Supplementary: Embodied Human Activity Recognition

Sha Hu Yu Gong Greg Mori
Simon Fraser University
8888 University Drive Burnaby, B.C. Canada. V5A 1S6
hushah@sfu.ca, gongyug@sfu.ca, mori@cs.sfu.ca

Additional information presented in this supplementary:

1. Details on the training hyperparameters and hidden layers of policy network (Sec. 1).
2. Details on the processing datasets (Sec. 2).

1. Hyperparameters and architectures

Table. 1 shows the hyperparameters for Imitation Learning phase, PPO training phase, and hidden layers of policy network. The code will be released upon acceptance.

2. Data Processing

As introduced in the main paper Sec.5.1, the episode dataset is constructed as combinations of human scenarios and agent starting positions. In this section, we provide additional details on processes and statistics in obtaining diverse dynamic human scenarios from two existing real motion capture datasets. The instructions to download the original datasets and implementations of the data process will be provided under the code.

ExPI [1]. ExPI has three types of data splits, and we follow the split setting named Common Action split. In this split, there are 7 classes of common activities performed by different couples of dancers. Each category has 10 mocap sequences, split in 5/5 for training/test set. Training and test sets share the same categories of activities but are performed by different people. Each mocap sequence is labeled with one activity category, and each sequence is at 25FPS. These dancers are instructed to perform professional dancing actions in a controllable lab studio instead of acting spontaneously. Thus, we assume that the annotated activity label for each sequence can represent the activity category of each frame from the start to the end of that sequence. To obtain episodes of equal temporal horizon, we take sub-sequences of length L from each sequence by a sliding window of stride s. We set $L = 60$. $s = 1/60$ for train/test, respectively. This results in 6656/135 human activity scenarios for train/test sets. We further split the training set into train/validation sets in a ratio of 0.85:0.15. Finally, we have $|\mathcal{H}_{train}| = 5658$, $|\mathcal{H}_{val}| = 998$, and $|\mathcal{H}_{test}| = 135$. The training set H_{train} is divided into two subsets: one for activity recognition network training and the other for policy training, with the allocation ratio being 0.4:0.6.

IL Training Hyperparameters	
learning rate	0.0001
batch size	32
PPO Training Hyperparameters	
discount factor	0.99
generalized advantage estimation parameter	0.98
ppo clip	0.2
rollout length	60
number of environments	32
ppo epochs	4
number of mini-batches per epoch	4
linear learning rate schedule for actor network	(initial lr=0, end lr=0.0001, start step=3200, end step=6400)
linear learning rate schedule for value network	(initial lr=0.001, end lr=0.0001, start step=3200, end step=6400)
learning rate for f^A, f^V, f^P, f^S	the lower of the two current learning rates of actor network and value network
actor loss coefficient	1
value loss coefficient	0.5
entropy regularization coefficient	0.001
max gradient norm	0.5
Policy Network	
accumulated recognition state encoder f^A hidden layers	[32, 32]
visual encoder f^V hidden layers	[64, 32]
position encoder f^P hidden layers	[32]
state encoder f^S hidden layers	[64, 32]
actor network hidden layers	[16, 7]
value network hidden layers	[16, 1]

Table 1. Hyperparameters for IL training, hyperparameters for PPO training, and hidden layers of policy network.

AIST++ [2]. AIST++ reconstructs 3D motion from the AIST [3] dataset. We use a subset of motion sequences

named Advanced Choreography, created by AIST's video classification benchmark. In this subset, there are 10 classes of common activities. Each category has 21 mocap sequences performed by 3 actors, each of whom performs 7 sequences. In total, the dataset covers 210 mocap sequences by 30 different actors. We follow the split from AIST. For each category, 2 actors are randomly selected for the training set, i.e., $2 \times 7 = 14$ sequences for each category. The remaining dancer is used for the test set, i.e., 7 sequences for each category. Each mocap sequence is at 60FPS. Again, we assume that the annotated activity label for each sequence represents the activity category of each frame of that sequence. Each sequence is first subsampled at intervals of 16 frames to construct episodes of extended elapsed times. Then, we take sub-sequences of length $L = 60$ from each sequence by a sliding window of stride $s = 1/60$ for train/test sets. This results in 10912/113 human activity scenarios for train/test sets. We further split the training set into train/validation sets in a ratio of 0.85:0.15. Finally, we have $|\mathcal{H}_{train}| = 9276, |\mathcal{H}_{val}| = 1636$, and $|\mathcal{H}_{test}| = 113$. Similar to ExPI, the training set H_{train} is divided into two subsets: one for activity recognition network training and the other for policy training, with the allocation ratio being 0.4:0.6.

References

[1] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13053–13064, 2022. 1

[2] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[3] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 501–510, Delft, Netherlands, Nov. 2019. 1