Social Diffusion: Long-term Multiple Human Motion Anticipation Supplementary material

Julian Tanke*[1], Linguang Zhang[2], Amy Zhao[2], Chengcheng Tang[2], Yujun Cai[2], Lezi Wang[2], Po-Chen Wu[2], Juergen Gall[1,3], and Cem Keskin[2]

[1]University of Bonn
[2]Reality Labs Research
[3]Lamarr Institute for Machine Learning and Artificial Intelligence
{tanke,gall}@iai.uni-bonn.de
{linguang,xamyzhao,chengcheng.tang,yujuncai,wanglezi,pochenwu,cemkeskin}@meta.com

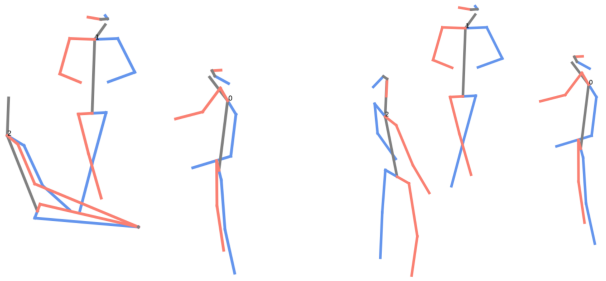

Figure 1: Example frame from sequence **170224_haggling_a2** of the Haggling dataset [3]. Left: Original poses from [3]. Right: Poses after our cleaning step.

1. Cleaning the Haggling Dataset

Since the 3D human poses in the Haggling dataset [3] have been estimated [2, 4], the original dataset contains many artifacts. We thus cleaned the 3D human poses by marking errors and interpolation. We removed a few sequences which we were not able to correct. An example of our manual pose correction is shown in Figure 1. For a video, we refer to https://github.com/jutanke/social_diffusion.

2. Symbolic Social Cues

As mentioned in Section 5.1 of the paper, we define the following states:

- ø: nobody talks
- **L**: the left seller is talking
- **R**: the right seller is talking
- **B**: the buyer is talking

as well as the following attention states:

- $\hat{\text{L}}$: the left seller has the buyers attention
- $\hat{\text{R}}$: the right seller has the buyers attention

This way we get 16 interaction states:

1. ø $\hat{\text{L}}$ (see social_states/label01.mp4)
2. ø, $\hat{\text{R}}$
3. **L**, $\hat{\text{L}}$ (see social_states/label03.mp4)
4. **L**, $\hat{\text{R}}$ (see social_states/label04.mp4)
5. **R**, $\hat{\text{L}}$
6. **R**, $\hat{\text{R}}$
7. **LR**, $\hat{\text{L}}$ (see social_states/label07.mp4)
8. **LR**, $\hat{\text{R}}$
9. **B**, $\hat{\text{L}}$ (see social_states/label09.mp4)
10. **B**, $\hat{\text{R}}$
11. **BL**, $\hat{\text{L}}$ (see social_states/label11.mp4)
12. **BL**, $\hat{\text{R}}$
13. **BR**, $\hat{\text{L}}$
14. **BR**, $\hat{\text{R}}$ (see social_states/label14.mp4)
15. **BLR**, $\hat{\text{L}}$
16. **BLR**, $\hat{\text{R}}$ (see social_states/label16.mp4)

3. Interaction with Objects

Our method can not only be applied to multiple persons, but also to persons interacting with an object. Figure 2

*Work done partially while Julian was at Reality Labs Research.

Figure 2: Forecast motion on the *Sports* sequence from [2, 4], consisting of two persons throwing a ball. The ball and its past trajectory are marked in red and each column represents 1 second of motion. On the right hand side, we show the global view, which shows that the distance between the persons is correctly maintained.

Figure 3: Various degrees of interactions in 3DPW [6].

shows an example from the Sports dataset from Panoptic Studio [2, 4].

4. Other Datasets

3DPW [6]: The 3DPW dataset contains recordings of 1 to 2 persons in various settings. In total (train, test, validation), it has 21.967 frames (about 6 minutes) that contain two persons, split into 27 sequences. The sequences cover different levels of interactions like one person watching another one performing a task, two persons walking alongside, or dancing as shown in Figure 3. Compared to the revised Haggling dataset with over 300.000 frames, the dataset is very small, it is limited to only 2 persons, and the 3D poses contain very strong artifacts like sliding as mentioned in [7].

MuPoTS-3D [5]: MuPoTS-3D contains recordings of 2 to 3 persons in workout settings. The dataset is even smaller than 3DPW and contains only 4.4 minutes of multi-person motion recordings. Most of the time, the persons perform their workout and there are no interactions between the persons, see Figure 4.

CMU [7]: The dataset consists of mixed sequences of the CMU-Mocap dataset [1]. Each sequence contains 3 persons but the motions are sampled from 2 different sequences. Most of the interactions are thus unrealistic.

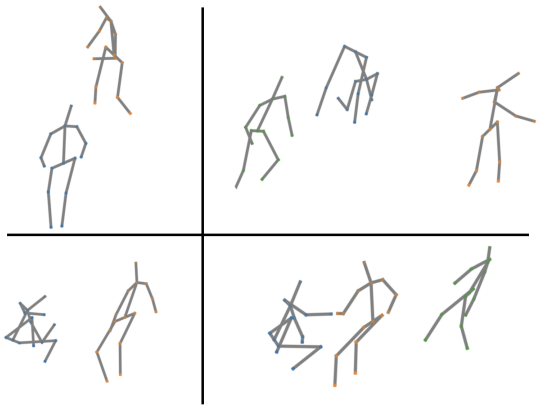

Figure 4: Various sequences of 2-3 persons in the MuPoTS-3D [5] dataset. The dataset contains sports sequences where persons perform their workout and interact only seldom. The bottom right shows a rare example of interaction.

2 seconds			10 seconds		
gt	ours	MRT [7]	gt	ours	MRT [7]
0.948	0.567	0.520	0.953	0.582	0.01

Table 1: User study on the Haggling dataset.

5. User Study

We provide a user study in Table 1 were 11 subjects were asked to judge short (2 seconds) or long (10 seconds) sequences of forecast human motion of 24 randomly selected sequences. We showed randomly the ground-truth (gt), results from our approach, or from [7]. The subjects rated the sequences by realistic (1), unsure (0.5), or unrealistic (0). We report the mean rating (higher is better). The results are consistent with the results in the paper. MRT [7] performs well for short time horizons but accumulates artifacts over time, making it look unrealistic. Since 10s are slightly better to judge for humans than 2s, the rating for ground-truth

steps	1	10	100	500	800	1000
NDMS	0.103	0.170	0.176	0.207	0.210	0.230

Table 2: Impact of the number of diffusion steps on the Haggling dataset.

and ours is slightly higher for 10s.

5.1. Impact of Number of Diffusion Steps

We evaluate the impact of the number of diffusion steps in Table 2. We follow the original diffusion implementation and train our method with 1000 diffusion steps. We observe that the motion quality saturates after 500 diffusion steps.

References

[1] CMU. Carnegie-Mellon Mocap Database. 2

[2] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In *International Conference on Computer Vision*, 2015. 1, 2

[3] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1

[4] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1, 2

[5] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *International Conference on 3D Vision*, 2018. 2

[6] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European conference on computer vision*, 2018. 2

[7] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. *Advances in Neural Information Processing Systems*, 2021. 2