Action-Conditionned 3D Human Pose Motion Generation

Victoria Brami
Master MVA
Ecole Normale Supérieure Paris-Saclay
victoria.brami@eleves.enpc.fr

Abstract

We study a robust method to generate realistic 3D-human motions which relies on the training of a Variational Auto-Encoder. We intent to improve the proposed model through the use of PARE, another pose inference model. We show that exploiting PARE renders better synthetic motions on NTU RGB dataset. We then test the model on more challenging motions like gymnastic floor and beam exercises on a hand-made dataset built from FineGym.

1. Introduction

Being able to generate an infinite number of actions motions represent an important breakthrough for tackling lack of data correlated issues, but also for gaining variety in motions. A promising approach is proposed by ACTOR, a VAE-transformer model capable to generate wide-range different motions. ACTOR learns on motions estimated on videos by a previous model VIBE. However, a more recent motion estimator model called PARE tends to make better motion estimations than VIBE. Therefore using PARE inference in a first step may allow ACTOR to be trained on more realistic estimated motions. We investigate this question throughout this project. The main objectives covered in this project are:

- Replicate the results of ACTOR article on NTU13 dataset, using VIBE estimator in the first step and ensure the goodness of the quantitative results.
- Replace VIBE by PARE in the first step then retrain ACTOR on the newly estimated motions and make a comparative analysis of the qualitative results.
- Test motions' generation on faster movements like gymnastic moves.

In Section 2 we describe briefly the pipeline used for motion generation. We present the results of the experiments carried out on motion synthesis from NTU13 dataset with VIBE and PARE motions estimators in Section 3. In Section 4 and Section 5 we show the motions generated by ACTOR when trained on FineGym dataset before concluding with a critical analysis.

2. Action Conditioned motion Generation

2.1. Principles

Motions generation pipeline consists of two steps. Given some videos showing an action a, we first infer a 3D-SMPL pose thanks to a first trained Network. In our work, we consider two different model to perform the inference: VIBE [2] used in ACTOR related paper, then PARE [3] a more recent model. PARE outperforms VIBE in motion estimation: it demonstrated it handles better occlusions as well as fast motions in videos. This first model outputs a sequence of 3D SMPL poses [4], a disentangled body representation used in input for the motion generative model.

2.2. ACTOR Generative Model Overview

ACTOR model [5] is a VAE-transformer based model allowing the synthesis of any desired motion. The model is split into two parts as shown in 3:

- The **encoder**. Given a sequence of T successive 3D SMPL poses and the action label a, the encoder represents the movement in a latent space, by forwarding the inputs in a transformer-based architecture network. It outputs distribution parameters (μ, Σ).
- The **decoder**: it takes in input a vector z sampled from the former latent space, an action label a and a sequence length T, and restore a complete 3D motion through another transformer-based network.

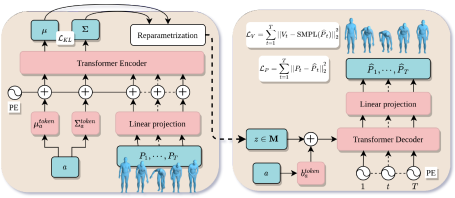

Figure 1: ACTOR Model Architecture

Training. Several sources of loss are used for the training in order to render the most realistic motions possible:

- A L2-loss $\mathcal{L}_P$ between ground truth $\{P_t\}$ and predicted pose parameters $\{P_t\}$.
- A L2-loss $\mathcal{L}_V$ between ground truth $\{V_t\}$ and predicted vertex parameters $\{V_t\}$.
- A Kl divergence loss $\mathcal{L}_{KL}$to enforce the encoder to ouputs parameters similar to a Gaussian distribution. We weighted the $\mathcal{KL}$-loss with a coefficient $\lambda_{\mathcal{KL}}$. We set this value to 10^{-5}. Greater values of $\lambda_{\mathcal{KL}}$ did not return satisfying results.

We remarked the loss values kept decreasing during training phases, however we restricted the number of epochs to 1800-2000 on NTU RGB+D dataset and 30k on FineGym2 dataset.

3. Experiments

3.1. Dataset and metrics

NTU RGB+D dataset [1] As described in ACTOR article [5], we choose to test motion generation on NTU RGB+D subset, which consists of 3902 videos of 13 different common human motions: *Foot Kicking*, *Sitting*, *Standing up*, etc.

Evaluation metrics. With VIBE motions estimations we compute the FID score a characteristic metric for generative model and the motion classification accuracy thanks to a pre-trained action recognition model on VIBE features.

