Generative Adversarial Graph Convolutional Networks for Human Action Synthesis

Bruno Degardin[1,4,5], João Neves[2,4], Vasco Lopes[2,4,5], João Brito[5], Ehsan Yaghoubi[3], Hugo Proença[1,4]
[1]IT - Instituto de Telecomunicações, [2]NOVA LINCS, [3]C4-Cloud Computing Competence Center
[4]Universidade da Beira Interior, Portugal [5]DeepNeuronic
bruno.degardin@ubi.pt

A. Hyperparameters and training configurations

Datasets settings. For global movement experiments in Section 4.2 , which included NTU RGB+D [6] and NTU-120 RGB+D [5] datasets, the temporal length of the skeleton sequences was normalized to $t = 64$ frames. The reason behind the chosen temporal length resides in the action execution average of the dataset (64 frames). Despite both datasets containing some annotation errors (some inaccurate 3D joints position), no sample filtering was applied. We confirm the superiority of our method in action conditioning by using every action class in both datasets (60 for NTU RGB+D and 120 for NTU-120 RGB+D). For local movement experiments in Section 4.3 , which included Human3.6M [3] and NTU-2D RGB+D [6] datasets, the same settings were applied as previous approaches [2, 7, 8]. Specifically, the temporal length was normalized to $t = 50$ frames, the number of action classes used are 10, and both datasets were normalized from real/global movement to local movement, which facilitates the generation process. In Human3.6M [3] dataset, the following action classes are used: *sitting, sitting down, discussion, walking, greeting, direction, phoning, eating, smoking and posing*. In NTU-2D RGB+D [6] dataset, the following action classes are used: *drinking water, jump up, make phone call, hand waving, standing up, wear jacket, sitting down, throw, cross hand in front and kicking something*. Also, for a fair comparison, training samples from NTU-2D RGB+D [6] were carefully selected from each class on NTU RGB+D [6] similar to previous methods [2, 7, 8].

Training configurations. We train the networks using Adam [4] optimizer with $\alpha = 2 \times 10^{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$ for all datasets with a minibatch size of 32. Since we rely on the WGAN-GP loss [1], we set $n_{critic} = 5$, which sets the number of iterations of the discriminator per generator iteration.

Upsampling and downsampling details. As illustrated in Fig. 3 (paper), the spatial resolution of the skeleton is increased from the intermediate latent point as $1 \rightarrow 5 \rightarrow 11 \rightarrow 25$ joints for the NTU RGB+D [6], NTU-2D RGB+D [6] and NTU-120 RGB+D [5] datasets. For the Human3.6M [3] dataset the spatial resolution is increased as $1 \rightarrow 2 \rightarrow 7 \rightarrow 15$ joints. In all datasets, the temporal resolution is increased by doubling $t/16$ until reaching the dataset's temporal length t. The same resolutions reversed are applied for the downsampling paths in the discriminator.

Mapping network structure. Our non-linear mapping network comprises fully connected layers with 512 as the dimensionality of the input and output activations. As demonstrated in Table 2 , the increasing number of different subjects in the training data results in a more complex latent representation requiring a deeper mapping network. For this reason, we set 6 layers for the Human3.6M [3], and 8 layers for the NTU-120 RGB+D [5] dataset. NTU RGB+D and NTU-2D RGB+D [6] datasets follow the same settings as studied in Table 2 .

Noise injection details. The noise injector described in Section 3.4.1 samples a random noise $\boldsymbol{r}_l$ using $\mathcal{N}(0, 1)$. Each joint at resolution level l has a respective weight to each channel and receives a different noise added channel-wise. This operation is applied to every generator's layer.

B. Action complexity

We include several action samples synthesised by our graph convolutional generator that demonstrate various aspects related to action complexity (see also accompanying video). Apart from the ability to generate up to 120 different action classes, we are able to generate global (real) body movement in 3D space, which, to the best of our knowledge, such complex actions under global movement settings had proven to be uncharted territory for previous methods. Figure 1 shows different action examples illustrating the detail and expressiveness achievable using our method in NTU RGB+D [6]. In Figure 2, we demonstrate the ability to generate desired actions among 120 different classes from NTU-120 RGB+D [5].

