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

This section describes references

2.1. Pose Estimation

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

2.2. Motion Priors

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).

I would see the success of HuMoR in occlusion situations to be largly due to the SMPL prior.

HuMoR discussions:

- They consider extending the method to include body shape parameters in the state an important direction for improved generalisation.
- They claim normalising flows and neural ODEs show potential but they only link 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)
- They claim 'MVAE' does not work well
- The SMPL regularisation and the learned contional prior are important during training
- 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 [TODDO] 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

- MotionVAE is used with Deep RL with the action space taken to be the latent space of the CVAE, with a reward function that defines goals of a character, the control policy walks through the actions space to guide the generative model in accordance with these goals. Could be interesting for interactive character animation
- Their state representation has some differences to HuMoR, notably that the root position his projected onto the ground.
- Contains a nice overview of motion prediction methods
- Latent dimension size: 32 (typical physics based humanoid degrees of freedom)
- 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
  - Use of SMPL by HuMoR
  - Difference in network architectures
    * HuMoR just uses MLPs and MVAE decoder is a 'MANN-style mixture-of-expert neural network' (6 networks, gating network weighting their outputs)
    * RELU in HuMoR, ELU in MVAE
    * MVAE decoder has latent variable input at each layer (not sure about HuMoR)
  -

2.2.1. Overview of Approaches

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.

- MVAE [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
- 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 MVAE)
  - Trained without rollout (I beleive?)
  - State positions referenced as in SMPL model (to $(0, 0)$?)

- **TODO: Things I haven't looked into so deeply**
- Mixture-density network RNNs (MDN-RNNS)
  - Referenced in [LZCvdP21]
  - Output a distribution as a gaussian mixture model
- Time-convolutional autoencoders
  - Referenced in [LZCvdP21]
  - Learns a latent motion manifold
  - Followup paper also referenced
- Humor claims normalising flows and neural ODEs show potential but they only link to papers explaining these concepts and not actually using them for this purpose so not sure

2.2.2. General notes

Training VAEs:

- Posterior collapse
  - Decoder ignores latent variable and overfits to the training sequences
  - MVAE: input latent variable at each stage of the decoder
  - Weighting of β-VAE
- Quality vs. generalisation
  - Depends on weighting of KL and Reconstructions Losses
  - MVAE find empirically that: *MVAE reconstruction and KL losses being within one order of magnitud*
- Stable sequence prediction
  - Not sure we care so much as we probably won't be extrapolating new sequences
  - MAVE: Scheduled sampling during training - model progressively learns to deal with it's own predictions

Your Central Work

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero et accumsan et iusto odio dignissim qui blandit.

3.1. First Section

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero et accumsan et iusto odio dignissim qui blandit.

Delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis

Quant.	*Ingredient*
200g	Weißmehl
1/4	Packung Frischhefe
4EL	lauwarme Milch
4EL	ol
1TL	Zucker
1TL	Salz
	lauwarmes Wasser

Table 3.1.: *Flammkuchenteig. The ingredients have to be carefully chosen.*

nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero et accumsan et iusto odio dignissim qui blandit.

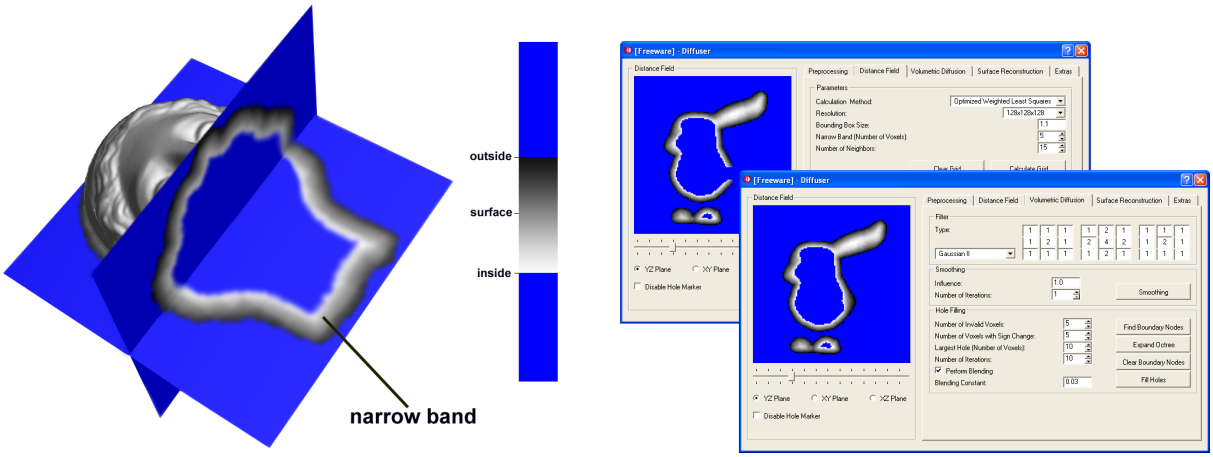
3.1.1. Fist Subsection

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam

3.1.2. Another Subsection

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat, see Table 3.1. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam, see Figure 3.1 (a). Isn't it?

