Towards Natural and Accurate Future Pose Prediction for Human and Animals

Shuang Wu *
SCSE, Nanyang Technological University
wushuang@bii.a-star.edu.sg

Zhenguang Liu *
SCIE, Zhejiang Gongshang University
liuzhenguang2008@gmail.com

Shuyuan Jin
CEG, National University of Singapore
shuyuan_jin@u.nus.edu

Qi Liu
SoC, National University of Singapore
leuchine@gmail.com

Shijian Lu
SCSE, Nanyang Technological University
shijian.lu@ntu.edu.sg

Roger Zimmermann
SoC, National University of Singapore
textttrogerz@comp.nus.edu.sg

Li Cheng
BII, A*STAR Singapore
chengli@bii.a-star.edu.sg

Abstract

This work addresses the problem of future 3D pose prediction for articulate objects given their observed skeleton sequence. Current methods represent the skeletons of articulate objects as a set of 3D joints, which ignore the relationship between joints and fails to encode fine-grained anatomical constraints. Moreover, conventional Recurrent Neural Networks (RNNs) are employed to model motion context, which inherently have difficulty in capturing long-term dependencies. To address these problems, we propose a Lie algebra representation that models the skeleton as a whole and encodes anatomical constraints explicitly. In addition, a novel RNN structure is designed for motion context modeling, capturing both local contexts for individual frames and global contexts for entire sequences. We explore the possibility to apply our method to a range of objects including human, fish, and mouse. Extensive experiments show that our approach achieves superior pose predictions over the state-of-the-art methods.

1 Introduction

Anticipating the movements of articulate objects, especially human and animals, is crucial for a machine to adjust its behavior, plan its action, and properly allocate its attention when interacting with humans and animals. Natural and accurate future motion prediction is also extremely valuable for a wide range of applications including high-fidelity animal simulation in games and movies, human and animal tracking, human-machine interaction, and intelligent driving [2, 7, 6, 8].

In this work, we concern the specific problem of predicting future 3D poses of an articulate object based on its past skeleton sequences. The problem is challenging due to the non-linear dynamics, high dimensionality, and stochastic nature of human or animal movements. Recently, a family of methods

*Denotes equal contribution

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

based on recurrent neural networks (RNNs) have attracted increasing interests due to their superior performance. Following the released implementations of existing methods, we have empirically observed that current methods often have significant difficulty in obtaining *natural* and *accurate* future motion prediction. For long-term future prediction, existing methods tend to degrade into motionless states or drift away to non-human like motion.

We believe these issues are mainly due to the following two reasons. *First*, current algorithms do not respect the hierarchical nature of the skeletal anatomy. This often leads to strange distortions in the predicted skeleton. *Second*, in temporal motion dynamic modeling, current approaches rely on conventional recurrent units, such as LSTM and GRU, where the hidden state sequentially reads a frame and updates its value. This hidden state tends to be overwhelmed by the input at the current time step and such recurrent units are known to have difficulty in capturing long-term dependencies [1]. Besides, the state must be updated frame by frame, this inherently limits these methods to be computationally non-parallel.

2 Our Approach

To tackle these problems, we propose an approach that consists of two key components: 1) an unified Lie algebra representation formalism, and 2) hierarchical motion recurrent network (HMR).

Specifically, we develop a unified Lie algebra representation for articulate objects, which follows the kinematic structure of the body and explicitly encodes the actual DoFs of individual joints and geometric constraints. The DoFs are encoded as $\mathfrak{se}(3)$ Lie algebra parameters. The matrix exponential maps $\mathfrak{se}(3)$ parameters to 3D rigid transformations that relate the local coordinate systems defined along successive joints in the kinematic chain.

A temporal sequence of the Lie algebra parameters is then collectively fed into the proposed HMR network to encode dynamic evolution of poses. The future pose prediction problem can now be formulated as follows: Given as input the sequence of Lie-algebra parametrized poses, $\langle \mathbf{p}_1, \cdots, \mathbf{p}_t \rangle$, generate predictions for $\langle \mathbf{p}_{t+1}, \cdots, \mathbf{p}_{t+T} \rangle$.

