Gravity-Aware Monocular 3D Human-Object Reconstruction
Supplementary Material

Rishabh Dabral[1,2] Soshi Shimada[2] Arjun Jain[3,4] Christian Theobalt[2] Vladislav Golyanik[2]

[1]IIT Bombay [2]MPI for Informatics, SIC [3]IISc Bangalore [4]Fast Code AI

This supplementary material contains further details on GRAVICAP. We also provide a supplementary video with further analysis of our method, in-the-wild 3D reconstructions and qualitative comparisons with other methods.

1. In-the-Wild Results

To verify the accuracy of visual metrology achievable by our method, we test it on several real-world scenarios where the distance references are available.

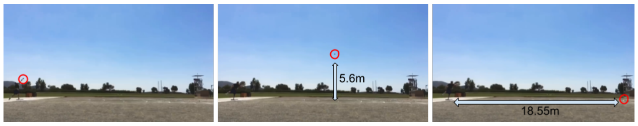

Figure 1: Shot put sequence

Professional Shotput Throw: A professional shot put thrower can throw in the range of 20 meters. The image in Fig. 1 is a reference to a throw by the world record holder Ryan Crouser[1]. Since the person is extremely blurred in the clip, we test our method with only object-related constraint E_b, while ensuring that the magnitude of gravity vector is $9.81\,m/s^2$. Our method estimates the throw to be 18.557 m long. Further, we estimate the maximum point of the object's trajectory to be $5.46\ m$. Finally, the estimated gravity direction indicates an upward tilt of 13° of the camera.

Basketball Throw: We measure the final position of the ball in the trajectory of the throw depicted in Fig. 2. We note the absolute y-position of the ball when it touches the hoop and compare it with the absolute y-position of the feet (as estimated by VNect). The difference between the two gives us an estimate of the height of the hoop. The trajectory estimates show a height of $3.03\ m$, which is close to the actual height of the hoop ($3.05\ m$).

2. Noise Sensitivity Analysis

To test the sensitivity of GRAVICAP to noise, we perform a sensitivity analysis and summarise the results in Table 1.

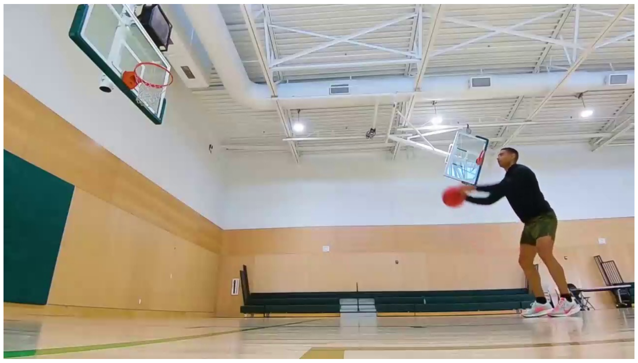

Figure 2: Basketball throw sequence

	GT	$\sigma = 10$	$\sigma = 30$	$\sigma = 50$	$\sigma = 100$
Pose, 6 DoF	11.7	23.2	39.9	88.3	227.2
Pose, 7 DoF	8.9	13.4	31.1	69.8	224
Pose, 10 DoF	26.3	60.87	126.5	150.4	225.0
Object	12.4	76.3	134.4	155.8	>400

Table 1: Comparing the effect of adding Gaussian noise to the ground-truth 3D poses and 2D object trajectories on root translation predictions. The unit of σ is mm for poses and pixels for 2D object trajectories.

As expected, the performance is affected by strong 2D object trajectory perturbations, as high σ, *i.e.,* the standard deviation of Gaussian noise, disrupts its parabolic nature.

3. Detection of Trajectory Breaks

To detect an episode switch in multi-episode sequences, we traverse the 2D trajectory with a sliding window of five frames. During each slide, we measure the position difference between the positions of the objects in adjacent frames. For a switch to have occurred, the direction of position differences of the first half must be opposite to that of the second half. This, however, is not sufficient for detecting a switch, as the same happens when the object reaches its peak. To confirm a switch, we measure the change in magnitude of velocity during the inversion. We set a threshold of 10 pixels per frame to confirm the episode switch.

[1] https://www.youtube.com/watch?v=TZANFlvsXv4

4. Handling Multi-Episodes and Two Persons

Multi-Episodes: Whenever we have a multi-episode sequence, we perform joint optimisation on all the episodes. This leads to coherent reconstruction in the sense that the trajectory is continuous and jitter-free. For this, we impose the continuity constraint, E_{co}, that ensures that the last position of the previous episode is the same as the first position of the current episode. Specifically, if $i = 2, 3, \dots$ refers to an episode in a multi-episode sequence, then

$$E_{co} = \left\| B_T^{i-1} - B_0^i \right\|_2^2 , \tag{1}$$

where T is the number of frames in the i^{th} episode.

Two Persons: For the case with two persons, while the pose projection constraint remains the same (only this time, applied to multiple poses), the contact term needs to accommodate information about which person the object is in contact with:

$$\arg\min E_c(P) = \arg\min \sum_{(c,t)\in\mathcal{C}} \left\| P_t^{c,\delta} - B_t \right\|_2^2 . \tag{2}$$

In this equation, $P_t^{c,\delta}$ indicates the 3D position of the joint c of the person δ ($|\delta|$ is the number of people in the scene) at time of contact t.

```
Root
  |- S1
    |- Cam1
      |- Eps1
         annot_e1.pkl
         |- Images
            s_01_c_01_e_01_00001.jpg
            s_01_c_01_e_01_00002.jpg
            ...
      |- Eps2
         annot_e2.pkl
         |- Images
            s_01_c_01_e_02_00001.jpg
            s_01_c_01_e_02_00002.jpg
            ...
```

Figure 3: Dataset structure tree

5. Dataset Structure

Our dataset consists of nine activity sequences (eight single-person and one with two persons) performed by four subjects. For each sequence, we have two-three multi-episodes involving one or more episodes in succession. We provide annotations and images for up to three camera views. The annotations include the 2D and 3D human poses, 2D and 3D object trajectories, the camera calibration parameters and point-of-contact information (frame numbers and joints which are closest to the body at the time of contact). The images are of size 1200x877 px. The structure of the dataset is demonstrated in Fig. 3.