JUMPS: Joints Upsampling Method for Pose Sequences

Lucas Mourot, François Le Clerc, Cédric Thébault and Pierre Hellier
InterDigital, Inc

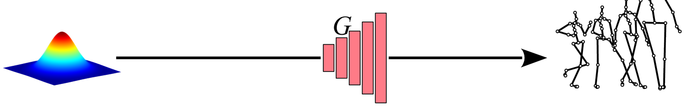
Abstract

Human Pose Estimation is a low-level task useful for surveillance, human action recognition, and scene understanding at large. It also offers promising perspectives for the animation of synthetic characters. For all these applications, and especially the latter, estimating the positions of many joints is desirable for improved performance and realism. To this purpose, we propose a novel method called JUMPS for increasing the number of joints in 2D pose estimates and recovering occluded or missing joints. We believe this is the first attempt to address the issue. We build on a deep generative model that combines a Generative Adversarial Network (GAN) and an encoder. The GAN learns the distribution of high-resolution human pose sequences, the encoder maps the input low-resolution sequences to its latent space. Inpainting is obtained by computing the latent representation whose decoding by the GAN generator optimally matches the joints locations at the input. Post-processing a 2D pose sequence using our method provides a richer representation of the character motion. We show experimentally that the localization accuracy of the additional joints is on average on par with the original pose estimates.

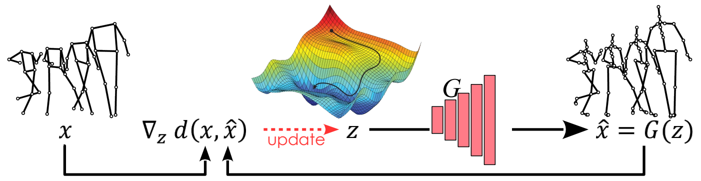
Overview

We improve and enrich 2D human poses, e.g. human pose estimation outputs, with spatio-temporal priors by:

1) Learning the human motion distribution with a deep generative model:



2) Upsampling and completing joints through optimization in the model's latent space:



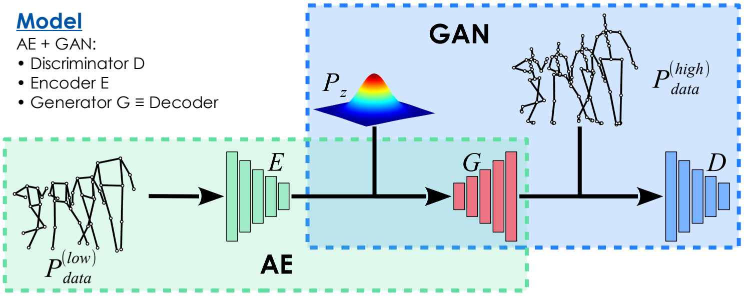
Contributions

- A novel method based on a deep generative model for inpainting 2D pose sequences and upsampling the joints, relying on temporal analysis.
- A hybrid GAN/autoencoder architecture; we show that the autoencoder is crucial for a better convergence and accuracy.
- We show that optimization in latent space is greatly improved by adding a Procrustes alignment at each iteration.
- We provide qualitative and quantitative assessments of the effectiveness of our method on the MPI-INF-3DHP [1] human pose dataset.

Model

AE + GAN:

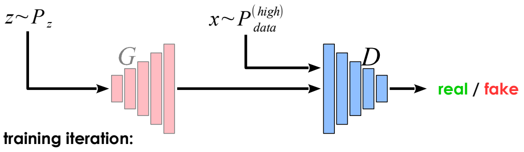
- Discriminator D
- Encoder E
- Generator G ≡ Decoder



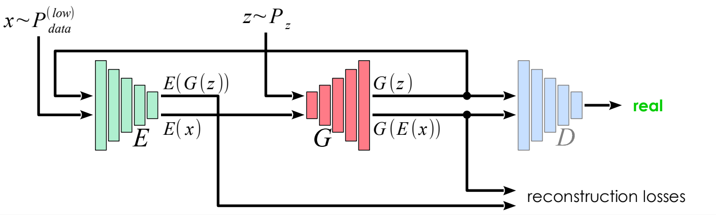
Training

- Alternating between discriminator, and encoder & generator.
- GAN framework: Wasserstein GAN with gradient penalty.
- AE framework: reconstruction & backward reconstruction losses.

