Anticipating Human Intention for Full-Body Motion Prediction

Philipp Kratzer[1], Niteesh Balachandra Midlagajni[2] and Jim Mainprice[1]
Machine Learning and Robotics Lab, University of Stuttgart, Germany
[1]firstname.lastname@ipvs.uni-stuttgart.de, [2]niteesh.m.92@gmail.com

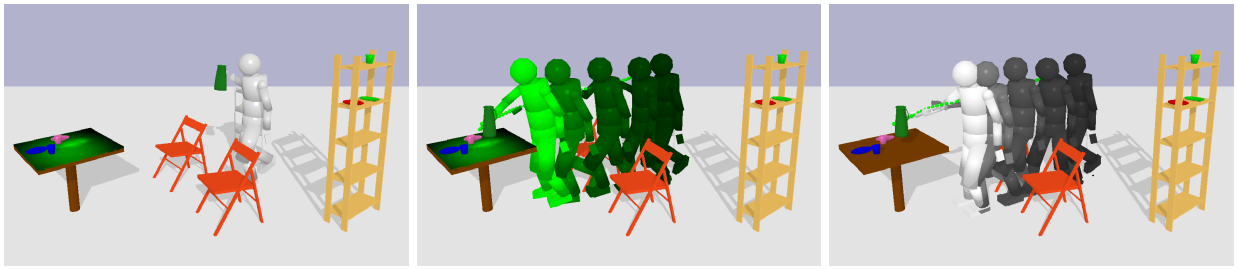

Fig. 1: Prediction for placing a jug on the table. a placement affordance is predicted as a probability density function on the table (a), depicted in green a full-body motion is optimized (b), which is compared to ground truth motion (c).

***Abstract*— Motion prediction in unstructured environments is a difficult problem and is essential for safe and efficient human-robot space sharing and collaboration. We propose an algorithmic framework that accounts explicitly for the environment geometry based on a model of affordances and a model of short-term human dynamics both trained on motion capture data. We perform experiments on place trajectories and show that we achieve similar performance for full-body motion predictions as using oracle place locations.**

I. INTRODUCTION

When interacting with their environment, humans model the action possibilities directly in the product space of their own capabilities and the environment. This idea of the existence of an intuitive and perceptual representation of the possibilities in an environment is known as affordances [1], [2]. In Robotics, affordances can be used to model the actions a robot is able to perform [3], [4], [5]. For example, Koppula and Saxena define object affordance as potential functions depending upon how the object will be interacted with [6].

In prior work graphical models, such as Hidden Markov Models (HMM) and Conditional Random Fields (CRF), have been used in order to predict human motion or intention. [7], [8], [9]. While these approaches are sound they generally do not scale to large databases of motion capture or are limited to predict 2d motion of humans and do not deal with the full-body case.

In this paper, we propose a neural network framework to learn and encode affordances from data. We focus on placeability prediction and model it as probability density functions conditioned on the environment and the kinematic state of the human. We combine this intention with a full-body motion prediction system [10] to produce accurate predictions as seen in Fig. 1.

We gathered a dataset with 5 participants using a motion capture system. Affordances and short-term motion models were trained on this dataset. Our results demonstrate superiority of our affordance densities for predicting placement locations. Finally, we show that combining goalset predictions and motion predictions compares similarly to using oracle goal locations.

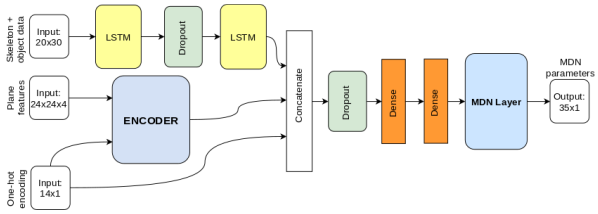

Fig. 2: Placeability network architecture

II. COMBINED INTENTION AND FULL-BODY PREDICTION

The framework works as follows: Offline we train the short-term prediction and the placeability affordance model. Online we use the affordance model to predict a density over place locations. We extract the the maximum likelihood place location and optimize the short-term prediction to end at the extracted location. Thus, we receive a full-body trajectory towards the place location.

