2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
September 24–28, 2017, Vancouver, BC, Canada

Socially Aware Motion Planning with Deep Reinforcement Learning

Yu Fan Chen, Michael Everett, Miao Liu†, and Jonathan P. How

***Abstract*— For robotic vehicles to navigate safely and efficiently in pedestrian-rich environments, it is important to model subtle human behaviors and navigation rules (e.g., passing on the right). However, while instinctive to humans, socially compliant navigation is still difficult to quantify due to the stochasticity in people's behaviors. Existing works are mostly focused on using feature-matching techniques to describe and imitate human paths, but often do not generalize well since the feature values can vary from person to person, and even run to run. This work notes that while it is challenging to directly specify the details of what *to* do (precise mechanisms of human navigation), it is straightforward to specify what *not to* do (violations of social norms). Specifically, using deep reinforcement learning, this work develops a time-efficient navigation policy that respects common social norms. The proposed method is shown to enable fully autonomous navigation of a robotic vehicle moving at human walking speed in an environment with many pedestrians.**

Fig. 1: A robotic vehicle navigating autonomously in a pedestrian-rich environment. Accounting for social interactions is important for operating such vehicles safely and smoothly.

I. Introduction

Recent advances in sensing and computing technologies have spurred greater interest in various applications of autonomous ground vehicles. In particular, researchers have explored using robots to provide personal mobility services and luggage carrying support in complex, pedestrian-rich environments (e.g., airports and shopping malls) [1]. These tasks often require the robots to be capable of navigating efficiently and safely in close proximity of people, which is challenging because pedestrians tend to follow subtle social norms that are difficult to quantify, and pedestrians' intents (i.e., goals) are usually not known [2].

A common approach treats pedestrians as dynamic obstacles with simple kinematics, and employs specific reactive rules for avoiding collision [3]–[6]. Since these methods do not capture human behaviors, they sometimes generate unsafe/unnatural movements, particularly when the robot operates near human walking speed [2]. To address this issue, more sophisticated motion models have been proposed, which would reason about the nearby pedestrians' hidden intents to generate a set of predicted paths [7], [8]. Subsequently, classical path planning algorithms would be employed to generate a collision-free path for the robot. Yet, separating the navigation problem into disjoint prediction and planning steps can lead to the *freezing robot problem*, in which the robot fails to find any feasible action because the predicted paths could mark a large portion of the space untraversable [9]. A key to resolving this problem is to account for cooperation, that is, to model/anticipate the impact of the robot's motion on the nearby pedestrians.

Existing work on cooperative, socially compliant navigation can be broadly classified into two categories, namely *model-based* and *learning-based*. Model-based approaches are typically extensions of multiagent collision avoidance algorithms, with additional parameters introduced to account for social interactions [7], [10]–[13]. For instance, to distinguish between human–human and human–robot interactions, the extended social forces model [11], [12] augments the potential field algorithm with additional terms that specify the repulsive forces (e.g., strength and range) governing each type of interaction. Model-based methods are designed to be computationally efficient as they often correspond to intuitive geometric relations; yet, it is unclear whether humans do follow such precise geometric rules. In particular, the force parameters often need to be tuned individually, and can vary significantly for different pedestrians [12]. Also, it has been observed that model-based methods can lead to oscillatory paths [2], [14].

In comparison, learning-based approaches aim to develop a policy that emulates human behaviors by matching feature statistics, such as the minimum separation distance to pedestrians. In particular, Inverse Reinforcement Learning (IRL) [15] has been applied to learn a cost function from human demonstration (teleoperation) [16], and a probability distribution over the set of joint trajectories with nearby pedestrians [2], [17]. Compared with model-based approaches, learning-based methods have been shown to produce paths that more closely resemble human behaviors, but often at a much higher computational cost. This is because computing/matching trajectory features often requires anticipating the joint paths of all nearby pedestrians [2], and might depend

Laboratory of Information and Decision Systems, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA, USA
{chenyuf2, mfe, jhow}@mit.edu
†IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA miao.liu1@ibm.com

on some unobservable information (e.g., pedestrians' goals). More importantly, since human behaviors are inherently stochastic, the feature statistics calculated on pedestrians' paths can vary significantly from person to person, and even run to run for the same scenario [2], [16]. This raises concerns over whether such feature-matching methods are generalizable to different environments [13].

In short, existing works are mostly focused on modeling and replicating the detailed mechanisms of social compliance, which remains difficult to quantify due to the stochasticity in people's behaviors. In comparison, humans can intuitively evaluate whether a behavior is acceptable. In particular, human navigation (or teleoperation) is time-efficient and generally respects a set of simple social norms (e.g., "passing on the right") [2], [16], [18]. Building on a recent paper [14], we characterize these properties in a reinforcement learning framework, and show that human-like navigation conventions emerge from solving a cooperative collision avoidance problem.

The main contributions of this work are (i) introducing socially aware collision avoidance with deep reinforcement learning (SA-CADRL) for explaining/inducing socially aware behaviors in a RL framework, (ii) generalizing to multiagent ($n > 2$) scenarios through developing a symmetrical neural network structure, and (iii) demonstrating on robotic hardware autonomous navigation at human walking speed in a pedestrian-rich environment.

