Special Topics in Object Detection summary

V. Nousi, C. Symeonidis, Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr
Version 2.5

Artificial Intelligence & Information Analysis Lab

Advanced Object Detection

- **Embedded object detection**
- Small object detection
- Person detection from aerial views

Object Detection

- Object detection = classification + localization:
- Find **what** is in a picture as well as **where** it is.

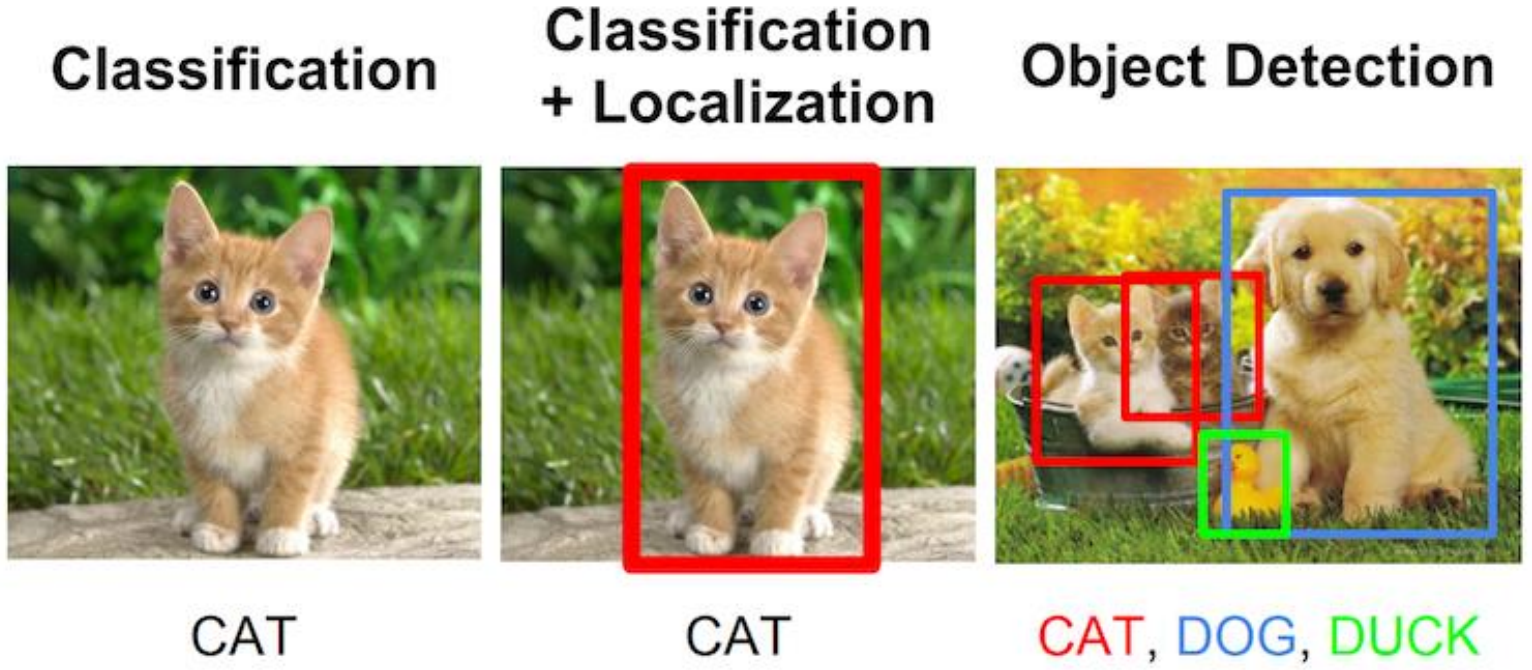
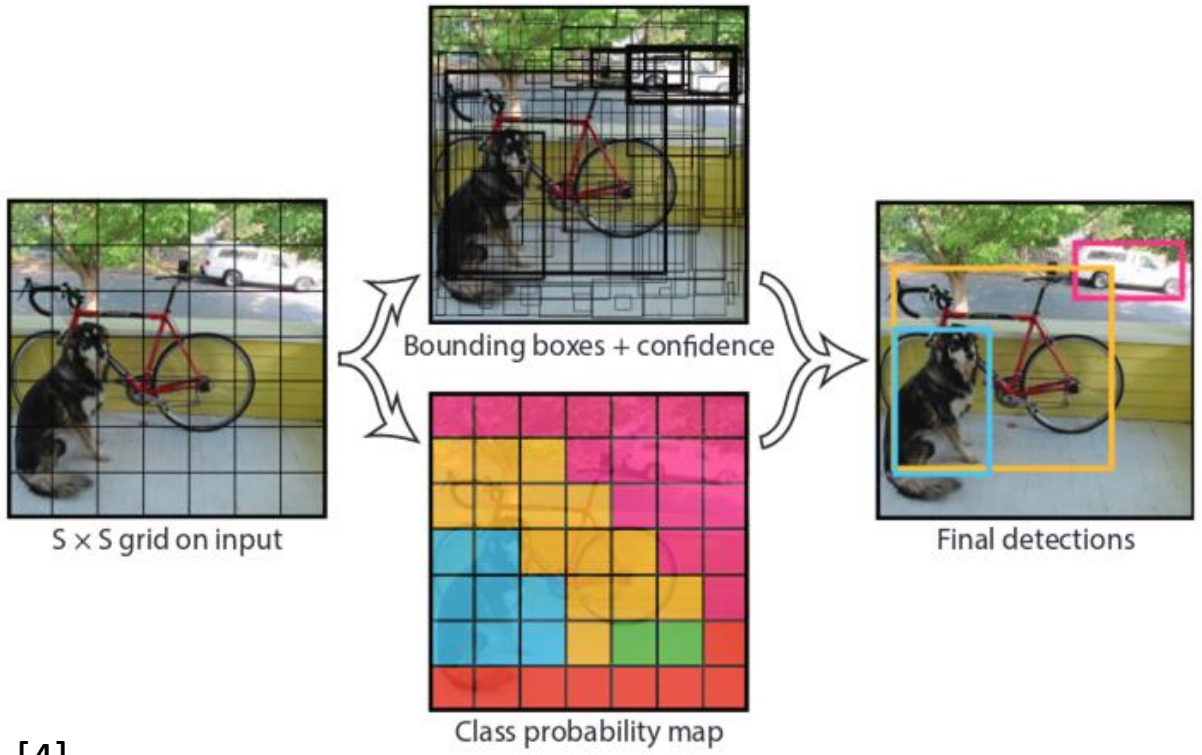


