Supplementary Material *for*

Beyond Static Features for Temporally Consistent 3D Human Pose and Shape from a Video

Hongsuk Choi[1] Gyeongsik Moon[1] Ju Yong Chang[2] Kyoung Mu Lee[1]

[1]ECE & ASRI, Seoul National University, Korea [2]ECE, Kwangwoon University, Korea

{redarknight,mks0601,kyoungmu}@snu.ac.kr, juyong.chang@gmail.com

S1. More qualitative results

We provide more qualitative results in the online video [1], which consists of three parts. The first part shows the qualitative results of our TCMR on in-the-wild videos that have fast and diverse motions from 3DPW [14]. We also provide the outputs rendered from the opposite view. The second part compares the proposed TCMR with VIBE [8] and MEVA [10]. The results are rendered on a plain background with a fixed camera to clearly compare the temporal consistency and smoothness of 3D human motion following MEVA [10]. The fixed camera has the fixed weak-perspective camera parameters s and t, which are set to one and zero, respectively. The last part provides the results of TCMR on Internet videos. The bounding boxes of people in the videos are tracked by a multi-person tracker that uses YOLOv3 [12]. With the cropped images from the bounding boxes, our TCMR processes 41 frames per second (fps) for the video [2] with 5 people. A single NVIDIA RTX 2080Ti GPU is used for the test.

S2. Human evaluation.

We surveyed 50 people to pick the most realistic motion from TCMR, MEVA, and VIBE outputs on 20 sequences of 3DPW [14] validation and test sets. TCMR, MEVA, and VIBE got 69%, 26%, and 5% votes, respectively. The result is coherent with the acceleration error results of the three methods in the main manuscript.

S3. Attention values in feature integration.

During the temporal feature integration, the past and future temporal features are weighted more than the current temporal feature, and the variation range of each attention value is $\pm 20\%$. The past and future temporal features' attention values tend to become larger when the current pose is difficult or the motion is fast. The attached videos plot the attention values of the past, future, and current temporal features on two sequences of 3DPW [14]. The values are written at the top-right of frames, and the sum is always 1. As the video shows, the attention value of the current temporal feature does not drop below 0.4 when a subject is walking in slow motion, whereas the value overall stays below 0.4 when a subject is playing basketball with fast movement and complex poses.

S4. Datasets

3DPW. 3DPW [14] is captured from in-the-wild and contains 3D human pose and shape annotations. It consists of 60 videos and 51K video frames in total, which are captured with a phone at 30 fps. IMU sensors are leveraged to acquire the groundtruth 3D human pose and shape. We follow the official split protocol to train and test our model, where train, validation, test sets consist of 24, 12, 24 videos, respectively. Also, we report MPVPE on 3DPW because it only has groundtruth 3D shape among the datasets we used. We use 14 joints defined by Human3.6M [5] for evaluating PA-MPJPE and MPJPE following the previous works [6–9].

Human3.6M. Human3.6M [5] is a large-scale indoor 3D human pose benchmark, which consists of 15 action categories and 3.6M video frames. Following [8], our TCMR is trained on 5 subjects (S1, S5, S6, S7, S8) and tested on 2 subjects (S9, S11). We subsampled the dataset to 25 fps (originally 50 fps) for training and evaluation on the acceleration error. 14 joints defined by Human3.6M are used for computing PA-MPJPE and MPJPE.

MPI-INF-3DHP. MPI-INF-3DHP [11] is a 3D benchmark mostly captured from indoor environment. The train set has 8 subjects, 16 videos per subject, and 1.3M video frames captured at 25 fps in total. It exploits a marker-less motion capture system and provides 3D human pose annotations.

The test set contains 6 subjects performing 7 actions in both the indoor and outdoor environment. The positional errors (*i.e.*, PA-MPJPE and MPJPE) of TCMR are measured

[1] https://www.youtube.com/watch?v=WB3nTnSQDII

[2] https://www.youtube.com/watch?v=Opry3F6aB1I

Table S1: Comparison between different models using ResNet with different initialization to extract static features. All models use the same SMPL parameter regressor pretrained by SPIN [9].

ResNet initialization	remove residual	PoseForecast	PA-MPJPE↓	Accel↓
ResNet with random initialization	✗	✗	126.5	24.3
ResNet pretrained on ImageNet [13]	✗	✗	103.7	65.5
ResNet from SPIN [9]	✗	✗	55.6	29.2
ResNet from SPIN [9] (TCMR. Ours.)	✓	✓	**53.9**	**7.7**

on the valid frames, which are composed of every 10th frame approximately, using 17 joints defined by MPI-INF-3DHP. The acceleration error is computed using all frames.

