

# Social Behavior Learning with Realistic Reward Shaping

Yuan Gao<sup>1</sup>, Fangkai Yang<sup>2</sup>, Martin Frisk<sup>1</sup>, Daniel Hernandez<sup>3</sup>, Christopher Peters<sup>2</sup>, Ginevra Castellano<sup>1</sup>

**Abstract**—Deep reinforcement learning has been widely applied in the field of robotics recently to study tasks like locomotion and grasping, but applying it to social robotics remains a challenge. In this paper, we present a deep learning scheme that acquires a prior model of robot behavior in a simulator as a first phase to be further refined through learning from subsequent real-world interactions involving physical robots. The scheme, which we refer to as Staged Social Behavior Learning (SSBL), considers different stages of learning in social scenarios. We implement robot approaching behavior towards small groups based on this scheme and evaluate the performance of different configurations using objective and subjective measures. We found that our model generates more socially-considerate behavior compared to a state-of-the-art model. We also suggest that SSBL could be applied to a wide class of social robotics applications.

## I. INTRODUCTION

Deep reinforcement learning (DRL) algorithms provide a framework for automatic robot perception and control [1] [2]. In recent years, methods based on DRL have achieved great performance in different control tasks like grasping and locomotion [3]. However, the question of how to make robots learn appropriate social behaviors under modern frameworks remains underexplored, partly due to the lack of cross-disciplinary synergies in human-robot interaction studies. As a consequence, the interaction scenarios studied in previous research have been limited to simplified cases and the algorithms studied to relatively simple ones [4].

Our approach is to learn a prior model in a simulator first and then refine the policy using model-based reinforcement learning (RL) in the real world. Learning a prior model in simulator for human-robot interaction (HRI) has a lot of potential benefits. First, it can save a significant amount of real world interactions. Several works [5] [6] have shown that learning a model for physical interactions can help robots learn faster in the real world. Secondly, in social interactions, humans have little tolerance for random behaviors [7], and lose interest quickly if the model deviates too much from social norms. Additionally, the mathematical modeling of social interactions in a simulated setting allows researchers to control factors more rigorously, which can help with the issue of replicability. Finally, learning a prior model can make use of previously-developed models, such as DeepPose [8] or

Emonets [9]. However, unlike simulating physical interactions [10], simulating HRI poses a different set of challenges. One of the main challenges is that it is hard for the simulator to accurately model human relevant behavior. Simulators of physical interactions are based on physical laws which are well understood, while human behavior is less predictable. Nevertheless, two ways have been considered to simulate social feedback based on real world signals. The first one is to use computational models [11] that have been studied in experiments [12] and the second one is to use machine learning methods [13].

In this paper, we consider how to efficiently learn a prior social interaction model in the simulator and propose a Staged Social Behavior Learning (SSBL) scheme for learning appropriate social behavior with continuous actions. Specifically, we consider a task in which the robot moves towards groups based on their simulated social feedback. The task is learned in an end-to-end fashion i.e. from vision to social behaviors in a virtual environment. SSBL involves a pipeline for simulated social robot learning that deconstructs a social task into three steps. In the first step, the robot learns a compressed representation of the world from vision or other modalities. This step is important because it significantly reduces the complexity of the DRL problem. After the compressed information is learned, the algorithm learns a dynamical model in the environment. The last step is to make sure that the learned behavior follows the social standard by using simulated social feedback as realistic reward.

## II. LITERATURE REVIEW

### A. Deep Reinforcement Learning in HRI

Reinforcement learning (RL) has been used since the early days of HRI. One of the few early works that considered using social feedback as accumulative rewards were conducted by Bozinovski [14] [15]. After that, many papers in HRI started to investigate the effect of RL algorithms like Exp3 [16] or Q-learning in social robotics settings. However, since such algorithms lack the ability to capture important features from high-dimensional signals [17], their applicability to solve HRI problems remains limited. After the era of deep learning started in 2006 [18], many different algorithms were proposed to understand different modalities in HRI, for example, ResNet [19] for image processing and Long Short-Term Memory [20] (LSTM)-based solutions for text processing. As a consequence, some HRI researchers started to investigate deep learning's role in the area of HRI. A pioneer work was conducted by Qureshi in 2017 [21]. In this work, a Deep Q-Network (DQN) [17] was used to learn a mapping from visual input to one of several

\*This work is partly supported by the Swedish Research Council (grant n. 2015-04378) and partly by the COIN project (RIT15-0133) funded by the Swedish Foundation for Strategic Research.

<sup>1</sup> Author is with Department of Information Technology, Uppsala University, Uppsala, Sweden. alex.yuan.gao@it.uu.se

<sup>2</sup> Author is with Department of Computational Science and Technology, Royal Institute of Technology, KTH, Stockholm, Sweden.

<sup>3</sup> Author is with Department of Computer Science, University of York, York, United Kingdom.

predefined actions for greeting people. Another work was conducted by Madson [22], where a DQN was used for learning generalized, high-level representations from both visual and auditory signals.