II. Background

A. Collision Avoidance with Deep Reinforcement Learning

A multiagent collision avoidance problem can be formulated as a sequential decision making problem in a reinforcement learning framework [14]. Let $\mathbf{s}_t$, $\mathbf{u}_t$ denote an agent's state and action at time t, and let $\tilde{\mathbf{s}}_t$ denote the state of a nearby agent. To capture the uncertainty in nearby agents' intents, the state vector is partitioned into observable and unobservable parts, that is $\mathbf{s}_t = [\mathbf{s}_t^o, \mathbf{s}_t^h]$. Let the observable states be the agent's position, velocity, and radius (size), $\mathbf{s}^o = [p_x, p_y, v_x, v_y, r] \in \mathbb{R}^5$; let the unobservable states be the agent's intended goal position, preferred speed, and orientation, $\mathbf{s}^h = [p_{gx}, p_{gy}, v_{pref}, \psi] \in \mathbb{R}^4$; and let the action be the agent's velocity, $\mathbf{u}_t = \mathbf{v}_t$. It will be explained in Section IV that the observable states can be readily obtained from sensor measurements. The objective is to develop a policy, $\pi : (\mathbf{s}_t, \tilde{\mathbf{s}_t}^o) \mapsto \mathbf{u}_t$, that minimizes the expected time to goal $\mathbb{E}[t_g]$ while avoiding collision with nearby agents,

$$\underset{\pi(\mathbf{s}, \tilde{\mathbf{s}}^o)}{\operatorname{argmin}} \quad \mathbb{E}\left[t_g | \mathbf{s}_0, \tilde{\mathbf{s}}_0^o, \pi\right] \tag{1}$$

$$s.t. \quad ||\mathbf{p}_t - \tilde{\mathbf{p}}_t||_2 \geq r + \tilde{r} \qquad \forall t \tag{2}$$

$$\mathbf{p}_{t_g} = \mathbf{p}_g \tag{3}$$

$$\mathbf{p}_t = \mathbf{p}_{t-1} + \Delta t \cdot \pi(\mathbf{s}_{t-1}, \tilde{\mathbf{s}}_{t-1}^o)$$

$$\tilde{\mathbf{p}}_t = \tilde{\mathbf{p}}_{t-1} + \Delta t \cdot \pi(\tilde{\mathbf{s}}_{t-1}, \mathbf{s}_{t-1}^o), \tag{4}$$

where (2) is the collision avoidance constraint, (3) is the goal constraint, (4) is the agents' kinematics, and the expectation in (1) is with respect to the other agent's unobservable states (intents) and policy.

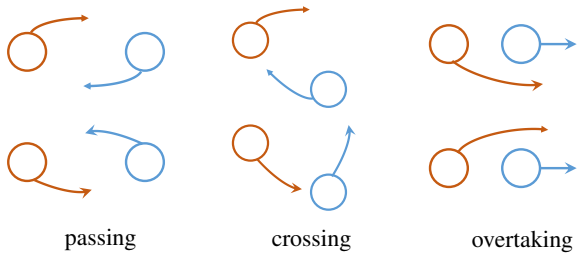


Fig. 2: Symmetries in multiagent collision avoidance. Left to right show two equally time-efficient ways for the red agent to pass, cross and overtake the blue agent. The top row is often called the left-handed rules, and bottom row the right-handed rules.

This problem can be formulated in a reinforcement learning (RL) framework by considering an agent's joint configuration with its neighbor, $\mathbf{s}^{jn} = [\mathbf{s}, \tilde{\mathbf{s}}^o]$. In particular, a reward function, $R_{col}(\mathbf{s}^{jn}, \mathbf{u})$, can be specified to reward the agent for reaching its goal and penalize the agent for colliding with others. The unknown state-transition model, $P(\mathbf{s}_{t+1}^{jn}, \mathbf{s}_t^{jn} \mid \mathbf{u}_t)$, takes into account the uncertainty in the other agent's motion due to its hidden intents ($\tilde{\mathbf{s}}^h$). Solving the RL problem amounts to finding the optimal value function that encodes an estimate of the expected time to goal,

$$V^*(\mathbf{s}_0^{jn}) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^{t \cdot v_{pref}} R_{col}(\mathbf{s}_t^{jn}, \pi^*(\mathbf{s}_t^{jn})) \mid \mathbf{s}_0^{jn}\right], \tag{5}$$

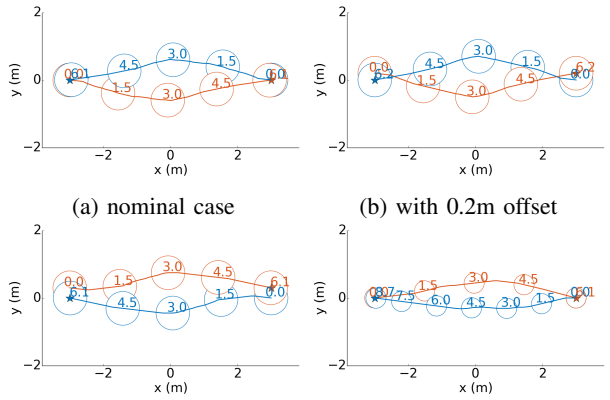
where $\gamma \in [0, 1)$ is a discount factor. The optimal policy can be retrieved from the value function, that is

$$\pi^*(\mathbf{s}_{t+1}^{jn}) = \underset{\mathbf{u}}{\operatorname{argmax}} R_{col}(\mathbf{s}_t, \mathbf{u}) + \gamma^{\Delta t \cdot v_{pref}} \int_{\mathbf{s}_{t+1}^{jn}} P(\mathbf{s}_t^{jn}, \mathbf{s}_{t+1}^{jn} | \mathbf{u}) V^*(\mathbf{s}_{t+1}^{jn}) d\mathbf{s}_{t+1}^{jn}. \tag{6}$$

A major challenge in finding the optimal value function is that the joint state $\mathbf{s}^{jn}$ is a continuous, high-dimensional vector, making it impractical to discretize and enumerate the state space. Recent advances in reinforcement learning address this issue by using deep neural networks to represent value functions in high-dimensional spaces, and have demonstrated human-level performance on various complex tasks [19]–[21]. While several recent works have applied deep RL to motion planning [22], [23], they are mainly focused on single agent navigation in unknown static environments, and with an emphasis on computing control inputs directly from raw sensor data (e.g., camera images). In contrast, this work extends the collision avoidance with deep reinforcement learning framework (CADRL) [14] to characterize and induce socially aware behaviors in multiagent systems.

