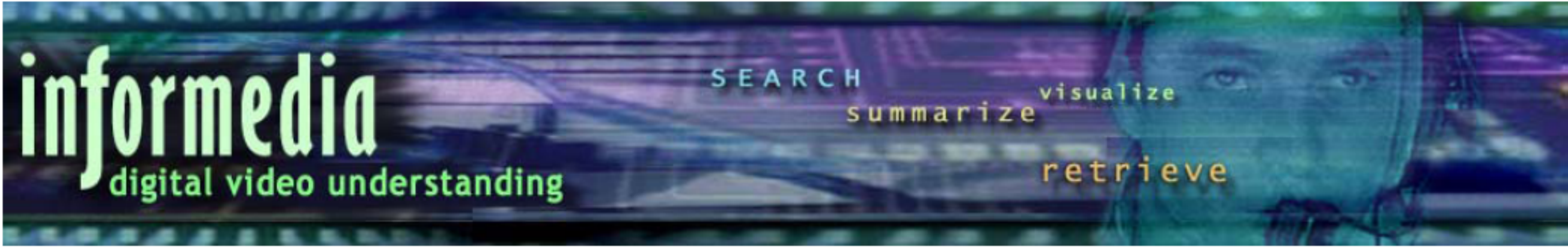


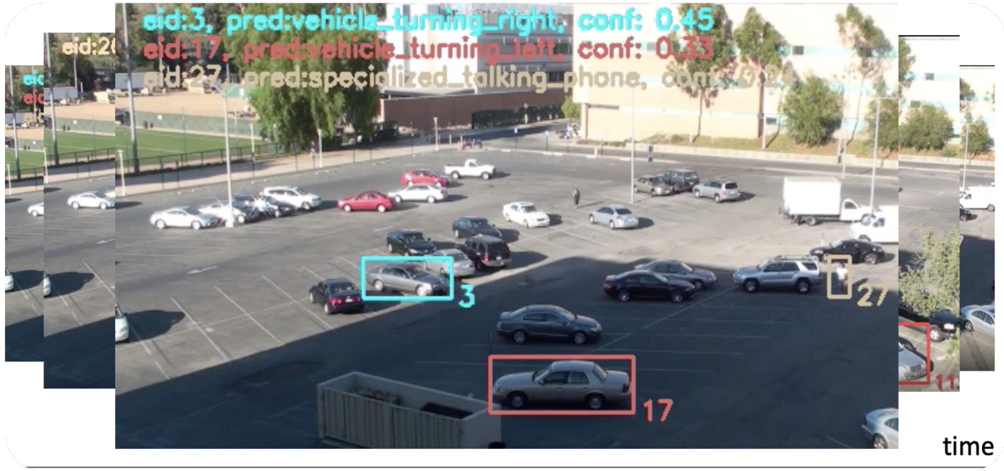
Real-time Activity Detection in Unknown Facilities with Dense Spatio-temporal Proposals

