以人为中心的复杂视频分析挑战

林巍峣

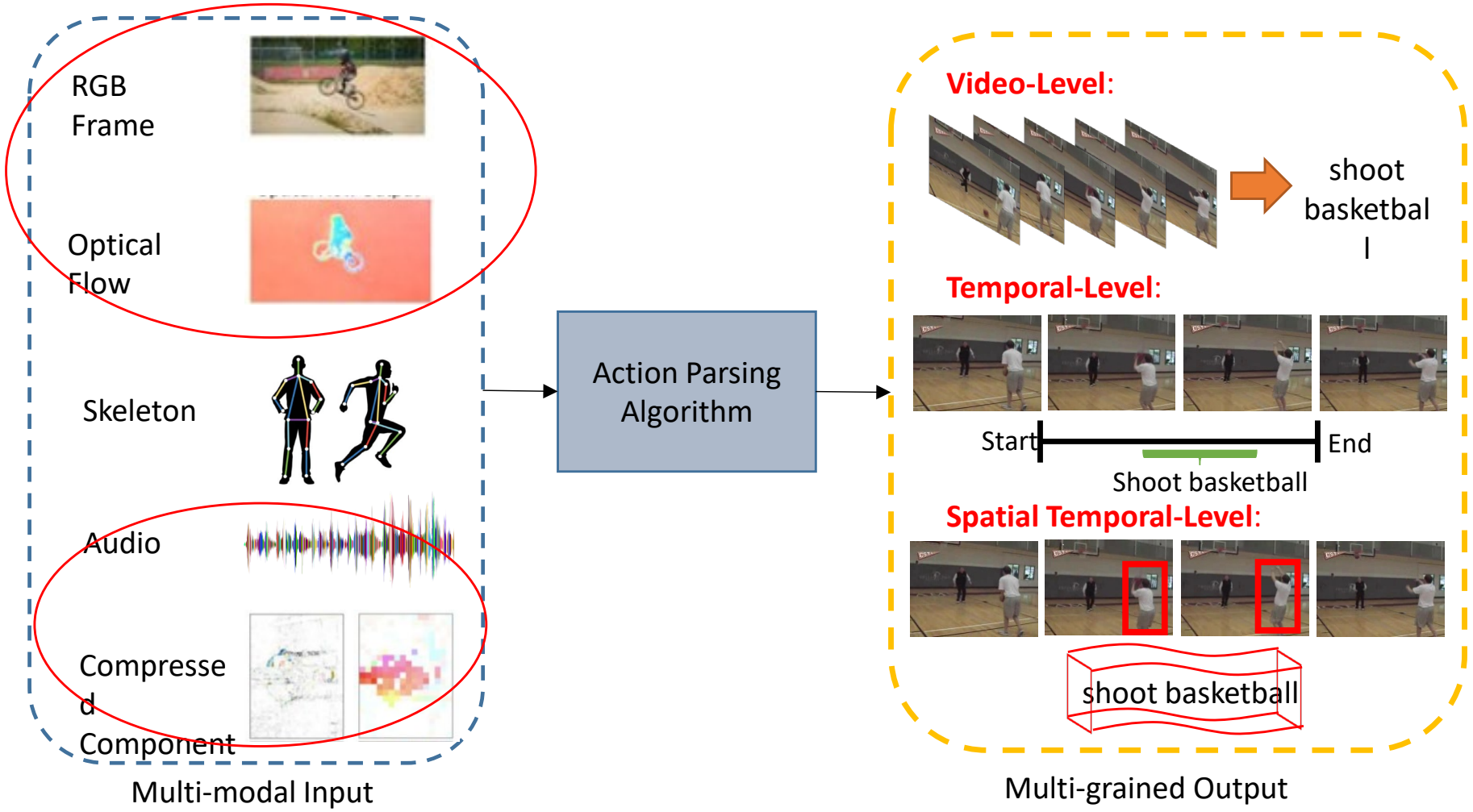
上海交通大学

2020. 11. 8

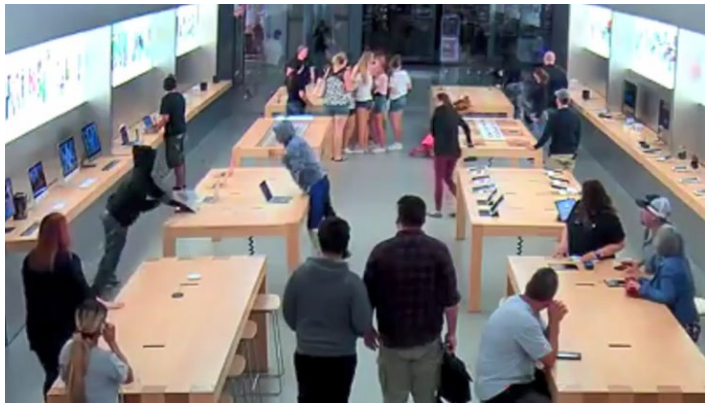
Outline

- Background
- Dataset
- Challenge
- Awards

Action Parsing with Different Settings



Typical Internet Videos



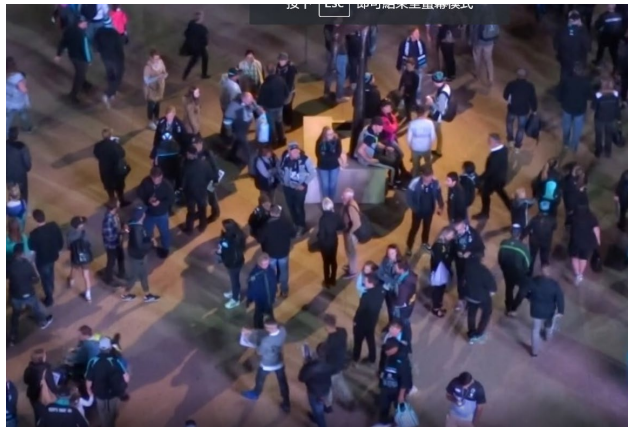
Typical Surveillance Videos

Background

- The development of modern intelligent city highly relies on the advancement of human-centric multimedia analysis technologies and datasets.

