以人为中心的复杂视频分析挑战



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)







Pose Track



MOT 20



APPLAUDING



APPLYING CREAM



Left: Stand, Carry/Hold, Listen to; Middle: Stand, Carry/Hold, Talk to; Right: Sit, Write

Kinetics





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://humaninevents.org/

[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	<pre># traj.(avg)</pre>	# action	pose track	surveillance	complex events
MSCOCO [1]	105,698	105,698	NA	NA	×	×	X
MPII ^[4]	14,993	14,993	NA	410	×	×	×
CrowdPose [5]	$\sim \! 80,\! 000$	$\sim \! 80,\! 000$	NA	NA	×	×	×
PoseTrack [2]	$\sim 267,000$	$\sim 26,000$	5,245(49)	NA	\checkmark	×	×
MOT16[6]	NA	292,733	1,276(229)	NA	×	\checkmark	×
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	×		\checkmark
Ours	1,099,357	1,302,481	2,687(485)	56,643	\checkmark		$\overline{\checkmark}$

Comparison—Person number







Comparison—Crowd Index



Crowd Index distribution

[1] http://humaninevents.org/

[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://humaninevents.org/

[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://humaninevents.org/

[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



Introduction

The development of modern intelligent city highly relies on the advancement of human-centric analysis technologies. Intelligent multimedia understanding is one of the essential technologies for visual analysis which requires many human-centered and event-driven visual understanding tasks such as human pose estimation, pedestrian tracking and action recognition.

In this grand challenge, we focus on very challenging and realistic tasks of human-centric analysis in various crowd & complex events, including subway getting on/off, collision, fighting, and earthquake escape (cf. Figure 1). To the best of our knowledge, few existing human analysis approaches report their performance under such complex events. With this consideration, we further propose a dataset (named as Human-in-Events or HiEve) with large-scale and densely-annotated labels covering a wide range of tasks in human-centric analysis.

Our HiEve dataset includes the currently largest number of poses (>1M), the largest number of complex-event action labels (>56k), and one of the largest number of trajectories with long terms (with average trajectory length >480). More information and details about our dataset can be found here.

Four challenging tasks are established on our dataset, which aims to bring together researchers in the multimedia and computer vision communities to enhance the performance of human motion, pose, and action analyzing methods in 3 aspects:

• Organize challenges on our large-scale dataset with a comprehensive tasks of human-centric analysis and facilitate the multimedia & AI researches &



Challenge and Dataset on Large-scale Human-centric Video Analysis in Complex Events (HiEve)

Track-4: Person-level Action Recognition in Complex Events

(f-mAP@a denotes the frame mAP value calculated under given IOU threshold a; the prefix 'w' indicates the value is calculated with frame weight ; the suffix '@avg' denotes the value is computed by averaging all f-mAP or wf-mAP values under different threshold.)

Team Name	wf-mAP@avg	wf-mAP@50	wf-mAP@60	wf-mAP@75	f-mAP@avg	f-mAP@50	f-mAP@60	f-mAP@75
MSF (MM'20 GC Submission)	0.2605	0.3148	0.2895	0.1772	0.3312	0.4044	0.3673	0.2219
	YiTu-NUS. Team: YiTu-NUS. https://yitutech.com/en							
VM (MM'20 GC Submission)	0.2548	0.3493	0.2922	0.1230	0.2772	0.3616	0.3208	0.1493
		Seedland. Seedland. http://www.seedland.cc						
CF (MM'20 GC Submission)	0.1531	0.1988	0.1797	0.0807	0.2063	0.2645	0.2414	0.1130
				Уа	Yanbin Hao, Zi-Niu Liu and Hao Zhang, VIREO-BigVid, http://vireo.cs.cityu.edu.hk/			

[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







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://humaninevents.org/