THE ANNUAL TRECVID WORKSHOP

Analysis of tracked objects for detecting activities in extended videos using 3D-CNN architectures

part of “CERTH-ITI participation in TRECVID 2020”

Konstantinos Gkountakos (presenter), Damianos Galanopoulos, Marios Mpakratsas, Despoina Touska, Anastasia Moumtzidou, Konstantinos Ioannidis, Ilias Gialampoukidis, Stefanos Vrochidis, Vasileios Mezaris, Ioannis Kompatsiaris

This work was partially supported by the European Commission under contracts H2020-832876 aqua3S, H2020-786731 CONNEXIONs, H2020-833115 PREVISION and H2020-780656 ReTV.

Objective of the submission

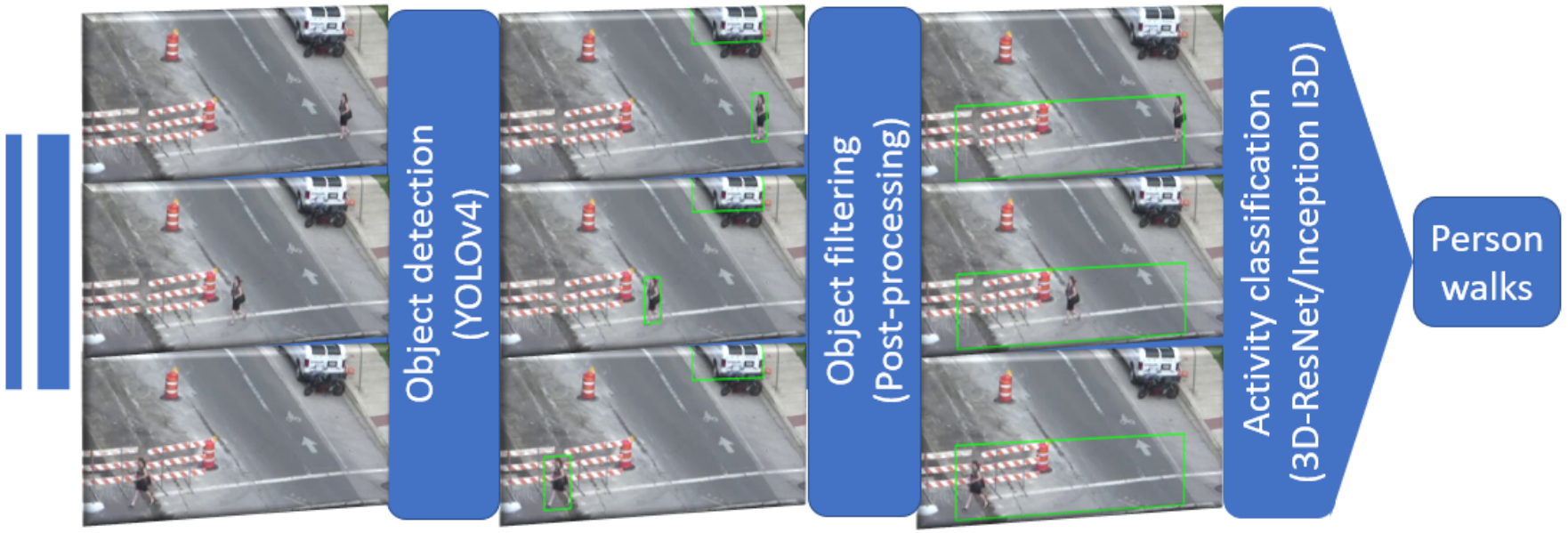
- Detecting Activities in Extended Video (ActEV) task aims to:
  - **Detect** activities of interest over time
  - **Recognize** the detected activities
- Challenges:
  - Camera's **field of view** is extremely large and includes multiple activities in very small areas
  - Detect **multiple activities at the same time**
  - **Untrimmed** videos are difficult to be watched, analyzed, and observed by humans