B. Characterization of Social Norms

It has been widely observed that humans tend to follow simple navigation norms to avoid colliding with each other, such as passing on the right and overtaking on the left [18]. Albeit intuitive, it remains difficult to quantify the precise



(a) nominal case (b) with 0.2m offset

(c) with 0.4m offset (d) with diff. radii and speeds

Fig. 3: Indications of a navigation convention from the CADRL policy. Circles show each agent's position at the labeled time, and stars mark the goals. (a) CADRL shows a preference to the right as two agents pass each other in a symmetrical test case (swapping position). (b) This passing direction is robust to a small offset in the initial condition. (c) The passing direction changes at a larger offset (balancing with time to reach the goal). (d) The convention is not consistent with human social norms, as two agents of different sizes and velocities exhibit a preference to the left.

mechanisms of social norms, such as when to turn and how much to turn when passing another pedestrian; and the problem exacerbates as the number of nearby pedestrians increases. This is largely due to the stochasticity in people's motion (e.g., speed, smoothness), which can vary significantly among different individuals [2].

Rather than trying to quantify human behaviors directly, this work notes that the complex normative motion patterns can be a consequence of simple local interactions. For instance, an intuitive pairwise collision avoidance rule [10] can cause simulated agents moving in the same direction to form lanes in long corridors. Thus, we conjecture that rather than a set of precisely defined procedural rules, social norms are the emergent behaviors from a time-efficient, reciprocal collision avoidance mechanism. Reciprocity implicitly encodes a model of the other agents' behavior, which is the key for enabling cooperation without explicit communication. Also, note that reciprocity does not require a unique set of navigation rules, since both the left-handed and the right-handed rules can resolve path conflicts as shown in Fig. 2. Similarly, human navigation conventions are not unique, as the strength (e.g., separation distance) and passing direction vary in different countries [24].

Existing works have reported that human navigation (or teleoperation of a robot) tends to be cooperative and time-efficient [2], [16]. This work notes that these two properties are encoded in the CADRL formulation through using the min-time reward function and the reciprocity assumption ($\tilde{\pi} = \pi$). Furthermore, it was interesting to observe that while no behavioral rules (e.g., function forms) were imposed in the problem formulation, CADRL policy exhibits certain navigation conventions, as illustrated in Fig. 3. In particular, Fig. 3a illustrates two CADRL agents passing on the right of each other, showing signs of conforming to mutually agreed rules. More importantly, this preference in passing direction is robust to small deviations in the initial condition, as shown in Fig. 3b. As the offset increases, the CADRL agents eventually change passing direction in favor of shorter, smoother paths (Fig. 3c). Recall no communication took place and each agent's intent (e.g., goal) is not known to the other.

However, the cooperative behaviors emerging from a CADRL solution are not consistent with human interpretation. For instance, two CADRL agents with different sizes and preferred speeds show a preference to pass on the left of each other (Fig. 3d). This is because an agent's state $\mathbf{s}$ is defined to be the concatenation of its position, velocity, size and goal, so it is unlikely that an emerging tie breaking navigation convention would solely depend on the relative position (as human social norms). Moreover, the cooperative behaviors of CADRL cannot be controlled – they are largely dependent on the initialization of the value network and set of randomly generated training test cases. The next section will address this issue and present a method to induce behaviors that respect human social norms.

III. Approach

The following presents the socially aware multiagent collision avoidance with deep reinforcement learning algorithm (SA-CADRL). We first describe a strategy for shaping normative behaviors for a two-agent system in the RL framework, and then generalize the method to multiagent scenarios.

A. Inducing Social Norms

Recall the RL training process seeks to find the optimal value function (5), which maps from an agent's joint state with its neighbor, $\mathbf{s}^{jn} = [\mathbf{s},\ \tilde{\mathbf{s}}^o]$, to a scalar value that encodes the expected time to goal. To reduce redundancy (up to a rotation and translation), this work uses a local coordinate frame with the x-axis pointing towards an agent's goal, as shown in Fig. 4. Specifically, each agent's state is parametrized as

$$\mathbf{s} = [d_g,\ v_{pref},\ v_x,\ v_y,\ \psi,\ r] \tag{7}$$

$$\tilde{\mathbf{s}}^o = [\tilde{p}_x,\ \tilde{p}_y,\ \tilde{v}_x,\ \tilde{v}_y,\ \tilde{r},\ \tilde{d}_a,\ \tilde{\phi},\ \tilde{b}_{on}]\ , \tag{8}$$

where $d_g = ||\mathbf{p}_g - \mathbf{p}||_2$ is the agent's distance to goal, $\tilde{d}_a = ||\mathbf{p} - \tilde{\mathbf{p}}||_2$ is the distance to the other agent, $\tilde{\phi} = \tan^{-1}(\tilde{v}_y/\tilde{v}_x)$ is the other agent's heading direction, and $\tilde{b}_{on}$ is a binary flag indicating whether the other agent is real or virtual (details will be provided in Section III-B).

This work notes that social norms are one of the many ways to resolve a symmetrical collision avoidance scenario, as illustrated in Fig. 2. To induce a particular norm, a small bias can be introduced in the RL training process in favor of one set of behaviors over others. For instance, to encourage passing on the right, states (configurations) with another agent approaching from the undesirable side can be penalized (green region in Figure 4). The advantage of this approach is that violations of a particular social norm are usually easy to specify; and this specification need *not* be precise. This is because the addition of a penalty breaks

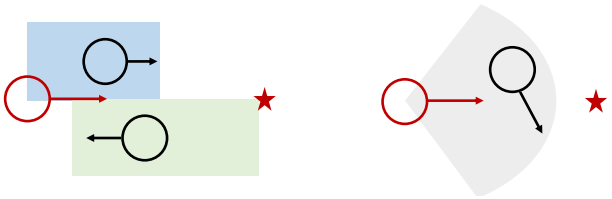
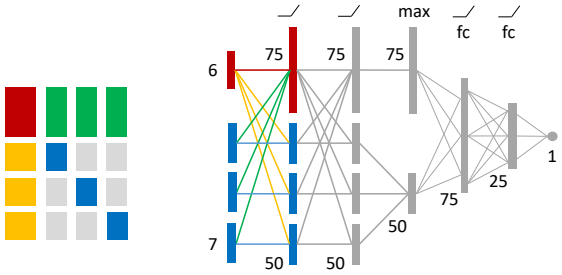


Fig. 4: Norm inducing reward function (depiction of (10)-(12)). The red agent is penalized if there is another agent in the blue, green or gray shaded regions, corresponding to overtaking, passing and crossing, respectively. This induces the right-handed rules as shown in Fig. 2.