### B. Learning Representations

RL based solely on visual observations has been used to solve complex tasks such as playing ATARI games [23], driving simulated cars [24] and navigating mazes [25]. However, learning policies directly from high-dimensional data such as images requires a large amount of samples, which makes it intractable in social robot learning [26]. One solution is to use low-dimensional hand-crafted features as the state, but this would reduce learning autonomy.

Prior works have utilized deep autoencoders (AEs) to learn a state representation, including Lange et al. [27]. Several variants of AEs have been applied as well, including attempts by Böhmer et al. [26] to learn the dynamics of the environment by constructing an AE predicting the next image, and Finn et al. [28] who adopted a spatial AE (SAE) to learn an intermediate representation consisting of image coordinates of relevant features. The latter suggested that this intermediate representation made it particular well suited for high-dimensional continuous control.

### C. Group Behavior

Numerous works have been done in group dynamic behaviors. Particle-based methods [29] [30] simulate global collective behaviors of large scale groups or crowds. For modeling small scale groups, agent-based methods [31] [32] are adopted to simulate the behavior of each individual based on rules of behavior. Specifically, in a small multi-party conversation group, Kendon [33] proposed the *F-formation* system to define the positions and orientations of individuals within a group, which characterized dynamic group behaviors. Several studies have been carried out that concern robot approaching behaviors towards small groups i.e. in which an agent moves towards a group in an attempt to join an ongoing task or conversation. Ramírez et al. [34] adopted inverse reinforcement learning, involving several participants demonstrating approaching behaviors for a robot to learn. Pedica et al. [35] integrated behavior trees in their reactive method to simulate lifelike social behaviors, including robot approaching behavior towards groups. Both approaching and leaving behaviors are considered in [36], where a finite state machine is utilized in the transitions between different social behaviors. Jan et al. [11] presented an algorithm for simulating movement of agents, such as an agent joining the conversation. The agents dynamically move to new locations, but without proper orientations.

## III. METHODOLOGY

In the following sections, we introduce the fundamental concepts needed to train a prior model for robot approaching behaviors in accordance with SSBL. Section III-A introduces some basic concepts and details on how the environment was set up. Sections III-B to III-D describe the three stages of

SSBL training. Specifically Section III-B pertains to the state representation and its training procedure. It also details the various architectures used to evaluate this step. Section III-C shows how one can formulate the training of a dynamical model within an RL framework. Section III-D shows how social norms can be acquired by utilizing concepts from the Social Force Model (SFM) [33]. In the original work, SFM was also used to generate social robot behaviors, which will be used as a baseline for evaluating our learned policy.

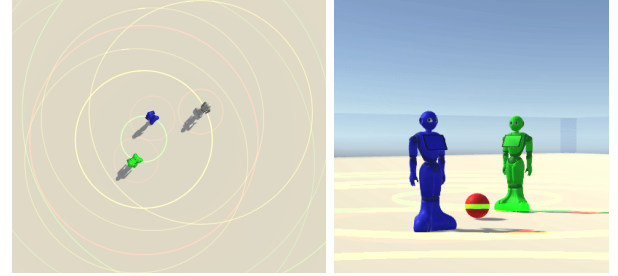


Fig. 1: Two views of the environment setup. The left image shows a top-down view of the environment. The right image shows the first-person view from the robot's perspective. The environment contains two Simulated Human Agents (SHAs: green and blue), and a robot agent (gray). Circles around the human agents and the robot represent different levels of space as discussed in Section III-C.

### A. Environment Setup

In order to simulate robot approaching behaviors, we first build a simulator using *Unity 3D*<sup>1</sup> game engine. The environment consists of a square floor surrounded by four walls. A conversation group which contains two Simulated Human Agents (SHAs) is spawned at a random position within this domain. The robot agent is spawned outside the group and performs approaching behaviors. The virtual agents (agents in the group and robot agent) are pre-defined assets which resemble the SoftBank Pepper robot. Figure 1 shows one example of environment's top-down and first-person view. The blue and green agents are SHAs and the gray one at the top right of the top-down view is the robot agent. The first-person view (Figure 1, right) is from the robot agent's perspective.

In this paper, we are mainly concerned with the task of learning a prior model in the simulator for a robot's approaching behaviors towards small groups of individuals. The task can be formulated as an RL problem. Let us consider  $s_t$  and  $a_t$  as the state and action of the robot agent at time  $t$ , respectively. Learning the dynamic behavior for approaching a group can be viewed as maximizing the expected cumulative reward  $E_{\tau \sim \pi}[\mathcal{R}(\tau)]$  over trajectories  $\tau = \{s_1, a_1, \dots, s_T, a_T\}$ , where  $\mathcal{R}(\tau) = \sum_{t=1}^T \mathcal{R}(s_t, a_t)$  is the cumulative reward over  $\tau$ . The expectation is under distribution  $p(\tau) = p(s_1) \prod_{t=1}^T p(s_{t+1}|s_t, a_t)p(a_t|s_t)$ , where  $\pi(s_t) = p(a_t|s_t)$  is the policy we would like to train