- Limitations of the existing human-centric multimedia analysis datasets
  - focus on normal or relatively simple scenes (easy-predictable motions/poses)
  - the coverage and scale are limited (with coarse labels or with simple scenes)

COCO



MOT 20

Pose Track

Kinetics

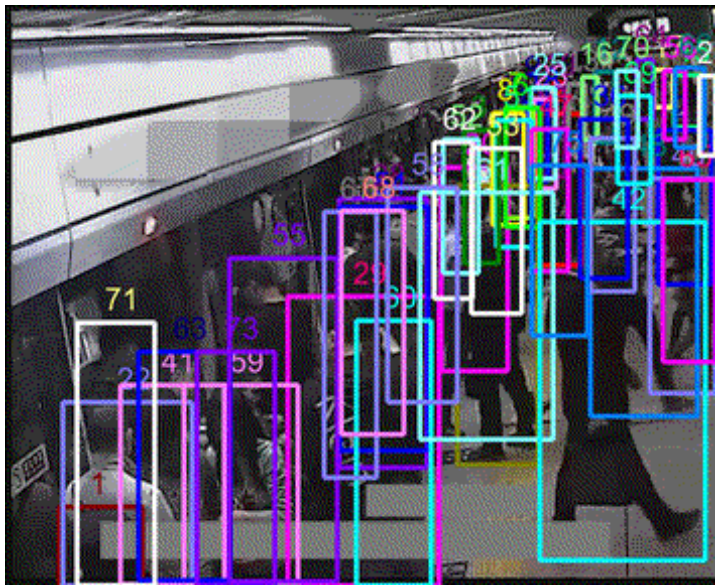
AVA

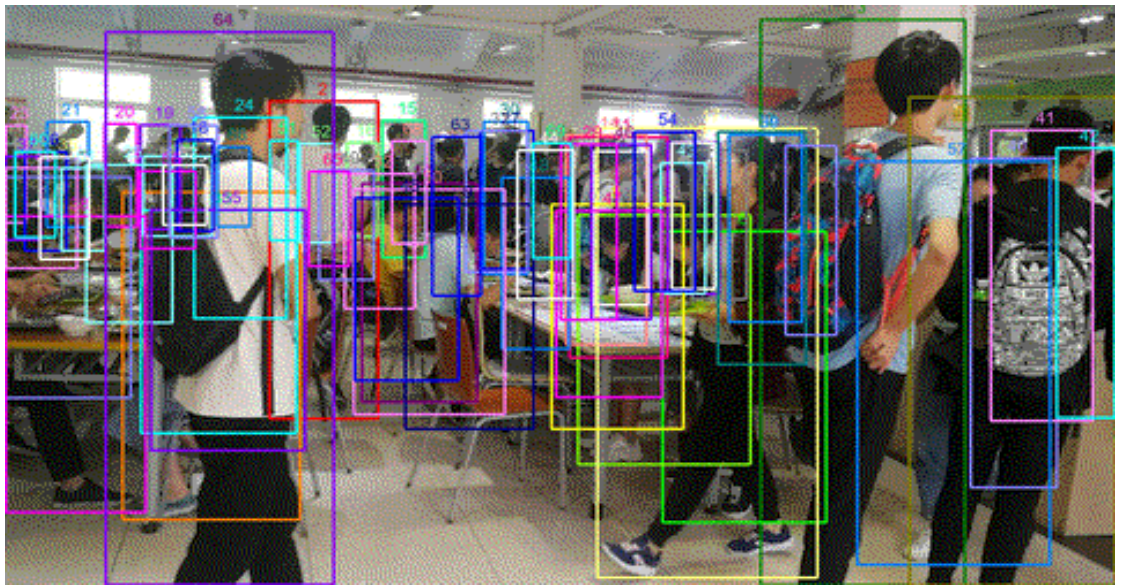
UCF-Crime

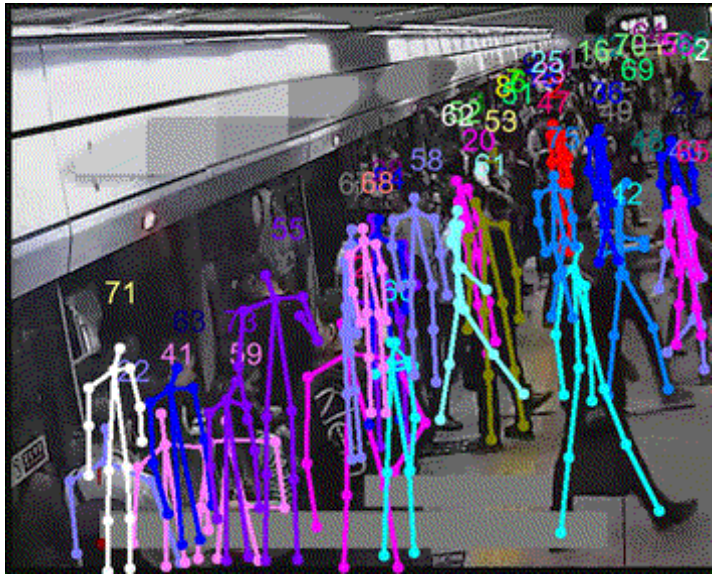
Our dataset

- Focus on very challenging and realistic tasks of human-centric analysis in various crowd & complex events.
  - Large-scale Human-centric Video Analysis in Complex Events (HiEve)
- Features:
  - Covers a wide range of human-centric understanding tasks including motion, pose, and action
  - Has substantially larger data scales, which includes the largest # of poses (>1M), the largest # of complex-event action labels (>56k), and one of the largest # of trajectories with long terms (with average trajectory length >480).
  - Focuses on the challenging scenes under various crowd & complex events (such as dining, earthquake escape, subway getting-off, and collision)