(a) weight matrix (b) symmetric multiagent net

Fig. 5: Network structure for multiagent scenarios. (a) illustrates the weight matrix used in the first layer, where block partitions with the same color are constrained to be the same. (b) shows the overall network structure, with two symmetric layers, one max-pooling layer, and two fully connected layers. The colored connections in the first layer correspond to (a), where red are the weights associated with the agent itself, and blue are that of the nearby agents.

the symmetry in the collision avoidance problem, thereby favoring behaviors respecting the desired social norm. This work uses the following specification of a reward function R_{norm} for inducing the right-handed rules (Fig. 4),

$$R_{norm}(\mathbf{s}^{jn}, \mathbf{u}) = q_n I(\mathbf{s}^{jn} \in \mathcal{S}_{norm}) \tag{9}$$

$$s.t. \quad \mathcal{S}_{norm} = \mathcal{S}_{pass} \cup \mathcal{S}_{ovtk} \cup \mathcal{S}_{cross}$$

$$\mathcal{S}_{pass} = \{\, \mathbf{s}^{jn} \mid d_g > 3, \quad 1 < \tilde{p_x} < 4, \quad -2 < \tilde{p_y} < 0, \quad |\tilde{\phi} - \psi| > 3\pi/4 \,\} \tag{10}$$

$$\mathcal{S}_{ovtk} = \{\, \mathbf{s}^{jn} \mid d_g > 3, \quad 0 < \tilde{p_x} < 3, \quad |\mathbf{v}| > |\tilde{\mathbf{v}}| \quad 0 < \tilde{p_y} < 1, \quad |\tilde{\phi} - \psi| < \pi/4 \,\} \tag{11}$$

$$\mathcal{S}_{cross} = \{\, \mathbf{s}^{jn} \mid d_g > 3, \quad \tilde{d}_a < 2, \quad \tilde{\phi}_{rot} > 0, \quad -3\pi/4 < \tilde{\phi} - \psi < -\pi/4 \,\}, \tag{12}$$

where q_n is a scalar penalty, $I(\cdot)$ is the indicator function, $\tilde{\phi}_{rot} = \tan^{-1}((\tilde{v}_x - v_x)/(\tilde{v}_y - v_y))$ is the relative rotation angle between the two agents, and the angle difference $\tilde{\phi} - \psi$ is wrapped between $[-\pi, \pi]$. An illustration of these three penalty sets is provided in Fig. 4.

The parameters defining the penalty set $\mathcal{S}_{norm}$ affect the rate of convergence. With (10)-(12), the SA-CADRL policy converged within 700 episodes (exhibiting the desired behaviors such as passing on the right on all validation test cases). With a 30% smaller penalty set (i.e., shrinking the shaded regions in Fig. 4), convergence occurred after 1250 episodes. Larger penalty sets, however, could lead to instability or divergence. Also, as long as training converges, the penalty sets' size does not have a major effect on the learned policy. This is expected because the desired behaviors are not in the penalty set. Similarly, (9)-(12) can be modified to induce left-handed rules. We trained two SA-CADRL policies to learn left-handed and right-handed norms starting from the same initialization, the results of which are shown in Fig. 6. The learned policies exhibited similar qualitative behaviors as shown in Fig. 2. Also note that training is performed on randomly generated test cases, and not on the validation test cases.

B. Training a Multiagent Value Network

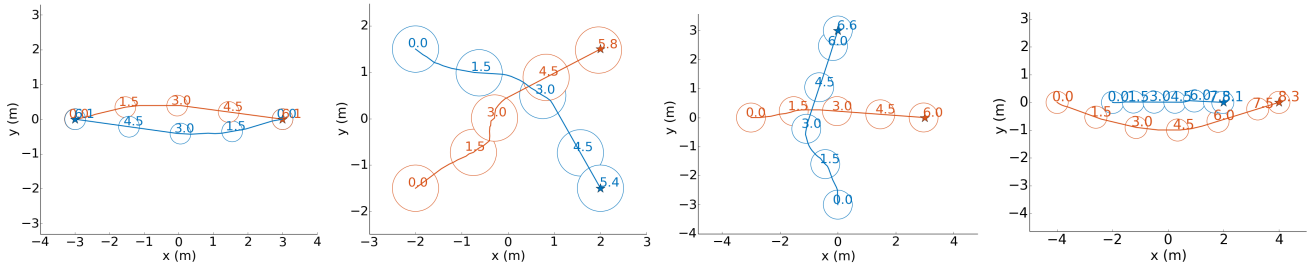
The CADRL work [14] trained a two-agent network with three fully connected hidden layers, and used a minimax scheme for scaling up to multiagent ($n > 2$) scenarios. Since training was solely performed on a two-agent system, it was difficult to encode/induce higher order behaviors, such as accounting for the relations between nearby agents. This work addresses this problem by developing a method that allows for training on multiagent scenarios directly.

To capture the multiagent system's symmetrical structure, a neural network with weight-sharing and max-pooling layers is employed, as shown in Fig. 5. For a four-agent network shown in Fig. 5b, the three nearby agents' observed states can be swapped (blue input blocks) without affecting the output value. This condition is enforced through weight-sharing, as shown in Fig. 5a. Two of such symmetrical layers are used, followed by a max-pooling layer for aggregating features and two fully-connected layers for computing a scalar value. This work uses the rectified linear unit (ReLU) as the activation function in the hidden layers.