Figure: http://cs231n.stanford.edu/slides/2016/winter1516_lecture8.pdf

YOLO v2

- YOLO **divides** the input image into an $S \times S$ grid.
- If the **center** of an object falls within a cell of the grid, that cell is responsible for detecting that object.
- N is the maximal number of bounding boxes that each grid cell can detect.
- Each cell predicts N **bounding boxes** and confidence and classification scores for those boxes.



[4]

Using object detectors for drone-based shooting

- **Reducing the input image size can also increase the detection speed**
  - However, this can **significantly impact the accuracy** when detecting very small objects (which is the case for drone shooting)

Model	Input Size	Pascal 2007 test mAP*
YOLO v.2	544x544	77.44
YOLO v.2	416x416	74.60
YOLO v.2	288x288	67.12
YOLO v.2	160x160	48.72
YOLO v.2	128x128	40.68

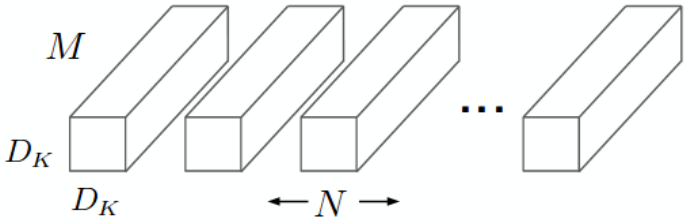
Using object detectors for drone-based shooting

- **We evaluated the faster detector (YOLO) on an GPU accelerated embedded system (NVIDIA TX-2)** that is available on our drone
- Adjusting the input image size allows for increasing the throughput
- Detection with satisfactory accuracy is not realtime

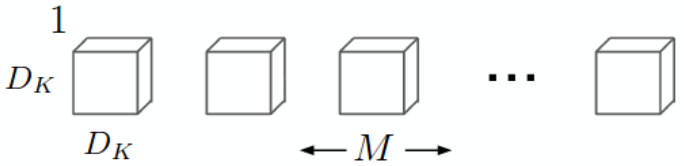
Model	Input Size	FPS
YOLO v.2	604x604	3
YOLO v.2	544x544	4
YOLO v.2	416x416	7
YOLO v.2	308x308	10
Tiny YOLO	604x604	9
Tiny YOLO	416x416	15

