Motion Priors for Pose Estimation and Animation Workflows

Luke Smith

Master Thesis
April 2022

Prof. Dr. Robert W. Sumner

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

cgl
computer graphics laboratory

Abstract

TODO

Zusammenfassung

TODO: translate to German

Contents

List of Figures

List of Tables

Introduction

Introduction

- Existing pipeline description - 2d keypoints - 3d keypoints - Optimisation of cameras, ground plane, etc. - Existing pipeline problems - Robustness - Investigation: Motion Priors - Use for plausible motion

Related Work

To begin with, a review of Pose Estimation techniques is presented, as this forms an important part of the motion learning pipeline, which often operates on 2D/3D pose estimations. Next the motion prior literature is explored and related works for learning human motion are presented.

2.1. Pose Estimation

The existing pipeline for 2D pose estimation is based on Open Pose [CHS$^+$19]. OpenPose TODO

2.2. Motion Priors

Classical methods are presented in [SMK22], the involve matching/interpolating from existing databases.

The authors of HuMoR [RBH$^+$21] presented a novel approach for learning and using a plausible motion prior. They train a conditional VAE that learns a distribution over latent transitions, in a canonical reference frame, between *states* that consist of a root translation, 3D joint positions, joint angles, and the respective velocities. They most notably use this model as a prior in a 'test time optimisation', which generates plausible sequence motions optimising for an initial state and a sequence of transitions starting from frame by frame estimates (2D/3D joints or points clouds). This optimisation includes, alongside others, a motion prior term based upon the conditional distribution $p(z_t|x_{t-1})$ that encourages plausible motion for the learned sequence. Note that the CVAE decoder also predicts ground plan contact alongside change in state, which are used as regularisers during their main use case 'test time optimisation'. The test time optimisation can operate on many modalities, 2D/3D joints, point clouds, etc., as the optimisation contains a Data Term ϵ_{data} that can be tailored to the modality as the HuMoR state is information rich, containing 3D joints (hence can fit to 2D joints through projection or directly to 3D) and can parametrise the SMPL model (hence the SMPL mesh can be correlated to point clouds). The initialisation for the test time optimisation is based upon VPoser TODO.

HuMoR discussions:

- Assumptions:
  - The method necessitates knowledge of the ground plane, which is presently needed (empirical observation) for convergence during training (as the dataset is of motions with a flat ground), and thus also at test time even though it is not conceptually necessary
  - Assumes static camera
- Limitations:
  - Single person formulation

The authors of HuMoR [RBH^{+}21] were inspired by the Motion VAE [LZCvdP21] paper. This paper uses an Conditional VAE (with assumed standard normal prior conditioning (vs. NN in HuMoR)) that directly outputs the next state (rather than the change in state in HuMoR). The model is used Autoregressively to predict motion (rather than the main presented use of HuMoR which is to fit motion to a sequence of existing 2D/3D joint predictions, though HuMoR can equally well be used autoregressively), and is trained with the typical ELBO in a supervised manner.
Some notes to self about MotionVAE

- RL algo trained to walk the latent space
- Some notes about things they mention in the related work section:
  - They cite [Wang et al. 2019] who train a stochastic generative model with output *processed by a refiner network to remove foot skating and add robustness.*
- Main differences to HuMoR
  - c.f discussion section in HuMoR
  - Conditional prior
  - Predict change in motion
  - Predict ground contacts
  - Much additional regularisation in training
  - Difference state representation (root projected to ground)
  - Use of SMPL by HuMoR
  - Difference in network architectures
    - HuMoR just uses MLPs and MotionVAE decoder is a 'MANN-style mixture-

    of-expert neural network' (6 networks, gating network weighting their outputs)

    - RELU in HuMoR, ELU in MotionVAE
    - MotionVAE decoder has latent variable input at each layer (not sure about HuMoR)

  –

The learned-inbetweenings [TWH^{+}22] paper is very similar to MotionVAE. They present a similar architecture except that they predict transitions as in HuMoR [RBH^{+}21], and that they seem to only predict the lower body joint velocities and rotations. The decoder architecture is similar with a gating network and multiple expert networks. They also train a sampler to sample from the latent space, similar to MotionVAE which trains a RL model. Some notes to self about Learned-Inbetweenings

- Contains a nice probabilistic formulation of the motion generation problem

There seem to be quite a number of works that present the CVAE architecture as a base with varying state representations, conditioning variables and loss terms, [RBH^{+}21, TWH^{+}22, LZCvdP21, MWFD21]. [RBH^{+}21] has a learned prior, , [TWH^{+}22] has a normal prior and predicts

DeepPhase [SMK22] proposes a convolutional autoencoder that operates on fixed length (TODO: I beleive fixed length) sequences of 3D joint velocities, learning a latent space that it encourages to represent sinusoidal functions (through phase/frequency/amplitude/offsets) that represent periodic features of motion. The auto-encoder maps to and from sequences of 3D joint velocities, and the latent variables represent a sequence of phase/etc. values over the whole motion, hence the change in parameters can represent a shift between different periodic motions and thus can describe non-periodic motions.