Example Videos in our HiEve challenge & dataset





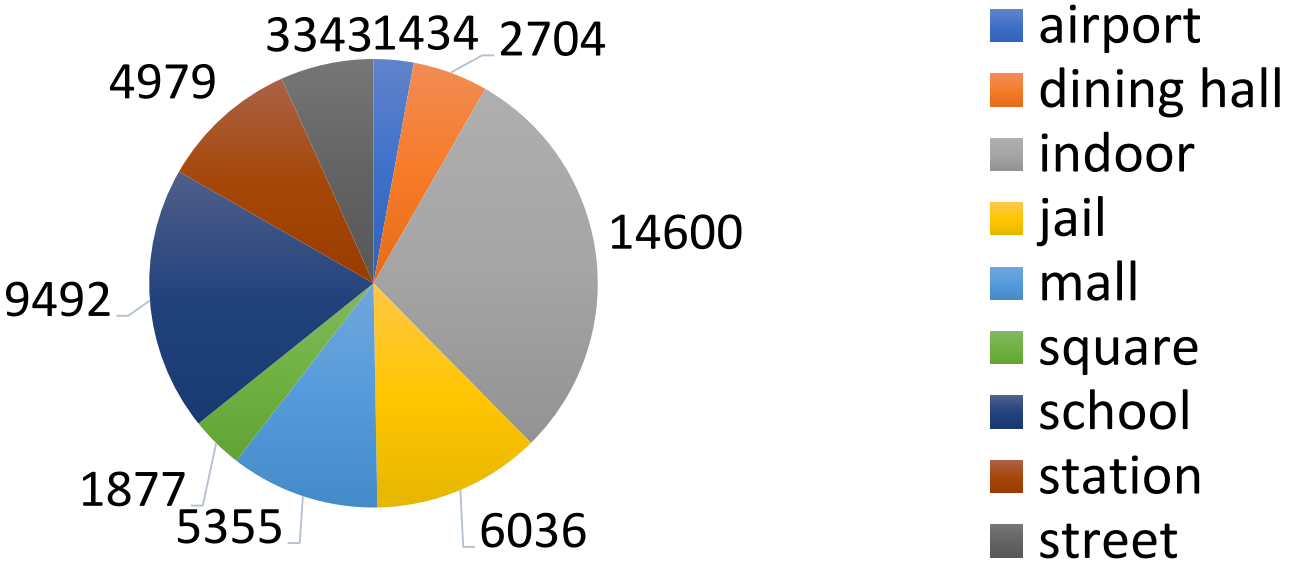


Outline

- Background
- Dataset
- Challenge
- Awards

Data Collection

- Select several crowded places with complex and diverse events
- Collected from 9 different scenes

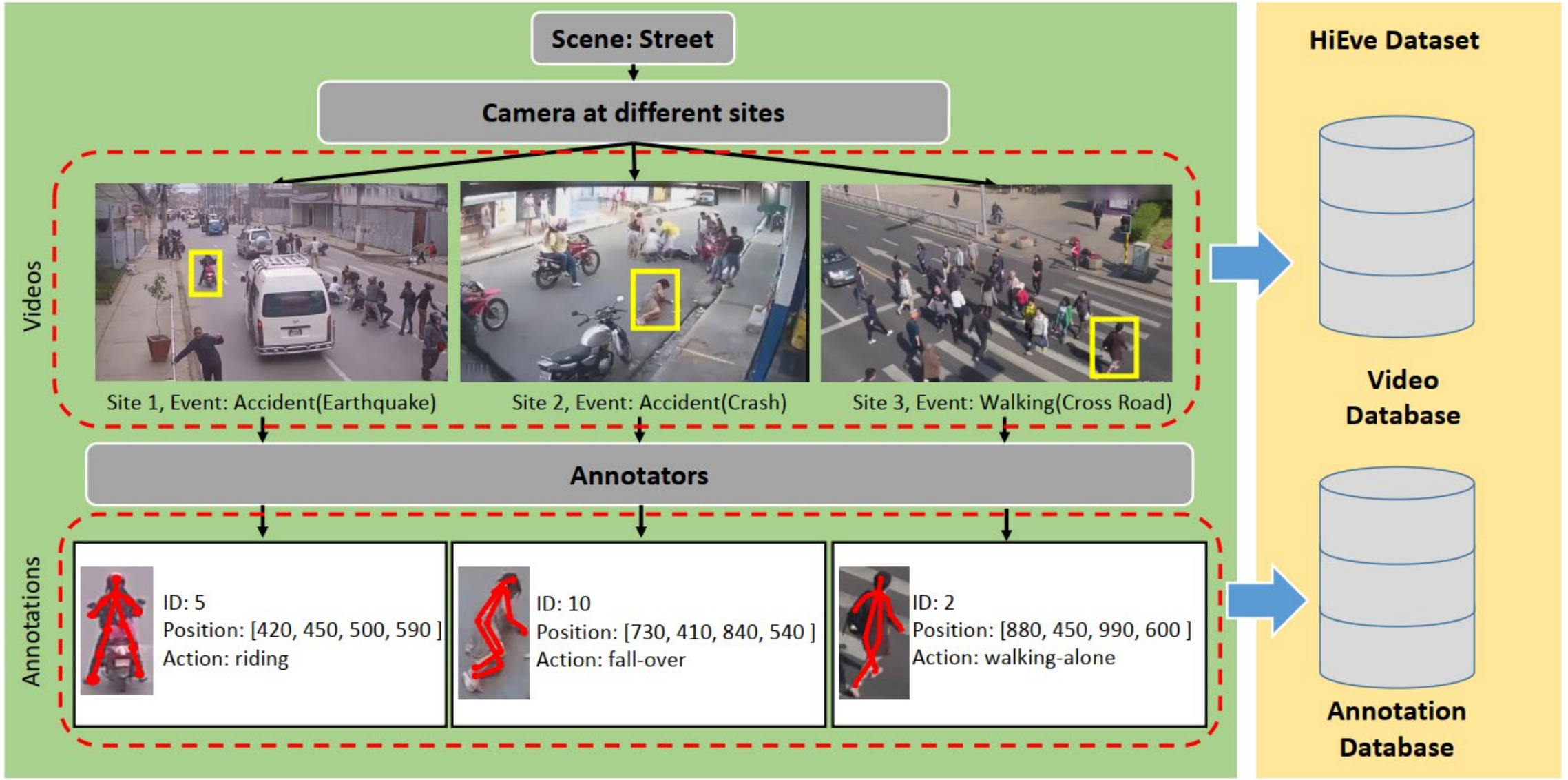


Distribution of different scenes in HiEve dataset

[1] http://humanineventsorg/
[2] Human in Events: A Large-Scale Benchmark for Human-centric Video Analysis in Complex Events, https://arxiv.org/abs/2005.04490

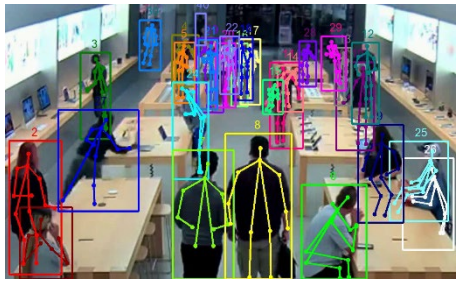
Example video snapshots

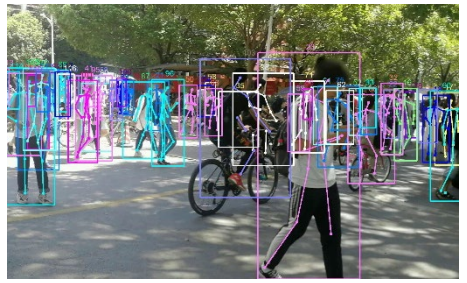
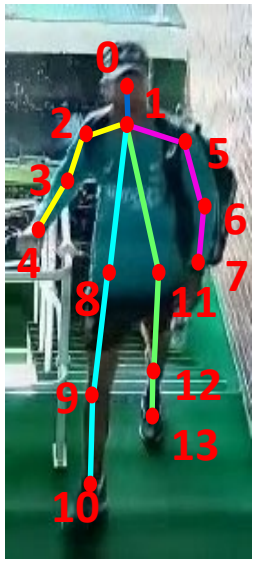
Data Construction Pipeline



