Structured Prediction Helps 3D Human Motion Modelling

-ICCV 2019
-ETH Zürich

June 21, 2020

Chuan Guo

Preliminary

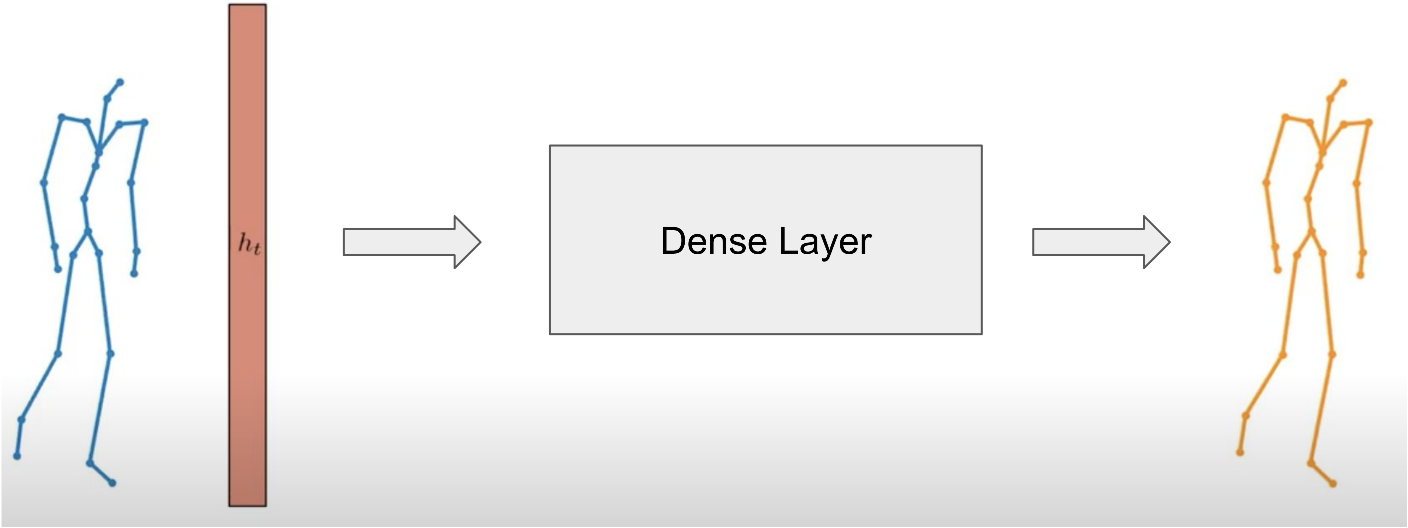
Method

Datasets and Models

Evaluation on Human3.6M

Evaluation on AMASS

Motivation



Contribution

- Main contribution:
  - Novel structured prediction layer which incorporate skeleton hierarchy.
  - This prediction layer is agnostic to the underlying network.
- Others:
  - Evaluations on Human3.6 and AMASS datasets.

Preliminary

Method

Datasets and Models

Evaluation on Human3.6M

Evaluation on AMASS

Structured Prediction Layer

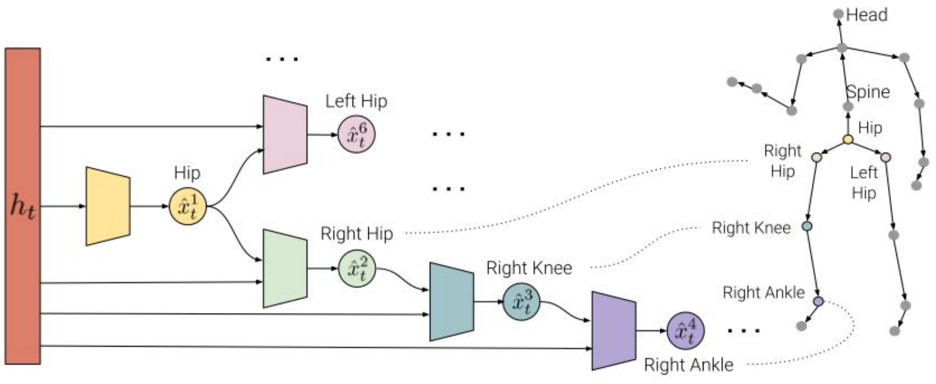


Figure 2: **SPL overview**. Given the context $\boldsymbol{h}_t$ of past frames, joint predictions $\hat{\boldsymbol{x}}_t^{(k)}$ are made hierarchically by following the kinematic chain defined by the underlying skeleton. Only a subset of joints is visualized for clarity.

