**PoseTrack: A Benchmark for Human Pose Estimation and Tracking**

**The 7 W's**

Q1. Existing systems for video-based pose estimation and tracking struggle to perform well on realistic videos with multiple people and often fail to output body-pose trajectories consistent over time. To address this shortcoming this paper introduces PoseTrack which is a new large-scale benchmark for video-based human pose estimation and articulated tracking.

Q2. It is relevant because it proposes a new benchmark for human pose estimation and articulated tracking that is significantly larger and more diverse in terms of data variability and complexity compared to existing pose tracking bench- marks.

Q3. The most popular benchmarks to date for evaluation of single person pose estimation are “LSP” [25] (+ “LSP Ex- tended” [26]), “MPII Human Pose (Single Person)” [1] and MS COCO Keypoints Challenge [28]. LSP and LSP Ex- tended datasets focus on sports scenes featuring a few sport types. Although a combination of both datasets results in 11,000 training poses, the evaluation set of 1000 is rather small. More recently, MS COCO Keypoints Challenge [28] has been introduced to provide a new large- scale benchmark for single frame based multi-person pose estimation.

Q4. All these datasets are only limited to single- frame based body pose estimation. In contrast, our dataset also focuses on a more challenging task of multi-person pose estimation in video sequences containing highly articulated people in dense crowds. Although there exist training sets for special scenarios, such as sports [51, 23] and upright frontal people [6], these benchmarks focus on single isolated individualsand are still limited in their scope and variability of represented activities and body motions.

Q5. Our new benchmark encompasses three tasks focusing on i) single-frame multi-person pose estimation, ii) multi-person pose estimation in videos, and iii) multi-person articulated tracking. To establish the bench- mark, we collect, annotate and release a new dataset that features videos with multiple people labeled with person tracks and articulated pose.

Q6. In this work, we aim to fill this gap by establishing a new large-scale, high-quality benchmark for video-based multi-person pose estimation and articulated tracking.

Q7. Our analysis shows that current methods perform well on easy sequences with well separated upright people, but are severely challenged in the presence of fast camera motions and complex articulations. Addressing these challenges remains an important direction for the future work.