MobileNets

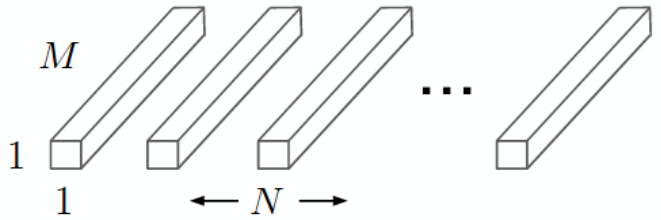
- **Standard convolutional filters:**
- N $D_k \times D_k$ filters, depth M ($M = 3$ for RGB images).
- **Depth-wise separable convolutional filters (Mobilenet):**
- M $D_k \times D_k$ filters of depth 1, one per input channel ($M = 3$ for RGB images) **and**
- N 1×1 filters (essentially weighted averaging of the M channels produced by the previous step).



(a) Standard Convolution Filters



(b) Depthwise Convolutional Filters



(c) 1×1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

[8]

Object Detection on NVIDIA Jetson TX2

- YOLO: good precision in general, but too heavyweight:
  - small objects are more challenging.
- Evaluation on VOC:

Input Image Size	FPS	mAP	Forward time (ms) No TensorRT	Forward time (ms) TensorRT	Forward time (ms) FP16
608x608	2.9	71.26	241.5	128.8	69.3
544x544	3.2	73.64	214.4	121.2	64.3
480x480	5.4	74.50	155.4	62.3	35.7
416x416	6.4	73.38	155.3	56.5	32.5
352x352	7.8	71.33	111.0	45.0	24.3
320x320	8.5	70.02	103.0	40.4	22.8

Object Detection on NVIDIA Jetson TX2

- Tiny YOLO: low precision, but very lightweight.
- Evaluation on VOC:

Input Image Size	FPS	mAP	Forward time (ms) No TensorRT	Forward time (ms) TensorRT	Forward time (ms) FP16
608x608	6.5	51.28	76.5	37.5	22.1
544x544	8.2	52.93	68.4	34.8	20.5
480x480	13.4	55.00	50.1	17.2	11.7
416x416	16.5	56.28	49.9	15.7	10.3
352x352	20	55.05	37.1	13.0	7.9
320x320	23	53.81	34.0	11.7	7.2

Advanced Object Detection

- Embedded object detection
- **Small object detection**
- Person detection from aerial views

Small Object Detection

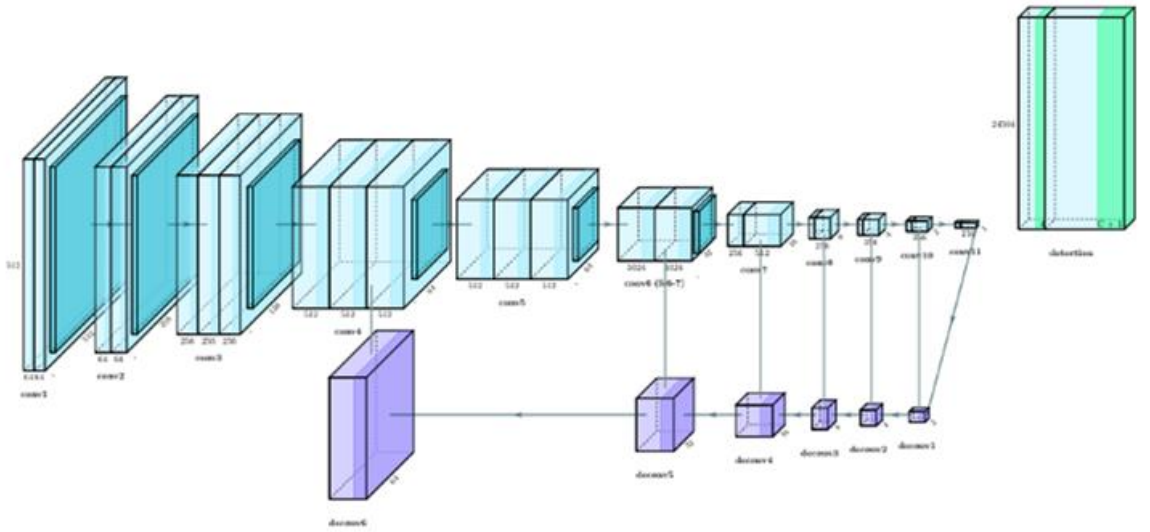
- Despite recent advances, state-of-the-art detectors face difficulties when it comes to detecting small objects
- Modern benchmarks feature images shot by UAVs which typically contain small and even tiny objects
- Research has focused on improving the performance of object detectors on such object sizes

Small Object Detection

Zhu, P., Wen, L., Bian, X., Ling, H. and Hu, Q., 2018. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*.