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero et accumsan et iusto odio dignissim qui blandit



(a) (b)

Figure 3.1.: *Volumetric diffusion.* (a) *Slices of the distance volume reveal the narrow band.* (b) *The user interface of the automatic hole filling tool allows to fine-tune the algorithm. The volumetric representation can be previewed before surface reconstruction.*

(a) Caption first. *(b) Caption second.*

Figure 3.2.: *Caption of both (a), (b).*

praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam

3.2. Second Section

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat

nulla facilisis at vero et accumsan et iusto odio dignissim qui blandit

Conclusion and Outlook

TODO

Information For The Few

Nein, meine Texte les ich nicht, so nicht, st?hnte Oxmox. Er war mit Franklin, Rockwell und dem halbtaxgrauen Panther Weidemann in Memphis (Heartbreak Hotel) zugange. Sie warteten auf die fette Gill, um bei der Bank of Helvetica die Kapit?lchen in Kapital umzuwandeln. Oxmox liess nicht locker. Ich fleh euch an, rettet meine Copy, gebt meinem Body nochn Durchschuss! Kein Problem, erbarmte sich Old Face Baskerville, streichelte seinen Hund, zog seine einspaltige Poppl, legte an und traf! (Zeidank nichts Ernstes — nurn bisschen Fraktur.) Oxmox: Danke, ist jetzt mit Abstand besser. Derweil jumpte der Fox leise over the Buhl, die sich mal wieder immerdar wie jedes Jahr gesellte. Diesmal war Guaredisch ihr Erw?hlter, weil seine Laufweite einem vollgetankten Bodoni entsprach und seine ungez?gelte Unterl?nge ihre Serifen so serafisch streifte, dass sie trotz Techtelmechtelei die magere Futura, jene zuverl?ssige und gern eingesetzte Langstreckenl?uferin, rechtsb?ndig ?berholen konnten.

A.1. Foo Bar Baz

Nein, meine Texte les ich nicht, so nicht, st?hnte Oxmox. Er war mit Franklin, Rockwell und dem halbtaxgrauen Panther Weidemann in Memphis (Heartbreak Hotel) zugange. Sie warteten auf die fette Gill, um bei der Bank of Helvetica die Kapit?lchen in Kapital umzuwandeln. Oxmox liess nicht locker. Ich fleh euch an, rettet meine Copy, gebt meinem Body nochn Durchschuss! Kein Problem, erbarmte sich Old Face Baskerville, streichelte seinen Hund, zog seine einspaltige Poppl, legte an und traf! (Zeidank nichts Ernstes — nurn bisschen Fraktur.) Oxmox: Danke, ist jetzt mit Abstand besser. Derweil jumpte der Fox leise over the Buhl, die sich mal wieder immerdar wie jedes Jahr gesellte. Diesmal war Guaredisch ihr Erw?hlter, weil seine Laufweite einem vollgetankten Bodoni entsprach und seine ungez?gelte Unterl?nge ihre Serifen so serafisch streifte, dass sie trotz Techtelmechtelei die magere Futura, jene zuverl?ssige und gern eingesetzte Langstreckenl?uferin, rechtsb?ndig ?berholen konnten.

A.2. Barontes

Nein, meine Texte les ich nicht, so nicht, st?hnte Oxmox. Er war mit Franklin, Rockwell und dem halbtaxgrauen Panther Weidemann in Memphis (Heartbreak Hotel) zugange. Sie warteten auf die fette Gill, um bei der Bank of Helvetica die Kapit?lchen in Kapital umzuwandeln. Oxmox liess nicht locker. Ich fleh euch an, rettet meine Copy, gebt meinem Body nochn Durchschuss! Kein Problem, erbarmte sich Old Face Baskerville, streichelte seinen Hund, zog seine einspaltige Poppl, legte an und traf! (Zeidank nichts Ernstes — nurn bisschen Fraktur.) Oxmox: Danke, ist jetzt mit Abstand besser. Derweil jumpte der Fox leise over the Buhl, die sich mal wieder immerdar wie jedes Jahr gesellte. Diesmal war Guaredisch ihr Erw?hlter, weil seine Laufweite einem vollgetankten Bodoni entsprach und seine ungez?gelte Unterl?nge ihre Serifen so serafisch streifte, dass sie trotz Techtelmechtelei die magere Futura, jene zuverl?ssige und gern eingesetzte Langstreckenl?uferin, rechtsb?ndig ?berholen konnten.