Data Annotation

- Bounding Box
- ID
- Key-points
- Action



Tasks

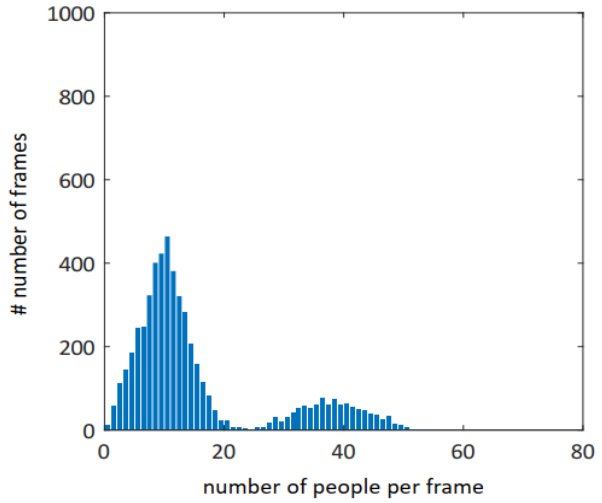
- Object Tracking
- Pose Estimation
- Pose Tracking
- Action Recognition

Comparison—Scale

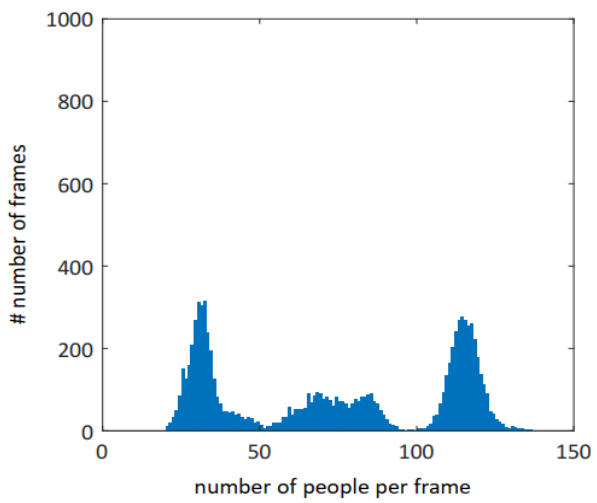
- HiEve is the largest, human-centric dataset
  - Largest number of poses (>1M), actions (>56k), and trajectories with long terms (L_{avg}> 480)
  - Various human-centric visual tasks (MOT, Pose, Action)
  - Crowd & complex events

Dataset	# pose	# box	# traj.(avg)	# action	pose track	surveillance	complex events
MSCOCO [1]	105,698	105,698	NA	NA	×	×	×
MPII [4]	14,993	14,993	NA	410	×	×	×
CrowdPose [5]	~80,000	~80,000	NA	NA	×	×	×
PoseTrack [2]	~267,000	~26,000	5,245(49)	NA	√	×	×
MOT16[6]	NA	292,733	1,276(229)	NA	×	√	×
MOT17	NA	901,119	3,993(226)	NA	×	√	×
MOT20 [7]	NA	1,652,040	3457(478)	NA	×	√	×
Avenue [8]	NA	NA	NA	15	×	√	×
UCF-Crime [3]	NA	NA	NA	1,900	×	√	√
Ours	**1,099,357**	1,302,481	2,687(485)	**56,643**	√	√	√

