EventPoseFormer: Event-Based Real-time 3D Human Pose Estimation

Melica Omer Ali[1,2], Francesca Odone[1], Arren Glover[2], Chiara Bartolozzi[2] and Gaurvi Goyal[3]

[1] University of Genova, [2] Italian Institute of Technology, [3] Maastricht University

3D Human Pose Estimation (HPE) is an elementary module for human centered applications, including medical rehabilitation, autonomous driving, and human robot interaction. While 2D pose is sufficient for some tasks, many fields require 3D spatial information to be effective. Approached based on intensity cameras are often computationally heavy and have an lower bound on the latency due to limitations of the sensor, making them unsuitable for applications with limited computational capacity or requiring real-time output such as mobile robotics and augmented reality. Event cameras provide visual information at high temporal resolution, overcoming these limitations. Nevertheless, there are limited works that estimate 3D HPE from event cameras, and none that do so in real-time.

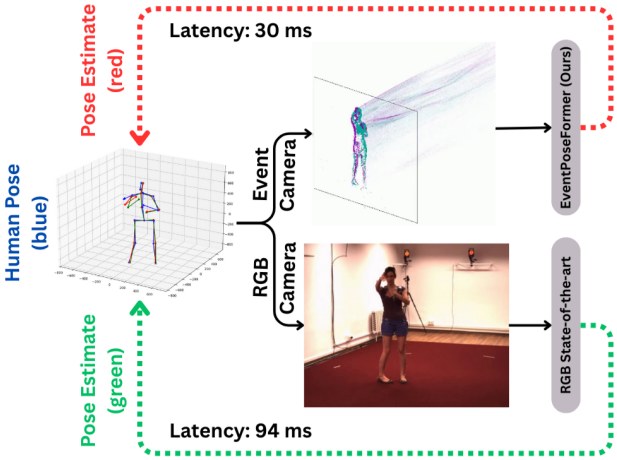

Figure 1: The EventPoseFormer pipeline converts event-camera data into 3D poses with 30ms latency, significantly faster than RGB state-of-the-art.

To address these challenges, we propose EventPoseFormer, an end-to-end monocular pipeline designed for real-time 3D HPE using event cameras. Our modular system takes MoveEnet [1] as a 2D backbone to generate high-frequency 13-joint 2D poses from event streams. More complex skeletons can be more informative to a lifting model, therefore, we introduced PoseBridge, a lightweight module that extrapolates these 13 joints into a 17-joint format using linear combinations of predicted points. Finally, a modified transformer-based lifter, PoseFormerV2[2], brings these joints to the 3D space. We adapted the lifting model for causal inference, by constraining the model to use only past and present data rather than future frames, ensuring the system is capable of low-latency, real-time processing.

The proposed system was evaluated on the Event-Human 3.6M dataset, demonstrating a significant performance-to-speed advantage. EventPoseFormer achieves an end-to-end latency of 30ms, a real-time frequency of 38Hz, and a 3D Mean Per Joint Position Error (MPJPE) of 85.7mm. In comparison, state-of-the-art RGB model HRNet-PoseFormerV2 provides similar but with 94ms latency, creating a noticeable lag for users. Our results highlight that while RGB methods may excel in static precision, EventPoseFormer provides a temporally closer match to ground truth in real-time scenarios. This work demonstrates that combining event-based sensing with latest neural networks is a viable and promising solution for high-speed event based sensing.

[1] Gaurvi Goyal et al. “MoveEnet: Online High-Frequency Human Pose Estimation with an Event Camera”. In: *IEEE/CVF CVPRW* (2023), pp. 4024–4033. URL: https://api.semanticscholar.org/CorpusID:260914713.

[2] Qitao Zhao et al. “PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation”. In: *IEEE CVPR*. 2023, pp. 8877–8886. DOI: 10.1109/CVPR52729.2023.00857.