# Efficient Parallel Activity Detection System for Extended Video Analysis (TRECVID Leaderboard)

Wenhe Liu, Xiaojun Chang, Guoliang Kang,
Lijun Yu, Yijun Qian, Alexander G. Hauptmann, et al.

THE ROBOTICS INSTITUTE
ML
Carnegie Mellon University
MONASH University

**Presenter: Lijun Yu**

2020.03.05

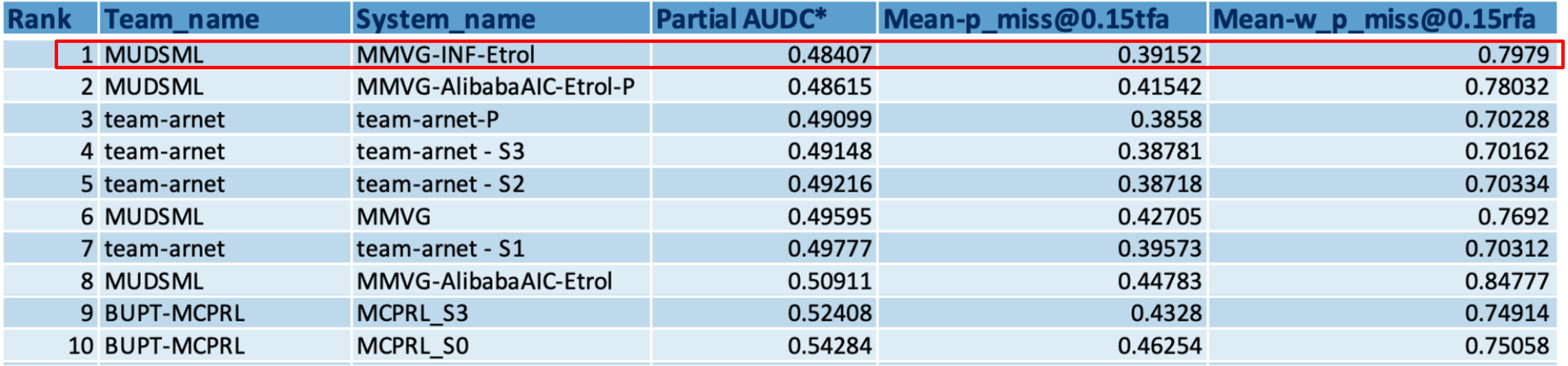
WACV 2020

Overview

System

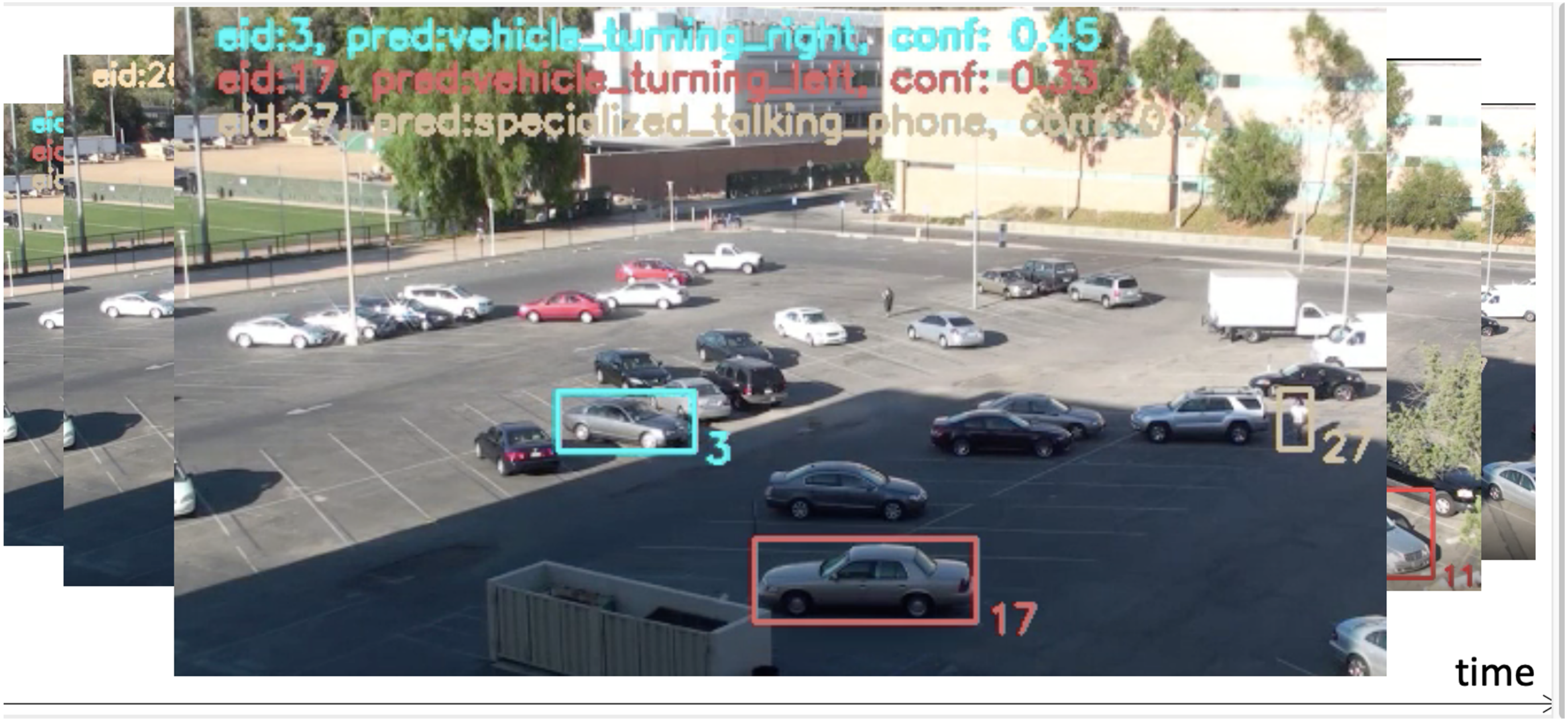
New Works

# Overview: Trecvid 2019 Actev Leaderboard: 1st place

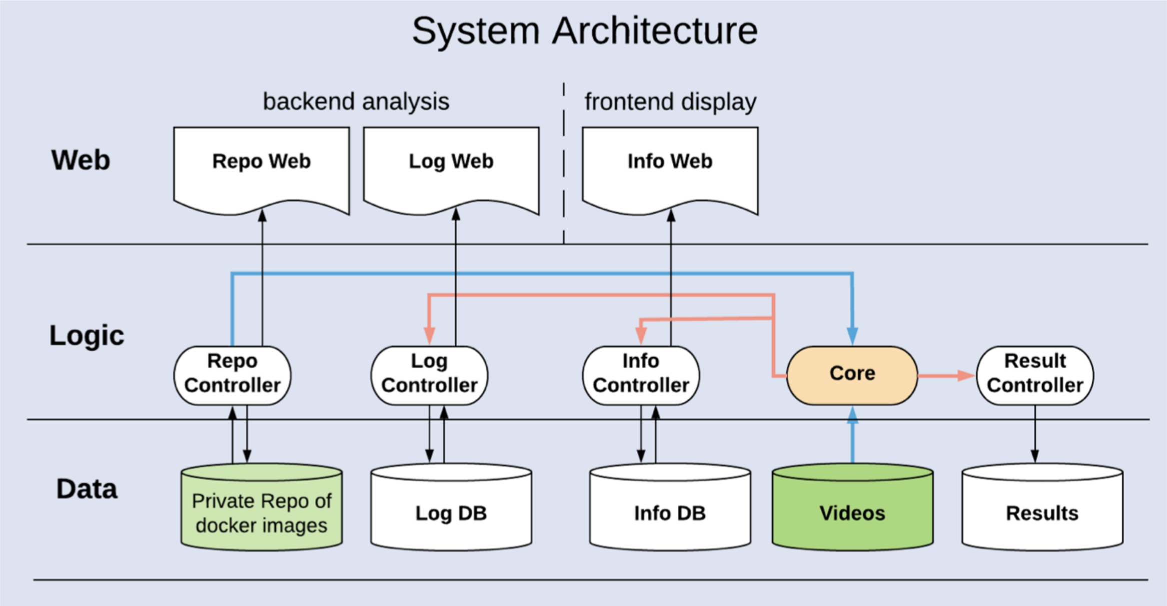


| Rank | Team_name | System_name | Partial AUDC* | Mean-p_miss@0.15tfa | Mean-w_p_miss@0.15rfa |
|---|---|---|---|---|---|
| 1 | MUDSML | MMVG-INF-Etrol | 0.48407 | 0.39152 | 0.7979 |
| 2 | MUDSML | MMVG-AlibabaAIC-Etrol-P | 0.48615 | 0.41542 | 0.78032 |
| 3 | team-arnet | team-arnet-P | 0.49099 | 0.3858 | 0.70228 |
| 4 | team-arnet | team-arnet - S3 | 0.49148 | 0.38781 | 0.70162 |
| 5 | team-arnet | team-arnet - S2 | 0.49216 | 0.38718 | 0.70334 |
| 6 | MUDSML | MMVG | 0.49595 | 0.42705 | 0.7692 |
| 7 | team-arnet | team-arnet - S1 | 0.49777 | 0.39573 | 0.70312 |
| 8 | MUDSML | MMVG-AlibabaAIC-Etrol | 0.50911 | 0.44783 | 0.84777 |
| 9 | BUPT-MCPRL | MCPRL_S3 | 0.52408 | 0.4328 | 0.74914 |
| 10 | BUPT-MCPRL | MCPRL_S0 | 0.54284 | 0.46254 | 0.75058 |

# Overview: Task Description


eid:3, pred:vehicle_turning_right, conf: 0.45
eid:17, pred:vehicle_turning_left, conf: 0.33
eid:27, pred:specialized_talking_phone, conf: 0.24
3
27
17
11
time

