Trajectory Optimization for Physics-Based Reconstruction of 3d Human Pose from Monocular Video

Supplementary Material

Erik Gärtner[1,2] Mykhaylo Andriluka[1] Hongyi Xu[1] Cristian Sminchisescu[1]

[1]Google Research, [2]Lund University

erik.gartner@math.lth.se

{mykhayloa,hongyixu,sminchisescu}@google.com

This supplementary material provides further details on our methodology and the data we used. §1 presents details on our physical human body model, §2 provides details regarding our simulation parameters, §3 presents our physics metrics, in §4 we present the datasets used in our experiments, §5 provides details about our method's hyperparameters, and lastly §6 summarizes our computational setup. When referring to equations or material in the main paper we will denote this by *(mp)*. Finally, please see our supplemental video for qualitative results of our method at tiny.cc/traj-opt.

1. Physical Body Model

Given a GHUM [11] body mesh $\mathbf{M}(\boldsymbol{\beta}, \boldsymbol{\theta}_0)$ associated with the shape parameters $\boldsymbol{\beta}$ and the rest pose $\boldsymbol{\theta}_0$, we build a simulation-ready rigid multibody human model that best approximates the mesh with a set of parameterized geometric primitives (*cf*. fig. 1). The hands and feet are approximated with boxes whereas the rest of the body links are approximated with capsules. The primitives are connected and articulated with the GHUM body joints.

Inspired by [1], we optimize the primitive parameters by minimizing

$$L(\boldsymbol{\psi}) = \sum_{b \in \mathbf{B}} \sum_{\mathbf{v}_g \in \mathbf{M}_b} \min_{\mathbf{v}_p \in \hat{\mathbf{M}}_b} ||\mathbf{v}_g - \mathbf{v}_p|| + \\ + \sum_{b \in \mathbf{B}} \sum_{\mathbf{v}_p \in \hat{\mathbf{M}}_b} \min_{\mathbf{v}_g \in \mathbf{M}_b} ||\mathbf{v}_p - \mathbf{v}_g||, \quad (1)$$

where $\boldsymbol{\psi}$ are the size parameters for the primitives, i.e. length and radius for the capsules, and depth, height and width for the boxes. The loss penalizes the bi-directional distances between pairs of nearest points on the GHUM mesh $\mathbf{M}_b$ and surface of the primitive geometry $\hat{\mathbf{M}}_b$ associated with the body link b.

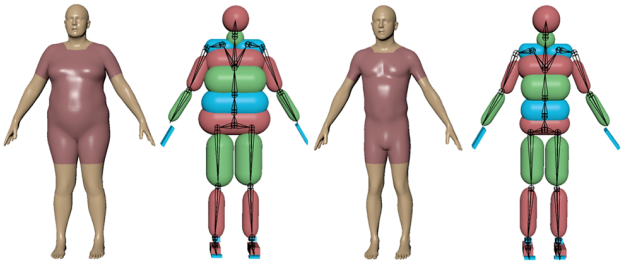

Figure 1. The physical body model's shape and mass parameters are based on an associated GHUM [11] mesh.

Furthermore, we learn a nonlinear regressor $\boldsymbol{\psi}(\boldsymbol{\beta})$ with an MLP that performs fast shape approximation at run time. The regressor consists of two 256-dimensional fully connected layers, and is trained with 50K shapes generated with Gaussian sampling of the latent shape space $\boldsymbol{\beta}$ together with the paired optimal primitive parameters using (1).

Our physical model share an identical skeleton topology with GHUM but does not model the face and finger joints, due to the focused interest on the body dynamics in this work. Extending with finger joints, however, would enable simulation of hand-object interactions which would be interesting, but we leave this for future work. We note that there is a bijective mapping for the shared 16 body joints between our model and GHUM, which allows for fast conversion between the physical and stastical representation.

2. Simulation Details

We run the Bullet simulation at 200 Hz, with friction coefficient $\mu = 0.9$ and gravitational acceleration constant 9.8 m/s^2. The PD-controllers controlling each torque motor is tuned with position gain $k_p = 4.0$, velocity gain $k_d = 0.3$, and torque limits similar to those presented in [5].

Weight	H36M	AIST	HumanEva-I	Grid
w_{COM}	15.0	15.0	15.0	{1, 2, 5, 10, 15, 25 }
w_{pose}	0.5	0.5	0.5	{0.1, 0.5, 1, 2 }
w_{2d}	4.0	4.0	4.0	{1, 2, 4, 8, 10 }
w_{nf}	1.0	1.0	1.0	{0.001, 0.1, 1, 10}
w_{TV}	1.0	1.0	1.0	{0.1, 1, 10}
w_{lim}	1.0	1.0	1.0	{0.1, 1, 10}

Table 1. Weights of the objective function described in §3.3 *(mp)*and (3) for our three main datasets: Human3.6M [3], AIST [9], and HumanEva-I [8]. "Grid" specifies the values evaluated while selecting hyperparameter values. Note that we did not exhaustively explore all combination.