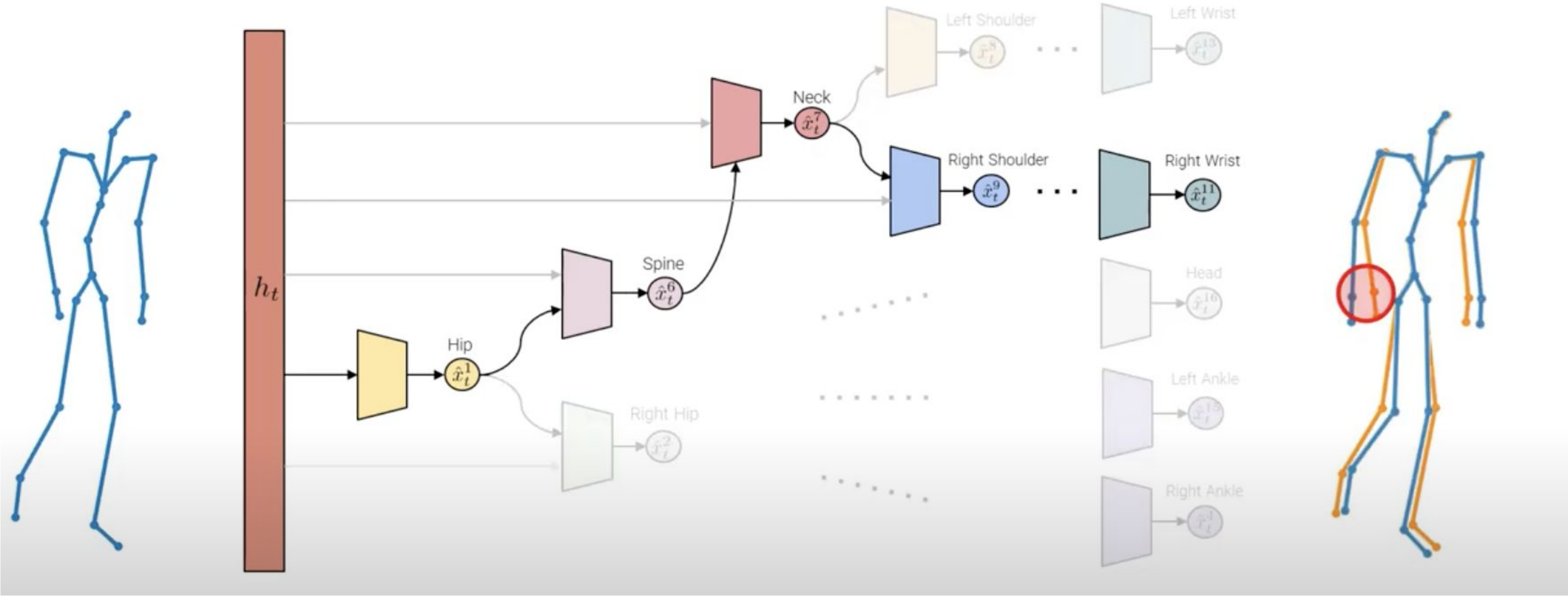
x_t is pose vector at t step.

K is the number of joints.

$$p_\theta(\boldsymbol{x}_t) = \prod_{k=1}^{K} p_\theta(\boldsymbol{x}_t^{(k)} \mid \text{parent}(\boldsymbol{x}_t^{(k)}), \boldsymbol{h}_t)$$

$$p_\theta(\boldsymbol{X}) = \prod_{t=1}^{T} \prod_{k=1}^{K} p_\theta(\boldsymbol{x}_t^{(k)} \mid \text{parent}(\boldsymbol{x}_t^{(k)}), \boldsymbol{h}_t)$$

Structured Prediction Layer



Per joint Loss

Ground truth

$$\mathcal{L}(\boldsymbol{X}, \hat{\boldsymbol{X}}) = \frac{1}{T \cdot N} \sum_{t=1}^{T} f(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_t)$$

Frames

Dimension

Prediction

k_th joint

$$\mathcal{L}(\boldsymbol{X}, \hat{\boldsymbol{X}}) = \sum_{t=1}^{T} \sum_{k=1}^{K} f(\boldsymbol{x}_t^{(k)}, \hat{\boldsymbol{x}}_t^{(k)})$$

Typically, the loss is calculated on pose vector space.

Here, loss is calculated for each joint first, and then summed up for the entire motion.