# Overview: System Architecture



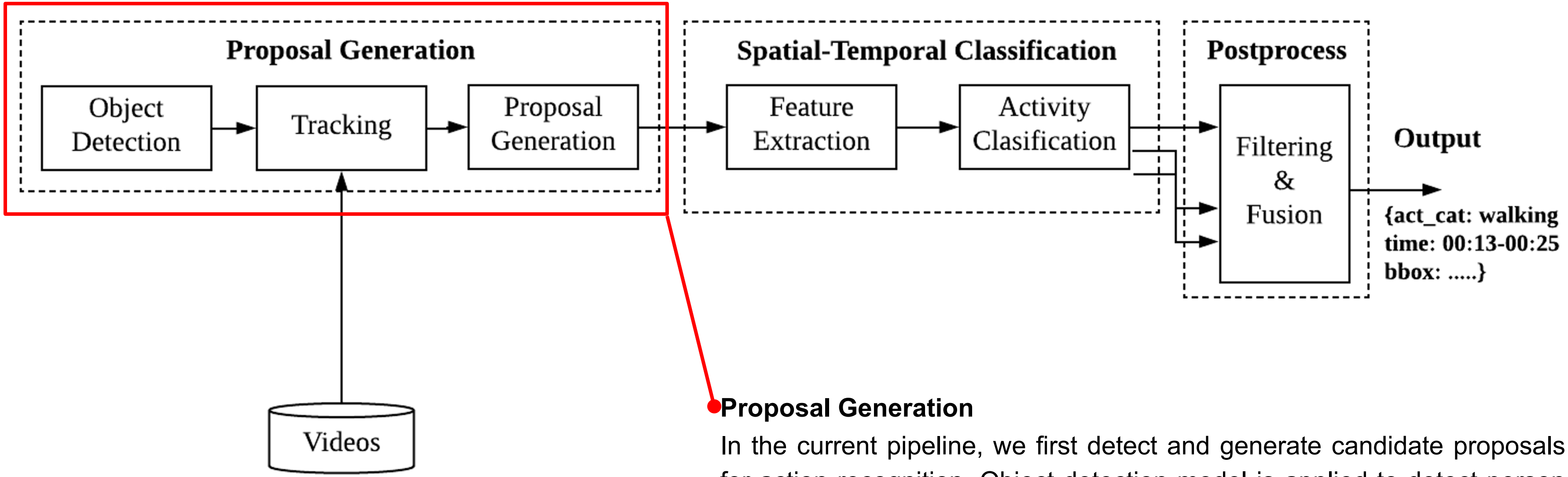
We implemented a three-tier architecture for our system:

**Data** tier: We manage the input videos and results, log and information of pipelines, repository that saves all the docker images used in the pipeline. (We omit the docker repository management part in our DIVA CLI  system.)

**Logic** tier: There is one core which manages the parallel video analysis in the system. It controls multiple controllers which manage the data.

**Web** tier:  Status of  docker images, logs and information of  pipeline are displayed for backend and frontend analysis. (We omit the web level and Log/info Database in our DIVA CLI  system.)

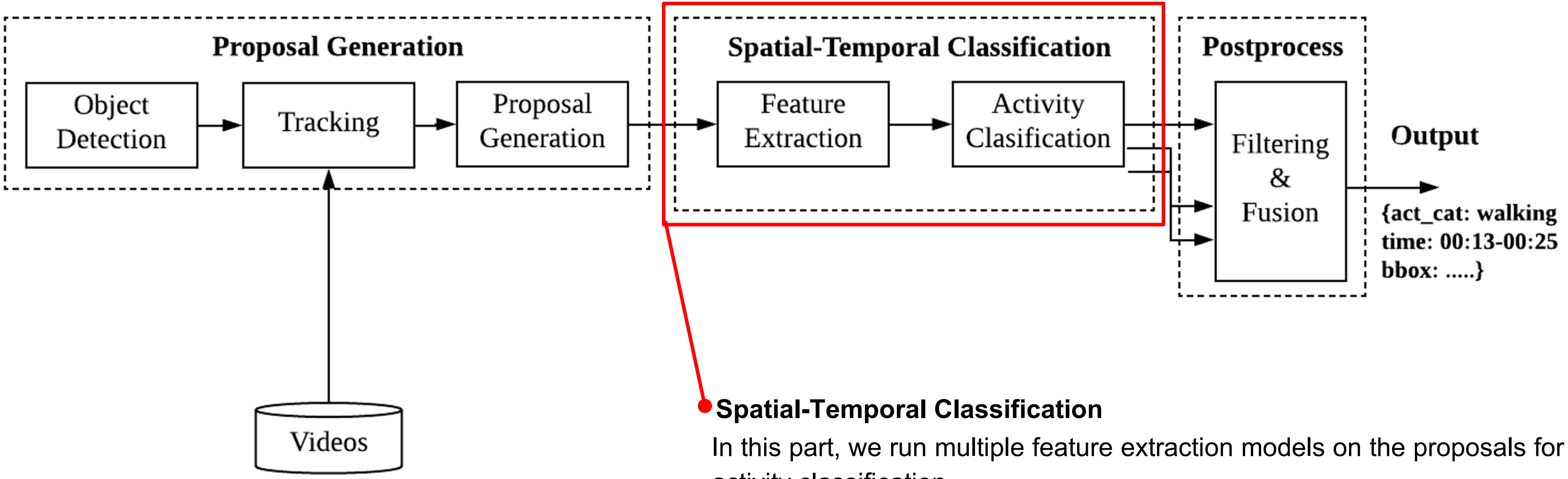
# Overview: The Framework (1/3)



**Proposal Generation**

In the current pipeline, we first detect and generate candidate proposals for action recognition. Object detection model is applied to detect person and vehicle objects. Then, we apply tracking model to create tracklets for all objects. After that, we run a proposal generation model to filter the useless tracklets and provided proposals to the next part.