3.2. Motion generation using VIBE

Training. We train the generative model on NTU13 dataset on fixed size 60-frames input. Due to GPU time limit issues, we restrain the training for 1800 epochs. We fix the weight for the KL loss at 10^{-5}. The quantitative results of VIBE training are shown in Table 1.

NTU13 dataset				
Model	FID$_{tr}$↓	Acc.↑	Div.→	Multim.→
Paper				
GT	$0.02^{\pm 0.00}$	$99.8^{\pm 0.01}$	$7.09^{\pm 0.03}$	$2.27^{\pm 0.01}$
Gen	$0.14^{\pm 0.00}$	$97.4^{\pm 0.14}$	$7.09^{\pm 0.04}$	$2.05^{\pm 0.01}$
Ours				
GT	$0.02^{\pm 0.00}$	$98.8^{\pm 0.04}$	$7.04^{\pm 0.02}$	$2.38^{\pm 0.01}$
Gen	$0.15^{\pm 0.00}$	$96.7^{\pm 0.31}$	$7.13^{\pm 0.02}$	$2.12^{\pm 0.01}$

Table 1: Comparison with the training done in ACTOR [5] paper: the hyperparameters used are the same although our model is trained for 1800 epochs (2000 in the paper).

3.3. Comparison with PARE

An axis for a model improvement consists of replacing the SMPL body inference model VIBE to be able to get even more realistic generated motions. We use PARE model [3], an acronym for Part Attention Regressor for 3D Human body Estimation. We apply PARE model on all NTU13 videos to get the corresponding SMPL poses.

Training. We re-train ACTOR on PARE estimations. We consider the same hyperparameters as used in VIBE for the training: we use AdamW optimizer with a learning rate set to 0.0001. Again, we take a mini-batch size at 20. ACTOR is re-trained for 2000 epochs with sequences of 60 successive poses in input.

Evaluation We are restricted to a qualitative motion analysis with PARE as we do not detain a pre-trained action-recognition model. With ACTOR/PARE combined model, we obtain more finished movements than with VIBE, especially on movements using the bottom part of the body: an illustration of the observation is displayed on figure 3. The sequences differ also from their point of view, or the arm/leg used (see figure 2). Using PARE allows the generation of more fine-grained motions than using VIBE. We provide other generated samples in SE

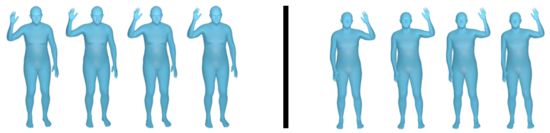

Figure 2: An example of diversity in the generated movement. Either left or right arm is shaken (5-frames space between each frame)

4. Experiments on New Action Motions

4.1. Choice of the dataset

ACTOR generator model combined with PARE motion estimator seems to achieve more realistic and completed moves generation. A further step consists in testing more challenging motions generation. We briefly describe the dataset below.

FineGym[6] is a dataset built on top of gymnastic videos. It provides annotations for fine-grained human action recognition on gymnastics. More than 95% of the videos are high resolution (either 720p or 1080p) as they come from top-level competitions. The train set contains 26320 movements labelled in 99 different categories.

4.2. FineGym2 subset construction

We chose 2 common gymnastic moves on the balanced beam: *Split Jump* and *Back Handspring*. We extract various

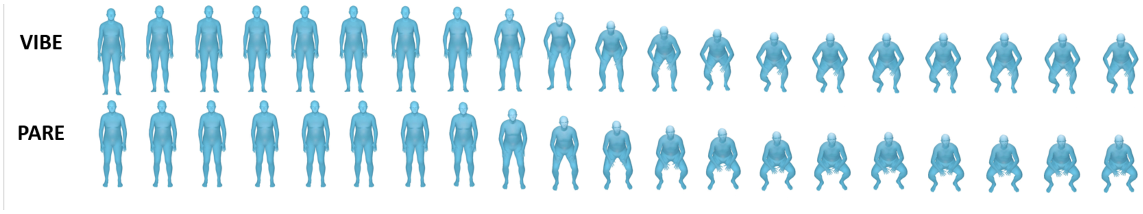

Figure 3: Comparison of the motions synthesized with VIBE (above) and PARE (below): Knee bending better rendered using PARE. The motion seems more complete. This is observed on other action labels like standing up. We take a move every 3 frames.

sequences manually for each label as we have to assert that:

1. The whole movement is contained in the video and it is the unique one. For example back-handspring move can come just after a round-off, and we must only keep the second part of the move.
2. The Gymnast is alone on the video to simplify preprocessing.
3. The Camera point of view remains the same during all the sequence.
4. The video's resolution has a good quality.