A. Placeability Affordance

We define the placeability affordance as a probability distribution over possible place locations on a surface. Figure 2 depicts the network architecture. The inputs to the model are the human skeleton and object states in positions over a trajectory of 1sec (20 timesteps), a 14 dimensional one-hot encoding of object type and surface we compute the affordance for, and a grid that covers the plane state (occupancy of objects).

a) Multi-modal placements: Placeability is fundamentally multi-modal. For instance in our experiments we consider a table setting scenario such as found in a home or restaurant, four people can sit next to the table, therefore there are four possible locations where the human can place a plate.

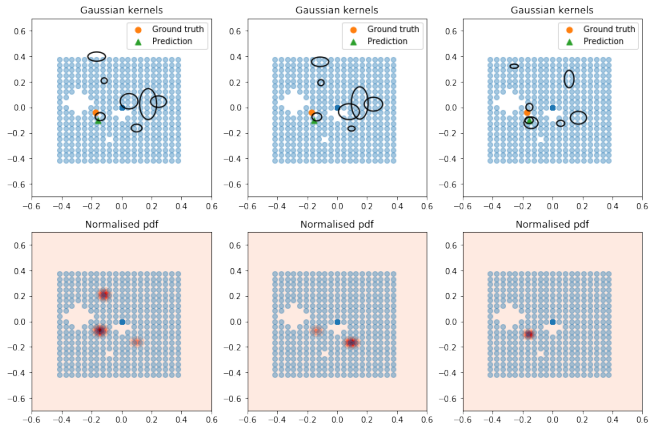

Fig. 3: MDN predictions at 4s, 1s and 0.5s respectively. The top images show all 7 Gaussian kernels, though most of them have low probability as verified by the density images below.

A standard approach to model multi-modal distributions are Mixture Density Networks (MDN) [11], which we make use of for modeling placement distributions:

$$p(x|d) = \sum_{i=1}^{m} \alpha_i \phi_i(x|d) \tag{1}$$

where m indicates the count of the components in the mixture model, α_i are the mixing coefficients. ϕ_i are functions representing conditional densities for the i^{th} kernel.

We use multivariate Gaussian kernels with diagonal covariance. We use 7 kernels in output, which gave good empirical results on our dataset. The network is trained using a neg-log likelihood (NLL) loss with the 2d place position on the surface as ground truth.

b) Constraining affordances to free regions: We improve our placeability model with the intention of making it more robust against violating regions where objects are already placed. This is done by building an autoencoder network with convolutional layers that is trained to output a binary occupancy map of the plane after the object is placed. The encoder is used to generate distinctive features for the placeability network (see Figure 2).

B. Full-Body Prediction

We aim to find a trajectory of human motion $h_{t+1:T}$, given a trajectory $h_{0:t}$ of already observed states. For this purpose we use a trajectory prediction framework introduced in [10]. The framework has 2 phases: 1) Offline, a VRED model f [12] is trained to predict purely kinematic trajectories $f(h_{0:t}, \delta) = h_{t+1:T}$. 2) Online, trajectory optimization techniques are used to adapt to environmental objectives. This is done by changing additional controls δ that are added to the VRED architecture. In this paper we use the *goalset* objective:

$$c_{\text{goalset}}(\delta) = \|\phi_{\text{FK}}(f(h_{0:t}, \delta)_T) - p^*\|^2 \tag{2}$$

which enforces the hand of the human to end up close to position p^*, with ϕ_{FK} being the forward kinematics map, mapping the last human state to the hand position.

TABLE I: Error of state prediction per time step.

ms	250	500	750	1000	1250	1500
Zerovel	2.38	5.18	7.76	9.68	11.09	11.94
VRED	0.88	1.70	2.82	4.01	5.27	6.30
ours	0.86	1.43	2.00	2.42	2.70	2.80
oracle	0.86	1.44	1.84	2.03	2.19	2.18

In order to account for our affordance model, we compute the expected prediction position p^* from the affordance model $P_{o,a}(x|h, s)$. Thus, the trajectory will be optimized to end up at this position.