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

Carnegie Mellon University
Language Technologies Institute

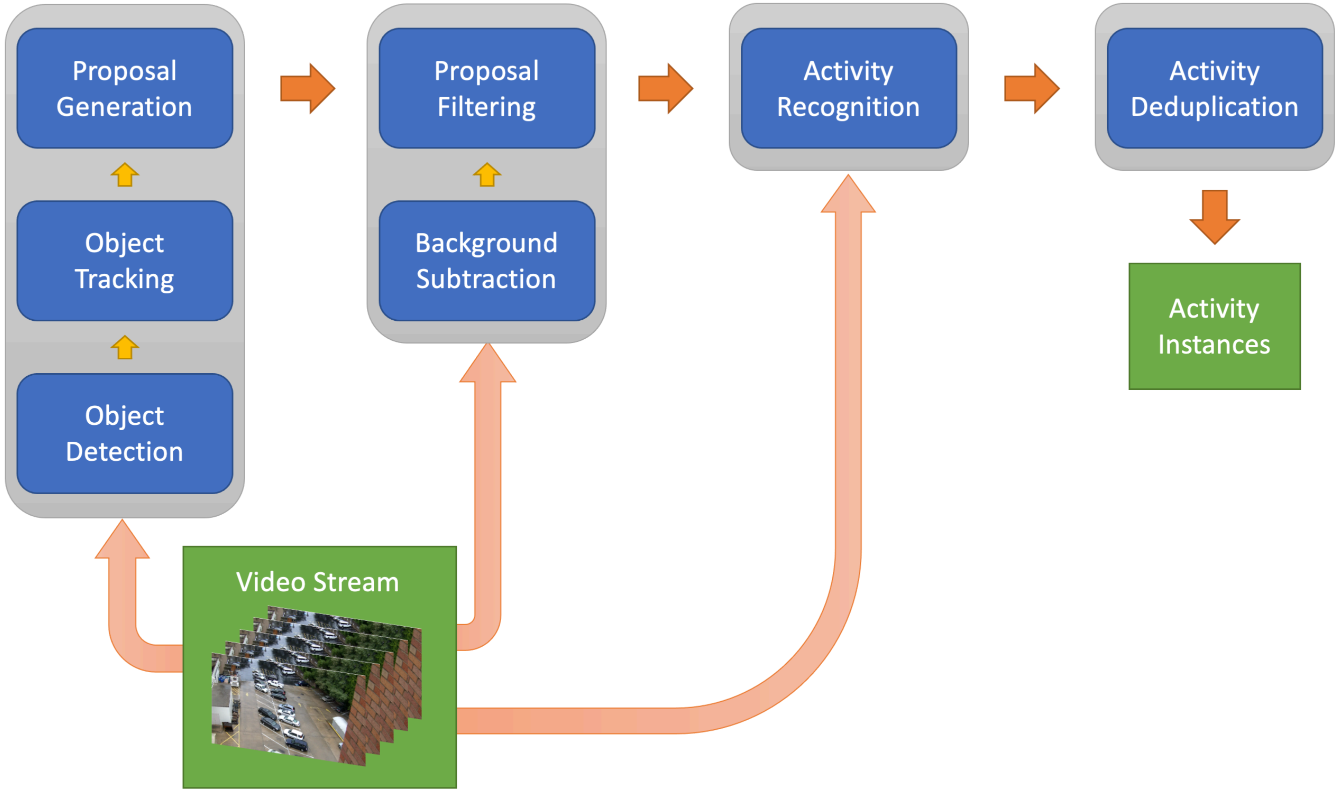
Introduction

- Task: Activities in Extended Videos (ActEV) Sequestered Data Leaderboard (SDL) Unknown Facility (UF)
- New techniques:
  - Dense spatio-temporal cube proposal paradigm
  - Real-time concurrent framework ***Pyturbo***
- Achievements:
  - **1st place** in ActEV SDL UF with $nAUDC@0.2T_{fa}$= 0.428 **22.3% ahead** of the runner up system
  - **1st place** in ActEV SDL Known Facilities (KF) (**32.4% ahead**)
  - **1st place** in TRECVID ActEV (**22.8% ahead**)



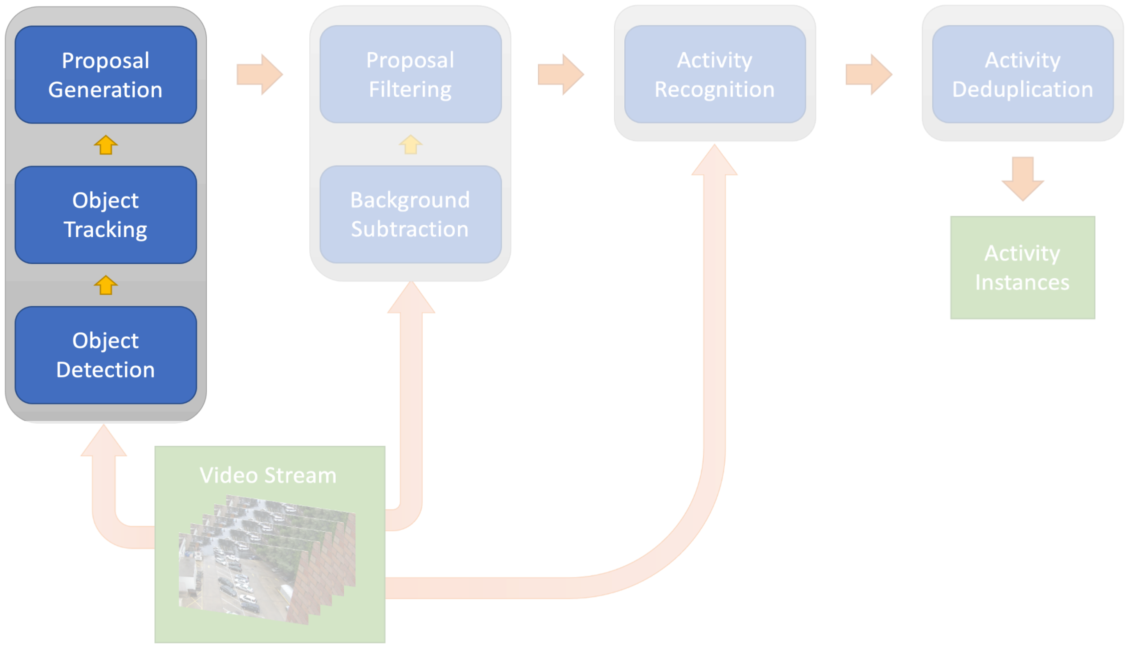
Scan and star at:
https://github.com/CMU-INF-DIVA/pyturbo