A.3. A Long Table with Booktabs

Table A.1.: *A sample list of words.*

ID	Word	Word Length	WD	ETL	PTL	WDplus
1	Eis	3	4	0.42	1.83	0.19
2	Mai	3	5	0.49	1.92	0.19
3	Art	3	5	0.27	1.67	0.14
4	Uhr	3	5	0.57	1.87	0.36
5	Rat	3	5	0.36	1.71	0.14
6	weit	4	6	0.21	1.65	0.25
7	eins	4	6	0.38	1.79	0.26
8	Wort	4	6	0.30	1.62	0.20
9	Wolf	4	6	0.18	1.54	0.19
10	Wald	4	6	0.31	1.63	0.19
11	Amt	3	6	0.30	1.67	0.14
12	Wahl	4	7	0.36	1.77	0.42
13	Volk	4	7	0.45	1.81	0.20
14	Ziel	4	7	0.48	1.78	0.42
15	vier	4	7	0.38	1.81	0.42
16	Kreis	5	7	0.26	1.62	0.33
17	Preis	5	7	0.28	1.51	0.33
18	Re-de	4	7	0.22	1.56	0.33
19	Saal	4	7	0.75	2.10	0.43
20	voll	4	7	0.48	1.82	0.24
21	weiss	5	7	0.21	1.59	0.36
22	?r-ger	5	7	1.16	2.69	0.59
23	bald	4	7	0.18	1.56	0.19

continued on next page

Table A.1.: *(Continued)*

ID	Word	Word Length	WD	ETL	PTL	WDplus
24	hier	4	7	0.40	1.70	0.43
25	neun	4	7	0.17	1.52	0.26
26	sehr	4	7	0.36	1.85	0.43
27	Jahr	4	7	0.50	1.82	0.43
28	Gold	4	7	0.04	1.35	0.20
29	T?-ter	5	8	0.15	1.39	0.59
30	Tei-le	5	8	0.30	1.71	0.46
31	Na-tur	5	8	0.18	1.59	0.41
32	Feu-er	5	8	0.30	1.71	0.45
33	Rol-le	5	8	0.15	1.46	0.45
34	Rock	4	8	0.29	1.68	0.25
35	Spass	5	8	0.28	1.64	0.32
36	G?s-te	5	8	0.49	1.75	0.66
37	En-de	4	8	0.36	1.72	0.33
38	Kunst	5	8	0.26	1.59	0.35
39	Li-nie	5	8	0.45	1.88	0.63
40	B?u-me	5	8	0.48	1.92	0.45
41	B?h-ne	5	9	0.94	2.48	0.62
42	Bahn	4	9	0.21	1.62	0.42
43	B?r-ger	6	9	0.38	1.70	0.65
44	Druck	5	9	0.60	2.03	0.31
45	zehn	4	9	0.41	1.84	0.42
46	Va-ter	5	9	0.36	1.78	0.40
47	Angst	5	9	0.29	1.56	0.35
48	lei-der	6	9	0.13	1.47	0.52
49	h?u-fig	6	9	0.82	2.31	0.52
50	le-ben	5	9	0.38	1.85	0.40
51	aus-ser	6	9	1.20	2.26	0.57
52	be-vor	5	9	1.28	2.75	0.39
53	Kai-ser	6	9	0.92	2.37	0.53
54	Markt	5	9	0.23	1.58	0.28
55	Os-ten	5	9	0.21	1.54	0.48
56	Krieg	5	9	0.33	1.67	0.50
57	Mann	4	9	0.31	1.47	0.25
58	Hal-le	5	9	0.24	1.65	0.45
59	heu-te	5	9	0.44	1.87	0.46
60	in-nen	5	10	0.36	1.80	0.45
61	Na-men	5	10	0.28	1.72	0.41
62	jetzt	5	10	0.70	2.07	0.32
63	kei-ner	6	10	0.28	1.62	0.53
64	Schu-le	6	10	1.02	2.12	0.48

continued on next page

Table A.1.: *(Continued)*

ID	Word	Word Length	WD	ETL	PTL	WDplus
65	Ar-beit	6	10	0.34	1.70	0.52
66	An-teil	6	10	0.27	1.63	0.53
67	di-rekt	6	10	0.67	2.04	0.47
68	vor-her	6	10	0.78	2.25	0.47
69	wol-len	6	10	0.44	1.85	0.51
70	Kampf	5	10	0.70	1.96	0.27
71	?n-dern	6	10	1.18	2.62	0.65
72	lau-fen	6	10	0.21	1.64	0.52
73	Eu-ro-pa	6	10	0.23	1.53	0.66
74	statt	5	10	1.61	2.86	0.39
75	Wes-ten	6	10	0.29	1.60	0.54

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.

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

[TODDO] TODO. Todo. *TODO*, TODO(TODO), TODO.