<sup>1</sup><https://unity3d.com/>

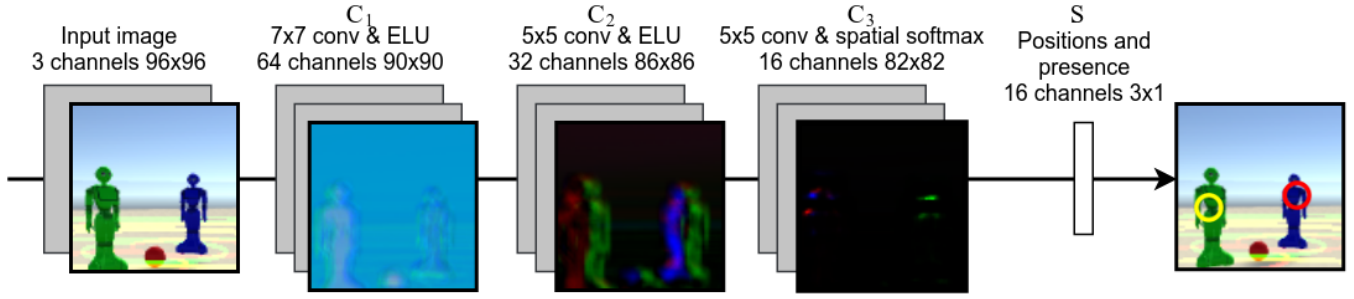


Fig. 2: Schematic view of the encoder used both by the Finn-architecture and ours. The rightmost image is not a part of the network, but a visualization of two of the positions computed in  $S$ . The slight discrepancy in position between the position of the activation peaks in  $C_3$  and the circles in the output-image is due to the convolutions not using any padding. This makes it so that  $C_3$  only represents the center  $82 \times 82$  pixels of the input.

and  $p(s_{t+1}|s_t, \mathbf{a}_t)$  is the forward model determined by the environment.

### B. State representations

In our experiments, we try three modes of representing the environment state to the robot agent: *Vector*, *CameraOnly* and *CameraSpeed*. The first mode is a vector-based representation, consisting of the positions and velocities of all the agents, together with the positions of the walls. This representation is ideal for learning, so it serves as an upper bound on the performance of this task.

The second and third modes are designed to resemble two common robotic settings: one where the robot is equipped with a camera, and one where the robot has both a camera and the ability to estimate its speed. In these modes, the full states are given as  $\mathbf{s}_t = \mathbf{I}_t$  and  $\mathbf{s}_t = (\mathbf{I}_t, \mathbf{v}_t)$  respectively. Here  $\mathbf{I}_t$  is the visual information from the robot's first-person view rendered by the Unity engine, and  $\mathbf{v}_t$  the velocity of the robot.

The method employed in this work is to learn a mapping from input images to simplified low-dimensional state representations, thus circumventing some of the problems associated with RL from high dimensional input [23]. To do this, we utilize an autoencoder (AE) [37], a neural net  $\phi$  that maps inputs to itself, s.t.  $\phi(x) \approx x$ . An AE can be decomposed into an encoder and a decoder,  $\phi \equiv \phi_{dec} \circ \phi_{enc}$ . By choosing the intermediate representation  $\phi_{enc}(\cdot)$  to be comparatively low-dimensional,  $\phi_{enc}(\mathbf{I}_t)$  or  $(\phi_{enc}(\mathbf{I}_t), \mathbf{v}_t)$  could serve as a simplified but sufficient representation of the state, facilitating accelerated learning.

We implemented and evaluated two different AEs. The first one is a regular convolutional AE. It uses the following encoder and decoder:

$$\phi_{enc}^{conv} \equiv D_1 \circ C_3 \circ C_2 \circ C_1 \quad (1)$$

$$\phi_{dec}^{conv} \equiv C_6 \circ C_5 \circ C_4 \circ D_2 \quad (2)$$

where the  $C_i$  are convolutional layers and the  $D_i$  are fully connected layers.

The second AE is based on the deep SAE described in [28], but with some significant variations. In the following

sections, we refer it as Spatial Auto-encoder Variant (SAEV). The SAEV uses the encoder

$$\phi_{enc}^{sae} \equiv S \circ C_3 \circ C_2 \circ C_1 \quad (3)$$

where  $C_i$  are convolutional layers.  $C_1, C_2$  using exponential linear units (ELU) activation [38], while  $C_3$  uses a spatial softmax-activation:

$$\text{softmax}(z)_{i,j,c} = \frac{e^{z_{i,j,c}}}{\sum_{w=0}^W \sum_{h=0}^H e^{z_{w,h,c}}} \quad (4)$$