The System



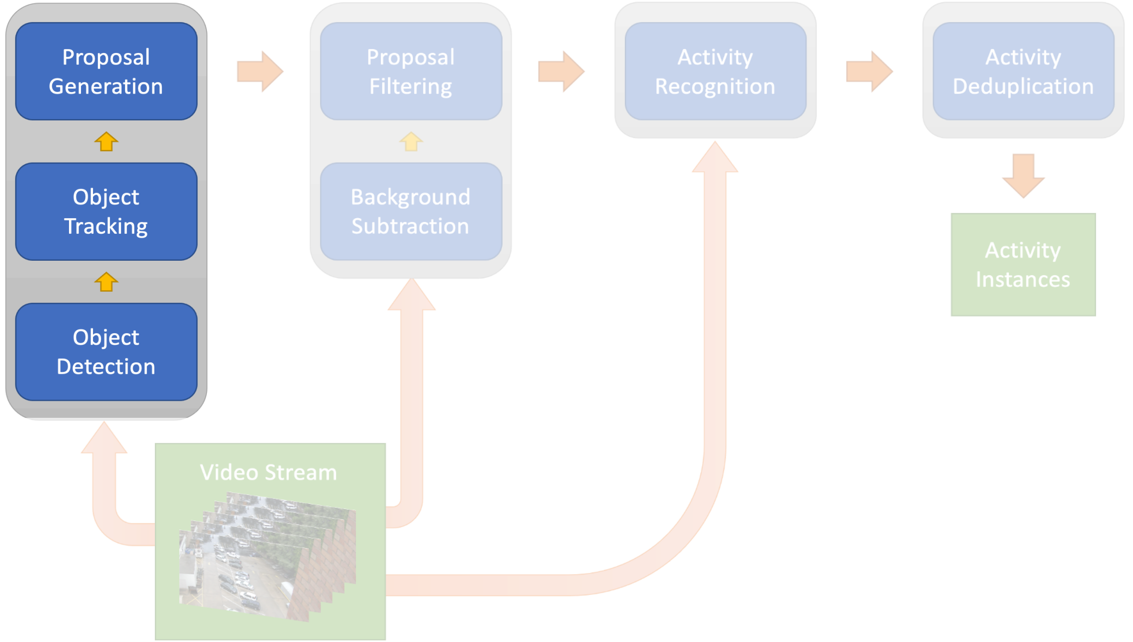
- Key intermediate concept: *spatio-temporal cube proposal*
- Unified approach for all types of activities
- Maximized speed via concurrent processing on CPUs and GPUs

Proposal Generation



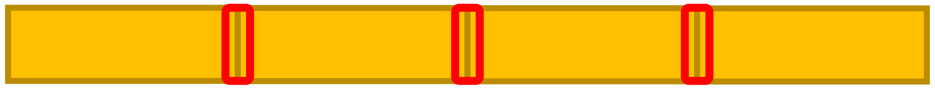
- Detection and Tracking
  - Frame-level detector
  - Process down-sampled frame sequence
- Proposal Paradigm
  - Previous: *spatial-temporal* **tube** *proposals*
    - Use whole trajectory of each tracked object
    - Still require temporal localization
    - Object's shape changes when resized for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