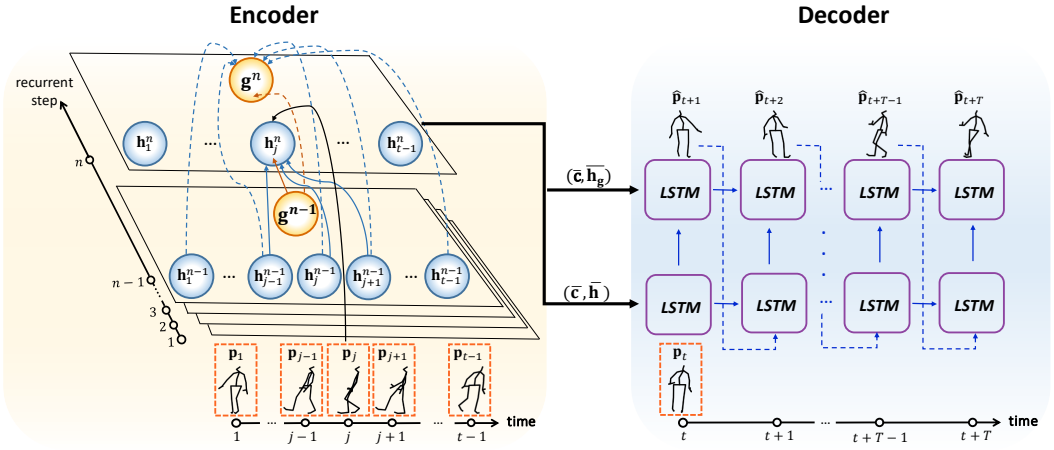

Figure 1: Our HMR model consists of a hybrid encoder-decoder network. The HMR encoder is unfolded over recurrent steps while the decoder is a two-layer stacked LSTM network.

2.1 Hierarchical Motion Context Modeling

To effectively model the dynamics of the entire input sequence, a Hierarchical Motion Recurrent (HMR) network is proposed as the encoder, where the entire input sequence of poses is fed one-shot instead of successively. Fig. 1 illustrates the HMR encoder and stacked LSTM decoder in our network. The motion contexts are modeled by $t-1$ frame-level hidden and cell states, $\{(\mathbf{h}_j, \mathbf{c}_j)\}_{j=1}^{t-1}$, each for one individual frame, as well as a global state, $(\mathbf{g}, \mathbf{c}_g)$. $(\mathbf{h}_j, \mathbf{c}_j)$ and $(\mathbf{g}, \mathbf{c}_g)$ capture frame level (i.e. local) and sequence level (i.e. global) motion contexts, respectively.

The state transition equations of the HMR encoder are formulated in Fig. 2.

Update of frame-level state $\mathbf{h}_j^n$ $\mathbf{h}_j^n$ is updated by exchanging information with its neighboring frames and with $\mathbf{g}^n$. There are a total of 4 types of *forget* gates: $\mathbf{f}^n$, $\mathbf{l}^n$, $\mathbf{r}^n$, and $\mathbf{q}^n$ (forward, left, right, and global forget gates), which respectively control the information flows from the current cell state $\mathbf{c}_j^{n-1}$, left cell state $\mathbf{c}_{j-1}^{n-1}$, right cell state $\mathbf{c}_{j+1}^{n-1}$, and global cell state $\mathbf{c}_g^{n-1}$ to the final cell state $\mathbf{c}_j^n$. The *input* gate $\mathbf{i}^n$ controls the information flow from the pose input $\mathbf{p}_j$. Finally, $\mathbf{h}_j^n$ is updated by a Hadamard product of the *output* gate $\mathbf{o}_j^n$ with the $\tanh$ activated cell state $\mathbf{c}_j^n$.

Update of sequence-level state $\mathbf{g}^n$ $\tilde{\mathbf{f}}_g^n$ and $\tilde{\mathbf{f}}_j^n$ are the respective *forget* gates that filter information from $\mathbf{c}_g^{n-1}$ and $\mathbf{c}_j^{n-1}$ to global cell state $\mathbf{c}_g^n$. The global state $\tilde{\mathbf{g}}^n$ at recurrent step n is updated by a Hadamard product of the *output* gate $\tilde{\mathbf{o}}_j^n$ with the $\tanh$ activated $\mathbf{c}_g^n$.