The mapping  $S$  takes a number of feature maps, which it treats as bivariate probability distributions. For each, a feature location is estimated by the expectation values:

$$\begin{aligned} x_c &= \mathbb{E}_{(i,j) \sim P_c} [i] \\ y_c &= \mathbb{E}_{(i,j) \sim P_c} [j] \end{aligned} \quad (5)$$

where  $P_c(i, j)$  is the  $(i, j)$  coordinate of the  $c^{\text{th}}$  feature-map of the input. The *presence* of a feature is defined as the weighted sum

$$\rho_c = \sum_{i=0}^W \sum_{j=0}^H P_c(i, j) \cdot \mathcal{N}(i, j | \boldsymbol{\mu} = (x_c, y_c), \boldsymbol{\Sigma} = k \cdot \mathbf{I}) \quad (6)$$

Intuitively, a feature map which is highly localized around the estimated position has a presence near 1, whereas one that is very spread out will have presence close to 0. The output from  $S$  is the concatenation of the  $(x_c, y_c, \rho_c)$  of each feature map. In other words, the intermediate representation contains actual image-coordinates of the features.

The main difference between our SAEV architecture and the SAE described in [28] is the decoder. The decoder we use is

$$\phi_{dec}^{sae} \equiv B \circ C_6 \circ C_5 \circ C_4 \circ \Delta \quad (7)$$

where  $C_i$  are convolutional layers,  $C_4, C_5$  uses ELU-activations, while  $C_6$  uses a sigmoid activation.  $\Delta: \mathbb{R}^{N \times 3} \rightarrow \mathbb{R}^{W \times H \times N}$  is a transformation that takes the  $N$   $(x_c, y_c, \rho_c)$ -tuples and maps each to a feature map:

$$\Delta(x_1, \dots, x_C, y_1, \dots, y_C, \rho_1, \dots, \rho_C)_{i,j,c} = \text{ELU}(\rho_c - \|(i, j) - (x_c, y_c)\|_2) \quad (8)$$

This creates  $N$  feature maps, with peaks at  $(i, j) = (x_c, y_c)$  that decrease radially outwards according to the ELU [38]. To the output of  $\Delta$ , three convolutional layers are applied, followed by an addition operation with a trainable constant to complete the decoder. The constant addition operation frees up the prior stages of the architecture to focus on learning positions of things that are not always in the same place.

All models are trained using the Adam-optimizer [39] on a loss function consisting of three components: reconstruction error  $L_{err} = \|\phi(s_t) - s_t\|_2$ , a presence based loss  $L_{pre} = 1 - \rho(s_t)$  that encourages localized features, and the smoothness loss  $L_{smooth} = (\phi_{enc}(s_{t+1}) - \phi_{enc}(s_t)) - (\phi_{enc}(s_t) - \phi_{enc}(s_{t-1}))$  defined in [28]. For the convolutional AE, the presence loss is ill-defined and thus that term was omitted. One can now use the intermediate representation  $\phi_{enc}(s_t)$  as input to the RL framework, or to visualize the corresponding image coordinates, as is shown in Figure 2.

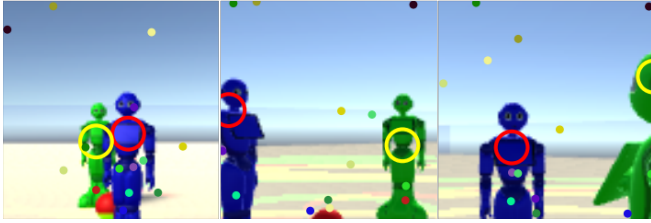


Fig. 3: Positions extracted from  $\phi_{enc}(s_t)$  visualized on top of their corresponding input images. The coloring of a feature is consistent across the three images. Most features are visualized as dots, except two features that have been chosen to be visualized as circles to more clearly show how they track their intended objects.

### C. Modeling group behavior

In a realistic multi-party conversation group, the individuals within it stand in appropriate positions with respect to others. This positional and orientational relationship has been defined as an *F-formation* as proposed by Kendon [33]. It characterizes a group of two or more individuals, typically in a conversation, to share information and interact with each other. Most importantly, it defines the *o-space* which is a common focused space in the group in which all individuals look inward and is exclusive to those external. When conditions change, such as a new individual joining the group, the group members should change position or orientation in order to form a new group including the newcomer. Jan et al. [11] proposed a group model which simulates these behaviors by a social force field. In this paper, we use an extended Social Force model which maintains F-formation through repositioning and reorientating by a conversation force field. This force field is produced and updated by three forces: a repulsion force, an equality force, and a cohesion force. The details of social force fields are described in [40]. In order to better model conversation groups, Hall's proxemics theory [41] is adopted when generating social

force fields, i.e. the repulsion, equality and cohesion forces occurring in personal, social and public spaces, respectively.

