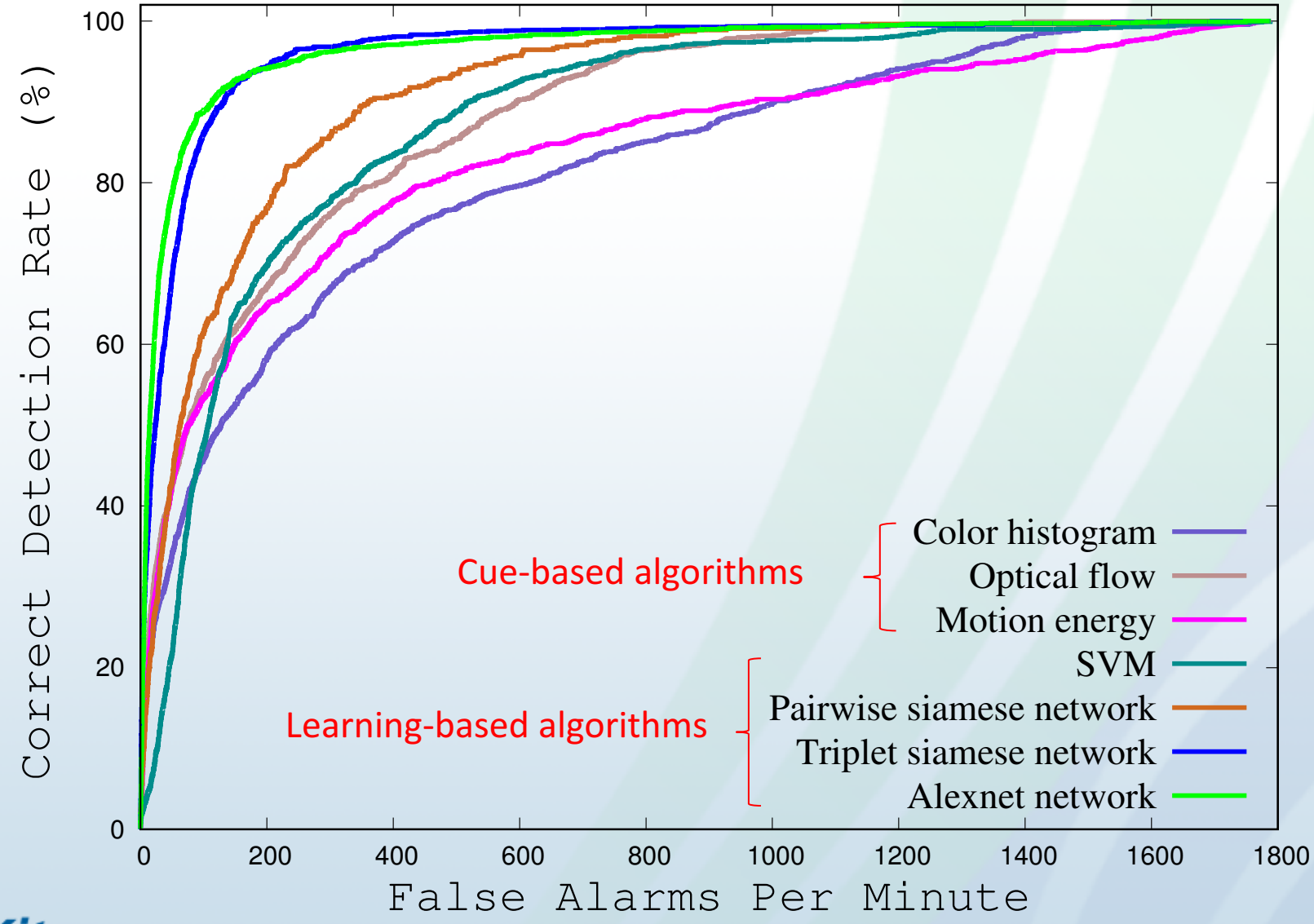


Kitware

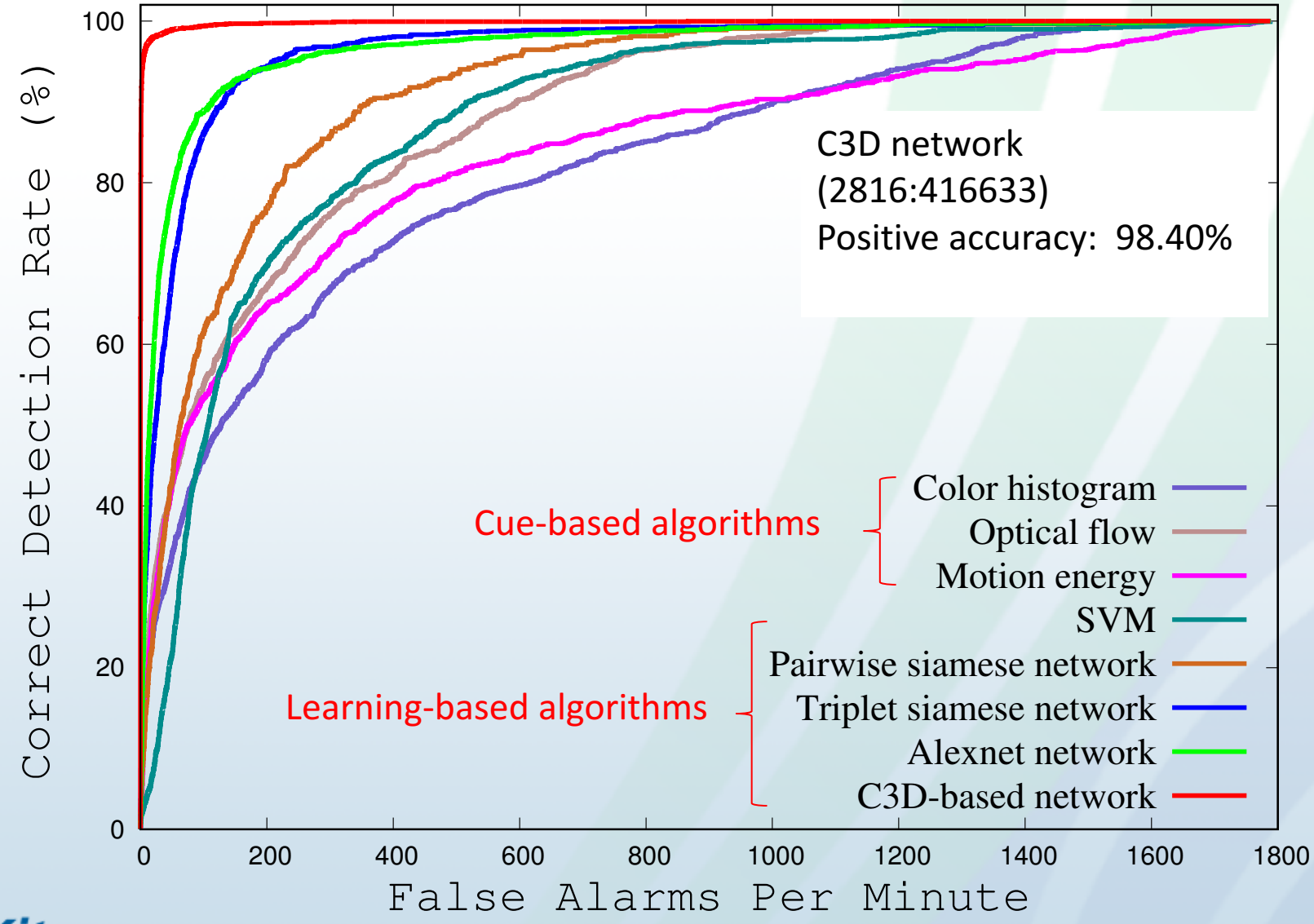
Performance – 0.5s drop duration



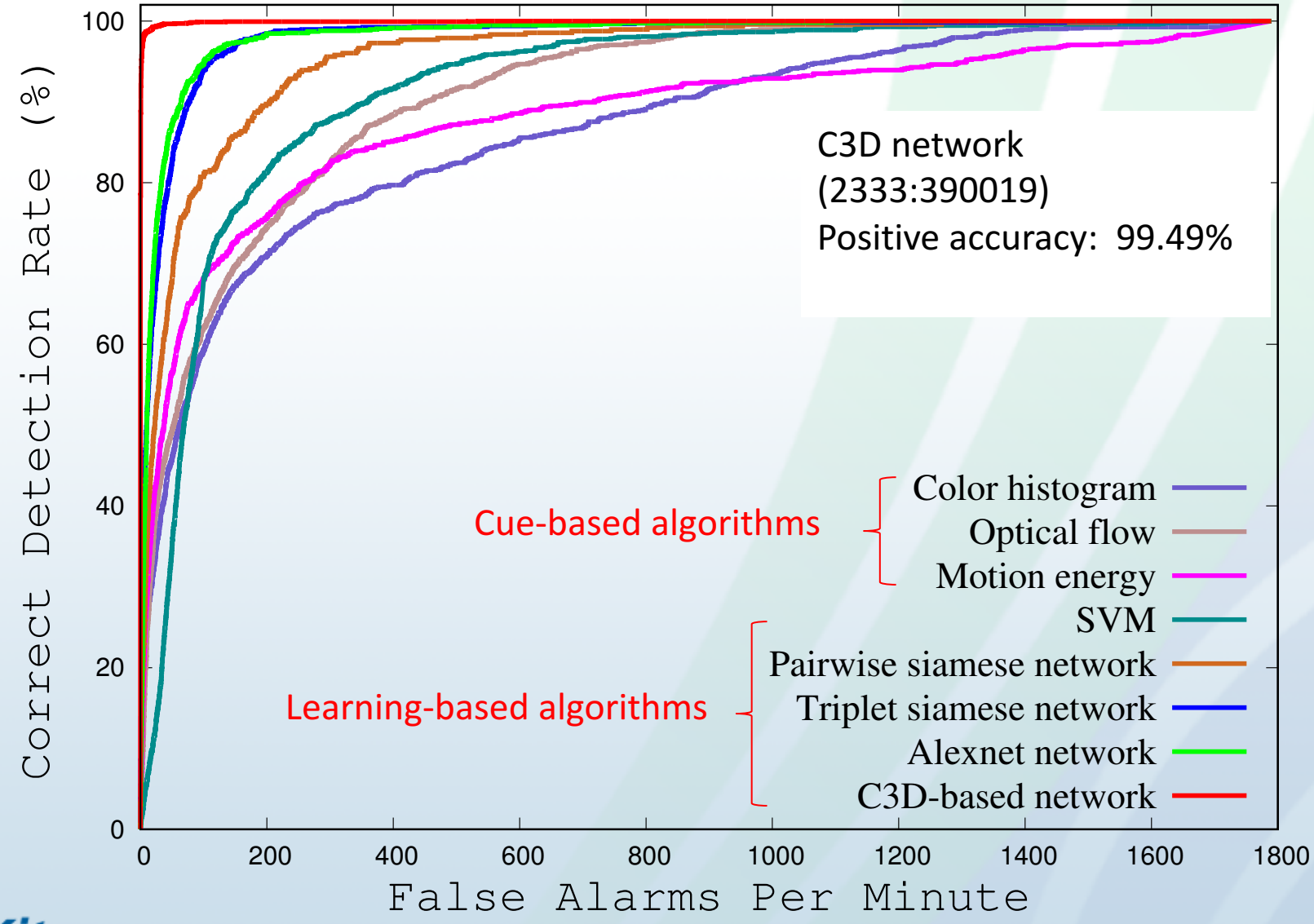
Performance – 0.5s drop duration



Performance – 0.5s drop duration

