TRECVID WORKSHOP 2022 - Virtual, December 6 - 9, 2022

ITI-CERTH participation in ActEV and AVS Tracks of TRECVID 2022

Human and vehicle activity detection and recognition from untrimmed videos

Konstantinos Gkountakos, Damianos Galanopoulos, Despoina Touska, Konstantinos Ioannidis, Stefanos Vrochidis, Vasileios Mezaris, Ioannis Kompatsiaris

Presenter: Despoina Touska

CERTH
CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

Problem Statement

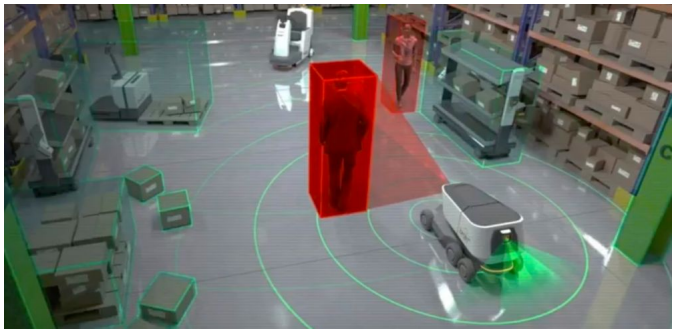
“Detection and recognition of human and vehicle-related activities from untrimmed video sequences”

- Input:
  - RGB frames
  - Indoor and outdoor environment
  - Multiple objects
- Output:
  - Type of activity
  - Activity duration
  - Confidence score

Label	Start Frame	End Frame	Score
person talks to person	5	240	91%

Activity Recognition Applications

- Surveillance scenarios
  - Traffic control
  - Abnormal event detection
  - Elderly patient monitoring
- Video indexing
  - YouTube videos
- Robotics
  - Robot navigation in an unknown environment
  - Effective human-robot interaction