The repulsion force prevents other agents from stepping inside its personal space and generates a repulsion force to push them away. Let  $N_p$  be the number of other agents inside the personal space of agent  $i$ , and  $\mathbf{p}_i$  is the corresponding position of agent  $i$ . The repulsion force is shown in equation 9.

$$\mathbf{F}_r = -(d_p - d_{min})^2 \frac{\mathbf{p}_r}{\|\mathbf{p}_r\|} \quad (9)$$

where  $\mathbf{p}_r = \sum_i^{N_p} (\mathbf{p}_i - \mathbf{p})$ ,  $\mathbf{p}$  is the position of the agent currently being evaluated.  $d_p$  is the radius of its personal space, and  $d_{min}$  is the distance to its closet agent inside the personal space.

The equality force keeps o-space shared to all group members by generating an attraction or a repulsion force towards a point in o-space. Also, an orientation force towards o-space is generated to change body orientation. Let  $N_s$  be the number of other agents inside the social space. The equality force  $\mathbf{F}_e$  and equality orientation  $\mathbf{d}_e$  are shown in equation 10.

$$\begin{aligned} \mathbf{F}_e &= (1 - \frac{m}{\|\mathbf{c} - \mathbf{p}\|}) (\mathbf{c} - \mathbf{p}) \\ \mathbf{d}_e &= \sum_i^{N_s} (\mathbf{p}_i - \mathbf{p}) \end{aligned} \quad (10)$$

where  $\mathbf{c}$  is the centroid, i.e.  $\mathbf{c} = (\mathbf{p} + \sum_i^{N_s} \mathbf{p}_i) / (N_s + 1)$ , and  $m$  is the mean distance of the members from the centroid.

The cohesion force prevents an agent to be isolated from a group and keeps agents close to each other by generating an attraction force. Let  $N_a$  be the number of other agents inside the public area,  $\mathbf{o}$  is the conversation center and  $s$  is the radius of the o-space. The cohesion force  $\mathbf{F}_c$  and cohesion orientation  $\mathbf{d}_c$  are shown in equation 11.

$$\begin{aligned} \mathbf{F}_c &= \alpha (1 - \frac{s}{\|\mathbf{o} - \mathbf{p}\|}) (\mathbf{o} - \mathbf{p}) \\ \mathbf{d}_c &= \sum_i^{N_a} (\mathbf{p}_i - \mathbf{p}) \end{aligned} \quad (11)$$

where  $\alpha = N_a / (N_s + 1)$ , which is the scaling factor for the cohesion force used to reduce the magnitude of the cohesion force if the agent is surrounded by other agents in its social area.

In order to include a component in reward function to drive the robot to approach the group. We incorporate the extended social force model described previously and consider a line integral  $r_1$  over a path  $L$  in aforementioned force fields, namely force fields in personal, social and public spaces, to be the *group forming reward*. Mathematically, the group forming reward for the robot agent is defined as follows

$$R_1 = \int_L r_1(\mathbf{u}) \cdot d\mathbf{u} \quad (12)$$

where  $\mathbf{u}$  is the position of the robot along the trajectory  $L$ , and  $r_1(\mathbf{u}) = \sum_{i \in \{r,e,c\}} \mathbf{F}_i(\mathbf{u})$  is the combined force on the robot agent. Note that the force fields  $\mathbf{F}_i$  depend on the positioning of all agents, including the SHAs, but for notational simplicity, this is not made explicit in the formulae.

Together with the group forming reward, another reward function called *non-increasing reward* is added to ensure the the energy in the force field is non-increasing. Mathematically, it is defined as

$$R_2 = \int_{t_0}^{t_1} \mathbb{1}_{\mathcal{A}}(\mathbf{u}(t)) dt \quad (13)$$

where  $\mathbb{1}$  is the indicator function and  $\mathcal{A}$  is the set of points along the robot's trajectory where  $dr_1(\mathbf{u}(t))/dt \geq 0$ . These two reward functions help the robot agent to approach the group center. To add further incentive to complete the task, two other other reward components are added. They are a time-penalty  $R_3 = -\int_{t_0}^{t_1} dt$  ( $t_0, t_1$  are the times an episode starts and ends), together with a bonus reward  $R_4$  for successful approaching behavior within the required number of time steps.

#### D. Following Social Norms

In order to make the robot adhere to social norms when it is approaching the group, simulated feedback from other agents is taken into consideration. Therefore, the robot agent considers the impact of its own behavior on others, which is important in generating appropriate real-world robot approaching behaviors. Here, we define summation of all the line integrals of SHAs' paths in the force fields,

$$R_5 = -\sum_{j=0}^{N_p} \int_{L_j} \sum_{i \in \{r,e,c\}} \mathbf{F}_{ij} \cdot d\mathbf{u}_j \quad (14)$$

where  $N_p$  means the total number of the SHAs.

