Deep representation learning for human motion prediction and classification

Judith Bütepage, Michael J. Black, Danica Kragic and Hedvig Kjellström

butepage@kth.se, black@tuebingen.mpg.de, dani@kth.se, hedvig@kth.se

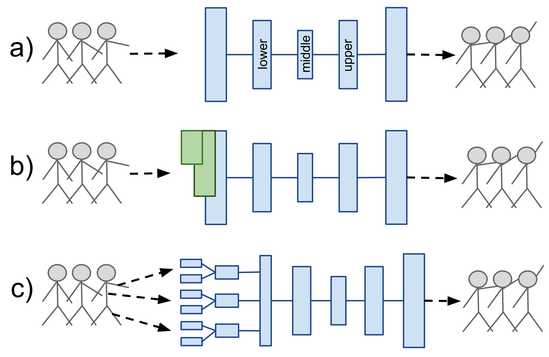
IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR July 21-26 2017 Honolulu

Motivation

Generative models of 3D human motion are often restricted to a small number of activities and can therefore not generalize well to novel movements or applications. In this work we propose a deep learning framework for human motion capture data that learns a generic representation from a large corpus of motion capture data and generalizes well to new, unseen, motions. Using an encoding-decoding network that learns to predict future 3D poses from the most recent past, we extract a feature representation of human motion. Most work on deep learning for sequence prediction focuses on video and speech. Since skeletal data has a different structure, we present and evaluate different network architectures that make different assumptions about time dependencies and limb correlations. To quantify the learned features, we use the output of different layers for action classification and visualize the receptive fields of the network units. Our method outperforms the recent state of the art in skeletal motion prediction even though these use action specific training data.

Models



a) symmetric temporal encoder (S-TE)

b) convolutional temporal encoder (C-TE)

c) hierarchical temporal encoder (H-TE)

We train the model on a large portion of the CMU mocap dataset, producing a generic representation.

Acknowledgement

This work was partly supported by the EU through the project socSMCs (H2020-FETPROACT-2014) and Swedish Foundation for Strategic Research.

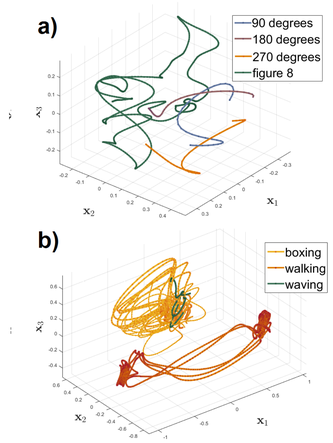
References

[1] Fragkiadaki, Katerina and Levine, Sergey and Felsen, Panna and Malik, Jitendras. *Recurrent network models for human dynamics.* IEEE International Conference on Computer Vision (2015): 346–4354.

[2] Jain, Ashesh and Zamir, Amir R and Savarese, Silvio and Saxena, Ashutosh. *Structural-RNN: Deep Learning on Spatio-Temporal Graphs.* IEEE Conference on Computer Vision and Pattern Recognition (2016): 5308–5317.

[3] Liu, Hailong and Taniguchi, Tadahiro. *Feature extraction and pattern recognition for human motion by a deep sparse autoencoder.* IEEE International Conference on Computer and Information Technology (2014): 173–181.

Representational manifold



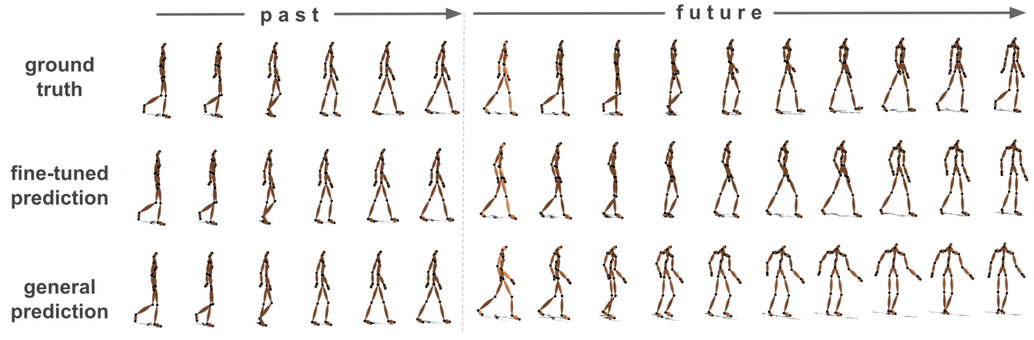
Projecting the activity of the units onto three dimensions reveals a low-dimensional representation of human motion dynamics.

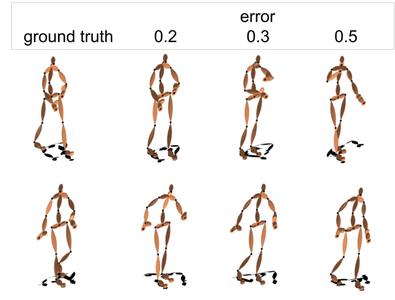
Classification

Method	Classification rate		
Data (1.6 s)	0.76		
PCA	0.73		
	Lower Layer	Middle Layer	Upper Layer
DSAE [3]	0.72	0.65	0.62
S-TE	**0.78**	0.74	0.67
C-TE	**0.78**	0.74	0.73
H-TE	0.77	0.73	0.69

We classify the actions *walk, run, punching, boxing, jump, shake hands, laugh and drink* based on the output of units of different layers. Deep sparse autoencoders (DSAE) reconstruct the input data and do not predict.

Prediction





Method	Short Term			Long Term		
	80ms	160ms	320ms	560ms	1000ms	1600ms
H3.6M (average over four action-specific models)						
ERD [1]	0.39	0.44	0.53	0.6	0.67	0.7
S-RNN [2]	0.34	0.38	0.45	0.52	0.59	0.63
LSTM3L	0.27	0.33	0.42	0.49	0.57	0.62
H3.6M						
S-TE	0.36	0.37	0.37	0.38	0.43	0.45
C-TE	**0.21**	0.23	0.27	0.34	0.42	0.45
H-TE	**0.21**	**0.22**	**0.25**	**0.3**	**0.36**	**0.39**
CMU						
S-TE	0.3	0.31	0.33	0.35	0.37	0.37
C-TE	**0.18**	0.21	0.25	0.30	0.33	0.35
H-TE	**0.18**	**0.20**	**0.24**	**0.28**	**0.31**	**0.33**

Compared to recurrent, action specific models, our approach is able to generalize to unseen action types.
Even for long-term predictions, the motion prediction error remains smaller.

Conclusion

In this work we develop an unsupervised representation learning scheme for long-term prediction of everyday human motion that is not confined to a small set of actions. We demonstrate that our learned low-dimensional representation can be used for action classification and that we outperform more complex deep learning models in terms of motion prediction. Our approach can be viewed as a generative model that has low computational complexity once trained, which makes it suitable for online tasks.