Towards Context-Preserving Human to Robot Motion Mapping

Taras Kucherenko Hedvig Kjellström
Department of Robotics, Perception and Learning,
KTH Royal Institute of Technology, Sweden
{tarask,hedvig}@kth.se

***Abstract*—We aim to learn a mapping of human communicative behavior to a robot with fewer degrees of freedom, which would preserve its high-level characteristics, such as attitude and intention. We take a deep learning approach, using the encoding-decoding idea. Our preliminary results of the human motion representation learning indicate that we can benefit from having recurrent connections only in some layers of the neural network. The best-performing representation size is roughly two times less than the original input data dimensionality.**

I. Introduction

A step towards robot autonomy is telepresence, where a robot is steered by a human, e.g. representing the human in a meeting. A central problem is then motion mapping. Mapping from human to robot motion is not a straightforward task, because most robots have fewer degrees of freedom (DoF), as well as other physical limitations.

Most of the work in this area is restricted to the mapping on an activity level [1], [2]. In contrast, we aim to transmit the "content" of the motion and its high-level characteristics, e.g. attitude and intention.

We propose a deep learning approach, in which we map the human embodiment, via the abstract representation of the motion, to the robot embodiment, as shown in Fig.1.

Our system includes encoding and decoding of the motion, which resembles the idea of autoencoders (AE). Hence, we choose AE as a network architecture. In particular, we use Stacked Denoising AE [3] with LSTM [4] for recurrency, as described in Section III. We are currently collecting training data for human-to-robot maping. In this paper we report on initial experiments with the CMU Mocap dataset, where we evaluate different design choices. We aim to design an architecture with the most generalizable representation. So we experiment with two important decisions, as described in Section IV:

1) Which layers of the network should be recurrent and which fully-connected?
2) What is the effect of the size of the middle layer?

II. Related work

An early work on deep representation learning for human motion had no recurrency in their networks. Lie and Taniguchi [5] have used sliding window over time-series and random forest on low-dimensional representation in order to classify the motion. They showed that reducing dimensionality of the human motion data drastically improves classification performance and that deep sparse autoencoder (SAE) can extract better low-dimensional representations than PCA or a shallow SAE. Recently, this approach was extended [6] to prediction. It outperformed previous work and yielded better generalization. In contrast to these papers, we are processing the motion frame by frame and use an LSTM to store relevant information about the past frames.

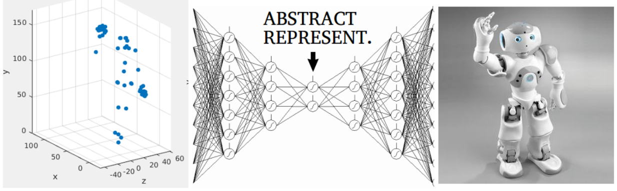


Fig. 1: Illustration of our approach to mapping from human motion to robot motion: use latent representation, abstracted from the embodiment. To the left we show a human pose obtained through 3D motion capture. To the right is a NAO robot: humanoid with much fewer DoF.

Plenty of research has been done on using recurrent neural networks for representation learning as well. One approach [7] was to use recurrent connections at each layer. To reduce the number of parameters and avoid overfitting, an LSTM was used only at the last layer, and a standard RNN in all the other layers. An alternative approach [8] used encoding-recurrent-decoding network, where recurrency was applied only at the middle layer. They was focused on motion prediction or classification, while we are aiming for human-to-robot mapping.

To the best of our knowledge, there is a lack of systematic analysis of the effect of adding recurrent connections and the best place to add such connections. This analysis is the main topic of our experiments.

III. Deep Network Architecture

We use a deep neural network, which has 5 hidden layers (Figure 2b) with 100 neurons in each, combined as a Stacked Denoising Autoencoder (SDAE) [3]. This network is designed for unsupervised representation learning and is regularized by adding noise to the input data and learning to reconstruct the noise-free input. We modify the original SDAE by adding a recurrency (LSTM), unrolled over 64 time-steps.