Activities recognition methods: Traditional approaches

- Try to classify short (in terms of duration) clips (trimmed) to predefined categories
  - **One or more** activities that occur in parallel
  - Classification of a **frame sequence**
  - Different **modalities** such as depth and optical flow contribute to enhance the performance
  - You only have to do with the **type** of activity and dismiss the temporal boundaries

Proposed approach

- Two-step approach:
  - Detection of objects in order to extract the **Extended Activity Bounding Boxes** (**EABBox**) using YOLOv4
  - Activity recognition for the objects identified in each **EABBox** using **3D-CNNs, 3D-ResNet** and **Inception I3D**

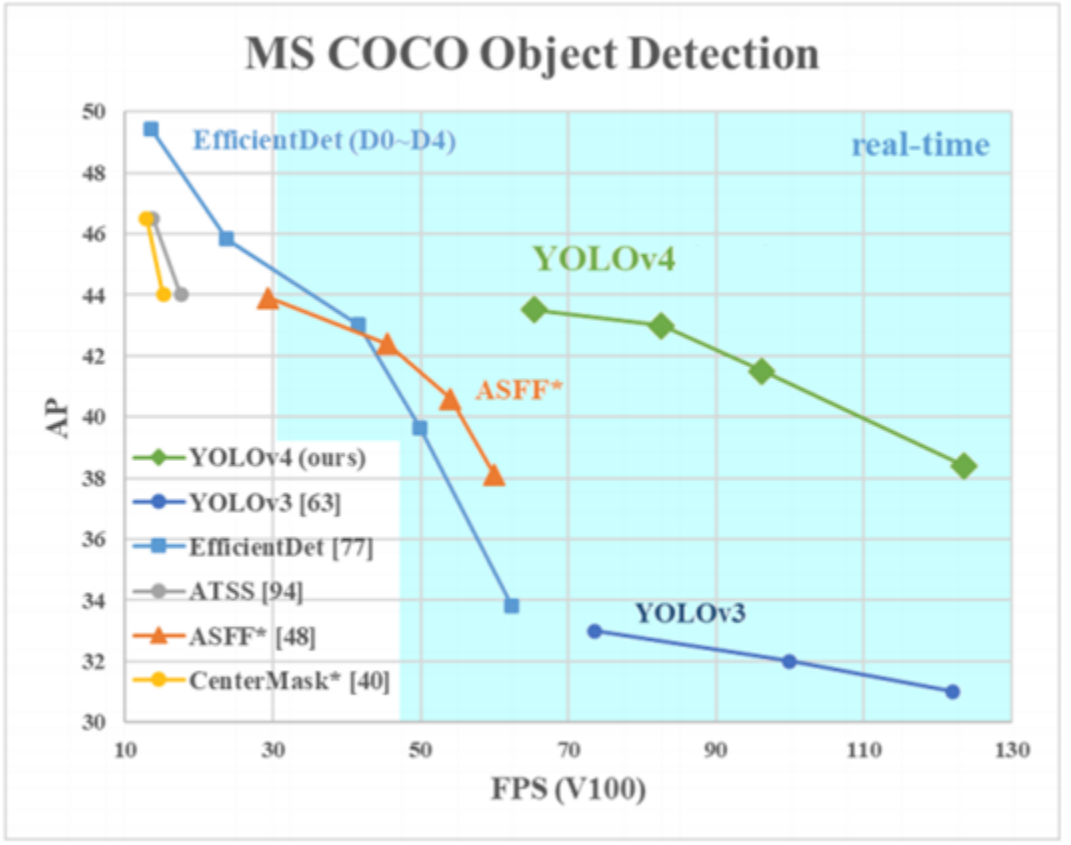


Challenge dataset

- Dataset:
  - The dataset consists of **35 person and vehicle related activities**
  - **Outdoor** environment
  - **Natural** light condition

Object detection
YOLOv4

- State-of-the-art **real-time object detector**
- **Pre-trained** using Microsoft COCO dataset
  - Include objects such as "person", "car", "truck
- Applied on **validation** and **evaluation** sets
- Detected objects are described by a **bounding** box and a **confidence score**
- **Object tracker** based on **Euclidean distance**



Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv preprint arXiv:2004.10934*.