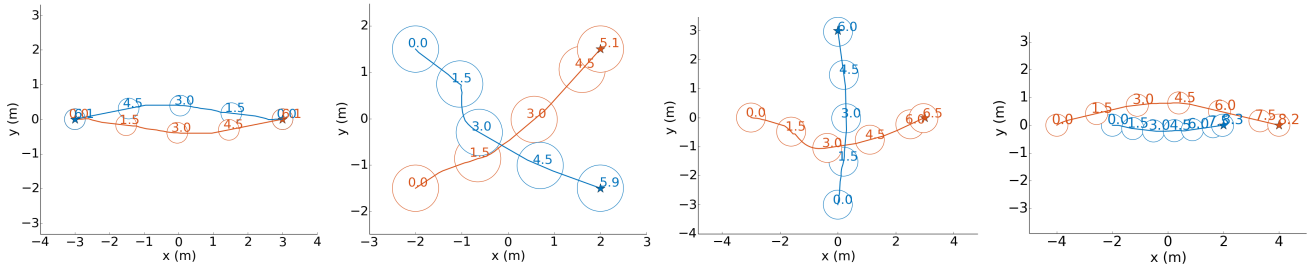
The input to the n-agent network is a generalization of (7)-(8), that is, $\mathbf{s}^{jn} = [\mathbf{s}, \tilde{\mathbf{s}}^{o,1}, \ldots \tilde{\mathbf{s}}^{o,n-1}]$, where the superscripts enumerate the nearby agents. The norm-inducing reward function is defined similarly as (9), where a penalty is given if an agent's joint configuration with the *closest* nearby agent belongs to the penalty set $\mathcal{S}_{norm}$. The overall reward function is the sum of the original CADRL reward and the norm-inducing reward, that is, $R(\cdot) = R_{col}(\cdot) + R_{norm}(\cdot)$.

The procedure for training a multiagent SA-CADRL policy is outlined in Algorithm 1, which follows similarly as in [14], [19]. A value network is first initialized by training on an n-agent trajectory dataset through neural network regression (line 1). Using this value network (6) and following an ϵ-greedy policy, a set of trajectories can be generated on random test cases (line 5-7). The trajectories are then turned into state-value pairs and assimilated into the experience sets E, E_b (line 10-11). A subset of state-value pairs is sampled from the experience sets, and subsequently used to update the value network through back-propagation (line 12-13). The process repeats for a pre-specified number of episodes (line 3-4).

Compared with CADRL [14], two important modifications are introduced in the training process. First, two experience sets, E, E_b, are used to distinguish between trajectories



(a) left-handed rule



(b) right-handed rule

Fig. 6: SA-CADRL policies exhibiting socially aware behaviors. Circles show each agent's position at the labeled time, and stars mark the goals. (a) and (b) show the trajectories generated by SA-CADRL policies trained to learn left-handed and right-handed rules, respectively. These behaviors are time-efficient and agree with the qualitative characterization of social norms shown in Fig. 2. Recall training is performed using randomly generated test cases.

Algorithm 1: Deep V-learning for SA-CADRL

1 initialize and duplicate a value net with n agents $V(\cdot;\theta,n)$, $V' \leftarrow V$
2 initialize experience sets $E \leftarrow \emptyset$, $E_b \leftarrow \emptyset$
3 **for** *episode*=$1,\ldots,N_{eps}$ **do**
4 **for** *m times* **do**
5 $p \sim$ Uniform(2, n)
6 $\mathbf{s}_0^1, \mathbf{s}_0^2, \ldots, \mathbf{s}_0^p \leftarrow$ randomTestcase(p)
7 $\mathbf{s}_{0:t_f}^1, \mathbf{s}_{0:t_f}^2, \ldots, \mathbf{s}_{0:t_f}^p \leftarrow$ SA-CADRL(V)
8 with prob ϵ_f, mirror every traj $\mathbf{s}_{0:t_f}^i$ in the x-axis
9 **for** *every agent i* **do**
10 $y_{0:T}^i \leftarrow \text{findValues}\left(V',\ \mathbf{s}_{0:t_f}^{jn,i}\right)$
11 $E,\ E_b \leftarrow \text{assimilate}\left(E,\ E_b,\ (y^i, \mathbf{s}^{jn,i})_{0:t_f}\right)$
12 $e \leftarrow \text{randSubset}(E) \cup \text{randSubset}(E_b)$
13 $\theta \leftarrow \text{RMSprop}(e)$
14 **for** *every C episodes* **do**
15 Evaluate(V), $V' \leftarrow V$
16 **return** V

that reached the goals and those that ended in a collision (line 2, 11). This is because a vast majority ($\geq 90\%$) of the random generated test cases were fairly simple, requiring an agent to travel mostly straight towards its goal. The bad experience set E_b improves the rate of learning by focusing on the scenarios that fared poorly for the current policy. Second, during the training process, trajectories generated by SA-CADRL are reflected in the x-axis with probability ϵ_f (line 8). By inspection of Fig. 2, this operation flips the paths' topology (left-handedness vs right-handedness). Since a trajectory can be a few hundred steps long according to (6), it could take a long time for an ϵ-greedy policy to explore the state space and find an alternative topology. In particular, empirical results show that, without this procedure, policies can still exhibit the wrong passing side after 2000 training episodes. This procedure exploits symmetry in the problem to explore different topologies more efficiently.

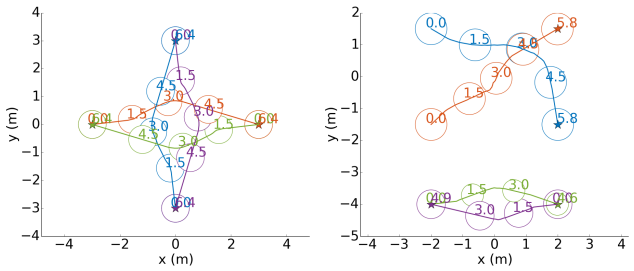
Furthermore, an n-agent network can be used to generate trajectories for scenarios with fewer agents (line 5). In particular, when there are $p \leq n$ agents, the inputs in Fig. 5b corresponding to the non-existent agents[1] can be filled by adding virtual agents – replicating the states of the closest nearby agent and set the binary bit $\tilde{b}_{on}$ to zero (8). The use of this parametrization avoids the need for training many different networks. A left-handed and a right-handed four-agent SA-CADRL policies are trained using the network structure shown in Fig. 5. Sample trajectories generated by these policies are shown in Fig. 7, which demonstrate the preferred behaviors of each respective set of social norms.