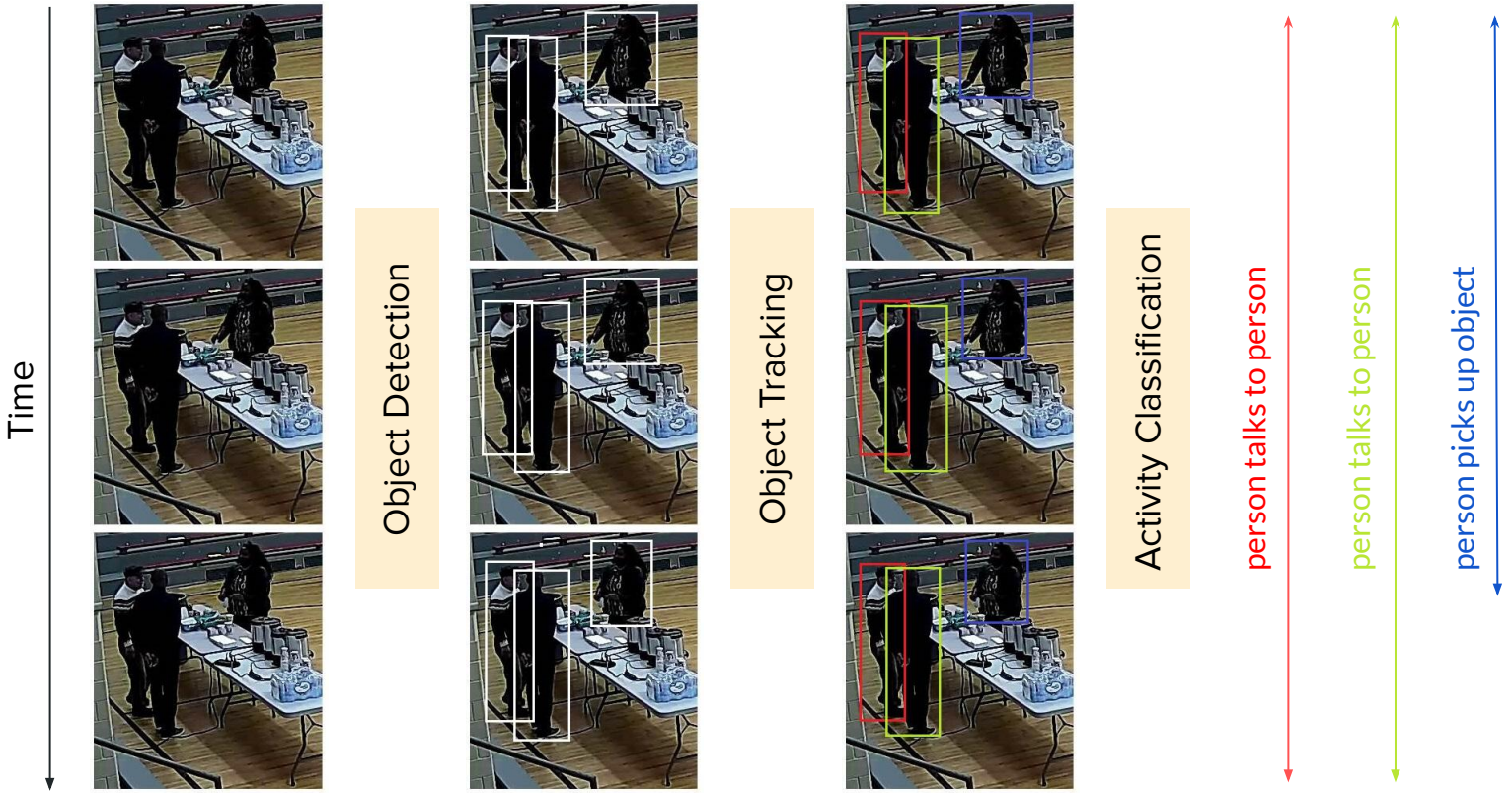


1. https://www.therobotreport.com/autonomous-navigation-design-challenges/

Challenges in Surveillance Scenarios

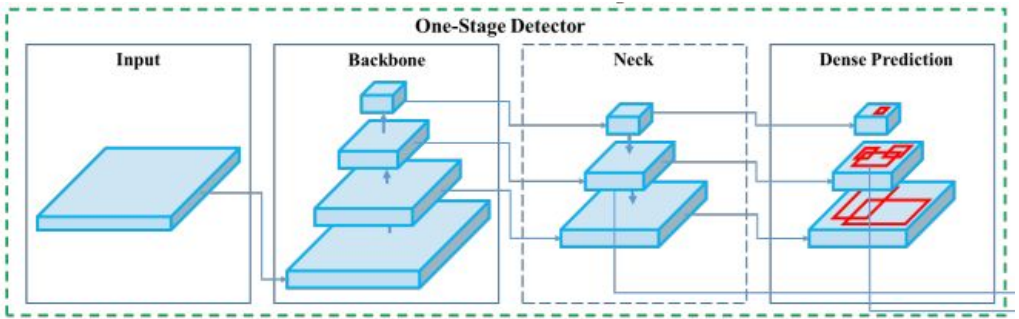
- Untrimmed video resources
- Multiple simultaneous activities
- Multi-tasking objects
- Interaction between objects
- Defining the exact spatiotemporal boundaries of activities

Proposed Approach



Object Detection - YOLOv4

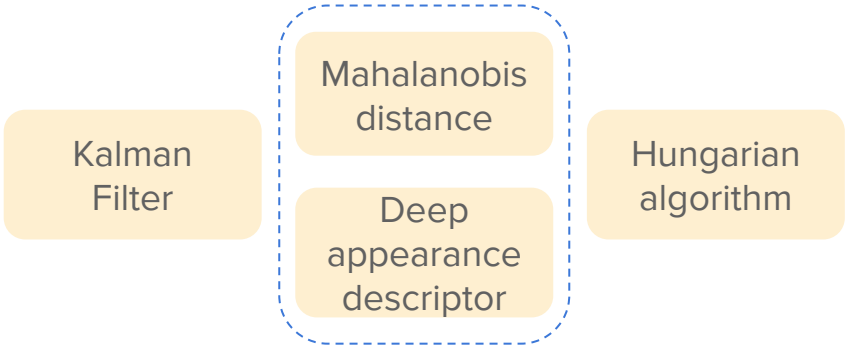
- Real time speed on the MS COCO dataset
  - 43.5 % AP running at 65 FPS on a Tesla V100
- CSPDarknet53 as backbone
- Concatenated Path Aggregation Networks (PAN) with Spatial Pyramid Pooling (SPP) Modules as neck
- Bag of freebies (BoF) and Bag of specials (BoS) methods as optimization procedures



1. Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.
2. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

Object Tracking - DeepSORT

- Extension of Simple Online Realtime Tracking (SORT)
- Associations metrics:
  - Mahalanobis distance
  - Deep appearance descriptor
    - Trained on a large-scale re-identification dataset using the cosine metric learning approach
- Interpolation to fill trajectory gaps



1. Wojke, N., Bewley, A., & Paulus, D. (2017, September). Simple online and realtime tracking with a deep association metric. In 2017 IEEE international conference on image processing (ICIP) (pp. 3645-3649). IEEE.
2. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., & Tian, Q. (2016, October). Mars: A video benchmark for large-scale person re-identification. In European conference on computer vision (pp. 868-884). Springer, Cham.