The final reward is a combination of all five rewards. Each is associated with a weight  $w_i \geq 0$  to indicate the importance of that reward category. On top of the weights considered for each category of rewards, two other weights are used to influence the behavior of the robot. One weight is called *egoism wight*  $w_e$ , which decides how much the robot agent considers achieving its own goal of approaching the group center. The other weight, *altruism weight*  $w_a$  decides how much it cares about other agents, meaning avoiding pushing other SHAs around. The final reward is defined as follows:

$$R = w_e \cdot (w_1 \cdot R_1 + w_2 \cdot R_2 + w_3 \cdot R_3 + w_4 \cdot R_4) + w_a \cdot w_5 \cdot R_5 \quad (15)$$

By balancing the different weights, we produce a realistic reward function that captures important notions from human social interaction, such as respecting the private space of others.

## IV. RESULTS

We used a DRL algorithm called Proximal Policy Optimization (PPO) [42] to learn an appropriate behavior for the robot agent. We selected PPO due to its stability advantages [43] over DQN-based RL algorithms. We used ML-Agents Toolkit<sup>2</sup> [44] to carry out our experiments.

### A. Models Configurations

To determine what state representation and type of network structure for the value and policy networks are the most suitable for robot approaching behavior, we evaluate combinations of state representations, and network architectures. For the state representations containing visual information, we evaluate both AEs (*conv* and *SAEV* from section III-B). The network structures considered are Feed-Forward (FF) networks and LSTM networks. The reported results are the best cumulative reward after one million steps (measured as an exponentially weighted running average) of the agent. Table I shows the model configurations and their corresponding performance.

Model	Reward	Percentage
Vector + LSTM (Baseline)	<b>-0.256</b>	<b>100.00%</b>
CameraOnly + SAEV + FF	-0.869	57.06%
CameraOnly + SAEV + LSTM	-0.804	61.63%
CameraOnly + conv + FF	-0.81	61.18%
CameraOnly + conv + LSTM	-1.091	41.51%
CameraSpeed + SAEV + LSTM	<b>-0.544</b>	<b>79.80%</b>
CameraSpeed + conv + LSTM	-0.709	68.22%
Random policy	-1.684	0.00%

TABLE I: Results of the different configurations.

Performance is measured both as cumulative reward, described in Section III-A, and as percentages. Percentages express relative performance, such that 100% correspond to the best performance, and 0% to the mean performance of a uniformly random agent. Figure 4 shows the learning curves of the best model with only image input, SAEV and a LSTM policy network.

### B. Approaching Behaviour: Perceptual Study

We compare the robot approaching behavior learned by our model with the one generated by SFM [40]. In a study conducted by Pedica et al. [12], it was shown that SFM increased believability of static group forming. A major drawback of SFM is that it is directly controlled by the social forces and therefore does not act according to current situations of the environment. We hypothesize that a learned agent that is able to accelerate and decelerated based on the simulated social feedbacks in RL framework can introduce more believability and social appropriateness. In order to compare the behaviors of the SFM and our model, we implemented a version of SFM and compared it with a model learned with the reward function defined in Section III-D. Figure 5 show paths sampled from our trained model and paths sampled from the SFM with the same initial positions.

<sup>2</sup><https://github.com/Unity-Technologies/ml-agents>



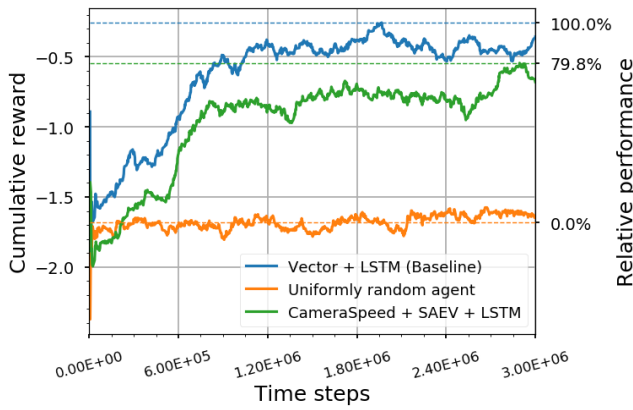


Fig. 4: Learning curves of the best model compared to a uniformly random agent and the baseline.

In order to subjectively assess the behavior of our learned model compared to the behavior generated by SFM, we created a questionnaire to evaluate the approaching behaviors using subjective measures. In this study, the social behaviours are considered from three dimensions, namely *polite*, *sociable* and *rude*, suggested by suggested by Okal and Arras [45].

The questionnaire contains in total six videos of approaching behaviour. The videos show six different approaching behaviors of the robot towards groups from three starting locations by both our model and the SFM. The participants observe the behaviors from a top-down view. For each video, the participant needs to answer four questions. Among the four questions, the first three questions ask the participants to evaluate how much do they think the behaviours are *polite*, *sociable* or *rude* respectively and fourth question asks how much do they think the robot is human-like. For each question, the participants need to select a point from a 1-7 Likert scale, where 1 means "not at all" and 7 means "very". The videos and their corresponding questions are given to the participants in a random order.