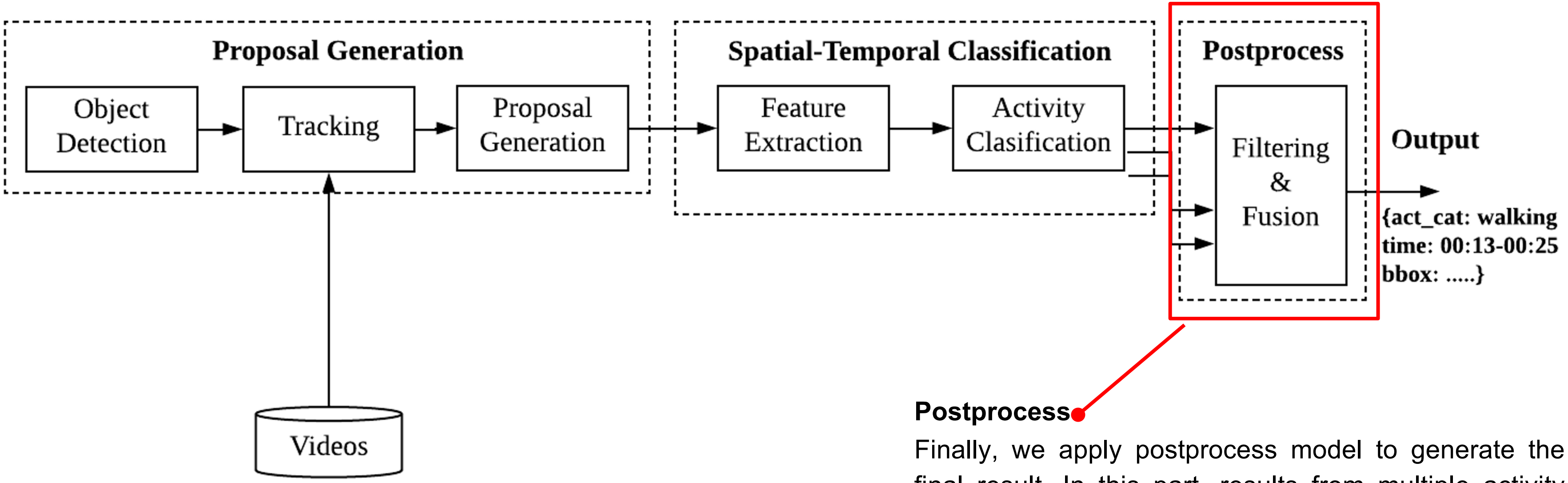
# Overview: The Framework (2/3)



**Spatial-Temporal Classification**

In this part, we run multiple feature extraction models on the proposals for activity classification.

# Overview: The Framework (3/3)



**Postprocess**

Finally, we apply postprocess model to generate the final result. In this part, results from multiple activity classifier feeded with different features will be filtered and provide an ensembled result.

# Overview: Models

| Models Abbr. | Type | Description |
|---|---|---|
| IOD | Object Detector | Image object detection, it could skip frames by fixed steps |
| VOD | Object Detector | Video object detection, it adaptively skips frames in a batch of frames |
| I3D_RGB | Feature Extractor/classifier | Inflated 3D Convnet Model using RGB image as input |
| I3D_FLOW | Feature Extractor/classifier | Inflated 3D Convnet Model using Optical Flow image as input |
| BI-RNN | Classifier | Bi-directional RNN, it requires feature as input |

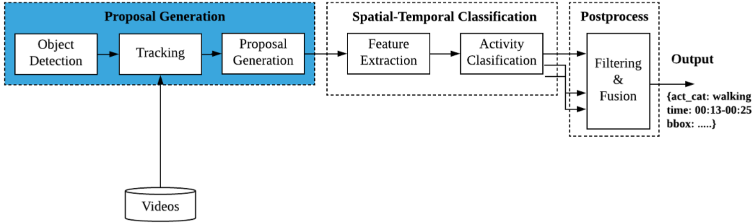
Overview

System

New Works

# Proposal Generation - Objective Detection

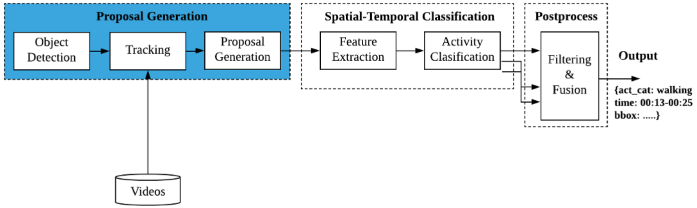
| **Resnet101-FPN-dilation / train on 1920x1080 / test on 1920x1080 / 3.5 fps** | | | | | | | |
|---|---|---|---|---|---|---|---|
| metric | Bike | Construction Vehicle | Door | Person | Prop | Push_Pulled_ Object | Vehicle |
| AP | 0.59 | 0.645 | 0.603 | 0.836 | 0.448 | 0.702 | 0.984 |
| AR | 0.91 | 0.906 | 0.64 | 0.911 | 0.91 | 0.895 | 0.985 |



We utilize faster RCNN Ren et al. (2015) with feature pyramid network Lin et al. (2017) on ResNet-101 He et al.(2016) as the backbone for object detection, in which RoIAlign is used to extract features for Region-of-Interest. We apply object detection on every k frame from the videos. Full resolution images are input to the model and we fine-tune our model using the full 15 object class annotation in the the VIRAT dataset.