IV. Results

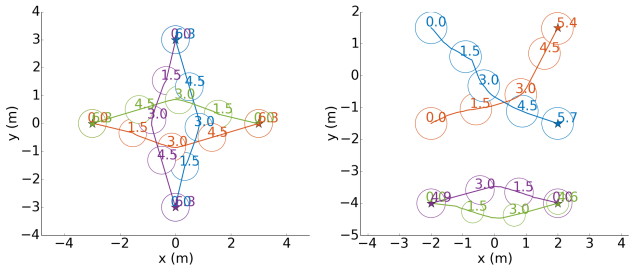
A. Computational Details

The size and connections in the multiagent network shown in Fig. 5 are tuned to obtain good performance (ensure convergence and produce time-efficient paths) while achieving real-time performance. In particular, on a computer with an i7-5820K CPU, a Python implementation of a four-agent SA-CADRL policy takes on average 8.7ms for each query of the value network (finding an action). Furthermore, offline

[1]Consider an agent with two nearby agents using a four-agent network.



(a) left-handed rule



(b) right-handed rule

Fig. 7: SA-CADRL policies generalized to multiagent scenarios using the network structure in Fig. 5. Circles show each agent's position at the labeled time, and stars mark the goals. (a) and (b) show trajectories corresponding to the left-handed and right-handed rules, respectively.

training (Algorithm 1) took approximately nine hours to complete 3,000 episodes. In comparison, a two-agent network took approximately two hours for 1,000 training episodes. The four-agent system took much longer to train because its state space is much larger (higher dimensional) than that of the two-agent system. The training process was repeated multiple times and all runs converged to a similar policy – exhibiting the respective desired social norms on all test cases in an evaluation set. Furthermore, this work generated random test cases with sizes and velocities similar to that of normal pedestrians [25], such that $r \in [0.2, 0.5]$m, and $v_{pref} \in [0.3, 1.8]$m/s. Also, a desired minimum separation of 0.2m is specified through the collision reward R_{col}, which penalizes an agent for getting too close to its neighbors.

B. Simulation Results

Three copies of four-agent SA-CADRL policies were trained, one without the norm inducing reward R_{norm}, one with the left-handed R_{norm}, and the other with the right-handed R_{norm}. On perfectly symmetrical test cases, such as those shown in Figs. 3 and 6, the left and right-handed SA-CADRL policies always select the correct passing side according to the respective norm. To demonstrate that SA-CADRL can balance between finding time-efficient paths and respecting social norms, these policies are further evaluated on randomly generated test cases. In particular, we compute the average extra time to reach the goals[2], the minimum separation distance, and the relative preference between left-handedness and right-handedness. Norm preference is calculated by counting the proportion of trajectories that violate the left-handed or the right-handed version of (10)-(12) for more than 0.5 second. To ensure the test set is left-right balanced, each random test case is duplicated and reflected in the x-axis. Evidently, the optimal reciprocal collision avoidance (ORCA) [6] algorithm – a reactive, rule-based method that computes a velocity vector based on an agent's joint geometry with its neighbors – attains nearly 50-50 left/right-handedness on these test sets (first row of Table I).

The same four-agent SA-CADRL policies are used to generate trajectories for both the two-agent and the four-agent test sets. On the two-agent test set, all RL-based methods produced more time-efficient paths than ORCA[3]. CADRL exhibited a slight preference to the right (40-60 split). The four-agent SA-CADRL (none) policy, in comparison, exhibited a stronger preference than ORCA and CADRL in each of the passing, crossing, and overtaking scenarios (third row in Table I). This observation suggests that (i) certain navigation conventions could emerge as a means of resolving symmetrical collision avoidance scenarios, and (ii) the conventions don't always correspond to human social norms. For instance, SA-CADRL (none) prefers passing on the right but also overtaking on the right, which is a mix between right-handed and left-handed rules. In contrast, the SA-CADRL policies trained with R_{norm} exhibited a strong preference (85-15 split) of the respective social norm. Recall that this ratio is not 1 because there is a tradeoff between time-optimality and social compliance, as illustrated in Fig. 3. This tradeoff can be controlled by tuning q_n in (9). Evidently, SA-CADRL (lh/rh) achieves better social compliance at a cost of an approximately 20% larger $\bar{t}_e$, because satisfying the norms often requires traveling a longer path.

Similarly, the bottom rows of Table I show that in the four-agent test set, all RL-based methods outperformed ORCA, and SA-CADRL (lh/rh) exhibited behaviors that respect the social norms. CADRL produced paths that are closer to time-optimal than the other algorithms, but sometimes came very close (within 0.1m) to other agents. This close proximity occurred because CADRL was trained on a two-agent system, so its action choice is dominated by the single closest neighbor; possibly leading CADRL to select an action that avoids the closest neighbor but drives towards a third agent. In contrast, all SA-CADRL policies were trained on four-agent systems and they all maintained a larger average separation distance.

