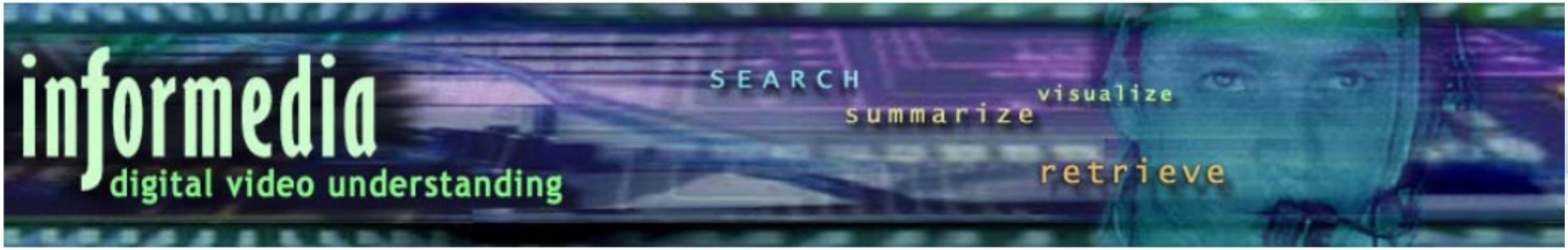


CMU Informedia at TRECVID 2021: Towards Better Spatial-Temporal Activity Detection

Lijun Yu, **Yijun Qian**, Wenhe Liu
and Alexander G. Hauptmann

NIST ActEV

Carnegie Mellon University
Language Technologies Institute

DIVA

Overview

System/Team	$nAUDC@0.2T_{fa}\downarrow$	Mean $P_{miss}@0.15T_{fa}\downarrow$	Mean $wP_{miss}@0.15R_{fa}\downarrow$	
Argus++ (Ours)	**0.39607**	**0.30622**	0.81080	1st
BUPT	0.40853	0.32489	**0.79798**	
UCF	0.43059	0.34080	0.86431	
M4D	0.84658	0.79410	0.88521	
TokyoTech_AIST	0.85159	0.81970	0.94897	
Team UEC	0.96405	0.95035	0.95670	

TRECVID 2021 Leaderboard*

*Snapshot: *Updated: 2021-10-25 11:02:47 -0400 https://actev.nist.gov/trecvid21#tab_leaderboard*