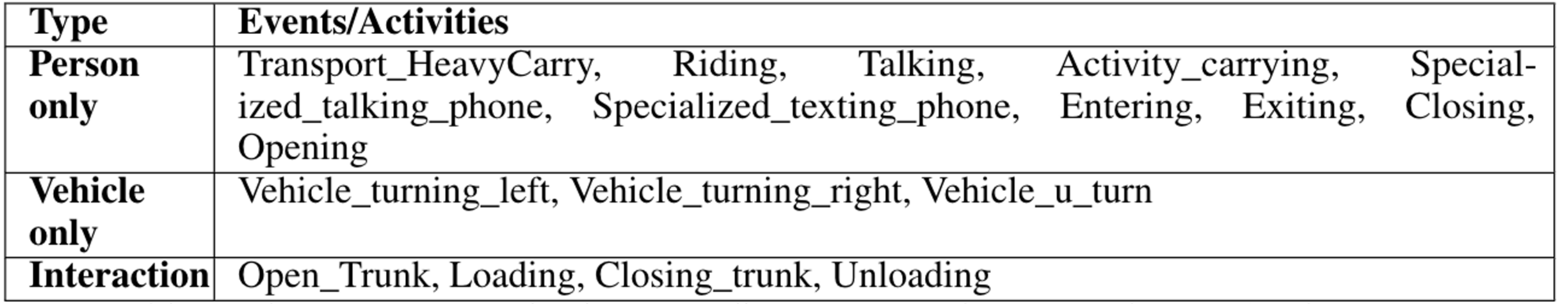
# Proposal Generation - Tracking

| | Recall (%) | Precision (%) | ID Switches | MOTA (%) | MOTL (%) |
|---|---|---|---|---|---|
| KCF | 93.5 | **97.1** | 2519 | 91.3 | 90.5 |
| Deep-sort | **95.2** | 96.5 | **909** | **91.7** | **91.8** |



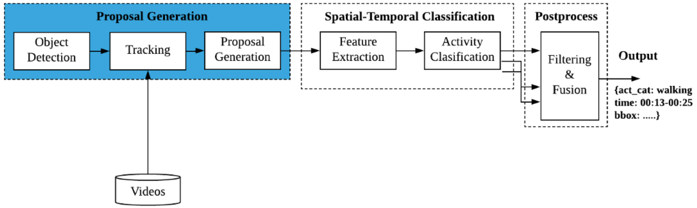
We utilize deep SORT Wojke et al.(2017) to generate tracklets by associating detected objects across frames. We follow a similar track handling and Kalman filtering framework Wojke et al.(2017). We use bounding box center position (u,v), aspect ratio γ, height h and their respective velocities in image coordinates as Kalman states. We compute the Mahalanobis distance between predicted Kalman states and newly arrived measurement to incorporate motion information. For each bounding box detection, we use the feature obtained from object detection module as a appearance descriptor. We compute the cosine distance between tracks and detections in appearance space.

# Proposal Generation - Event Proposal Generation



| Type | Events/Activities |
|---|---|
| **Person only** | Transport_HeavyCarry, Riding, Talking, Activity_carrying, Specialized_talking_phone, Specialized_texting_phone, Entering, Exiting, Closing, Opening |
| **Vehicle only** | Vehicle_turning_left, Vehicle_turning_right, Vehicle_u_turn |
| **Interaction** | Open_Trunk, Loading, Closing_trunk, Unloading |

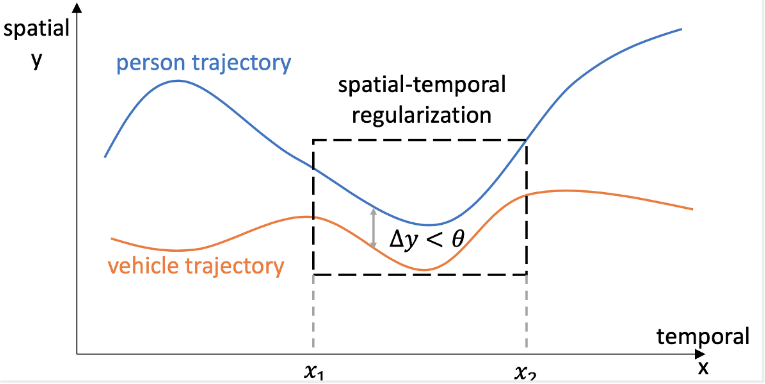
Table 1: The events categorization according to proposal types on the VIRAT dataset.