The gradient based optimization algorithm L-BFGS [13] is used to optimize the trajectory. The gradients are calculated using automatic differntiation functionalities from tensorflow.

III. Results

In our setup a Motion capture system was used to capture full-body human motion. We also track moveable objects, such as cups, and stationary objects, such as tables and shelves. Participants were asked to perform tasks related to setting up the table and clearing it. A total of 5 users participated in the recording session. In total we recorded 120min. We used data of 3 of the subjects for training and 2 for testing.

Figure 3 visualizes the mixture components on a test set example for placing a cup on the table at 3 time instances. It can be seen that when the subject is far away from the table, there are multiple possibilities of potential placeable regions and as the subject moves towards the table, that uncertainty reduces and confines to one dense most likely region.

In order to test the full-body prediction we extract 27 trajectories for placing on the table. Table I shows the distance to the ground truth at different times in the future for our method and several baselines. The sum over distances of key joints (wrists, elbows, knees, ankles and pelvis) is shown. Values are averaged over the 27 trajectories.

The zero velocity baseline just keeps the current state as prediction for future timesteps. The VRED baseline just unrolls the recurrent neural network. Our method takes the affordance prediction into account and optimizes to end up at p^*. The oracle has additional oracle information about the true endposition of the wrist.

It can be seen that the oracle prediction performs best, which is not surprising, as it uses information that is not available at prediction time. Our method using the place point prediction performs second best and outperforms the prediction without any optimization at all time steps.

IV. Conclusions

We presented a system to learn human object affordances for human motion prediction. We demonstrate that the method can be used to predict full-body trajectories.

A user study was conducted to collect a dataset in a motion-capture setup on a table setup task. Our experiments proof that the affordances can be used to improve full-body motion prediction within a state-of-the-art motion prediction framework.

REFERENCES

[1] J. J. Gibson, "The senses considered as perceptual systems." 1966.

[2] J. Gibson, *The Ecological Approach to Visual Perception*, ser. Resources for ecological psychology. Lawrence Erlbaum Associates, 1979.

[3] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: From sensory–motor coordination to imitation," *IEEE Trans. Robotics*, vol. 24, no. 1, pp. 15–26, 2008.

[4] A. Gonalves, G. Saponaro, L. Jamone, and A. Bernardino, "Learning visual affordances of objects and tools through autonomous robot exploration," in *IEEE Int. Conf. on Autonm. Robot Systems and Competitions (ICARSC)*, 2014, pp. 128–133.

[5] A. Dehban, L. Jamone, A. Kampff, and J. Santos-Victor, "Denoising auto-encoders for learning of objects and tools affordances in continuous space," in *IEEE Int. Conf. Robotics And Automation (ICRA)*, 2016, pp. 4866–4871.

[6] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE Trans. on Pattern Analysis and Machine Intell.*, vol. 38, no. 1, pp. 14–29, 2016.

[7] M. Bennewitz, W. Burgard, G. Cielniak, and S. Thrun, "Learning motion patterns of people for compliant robot motion," *The Int. Journal of Robotics Research*, vol. 24, no. 1, pp. 31–48, 2005.

[8] D. Kulić, C. Ott, D. Lee, J. Ishikawa, and Y. Nakamura, "Incremental learning of full body motion primitives and their sequencing through human motion observation," *The Int. Journal of Robotics Research*, vol. 31, no. 3, pp. 330–345, 2012.

[9] J. Elfring, R. Van De Molengraft, and M. Steinbuch, "Learning intentions for improved human motion prediction," *Robotics and Autonm. Systems*, vol. 62, no. 4, pp. 591–602, 2014.

[10] P. Kratzer, M. Toussaint, and J. Mainprice, "Prediction of human full-body movements with motion optimization and recurrent neural networks," in *IEEE Int. Conf. Robotics And Automation (ICRA)*, 2020.

[11] C. M. Bishop, "Mixture density networks," 1994.

[12] H. Wang and J. Feng, "Vred: A position-velocity recurrent encoder-decoder for human motion prediction," *arXiv preprint arXiv:1906.06514*, 2019.

[13] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, 1995.