C. Hardware Experiment

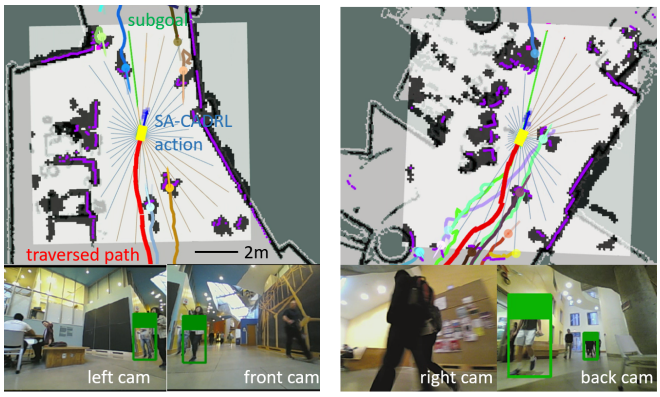
The SA-CADRL policy is implemented on a robotic vehicle for autonomous navigation in an indoor environment with many pedestrians, as shown in Fig. 1. The differential-drive vehicle is outfitted with a Lidar for localization, three Intel Realsenses for free space detection, and four webcams for pedestrian detection. Pedestrian detection and tracking is

[2] $\bar{t}_e = \frac{1}{n}\sum_{i=1}^{n}[t_g^i - ||\mathbf{p}_0^i - \mathbf{p}_g^i||_2 \,/\, v_{pref}^i]$, where t_g^i is the ith agent's time to reach its goal, and the second term is a lower bound of t_g^i (straight toward goal at the preferred speed).

[3] ORCA specifies a reactive, geometric rule for computing a collision-free velocity vector, but it does not anticipate the evolution of an agent's state with respect to other agents nearby. Thus, ORCA can generate shortsighted actions and oscillatory paths (see [14] for a detailed explanation).

TABLE I: SA-CADRL's performance statistics on 200 randomly generated test cases. SA-CADRL policies were trained (i) without the norm inducing reward R_{norm}, (ii) with left-handed R_{norm}, and (iii) right-handed R_{norm}, which are abbreviated as none, lh, rh, respectively. Results show that SA-CADRL policies produced time-efficient paths and exhibited behaviors that respect the corresponding social norm.

Number of agents	Method	Extra time to goal $\bar{t}_e$(s) [Avg / 75th / 90th pctl]	Min separation dist. (m) [10th pctl / avg]	Norm preference (%) [left-handed / right-handed]		
				passing	crossing	overtaking
2	ORCA [6]	0.46 / 0.49 / 0.82	0.108 / 0.131	45 / 55	51 / 49	50 / 50
	CADRL [14]	**0.25 / 0.30 / 0.47**	0.153 / 0.189	37 / 63	38 / 62	43 / 57
	SA-CADRL(none)	0.27 / 0.28 / 0.54	**0.169 / 0.189**	10 / 90	32 / 68	63 / 37
	SA-CADRL(lh)	0.30 / 0.36 / 0.67	0.163 / 0.192	**98** / 2	**85** / 15	**86** / 14
	SA-CADRL(rh)	0.31 / 0.38 / 0.69	0.168 / 0.199	2 / **98**	15 / **85**	17 / **83**
4	ORCA [6]	0.86 / 1.14 / 1.80	0.106 / 0.125	46 / 54	50 / 50	48 / 52
	CADRL(minimax) [14]	**0.41 / 0.54 / 0.76**	0.096 / 0.173	31 / 69	41 / 59	46 / 54
	SA-CADRL(none)	0.44 / 0.63 / 0.85	**0.162 / 0.183**	33 / 67	33 / 67	62 / 38
	SA-CADRL(lh)	0.49 / 0.69 / 1.00	0.155 / 0.178	**83** / 17	**67** / 33	**73** / 27
	SA-CADRL(rh)	0.46 / 1.63 / 1.02	0.155 / 0.180	12 / **88**	29 / **71**	30 / **70**



(a) passing on the right (b) overtaking on the left

Fig. 8: Visualization of a robotic vehicle's sensor data. The vehicle (yellow box) uses the right-handed SA-CADRL policy to navigate autonomously in an indoor environment at a nominal speed of 1.2m/s. (a) shows the vehicle passing a pedestrian on the right, where the person in the front camera image is detected and shown as the blue cylinder in the rviz view. (b) shows the vehicle overtaking a pedestrian on the left, where the person in the right camera image corresponds to the teal cylinder in the rviz view.

achieved by combining Lidar's pointcloud data with camera images [26]. The speed, velocity, and size (radius) of a pedestrian are estimated by clustering the pointcloud data [27]. The estimated radius includes a buffer (comfort) zone as reported in [2], [16]. Obstacles within a 10m $\times$ 10m square (perception range) centered at vehicle are detected and used to populate an occupancy map, shown as the white square in Fig. 8a. Interested readers are referred to [28] for more details on the perception system and hardware construction of the robot.

Motion planning uses the diffusion map algorithm [29] for finding global paths and SA-CADRL for local collision avoidance. In particular, the diffusion map algorithm considers *static* obstacles in the environment to find a subgoal within the vehicle's planning horizon (5m), which is shown at the end of the green line in Fig. 8a. Also, a set of feasible directions (heading and range) is computed, which corresponds to the free space in the occupancy map (white square). The feasible directions are visualized as the thin blue lines emanating from the vehicle. SA-CADRL takes in the set of detected pedestrians shown as the colored disks, and chooses an action (velocity vector) from the feasible directions to move the vehicle toward the subgoal. SA-CADRL's decision is shown as the blue arrow in Fig. 8a, which does not necessarily line up with the subgoal. Note that pedestrians can be detected beyond the 5m static planning horizon, thus allowing socially aware interaction at a longer range. This whole sense-plan-execute cycle is fast enough to operate in real-time at 10Hz on a Gigabyte Brix computer onboard the vehicle.

Using this motion planning strategy, the vehicle was able to navigate fully autonomously in a dynamic indoor environment. In particular, the vehicle is issued randomly generated goals ten times, with an average distance between successive goals of more than 50 meters. During the experiment, an average of 10.2 persons came within 2m of the vehicle each minute. All encountered pedestrians are part of the regular daily traffic, and not testers/personnel associated with this work. For example, there were undergraduate students going between lectures and tourists visiting the campus. At a nominal speed of 1.2m/s, which is approximately the average human walking pace [25], the vehicle maintained safe distance to the pedestrians and generally respected social norms. While a safety driver was walking approximately five meters behind vehicle, he never had to intervene or take over control at any time during the ten runs. Since people in North America follow the right-handed rule, the SA-CADRL policy with right-handed norms is used for the hardware experiment, which causes the vehicle to generally pass pedestrians on the right and overtake on the left. For examples, snippets of the experiment are shown in Fig. 8. A hardware demonstration video can be found at https://youtu.be/CK1szio7PyA.