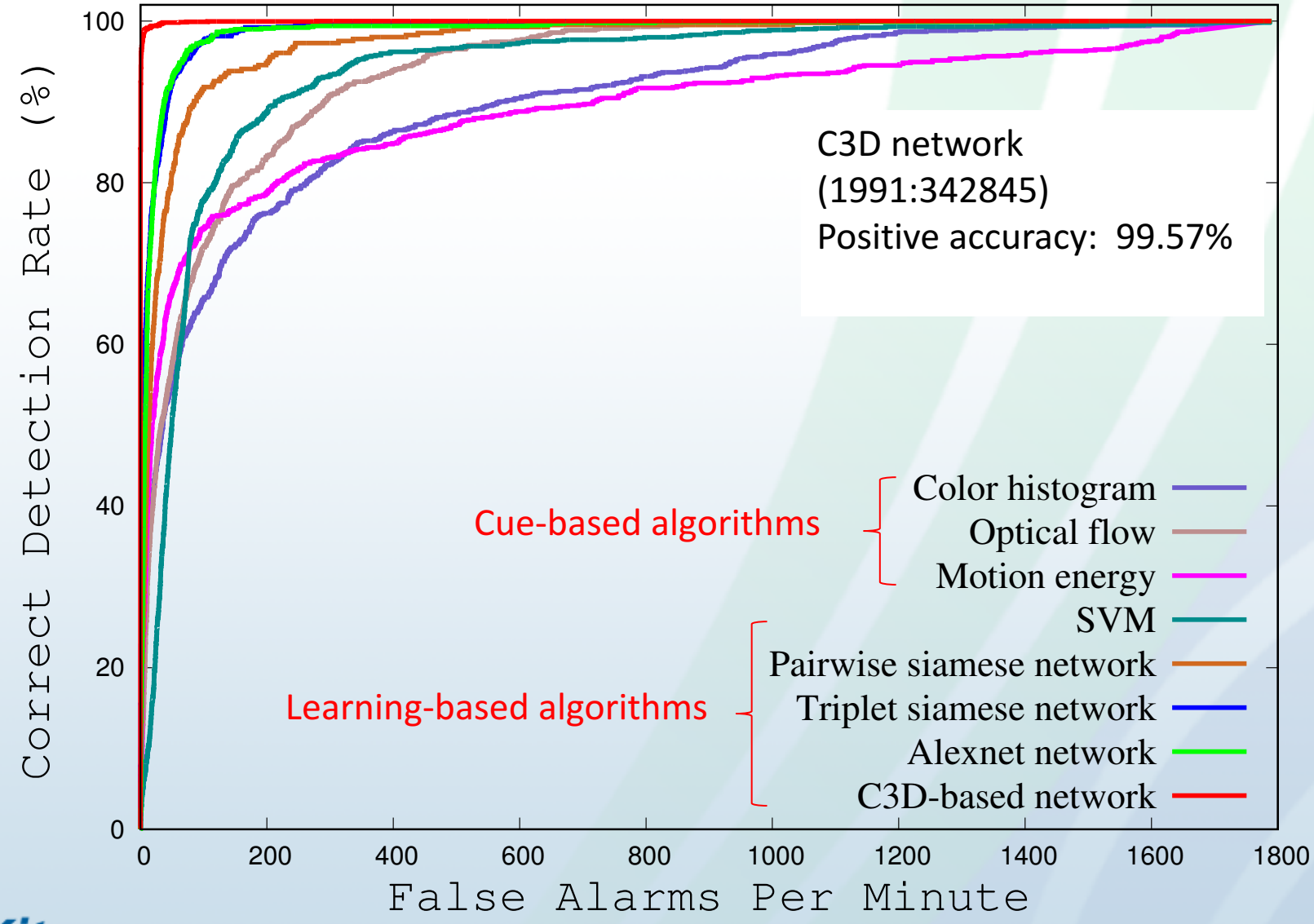


Performance – 1s drop duration



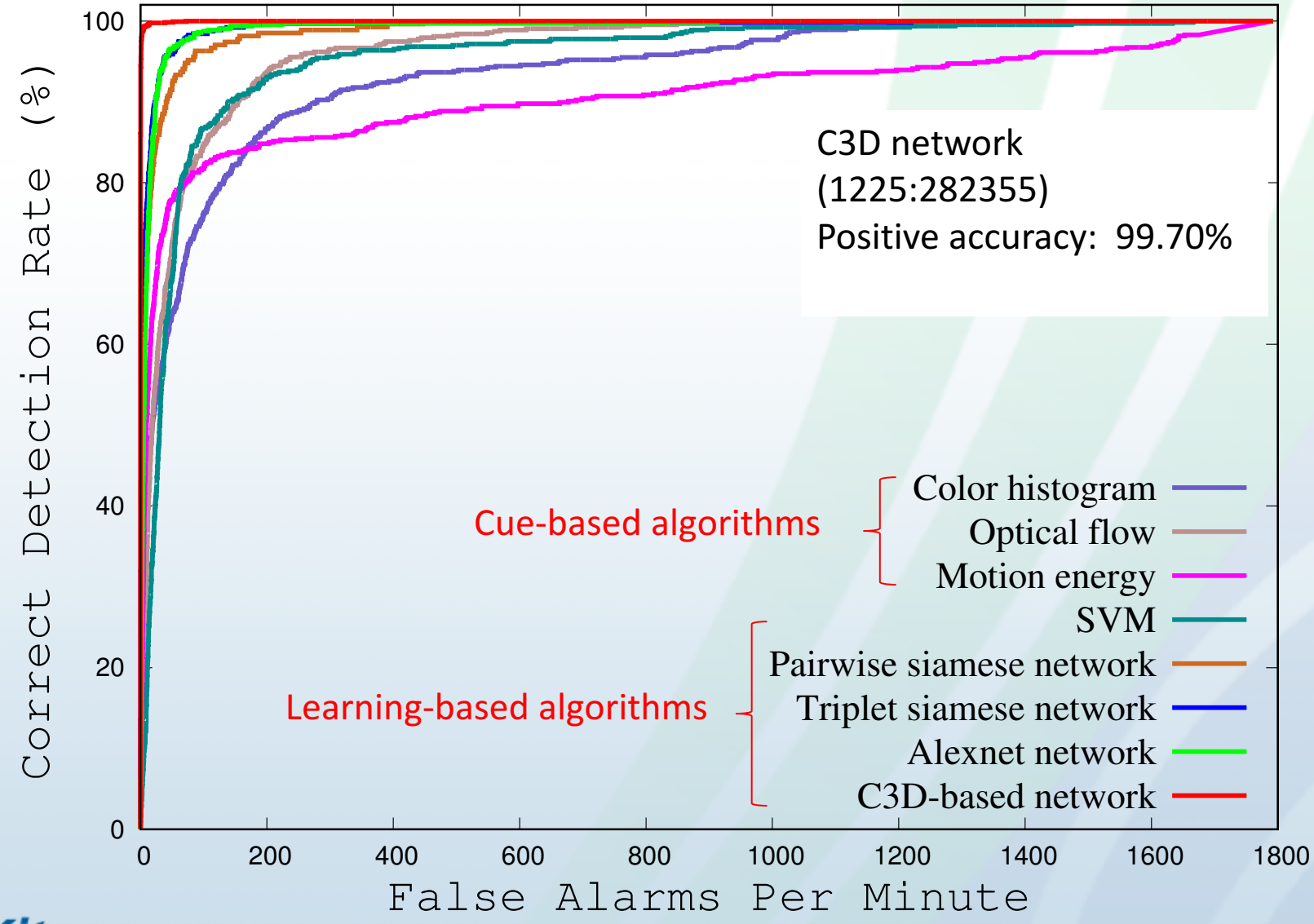
Kitware

Performance – 2s drop duration



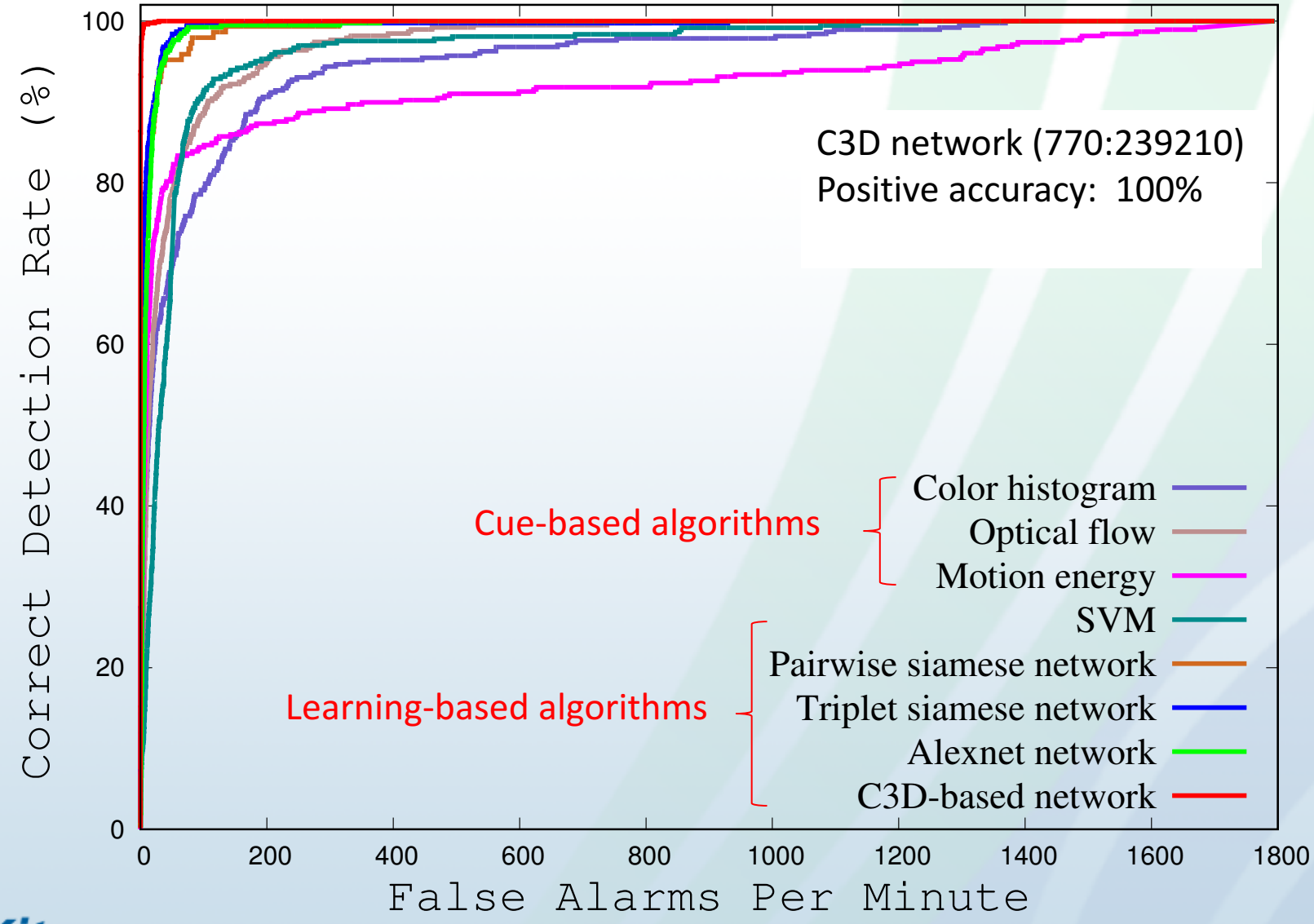
Kitware

Performance – 5s drop duration