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

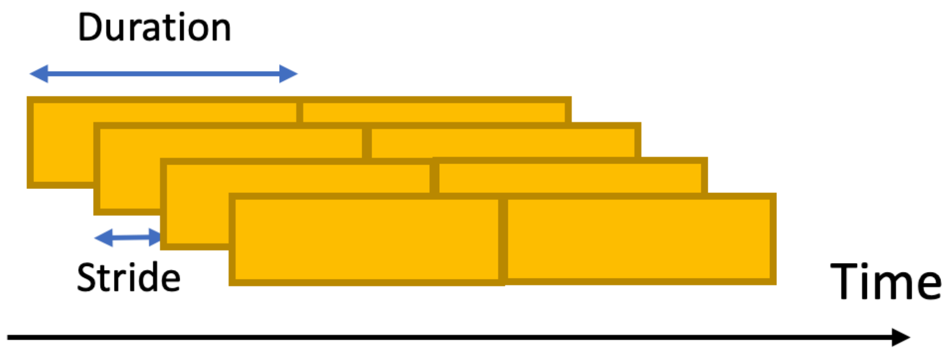


Proposal Sampling

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

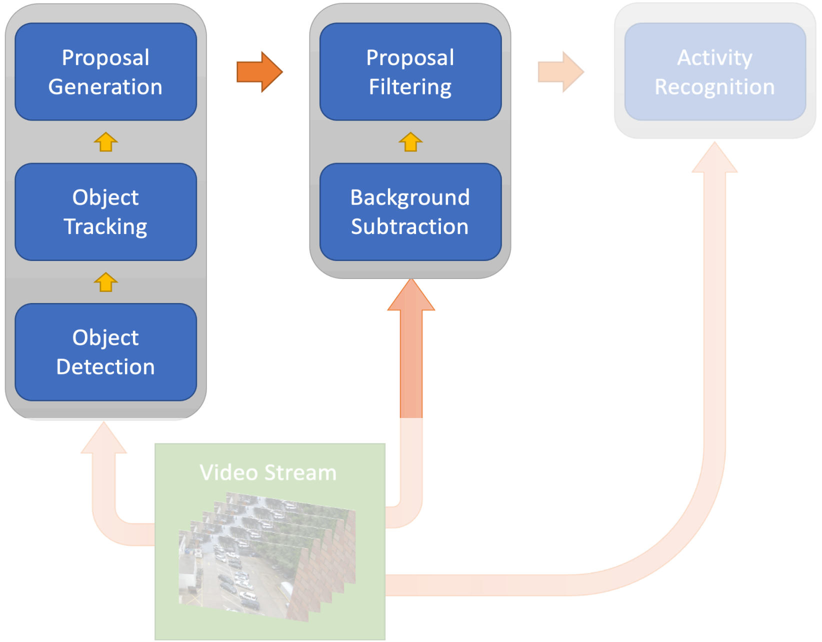


  - Stride = Duration
  - Significant performance drop at boundaries
- New: **dense overlapping** proposal sampling
  - No *boundary*
  - Stride ≤ Duration