We extract the moves with respect to these constraints from the 743 and 1121 available sequences. The sequences are slightly shorter, they contain at least 55 frames. We provide the code related to FineGym dataset preprocessing. here[1].

Label	Number of samples
Split Jump	117
Back Handspring	147

Table 2: FineGym2 dataset: 2 labels are chosen for motion generation and we extract manually each of the video from different gym competitions.

4.3. Training

As FineGym2 dataset is much smaller than NTU13 dataset, we choose to train during 30000 epochs on PARE estimations. We train ACTOR on video sequences of 50 frames. The other hyperparamters (optimizer, learning rate, etc.) remain the same as those chosen for NTU13.

4.4. Qualitative Results

An batch of synthesized motions is shown in figure 5. For "Split-Jump" moves, ACTOR manages to generate movements with varying points of view. Some of them look realistic, while other are physically impossible to do.

We also frequently observe some noise: the movements are less smooth than those from NTU13. This may be due to the non-perfect stability of the cameras in FineGym2 videos.

ACTOR has much more difficulties with the second label, especially on the part of the movement where the back is stretched. The preparation phase seems realistic.

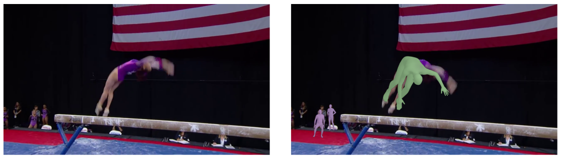

Figure 4: Motion wrongly estimated by PARE

The inaccurate synthesis of the "Back Handspring" movement is mainly due to the movement speed. The resulting frames during this phase are sometimes blurred: this makes PARE estimations erroneous. Some frames lack of estimations. Besides, PARE fails to infer postures where the back is too stretched as illustrated on Figure 4.

5. Conclusions

We presented in this project ACTOR, a transformer-based VAE model capable to synthesize a wide range of motions. When retraining ACTOR on VIBE pose estimations, we almost retrieved the same scores as announced in the paper. Besides, we demonstrated qualitatively that replacing VIBE by PARE in motion estimation allowed to obtain even better synthetic motions on NTU dataset: PARE makes better inferences on movements involving articulations bending like *Sitting down, Standing up*. When using ACTOR on gymnastics movements synthesis, it manage to generate a *Split Jump*. However it fails to synthesize correct movements for *Backhandspring*. This failure mainly comes from difficulties from PARE and VIBE to infere a correct pose on back streching movements.

References

[1] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Ac-

[1] https://github.com/Victoria-brami/action_conditioned_motion_generator.git

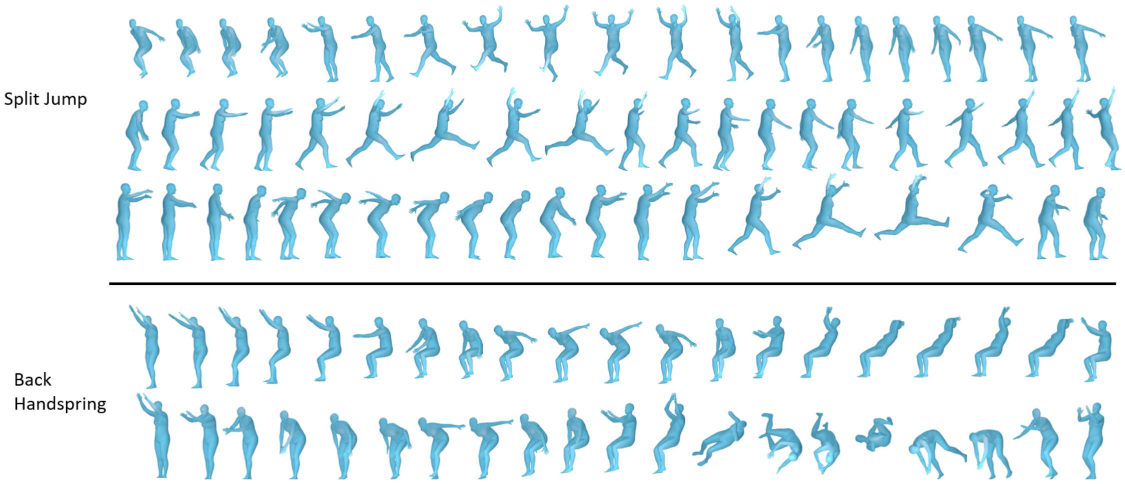

Figure 5: Motion generated with PARE: We take a move every 3 frames. We obtain different points of view for the same motion. The move on the 3^{rd} line is well generated while the 2^{nd} one is not physically possible: the man splits her legs while jumping, then re-splits again. On the 4^{th} row the woman is stuck in the preparation motion before the back-handspring

tion2motion: Conditioned generation of 3d human motions. *CoRR*, abs/2007.15240, 2020.