Activity Classification - Sets of Activities Groups

PR Classes	VR & PVR Classes
person reads document	person closes vehicle door
person enters scene through structure	person enters vehicle
person enters scene through structure	person exits vehicle
person stands up	person opens vehicle door
person sits down	vehicle starts
person talks to person	vehicle stops
person picks up object	vehicle turns left
person puts down object	vehicle turns right
person opens facility door	
person texts on phone	
person interacts with laptop	
person transfers object	

Activity Classification - Activity Classifier

- 3D-ResNet
- Four sequential bottleneck blocks
- Initialization using the Kinetics-400 dataset
- Training with a multi-label manner using the MEVA dataset
- Two separate activity classifiers trained using two sets of activities groups
- Weighted binary cross entropy loss
  - Balance activity-wise
  - n=sampleNumber, N=numberOfSamples, c=classNumber, p_c=classWeight, σ=sigmoidFunction

$$l_{n,c} = -[p_c y_{n,c} log(\sigma(x_{n,c})) + (1 - y_{n,c}) log(1 - \sigma(x_{n,c}))],\ p_c = \frac{N - \sum_{i=1}^{N} y_{i,c}}{\sum_{i=1}^{N} y_{i,c}}$$

1. Hara, K., Kataoka, H., & Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 6546-6555).

Activity Classification - Activities Refinement

- Two thresholds for the activities' scores
  - T_{low} excludes activity proposals with lower score
  - T_{high} includes:
    - Frame batches with higher scores in an activity proposal
    - Frame batches with lower scores among high-scored frame-batches
- Non-Maximum Suppression (NMS)
- Semantic rules of mutually exclusive groups

Group 1	vehicle starts, vehicle stops, person closes vehicle door, person opens vehicle door, person enters vehicle, person exits vehicle
Group 2	vehicle turns left, vehicle turns right, person closes vehicle door, person opens vehicle door, person enters vehicle, person exits vehicle
Group 3	person stands up, person sits down, person enters scene through structure, person exits scene through structure
Group 4	person picks up object, person puts down object, person reads document

Submissions

Baseline system includes:

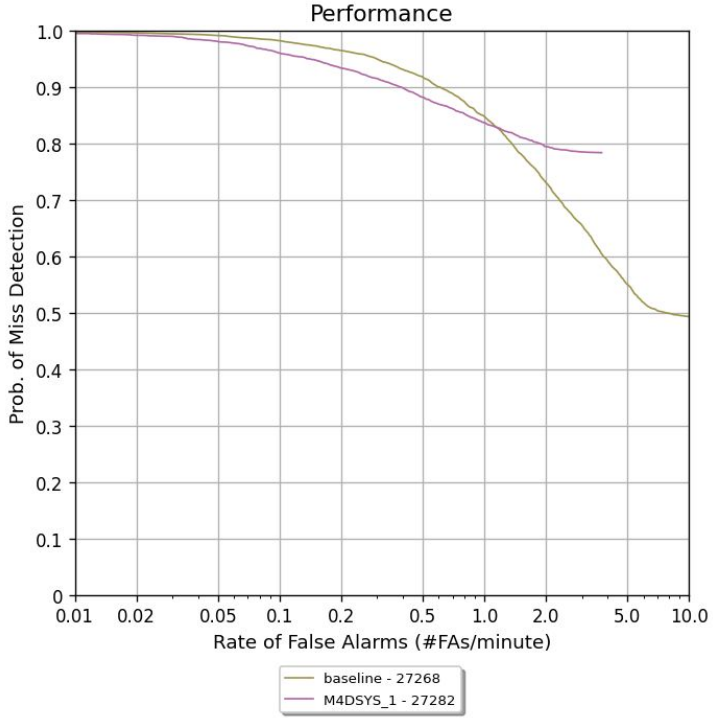
- YOLOv4
- DeepSORT
- two 3D-ResNet classifiers
- T_{high}=40%
- T_{low}=0%

M4DSYS_1 system includes:

- YOLOv4
- DeepSORT
- two 3D-ResNet classifiers
- T_{high}=65%
- T_{low}=10%
- NMS
- Semantic rules

Aggregated DET Curves

Graphical representation of Baseline vs M4DSYS_1 system performance in MEVA test set



Evaluation Results

Activity instances evaluation in MEVA validation and test sets using both Baseline and M4DSYS_1 systems

Dataset	Validation Set		Test Set	
Metric	**Baseline**	**M4DSYS_1**	**Baseline**	**M4DSYS_1**
pmiss@0.1rfa	0.9787	**0.9513**	0.9823	**0.9603**
nAUDC@0.2rfa	0.9802	**0.9528**	0.9819	**0.9639**
Correct Detections	**3142**	1233	-	-
False Detections	198059	**23269**	-	-
Missed Detections	**2670**	4579	-	-
Number of Activities	201201	24502	144071	23572

Conclusion and Future Work

- M4DSYS_1 outperforms Baseline system
  - Significant reduction in the number of false detections
- Metrics’ values are still high as:
  - Activity labels are assigned to the whole object’s trajectory; not in parts of it as annotated
  - Activity classifiers are misled
- Filtering activity proposals as future work

CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

Thank you

Despoina Touska
destousok@iti.gr