Preliminary

Method

Datasets and Models

Evaluation on Human3.6M

Evaluation on AMASS

Experiment(dataset)

Input sequences are 2 seconds(120 frames), targets are 400ms(24 frames)

- Human3.6
  - 632, 894 frames
  - 120 test samples across 15 categories
  - 21 joints

- AMASS
  - 9, 084, 918 frames
  - 3,304 test samples
  - 15 joints

Models

- Seq2seq: input poses are represented as **axis angle**(exponential map);
  - On human motion prediction using recurrent neural networks.(CVPR 2017)
- QuarterNet: inputs are **quaternion representation**.
  - Modeling human motion with quaternion based neural networks.(IJCV 2020)
- RNN: inputs are **rotation matrices**.
  - Single layer RNN network.

Preliminary

Method

Datasets and Models

Evaluation on Human3.6M

Evaluation on AMASS

Evaluation on Human3.6M

	Walking				Eating				Smoking				Discussion			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
LSTM-3LR [7]	0.77	1.00	1.29	1.47	0.89	1.09	1.35	1.46	1.34	1.65	2.04	2.16	1.88	2.12	2.25	2.23
SRNN [14]	0.81	0.94	1.16	1.30	0.97	1.14	1.35	1.46	1.45	1.68	1.94	2.08	1.22	1.49	1.83	1.93
Zero-Velocity [20]	0.39	0.68	0.99	1.15	0.27	0.48	0.73	0.86	0.26	0.48	0.97	0.95	0.31	0.67	0.94	1.04
AGED [33]	0.22	0.36	0.55	0.67	**0.17**	**0.28**	**0.51**	**0.64**	0.27	**0.43**	**0.82**	**0.84**	0.27	**0.56**	**0.76**	**0.83**
Seq2seq-sampling-sup [20]	0.28	0.49	0.72	0.81	0.23	0.39	0.62	0.76	0.33	0.61	1.05	1.15	0.31	0.68	1.01	1.09
Seq2seq-sampling-sup-SPL	0.23	0.37	**0.53**	**0.61**	0.20	0.32	0.52	0.67	0.26	0.48	0.92	0.90	0.29	0.63	0.90	0.99
Seq2seq-sampling [20]	0.27	0.47	0.70	0.78	0.25	0.43	0.71	0.87	0.33	0.61	1.04	1.19	0.31	0.69	1.03	1.12
Seq2seq-sampling-SPL	0.23	0.38	0.58	0.67	0.20	0.32	0.52	0.66	0.26	0.48	0.92	0.90	0.30	0.64	0.91	0.99
QuaterNet [25]	**0.21**	**0.34**	0.56	0.62	0.20	0.35	0.58	0.70	**0.25**	0.47	0.93	0.90	**0.26**	0.60	0.85	0.93
QuaterNet-SPL	0.22	0.35	0.54	**0.61**	0.20	0.33	0.55	0.68	**0.25**	0.47	0.91	0.88	**0.26**	0.59	0.84	0.91
RNN	0.30	0.48	0.78	0.89	0.23	0.36	0.57	0.72	0.26	0.49	0.97	0.95	0.31	0.67	0.95	1.03
RNN-SPL	0.26	0.40	0.67	0.78	0.21	0.34	0.55	0.69	0.26	0.48	0.96	0.94	0.30	0.66	0.95	1.05

Euler angle metric

Preliminary

Method

Datasets and Models

Evaluation on Human3.6M

Evaluation on AMASS

Evaluation on AMASS

Report accumulated error util time step t, instead of error at time step t.