[2] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[3] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11127–11137, October 2021.

[4] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[5] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *International Conference on Computer Vision (ICCV)*, 2021.

[6] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

APPENDIX

Generated poses with VIBE motion estimator

Test with bigger λ_{KL} values on PARE estimations

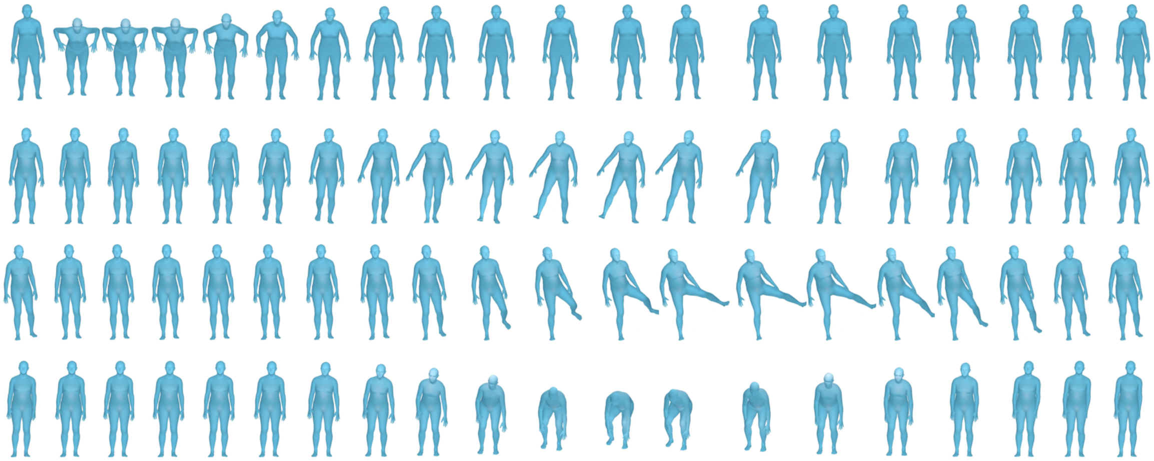

Figure 6: Motions generated by ACTOR using VIBE motion estimator. The movements were taken every 3 frames. The motions are all smooth although some of the poses like *Picking* on the first row seem to be incomplete

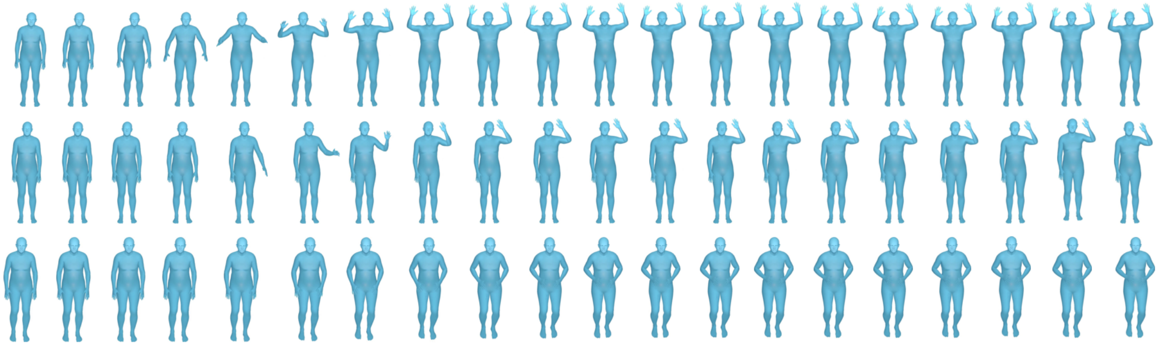

Figure 7: Motions generated on NTU13 dataset when action is trained on PARE estimations with $\lambda_{KL} = 5.10^{-5}$. For most labels of the labels, the motion tends to freeze,especially when raising an arm of a leg. On the last line, the man is supposed to be running and yet he is almost immobile. A smaller value is therefore more suitable for our problem.