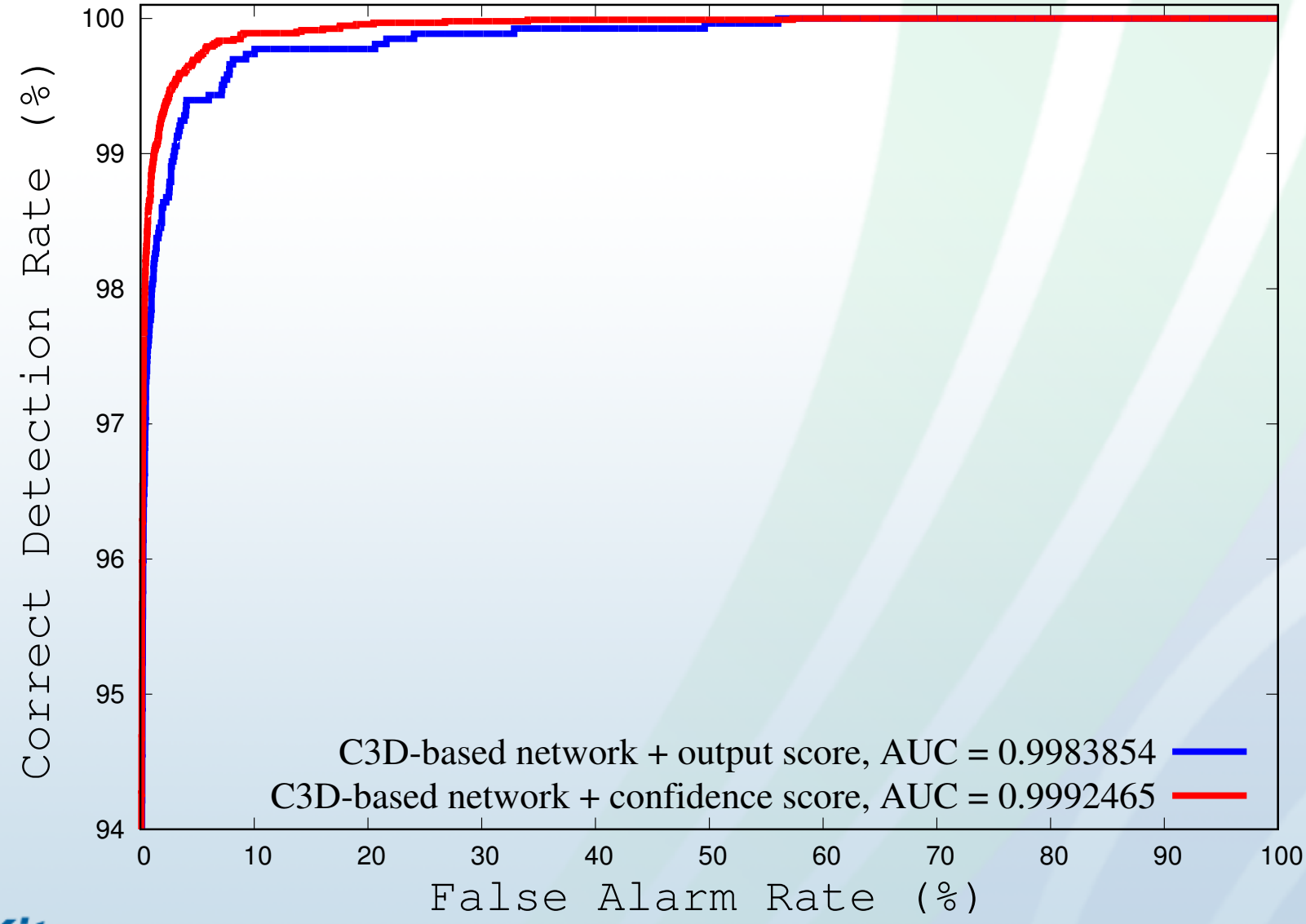
Kitware

Performance – 10s drop duration



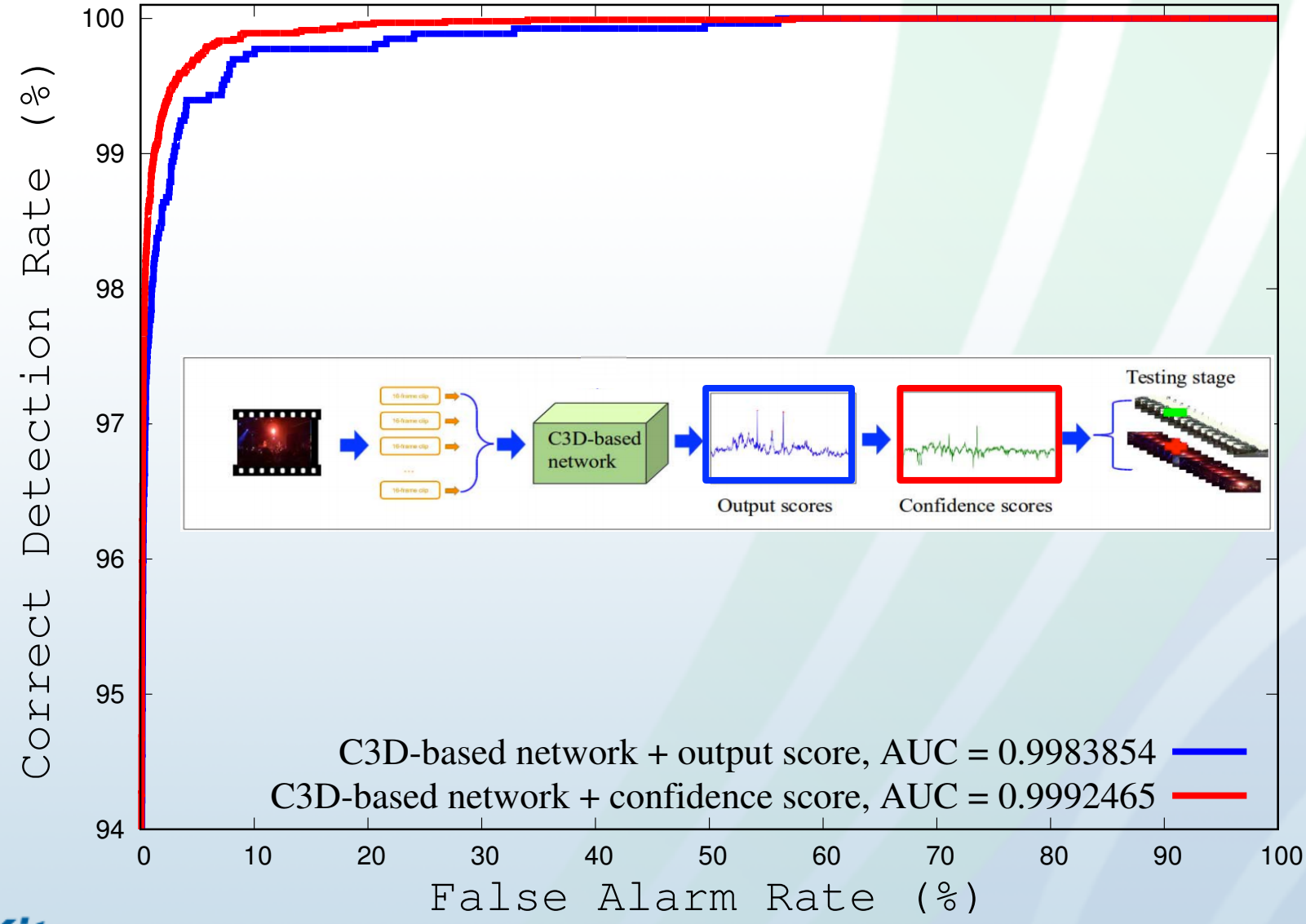
Kitware

Frame-level evaluation on YFCC100cm

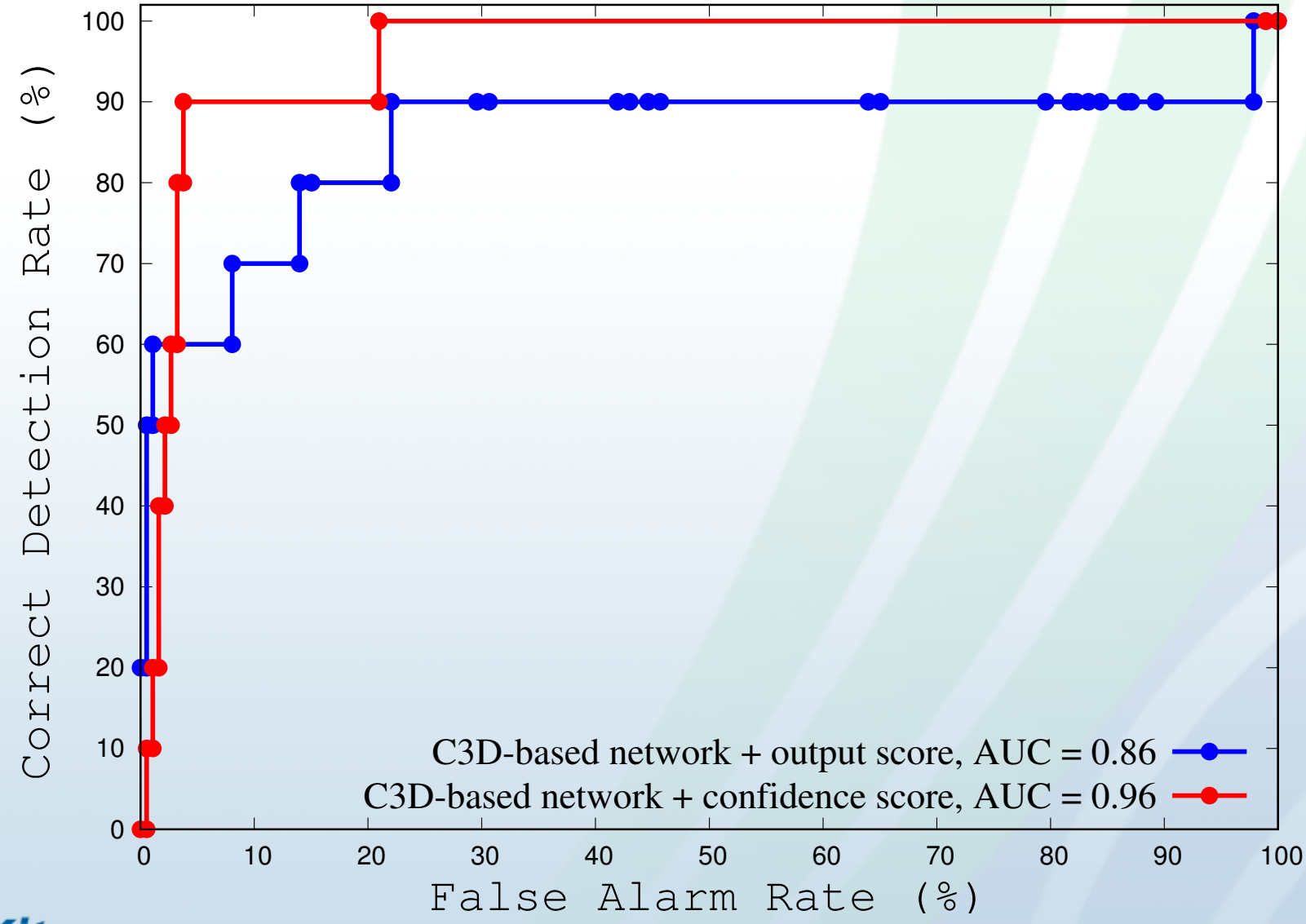


