

# Hierarchical Learning with Unsupervised Skill Discovery for Highway Merging Applications

Yigit Gurses, Kaan Buyukdemirci, and Yildiray Yildiz

**Abstract**—Driving in dense traffic with human and autonomous drivers is a challenging task that requires high level planning and reasoning along with the ability to react quickly to changes in a dynamic environment. In this study, we propose a hierarchical learning approach that uses learned motion primitives as actions. Motion primitives are obtained using unsupervised skill discovery without a predetermined reward function, allowing them to be reused in different scenarios. This can reduce the total training time for applications that need to obtain multiple models with varying behavior. Simulation results demonstrate that the proposed approach yields driver models that achieve higher performance with less training compared to baseline reinforcement learning methods.

**Index Terms**—Reinforcement learning, Motion Primitive, Skill Discovery, Hierarchical Learning

## I. INTRODUCTION

In the past decades, applications in which autonomous agents interact with humans have been rapidly growing. In order to ensure such interactions are safe and reliable, an agent needs to have a high level understanding of its environment and be capable of long term planning. To that extent, reinforcement learning (RL) is promising as a solution where agents learn to optimize a given reward function in both near and distant future. Deep RL has solved many challenging problems [1], [2] and achieved super-human performance in many tasks [3], [4]. There is also growing literature on using deep RL in autonomous driving [5]–[9]. A comprehensive review of the subject can be found in [10].

Although RL methods proved useful in solving complex decision making problems through optimizing a reward function, they encounter difficulties in environments with sparse rewards. Road traffic is an example for such an environment, where critical events like motor vehicle collisions occur rarely. One solution for this problem is incorporating domain knowledge through reward shaping [11], [12]. However, this approach introduces human bias into the process and can lead to sub-optimal performance.

Hierarchical RL (HRL) is a branch of RL that can be helpful in solving the sparse rewards problem. HRL deals with the task of decomposing tasks into sub-tasks and using the solutions of sub-tasks in a high level solution. This can be achieved by using a set of policies as actions of a high level agent [13]–[16], or using these policies to enhance a loss function to

improve the training process [17], [18]. In the most common form, HRL reduces the action spaces of high-level policies by employing low-level policies that can be pre-learned [19]. This can be useful in overcoming the difficulties posed by sparse rewards, since low-level policies can be trained using intrinsic dense rewards that are not dependent on the high level goals [20].

Low-level policies used in HRL can be obtained using several methods including unsupervised exploration and pre-training [21]–[24]. One approach that shows promise is unsupervised skill discovery (USD), which focuses on obtaining policies with high variance within themselves and with each other, without a predetermined reward function [25], [26]. One advantage of USD is yielding a transferable skill-set that can be reused in RL tasks with different reward functions. This provides a scalable approach for obtaining multiple models once the initial investment to obtain the skills is made. USD can also be employed in multi-agent settings [28], [29] which traffic is an example of.

In this paper, we propose to harness the potential of USD for obtaining road traffic driving strategies that can be used as driver models or autonomous driving algorithms. We achieve this by generating HRL algorithms that utilize USD to obtain low-level policies. We believe that in complex traffic scenarios like highway merging, which suffers from sparse rewards, this method provides driving strategies that 1) can be transferred to multiple scenarios with minimal effort, and 2) can be created with a reduced computational load.

To summarize, the main contributions of this study are the following.

- 1) An RL training method to obtain transferable latent skills for the highway merging environment.
- 2) An HRL model for driving strategies that utilizes skills as actions to achieve faster convergence and higher performance, compared to conventional RL approaches.

## II. METHOD

In this section, we first provide the minimum basic background that is necessary to grasp the main idea behind unsupervised skill discovery (USD), the foundations of which are laid out in [26], and then explain USD and its employment in hierarchical reinforcement learning (HRL).

### A. Markov Decision Process

A Markov Decision Process (MDP) is a tuple  $(S, A, P, r)$ , whose elements are defined below.

- $S$  is the set of all states.

Y. Gurses is with the Department of Computer Engineering at Bilkent University, K. Buyukdemirci is with the Department of Electrical and Electronics Engineering at Bilkent University, and Y. Yildiz is with the Department of Mechanical Engineering at Bilkent University.

- $A$  is the set of all actions.
- $P : S \times A \times S \rightarrow [0, 1]$  is the function where  $P(s, a, s')$  represents the probability of transitioning to state  $s'$  from state  $s$  if action  $a$  is taken. Note that  $\sum_{s' \in S} P(s, a, s') = 1$  for all  $(s, a) \in S \times A$ .
- $r : S \times A \rightarrow \mathbb{R}$  is the function where  $r(s, a)$  represents the immediate reward for taking action  $a$  in state  $s$ .