Post-processing 1
Examine the effect of merging overlapped EABBoxes

1. **Process** only the **selected frames** according to the **annotations** of the dataset
2. **Select the objects** with class name belong to one of the following **categories**:
   - "person", "car", "bus", "truck" and "bicycle"
3. **Select** the predictions that have **confidence score** above a certain **threshold**
   - The value of the threshold was set at **25% empirically** to deal with the wrong detections of YOLOv4
4. **Select one class for every object**
   - In the case that an object is described by **two or more similar to target labels** (i.e. "car" and "truck")
     - We keep the class with a higher confidence score
   - Considering the **Intersection over the Union** (IoU)
     - With a certain threshold as a metric for overlapping, experimentally was set at **70%**

Post-processing 1
Examine the effect of merging overlapped EABBoxes

5. **Select** only the **moving objects**
   - The **detection of moving objects** calculated by checking the **global** start and the end position
     - Short positions are excluded as **considered static**
   - The class "**person**" is excluded as the data consists of both **static** and **movable** person-related activities
6. **Merge objects**
   - In cases that **two detected** objects are involved in one activity
     - A pair has to **include the class "person"** and one of the vehicle classes i.e. "car"
     - **Temporal overlapping** with less than 30 frames difference
     - **Spatial overlapping** with IoU bigger than 70%
7. **Calculate the EABBox** for each object(s)

Post-processing 2
Examine the immobility of objects at short intervals

1. **Process** only the **selected frames** according to the **annotations** of the dataset
2. **Select the objects** with class name belong to one of the following **categories**:
   - "person", "car", "bus", "truck" and "bicycle"
3. **Select** only the **moving objects**
   - The **detection of moving objects** calculated by checking the start and the end position
     - Short positions are excluded as **considered static**
   - The class "**person**" is excluded as there are person-related activities both **static** and **dynamically moving**

Post-processing 2
Examine the immobility of objects at short intervals

4. **Examine the immobility** of an object in short time intervals
   - Experimentally set at 10 seconds
   - We keep only the slots of time that an object moves
     - i.e. vehicle turn right -> stop for a long time -> vehicle turn left
5. **Calculate the EABBox** for each object
6. **Sequences** with less than **20** frames are rejected

