History Repeats Itself: Human Motion Prediction via Motion Attention ——Supplementary Material——

Wei Mao[1], Miaomiao Liu[1], and Mathieu Salzmann[2]

[1] Australian National University, Canberra, Australia
[2] CVLab, EPFL, Lausanne, Switzerland
{wei.mao,miaomiao.liu}@anu.edu.au, mathieu.salzmann@epfl.ch

1 Datasets

Below we provide more details about the datasets used in our experiments.

Human3.6M. As in [3], we use the skeleton of the subject 1 (S1) of Human3.6M as standard skeleton to compute the 3D joint coordinates from the joint angle representation. After removing the global rotation, translation and constant angles or 3D coordinates of each human pose, this leaves us with a 48 dimensional vector and a 66 dimensional vector for human pose in angle representation and 3D position, respectively. As in [3, 2, 4], the rotation angles are represented as exponential maps. During training, we set aside subject 11 (S11) as our validation set to choose the model that achieves the best performance across all future frames, and the remaining 5 subjects (S1,S6,S7,S8,S9) are used as training set.

AMASS & 3DPW. The human skeleton in AMASS and 3DPW is defined by a shape vector. In our experiment, we obtain the 3D joint positions by applying forward kinematic on the skeleton derived from the shape vector of the CMU dataset. As specified in the main paper, we evaluate the model on BMLrub and 3DPW. Each video sequence is first downsampled to 25 frames per second, and evaluate on sub-sequences of length $M + T$ that start from every 5^{th} frame of each video sequence.

2 Implementation Details

We implemented our network in Pytorch [5] and trained it using the ADAM optimizer [1]. We use a learning rate of 0.0005 with a decay at every epoch so as to make the learning rate be 0.00005 at the 50^{th} epoch. We train our model for 50 epochs with a batch size of 32 for H3.6M and 128 for AMASS. One forward and backward pass takes 32ms for H3.6M and 45ms for AMASS on an NVIDIA Titan V GPU.

Table 1. Short-term prediction of joint angles on H3.6M. We report the results on 256 sub-sequences per action.

	Walking				Eating				Smoking				Discussion			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
LTD-10-25 [3]	0.26	0.47	0.73	0.80	0.21	0.45	0.71	0.82	0.26	0.43	0.74	0.86	0.48	**0.67**	1.10	1.28
LTD-10-10 [3]	0.25	0.45	0.72	0.78	**0.20**	**0.41**	0.70	0.82	0.25	0.41	**0.71**	**0.83**	0.47	0.68	**1.09**	**1.25**
Ours	**0.24**	**0.43**	**0.66**	**0.71**	**0.20**	**0.41**	**0.68**	**0.80**	**0.25**	**0.41**	**0.71**	**0.83**	**0.44**	0.68	**1.09**	**1.25**

	Directions				Greeting				Phoning				Posing				Purchases				Sitting			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
LTD-10-25 [3]	0.20	0.41	0.76	0.92	0.52	0.84	1.24	1.41	0.34	0.57	0.96	1.09	0.31	0.60	1.06	1.24	0.47	0.84	1.24	1.33	0.33	0.52	0.92	1.06
LTD-10-10 [3]	**0.19**	0.39	0.75	0.91	0.53	0.82	1.22	1.39	0 .33	**0.54**	**0.94**	**1.07**	0.30	0.61	1.02	**1.20**	0.45	0.80	1.22	**1.32**	0.28	**0.56**	**0.94**	1.08
Ours	**0.19**	**0.38**	**0.74**	**0.90**	**0.50**	**0.79**	**1.21**	**1.38**	**0.32**	**0.54**	**0.94**	**1.07**	**0.27**	**0.57**	**1.00**	1.22	**0.43**	**0.79**	**1.21**	**1.32**	**0.27**	**0.56**	**0.94**	**1.06**