### B. Reinforcement Learning

In RL, an agent’s behavior is represented in a policy function  $\pi : S \times A \rightarrow [0, 1]$  where  $\pi(s, a)$  represents the probability of the agent taking action  $a$  in state  $s$ . In an MDP, a tuple  $(s_t, a_t, r_t, s_{t+1})$  is defined as an experience sample and a sequence of experience samples is defined as a trajectory. Given a state-action pair  $(s_t, a_t)$ , the expected cumulative reward over all possible trajectories following  $(s_t, a_t)$  is calculated as

$$Q^\pi(s_t, a_t) = r(s_t, a_t) + \mathbb{E}_{a \sim \pi(s)} \left[ \sum_{i=1}^{\infty} \gamma^i r(s_{t+i}, a_{t+i}) | s_t, a_t \right], \quad (1)$$

where  $\gamma \in [0, 1]$  is the discount factor. The goal of reinforcement learning is to find an optimal policy  $\pi^*$  such that

$$\pi^* = \arg \max_{\pi} Q^\pi(s, a), \forall s \in S, \forall a \in A. \quad (2)$$

$\pi^*$  can be estimated using methods like Q-learning and deep Q-networks (DQN) [27].

### C. Unsupervised Skill Discovery (USD)

Concepts of mutual information and entropy, which are commonplace in derivations of USD, are explained below;

*Mutual Information (I):* Mutual information measures the average quantity of information gained about one variable by sampling the other variable. Intuitively, it measures how much one variable tells us about the other. Given two random variables  $X$  and  $Y$ , their marginal distributions  $p(X)$  and  $p(Y)$ , and joint distribution  $p(X, Y)$ ; the mutual information of  $X$  and  $Y$  is expressed as  $I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$ .

*Entropy (H):* Entropy is used to quantify the randomness or uncertainty of a random variable. Given a random variable  $X$  and its probability distribution  $p(X)$ , entropy of  $X$  is defined as  $H(X) = -\sum_{x \in X} p(x) \log p(x)$ .

The goal of USD is obtaining task agnostic skill policies that explore different subspaces of a state space. While obtaining these skills, the following goals are prioritized.

- We want skills to be inferable from given states. Maximizing the mutual information between skills ( $Z$ ) and states ( $S$ ), i.e. maximizing  $I(S; Z)$ , achieves this goal.
- We do not want skills to be inferable from actions, since there can be actions that do not have a significant effect on the environment. Minimizing mutual information between actions ( $A$ ) and skills given states, i.e. minimizing  $I(A; Z|S)$ , achieves this goal.
- We want each skill to act as randomly as possible and so explore as large of a state space as possible. Maximizing

the entropy ( $H$ ) of actions over states, i.e maximizing  $H[A|S]$ , achieves this goal.

These goals translate to maximizing the following objective function,

$$\begin{aligned} \mathcal{F}(\theta) &\triangleq I(S; Z) + H[A|S] - I(A; Z|S) \\ &= (H[Z] - H[Z; S]) + H[A|S] \\ &\quad - (H[A|S] - H[A|S; Z]) \\ &= H[Z] - H[Z; S] + H[A|S; Z], \end{aligned} \quad (3)$$

where  $\theta$  represents the parameters of our policy  $\pi_\theta$ . To maximize the first term  $H[Z]$ , the probability distribution of skills  $p(Z)$  is selected as uniform distribution.  $p(z)$  translates to the probability of the skill  $z$  being sampled at the beginning of an episode. To maximize the third term, a soft actor critic (SAC) [30] agent that maximizes the entropy of actions  $H[A|S; Z]$ , while simultaneously maximizing the expected reward, is trained. Training is done with the reward function,

$$r_z(s, a) = \log q_\phi(z|s) - \log p(z), \quad (4)$$

where  $q_\phi(z|s)$  is the output of a discriminator network that is trained concurrently with the SAC agent to predict  $p(z|s)$ . This reward increases the inferability of skills from states, and therefore minimizes the second term  $H[Z; S]$ . Once the training is complete, the policy  $\pi_\theta$ , which gives a probability distribution of actions given a state and skill, is obtained. The probability of the agent selecting the action  $a$  given a state-skill pair  $(s, z)$  is calculated as follows,

$$\mathbb{P}(A = a) = \pi_\theta(s, a|z) \quad (5)$$

By selecting a constant skill  $z$ , we we can define a policy  $\pi_z$  such that  $\forall z \in Z; \forall s \in S; \forall a \in A; \pi_z(s, a) = \pi_\theta(s, a|z)$ . In other words,  $\pi_z$  is the policy of skill  $z$  induced from  $\pi_\theta$ .

### D. Hierarchical Learning With Unsupervised Skill Discovery

In HRL the highest level task is defined as  $\Gamma$  and its corresponding task policy that maps the state space to subtasks of  $\Gamma$  is defined as  $\pi_\Gamma$ . A subtask  $\omega$  of  $\Gamma$  is constructed as follows.

- $\pi_\omega$  is the corresponding task policy of  $\omega$  that maps the state space to actions
- $r_\omega$  is the subtask reward function that is used for training  $\pi_\omega$

In this study, we define  $\Gamma$  as the task of reaching the end of the highway region without any crash accidents while staying as close as possible to the desired velocity and headway distances (the details are given in the following sections). Each skill  $z$  defines a subtask of  $\Gamma$  with the corresponding subtask reward  $r_z$  and policy  $\pi_z$ .