MEVA [LGK20] postulates that learning a single motion model results in smooth motion, as on average human motion is smooth (i.e we are not shaking while walking (their words)), hence they propose a two stage pipeline, in which the results of VAE that esimates coarse motion is passed into a human shape regressor that refines the poses, the inputs are temporally correlated features hence temporal consistency is maintained. The paper also presents some motion specific data augmentation techniques, speed variation through sampling, mirroring, and root rotations. Some notes to self about Learned-Inbetweenings

- TODO: very useful
- Nice related work section
- Nice dataset section

Another approach is presented in VIBE

2.3. Overview of Approaches

2.3.1. Motion Priors

We are most interested in models that learn plausible, task independent, human motion. These are refered to by [LZCvdP21] as *Motion-then-control* models. We limit our scope to parametric models.

- **Motion Priors**
- MotionVAE [RBH^{+}21]
  - Standard normal CVAE
  - Outputs next pose
  - Decoder is mixture of networks
  - Trained with rollout and scheduled sampling
  - State positions referenced to root projection onto ground
  - Nice investigation into using RL in the latent space for character control
  - NOTE: Latent dimension size: 32 (typical physics based humanoid degrees of freedom).
  - The state having velocities and the decoder predicting change in pose seems to implicitly model the time aspect of the motion, rather than explicitly modelling it like in [MWFD21].
- HuMoR [RBH^{+}21]
  - Parametrised conditional prior CVAE
  - Outputs change in state and person ground contacts
  - SMPL regularisers (a subset of their state parametrises the SMPL model)
  - Motion learned in a canonical reference frame (TODO: not sure about MotionVAE)
  - Trained without rollout (I beleive?)
  - State positions referenced as in SMPL model (to $(0, 0)$?)
  - Ground plane initialized with RCNN
  - Very nice feature of having velocities in the state and of predicting change in motion, this implicitly captures the direction of motion in time as well as in space
- Learned-inbetweenings paper [TWH^{+}22]
  - Bascially MotionVAE but outputs change in state like HuMoR
- Structed latent space for 4D motion [MWFD21]
  - VAE operating on a fixed number of frames but with 'varying duration' by including

a timestamp per frame (hence with wider spacing in the timestamp the movement is over a longer duration)

  - Conditions decoder with SMPL shape
  - It encorporates the timestamp to distinguish the direction of motion (i.e to avoid having you sample backwards in time when you walk the latent space. It's a direct next pose prediction and so would cluster close poses in the latent space regardless of time (I beleive))
  - They perform a comparison to other reconstruction methods to directly evaluate the VAE, not sure other papers did that much
  - Didn't find the paper so interesting

- DeepPhase [SMK22]
  - Autoencoder operating on fixed length sequences of 3D joint velocities
  - Latent space enforced to match sinusoidal functions that represent periodic motion
  - Periodic functions can change over the length of the sequence thus shift between periodic motions and represent non-periodic motions
- MEVA [LGK20]
  - Separates pipeline into learning coarse motion VAE and refining this prediction (they postulate the VAE can only learn smooth motion, then use a SMPL regressor to refine the predictions from temporally correlated features)
  - VAE operates on features extracted with temporal convolutions directly from the image rather than on SMPL/joint position/velocity based state
  - Presents nice augmentation techniques for AMASS, speed variation, mirroring and root rotations
  - Nice related work and dataset sections
- TransformerVAEPrior [CSY$^+$22]
  - Transformer based VAE motion prior
  - Operates on a **sequence** level, rather than just between 2 poses like HuMoR, latent code represents an entire sequence so the sequence can be decoded in one step => faster
  - Also uses AMASS
  - Can also be used as a motion rectifier like Humor, they say it's faster as it's a direct prediction of sequence
  - TODO: not sure about the length of the input sequence
- [HSKJ15]
  - Old paper but relevant ideas

  - Simple autoencoder (not variational), CNN based
  - Operates on sequence level, like TransformerVAEPrior
  - Denoising Autoencoding training (corrupting inputs with noise and recovering uncorrupted version), might be helpful for us as we want to de-corrupt data essentially
  - ConvAutoEnv2016 [HSK16] improves the architecture a little I beleive?