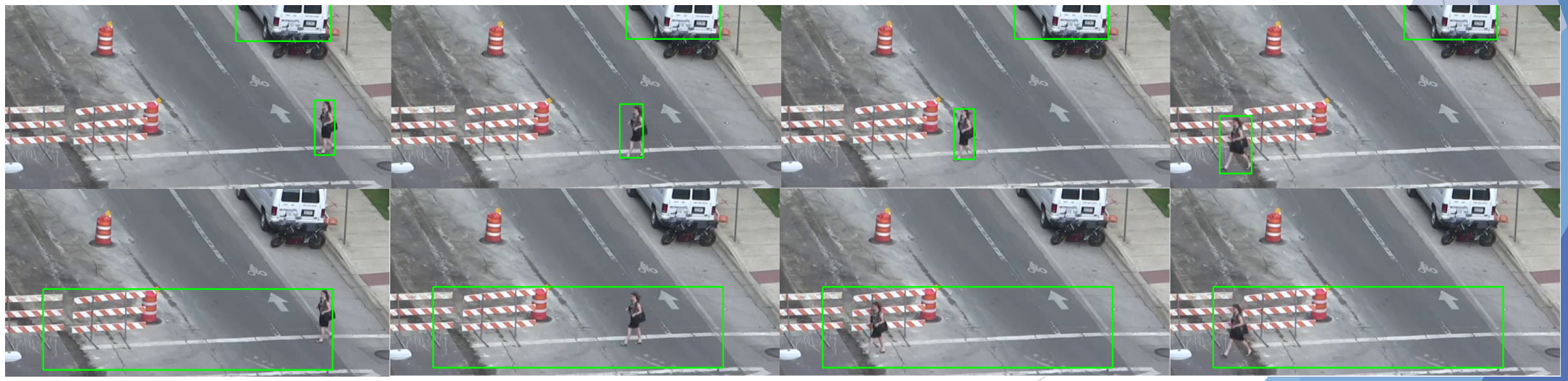
Post-processing 1 vs Post-processing 2

Commons

- **Process** only the **selected frames**
- **Select the objects categories**:
  - "person", "car", "bus", "truck" and "bicycle“
- **Select** only the **moving objects (global)**
- **Calculate the EABBox** for each object(s)

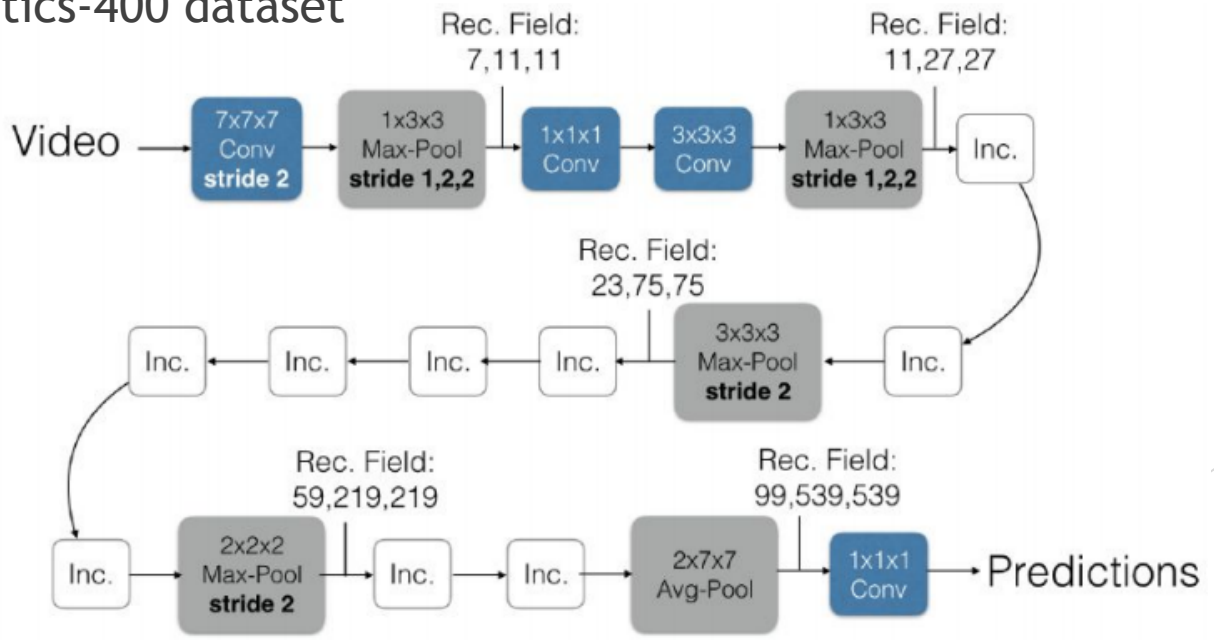
Differences

- **Ps1: Select** above a certain **threshold**
- **Ps1: Select one class for every object**
- **Ps1: Merge objects**
- **Ps2: Examine the immobility (local)**
- **Ps2: Sequences <20** frames are rejected