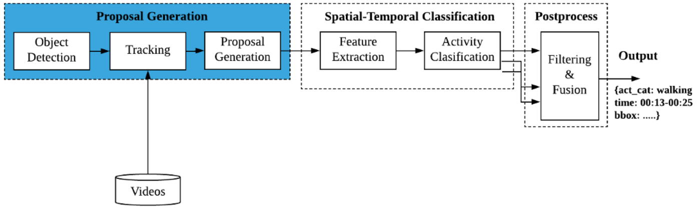


1. The person and vehicle only proposals contains only events happened on a single object (i.e., either a person or a vehicle).

1. To generate proposals of person-vehicle interaction, we associate individual person and vehicle to model their interactions.

# Proposal Generation - Event Proposal Generation

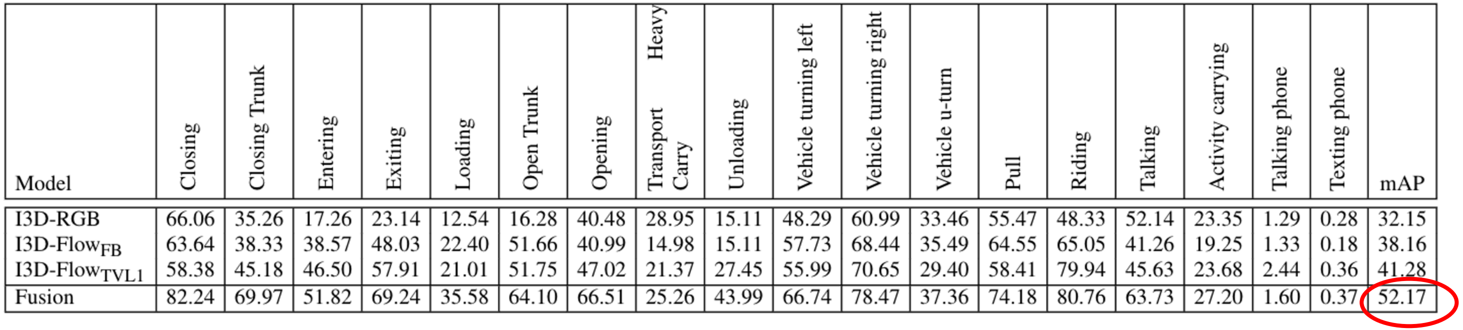




We use a spatial-temporal regularization schema to obtain the interaction proposals.

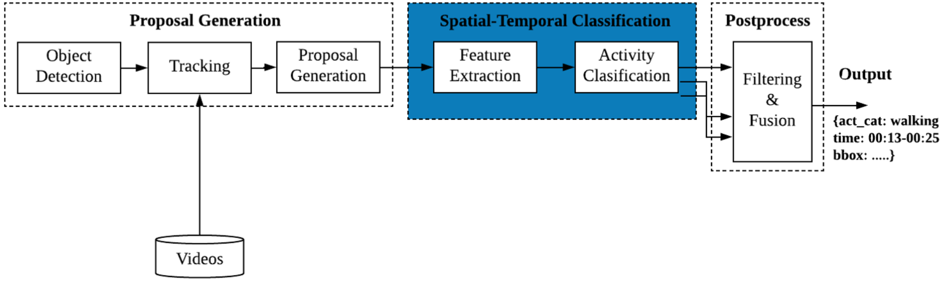
- Let the blue curve be the person trajectory and the red curve be the vehicle trajectory. The x-axis is the time dimension and the y-axis is the spatial dimension.
- In the black dashed line region, the spatial distance between person and vehicle trajectories are consistently close enough in space within the temporal window [x1, x2].
- Finally, we use this regularization to generate event proposals from two object trajectories.

# Spatial-Temporal Classification - Feature and Fusion



| Model | Closing | Closing Trunk | Entering | Exiting | Loading | Open Trunk | Opening | Transport Heavy Carry | Unloading | Vehicle turning left | Vehicle turning right | Vehicle u-turn | Pull | Riding | Talking | Activity carrying | Talking phone | Texting phone | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I3D-RGB | 66.06 | 35.26 | 17.26 | 23.14 | 12.54 | 16.28 | 40.48 | 28.95 | 15.11 | 48.29 | 60.99 | 33.46 | 55.47 | 48.33 | 52.14 | 23.35 | 1.29 | 0.28 | 32.15 |
| I3D-Flow$_{FB}$ | 63.64 | 38.33 | 38.57 | 48.03 | 22.40 | 51.66 | 40.99 | 14.98 | 15.11 | 57.73 | 68.44 | 35.49 | 64.55 | 65.05 | 41.26 | 19.25 | 1.33 | 0.18 | 38.16 |
| I3D-Flow$_{TVL1}$ | 58.38 | 45.18 | 46.50 | 57.91 | 21.01 | 51.75 | 47.02 | 21.37 | 27.45 | 55.99 | 70.65 | 29.40 | 58.41 | 79.94 | 45.63 | 23.68 | 2.44 | 0.36 | 41.28 |
| Fusion | 82.24 | 69.97 | 51.82 | 69.24 | 35.58 | 64.10 | 66.51 | 25.26 | 43.99 | 66.74 | 78.47 | 37.36 | 74.18 | 80.76 | 63.73 | 27.20 | 1.60 | 0.37 | 52.17 |

Table 2: Activity recognition results on the VIRAT testing set. (Higher is better)



We learn proposal-augmented I3D-Flow and I3D-RGB features by fine-tuning I3D Carreira and Zisserman (2017) models for activity recognition on VIRAT. The base models are pre-trained on ImageNet, Kinetics-600 Kay et al.(2017), and Charades Sigurdsson et al.(2016). We fine-tune on the VIRAT dataset with the annotated positive event proposals and 5-times non-trivial background proposal as the negatives. We extract raw RGB and two types of raw optical flow frames (TVL1and Farneback) from the spatial-temporal proposals for fine-tuning. The proposals are augmented by randomly scaling proposal in the temporal and spatial domain. After fine-tuning, we use the last convolutional layer as the feature for classification.