Sequence	Subject	Camera Id	Frames
Phoning	S11	55011271	400-599
Posing_1	S11	58860488	400-599
Purchases	S11	60457274	400-599
SittingDown_1	S11	54138969	400-599
Smoking_1	S11	54138969	400-599
TakingPhoto_1	S11	54138969	400-599
Waiting_1	S11	58860488	400-599
WalkDog	S11	58860488	400-599
WalkTogether	S11	55011271	400-599
Walking_1	S11	55011271	400-599
Greeting_1	S9	54138969	400-599
Phoning_1	S9	54138969	400-599
Purchases	S9	60457274	400-599
SittingDown	S9	55011271	400-599
Smoking	S9	60457274	400-599
TakingPhoto	S9	60457274	400-599
Waiting	S9	60457274	400-599
WalkDog_1	S9	54138969	400-599
WalkTogether_1	S9	55011271	400-599
Walking	S9	58860488	400-599

Table 2. The subset of Human3.6M used in the ablation experiments. Note that the data was downsampled from 50 to 25 FPS.

Sequence	Frames
gBR_sBM_c06_d06_mBR4_ch06	1-120
gBR_sBM_c07_d06_mBR4_ch02	1-120
gBR_sBM_c08_d05_mBR1_ch01	1-120
gBR_sFM_c03_d04_mBR0_ch01	1-120
gJB_sBM_c02_d09_mJB3_ch10	1-120
gKR_sBM_c09_d30_mKR5_ch05	1-120
gLH_sBM_c04_d18_mLH5_ch07	1-120
gLH_sBM_c07_d18_mLH4_ch03	1-120
gLH_sBM_c09_d17_mLH1_ch02	1-120
gLH_sFM_c03_d18_mLH0_ch15	1-120
gLO_sBM_c05_d14_mLO4_ch07	1-120
gLO_sBM_c07_d15_mLO4_ch09	1-120
gLO_sFM_c02_d15_mLO4_ch21	1-120
gMH_sBM_c01_d24_mMH3_ch02	1-120
gMH_sBM_c05_d24_mMH4_ch07	1-120

Table 3. Sequences used for evaluation on AIST.

3. Additional Metrics

In addition to the standard 2d and 3d joint position error metrics, we evaluate our reconstructions using physical plausibility metrics similar to those proposed in [6]. Since the authors were unable to share their code we implement our own versions the metrics which doesn't require footground contact annotations. A foot contact is defined as at least $N = 10$ vertices of a foot mesh being in contact with the ground plane. We set the contact threshold to $d = 0.005$ m for kinematics. To account for the modeling error when approximating the foot with a box primitive we set the contact threshold for dynamics to $d = -0.015$ m.

Footskate. The percentage of frames in a sequence where either foot joint moves more than 2 cm between two adjacent frames while the corresponding foot was in contact with the ground-plane.

Float. The percentage of frames in a sequence where at least one of the feet was not in contact but was within 2 cm of the ground-plane. This metric captures the common issue of reconstructions floating above the ground while not penalizing correctly reconstructed motion of e.g. jumps.

Velocity. The mean error between the 3d joint velocities in the ground-truth data and the joint velocity in the reconstruction. High error velocity indicates that the estimated motion doesn't smoothly follow the trajectory of the true motion. We define the velocity error as

$$e_v = \frac{1}{N}\sum_{i=1}^{N}\sum_{k\in K}|\dot{\tilde{\mathbf{x}}}_k^i - \dot{\mathbf{x}}_k^i|, \tag{2}$$

where $\dot{\tilde{\mathbf{x}}}_k^i$ is the magnitude of the ground-truth 3d joint velocity vector (in m/s) for joint k at frame i and where $\dot{\mathbf{x}}_k^i$ denotes the reconstructed joint. We estimate the velocity using finite differences from 3d joint positions and use first frame translation aligned joint estimates (as in MPJPE-G).

