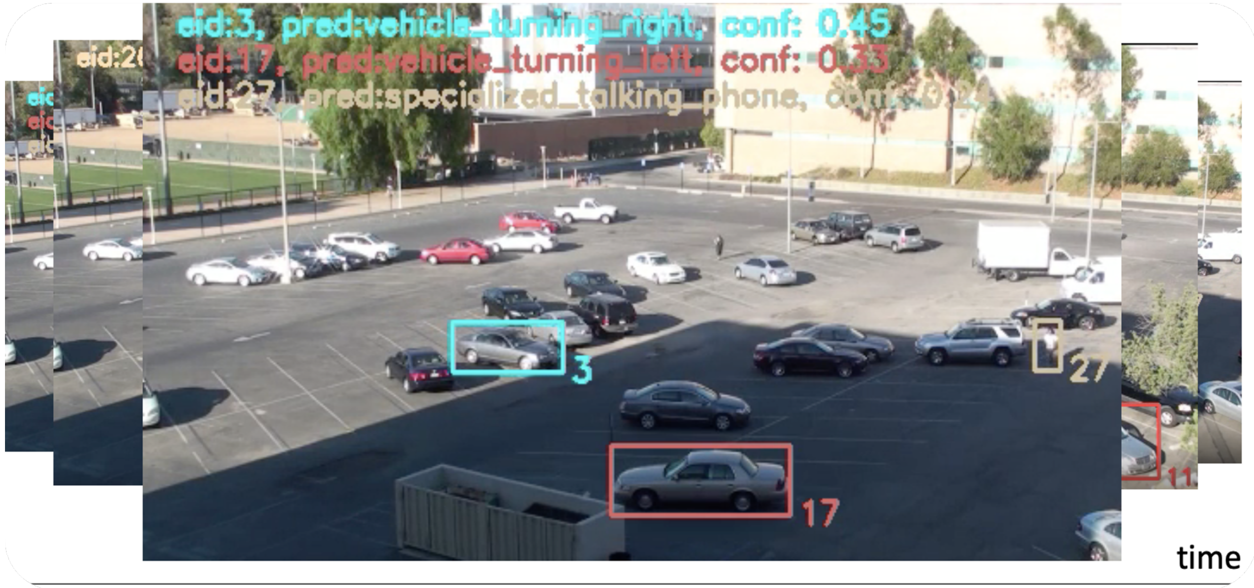
informedia
digital video understanding
SEARCH summarize visualize retrieve

CMU Informedia at TRECVID 2020: Towards Real-time Activity Detection with Dense Spatio-temporal Proposals

Lijun Yu, Yijun Qian, Wenhe Liu, Alexander G. Hauptmann
Carnegie Mellon University