# Spatial-Temporal Classification - Filtering

## Filtering

mAP 41.2 —> 40.9

mean w-pmiss 76.92—> 79.79

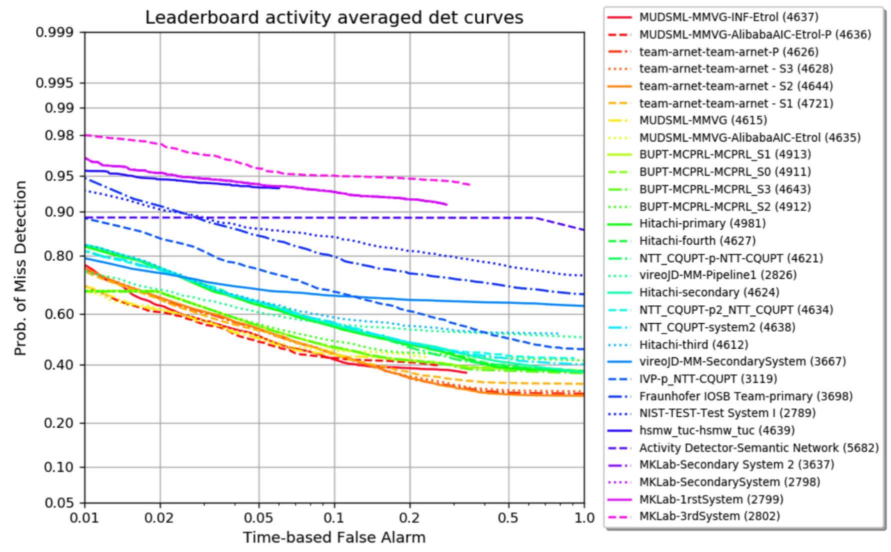
pAUDC 0.495 —> 0.48

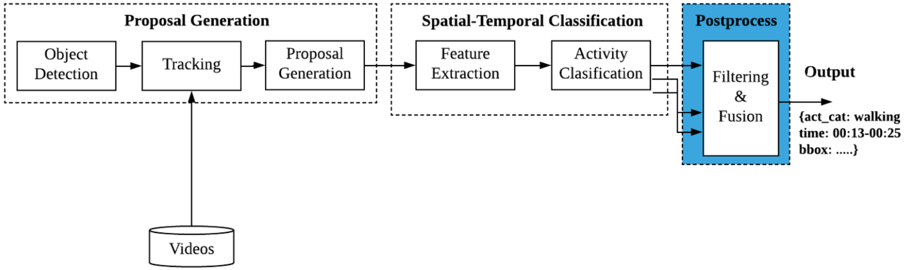
- Delete events that not match objective detection result
  (e.g., object: person vs. event: vehicle u-turn)

- Shorten events by increasing the threshold of confidence score

TRECVID

Updated: 2019-11-06 16:51:21 -05(





Partial AUDC is the area under the DET curve between a Time-based False Alarm rate of 0 and 0.2. Value of a perfect system is 0.

# Conclusion

1、Fusion can improve the performance through combining different information.

2、Using stricter filtering strategy can get more precise temporal boundary and decrease AUDC.

3、Quality is more important than quantity.