Proposal Sampling: An Example

Proposal Refinement

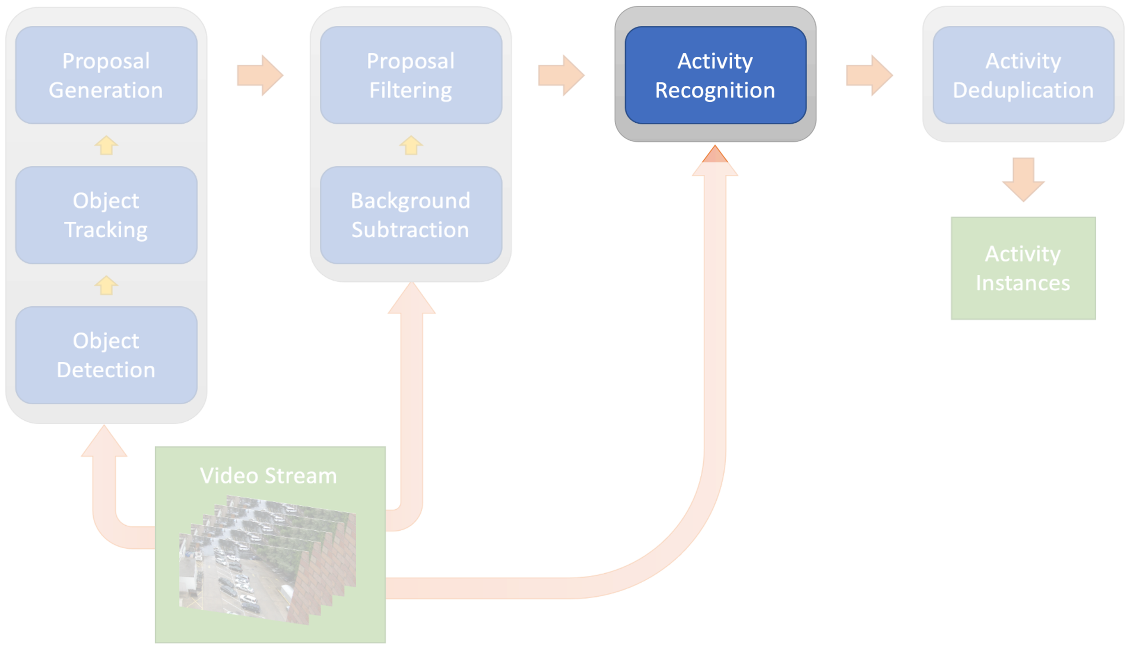


- Spatial localization
  - Extract seed track ids from the central frame
  - Enlarge the bounding boxes as the union of its track

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

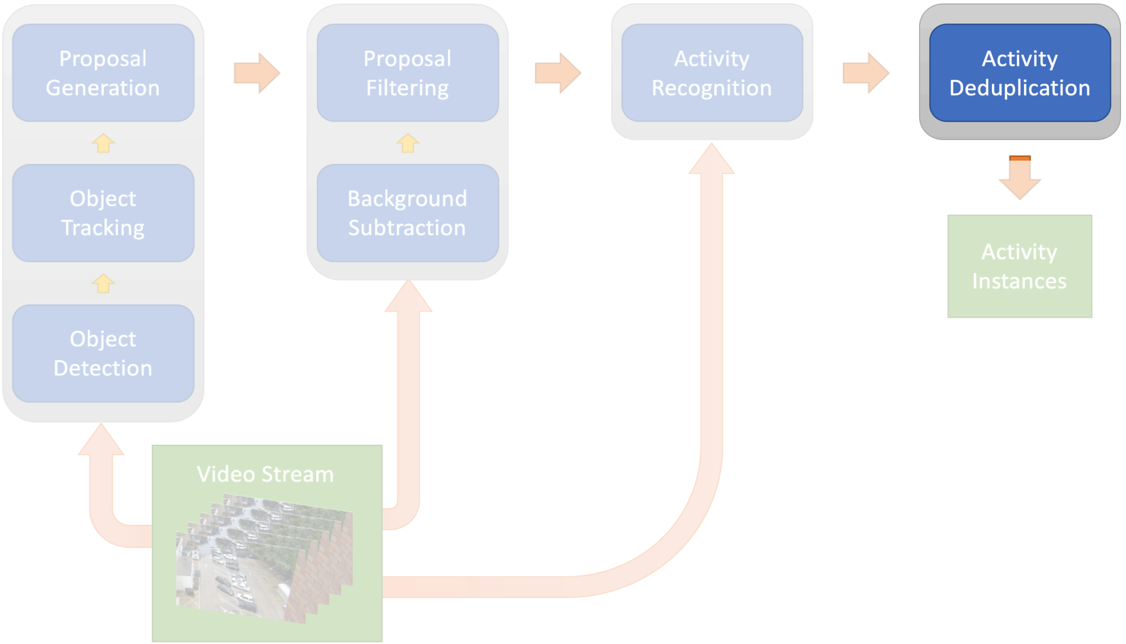
  - Robust through identity switch in the tracking algorithm
  - Ensures coverage of moving objects
- Proposal filtering
  - Leverage motion information, filter out stable objects
  - Binary frame masks from foreground segmentation
  - Proposal foreground score as the average value of pixel masks in its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