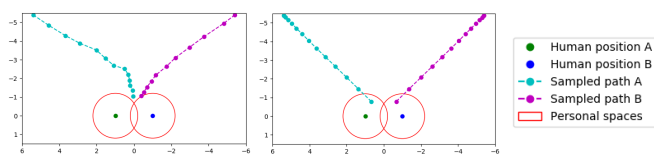


Fig. 5: Comparison of generated paths. The left image shows paths generated by our model and the right image shows paths generated by the social force model with the same initial positions.

In total, 20 participants answered the questions. They have an average age of 28.25 and mixed cultural backgrounds. Figure 6 shows their ratings on approaching behaviours generated by two models. We found that people consider the behavior generated by our model to be significantly more polite ( $t(19) = 6.45, p < .001$ ), less rude ( $t(19) = 6.46, p < .001$ ) and more sociable ( $t(19) = 2.65, p < .025$ ). However, we did not find the approaching behaviour generated by our

model to be significantly more human-like than the ones generated by SFM ( $t(19) = 1.01, p > .05$ ). This might be related to human-likeness is hard to measure when there are more than one factors involved [46].

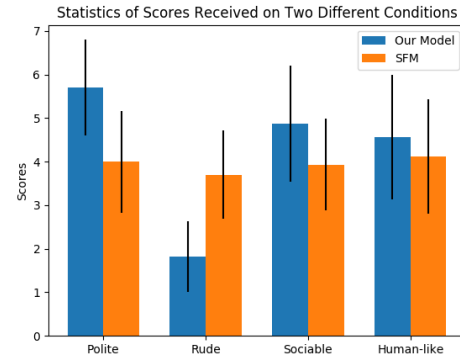


Fig. 6: Comparison of behaviors generated by two models. The behavior generated by our model is generally considered to be more polite, less rude and more sociable.

## V. CONCLUSION AND FUTURE WORK

In this work, we proposed a scheme (SSBL) that can be considered as a general framework for social robot learning. In order to demonstrate it, we implemented a robot approaching behavior task based on this scheme. For this task, we designed a reward function combining concepts from SFM and Hall's proxemics theory to enable the robot agent to learn a dynamical model which takes social norms into account. The training was repeated for three different state representations.

Our results show that we can generate more socially considerate behaviors than SFM, and the SAEV outperforms the vanilla convolutional AE on this task for both CameraOnly and CameraSpeed inputs.

In the future, we will conduct studies where humans are asked to qualitatively assess the behavior of our learned model compared to the behavior generated by SFM in real-world situations. Regarding the model configuration experiments, we will also investigate how to utilize more subtle real-world human feedback such as engagement to refine our learned model using model-based RL algorithms. As a result, the robot will exhibit comfortable approaching behavior by taking other affective measures into account. We will also conduct policy refinement experiments in the real world with the Pepper robot and actual humans.

## REFERENCES

- [1] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, *et al.*, "The limits and potentials of deep learning for robotics," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 405-420, 2018.
- [2] S. Chernova and A. L. Thomaz, "Robot learning from human teachers," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 8, no. 3, pp. 1-121, 2014.
- [3] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334-1373, 2016.