- **Motion aware but not focused on prior**
- VIBE
  - Operates directly on video
  - CNN features processed by GRUs (gated reccurence unit) to temporally correlate the features
  - Features fed into a NN regressor to estimate SMPL params as in [KBJM17]
  - Discriminator network jointly trained to introduce an extra loss to encourage plausible motion
- 3DDynamicsFromVideo [KZFM18]
  - Operates directly on video sequences
  - PRedicts pose at t-1,t,t+1 from all image features during training and also jointly trains a NN that predicts t-1,t,t+1 from just image at t
- Some more historic RGB methods presented in HuMoR [RBH+21] if needed
- Optimisation methods that refine predictions presented in HuMoR [RBH+21], e.g with smoothness priors or scene contact info
- **Pose estimation**
- HULC
  - Uses a scene point cloud to help generate dense contact estimation labels that are used to guide pose manifold sampling
- VPoser
  - Not actually a motion prior, it's a pose prior
  - Is used in Humor to help initialise the sequence of states
  - VPoser is used by optimising pose directly in the latent space, the latent space is trained to be a normal distribution hence if you penalise the norm of the latent vector you are encouraging it to be close to what you've learned to be viable human poses (i.e close to the normal dist)
- **TODO: Things I haven't looked into so deeply**
- RNNS
  - RNNs seem to be commonly used for generating future motion condition on control

    variable

  - Mixture-density network RNNs (MDN-RNNS)
    - Referenced in [LZCvdP21]
    - Output a distribution as a gaussian mixture model
  - SpatioTemporalRNN
    - (https://arxiv.org/pdf/1908.07214.pdf) cited by learned-inbetweening paper [TWH+22]
    - Learns a manifold through encoder/decoder
    - Separates 'spatial' and 'temporal' in encoder and decoder??
    - Predicts in batches with RNN, they claim this forces the model to capture mid and long term connections (I beleive the explicit velocity modelling in HuMoR should do the same thing)
  - Deep latent variable model
    - https://dl.acm.org/doi/pdf/10.1111/cgf.14116

- Other VAE motion prediction
  - Unified3DHumanMotionSynthesis
  - ActionConditionedTransformer
    - Throw a transformer at the problem
    - Direct pose sequence prediction
    - Condition on action
- Time-convolutional autoencoders
  - Referenced in [LZCvdP21]
  - Learns a latent motion manifold
- Humor claims normalising flows and neural ODEs show potential but then only links to papers explaining these concepts and not actually using them for this purpose so not sure
  - (Normalising flow: map to a simple distribution with an invertible function => tractable marginal likelihood (unlike with VAEs where we have to deal with an ELBO), but I'm not sure we care about the marginal likelihood in this case)
  - [HAB19] MoGlow referenced, it predicts next pose directly by sampling from a simple fixed distribution and then transforming to a pose via a normalising flow conditioned on conditioning variables and pose histories that are updated and introduced through a hidden lstm in the flow.

2.3.2. Pose Completion

The other possibile research direction related more to pose completion (or motion completion (though I assume motion completion would use similar methods to what was mentioned before)) in an animation setting.

- Protores [OBH$^+$21]
  - Learned inverse kinematics solution
  - Variable number of effector inputs processed and then mixed with a 'Proto' layer in the encoder
  - Decoder takes the pose embedding and decodes the full pose (contains several blocks to separate the semantically different parts of the decoding process)

Main

Intro to central work section

3.1. Section

TODO

Conclusion and Outlook

TODO

Appendix

Appendix intro

A.1. Section

TODO

Bibliography

[CHS$^+$19] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[CSY$^+$22] Xin Chen, Zhuo Su, Lingbo Yang, Pei Cheng, Lan Xu, Bin Fu, and Gang Yu. Learning variational motion prior for video-based motion capture, 2022.

[HAB19] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *CoRR*, abs/1905.06598, 2019.

[HSK16] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Trans. Graph.*, 35(4), jul 2016.

[HSKJ15] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs*, New York,NY,USA, 2015. Association for Computing Machinery.

[KBJM17] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. *CoRR*, abs/1712.06584, 2017.

[KZFM18] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. *CoRR*, abs/1812.01601, 2018.

[LGK20] Zhengyi Luo, S. Alireza Golestaneh, and Kris M. Kitani. 3d human motion estimation via motion compression and refinement. *CoRR*, abs/2008.03789, 2020.

[LZCvdP21] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. Character controllers using motion vaes. *CoRR*, abs/2103.14274, 2021.

[MWFD21] Mathieu Marsot, Stefanie Wuhrer, Jean-Sébastien Franco, and Stephane Durocher. Multi-frame sequence generator of 4d human body motion. *CoRR*, abs/2106.04387, 2021.

[OBH$^+$21] Boris N. Oreshkin, Florent Bocquelet, Félix G. Harvey, Bay Raitt, and Dominic Laflamme. Protores: Proto-residual architecture for deep modeling of human pose. *CoRR*, abs/2106.01981, 2021.

[RBH$^+$21] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021.

[SMK22] Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Trans. Graph.*, 41(4), jul 2022.

[TWH$^+$22] Xiangjun Tang, He Wang, Bo Hu, Xu Gong, Ruifan Yi, Qilong Kou, and Xiaogang Jin. Real-time controllable motion transition for characters. *ACM Transactions on Graphics*, 41(4):1–10, jul 2022.