Carnegie Mellon University
Language Technologies Institute

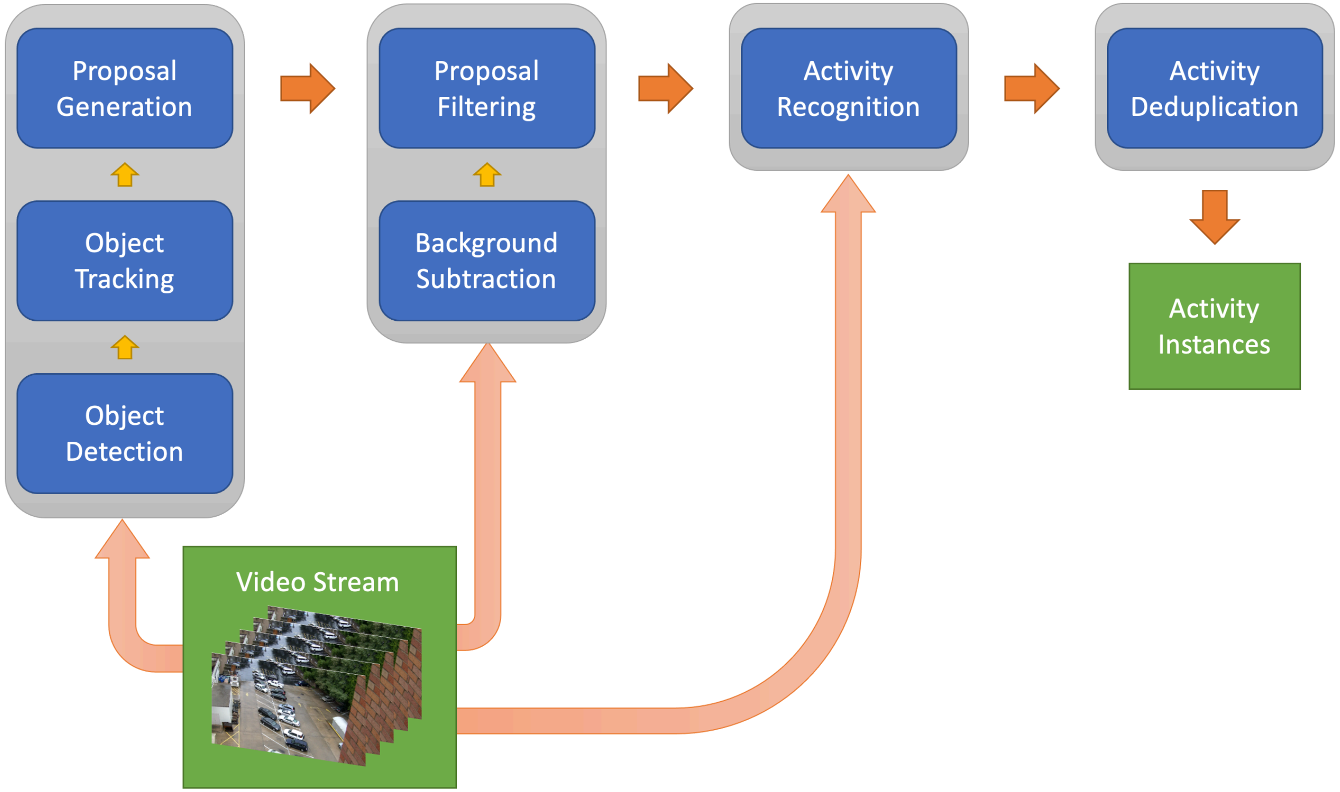
Introduction



- Task: Activities in Extended Videos (ActEV)
- New techniques:
  - Dense spatio-temporal cube proposal paradigm
  - Real-time concurrent framework ***Pyturbo***
  - Temporal Relocation Module (**TRM**) for action recognition
- Achievements:
  - **1st place** in TRECVID-ActEV 2020 with $nAUDC@0.2T_{fa}$=0.42
    23.8% ahead of the runner up system

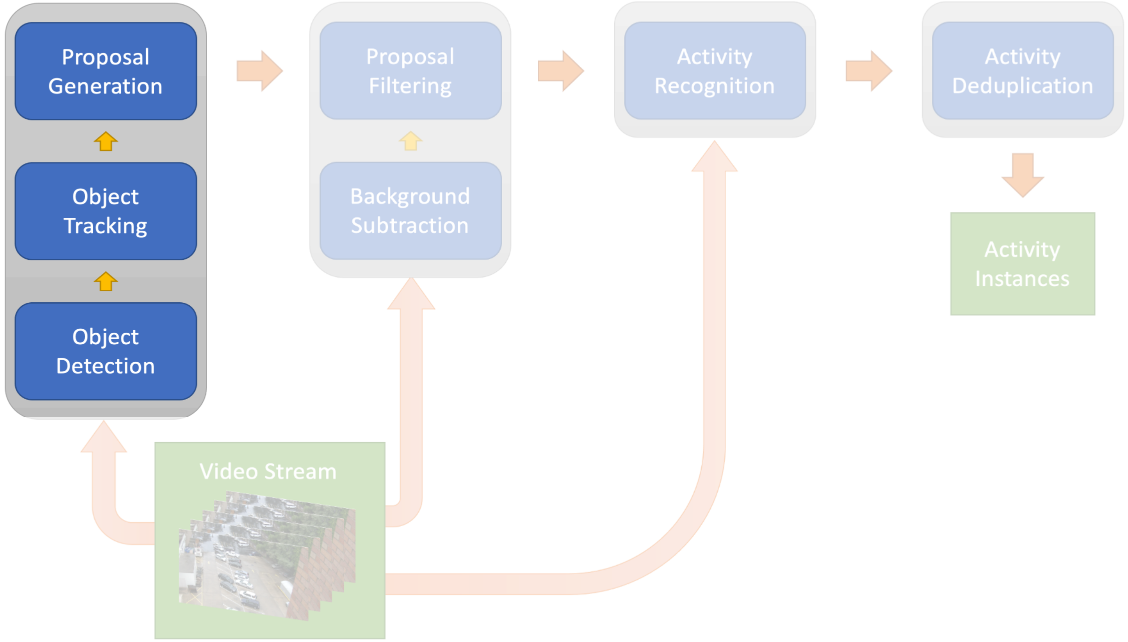
Scan and star at:
https://github.com/CMU-INF-DIVA/pyturbo