Activity classification Inception I3D

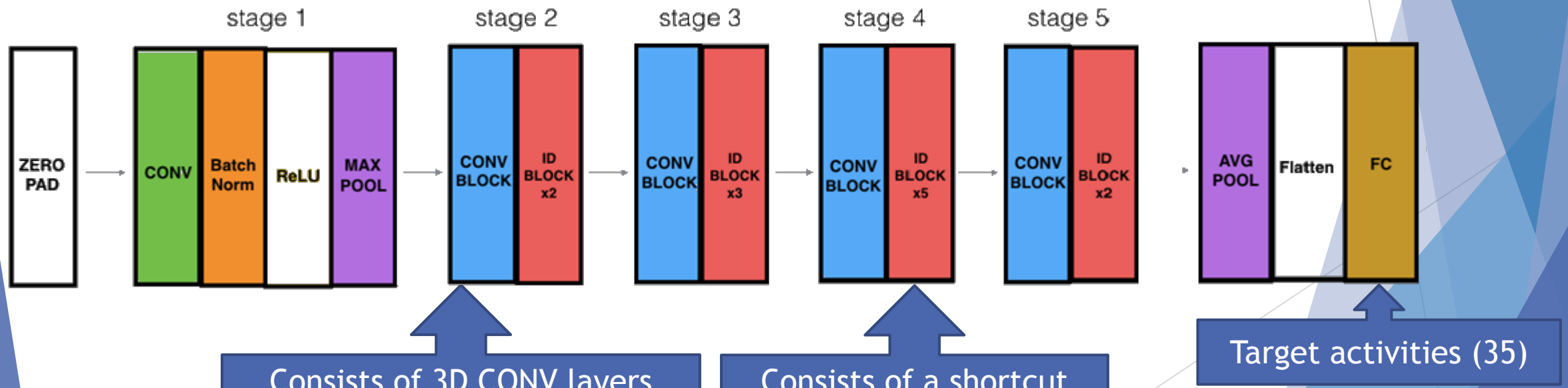
- Two-Stream Inflated 3D ConvNet (I3D)
  - 2D ConvNet inflation
- Inception model
- Loaded weights: Kinetics-400 dataset
- Total epochs: 1000
- Sample size: (32x224x224)(frames X width X height)
- Batch size: 8



https://www.cs.princeton.edu/courses/archive/spring18/cos598B/public/outline/Computer%20Vision%20Seminar%20Pres.pdf

Activity classification 3D-ResNet Training parameters

- Number of layers: 50
- Loaded weights: Kinetics-400 dataset
- Total epochs: 750
- Sample size: (16x112x112)(frames X width X height)
- Loss: categorical weighted cross-entropy
- Batch size: 64