Small Object Detection

- Feature map super resolution
  - During training only, thus deployment remains fast
  - Allows for information transfer from deeper layers to shallow lavers and vice versa



Advanced Object Detection

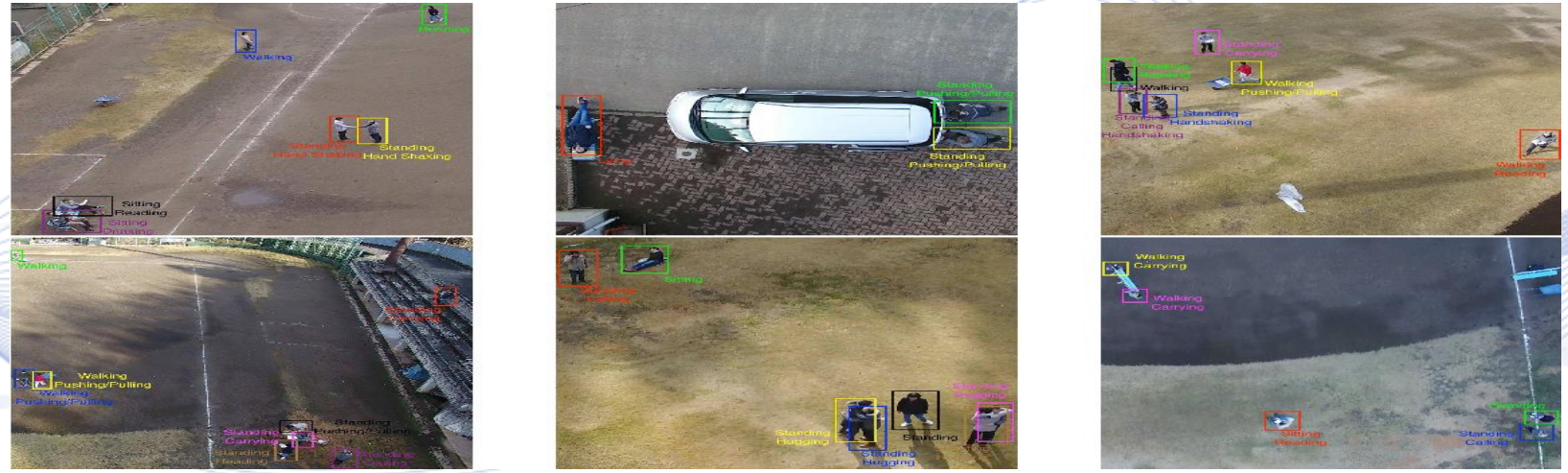
- Embedded object detection
- Small object detection
- **Person detection from aerial views**

Person Detection in Aerial Views

- Person detection is a vital part of any autonomous system regarding human safety.
- Though crowd detector achieves exceptional results in detecting dense crowds, it fails detecting sparse individuals.

Person Detection in Aerial Views

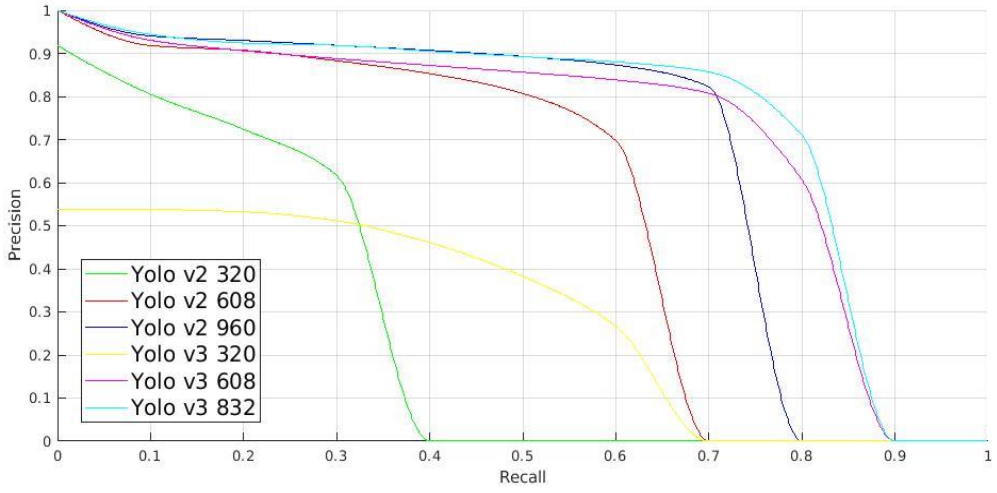
- Okutama-Action is a dataset which can be used for person detection and concurrent human action detection.
- It consists of 43 minute-long fully-annotated sequences with 12 action classes.
- It features many challenges missing in current datasets, including dynamic transition of actions, significant changes in scale and aspect ratio, abrupt camera movement, as well as multi-labeled actors.