Architecture



- Key intermediate concept: *spatio-temporal cube proposal*
- Unified approach for all types of activities
- Maximize speed via concurrent processing on CPUs and GPUs

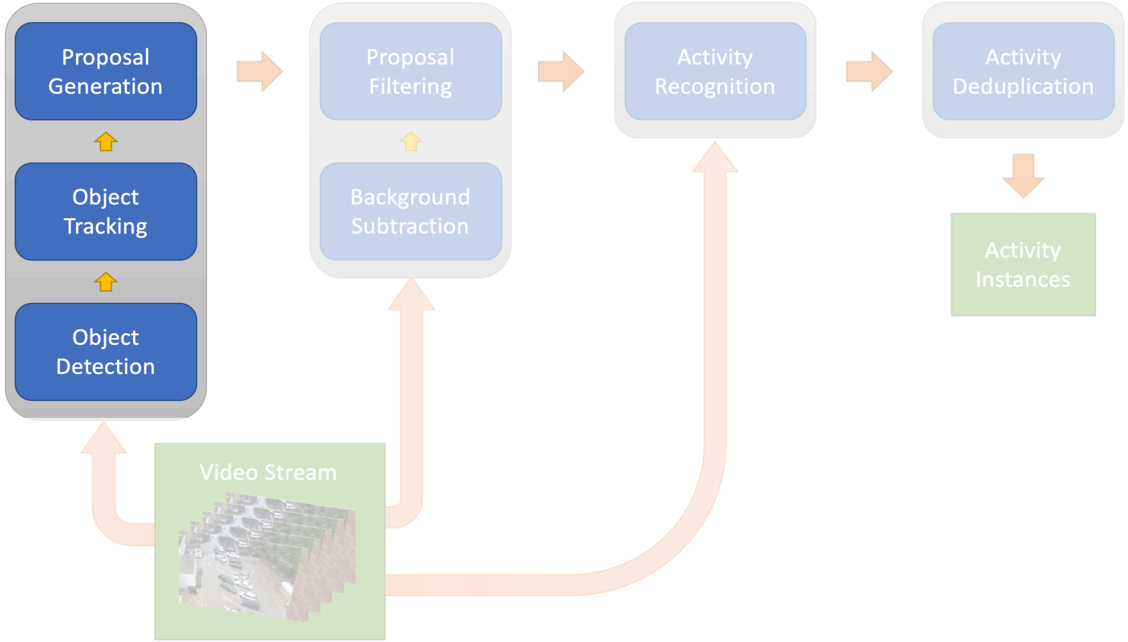
Proposal Generation



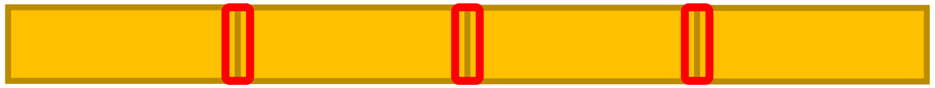
- Detection and Tracking
  - Pretrained frame-level detector
  - Process down-sampled frame sequence
- Proposal Paradigm
  - Previous: *spatio-temporal **tube** proposals*
    - Use whole trajectory of each tracked object
    - Still require temporal localization
    - Object's shape changes when resized for feature extraction
  - New: *spatio-temporal **cube** proposal:*
    - A simple six-tuple defining the boundaries in three dimensions