V. Conclusion

This work presented SA-CADRL, a multiagent collision avoidance algorithm that considers and exhibits socially compliant behaviors. In particular, in a reinforcement learning framework, a pair of simulated agents navigate around each other to learn a policy that respect human navigation norms, such as passing on the right and overtaking on the left in a right-handed system. This approach is further generalized to multiagent ($n > 2$) scenarios through the use of a

symmetrical neural network structure. Moreover, SA-CADRL is implemented on robotic hardware, which enabled fully autonomous navigation at human walking speed in a dynamic environment with many pedestrians. Future work will consider the relationships between nearby pedestrians, such as a group of people who walk together.

ACKNOWLEDGMENT

This work is supported by Ford Motor Company.

REFERENCES

[1] H. Bai, S. Cai, N. Ye, D. Hsu, and W. S. Lee, "Intention-aware online POMDP planning for autonomous driving in a crowd," in *Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 454–460.

[2] H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard, "Socially compliant mobile robot navigation via inverse reinforcement learning," *The International Journal of Robotics Research*, Jan. 2016.

[3] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *IEEE Robotics Automation Magazine*, vol. 4, no. 1, pp. 23–33, Mar. 1997.

[4] J. Van den Berg, M. Lin, and D. Manocha, "Reciprocal velocity obstacles for real-time multi-agent navigation," in *Proceedings of the 2008 IEEE International Conference on Robotics and Automation (ICRA)*, 2008, pp. 1928–1935.

[5] M. Phillips and M. Likhachev, "SIPP: Safe interval path planning for dynamic environments," in *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA)*, May 2011, pp. 5628–5635.

[6] J. van den Berg, S. J. Guy, M. Lin, and D. Manocha, "Reciprocal n-body collision avoidance," in *Robotics Research*, ser. Springer Tracts in Advanced Robotics, C. Pradalier, R. Siegwart, and G. Hirzinger, Eds. Springer Berlin Heidelberg, 2011, no. 70, pp. 3–19.

[7] S. Kim, S. J. Guy, W. Liu, D. Wilkie, R. W. Lau, M. C. Lin, and D. Manocha, "BRVO: Predicting pedestrian trajectories using velocity-space reasoning," *Int. J. Rob. Res.*, vol. 34, no. 2, pp. 201–217, Feb. 2015.

[8] V. V. Unhelkar, C. Pérez-D'Arpino, L. Stirling, and J. A. Shah, "Human-robot co-navigation using anticipatory indicators of human walking motion," in *Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 6183–6190.

[9] P. Trautman, J. Ma, R. M. Murray, and A. Krause, "Robot navigation in dense human crowds: Statistical models and experimental studies of human–robot cooperation," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 335–356, Mar. 2015.

[10] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Physical Review E*, vol. 51, no. 5, pp. 4282–4286, May 1995.

[11] G. Ferrer, A. Garrell, and A. Sanfeliu, "Robot companion: A social-force based approach with human awareness-navigation in crowded environments," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov. 2013, pp. 1688–1694.

[12] G. Ferrer and A. Sanfeliu, "Behavior estimation for a complete framework for human motion prediction in crowded environments," in *Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 5940–5945.

[13] D. Mehta, G. Ferrer, and E. Olson, "Autonomous navigation in dynamic social environments using multi-policy decision making," Oct. 2016.

[14] Y. Chen, M. Liu, M. Everett, and J. P. How, "Decentralized, non-communicating multiagent collision avoidance with deep reinforcement learning," in *Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, 2017.

[15] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the Twenty-First International Conference on Machine Learning*, ser. ICML '04. New York, NY, USA: ACM, 2004, pp. 1–.

[16] B. Kim and J. Pineau, "Socially adaptive path planning in human environments using inverse reinforcement learning," *International Journal of Social Robotics*, vol. 8, no. 1, pp. 51–66, June 2015.

[17] M. Kuderer, H. Kretzschmar, C. Sprunk, and W. Burgard, "Feature-based prediction of trajectories for socially compliant navigation," in *Robotics:Science and Systems*, 2012.

[18] R. A. Knepper and D. Rus, "Pedestrian-inspired sampling-based multi-robot collision avoidance," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, Sept. 2012, pp. 94–100.

[19] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[20] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[21] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," *arXiv:1602.01783 [cs]*, Feb. 2016.

[22] F. Sadeghi and S. Levine, "(CAD)^{2}RL: Real single-image flight without a single real image," *arXiv:1611.04201 [cs]*, Nov. 2016.

[23] T. Zhang, G. Kahn, S. Levine, and P. Abbeel, "Learning deep control policies for autonomous aerial vehicles with MPC-guided policy search," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 528–535.

[24] "The wisdom of crowds," *The Economist*, Dec. 2011.

[25] T. Gates, D. Noyce, A. Bill, and N. Van Ee, "Recommended walking speeds for timing of pedestrian clearance intervals based on characteristics of the pedestrian population," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1982, pp. 38–47, Jan. 2006.

[26] J. Miller, A. Hasfura, S. Y. Liu, and J. P. How, "Dynamic arrival rate estimation for campus Mobility On Demand network graphs," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2016, pp. 2285–2292.

[27] T. Campbell, M. Liu, B. Kulis, J. P. How, and L. Carin, "Dynamic clustering via asymptotics of the Dependent Dirichlet Process mixture," in *Advances in Neural Information Processing Systems (NIPS)*, May 2013.

[28] M. Everett, "Robot designed for socially acceptable navigation," Master Thesis, MIT, Cambridge, MA, USA, June 2017.

[29] Y. Chen, S.-Y. Liu, M. Liu, J. Miller, and J. P. How, "Motion planning with diffusion maps," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Daejeon, Korea, Oct. 2016.