A two-level representation of the entire input sequence is learned at the final recurrent step and passed to the decoder. This, together with the input pose $\mathbf{p}_t$, produces the pose prediction $\hat{\mathbf{p}}_{t+1}$. $\hat{\mathbf{p}}_{t+1}$ is then fed back as input to the subsequent cell of the decoder to recursively generate subsequent predictions.

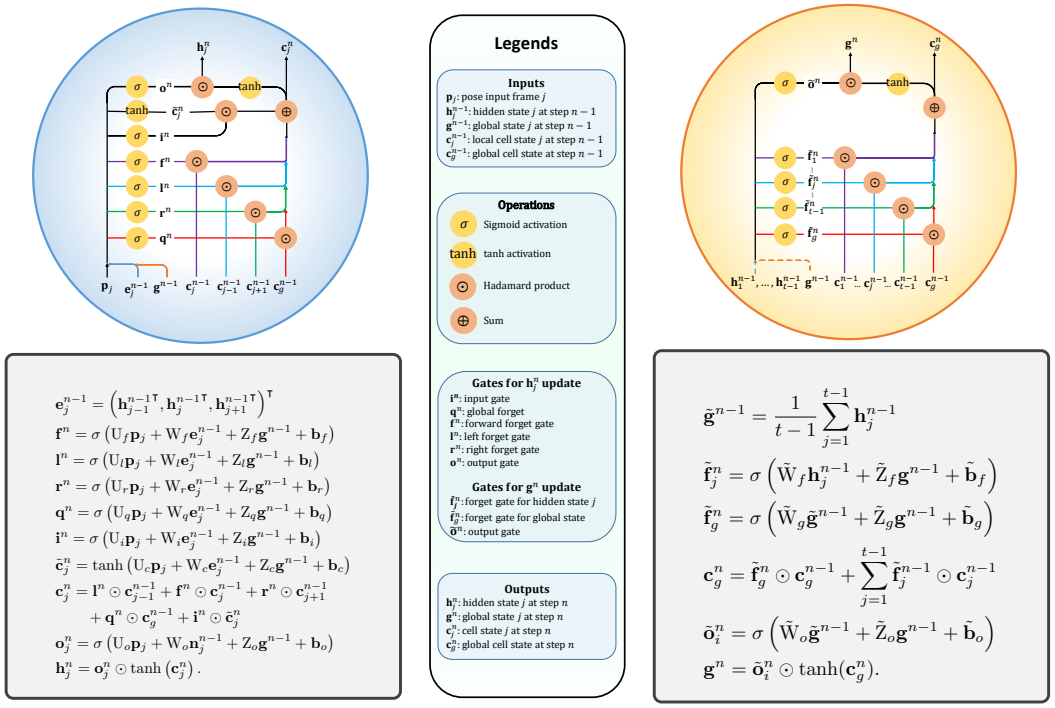

Figure 2: Recurrent update of the frame-level state $\mathbf{h}_j^n$ and sequence level-state $\mathbf{g}^n$.

3 Experiments

Datasets State-of-the-art systems typically focus on humans while animals are rarely studied. This motivates us to consider a more principled approach to address motion prediction across object categories. Experiments are conducted on three datasets of distinct articulate objects, namely human, fish, and mouse. For human, the 3D human full-body motion dataset H3.6m [4] is used. H3.6m contains 3.6 million 3D human poses with 15 activities performed by 7 subjects. For animals, we consider the fish and mouse datasets of [9].

Quantitative Evaluation We first benchmark our HMR against state-of-the-art methods on the H3.6m dataset of [4], employing the same mean angle error (MAE) metric as [5]. Training is performed over all activities with an input window size of $t = 50$ frames and training output window size of $T = 10$ frames. In Table 1, the performance of different methods are reported in terms of MAE for 4 complex activities from H3.6m dataset, namely "Discussion", "Greeting", "Posing" and "Walking Dog". Our HMR network delivers state of the art results for short-term prediction on complex activities of the H3.6m dataset. Evaluation on fish and mouse datasets are performed with the same protocol and reported in Table 2.