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

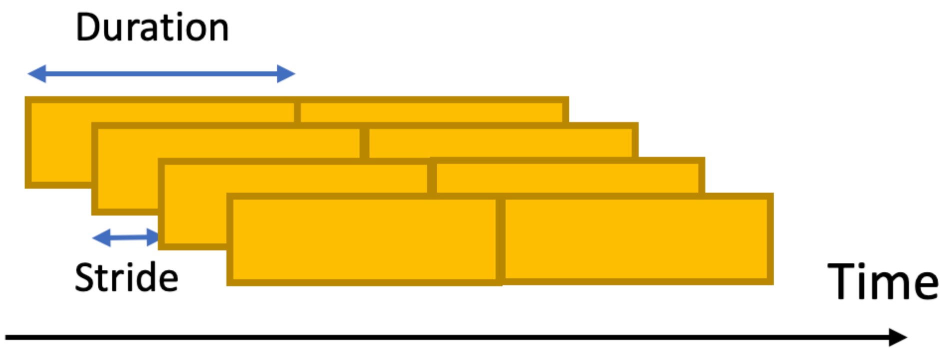
Proposal Sampling



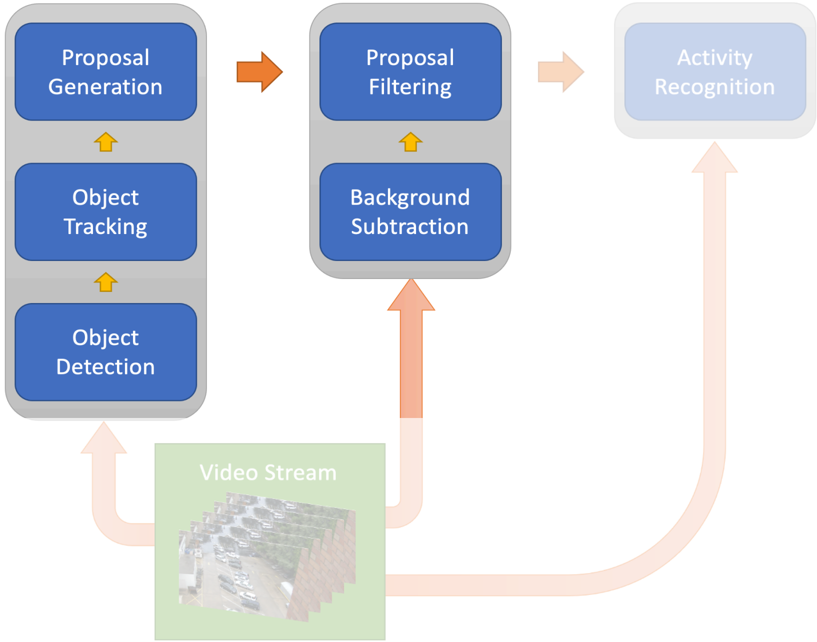
- How to handle untrimmed videos?
- Previous: cut into non-overlapping clips



  - Stride = Duration
  - Significant performance drop at boundaries
- New: **dense overlapping** proposal sampling
  - No *boundary*
  - Stride ≤ Duration



Proposal Refinement

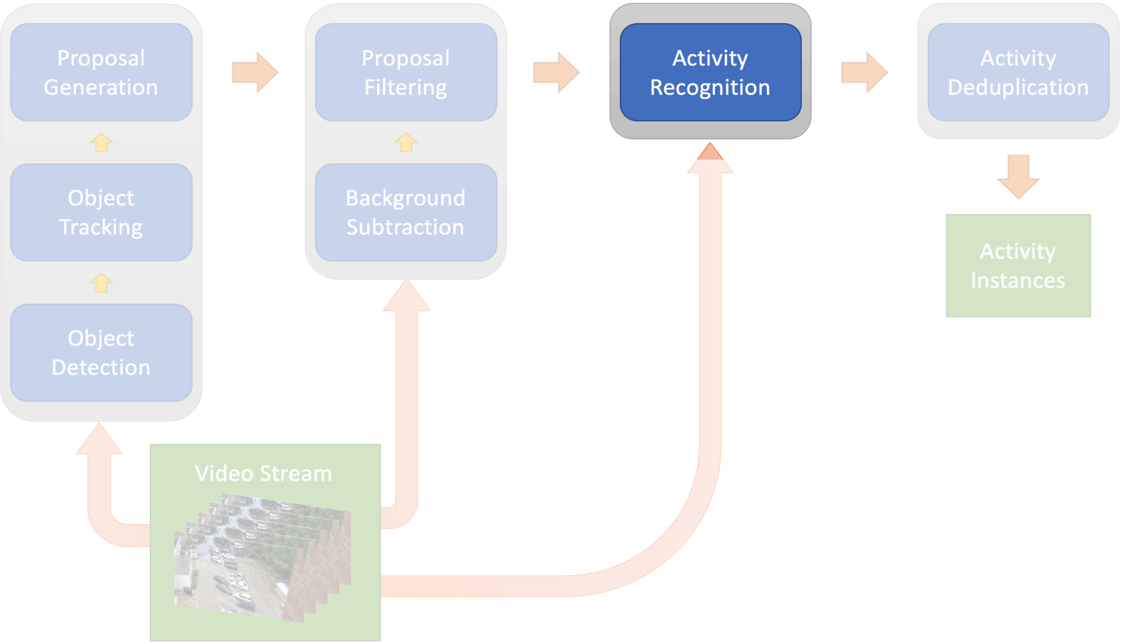


- Spatial localization
  - Extract seed track ids from the central frame
  - Enlarge the bounding boxes as the union of its track

