Multi-Person 3D Motion Prediction with Multi-Range Transformers

Jiashun Wang[1] Huazhe Xu[2] Medhini Narasimhan[1] Xiaolong Wang[1]
[1]UC San Diego [2]UC Berkeley

Background:

3D human motion prediction has drawn great attention these years

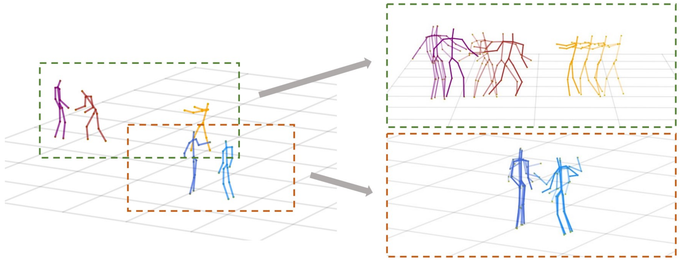
- Predicting future motion sequences given a sequence of history
- Focusing on single person motion
- Usually neglecting the movement of the root joint

Multi-person motion prediction is relatively under-explored and more challenging

- Considering multi-person interaction
- Modeling pose and trajectory jointly is needed, e.g. catching

Task:

Given a scene with N persons and their corresponding history motion, we aim to predict their future 3D motion.

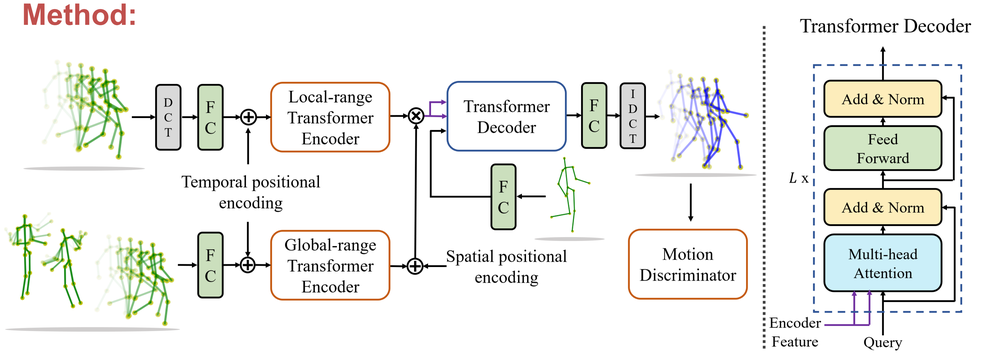


Representation:

Specifically, given $X_{1:k}^n = [x_1^n, \dots, x_k^n]$ representing the history motion of person n where n=1,…,N and k is the time step. We aim to predict the future motion $X_{k+1:T}^n$.

We use a vector $x_k^n \in R^{3J}$ containing the Cartesian coordinates of the J skeleton joints to represent the pose of the person n at time k.

Method:



Individual input motion is sent to the Local-range Transformer Encoder and all the person's motions are sent to the Global-range Transformer Encoder.

Spatial Positional Encoding (SPE)

- SPE encodes the spatial distance between the query token x_k^n and the tokens of every time step of each person $x_{1:k}^{1:N}$
- Helpful especially in a scene with a crowd of persons

$$\mathrm{SPE}(x_t^n, x_k) = \exp(-\frac{1}{3J}||x_t^n - x_k||_2^2)$$

Why local and global?

Local:

- The task of synthesizing a natural motion based on previous states itself is actually a challenging task
- To ensure the smoothness of the motion, the model requires dense sampling of the input sequence

Global:

- The interaction of all the persons in the whole scene, sparse sampling of the sequences are used
- Compute the global feature once

The separate processes of different aspects of data largely reduces the optimization difficulty.

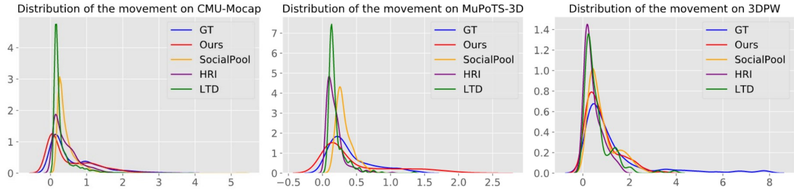
Experiment:

method	CMU-Mocap (3 persons)			MuPoTS-3D (2 ∼ 3 persons)			3DPW (2 persons)			Mix1 (9 ∼ 15 persons)			Mix2 (11 persons)		
	1 s	2s	3s	1 s	2s	3s	1 s	2s	3s	1 s	2s	3s	1 s	2s	3s
LTD [35]	1.37	2.19	3.26	1.19	1.81	2.34	4.67	7.10	8.71	2.10	3.19	4.15	1.72	2.58	3.45
HRI [34]	1.49	2.60	3.07	0.94	1.68	2.29	4.07	6.32	8.01	1.80	3.14	4.21	1.60	2.71	3.67
SocialPool [2]	1.15	2.71	3.90	0.92	1.67	2.51	4.17	7.17	9.27	1.85	3.39	4.84	1.72	3.06	4.26
Ours w/o Global	0.99	1.71	2.50	0.92	1.67	2.50	4.17	6.85	8.91	1.77	3.10	4.19	1.42	2.29	3.06
Ours w/o $\mathcal{D}$	1.13	1.84	2.57	0.92	1.62	2.26	4.17	6.41	8.09	1.75	3.00	4.00	1.34	2.19	2.95
Ours w/o SPE	1.05	1.68	2.37	0.92	**1.51**	2.23	3.92	6.18	**7.79**	1.75	3.09	4.13	1.31	2.15	2.92
Ours	**0.96**	**1.57**	**2.18**	**0.89**	1.59	**2.22**	**3.87**	**6.12**	7.83	**1.73**	**2.99**	**3.97**	**1.29**	**2.09**	**2.82**

In both cases with a small number and a large number of people, our method achieves state-of-the-art performance for different prediction time lengths.

method	CMU-Mocap	MuPoTS-3D	3DPW	Mix1	Mix2
LTD [44]	3.61±0.83	3.66±**0.93**	3.65±0.76	3.71±0.93	3.75±0.90
HRI [43]	3.36±0.96	3.59±1.25	3.76± **0.72**	3.67±0.89	3.71±0.90
SocialPool [2]	3.49±0.87	3.66±1.19	3.66±0.86	3.62±0.92	3.49±1.02
Ours	**3.62±0.78**	**3.68**±0.98	**3.78**±0.82	**3.74±0.83**	**3.77±0.82**
GT	3.78±0.76	3.85±0.96	3.77±0.81	3.77±0.87	3.88±0.79

We report the average and the standard error of the score. A higher average score means users think that the results are more natural looking. Our results get better reviews consistently cross all datasets.



It shows that other methods intend to predict a motion with less movement while our result is very similar to the ground truth.