Long Term Forecasting It is unrealistic to expect accurate forecasting in the long-term and a more reasonable goal is to achieve human-like motion. Exemplar visual results for long-term forecasting on the walking activity are illusttrated in Fig. 3 for a forecasting window of 50 seconds (1,250 frames). A total of 5 methods are compared, including ERD [3], LSTM-3LR [3], Res-GRU [6], and XYZ (a baseline method employing 3 stacked LSTM layers as the encoder and uses raw 3D joint coordinates instead of Lie algebra parameters as inputs), in addition to our HMR. Here, training is done over a single activity type for a longer training output window size of $T = 100$ frames. The competing methods demonstrate various types of deficiencies: LSTM-3LR converges to a motionless state within 1 sec; ERD exhibits jittery (nonsmooth) and unrealistic motion; Res-GRU converges to a motionless pose after 5 sec. XYZ yields good short term predictions but suffers from bone length deformation, leading to horrendous predictions in the long term. HMR is capable of producing natural pose predictions throughout the entire forecast window. In this regard, an important highlight of our architecture is the capability to generate long-term natural motion.

Methods	Discussion								Greeting							
	80ms	160ms	320ms	400ms	560ms	640ms	720ms	1000ms	80ms	160ms	320ms	400ms	560ms	640ms	720ms	1000ms
ERD ([3])	2.22	2.38	2.58	2.69	2.89	2.93	2.94	3.11	1.70	2.04	2.60	2.81	3.29	3.47	3.55	3.43
LSTM-3LR ([3])	1.80	2.00	2.13	2.13	2.29	2.32	2.36	2.44	0.93	1.51	2.27	2.54	2.97	3.05	3.12	3.09
SRNN ([5])	1.16	1.40	1.75	1.85	2.06	2.07	2.08	2.19	1.33	1.60	1.83	1.98	2.27	2.28	2.30	2.31
Res-GRU ([6])	0.31	0.69	1.03	1.12	1.52	1.61	1.70	1.87	**0.52**	0.86	1.30	1.47	1.78	1.75	1.82	1.96
Zero-velocity ([6])	0.31	0.67	0.97	1.04	1.41	1.56	1.71	1.96	0.54	0.89	1.30	1.49	1.79	1.74	1.77	1.80
MHU ([7])	0.31	0.66	0.93	1.00	1.37	1.51	1.66	1.88	0.54	0.87	1.27	1.45	1.75	1.71	1.74	1.87
HMR (Ours)	**0.29**	**0.55**	**0.83**	**0.94**	**1.35**	**1.49**	**1.61**	**1.72**	**0.52**	**0.85**	**1.25**	**1.40**	**1.65**	**1.62**	**1.67**	**1.73**
Methods	Posing								Walking Dog							
	80ms	160ms	320ms	400ms	560ms	640ms	720ms	1000ms	80ms	160ms	320ms	400ms	560ms	640ms	720ms	1000ms
ERD ([3])	2.42	2.77	3.26	3.39	3.43	3.42	3.45	3.87	1.58	1.78	2.02	2.10	2.31	2.37	2.48	2.60
LSTM-3LR ([3])	1.22	1.89	3.02	3.53	4.25	4.57	4.83	4.60	0.76	1.29	1.91	2.18	2.72	3.01	3.30	3.78
SRNN ([5])	1.74	1.89	2.23	2.43	2.67	2.73	2.79	3.42	1.57	1.73	1.93	1.96	2.13	2.17	2.23	2.20
Res-GRU ([6])	0.41	0.84	1.53	1.81	2.06	2.21	2.24	2.53	0.56	0.95	1.33	1.48	1.78	1.81	1.88	1.96
Zero-velocity ([6])	0.28	0.57	1.13	**1.37**	1.81	2.14	2.23	2.78	0.60	0.98	1.36	1.50	1.74	1.80	1.87	1.96
MHU ([7])	0.33	0.64	1.22	1.47	1.82	2.11	2.17	**2.51**	0.56	0.88	1.21	1.37	1.67	1.72	1.81	1.90
HMR (Ours)	**0.24**	**0.53**	**1.12**	1.42	**1.75**	**1.89**	**2.02**	**2.50**	**0.55**	**0.87**	**1.20**	**1.36**	**1.65**	**1.70**	**1.77**	**1.84**

Table 1: Short-term performance of different methods over 4 different activity types of the H3.6m dataset.