https://engmrk.com/residual-networks-resnets/

Submitted systems

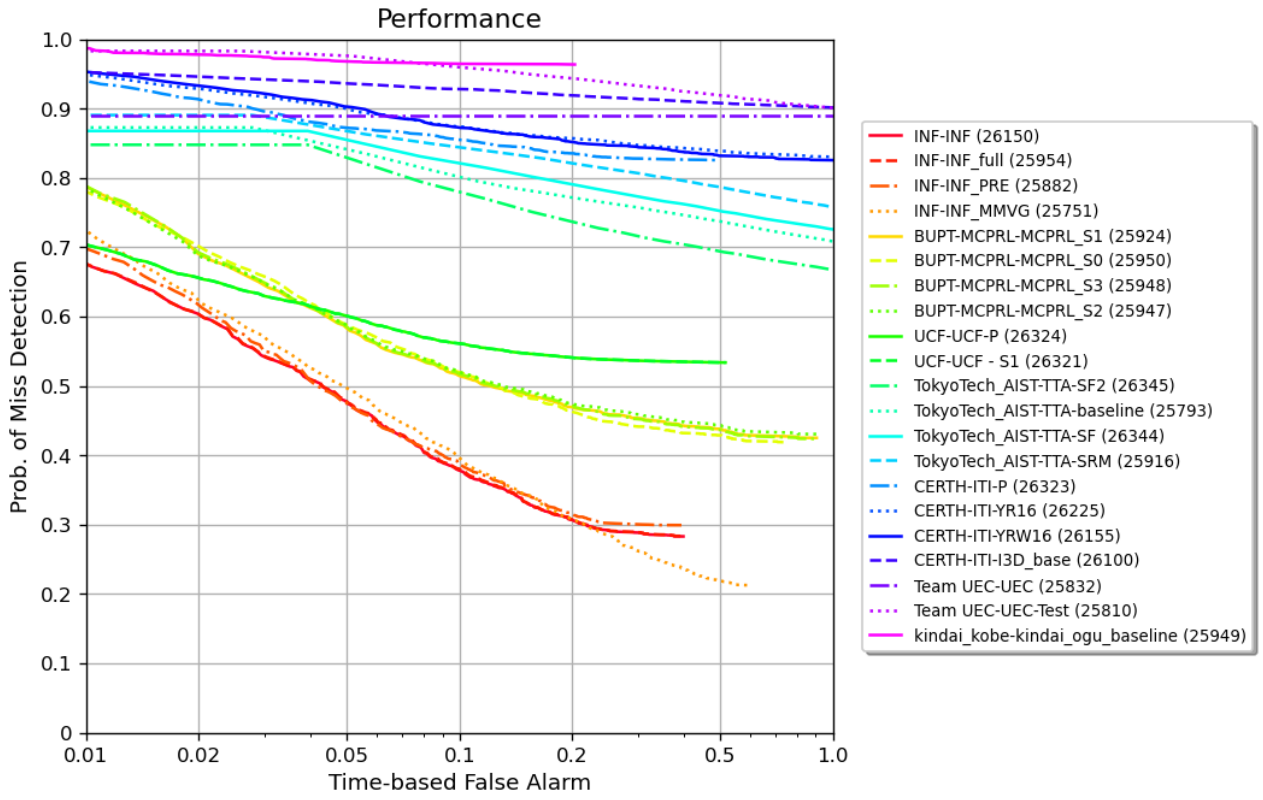
- **CERTH-ITI-I3D_base**: Ps1, I3D
- **CERTH-ITI-YRW16**: Ps1, 3D-ResNet-50, weighted cross-entropy loss
- **CERTH-ITI-YR16**: Ps1, **CERTH-ITI-YRW16**, considering activities more than 20 frames
- **CERTH-ITI-P**: Ps2, **CERTH-ITI-YR16**

