Ego-Body Pose Estimation via Ego-Head Pose Estimation

Jiaman Li (Stanford University; jiamanli@cs.stanford.edu)
C. Karen Liu† (Stanford University; karenliu@cs.stanford.edu)
Jiajun Wu† (Stanford University; jiajunwu@cs.stanford.edu)
DOI: 10.1145/3609468.3609473

Introduction

Estimating 3D human motion from an egocentric video, which records the environment viewed from the first-person perspective with a front-facing monocular camera, is critical to applications in VR/AR. However, naively learning a mapping between egocentric videos and full-body human motions is challenging for two reasons. First, modeling this complex relationship is difficult; unlike reconstruction motion from third-person videos, the human body is often out of view of an egocentric video. Second, learning this mapping requires a large-scale, diverse dataset containing paired egocentric videos and the corresponding 3D human poses. Creating such a dataset requires meticulous instrumentation for data acquisition, and unfortunately, such a dataset does not currently exist. As such, existing works have only worked on small-scale datasets with limited motion and scene diversity (yuan20183d; yuan2019ego; luo2021dynamics).

We introduce a generalized and robust method, EgoEgo, to estimate full-body human motions from only egocentric video for diverse scenarios. Our key idea is to use head motion as an intermediate representation to decompose the problem into two stages: head motion estimation from the input egocentric video and full-body motion estimation from the estimated head motion. For most day-to-day activities, humans have an extraordinary ability to stabilize the head such that it aligns with the center of mass of the body (keshner1988motor), which makes head motion an excellent feature for full-body motion estimation. More importantly, the decomposition of our method removes the need to learn from paired egocentric videos and human poses, enabling learning from a combination of large-scale, single-modality datasets (datasets with egocentric videos or 3D human poses only), which are commonly and readily available.

The first stage, estimating the head pose from an egocentric video, resembles the localization problem. However, directly applying the state-of-the-art monocular SLAM methods (teed2021droid) yields unsatisfactory results, due to the unknown gravity direction and the scaling difference between the estimated space and the real 3D world. We propose a hybrid solution that leverages SLAM and learned transformer-based models to achieve significantly more accurate head motion estimation from egocentric video. In the second stage, we generate the full-body motion based on a diffusion model conditioned on the predicted head pose. Finally, to evaluate our method and train other baselines, we build a large-scale synthetic dataset with paired egocentric videos and 3D human motions, which can also be useful for future work on visuomotor skill learning and sim-to-real transfer.

Our work makes four main contributions. First, we propose a decomposition paradigm, EgoEgo, to decouple the problem of motion estimation from egocentric video into two stages: ego-head pose estimation, and ego-body pose estimation conditioned on the head pose. The decomposition lets us learn each component separately, eliminating the need for a large-scale dataset with two paired modalities. Second, we develop a hybrid approach for ego-head pose estimation, integrating the results of monocular SLAM and learning. Third, we propose a conditional diffusion model to generate full-body poses conditioned on the head pose. Finally, we contribute a large-scale synthetic dataset with both egocentric videos and 3D human motions as a test bed to benchmark different approaches and showcase that our method outperforms the baselines by a large margin.
Method Overview

Our method, EgoEgo, estimates 3D human motion from a monocular egocentric video sequence. Our key idea is to leverage head motion: first estimating head motion from egocentric video, and then estimating full body motion from head motion. We show that head motion is an excellent feature for full-body motion estimation and a compact, intermediate representation that reduces the challenge into two much simpler sub-problems. Such a disentanglement also allows us to leverage a large-scale egocentric video dataset with head motion (but no full body motion) in stage one, and a separate 3D human motion dataset (but no egocentric videos) in stage two.

Head Pose Estimation from Egocentric Video

Estimating the head motion from an egocentric video can be viewed as a camera localization problem. However, we observed three issues that prevent us from directly applying the state-of-the-art monocular SLAM method (teed2021droid) to our problem. First, the gravity direction of the estimated head pose is unknown. Thus, the results cannot be directly fed to the full-body motion estimator, since it expects the head pose expressed in a coordinate frame where the gravity direction is \([0, 0, -1]^T\). Second, the estimated translation by monocular SLAM is not to scale when compared with the distance in the real world. Third, monocular SLAM tends to be less accurate in estimating relative head rotation than translation.