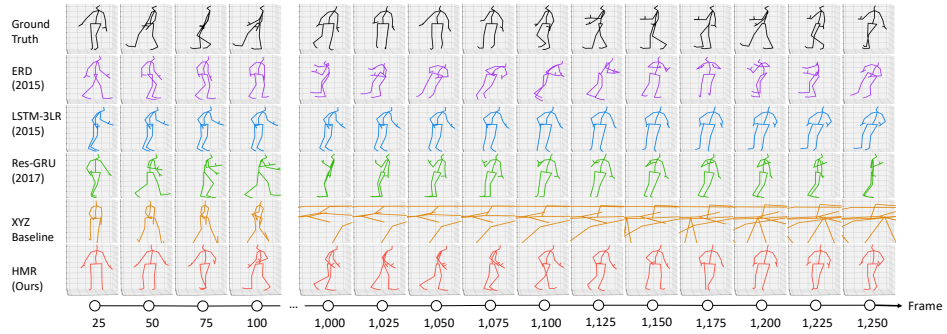

Figure 3: Long-term motion forecasting of walking activity by the comparison methods on the H3.6m dataset.

Methods	Fish								Mouse							
	80ms	160ms	320ms	400ms	560ms	640ms	720ms	1000ms	80ms	160ms	320ms	400ms	560ms	640ms	720ms	1000ms
ERD [3]	0.62	0.59	0.54	0.69	0.79	0.85	0.87	1.20	0.77	0.62	0.67	0.77	0.86	0.83	0.88	0.91
LSTM-3LR [3]	0.91	0.59	**0.42**	0.39	0.25	0.26	0.30	0.29	0.68	0.61	0.81	0.84	0.85	0.81	0.85	0.80
Res-GRU [6]	0.52	0.56	0.52	0.39	0.26	0.25	0.26	0.26	0.40	0.48	0.66	0.70	0.74	0.69	0.72	0.71
HMR (ours)	**0.40**	**0.48**	0.44	**0.28**	**0.13**	**0.12**	**0.13**	**0.11**	**0.39**	**0.44**	**0.56**	**0.63**	**0.69**	**0.68**	**0.67**	**0.70**

Table 2: Performance evaluation (in MAE) of the comparison methods for the Fish and Mouse datasets of [9].

4 Conclusion

For future pose prediction, this work considers a Hierarchical Motion Recurrent network, which is based on Lie algebra representation that naturally preserves the skeletal articulation of the underlying objects. Results on human and animal datasets demonstrate the competency of our approach in terms of both short-term and long-term motion predictions. Future work includes further investigation into group-level motion predictions as well as conditional and unconditional motion synthesis.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

[2] J. Butepage, M. Black, D. Kragic, and H. Kjellstrom. Deep representation learning for human motion prediction and classification. In *CVPR*, 2017.

[3] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik. Recurrent network models for human dynamics. In *ICCV*, 2015.

[4] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 36(7):1325–39, 2014.

[5] Ashesh Jain, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *CVPR*, pages 5308–5317, 2016.

[6] J. Martinez, M. Black, and J. Romero. On human motion prediction using recurrent neural networks. In *CVPR*, 2017.

[7] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamic. *CoRR*, abs/1805.02513, 2018.

[8] Jianwen Xie, Song-Chun Zhu, and Ying Nian Wu. Synthesizing dynamic patterns by spatial-temporal generative convnet. In *CVPR*, pages 1061–1069, 2017.

[9] C. Xu, L. Govindarajan, Y. Zhang, and L. Cheng. Lie-X: Depth image based articulated object pose estimation, tracking, and action recognition on Lie groups. *IJCV*, 123(3):454–78, 2017.