Training was done using Adam optimizer [9] and layer-wise pretraining. Error measure was an L2 reconstruction loss averaged over all the test examples. Early stopping

TABLE I: recurrency analysis

Recurrency	Train error	Test error
Everywhere	0.0063	0.096
No	0.0088	0.196
1st layer	0.0065	**0.021**
2st layer	0.0057	0.033
3st layer	0.0051	0.030
4st layer	0.0050	0.030
5st layer	0.0091	**0.021**

was used: the training was terminated, if the validation error droped by 8% from the best accuracy so far.

IV. Preliminary Experimental Results

We use CMU Mocap dataset: mean pose was substracted, the rotation and translation of all joints was mapped to the [-1,1] interval. Global coordinates were ignored, since they contain no redundancy.

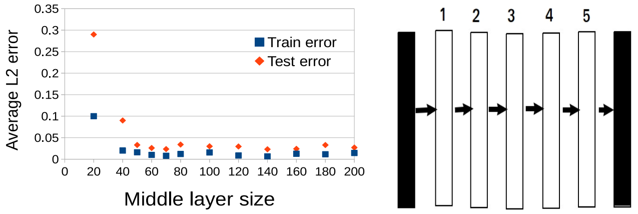
Our objective is to learn a representation with a good generalization to new types of motions.[1] In order to evaluate that we test our network on a type of motion not included in the training examples. We use the following motions for training : *kid on a playground, dance, kicking a ball, basketball dribble* and *running*. We use *jumping* for validation and *boxing* for testing.

A. Analysis of the amount of recurrency in the network

In the following experiment we test if recurrency is needed at every layer and at which layer are the recurrent connections the most important. We compare DAE without recurrency, with recurrency at each layer and with recurrency only at a particular layer. While other hyper-parameters are kept fixed, the learning rate is optimized by grid search for every architecture. Results in Table I clearly show that recurrency helps, but the network overfits to the training data when recurrency is added at each layer. The network performs best with recurrency added only to the first or the last layer. Over-fitting of fully-recurrent network likely happens because an LSTM layer has many more parameters than a FC layer and is harder to train when combined in a deep network.

B. Optimal middle layer size

In the following experiment we evaluate the choice of dimensionality of the representation. Keeping all the other parameters and architecture structure fixed, we vary only the middle (3^{rd}) layer size. Based on the previous experiment, we have chosen a network with the 1^{st} layer being LSTM and all the rest being FC. Fig. 2a shows that having a bottleneck in the network indeed decreases both train and test error. The best performance was achieved, when the representation size is 70, while the input data has 126 degrees of freedom. For a different task, such as mapping from human to robot, the best representation size may be different.



(a) Effect of the middle layer size. (b) Network architecture.

Fig. 2: Architecture analysis

V. Conclusions

This paper is a first step towards a method for human-to-robot mapping of communicative motion that preserves the essential high-level characteristics of the motion. We present initial analysis of design considerations for an architecture with the best generalization capability. We observe that it may be beneficial to have recurrent connections only in some layers of the network and that best representation size may be on the order of original dimensionality.

Acknowledgments

The authors would like to thank Rika Antonova, Vladimir Li, and Judith Bütepage for useful discussions and comments. The data used in this project was obtained from mocap.cs.cmu.edu. This work is funded by Stiftelsen för Strategisk Forskning.

References

[1] M. V. Liarokapis, P. Artemiadis, C. Bechlioulis, and K. Kyriakopoulos, "Directions, methods and metrics for mapping human to robot motion with functional anthropomorphism: A review," *School of Mechanical Engineering, National Technical University of Athens, Tech. Rep*, 2013.

[2] H. Kjellström, J. Romero, and D. Kragic, "Visual recognition of grasps for human-to-robot mapping," in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on.* IEEE Computer Society, 2008, pp. 3192–3199.

[3] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–1780, 1997.

[5] H. Liu and T. Taniguchi, "Feature extraction and pattern recognition for human motion by a deep sparse autoencoder," in *Computer and Information Technology (CIT), 2014 IEEE International Conference on.* IEEE Computer Society, 2014, pp. 173–181.

[6] J. Bütepage, M. Black, D. Kragic, and H. Kjellström, "Deep representation learning for human motion prediction and classification," *arXiv preprint arXiv:1702.07486*, 2017.

[7] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[8] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," 2015.

[9] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[1] In the human-to-robot mapping scenario, the method will instead generalize from human communicative motion to the corresponding robot motion