An Open-Source Multimodal Dataset for Modeling Human Pose Priors

Eni Halilaj[1,2] *, Soyong Shin[1], Eric Rapp[1], Donglai Xiang[2], Yaadhav Raj[2]
[1]Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA
[2] The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
Email: * ehalilaj@andrew.cmu.edu

The future of motion tracking is markerless. While the accuracy of marker-based motion tracking technology has remained the same in decades, the field of computer vision is experiencing unprecedented progress in recent years, largely due to the success of deep neural networks. Algorithms for joint detection from images and video have enabled convenient joint center estimation [1]. More recent advances involve the use of three-dimensional (3D) deformable meshed models that can track 3D movement, in addition to joint centers, from single or multiple images [2] and videos [3]. The purpose of this symposium is to start a discussion around how we can leverage rapid advances in computer vision to better understand and improve human movement biomechanics. Markerless motion tracking will not only free us from time-consuming set-up and post-processing protocols, but also enable movement analysis in natural environments, such as rehabilitation clinics, patient homes, athletic training facilities, and workplaces. Along with a review of recent advances in computer vision, this talk will discuss a new multimodal dataset of different activities of daily living that we are collecting to improve the accuracy of markerless motion tracking algorithms, which will be made publicly available.

While revolutionary, vision-based motion tracking algorithms still lack the accuracy needed for many clinical and sports biomechanics applications. One of the challenges of these techniques is that they are data driven. The status quo for fitting 3D deformable models to images and video involves two overarching steps. The first is identification of joint centers on a frame using a deep neural network (e.g., OpenPose [1]) learned from thousands of manually annotated images (e.g., COCO dataset). The second step involves fitting a deformable parametric 3D meshed model to the joint centers using a generative approach. A discriminator then decides if this proposal is acceptable using information from pose priors, which are defined as the space of physiologically plausible poses and learned from experimental motion capture data. Limited datasets that contain only healthy individuals with little variation on physiological characteristics have been used until now (e.g., AMASS dataset [4]). This is due to the difficulty of collecting and processing marker-based motion capture data at the large scale necessary for data-driven models. Thus, modeling of human pose priors remains a challenge.

The goal of our study was to establish the feasibility of collecting a large, diverse, and multi-modal dataset to improve the modeling of human pose priors and the accuracy of markerless motion tracking, without relying on labor-intensive marker-based technology. We propose to use CMU Panoptic Studio (Fig. 1), which is a massive multi-view geodesic dome consisting of 480 VGA, 31 HD cameras, and 10 Kinect sensors triggered simultaneously. After receiving Institutional Review Board approval and completing informed consent, 86 subjects between the ages of 18 and 22, with no history of gait pathologies were recruited for this feasibility study and recoded inside the Studio while performing 6.5 minutes of activities, including walking, jogging, stair climbing, jumping, sitting and standing, and standard range-of-motion activities. Subjects were also equipped with 15 full-body inertial measurement units (Fig. 1B; MC10, Cambridge, MA). Algorithms that fuse data from all the cameras were used to estimate joint centers (Fig. 1C). Next steps will focus on 3D pose reconstruction using a parametric body model (Fig. 1D&E).

Extended to a larger number of subjects, activities, and clinical populations, this dataset should contribute toward better modeling of human pose priors that can resolve ambiguities in motion tracking from single cameras, IMUs, or a combination of modalities. In addition to better modeling of pose priors, this dataset could also be used to develop and benchmark new algorithms for the prediction of 3D kinematics from IMUs, single or multiple cameras, or fusion of IMUs and cameras given that rendering of ground truth pose requires no human power. Ultimately, accurate markerless motion tracking will lower the barriers for motion analysis and streamline monitoring of kinematics in natural environments, where it is currently not common; in clinics, where assessments are subjective; and in research laboratories, where extensive human power is required to collect and process marker data. Toward that goal, it is imperative that we stay abreast of advances in computer vision, collectively brainstorm about how the biomechanics community can contribute, and consider open-sourcing our data and tools.

References

1. Wei et al. "Convolutional pose machines." *CVPR* (2016).
2. Bogo et al. "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image." *CCV* (2016).
3. Kocabas et al. "VIBE: Video Inference for Human Body Pose and Shape Estimation." *arXiv:1912.05656* (2019).
4. Mahmood et al. "AMASS" ICCV (2019).

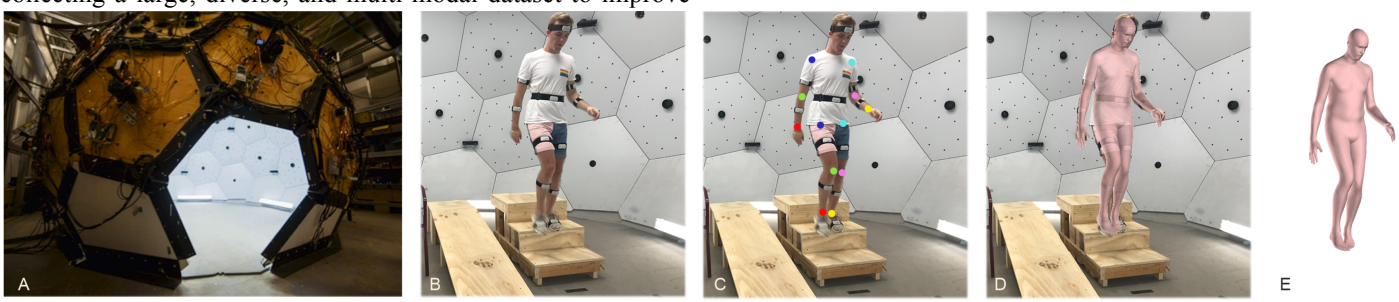

Figure 1. Experimental Set Up. A) CMU Panoptic Studio. B) Subjects were recorded with 480+ cameras and 15 IMUs while performing 6.5 minutes of activities. C) Joint center detection algorithms were used to track movement. D & E) Next steps will focus on 3D reconstruction of kinematics.