Activity Recognition

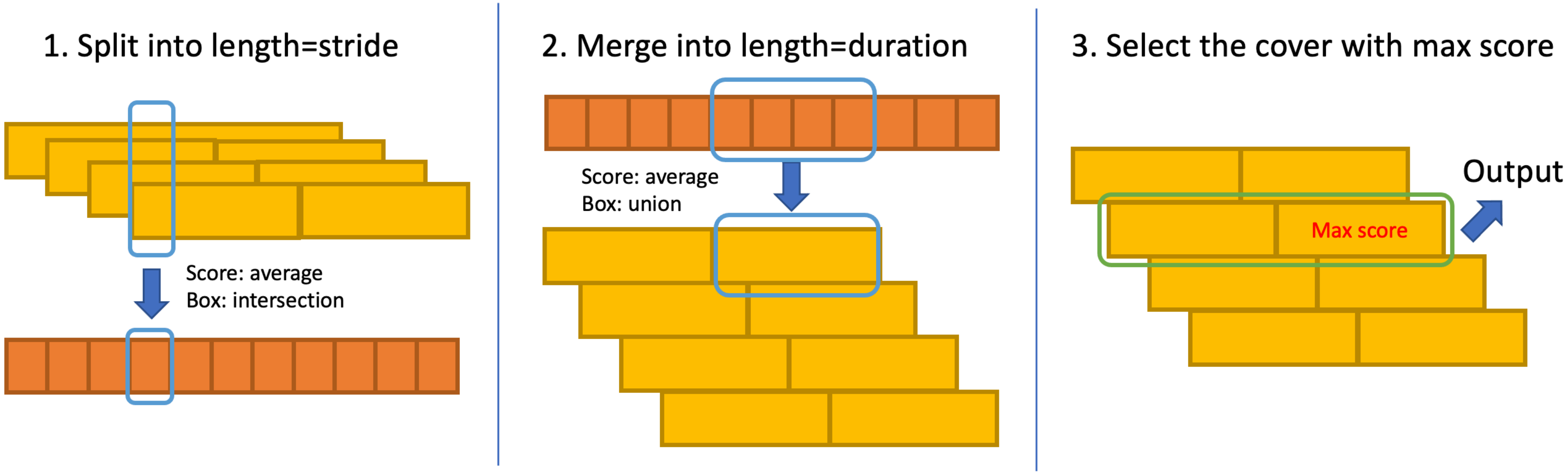


- Multi-label Classification
  - Binary cross entropy loss
  - Weighted by proposal scores
  - Balance activity-wise pos/neg samples
  - Balance samples of different activities
  - Balance samples of different datasets