Discriminator's training iteration:



Encoder & Generator's training iteration:



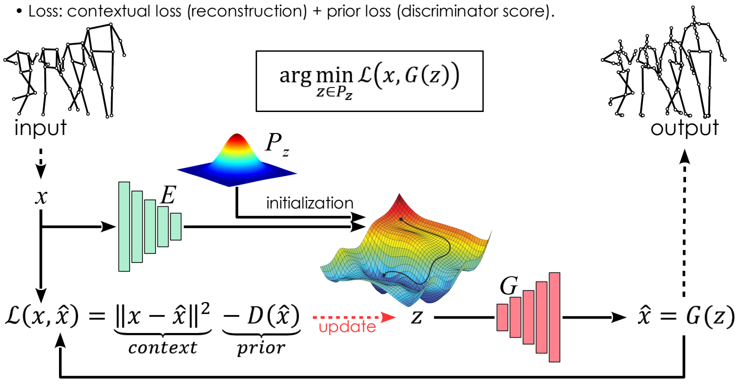
References

[1] Mehta et al., Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision, 3DV 2017.
[2] Fang et al., RMPE: Regional Multi-Person Pose Estimation, ICCV 2017.
[3] Li et al., CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark, arXiv:1812.00324, 2018.
[4] Xiu et al., Flow: Efficient Online Pose Tracking, BMVC 2018.

Improving Human Motion

We optimize z to recover the completed high-res. version of input x (low-res.) with $G(z)$.

- Several optimizations in parallel starting from different latent codes z, including $E(x)$.
- $G(z)$ is aligned (Procrustes Analysis) with x at each iteration.
- Loss: contextual loss (reconstruction) + prior loss (discriminator score).



$$\arg\min_{z \in P_z} \mathcal{L}(x, G(z))$$

$$\mathcal{L}(x,\hat{x}) = \underbrace{\|x-\hat{x}\|^2}_{context} \underbrace{- D(\hat{x})}_{prior}$$

Results

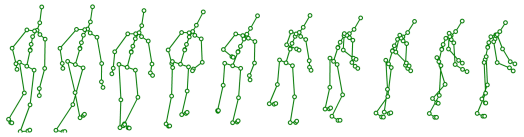
- Dataset: MPI-INF-3DHP [1]; Pose estimator: AlphaPose [2-4].
- Metrics: Percentage of Correct Keypoints PCK; Area Under the Curve (AUC).
- Ablations: without alignment "*JUMPS w/o P.A.*", without encoder "*JUMPS w/o ENC*".

1) Joints Upsampling

method	PCKh@0.1	PCKh@0.5	PCKh@1.0	AUC [0, 1]
JUMPS w/o P.A.	0.0368	0.4384	0.6814	0.3912
JUMPS w/o ENC.	0.1701	0.8259	0.9678	0.7005
JUMPS (ours)	**0.6096**	**0.9674**	**0.9965**	**0.8803**

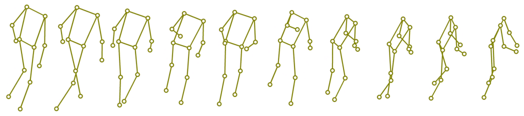
Ground truth

- MPI-INF-3DHP [1] annotations
- 28 joints



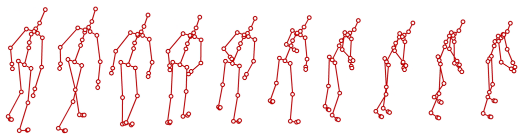
Ground truth

- 12 joints downsampled from 28 joints



JUMPS (ours)

- input: 12-joints ground truth
- output: 28 joints

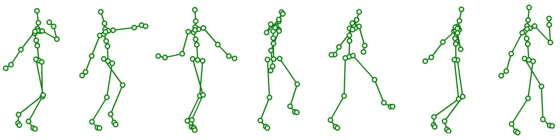


2) Human Pose Estimation Post-Processing

method	PCKh@0.1	PCKh@0.5	PCKh@1.0	AUC [0, 1]
AlphaPose [2-4]	**0.0941**	0.7659	0.9157	0.6310
JUMPS w/o P.A.	0.0207	0.3423	0.6304	0.3249
JUMPS w/o ENC.	0.0537	0.6801	0.9059	0.5692
JUMPS (ours)	0.0842	**0.7723**	**0.9276**	**0.6341**

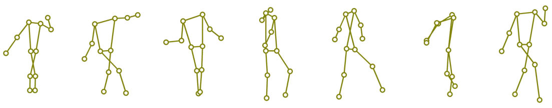
Ground truth

- 28 joints
- MPI-INF-3DHP [1] annotations



AlphaPose [2-4]

- input: images, frame by frame
- output: 12 joints



JUMPS (ours)

- input: AlphaPose outputs
- output: 28 joints