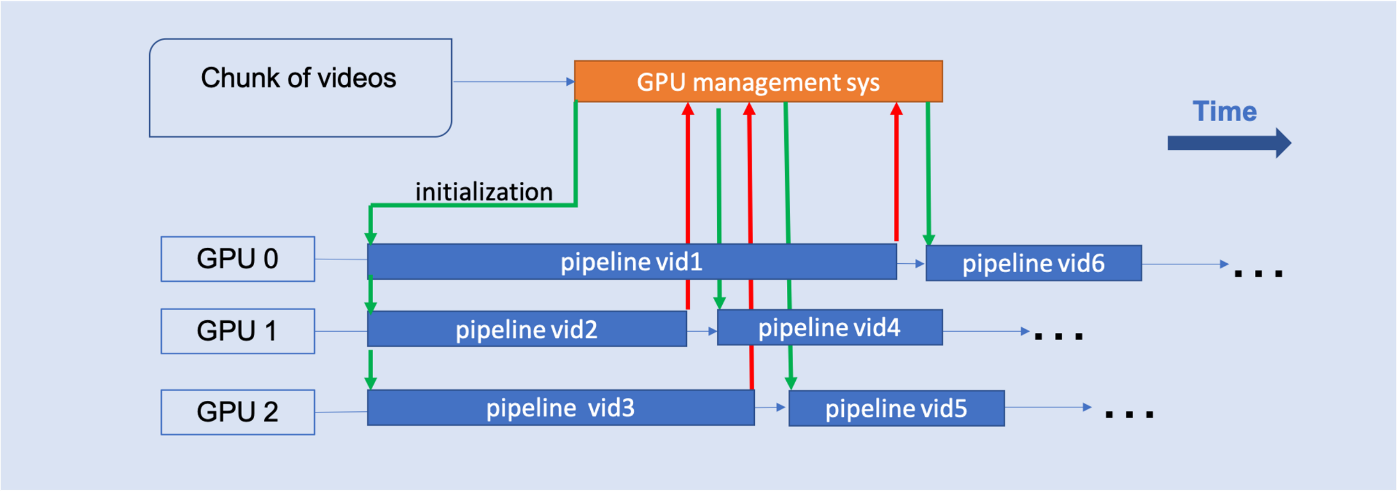
Overview

System

New Works

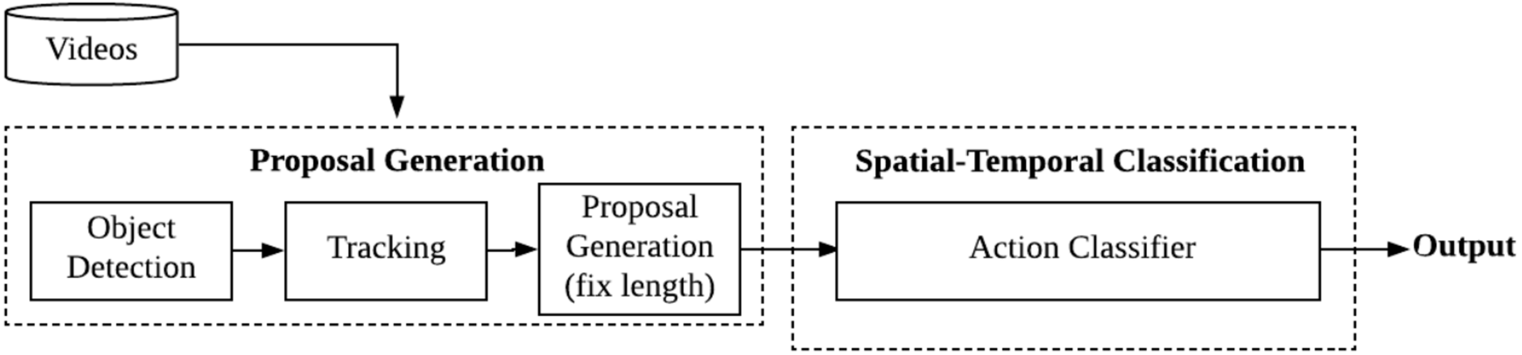
# Dynamic GPU Scheduler

**The GPU management system**



□ Free GPU A GPU Occupied by pipeline A. the length of the bar illustrates the running time.

# Faster System for SDL



**Performance**

- Performance achieved on SDL: pAUDC 0.48978 on SDL [MEVA], 0.646 x realtime.
- Fix-length proposals (90 frames) + action classifier is better than spatial-temporal classifier
  - SDL: [Fix-length+calssification] pAUDC 0.48 vs  [Vary-length + spatial-temporal calssification ] pAUDC 0.85

# Acknowledgement

This project is part of the Deep Intermodal Video Analytics (DIVA) project, supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00340.

# Q & A

## References

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

Xiaojun Chang, Zhigang Ma, Ming Lin, Yi Yang, and Alexander G. Hauptmann. 2017a. Feature Interaction Augmented Sparse Learning for Fast Kinect Motion Detection. *IEEE Trans. Image Processing* 26, 8 (2017), 3911–3920.

Xiaojun Chang, Yaoliang Yu, Yi Yang, and Eric P. Xing. 2017b. Semantic Pooling for Complex Event Analysis in Untrimmed Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 8 (2017), 1617–1632.

Jia Chen, Po-Yao Huang, Jiang Liu, Junwei Liang, Ting-Yao Hu, Wei Ke, Wayner Barrios, Vaibhav, Xiaojun Chang, Huang Dong, Alexander Hauptmann, Shizhe Chen, and Qin Jin. 2018. Informedia @ TRECVID 2018: Ad-hoc Video Search, Video to Text Description, Activities in Extended video. In *Proceedings of TRECVID 2018*. NIST, USA.

J. Chen, J. Liu, J. Liang, T. Hu, W. Ke, W. Barrios, D. Huang, and A. G. Hauptmann. 2019. Minding the Gaps in a Video Action Analysis Pipeline. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. 41–46. `https://doi.org/10.1109/WACVW.2019.00015`

Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S Davis, and Yan Qiu Chen. 2017. Temporal context network for activity localization in videos. In *Proceedings of the IEEE International Conference on Computer Vision*. 5793–5802.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. 2018. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 3–19.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2117–2125.

Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. 2011. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*. IEEE, 3153–3160.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*. Springer, 510–526.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *ICIP*.

Huijuan Xu, Abir Das, and Kate Saenko. 2017. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE international conference on computer vision*. 5783–5792.