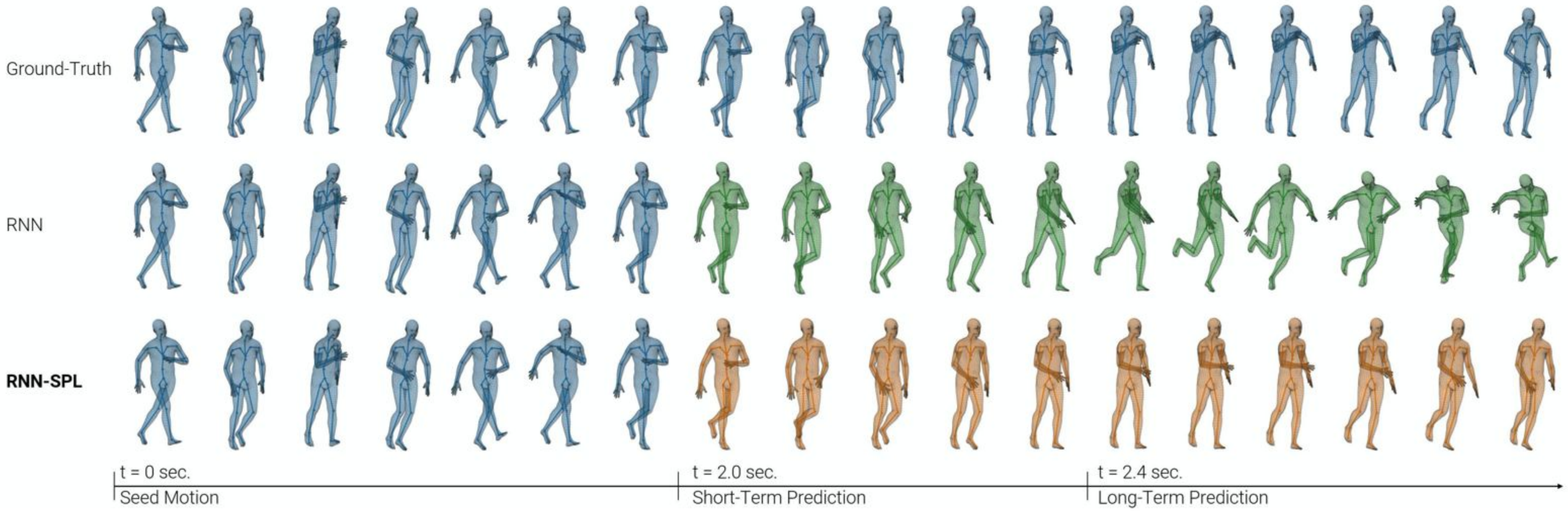
- Joint Angle Difference: error of rotation matrices

- Positional error: error of 3D joint positions

- PCK: percentage of predicted joints lying within a spherical threshold ρ around the target joint position

	Euler				Joint Angle				Positional				PCK (AUC)			
milliseconds	100	200	300	400	100	200	300	400	100	200	300	400	100	200	300	400
Zero-Velocity [20]	1.91	5.93	11.36	17.78	0.37	1.22	2.44	3.94	0.14	0.48	0.96	1.54	0.86	0.83	0.84	0.82
Seq2seq [20]*	1.46	5.28	11.46	19.78	0.24	0.95	2.16	3.87	0.09	0.35	0.80	1.41	0.91	0.87	0.87	0.83
Seq2seq-SPL	1.57	5.00	10.01	16.43	0.27	0.94	2.01	3.45	0.10	0.36	0.79	1.36	0.91	0.87	0.87	0.84
Seq2seq-sampling [20]*	1.71	5.15	9.71	15.15	0.32	1.00	1.97	3.14	0.12	0.39	0.77	1.23	0.88	0.86	0.87	0.85
Seq2seq-sampling-SPL	1.71	5.13	9.60	14.86	0.31	0.97	1.91	3.04	0.12	0.38	0.74	1.18	0.89	0.86	0.88	0.85
Seq2seq-dropout	1.26	4.41	9.24	15.46	0.23	0.84	1.82	3.13	0.09	0.33	0.71	1.21	0.92	0.88	0.88	0.85
Seq2seq-dropout-SPL	**1.26**	4.26	8.67	14.23	0.23	0.81	1.74	2.96	0.09	0.32	0.68	1.16	0.92	0.89	0.89	0.86
QuaterNet [25]*	1.49	4.70	9.16	14.54	0.26	0.89	1.83	3.00	0.10	0.34	0.71	1.18	0.90	0.87	0.88	0.85
QuaterNet-SPL	1.34	4.25	8.39	13.43	0.25	0.83	1.71	2.83	0.09	0.32	0.67	1.10	0.91	0.88	0.89	0.86
RNN	1.69	5.23	10.18	16.29	0.31	1.05	2.17	3.62	0.12	0.41	0.85	1.43	0.89	0.85	0.86	0.83
RNN-SPL	1.33	**4.13**	**8.03**	**12.84**	**0.22**	**0.73**	**1.51**	**2.51**	**0.08**	**0.28**	**0.57**	**0.96**	**0.93**	**0.90**	**0.90**	**0.88**

Even a single layer RNN could outperform state-of-art methods on the large and diverse dataset.


Ground-Truth
RNN
RNN-SPL
t = 0 sec.
Seed Motion
t = 2.0 sec.
Short-Term Prediction
t = 2.4 sec.
Long-Term Prediction

Thanks!