Social Behavior Learning with Realistic Reward Shaping

Yuan Gao\textsuperscript{1}, Fangkai Yang\textsuperscript{2}\textsuperscript{†}, Martin Frisk\textsuperscript{1\textdagger}, Daniel Hernandez\textsuperscript{3}, Christopher Peters\textsuperscript{2} and Ginevra Castellano\textsuperscript{1}

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. Based on this scheme, we implement robot approaching behaviors towards a small group generated from F-formation 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, i.e. social force 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 a small group generated from F-formation [14] 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 from a prior model which is built upon social forces [11] [15] in the environment. The last step is to make sure that the learned behavior follows the social standard by using simulated social norms 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 [16] [17]. After that, many papers in HRI started to investigate the effect of RL algorithms like Exp3 [18] or Q-learning [19] in social robotics settings. However, since such algorithms lack the ability to capture important features from high-dimensional signals [20], their applicability to solve HRI problems remains limited. After the era of deep learning started in 2006 [21], many different algorithms were proposed to understand different modalities in HRI, for example, ResNet [22] for image processing and Long Short-Term Memory [23] (LSTM)-based solutions for

\textsuperscript{†}Authors contributed equally.

\textsuperscript{1}This work is partly supported by the Swedish Research Council (grant n. 2015-04378) and partly by the COIN project (RTIT15-0133) funded by the Swedish Foundation for Strategic Research.

\textsuperscript{1} Author is with Department of Information Technology, Uppsala University, Uppsala, Sweden. alex.yuan.gao@it.uu.se

\textsuperscript{2} Author is with Department of Computational Science and Technology, Royal Institute of Technology, KTH, Stockholm, Sweden.

\textsuperscript{3} Author is with Department of Computer Science, University of York, York, United Kingdom.
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 [24]. In this work, a Deep Q-Network (DQN) [20] was used to learn a mapping from visual input to one of several predefined actions for greeting people. Another work was conducted by Madson [25], 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 [26], driving simulated cars [27] and navigating mazes [28]. However, learning policies directly from high-dimentional data such as images requires a large amount of samples, which makes it intractable in social robot learning [29]. 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. [30]. Several variants of AEs have been applied as well, including attempts by Böhmer et al. [29] to learn the dynamics of the environment by constructing an AE predicting the next image, and Finn et al. [31] 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 [32] [33] simulate global collective behaviors of large scale groups or crowds. For modeling small scale groups, agent-based methods [34] [35] are adopted to simulate the behavior of each individual based on rules of behavior. Specifically, in a small multi-party conversation group, Kendon [14] 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. Ramirez et al. [36] adopted inverse reinforcement learning, involving several participants demonstrating approaching behaviors for a robot to learn. Pedica et al. [37] 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 [38], 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 how the environment was built and how the training of a dynamical model within an RL framework was conducted. Section III-D shows how one can formulate the training of a dynamical model within an RL framework. Section III-C shows how social norms can be acquired by utilizing concepts from the Social Force Model (SFM) [14]. In the original SFM work [14], 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 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 $\tau = \{s_1, a_1, \ldots, s_T, a_T\}$, 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 and $p(s_{t+1}|s_t, a_t)$ 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 $s_i = \mathbf{I}_i$ and $s_i = (\mathbf{I}_i, v_I)$ respectively. Here $\mathbf{I}_i$ is the visual information from the robot’s first-person view rendered by the Unity engine, and $v_I$ 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 [26]. To do this, we utilize an autoencoder (AE) [39], 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, $\phi = \phi_{\text{dec}} \circ \phi_{\text{enc}}$.

By choosing the intermediate representation $\phi_{\text{enc}}(\cdot)$ to be comparatively low-dimensional, $\phi_{\text{enc}}(\mathbf{I}_i) \circ (\phi_{\text{enc}}(\mathbf{I}_i), v_I)$ 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:

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

$$
\phi_{\text{dec}}^\text{conv} = C_0 \circ C_5 \circ C_4 \circ D_2 \quad (2)
$$

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 [31], but with some significant variations. In the following sections, we refer it as Spatial Auto-encoder Variant (SAEV). The SAEV uses the encoder

$$
\phi_{\text{enc}}^\text{saev} = S \circ C_3 \circ C_2 \circ C_1
$$

where $C_i$ are convolutional layers. $C_1$,$C_2$ using exponential linear units (ELU) activation [40], while $C_3$ uses a spatial softmax-activation:

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

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:

$$
x_c = E_{i,j} \sim p_c(\cdot \mid i) \quad (5)
$$

$$
y_c = E_{i,j} \sim p_c(\cdot \mid j)
$$

where $p_c(i, j)$ is the $(i, j)$ coordinate of the $c$th feature-map of the input. The *presence* of a feature is defined as the weighted sum

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

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 [31] is the decoder. The decoder we
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 [15]. In order to better model conversation groups, Hall’s proxemics theory [42] 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 \( p_i \) be the corresponding position of agent \( i \). The repulsion force is shown in equation \( \text{(9)} \)

\[
\mathbf{F}_r = -(d_p - d_{\text{min}})^2 \frac{\mathbf{p}_r}{||\mathbf{p}_r||}
\]

where \( \mathbf{p}_r = \sum_{i=1}^{N_p} (p_i - p) \), \( p \) is the position of the agent currently being evaluated. \( d_p \) is the radius of its personal space, and \( d_{\text{min}} \) is the distance to its closest 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 \( \text{(10)} \)

\[
\mathbf{F}_e = (1 - \frac{m}{||c - p||})(c - p)
\]

\[
\mathbf{d}_e = \sum_{i=1}^{N_s} (p_i - p)
\]

where \( c \) is the centroid, i.e. \( c = (p + \sum_{i=1}^{N_s} p_i)/(N_s + 1) \), 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, \( o \) 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 \( \text{(11)} \)

\[
\mathbf{F}_c = \alpha (1 - \frac{s}{||o - p||})(o - p)
\]

\[
\mathbf{d}_c = \sum_{i=1}^{N_a} (p_i - p)
\]

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:

\[
R_1 = \int_{L} r_1(u) \cdot du
\]  

where \( u \) is the position of the robot along the trajectory \( L \), and \( r_1(u) = \sum_{i \in \{r,e,c\}} F_i(u) \) is the combined force on the robot agent. Note that the force fields \( F_i \) 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 energy in the force field is non-increasing. Mathematically, it is defined as

\[
R_2 = \int_{t_0}^{t_1} 1_A(u(t)) dt
\]

where \( 1 \) is the indicator function and \( A \) is the set of points along the robot’s trajectory where \( dr_1(u(t))/dt \geq 0 \). These two reward functions help the robot agent to approach the group center. To add further incentive to complete the task, two other 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.

\[R_4 = \sum_{j=0}^{N_p} \int_{L_j} \sum_{i \in \{r,e,c\}} F_{ij} \cdot du_j\]

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 weight \( w_e \), 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:

\[
R = w_e \cdot (w_1 \cdot R_1 + w_2 \cdot R_2 + w_3 \cdot R_3 + w_4 \cdot R_4) +
+ w_a \cdot w_5 \cdot R_5
\]

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) \[43\] to learn an appropriate behavior for the robot agent. We selected PPO due to its stability advantages \[44\] over DQN-based RL algorithms. We used ML-Agents Toolkit\[45\] 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. Table I shows the model configurations and their corresponding performance.

| Model                  | Reward | Percentage |
|------------------------|--------|------------|
| Vector + LSTM (Baseline)| -0.256 | 100.00%    |
| 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| -0.544 | 79.80%     |
| CameraSpeed + conv + LSTM| -0.709 | 68.22%     |
| Random policy          | -1.684 | 0.00%      |

TABLE I: Results of the different configurations. The reported results are the best cumulative reward of the model.

Performance is measured both as cumulative reward (an exponentially weighted running average is used to smooth the function.), described in Section III-A and as percentages. Percentages express relative performance, such that 100% correspond to the baseline performance, and 0% to the mean performance of a uniformly random agent. Figure 4 shows the learning curve of the best model, which uses image and robot’s speed as input, output of SAEV as learning state representation and a LSTM as policy network.

B. Approaching Behaviour: Perceptual Study

We compare the robot approaching behavior learned by our model with the one generated by SFM \[15\]. 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

\[https://github.com/Unity-Technologies/ml-agents\]
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 decelerate 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.

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 Okal and Arras [46]. 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".

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 [47].

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.

The code is publicly available at [https://github.com/gaoyuankidult/PepperSocial/tree/master](https://github.com/gaoyuankidult/PepperSocial/tree/master).
[47] 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.