Supplementary Material for Spatio-temporal Tendency Reasoning for Human Body Pose and Shape Estimation from Videos

Boyang Zhang
boyangchang@foxmail.com
Suping Wu*
pswuu@nxu.edu.cn
Hu Cao
caohu19980219@gmail.com
Kehua Ma
919056348@qq.com
Pan Li
18235448104@163.com
Lei Lin
2409937088@qq.com

School of information Engineering
Ningxia University
Ningxia, CHN

In the supplementary material, we present more experiments.

1 More Experiments

To verify the performance of our module in crowded scenarios, we also conduct quantitative comparisons with TCMR on 3DPW-CROWD. The sequence names of 3DPW-CROWD are courtyard_hug_00 and courtyard_dancing_00, a subset of the 3DPW validation set. Then we calculate and average the reconstruction accuracy of these two videos. Table 1 shows our performance on 3DPW-CROWD. The reconstruction accuracy of our method is much higher than TCMR, which also demonstrates the effectiveness of our method. Our method can focus on more distant frames, thereby complementing the human body that is not visible in the current frame.

Model	***MPJPE*↓**	***MPVPE*↓**	***PA-MPJPE*↓**
TCMR	97.3	125.1	62.3
Ours	**89.9**	**113.5**	**60.8**

Table 1: Quantitative comparison of our method under crowded scenes on 3DPW. Our method is able to reason about spatio-temporal trends in crowded scenes and obtain more human information.

* Corresponding Author: Suping Wu (*pswuu@nxu.edu.cn*)
© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

Model	*PA-MPJPE*↓	*Accel*↓
STR w/ GRU	62.7	9.0
STR w/ TTR	**61.6**	**8.4**

Table 2: Ablation experiments of the TTR module on the MPI-INF-3DHP dataset.

Model	*PA-MPJPE*↓	*Accel*↓
STR w/ SENet	62.5	8.6
STR w/ STE	**61.6**	**8.4**
STR w/ Common Aggregation	62.5	8.6
STR w/ Integration Strategies	**61.6**	**8.4**

Table 3: Ablation experiments of the STE module and Integration Strategies on the MPI-INF-3DHP dataset.

More Ablation Study. We do ablation experiments inside each module. First, as shown in Table 2, we use the original GRU to replace the TTR module, and we find that both the reconstruction error and the acceleration error increase. The original GRU is difficult to reason about the information of distant frames and the spatio-temporal tendency of the whole sequence. On the other hand, TTR performs tendency reasoning on each sub-segment by dividing the longer time series.

As shown in Table 3, we performed ablation studies on STE. We replace SENet with the STE module. Experiments show that compared with SENet, STE can better capture human motion features to enhance spatial tendency. SENet uses its own global features to calibrate different channels, while our STE is a spatial tendency enhancement module that captures motion-related features. We also investigate experiments on alternative integration strategies for common feature aggregation in Table 3. The integration strategy considers self-integration on the basis of ordinary aggregation. The aim is to introduce an integration of the temporal dimension, thereby enhancing human features that weaken over time.

Figure 1: The limitation case of our model. Human truncation and large occlusion can cause our method to fail to reason about the correct spatio-temporal tendency.

2 Limitations

As shown in Figure 1, we show the limitations of our approach. Due to continuous people entering the video screen, the phenomenon of human body truncation and large occlusion will inevitably occur in multi-frame images. This phenomenon results in the network not being able to capture all the features of the target and therefore unable to reason about the correct motion of the target.