- [4] E. Ferreira and F. Lefevre, "Reinforcement-learning based dialogue system for human-robot interactions with socially-inspired rewards," *Computer Speech & Language*, vol. 34, no. 1, pp. 256–274, 2015.
- [5] J. P. Hanna and P. Stone, "Grounded action transformation for robot learning in simulation," in *AAAI*, pp. 3834–3840, 2017.
- [6] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, et al., "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," *arXiv preprint arXiv:1709.07857*, 2017.
- [7] A. Abubshait and E. Wiese, "You look human, but act like a machine: agent appearance and behavior modulate different aspects of human-robot interaction," *Frontiers in psychology*, vol. 8, p. 1393, 2017.
- [8] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1653–1660, 2014.
- [9] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, et al., "Emonets: Multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.
- [10] R. Jonschkowski and O. Brock, "State representation learning in robotics: Using prior knowledge about physical interaction," in *Robotics: Science and Systems*, 2014.
- [11] D. Jan and D. R. Traum, "Dynamic movement and positioning of embodied agents in multiparty conversations," in *Proceedings of the Workshop on Embodied Language Processing*, pp. 59–66, Association for Computational Linguistics, 2007.
- [12] C. Pedica, H. H. Vilhjálmsón, and M. Lárusdóttir, "Avatars in conversation: the importance of simulating territorial behavior," in *International Conference on Intelligent Virtual Agents*, pp. 336–342, Springer, 2010.
- [13] Y. Gao, S. Wallkötter, O. Mohammad, and G. Castellano, "Investigating deep learning approaches for human-robot proxemics," in *Robot and Human Interactive Communication, 2018. ROMAN 2018. IEEE International Workshop on*, IEEE, 2018.
- [14] S. Bozinovski, G. Stojanov, and L. Bozinovska, "Emotion, embodiment, and consequence driven systems," in *Proc AAAI fall symposium on embodied cognition and action*, pp. 12–17, 1996.
- [15] S. Bozinovski and E. T. (Ed.), "A self-learning system using secondary reinforcement," in *Cybernetics and Systems Research*, pp. 397–402, 1982.
- [16] Y. Gao, W. Barendregt, M. Obaid, and G. Castellano, "When robot personalisation does not help: Insights from a robot-supported learning study," in *Robot and Human Interactive Communication, 2018. ROMAN 2018. IEEE International Workshop on*, IEEE, 2018.
- [17] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.
- [18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [21] A. H. Qureshi, Y. Nakamura, Y. Yoshikawa, and H. Ishiguro, "Robot gains social intelligence through multimodal deep reinforcement learning," in *Humanoid Robots (Humanoids), 2016 IEEE-RAS 16th International Conference on*, pp. 745–751, IEEE, 2016.
- [22] M. Clark-Turner and M. Begum, "Deep reinforcement learning of abstract reasoning from demonstrations," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 372–372, ACM, 2018.
- [23] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.
- [24] B. Tan, N. Xu, and B. Kong, "Autonomous driving in reality with reinforcement learning and image translation," *arXiv preprint arXiv:1801.05299*, 2018.
- [25] P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. J. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu, et al., "Learning to navigate in complex environments," *arXiv preprint arXiv:1611.03673*, 2016.
- [26] W. Böhmer, J. T. Springenberg, J. Boedecker, M. Riedmiller, and K. Obermayer, "Autonomous learning of state representations for control: An emerging field aims to autonomously learn state representations for reinforcement learning agents from their real-world sensor observations," *KI-Künstliche Intelligenz*, vol. 29, no. 4, pp. 353–362, 2015.
- [27] S. Lange, M. Riedmiller, and A. Voigtlander, "Autonomous reinforcement learning on raw visual input data in a real world application," in *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pp. 1–8, IEEE, 2012.
- [28] C. Finn, X. Y. Tan, Y. Duan, T. Darrell, S. Levine, and P. Abbeel, "Deep spatial autoencoders for visuomotor learning," *arXiv preprint arXiv:1509.06113*, 2015.
- [29] L. Heigheas, A. Luciani, J. Thollot, and N. Castagné, "A physically-based particle model of emergent crowd behaviors," *arXiv preprint arXiv:1005.4405*, 2010.
- [30] A. Treuille, S. Cooper, and Z. Popović, "Continuum crowds," *ACM Transactions on Graphics (TOG)*, vol. 25, no. 3, pp. 1160–1168, 2006.
- [31] S. R. Musse and D. Thalmann, "Hierarchical model for real time simulation of virtual human crowds," *IEEE Transactions on Visualization and Computer Graphics*, vol. 7, no. 2, pp. 152–164, 2001.
- [32] C. W. Reynolds, "Steering behaviors for autonomous characters," in *Game developers conference*, vol. 1999, pp. 763–782, Citeseer, 1999.
- [33] A. Kendon, *Conducting interaction: Patterns of behavior in focused encounters*, vol. 7. CUP Archive, 1990.
- [34] O. A. I. Ramírez, H. Khambhaita, R. Chatila, M. Chetouani, and R. Alami, "Robots learning how and where to approach people," in *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*, pp. 347–353, IEEE, 2016.
- [35] C. Pedica and H. H. Vilhjálmsón, "Lifelike interactive characters with behavior trees for social territorial intelligence," in *ACM SIGGRAPH 2012 Posters*, p. 32, ACM, 2012.
- [36] F. Yang, C. Li, R. Palmberg, E. Van Der Heide, and C. Peters, "Expressive virtual characters for social demonstration games," in *Virtual Worlds and Games for Serious Applications (VS-Games), 2017 9th International Conference on*, pp. 217–224, IEEE, 2017.
- [37] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.
- [38] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.
- [39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [40] C. Pedica and H. Vilhjálmsón, "Social perception and steering for online avatars," in *International Workshop on Intelligent Virtual Agents*, pp. 104–116, Springer, 2008.
- [41] E. T. Hall, R. L. Birdwhistell, B. Bock, P. Bohannon, A. R. Diebold Jr, M. Durbin, M. S. Edmonson, J. Fischer, D. Hymes, S. T. Kimball, et al., "Proxemics [and comments and replies]," *Current anthropology*, vol. 9, no. 2/3, pp. 83–108, 1968.
- [42] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [43] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," *arXiv preprint arXiv:1709.06560*, 2017.
- [44] A. Juliani, V.-P. Berges, E. Vckay, Y. Gao, H. Henry, M. Mattar, and D. Lange, "Unity: A general platform for intelligent agents," *arXiv preprint arXiv:1809.02627*, 2018.
- [45] B. Okal and K. O. Arras, "Learning socially normative robot navigation behaviors with bayesian inverse reinforcement learning," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pp. 2889–2895, IEEE, 2016.
- [46] K. F. MacDorman, "Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley," in *ICCS/CogSci-2006 long symposium: Toward social mechanisms of android science*, pp. 26–29, 2006.