Activity Deduplication

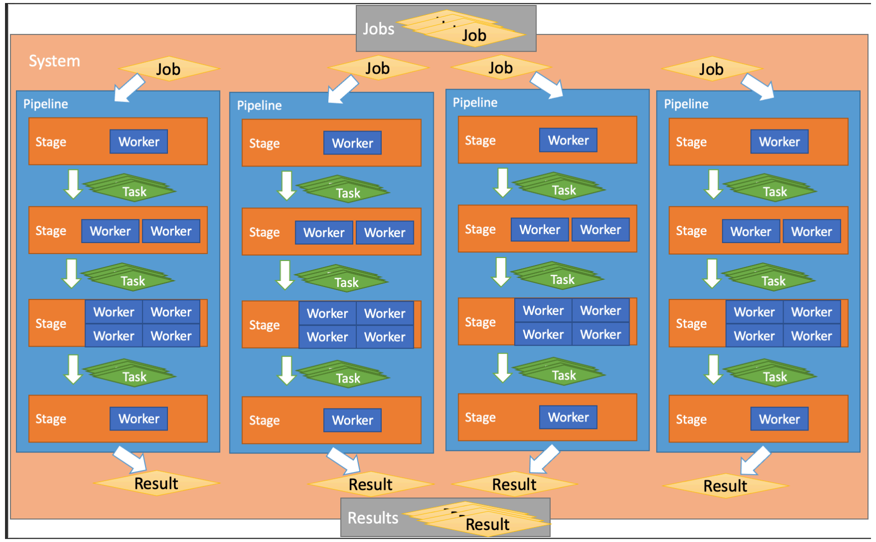


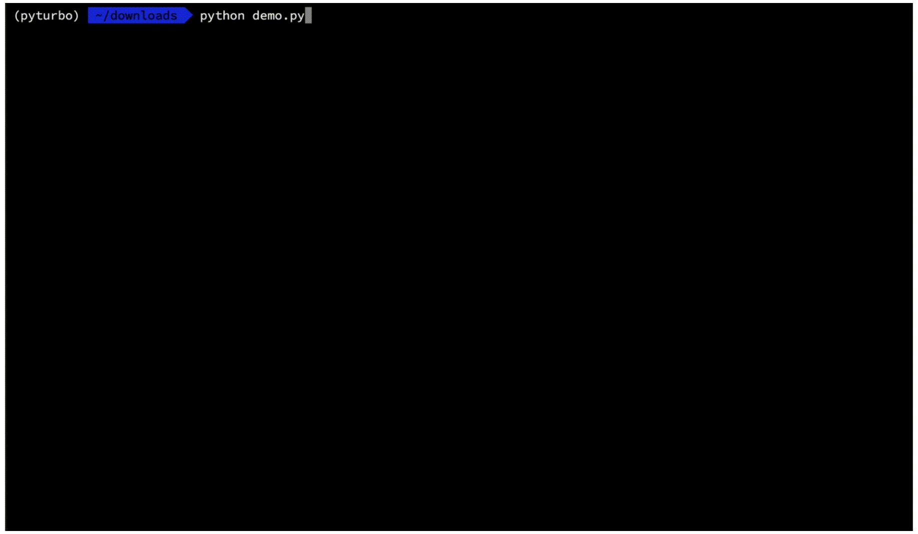
- Remove the duplicate activity instances from overlapping proposals
- Process all proposals in each activity type
- Perform interpolation upon overlapping cubes, maximizing information utilization



Efficiency: Concurrent Execution by 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

Experiments and Results

- Datasets
- Leaderboard Results
- Ablation Studies
- Reproducibility

Training Datasets

- Multiview Extended Video with Activities (MEVA) dataset Known Facility Release #1 (KF1)
  - Total: 257 EO videos annotated, 35 activity classes, 24 camera views
  - Instance Balancing: 158 for training and 99 for validation

- People in Public (PIP) dataset
  - 175k background stabilized clips annotated
  - 66 classes: mapped to the 37 MEVA classes
  - Only used to train activity recognition module

Benchmarks and Metrics

Benchmarks: Activities in Extended Videos (ActEV)

- ActEV’21 Sequestered Data Leaderboard (SDL): Unknown Facilities (UF)
- ActEV’21 SDL: Known Facilities (KF) – MEVA
- TREC Video Retrieval Evaluation (TRECVID) 2020 ActEV – VIRAT

Metrics

- $P_{miss}@0.02T_{fa}$: the recall of activity instances within a time limit of all positive frames plus 2% of negative frames. (TRECVID uses $P_{miss}@0.15T_{fa}$)
- $nAUDC@0.2T_{fa}$: the integration of P_{miss} on $T_{fa} \in [0, 0.2]$

ActEV21’ SDL UF Leaderboard

Rank	Team	Mean $nAUDC@0.2T_{fa}$	Mean $P_{miss}@0.02T_{fa}$	Relative Processing Time
1	CMU	**0.4280**	**0.6378**	0.66
2	IBM-MIT-Purdue	0.5507	0.7881	0.35
3	UCF	0.5625	0.7328	0.70
4	UMD	0.6612	0.7969	0.81

22.3% Better !

Lower is better.
https://actev.nist.gov/sdl#tab_leaderboard as of 01/01/2021.

ActEV21' SDL KF Leaderboard

Rank	Team	Mean $nAUDC@0.2T_{fa}$	Mean $P_{miss}@0.02T_{fa}$	Relative Processing Time
1	CMU	**0.2427**	**0.4620**	0.48
2	UCF	0.3589	0.5233	0.65
3	IBM-MIT-Purdue	0.3609	0.5975	0.13
4	UMD	0.4503	0.6657	0.75

32.4% Better!