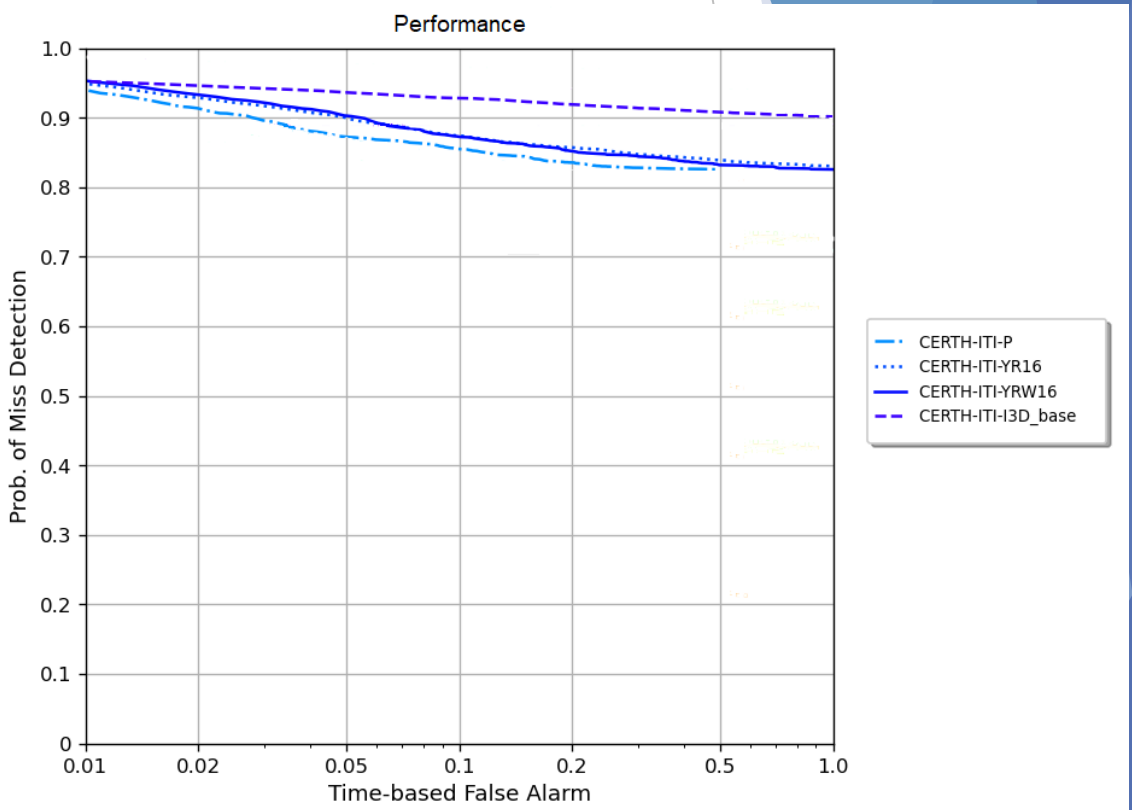
Evaluation results

System Name	PARTIAL AUDC	MEAN-P MISS@0.15TFA	MEAN-W_P MISS@ 0.15RFA
CERTH-ITI-I3D_base	0.93125	0.92318	0.92850
CERTH-ITI-YRW16	0.88530	0.86136	0.91187
CERTH-ITI-YR16	0.88511	0.86165	0.89439
CERTH-ITI-P	**0.86576**	**0.84454**	**0.88237**

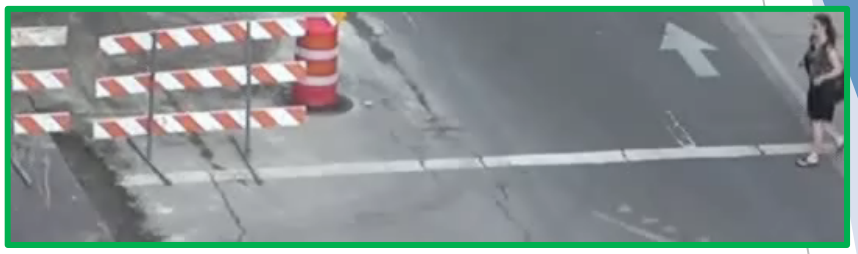
PARTIAL AUDC is the primary metric

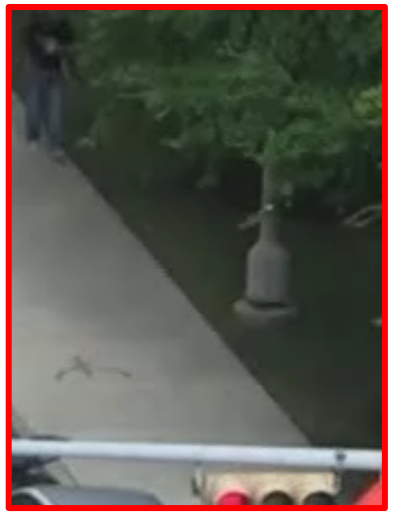
Experimental evaluation

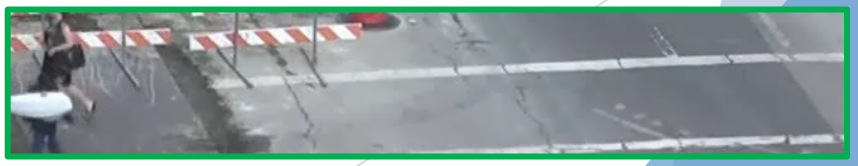




Indicative results
“Person walks”







Thank You

Konstantinos Gkountakos

gountakos@iti.gr

https://m4d.iti.gr/

MultiMoDal Data Fusion and Analytics Group

MKLab Multimedia Knowledge and Social Media Analytics Laboratory

Information Technologies Institute

CERTH CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

aqua3S

ReTV

CONNEXIONs

PREVISION

This work was partially supported by the European Commission under contracts H2020-832876 aqua3S, H2020-786731 CONNEXIONs, H2020-833115 PREVISION and H2020-780656 ReTV.