InstaVariety. InstaVariety is a 2D human dataset curated by HMMR [7], whose videos are collected from Instagram using 84 motion-related hashtags. There are 28K videos with an average length of 6 seconds, and OpenPose [2] is leveraged to acquire pseudo-groundtruth 2D pose annotations.

Penn Action. Penn Action [15] contains 2.3K video sequences of 15 different sports actions. It has a total of 77K video frames annotations for 2D human poses, bounding boxes, and action categories.

PoseTrack. PoseTrack [1] is a 2D benchmark for multi-person pose estimation and tracking in videos. It contains 1.3K videos and 46K annotated frames in total. The videos are captured at different fps, varying around 25 fps. We use 792 videos from the official train set, which has 2D pose annotations for 30 frames in the middle of the video.

S5. Effect of pretrained ResNet

Due to lack of video data, our TCMR and previous video-based methods [7, 8, 10] employ ResNet [4] pretrained by the single image-based 3D human pose and shape estimation methods [6, 9] to extract static features from input frames. The pretrained ResNet is trained on large-scale in-the-wild 2D human pose datasets and provides reliable static features. However, it is also one reason for the strong dependency of the system on the current static feature. The current static feature extracted by the pretrained ResNet already contains a strong cue on the current 3D human pose and shape, leading the system to leverage temporal information marginally.

In this regard, an alternative to our TCMR, one could train models from scratch without using the ResNet pretrained by [6, 9] to extract static features to reduce the strong dependency. Table S1 compares our TCMR, the baseline (the third row), and the models that do no use the ResNet pretrained by SPIN [9]. As the table shows, the models that do no use the ResNet pretrained by SPIN [9] reveal very high per-frame 3D pose errors. This indicates that training the models with only video data in the current literature is not sufficient for accurate 3D human pose estimation. The interesting part is that the model using ResNet with random initialization provides the highest 3D pose error but the lowest acceleration error among the models without our TCMR. While the high pose error attributes to the lack of train data, the low acceleration error implies that the strong cue of the current static feature adversely affects the temporal consistency of 3D human motion.

Table S2: Performance comparison between two networks taking different input fps on 3DPW [14]. The numbers in the second row are from Table 4 of the main manuscript.

input fps	PA-MPJPE↓	Accel↓
15	53.5	15.3
30	**52.7**	**7.1**

In summary, with the insufficient video data in the current literature, the proposed TCMR significantly improves the temporal consistency of 3D human motion by reducing the strong dependency on the current static feature. It also preserves the per-frame 3D pose accuracy by leveraging the ResNet pretrained on large-scale in-the-wild 2D human pose datasets to extract useful static features.

S6. Effect of input fps

Table S2 shows the effect of input fps. The acceleration error doubles when input fps reduces by half, whereas the accuracy remains relatively the same. The result indicates that TCMR can still fix invalid poses using relatively sparse temporal information. The result also implies that temporally dense information is critical for temporal consistency of outputs, which is intuitive.

S7. Pose2Mesh with temporal smoothing

We performed temporal smoothing on Pose2Mesh [3], the state-of-the-art single image-based 3D human pose and shape estimation method. Pose2Mesh wins the first in MPJPE, MPVPE, and acceleration error and the second in PA-MPJPE among single image-based methods according to Table 6 of the main manuscript. Pose2Mesh with euro-filter achieves PA-MPJPE 58.6, MPJPE 89.6, acceleration error 12.9 on 3DPW. TCMR still outperforms the smoothed Pose2Mesh by nearly twice in temporal consistency without any post-processing.

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. *CVPR*, 2018. 2

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *CVPR*, 2017. 2

[3] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. *ECCV*, 2020. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016. 2

[5] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2014. 1

[6] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. *CVPR*, 2018. 1, 2

[7] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. *CVPR*, 2019. 1, 2

[8] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. *CVPR*, 2020. 1, 2

[9] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. *ICCV*, 2019. 1, 2

[10] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3D human motion estimation via motion compression and refinement. *ACCV*, 2020. 1, 2

[11] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. *3DV*, 2017. 1

[12] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *IJCV*, 2015. 2

[14] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. *ECCV*, 2018. 1, 2

[15] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. *ICCV*, 2013. 2