Lower is better.
https://actev.nist.gov/sdl#tab_leaderboard as of 01/01/2021.

TRECVID 2020 ActEV Leaderboard

Rank	Team	Mean $nAUDC@0.2T_{fa}$	Mean $P_{miss}@0.15T_{fa}$
1	CMU	**0.4231**	**0.3324**
2	UCF	0.5483	0.5029
3	BUPT-MCPRL	0.5552	0.4878
4	TokyoTech-AIST	0.7975	0.7550

22.8% Better!

Lower is better.
https://actev.nist.gov/trecvid20#tab_leaderboard as of 01/01/2021.

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 MEVA KF1 validation set

(a) Non-overlapping proposals

Duration (# frame)	nAUDC@ 0.2Tfa
32	0.0431
64	0.0183
96	0.0170
128	0.0163
160	0.0186
192	0.0216

(b) **Overlapping** proposals

Duration / Stride (# frame)	16	32
32	0.0114	-
64	**0.0009**	0.0069
96	0.0190	0.0212

Performance of Proposal Filtering

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

Proposal statistics on MEVA KF1 validation set

Name	Unfiltered Proposals	Filtered Proposals
Number of proposals	568410	277511
Positive rate	0.0763	0.1538
Rate of unique label	0.8752	0.8749
Rate of two labels	0.9786	0.9789
Rate of three labels	0.9979	0.9979

Proposal quality metrics on MEVA KF1 validation set

nAUDC@0.2Tfa	IoU			Reference Coverage		
Threshold	Average	≥ 0	≥ 0.5	Average	≥ 0.5	≥ 0.9
Unfiltered Proposals	0.1969	0.0302	0.1133	0.1335	0.0855	0.4301
Filtered Proposals	0.2000	0.0408	0.1169	0.1470	0.0968	0.4468

Improvement from Proposal Filtering

- Proposal filtering **improves** the performance
- Proposal filtering **reduces** processing time (and scoring time !)

SDL UF Leaderboard results for proposal filtering. Lower is better.

Proposal Filter	Mean $nAUDC@0.2T_{fa}$	Mean $P_{miss}@0.02T_{fa}$	Relative Processing Time
Enabled	**0.4822**	**0.7171**	**0.58**
Disabled	0.5176	0.7647	0.93

Improvement from More Training Data

- MEVA: samples are weighted by proposal scores
- MEVA + PIP: samples not weighted

SDL Leaderboard results for different training data. Lower is better.

SDL UF	Mean $nAUDC@0.2T_{fa}$	Mean $P_{miss}@0.02T_{fa}$	Relative Processing Time
MEVA + PIP	**0.4280**	**0.6378**	0.66
MEVA	0.4657	0.6767	0.66

SDL KF	Mean $nAUDC@0.2T_{fa}$	Mean $P_{miss}@0.02T_{fa}$	Relative Processing Time
MEVA + PIP	0.2440	0.4594	0.48
MEVA	0.2427	0.4620	0.48

Training Speed and Reproducibility

- Training Set: **Only** MEVA KF1
- Three Stages:
  - Proposal generation
  - Label assignment and proposal filter learning
  - Classifier training
- Total Time:
  - Less than 48 hours on one standard SDL Machine (4x 2080Ti)
- State-of-the-art performance (without extra data)

Reference SDL Leaderboard results

	Mean $nAUDC@0.2T_{fa}$	Mean $P_{miss}@0.02T_{fa}$	Relative Processing Time
SDL KF	0.2427	0.4620	0.48
SDL UF	0.4657	0.6768	0.65

Take Away & Future Work

Lessons:

- Spatio-temporal cube proposal vs. tube proposal
- Dense overlapping proposal sampling vs. nonoverlapping sampling
- Balanced sampling strategy
- Weighted loss for classifier training
- More training data for action recognition

Prospects:

- Evaluation of spatial localization
- Evaluation of training time consumption

informedia
digital video understanding
SEARCH summarize visualize retrieve

Real-time Activity Detection in Unknown Facilities with Dense Spatio-temporal Proposals

Thanks for listening !

Carnegie Mellon University
Language Technologies Institute