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

  - Robust through identity switch in the tracking algorithm
  - Ensures coverage of moving objects
- Proposal filtering
  - Leverage motion information, filter out stable objects
  - Binary frame masks from foreground segmentation
  - Proposal foreground score as the average value of pixel masks in its cube
  - Learn the filter threshold at a tolerance level of lost positive samples

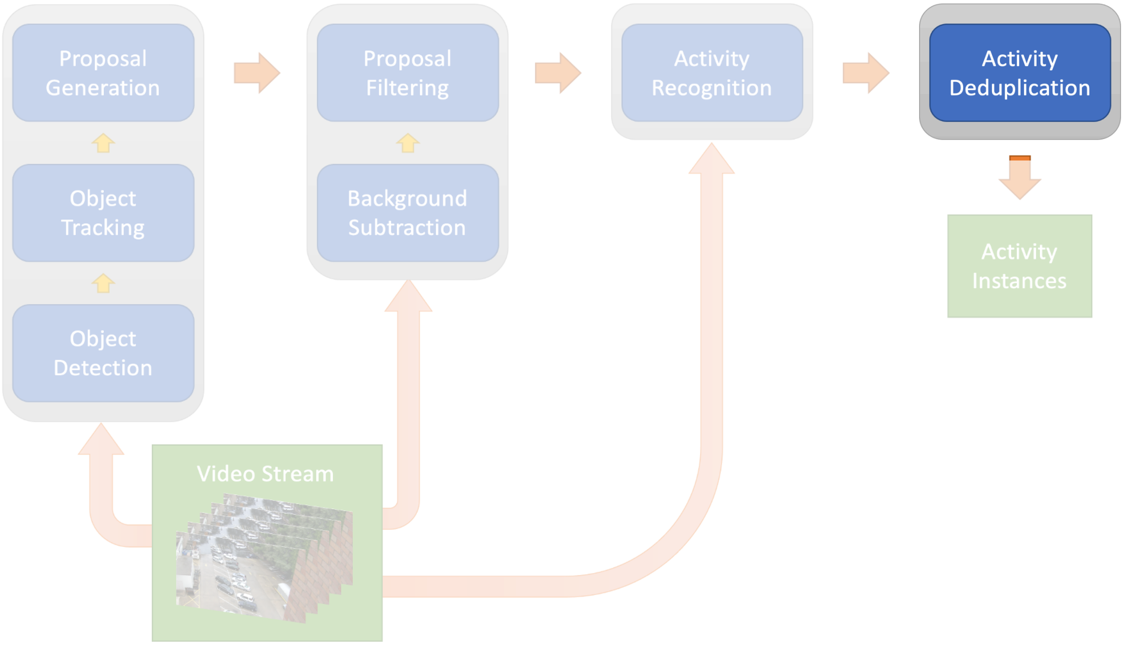
Activity Recognition

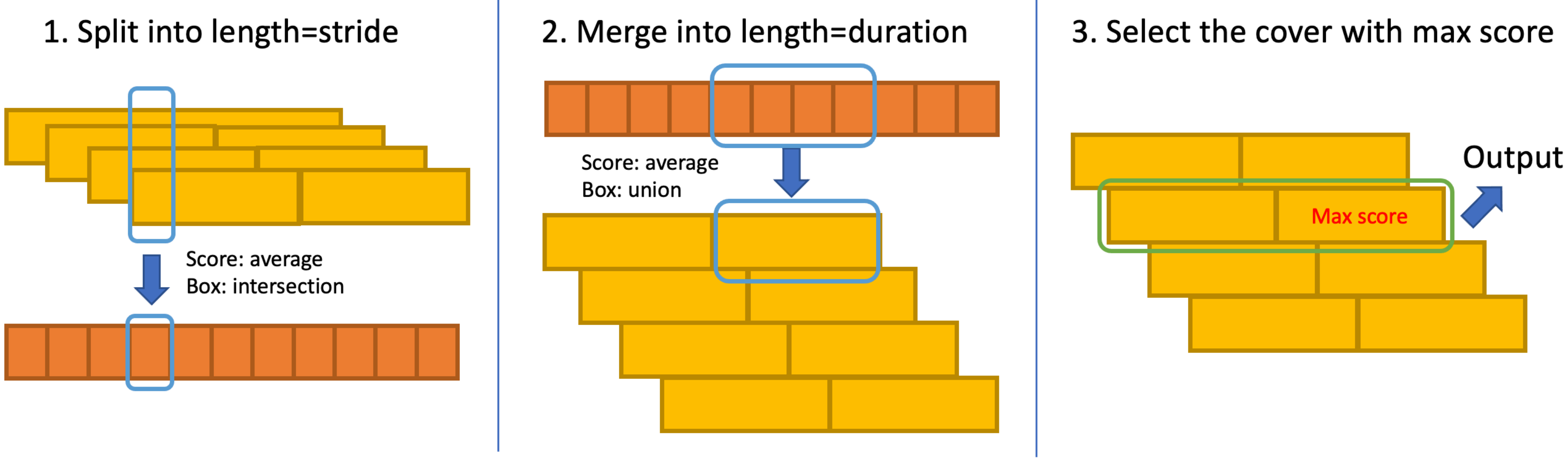


- Sparse frame sampling
- Multi-label Classification
  - Binary cross entropy loss
  - Balance pos/neg samples
  - Balance different classes