	Sitting Down				Taking Photo				Waiting				Walking Dog				Walking Together				Average			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
LTD-10-25 [3]	0.44	0.75	1.21	1.40	0.21	0.35	0.62	0.74	0.29	0.49	0.92	1.07	0.44	0.71	1.04	1.14	0.26	0.43	0.67	0.77	0.34	0.57	0.93	1.06
LTD-10-10 [3]	**0.43**	**0.74**	**1.20**	**1.38**	0.20	**0.34**	0.61	**0.72**	0.28	**0.47**	**0.90**	**1.05**	0.43	0.69	1.02	1.13	**0.24**	0.40	0.63	0.73	0.32	**0.55**	0.91	**1.04**
Ours	**0.43**	**0.74**	**1.20**	1.39	**0.19**	**0.34**	**0.60**	**0.72**	**0.27**	**0.47**	0.91	1.07	**0.42**	**0.68**	**1.01**	**1.12**	**0.24**	**0.39**	**0.62**	**0.71**	**0.31**	**0.55**	**0.90**	**1.04**

Table 2. Long-term prediction of joint angles on H3.6M.

	Walking				Eating				Smoking				Discussion			
milliseconds	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
LTD-10-25 [3]	0.92	0.97	1.03	1.05	0.99	1.16	1.26	1.33	1.07	1.26	1.41	1.55	1.48	**1.59**	**1.68**	**1.76**
LTD-10-10 [3]	0.95	1.03	1.09	1.12	**0.98**	1.15	1.28	1.36	1.04	1.21	**1.36**	1.51	**1.47**	**1.59**	1.71	1.79
Ours	**0.84**	**0.91**	**0.99**	**1.03**	**0.98**	**1.14**	**1.24**	**1.31**	**1.04**	**1.20**	1.38	**1.50**	1.49	1.62	1.72	1.82

	Directions				Greeting				Phoning				Posing				Purchases				Sitting			
milliseconds	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
LTD-10-25 [3]	1.10	1.23	1.35	**1.41**	1.63	1.81	1.95	2.01	1.29	**1.48**	**1.63**	**1.74**	**1.54**	1.81	**2.10**	**2.23**	1.51	1.66	1.80	1.87	1.34	1.60	**1.79**	**1.87**
LTD-10-10 [3]	1.09	**1.21**	**1.34**	**1.41**	1.63	1.82	1.99	2.06	1.29	1.50	1.67	1.78	1.53	1.81	2.12	2.25	1.52	1.68	1.83	1.91	1.34	1.60	**1.79**	1.89
Ours	**1.08**	1.22	1.35	1.42	**1.62**	**1.79**	**1.93**	**1.99**	**1.28**	1.49	1.65	1.76	1.55	**1.80**	**2.10**	2.24	**1.47**	**1.62**	**1.75**	**1.82**	**1.33**	**1.59**	**1.79**	1.88

	Sitting Down				Taking Photo				Waiting				Walking Dog				Walking Together				Average			
milliseconds	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
LTD-10-25 [3]	1.71	1.95	2.17	2.26	0.94	1.10	1.23	1.34	**1.30**	1.48	**1.63**	**1.74**	**1.30**	**1.45**	**1.55**	1.64	0.91	0.98	1.02	1.06	1.27	1.44	1.57	1.66
LTD-10-10 [3]	**1.68**	1.91	**2.13**	**2.22**	0.93	1.08	1.22	1.34	**1.30**	**1.47**	**1.63**	1.75	1.31	1.48	1.59	1.68	0.89	0.98	1.03	1.08	1.26	1.44	1.59	1.68
Ours	**1.68**	**1.90**	**2.12**	**2.22**	**0.92**	**1.07**	**1.21**	**1.33**	1.31	1.49	1.64	1.77	**1.30**	**1.45**	**1.55**	**1.63**	**0.86**	**0.94**	**1.00**	**1.04**	**1.25**	**1.42**	**1.56**	**1.65**

