International Journal of Computer Vision (2023) 131:2994–3018
https://doi.org/10.1007/s11263-023-01842-6

HiEve: A Large-Scale Benchmark for Human-Centric Video Analysis in Complex Events

Weiyao Lin[1] · Huabin Liu[1] · Shizhan Liu[1] · Yuxi Li[1] · Hongkai Xiong[1] · Guojun Qi[2] · Nicu Sebe[3]

Received: 19 August 2022 / Accepted: 15 June 2023 / Published online: 10 July 2023

Abstract

Along with the development of modern smart cities, human-centric video analysis has been encountering the challenge of analyzing diverse and complex events in real scenes. A complex event relates to dense crowds, anomalous individuals, or collective behaviors. However, limited by the scale and coverage of existing video datasets, few human analysis approaches have reported their performances on such complex events. To this end, we present a new large-scale dataset with comprehensive annotations, named human-in-events or human-centric video analysis in complex events (HiEve), for the understanding of human motions, poses, and actions in a variety of realistic events, especially in crowd and complex events. It contains a record number of poses (> 1 M), the largest number of action instances (>56k) under complex events, as well as one of the largest numbers of trajectories lasting for longer time (with an average trajectory length of >480 frames). Based on its diverse annotation, we present two simple baselines for action recognition and pose estimation, respectively. They leverage cross-label information during training to enhance the feature learning in corresponding visual tasks. Experiments show that they could boost the performance of existing action recognition and pose estimation pipelines. More importantly, they prove the widely ranged annotations in HiEve can improve various video tasks. Furthermore, we conduct extensive experiments to benchmark recent video analysis approaches together with our baseline methods, demonstrating HiEve is a challenging dataset for human-centric video analysis. We expect that the dataset will advance the development of cutting-edge techniques in human-centric analysis and the understanding of complex events. The dataset is available at http://humaninevents.org.

Keywords Complex events · Human-centric video analysis · Dataset and benchmark

1 Introduction

The development of smart cities highly relies on the advancement of fast and accurate visual understanding of multimedia (Xu et al., 2018; Mei et al., 2013; Chen et al., 2022). To achieve this goal, many human-centered and event-driven visual understanding problems have been raised, such as human pose estimation (Fang et al., 2017), pedestrian tracking (Dendorfer et al., 2020; Ren et al., 2018), and action recognition (Veeriah et al., 2015; Shu et al., 2019).

Recently, several public datasets (e.g., MSCOCO Lin et al. 2014; PoseTrack Andriluka et al. 2018; UCF-Crime Sultani et al. 2018) have been proposed to benchmark the aforementioned tasks. However, they have some limitations when applied to real scenarios with complex events such as dining, earthquake escape, subway getting-off and collisions.

Communicated by Dima Damen.

✉ Weiyao Lin
wylin@sjtu.edu.cn

Huabin Liu
huabinliu@sjtu.edu.cn

Shizhan Liu
shanluzuode@sjtu.edu.cn

Yuxi Li
lyxok1@sjtu.edu.cn

Hongkai Xiong
xionghongkai@sjtu.edu.cn

Guojun Qi
guojunq@gmail.com

Nicu Sebe
niculae.sebe@unitn.it

[1] Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China

[2] Machine Perception and Learning Lab, Orlando, USA

[3] University of Trento, Trento, Italy