- Activity-wise Late Fusion
  - Each of the classifier shows superiority on a subset of actions

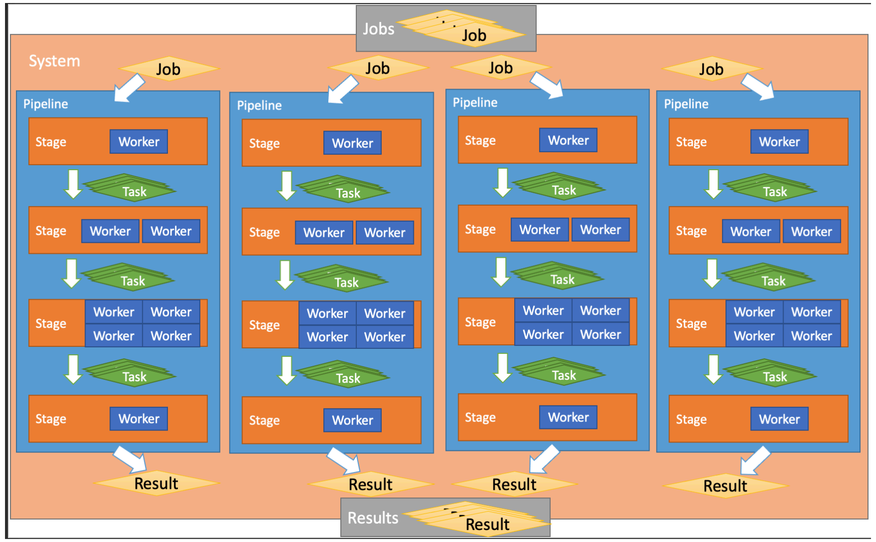
Activity Deduplication

- Remove the duplicate activity instances from overlapping proposals
- Process all proposals in each activity type
- Perform interpolation upon overlapping cubes, maximizing information utilization

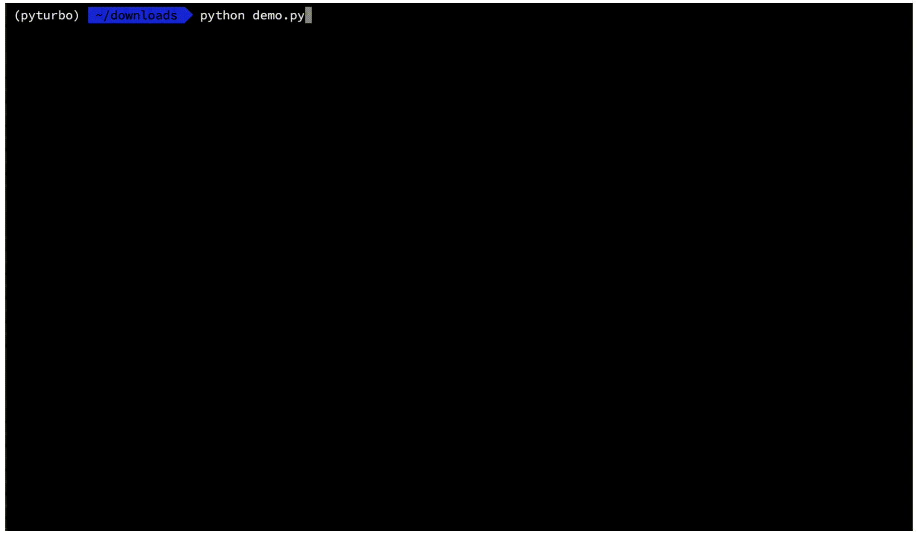




Efficient Concurrent Execution: Pyturbo



- Multiple level of abstraction:
  - worker/stage/pipeline/system
  - job/task/result
- Easy to implement
  Fast to execute
  - Automatic resource allocation
  - Retry and fail-safe mechanisms
  - Run your CPUs and GPUs all to 100%!