Based on these observations, we propose a hybrid method (shown in Figure 1) that leverages SLAM and learned models to achieve more accurate head pose estimation than the state-of-the-art SLAM alone. First, we develop a transformer-based model GravityNet to estimate the gravity direction from the rotation and the translation trajectories computed by SLAM. We rotate the SLAM translation by aligning the estimated gravity direction with the real gravity direction \([0, 0, -1]^T\) in the 3D world. Moreover, from the optical flow features extracted from the egocentric video, our method learns a model, HeadNet, to estimate head rotations and translation distance. The predicted translation distance of HeadNet is used to re-scale the translation estimated by SLAM. The predicted head rotation by HeadNet is directly used to replace the rotation estimated by SLAM.

Full-Body Pose Estimation from Head Pose

Predicting full-body pose from head pose is not a one-to-one mapping problem as different full-body motions may have the same head pose. Thus, we formulate the task using a conditional generative model. Inspired by the recent success of the diffusion model in image generation (rombach2022high), we deploy a diffusion model to generate full-body poses conditioned on head poses. We use the formulation proposed in the denoising diffusion probabilistic model (DDPM) (ho2020denoising).

Dataset and Benchmark

Our method does not need paired training data. Still, for benchmarking purposes, we develop a way to automatically synthesize a large-scale dataset with various paired egocentric videos and human motions.

Generate Motions in 3D Scenes

To generate a dataset with both egocentric video and ground truth human motions, we use a large-scale motion capture dataset AMASS (mahmood2019amass) and a 3D scene dataset Replica (straub2019replica). We convert the scene mesh from Replica to the signed distance field (SDF) for the penetration calculation. We divide each sequence
of AMASS (mahmood2019amass) into subsequences with 150 frames. For each subsequence, based on the semantic annotation provided by Replica (straub2019replica), we place the first pose in a random location with the feet in contact with the floor. Then we calculate penetration loss following (wang2021synthesizing) for each pose in this sequence. We empirically set a threshold and only keep the poses with penetration loss less than the threshold.

Synthesize Realistic Egocentric Images

The motion sequences produced by detecting penetration with 3D scenes provide the camera pose trajectories to render synthetic egocentric videos. AI Habitat (habitat19iccv; szot2021habitat) is a platform for embodied agent research that supports fast rendering given a camera trajectory and a 3D scene. We feed the head pose trajectories to the platform and synthesize realistic images in the egocentric view. We generate 1,664,616 frames with 30 fps, approximately 15 hours of motion in 18 scenes. We name the synthetic dataset AMASS-Replica-Ego-Syn (ARES) and show an example from our synthetic dataset in Figure 2.

Results

We compare the complete pipeline of EgoEgo with baseline methods PoseReg (yuan2019ego) and Kinpoly-OF (luo2021dynamics) on ARES as shown in Table 1. $O_{\text{head}}$, $T_{\text{head}}$ measures the head orientation and translation errors. MPJPE represents the mean joint position errors. Accel is to compute the acceleration error. We show that our EgoEgo outperforms all the baselines by a large margin. For qualitative comparisons and more results on ARES and real-world datasets, we encourage readers to check our project page. Our approach can generate more dynamic and realistic motions compared to the baselines.

Conclusion

We presented a generalized framework to estimate full-body motions from egocentric video. The key is to decompose the problem into two stages. We predicted the head pose from an egocentric video and fed the output from the first stage to estimate full-body motions in the second stage. In addition, we developed a hybrid solution to produce more accurate head poses on top of monocular SLAM. We also proposed a conditional diffusion model to generate diverse high-quality full-body motions from predicted head poses. To benchmark different methods in a large-scale dataset, we proposed a data generation pipeline to synthesize a large-scale dataset with paired egocentric videos and 3D human motions. We showcased superior results on both the synthetic and the real-captured dataset compared to prior work.

Jiaman Li is a PhD student in the Computer Science Department at Stanford University. Her research interest is the intersection of computer vision and graphics. She currently focuses on 3D human motion modeling including human motion estimation and human motion synthesis.
C. Karen Liu is an associate professor in the Computer Science Department at Stanford University. Her research interests span computer animation, robotics, reinforcement learning, physics simulation, optimal control, and computational biomechanics.

Jiajun Wu is an Assistant Professor of Computer Science at Stanford University, affiliated with the Stanford AI Lab (SAIL) and the Stanford Vision and Learning Lab (SVL). He studies machine perception, reasoning, and interaction with the physical world, drawing inspiration from human cognition.