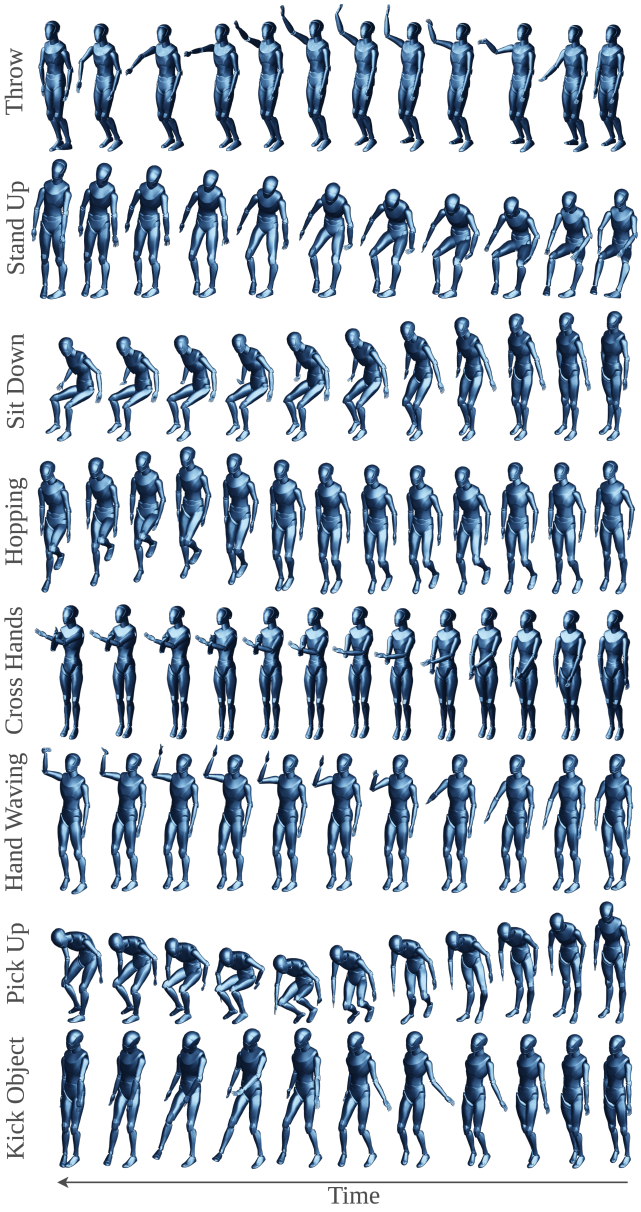

Figure 1: **Synthetic set of actions** generated by our graph convolutional generator trained on NTU RGB+D [6].

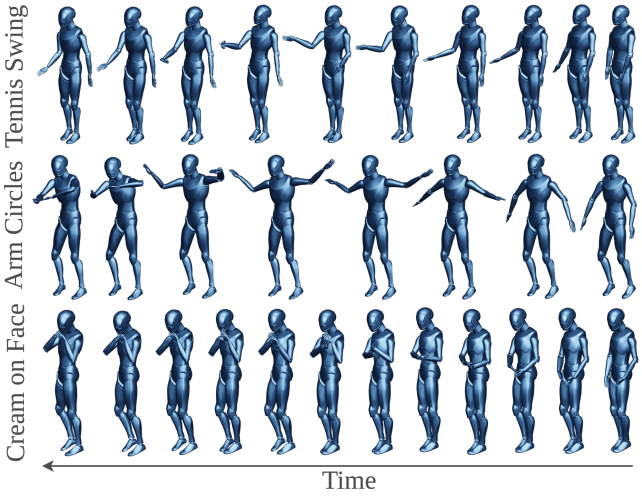

Figure 2: **Synthetic set of actions** generated by our graph convolutional generator trained on NTU-120 RGB+D [5].

References

[1] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

[2] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. A recurrent variational autoencoder for human motion synthesis. In *28th British Machine Vision Conference*, 2017.

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[5] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019.

[6] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[7] Zhenyi Wang, Ping Yu, Yang Zhao, Ruiyi Zhang, Yufan Zhou, Junsong Yuan, and Changyou Chen. Learning diverse stochastic human-action generators by learning smooth latent transitions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12281–12288, 2020.

[8] Ping Yu, Yang Zhao, Chunyuan Li, Junsong Yuan, and Changyou Chen. Structure-aware human-action generation. In *European Conference on Computer Vision*, pages 18–34. Springer, 2020.