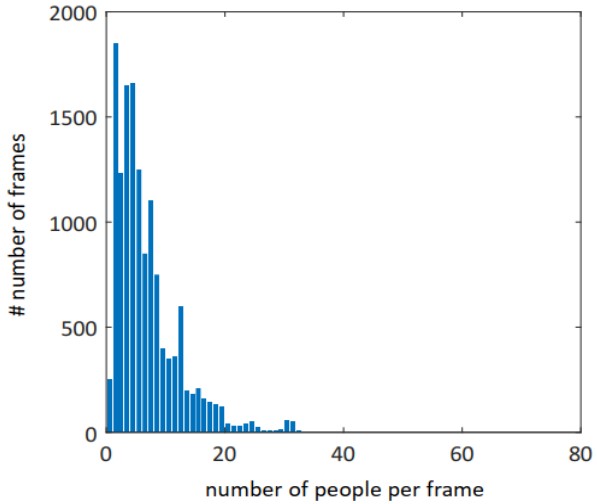
Comparison—Person number



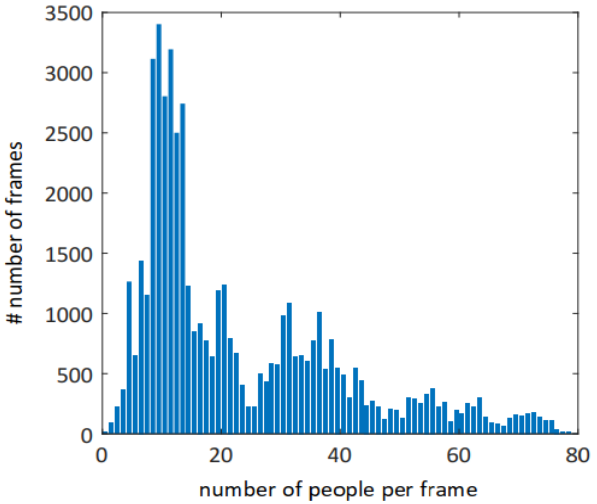
(a) MOT17



(b) MOT20

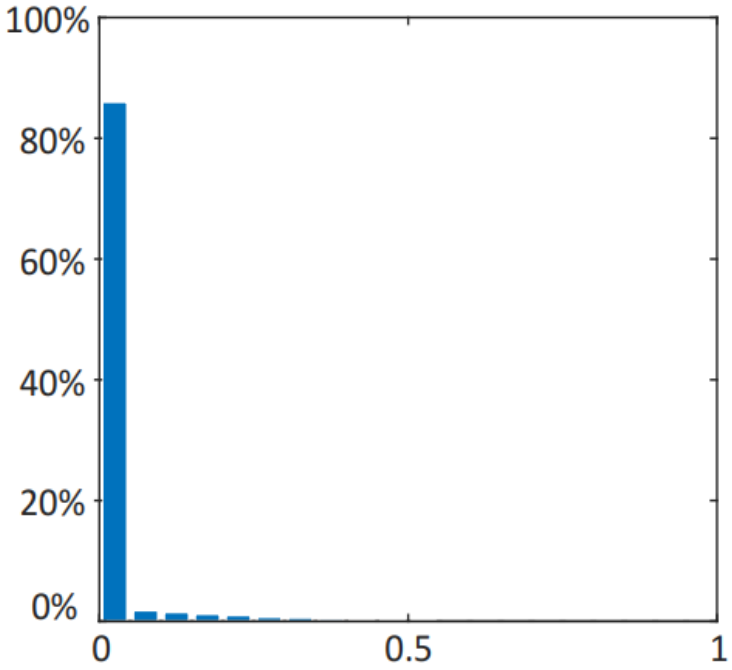


(c) PoseTrack

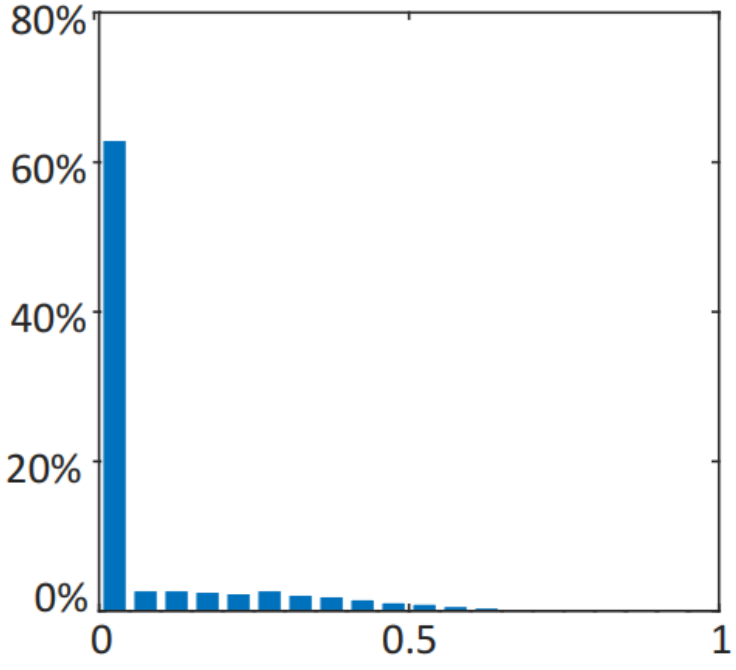


(d) HiEve

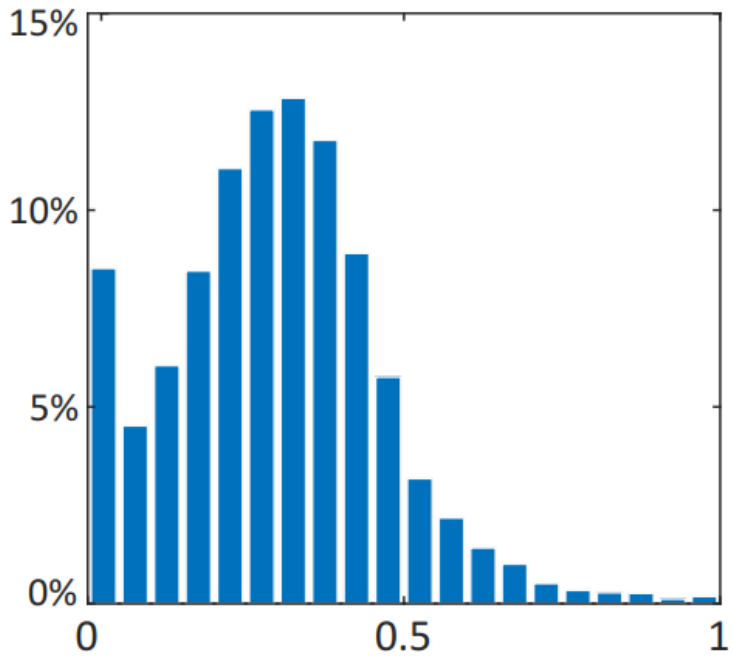
Comparison—Crowd Index



(a) MPII



(b) MSCOCO

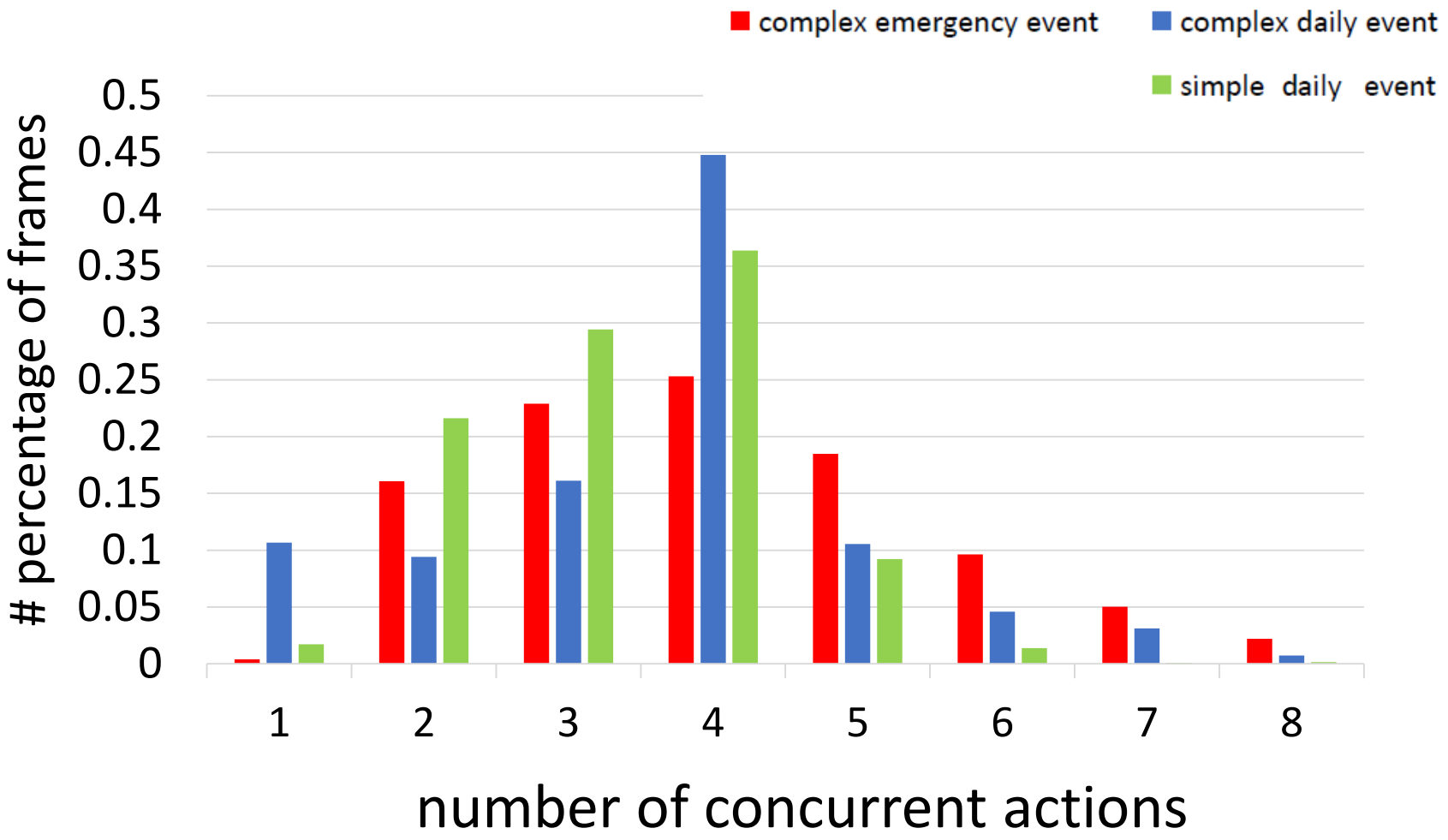


(c) HiEve

Crowd Index distribution

[1] http://humanineventsorg/
[2] Human in Events: A Large-Scale Benchmark for Human-centric Video Analysis in Complex Events, https://arxiv.org/abs/2005.04490

Comparison—Action



[1] http://humanineventsorg/

[2] Human in Events: A Large-Scale Benchmark for Human-centric Video Analysis in Complex Events, https://arxiv.org/abs/2005.04490

Outline

- Background
- Dataset
- Challenge
- Awards

Challenge

ACM MM Grand Challenge on
Large-scale Human-centric Video Analysis in Complex Events

