Supplementary Material: Exploiting temporal context for 3D human pose estimation in the wild

Anurag Arnab[1*†] aarnab@robots.ox.ac.uk
Carl Doersch[2*] doersch@google.com
Andrew Zisserman[1,2] zisserman@google.com
[1]University of Oxford [2]DeepMind

Section A1 lists the hyperparameters we used for our bundle adjustment, whilst Sec. A2 provides some more details about the dataset we automatically generated from Kinetics.

A1. Experimental Details

A1.1. Bundle adjustment hyperparameters

Table A1 shows the values of our bundle adjustment hyperparameters for our experiments.

Table A1. Bundle adjustment hyperparameters used for experiments

Hyperparameter	Human 3.6M [1]	Kinetics [4]
λ_R	1×10^{-3}	1×10^{-3}
λ_I	10	1
λ_β	0.2	0.05
λ_J	1×10^{-4}	1×10^{-4}
λ_1	5	0.2
λ_2	1×10^{-4}	1×10^{-3}
λ_3	2	20
τ_R	–	50
τ_I	–	2×10^{-2}

Note that the 2D joint positions, $\mathbf{x}$ are measured in pixels, and that the largest spatial dimensions of a video frame is typically around 450. On the other hand, the 3D joint positions, $\mathbf{X}$ and camera parameters are typically in the range $[-1, 1]$. As the range of the 2D joint positions is higher, the values of λ_R and λ_2, are small, even though they have a significant effect on the bundle adjustment.

λ_I and λ_β are higher on Human 3.6M than they are on Kinetics. These weights are used in the prior term that encourages the bundle adjustment result to stay close to the initialisation (Eq. 8 of main paper). The initialisation that we get from HMR [2] is far better on Human 3.6M than on Kinetics, which is why λ_I and λ_β are higher on Human 3.6M. It is expected that HMR performs better on Human 3.6M as it has been trained with 3D supervision from this dataset.

*Equal contribution.
†Work done during an internship at DeepMind

Table A2. Ablation study of our HMR retraining schemes. PA-only 3D means during our retraining of HMR, we discard the losses on SMPL joints and absolute 3D locations and only use losses on joints after Procrustes alignment. No 2D means disabling all HMR datasets that contain only 2D data (and therefore disabling the adversarial prior which is only used on 2D datasets).

	3DPW	HumanEVA
Original data, original training	77.2	85.7
Original data, PA-only 3D	78.7	86.2
Original data, PA-only 3D, no 2D	144.6	99.2
Original + Kinetics data, original training	91.1	90.0
Original + Kinetics data, PA-only 3D, no 2D	**72.2**	**82.1**

A1.2. HMR training modification and ablation

When training with Kinetics data, we find that it is beneficial to not use any of the original 2D data used by HMR, and thus also to not use the adversarial pose prior since it is only used on 2D pose datasets [2]. We suspect that this is because the adversarial pose prior encourages predictions that are closer to the mean pose, and since we use HMR to initialise our bundle adjustment, our Kinetics data may also have a slight bias towards this mean pose. Applying the same prior while retraining may aggravate this problem.

We also find it's important to train only on 3D keypoints after Procrustes alignment, rather than training directly on SMPL joint angles and absolute 3D keypoint locations. Note this means that HMR only learns to predict the camera orientation by minimizing 2D reprojection error. We suspect that this strategy is effective because Kinetics has a very large range of camera orientations, which may not match well with evaluation datasets that have less variety in camera pose.

Table A2 shows that our modifications to the HMR training procedure help only when we train with additional Kinetics data. When using the original training data, our modified training procedure does not improve results. Removing

the original 2D data from training also has a large negative impact on performance. This is because the original training data has a relatively small amount of 3D supervision (Human 3.6M [1] and MPI-3DHP [5]).

A1.3. 3DPW Evaluation Protocol

The 3DPW dataset contains 60 clips, consisting of outdoor videos captured from a moving mobile phone and 17 IMUs attached to the subjects [6]. The IMU data allowed the authors to accurately compute 3D poses which we use as ground truth. We evaluate on the test set comprising 24 videos, using the 14 keypoints that are common on both the MS-COCO and SMPL skeletons as also done by [3]. We only evaluate on frames where enough of the person is visible to estimate a 3D pose for it. This is performed by discarding examples where less than 7 ground-truth 2D keypoints are visible. We compute the Procrustes-aligned error independently for each pose, and then average errors for each tracked person within each video, before finally averaging over the entire dataset (thus videos with two people count twice as much as videos with one).

We follow the standard HMR pre-processing when evaluating input images: the bounding box around the person is scaled such that the height of the person is about 150 pixels. As we are not evaluating the 2D person bounding-box performance of our algorithm, we use the ground truth person bounding-box. The height of the person is estimated by taking the difference of the highest and lowest valid keypoints, where a keypoint is considered valid if its score is greater than 0.1.