Kitware

Frame-level evaluation on YFCC100cm

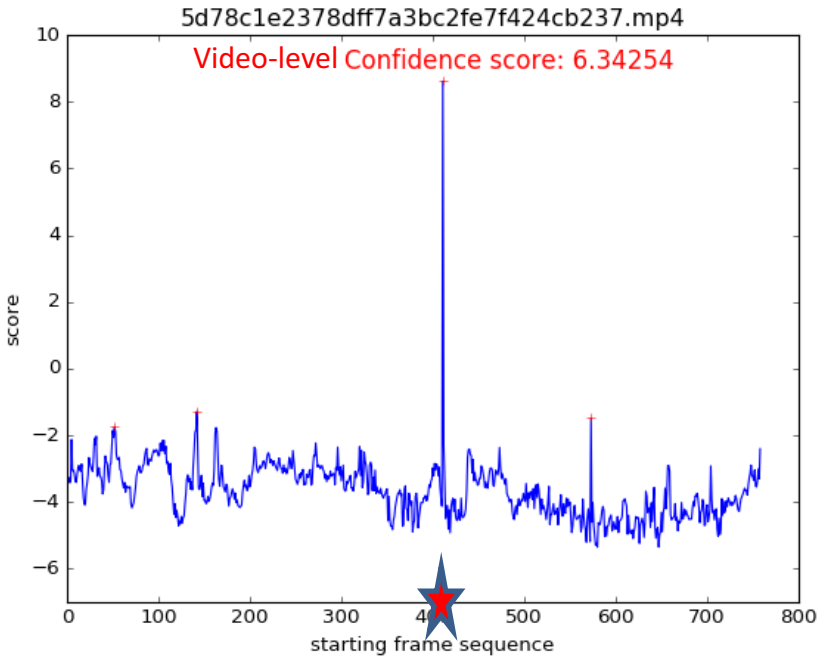


Video-level evaluation on NC17-Dev2-Beta1

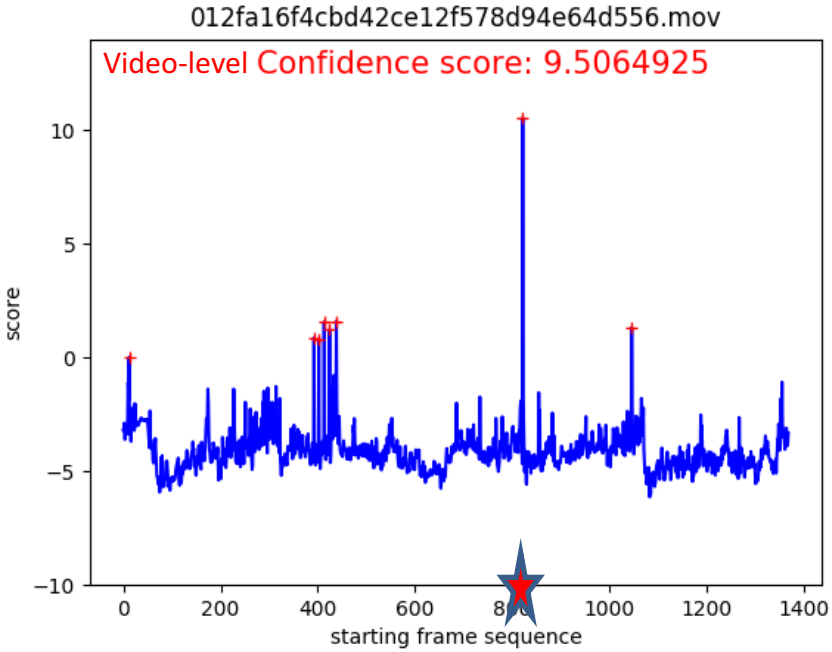


Kitware

Example (1)


5d78c1e2378dff7a3bc2fe7f424cb237.mp4
Video-level Confidence score: 6.34254
score
starting frame sequence
10
8
6
4
2
0
−2
−4
−6
0
100
200
300
500
600
700
800

Example (2)


012fa16f4cbd42ce12f578d94e64d556.mov
Video-level Confidence score: 9.5064925
score
10
5
0
−5
−10
0
200
400
600
1000
1200
1400
starting frame sequence

Conclusion

- Proposed a C3D-based network with confidence score defined with a peak detection and a scale term for frame dropping detection.
- Flexibly explore the underlying spatio-temporal relations within a video.
- Able to provide temporal localization of frame drops.

Future work

- Distinguish between shot breaks and frame drops.
- Expand the training dataset to include videos with zoom changes and fast camera motion.
- Use a Long Short-term Memory (LSTM) based network for faster run-time.
- Address other types of video manipulations, e.g. temporal duplication, looping, and frame rate changes.

Thanks!

chengjiang.long@kitare.com

Kitware