- Goal: Bring researches in the community together to advance human-centric analyzing methods from 3 aspects:
  - Contributed papers
  - Challenges
  - Foster new ideas & directions
- **Organizers**

Weiyao Lin
Shanghai Jiao Tong University, China

Guojun Qi
Futurewei Technologies, MAPLE Lab, USA

Nicu Sebe
University of Trento, Italy

Ning Xu
Adobe Research, USA

Hongkai Xiong
Shanghai Jiao Tong University, China

Mubarak Shah
University of Central Florida, USA

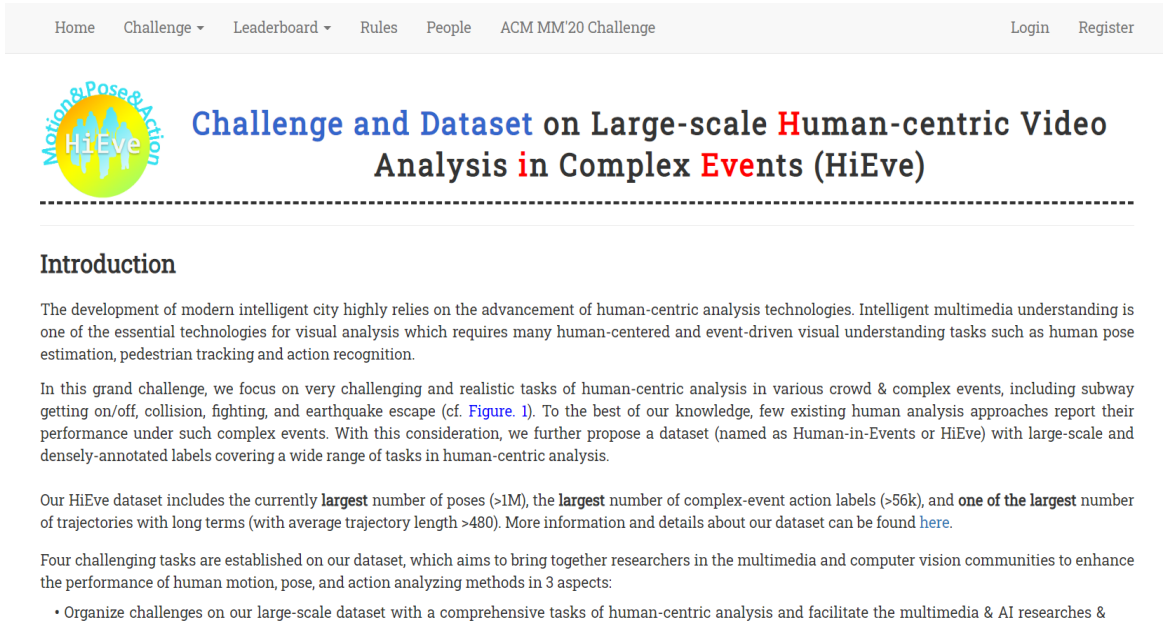
Challenge Task & Evaluation metrics

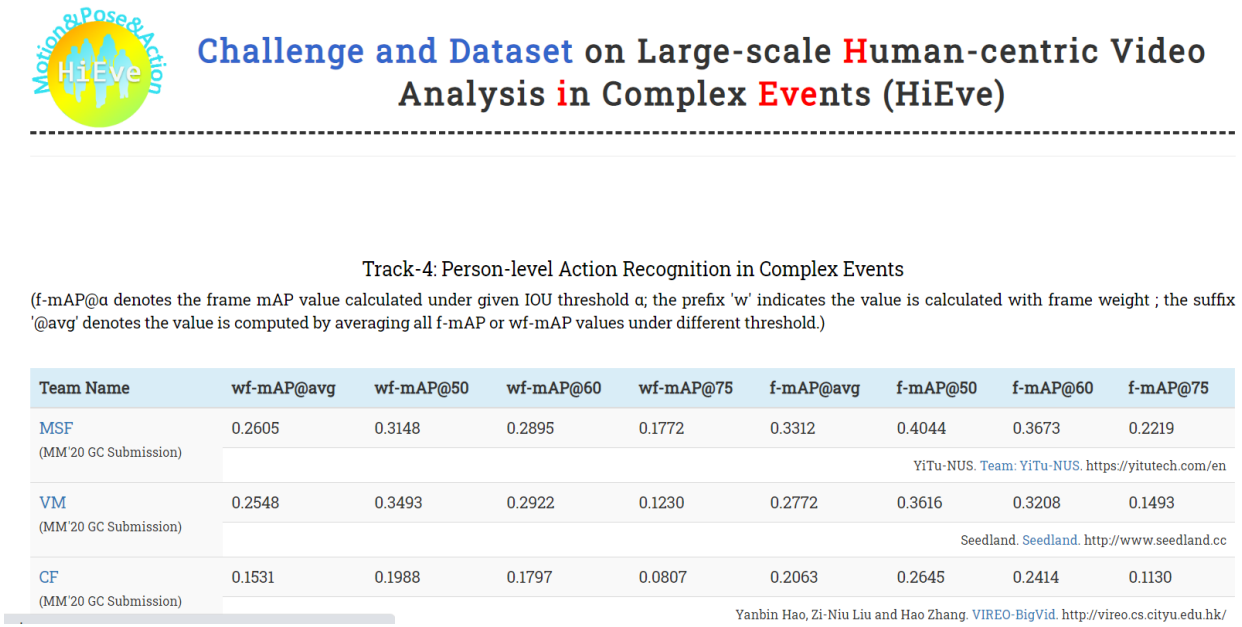
- Track-1 – MOT (subtrack: public)
  - Metric: MOTA, MOTP, IDF,IDS, w-MOTA
- Track-1 – MOT (subtrack: private)
  - Metric: MOTA, MOTP, IDF,IDS, w-MOTA
- Track-2 – Crowd Pose estimation
  - Metric: AP, AP@average, w-AP
- Track-3 – Crowd Pose tracking
  - Metric: MOTA, MOTP, AP
- Track-4 – Person-level Action recognition
  - Metric: frame-mAP, w-fmAP