Sequence	Subject	Camera Id
S11	Directions_1	60457274
S11	Discussion_1	60457274
S11	Greeting_1	60457274
S11	Posing_1	60457274
S11	Purchases_1	60457274
S11	TakingPhoto_1	60457274
S11	Waiting_1	60457274
S11	WalkDog_1	60457274
S11	WalkTogether_1	60457274
S11	Walking_1	60457274
S9	Directions_1	60457274
S9	Discussion_1	60457274
S9	Greeting_1	60457274
S9	Posing_1	60457274
S9	Purchases_1	60457274
S9	TakingPhoto_1	60457274
S9	Waiting_1	60457274
S9	WalkDog_1	60457274
S9	WalkTogether_1	60457274
S9	Walking_1	60457274

Table 4. The evaluation subset of Human3.6M used in the main evaluation. The subset is similar to the one used in [7]. We down-sampled the data from 50 FPS to 25 FPS.

4. Datasets

Human3.6M. We use two subsets for our experiments on Human3.6M [3]. When we compare our method to state-of-the-art methods we use a dataset split similar to the one used in [10]. See tab. 4 for the complete lists of sequences we use. Similarly to [7, 10], we down sample the sequences from 50 FPS to 25 FPS.

When perform ablations of our model we a smaller subset where we select 20 4-sec sequences from the test split of Human3.6M dataset (subjects 9 and 11). We selected sequences that show various dynamic motions such as walking dog, running and phoning (with large motion range), to sitting and purchasing (with occluded body parts). For each sequence, we randomly selected one of the four cameras. We list the sequences in tab. 2.

HumanEva-I. We evaluate our method on the subset of HumanEva-I walking sequences [8] as selected by [6], see tab. 5.

AIST. We select four second video sequences from the public dataset [4, 9], showing fast and complex dancing motions, picked randomly from one of the 10 cameras. We list our selected sequences in tab. 3.

"In-the-wild" internet videos. We perform qualitative evaluation of our model on videos of dynamic motions rarely found in laboratory captured datasets. These videos were made available on the internet under a CC-BY license which grants the express permission to be used for any purpose. Note that we only used the videos to perform qualitative analysis of our approach – the videos will not be redistributed as a dataset.

Sequence	Subject	Camera Id	Frames
Walking	S1	C1	1-561
Walking	S2	C1	1-438
Walking	S3	C1	1-490

Table 5. Sequences used for evaluation on HumanEva-I.

4.1. Human Data Usage

This work relies on recorded videos of humans. Our main evaluation is performed on two standard human pose benchmarks: Human3.6M[1] [3] and AIST[2] [9]. These datasets have been approved for research purposes according to their respective websites. Both datasets contain recordings of actors in laboratory settings. To complement this, we perform qualitative evaluation on videos released on the internet under creative commons licenses.

5. Hyperparameters

The most important hyperparameters are the weights of the weighted objected function described in §3.3 *(mp)*. Where combined loss function is given by

$$\begin{aligned} L = & w_{COM} L_{COM} + w_{pose} L_{pose} \\ & + w_{2d} L_{2d} + w_{nf} L_{nf} + w_{TV} L_{TV} \\ & + w_{lim} L_{lim}. \end{aligned} \tag{3}$$

We tuned the weights on sequences from the training splits. The goal was to scale the different components such that they have roughly equal magnitudes while minimizing the MPJPE-G error. See tab. 1 for details regarding the search grid and the chosen parameter values.

6. Computational Resources

For running small experiments we used a desktop workstation equipped with an "Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz" CPU, 128 GB system memory and two NVIDIA Titan Xp GPUs. We ran kinematics in the cloud using instances with a V100 GPU, 48 GB of memory and 8 vCPUs. In the dynamics experiments, we used instances with 100 vCPUs and 256 GB of memory for the CMA-ES [2] optimization. Optimizing a window of 1 second of video takes roughly 20 min using a 100 vCPUs instance.

[1] http://vision.imar.ro/human3.6m/

[2] https://aistdancedb.ongaaccel.jp/

References

[1] Mazen Al Borno, Ludovic Righetti, Michael J. Black, Scott L. Delp, Eugene Fiume, and Javier Romero. Robust Physics-based Motion Retargeting with Realistic Body Shapes. In *Computer Graphics Forum*, 2018. 1

[2] Nikolaus Hansen. *The CMA Evolution Strategy: A Comparing Review*, pages 75–102. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. 3

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 2, 3

[4] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation, 2021. 3

[5] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Trans. Graph.*, 37(4):143:1–143:14, July 2018. 1

[6] Davis Rempe, Leonidas J. Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[7] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics*, 39(6), dec 2020. 3

[8] L. Sigal, A. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1):4–27, Mar. 2010. 2, 3

[9] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 501–510, Delft, Netherlands, Nov. 2019. 2, 3

[10] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *Int. Conf. Comput. Vis.*, 2021. 3

[11] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3d human shape and articulated pose models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6184–6193, 2020. 1