Images for Okutama-Action dataset

Person Detection in Aerial Views

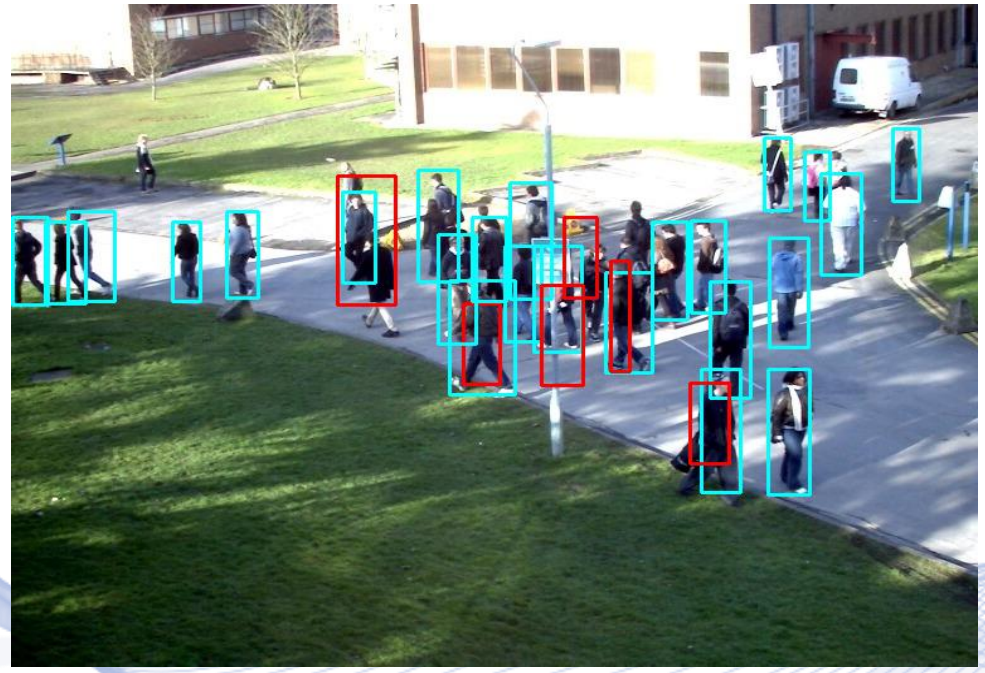
- YOLOv3 achieves higher (Mean Average Precision) MAP, compared to YOLOv2 using the same resolution. The only exception is the scale of images with resolution 308x308 where YOLOv2 is slightly better.
- Smaller resolutions (<608x608 pixels) had a very negative impact to the detector's accuracy. The main reason for this problem is the size of the detections, which are small due to UAV's high altitude.