[1] http://humanineventsorg/
[2] Human in Events: A Large-Scale Benchmark for Human-centric Video Analysis in Complex Events, https://arxiv.org/abs/2005.04490

Challenge Website

- Allow participants to register, login, and submit their results
- Automatically evaluate the results and manage the leaderboards
- Comprehensively manage user and result

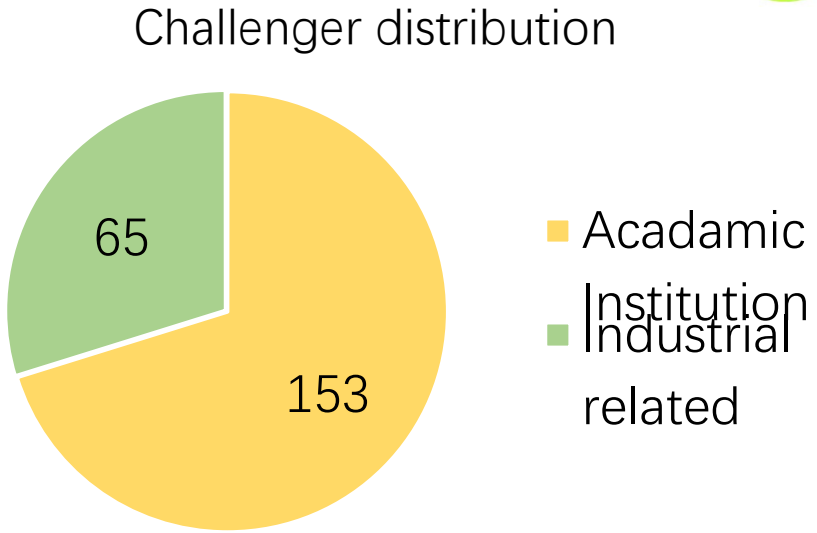


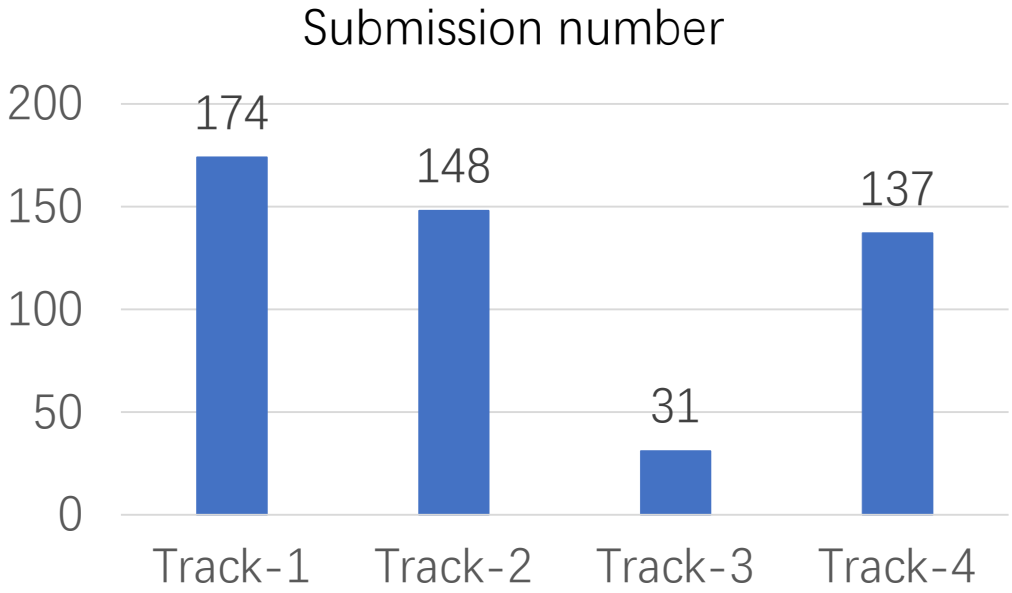


[1] http://humaninevents.org/

Challenge Statistics

- Timeline
  - Training data release: May 4, 2020
  - Testing data release: May 11, 2020
  - Result submission: June 22, 2020
  - Result notification: June 30, 2020
- Statistics
  - 218 register team members
  - Good coverage
    - European, Singapore, Austria, China, US
    - 65 industrial related challengers
  - 126 teams submitted their results

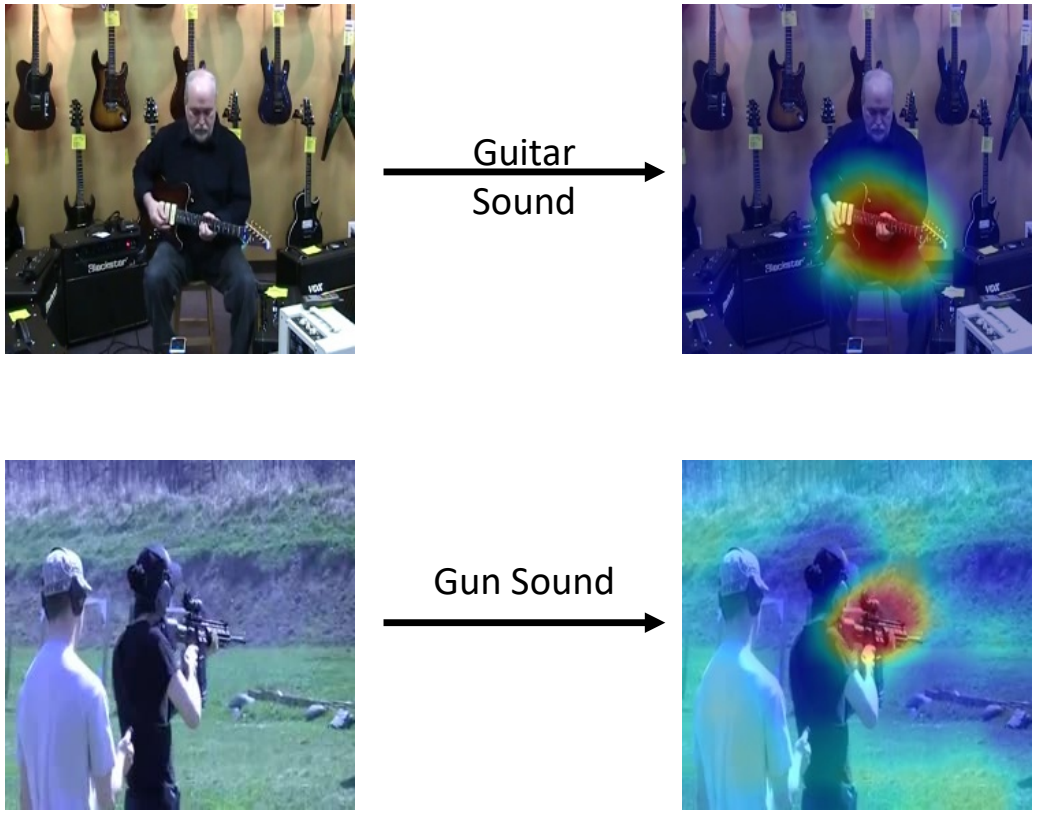




[1] http://humanineventsorg/ACM_welcome.html

Future Work

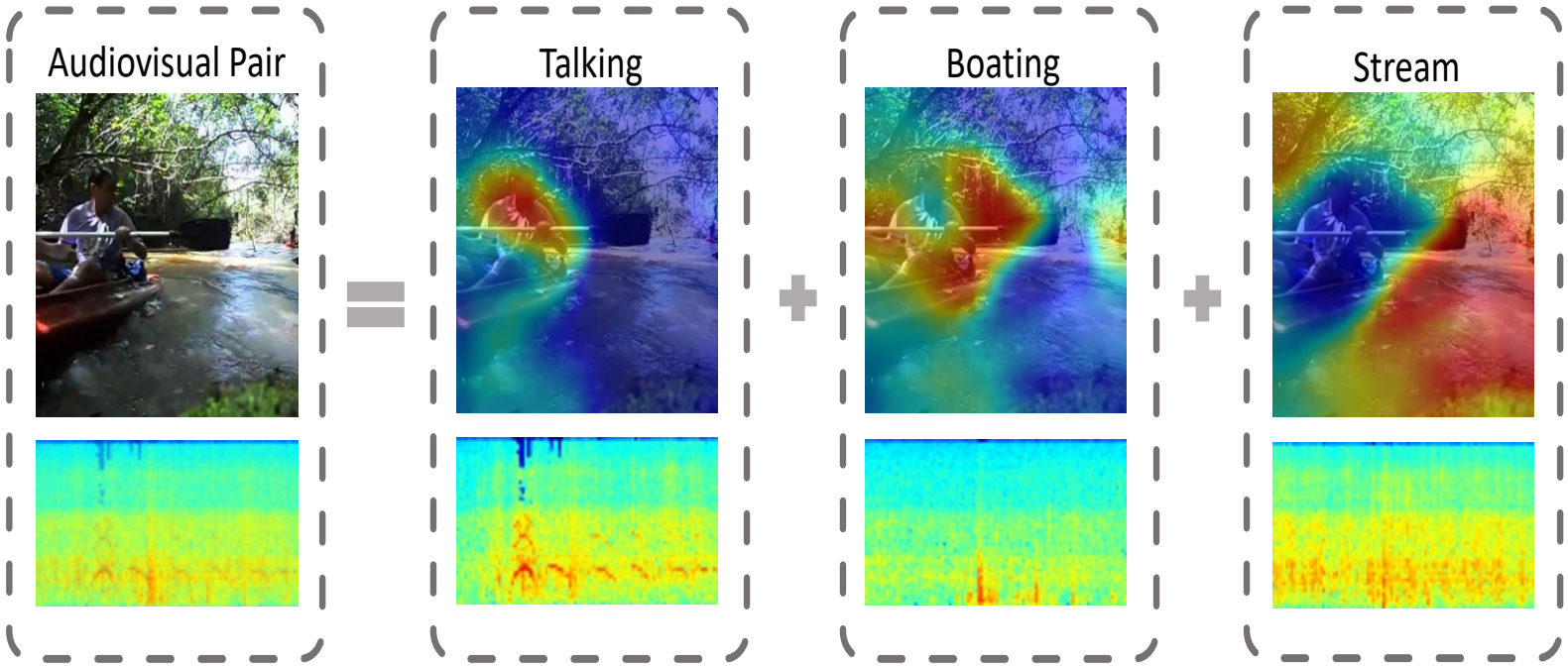
- Multi-modality, especially multi-source audio-visual

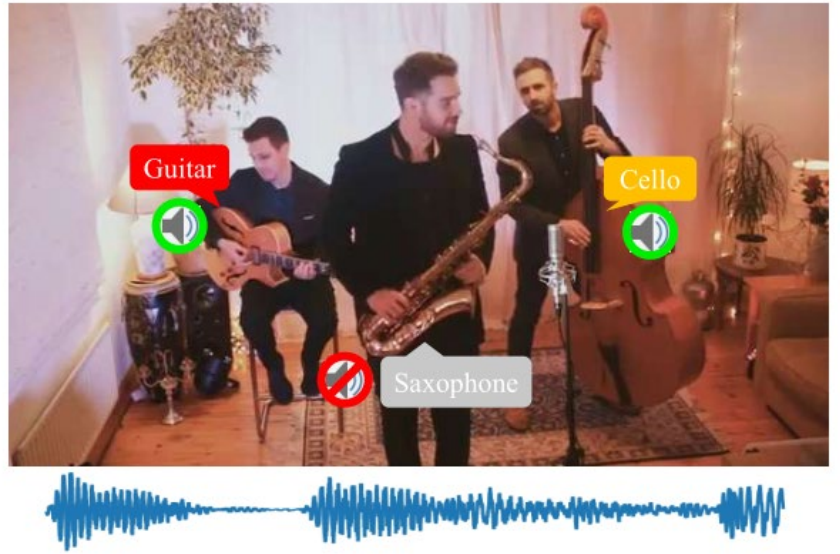


- Multiple audiovisual components
- Associate sound-object pairs without one-to-one annotations

[R. Qian et al, Multiple Sound Sources Localization from Coarse to Fine, ECCV 2020]

Multi-source sounding object localization





Thank You

https://weiyaolin.github.io/

http://humanineevents.org/