Scan and star at: https://github.com/CMU-INF-DIVA/pyturbo

Datasets and Metrics

In TRECVID 2020, a new partition of the VIRAT dataset is introduced.

- With augmented annotation of 35 activities.
- 64 videos for training, 54 videos for validation, 246 videos for testing.

The main metrics are $nAUDC@0.2T_{fa}$ and $P_{miss}@0.15T_{fa}$

- $P_{miss}@0.15T_{fa}$ measures the recall of activity instances within a time limit of all positive frames plus 15% of negative frames.
- $nAUDC@0.2T_{fa}$ is the integration of P-miss on $T_{fa} \in [0, 0.2]$.

Implementation Details

- *Object detector* : Mask R-CNN with ResNet-101 pretrained on MS COCO from Detectron2, applied every 8 frames
- *Object tracker* : reuse RoI feature from detector with associative algorithm from Towards-Realtime-MOT
- *Proposal generation*: duration 64 frames, stride 16 frames
- *Label assignment* : convert VIRAT annotation into cube format and match with spatio-IoU in each temporal window
- *Background filter* : tolerance at 5% positive proposals
- *Activity Classifiers* : R(2+1)D, X3D, Temporal Relocation Module (TRM)

Quality Analysis of Proposals

- Estimate the upper bound performance of proposals
  - Assume we have an ideal classifier
  - Test the capability of proposal paradigm
  - Directly convert the annotations into proposal format and get scored

Performance of proposals on VIRAT validation set

(a) Non-overlapping

Duration (# frame)	$nAUDC@0.2T_{fa}$
32	0.1208
64	**0.0673**
96	0.0688
128	0.0788

(b) Overlapping

Duration / Stride (# frame)	16	32
32	0.0705	-
64	**0.0127**	0.0621
96	0.0275	0.0504

Performance of Proposal Filtering

- Still assume an ideal classifier
- To evaluate spatial alignment of proposals, further filter at IoU and reference coverage levels from 0, 0.1, to 0.9 to get partial results

Proposal statistics on VIRAT validation set

Name	Unfiltered Proposals	Filtered Proposals
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

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

Performance of Classification and Fusion (1)

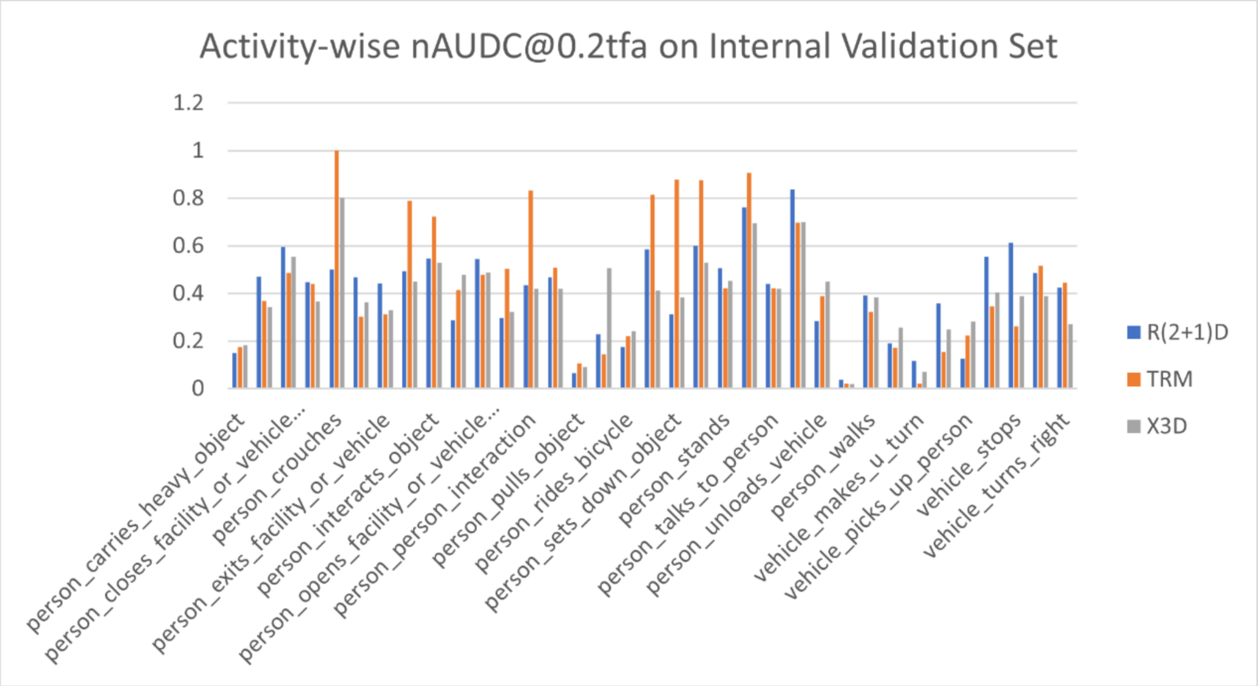
On validation set:

- R(2+1)D is the best
    - The pre-trained model
- Each of the classifiers has its best activities
    - Fusion could work

Proposal statistics on validation set

Model	Pretraining	Input	$nAUDC@0.2T_{fa}$	Mean $P_{miss}@0.15T_{fa}$
R(2+1)D	IG65M	32×112×112	**0.356**	**0.256**
X3D	Kinetics	16×312×312	0.383	0.284
TRM	Kinetics	8×224×224	0.394	0.303

Activity-wise results on validation set



Performance of Classification & Fusion (2)

On test set (leaderboard):

- Fusion helps
- More data is useful (not quite)

Performance on Leaderboard

Model	Training Data	$nAUDC@0.2T_{fa}$
R(2+1)D	Training set	0.438
R(2+1)D	Training+validation sets	0.436
R(2+1)D+TRM	Training set	0.431
R(2+1)D+TRM	Training+validation sets	0.429
R(2+1)D+TRM+X3D	Training set	0.424
R(2+1)D+TRM+X3D	Training+validation sets	**0.423**

Leaderboard Results

- Our $nAUDC@0.2T_{fa}$ is 23.8% better!
- Our $P_{miss}@0.15T_{fa}$ is 31.9% better!

TRECVID 2020 ActEV Leaderboard as of Nov. 20

Rank	Team	Best System	$nAUDC@0.2T_{fa}$	Mean $P_{miss}@0.15T_{fa}$
1	INF	INF (Ours)	**0.42307**	**0.33241**
2	BUPT-MCPRL	MCPRL_S1	0.55515	0.48779
3	UCF	UCF-P	0.58485	0.54730
4	TokyoTech_AIST	TTA-SF2	0.79753	0.75502
5	CERTH-ITI	P	0.86576	0.84454
6	Team UEC	UEC	0.95168	0.95329
7	kindai_kobe	kind_ogu_baseline	0.96820	0.96443

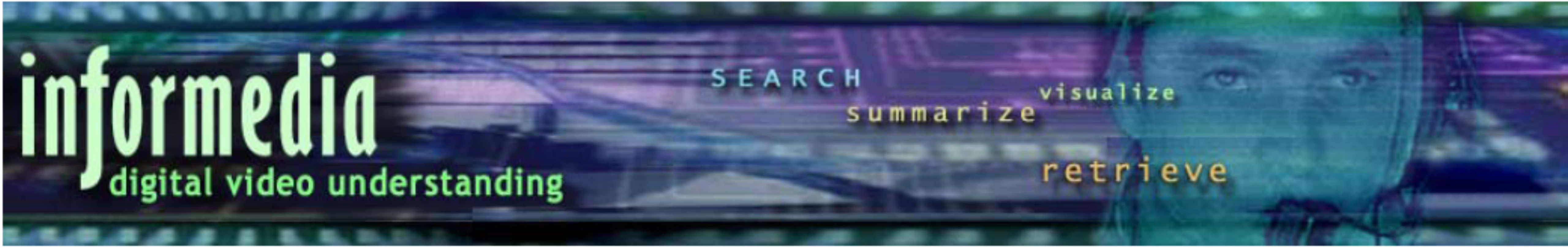
Take Away & Future Work

Lessons:

- Spatio-temporal cube proposal vs.. tube proposal
- Dense overlapping proposal sampling vs. nonoverlapping sampling
- Fusion of classifiers

Prospects:

- Use completely hidden test set to prevent possible bias
- Adopt online inference and evaluation of submitted systems
  - Evaluating both effectiveness and efficiency



CMU Informedia at TRECVID 2020: Towards Real-time Activity Detection with Dense Spatio-temporal Proposals

Thanks for listening!

Carnegie Mellon University
Language Technologies Institute