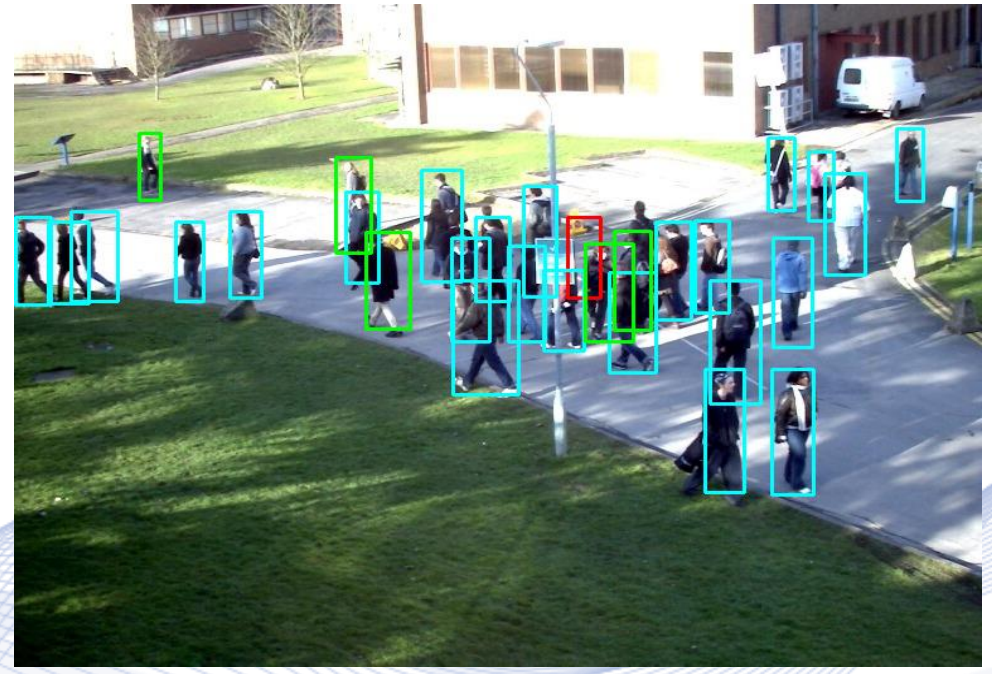


Detector	Resolution	MAP	FPS
YOLOv2	320x320	27.6%	7
YOLOv2	608x608	54.7%	4.5
YOLOv2	832x832	63.9%	2.2
YOLOv3	320x320	27.1%	4.7
YOLOv3	608x608	69.4%	1.8
YOLOv3	832x832	71.4%	1

Table 6: Results of person detectors on Okutama-Action dataset.

Improving neural NMS by exploiting interest-point detectors (Evaluation)

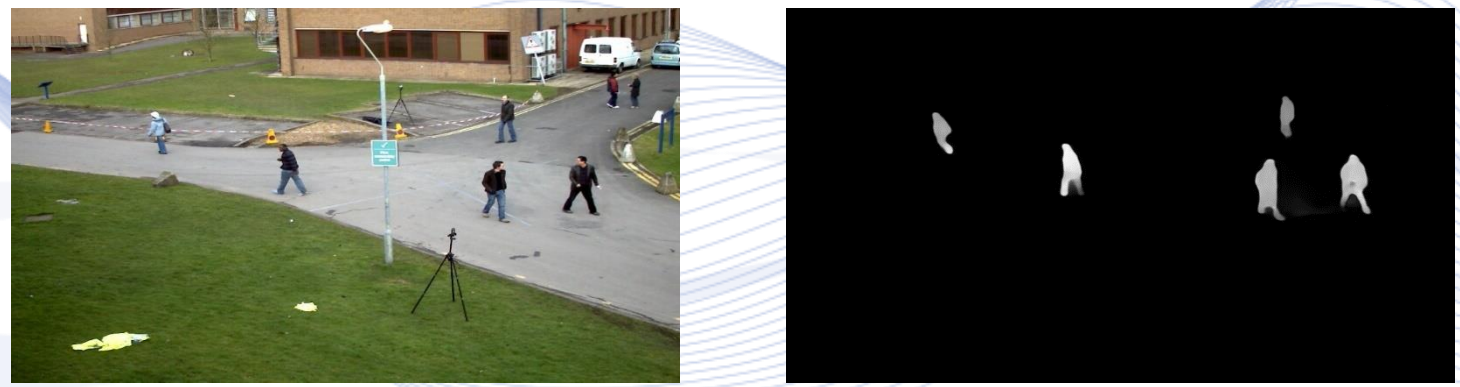




The highest (30) scoring detections using GreedyNMS (left) and the proposed method (right).

Improving neural NMS in videos

- In order to improve further neural NMS in videos, ROI representations based on interest-point maps were concatenated with representations derived by exploiting optical flow maps.
- The state-of-the-art FlowNet 2.0 was employed for the task of optical-flow estimation.
- This variation seems to improve the overall AP on PETS dataset.



RGB image (left), Optical flow map (right)

Method	AP
Greedy NMS IoU>0.4	76.4%
Default GossipNet_128	84.3%
FAST_FMoD_90	86.4%
FAST_FMoD_90/OpticalFlow_FMoD_90	86.9%
EdgeMap_FMoD_90/OpticalFlow_FMoD_90	87.1%
Improvement overBest Baseline	**2.8%**
Improvement over previous experiments	**0.7%**

Empirical evaluation on PETS dataset

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr