PoseScript: 3D Human Poses from Natural Language

Ginger Delmas[1,2], Philippe Weinzaepfel[1], Thomas Lucas[1], Francesc Moreno Noguer[2], Grégory Rogez[1]
[1]NAVER LABS Europe [2]IRII-UPC-CSIC

ECCV TEL AVIV 2022

paper ● code ● data

Institut de Robòtica i Informàtica Industrial

NAVER LABS Europe

Motivation

Using text to improve semantic understanding of human poses

Teaching poses

Pseudo 3D annotation when MoCap is difficult

Digital pose design

Contributions

- **PoseScript dataset**: descriptions of 3D human poses
- Text-to-Pose **retrieval** model
- Text-conditioned **generative** model

Downward Dog yoga pose[1]

The pose has the head down, ultimately touching the floor, with the weight of the body on the palms and the feet. The arms are stretched straight forward, shoulder width apart; the feet are a foot apart, the legs are straight, and the hips are raised as high as possible.

The PoseScript dataset

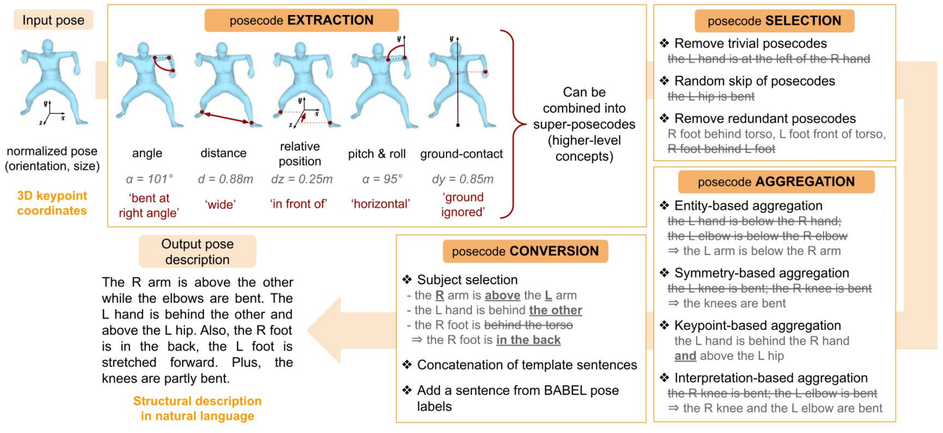
PoseScript dataset =  AMASS[2] 3D pose (20k) + Human-written description (4k) + Several automatic descriptions (6 x 20k)

Collection of human-written descriptions on Amazon Mechanical Turk[3] (AMT)

Someone is stretching, bent double forward, the head between the arms, and the arms relaxed **with hands touching at the feet**. The knees are slightly bent. The legs are close but not joined.

Automatically generate pose descriptions thanks to a randomized captioning pipeline

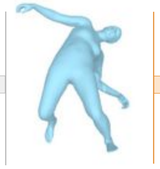
Get more training data at no cost: generate 60k descriptions in the time it takes to write 1!



Example

Data collection on AMT

The person is standing while bending backwards, as if they are dodging bullets in The Matrix. Both legs are bent backwards, and their arms are at their sides while not touching the ground.

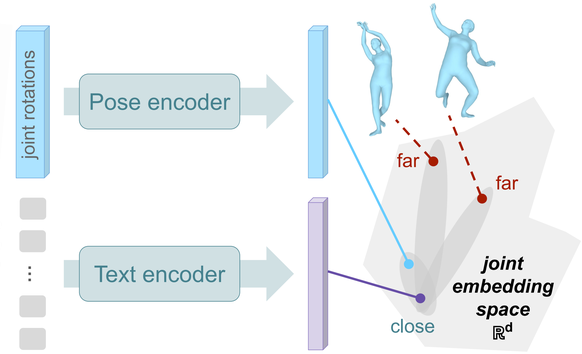


Automatic Captioning Pipeline

The figure is doing backwards movements and is in a inclined pose. The right knee is forming a L shape and the left foot is stretched forwards, the right elbow is barely bent, then the left shoulder is further down than the right. The subject is inclined backward and to the left of the pelvis. The left hand is further down than the left hip and behind the right hand and wide apart from the right hand, the right leg is behind the other. The right upper arm is parallel to the ground.

Text-to-Pose Retrieval model

The right arm is above the other while the elbows are bent [...] the right foot is in the back, the left foot is stretched forward. The knees are partly bent.

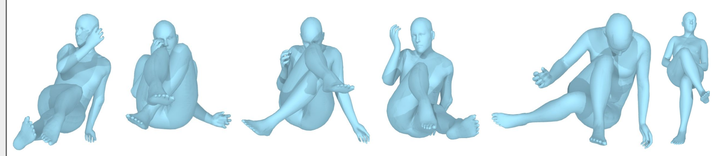


Performance on the human-written descriptions when pretraining on the automatic descriptions generated by our captioning pipeline:

	mRecall↑
without pretraining	12.8
with pretraining	**29.8**

Legs spread and both hands on hips.

Someone is sitting with their right leg crossed over their left. The back is reclined to a lounging position. Their head is upright, turned slightly to their left as the hands are folded on their lap.

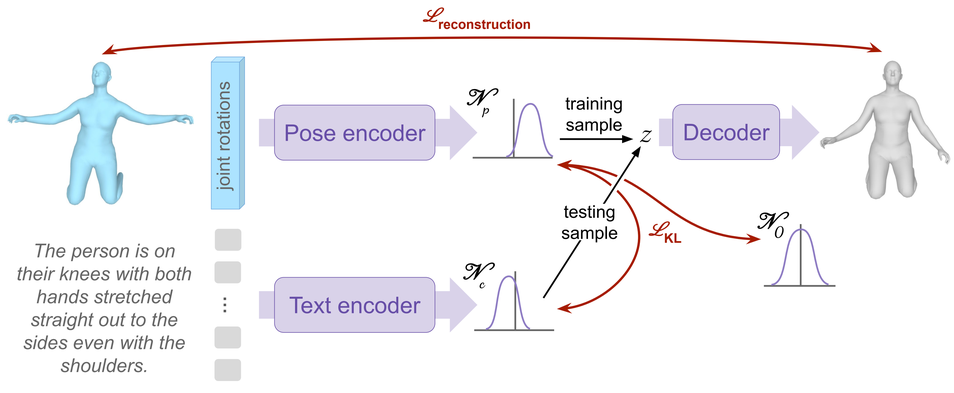


Pose retrieval from images using EFT[4] SMPL fits on MS-COCO[5]

Legs crossed, head facing forward.

The person is kneeling down like they are at the starting line of a race. The right knee is slightly off of the ground and the right hand is in front of it touching the ground.

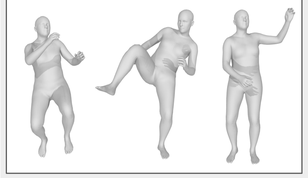
Text-conditioned Generative model of 3D Human Poses



The person is on their knees with both hands stretched straight out to the sides even with the shoulders.

He is bent forward with both knees bent and his face right above knee level looking down. His right arm is hanging down with his elbow bent and his hand in front between both of his legs and the palm of his hand is facing towards his legs. His left arm is extended out to his side at about waist height with the palm of his hand faced up.

The left elbow is bent at square angle.



The more detailed the description, the less diverse the different generated poses.

Take-home messages

- **PoseScript** is a dataset pairing **3D human poses** with both automatically generated and **human-written descriptions**.
- We use it to train a **text-to-pose retrieval model** and a **text-conditioned pose generative model**.
- We show that **better performance** on human data when **pretraining on automatic descriptions** generated by our captioning pipeline.

References

[1] https://en.wikipedia.org/wiki/Downward_Dog_Pose
[2] AMASS: Archive of Motion Capture as Surface Shapes, Mahmood et al., ICCV 2019
[3] https://www.mturk.com/
[4] Exemplar fine-tuning for 3D human model fitting towards in-the-wild 3D human pose estimation, Joo et al., 3DV 2020
[5] Microsoft COCO: Common Objects in Context, Lin et al., ECCV 2014