Activity detection

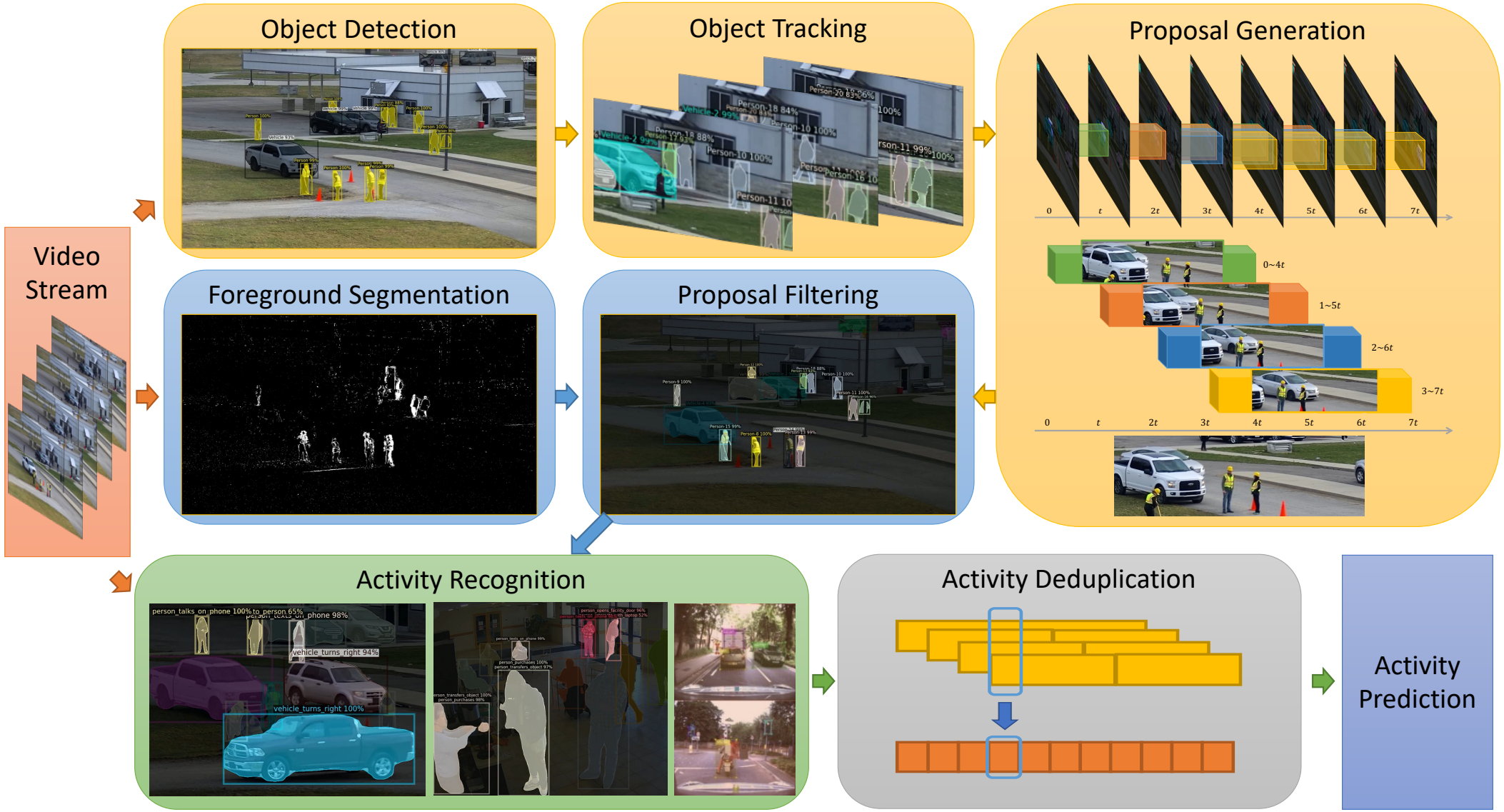
- In ***unconstrained*** videos:
  untrimmed and with large field-of-views
- Three aspects
  - Temporal localization
  - Spatial localization
  - Action classification

Target

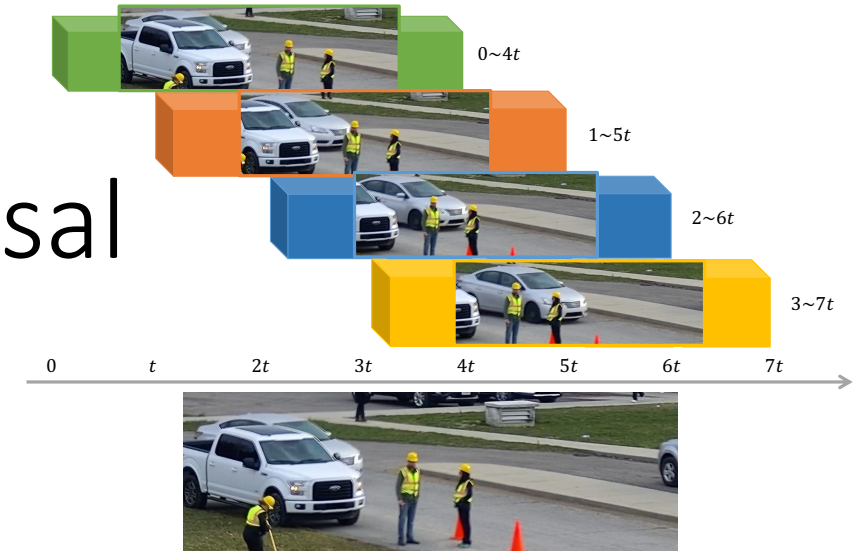
- Detect either atomic activities (e.g., standing up) or continuous repetitive activities (e.g., walking)
- Match multiple non-overlapping predictions to each ground truth

Argus++ Framework

Argus++ Architecture



Intermediate Concept: Cube Proposal



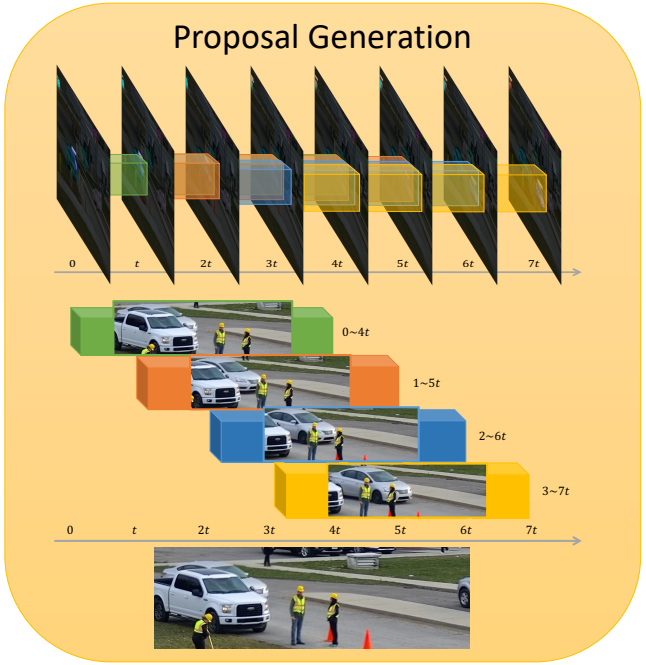
- Proposal
  - A candidate region where activity may occur
  - Processing element for activity recognition

- Spatio-temporal cube proposal
  - A simple six-tuple defining the boundaries in three dimensions

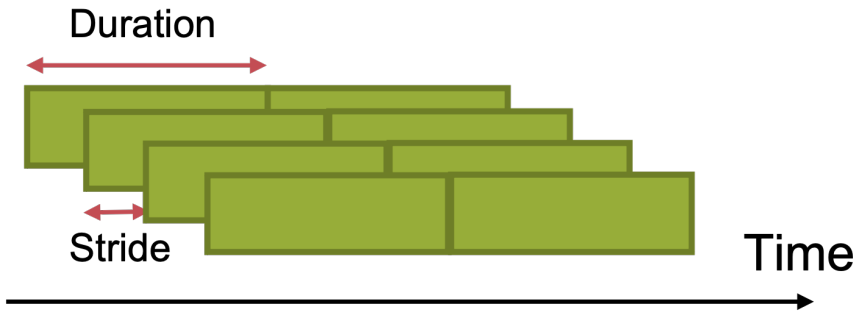
$$p_i = (x_0^i, x_1^i, y_0^i, y_1^i, t_0^i, t_1^i)$$

  - Fixed temporal duration when sampled
  - Much simpler than activity instances or tube proposals

Proposal Generation



- Proposal sampling
  - Dense overlapping sampling on untrimmed videos
  - Ensure completeness and coverage of any activity instance

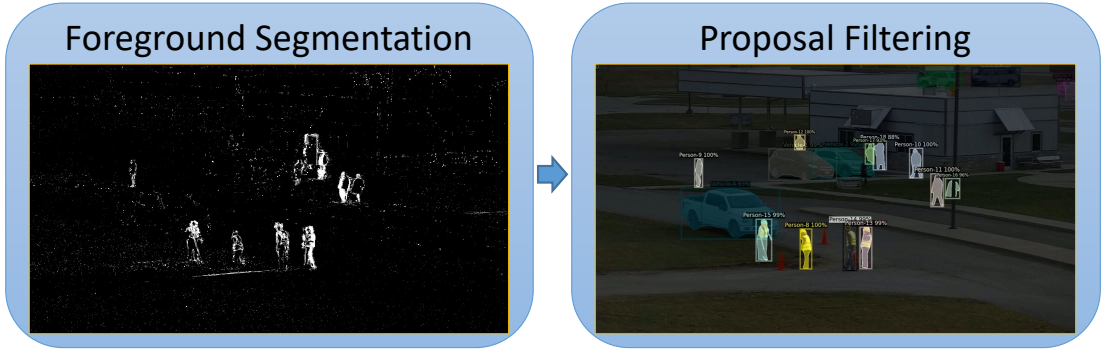


- Proposal refinement
  - Seed track ids from central frame in each temporal window
  - Enlarge bounding boxes as union across the window

$$(x_0, x_1, y_0, y_1)_k = \bigcup(\{(x_0, x_1, y_0, y_1)_{i,j} \mid t_0 \le i \le t_1, tr_{i,j} = tr_{t_c,k}\})$$
$$k = 1, \cdots, n_{t_c}$$

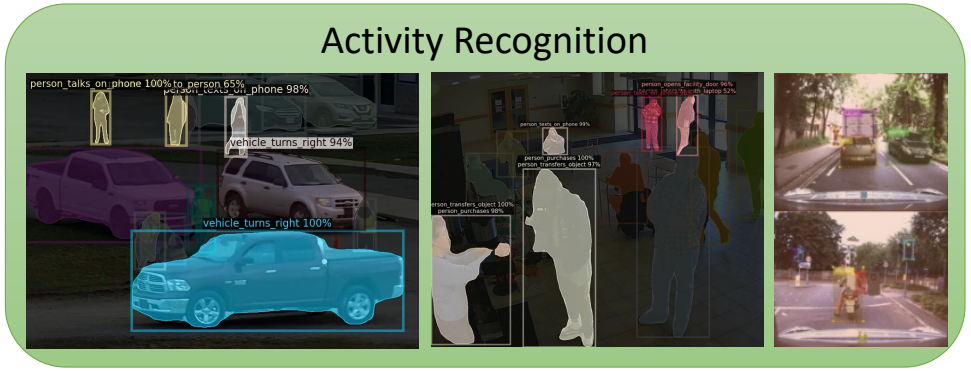
Proposal Generation: An Example

Proposal Filtering

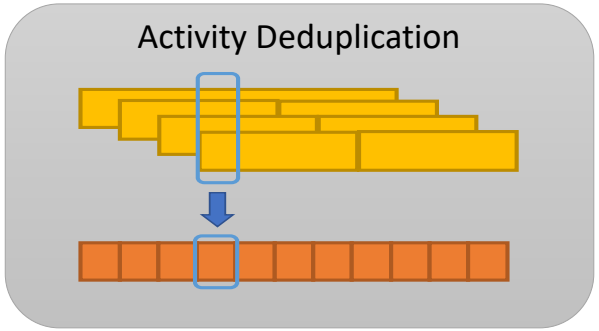


- Foreground segmentation
  - Frame-level binary mask for foreground pixels
  - Proposal foreground score as average value of pixel mask inside the cube
  - Learn a filtering threshold by allowing up to some sacrificed true positive
- Label assignment
  - Convert annotation into cube format by dense sampling
  - Estimate spatial IoU between proposal and ground truth cubes
  - Follow Faster R-CNN in selecting positive and negative samples
- Proposal evaluation
  - Assume perfect classifier by using assigned labels
  - Pass through following steps and use official metrics to estimate upper bound

Activity Recognition

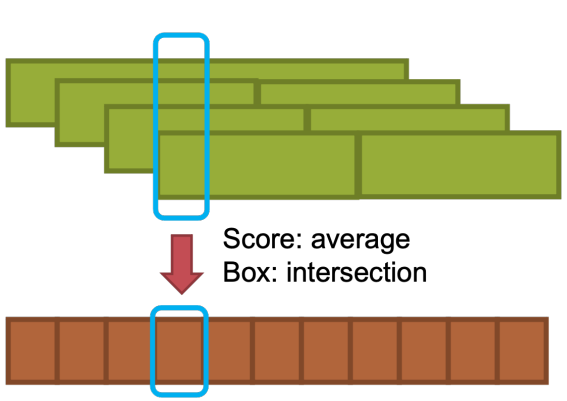


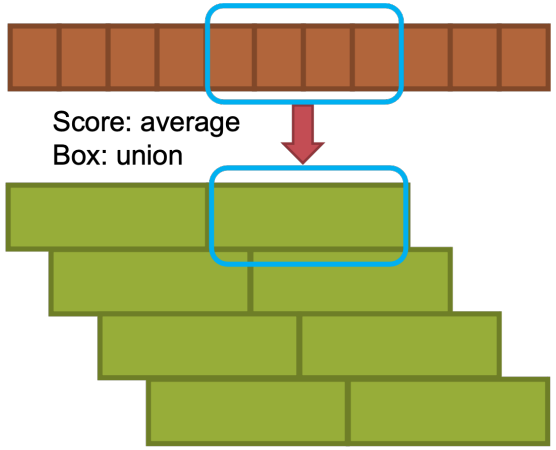
- Multi-label Classification
  - Binary cross entropy loss
  - Weighted by proposal scores
  - Balance activity-wise pos/neg samples
  - Balance samples of different activities
- Action-wise late fusion

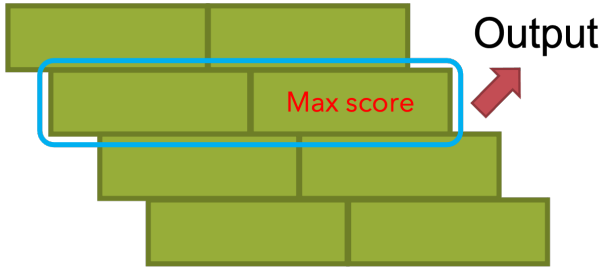


Activity Deduplication

- Overlapping instances







- Adjacent instances
  - Merge adjacent cubes above certain threshold, subject to a minimum duration

Experimental Results

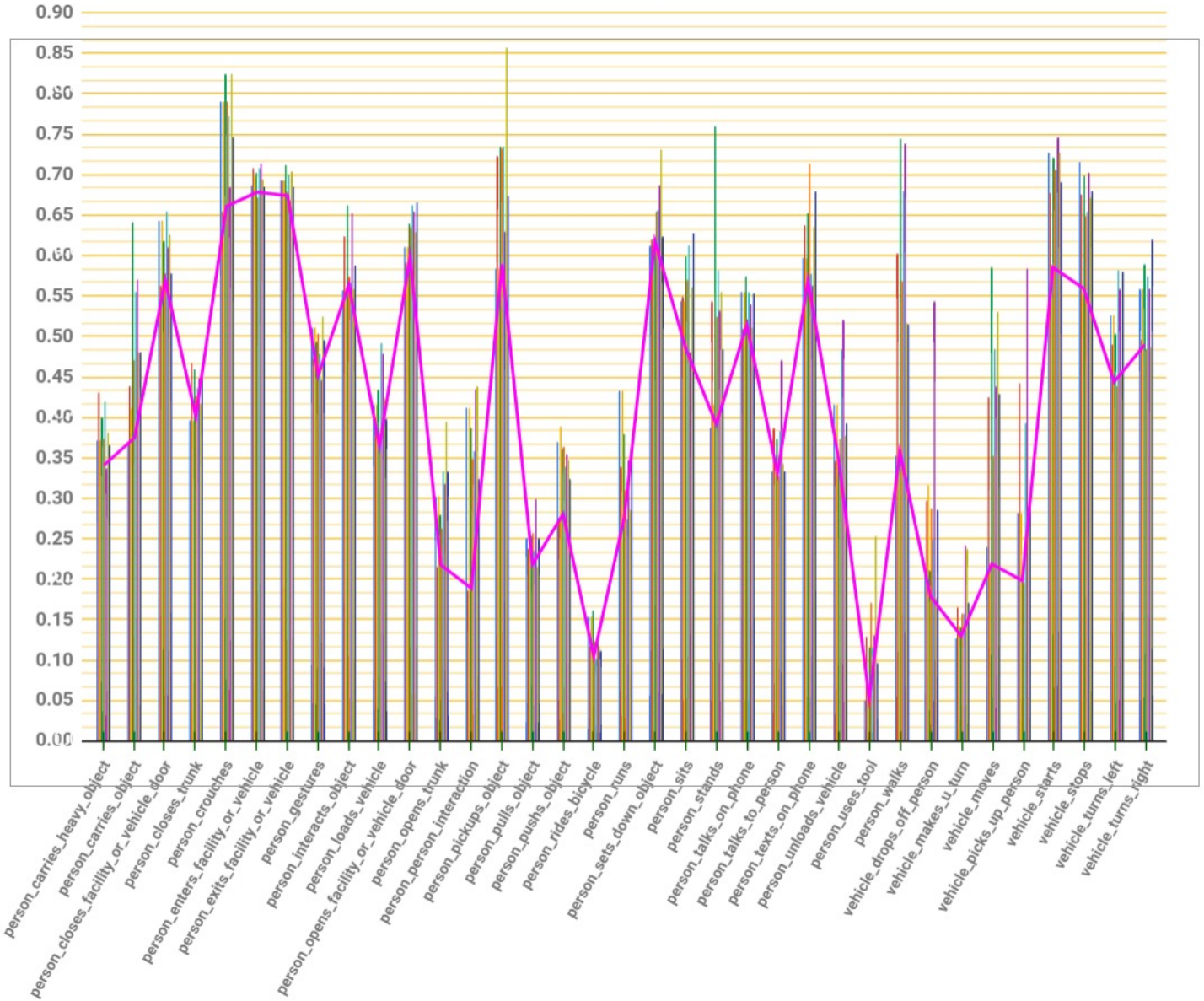
Implementation Details

- Object detection: Mask R-CNN with Resnet-101 on COCO, stride=8
- Multi-object tracking: Towards-Realtime-MOT
- Foreground segmentation: HoG
- Proposal: duration=64, stride=16
- Classifiers: R(2+1)D, X3D, TRM

NIST TRECVID 2021 ActEV

System/Team	$nAUDC@0.2T_{fa} \downarrow$	Mean $P_{miss}@0.15T_{fa} \downarrow$	Mean $wP_{miss}@0.15R_{fa} \downarrow$
Argus++ (Ours)	**0.39607**	**0.30622**	<u>0.81080</u>
BUPT	<u>0.40853</u>	<u>0.32489</u>	**0.79798**
UCF	0.43059	0.34080	0.86431
M4D	0.84658	0.79410	0.88521
TokyoTech_AIST	0.85159	0.81970	0.94897
Team UEC	0.96405	0.95035	0.95670

NIST TRECVID 2021 ActEV -Fusion



CMU best submission (26562) is an action-wise fusion system (the pink curve) and ranks first on the ActEV TRECVID21 Leaderboard.

Ablation Study

- Coverage of Proposal Formats

Table 8. Lower Bounds of $nAUDC@0.2T_{fa}$ on VIRAT Validation Set with different proposal formats. Italic values are non-overlapping proposals while the others are overlapping proposals. Duration and stride are in the unit of frames.

Duration / Stride	16	32	64	96
32	0.0705	*0.1208*	-	-
64	**0.0127**	0.0621	*0.0673*	-
96	0.0275	0.0504	-	*0.0688*

- Performance of Proposal Filtering

Table 9. Statistics of Proposals on VIRAT Validation Set

Name	Unfiltered	Filtered
Number of Proposals	211271	62831
Positive rate	0.1704	**0.5204**
Rate of unique label	0.4558	0.4415
Rate of two labels	0.4127	0.4252
Rate of three labels	0.1017	0.1060

Table 7. Proposal Quality Metrics on VIRAT Validation Set

$nAUDC@0.2T_{fa}$	IoU			Reference Coverage		
Threshold	Average	≥ 0	≥ 0.5	Average	≥ 0.5	≥ 0.9
Unfiltered Proposals	0.2358	0.0772	0.1518	0.1562	0.1125	0.4211
Filtered Proposals	0.2352	0.0772	0.1469	0.1563	0.1099	0.4280

Table 10. Proposal Filter on NIST ActEV'21 SDL Unknown Facility Micro Set

Proposal Filter	$nAUDC@0.2T_{fa} \downarrow$	Processing Time
Enabled	**0.4822**	0.582
Disabled	0.5176	0.925

Conclusion and Future Work

Conclusion and Future Work

- Argus++: Robust Real-time Activity Detection System
- Overlapping Spatio-temporal Cube proposal works
- Action-wise classifier fusion works
- Superior performance in TRECVID ActEV 2020/2021

- Extending to stricter settings: bipartite matching with spatial localization
- Generalizing to more scenarios such as UAV videos
- Zero-shot or Few-shot Activity Detection
- System submission for speed evaluation
- Code validation for reproduction

Acknowledgements

This research is supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00340. This research is supported in part through the financial assistance award 60NANB17D156 from U.S. Department of Commerce, National Institute of Standards and Technology. This project is funded in part by Carnegie Mellon University's Mobility21 National University Transportation Center, which is sponsored by the US Department of Transportation.


Thank You!