3 Additional Results on H3.6M

3.1 Results on 256 Random Sub-sequences

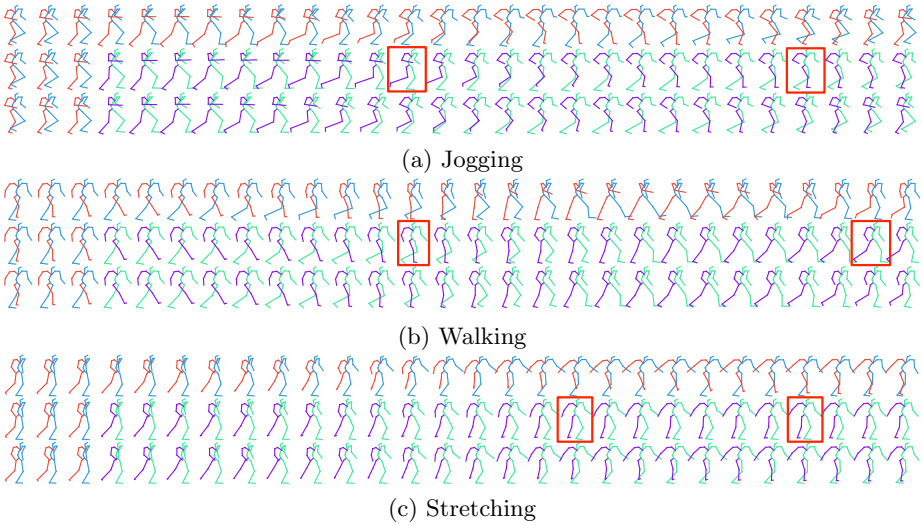
In Table 1 and 2, we report the Human3.6M results in angle representation for short-term and long-term prediction, respectively. Here, we average the error over 256 random sub-sequences per action, which was proven in [6] to be more stable than averaging over 8 random sub-sequences per action as is commonly done. Our conclusions remain unchanged: our approach achieves the state-of-the-art performance for both short-term and long-term prediction on average.

3.2 Generating Long Future for Periodical Motions

For periodical motions, such as "Walking", our approach can generate very long futures (up to 16 seconds). As shown in the supplementary video, such future predictions are hard to distinguish from the ground truth even for humans.

4 Additional Results on AMASS

In Fig. 1, we compare the results of LTD [3] and of our approach on the BMLrub dataset. Our results better match the ground truth.



(a) Jogging

(b) Walking

(c) Stretching

Fig. 1. Qualitative comparison on the BMLrub dataset. From top to bottom, we show the ground-truth motion, the prediction results of LTD [3] and of our approach on 3D position. The observed poses are shown as blue and red skeletons and the predictions in green and purple. As highlighted by the red boxes, our predictions better match the ground truth, in particular for the legs.

5 Motion Attention vs. Frame-wise Attention

To further investigate the influence of *motion* attention, where the attention on the history sub-sequences $\{\mathbf{X}_{i:i+M+T-1}\}_{i=1}^{N-M-T+1}$ is a function of the first M poses of every sub-sequence $\{\mathbf{X}_{i:i+M-1}\}_{i=1}^{N-M-T+1}$ (keys) and the last observed M poses $\mathbf{X}_{N-M+1:N}$ (query), we replace the keys and query with the last frame of each sub-sequence. That is, we use $\{\mathbf{X}_{i+M-1}\}_{i=1}^{N-M-T+1}$ as keys and $\mathbf{X}_N$ as query. We refer to the resulting method as *Frame-wise Attention*. As shown in Table 3, motion attention outperforms frame-wise attention by a large margin. As discussed in the main paper, this is due to frame-wise attention not considering the direction of the motion, leading to ambiguities.

Table 3. Comparison of frame-wise attention and with our motion attention.

milliseconds	80	160	320	400	560	720	880	1000
Frame-wise Attention	24.0	44.5	76.1	88.3	107.5	121.7	131.7	136.7
Motion Attention	**10.8**	**23.9**	**49.4**	**60.7**	**77.3**	**92.0**	**104.4**	**112.4**

References

1. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. ICLR (2015)
2. Li, C., Zhang, Z., Lee, W.S., Lee, G.H.: Convolutional sequence to sequence model for human dynamics. In: CVPR. pp. 5226–5234 (2018)
3. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: ICCV. pp. 9489–9497 (2019)
4. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: CVPR (July 2017)
5. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS-W (2017)
6. Pavllo, D., Feichtenhofer, C., Auli, M., Grangier, D.: Modeling human motion with quaternion-based neural networks. IJCV pp. 1–18 (2019)