A2. Dataset statistics

Figure A1 visualises the distribution of Kinetics action classes in our dataset. We can see that the distribution has a fairly long tail: Our bundle adjustment method works well for a variety of object classes, including many types of dancing and various outdoor activities, where there are usually not many people in the video clip and the whole body is visible. There are also many classes for which only a handful of videos are automatically selected. These are typically classes such as “tying tie”, “bending metal” and “knitting” where the person is usually not fully visible. Note that there are 400+ clips for each action in the Kinetics-400 dataset [4] that we use, and that we have always selected at least one video of each action class.

References

[1] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 36(7):1325–1339, 2014. 1, 2

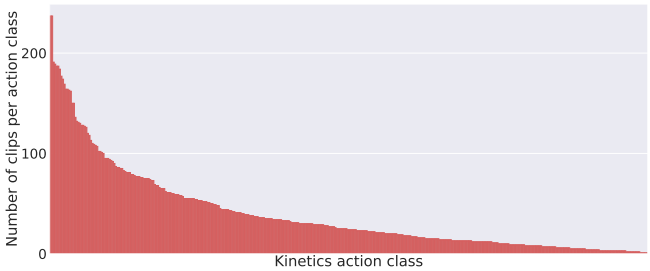
[2] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1

[3] A. Kanazawa, J. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 2

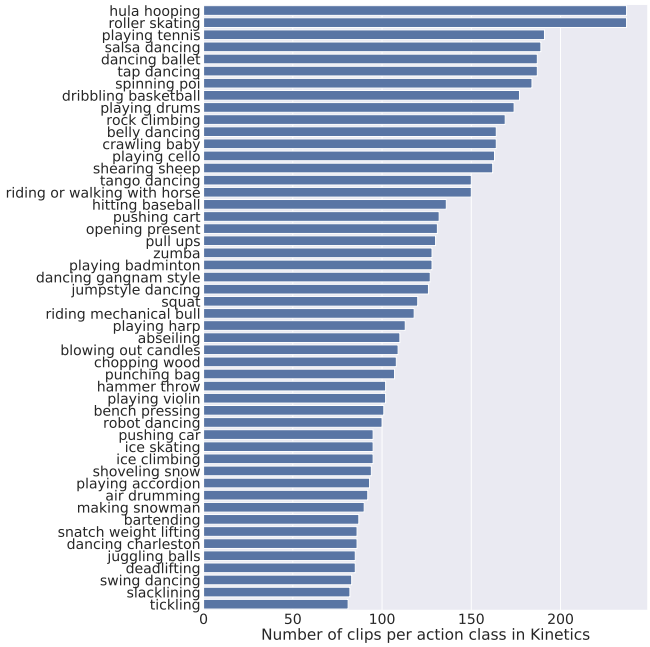
[4] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. In *arXiv preprint arXiv:1705.06950*, 2017. 1, 2

[5] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3D human pose estimation in the wild using improved cnn supervision. In *3DV*, 2017. 2

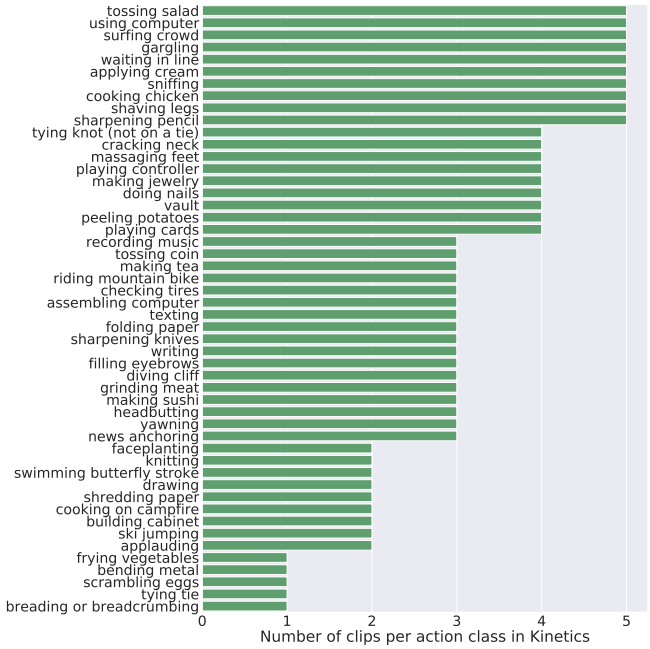
[6] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 2



(a) Number of clips selected per action class. For legibility, the action classes are not shown in the x axis, and the most- and least-common classes are shown below instead.



(b) The number of clips selected per class for the 50 most common Kinetics action classes.



(c) The number of clips selected per class for the 50 least common Kinetics action classes.

Figure A1. Number of video clips selected per action class in the Kinetics dataset. (a) shows the overall distribution of video clips selected per action class, whilst (b) and (c) show the most- and least-common Kinetics action classes respectively.