### III. TRAFFIC SCENARIO

In this study, we create driving strategies for a mandatory merging scenario in a highway environment. To build a high-fidelity simulation environment, we use real traffic-data that is available online. The details are explained in the following subsections.

#### A. Obtaining Environment Parameters From Real-Life Data

NGSIM I-80 is a data set consisting of vehicle trajectories collected on Eastbound I-80 in San Francisco Bay Area in Emeryville, CA, on April 13, 2005 [31]. The data is collected from a section of the road that is approximately 500 meters, includes six freeway lanes and an on-ramp lane that merges into the freeway (See Fig. 1). The data is recorded for three intervals of 15 minutes.

We use a reconstructed version of NGSIM I-80 [32] where the lateral vehicle coordinates are represented as discrete lane-ids, instead of real valued y-coordinates. We process the data collected on the rightmost lane and the ramp to obtain certain statistics to be used in generating our scenario: The mean and the standard deviation of the headway distance distribution are calculated to be 11.9m and 9.6m, respectively. The mean and the standard deviation of the velocity distribution considering both lanes are calculated as 5.9m/s and 3.3m/s, respectively. The mean and the standard deviation of the acceleration distribution on both lanes are calculated as  $-0.12\text{m/s}^2$  and  $-0.98\text{m/s}^2$ . The dimensions of the road section where the data is collected are used to define the dimensions of the road in the simulated environment, which is explained in the following subsection.

#### B. Environment Construction

The environment consists of a highway lane and an on-ramp lane merging into the highway (see Fig. 2). The total length of the road is 360m. The on-ramp ends at 240m, and the legal merging zone starts at 45m, before which the vehicles are not

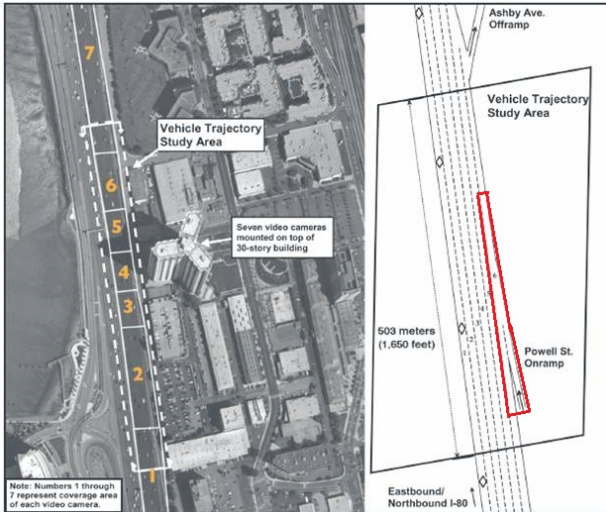


Fig. 1. NGSIM I-80 Study Area, Area of Interest Enclosed in Red Lines

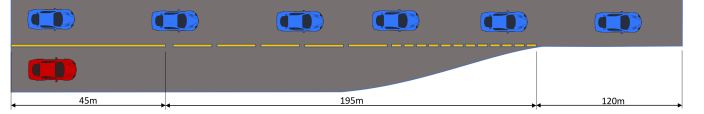


Fig. 2. Highway Merging Environment

allowed to merge. This results in a 195m-long legal merging zone. Lanes have a width of 3.7m. All vehicles are assumed to have a width of 2m and a length of 5m. At the start of an episode,

- the ego vehicle starts at the on-ramp at 0m with a velocity sampled from a uniform distribution in the range  $[2.3\text{m/s}, 3.3\text{m/s}]$ , and
- the highway lane is populated with  $n$  cars where each car has a velocity of 5.9m/s. For  $i \in \{1, 2, \dots, n\}$ ,  $i^{\text{th}}$  car starts at the coordinate  $x_i = 50\text{m} * (i - 1) + x_i^{\text{rand}}$  where  $x_i^{\text{rand}}$  is sampled from uniform distribution in the range  $[-10\text{m}, 10\text{m}]$ .

#### C. State Transition Model

At any time frame, there exists 6 environment vehicles and 1 ego vehicle on the road. Each has a binary variable  $l(t)$  representing their current lanes.  $l(t) = 0$  implies the vehicle is on the highway lane, and  $l(t) = 1$  implies the vehicle is on the on-ramp merging lane. Two continuous variables  $x(t)$  and  $v(t)$  represent x-coordinates and velocities, respectively. At each time frame, vehicles choose an acceleration  $a(t)$ , and a probability of lane change  $l_p(t)$ .

The evolution of longitudinal vehicle states is computed as

$$x(t + \Delta t) = x(t) + v(t) * \Delta t + 1/2a(t) * \Delta t^2, \quad (6)$$

$$v(t + \Delta t) = v(t) + a(t) * \Delta t. \quad (7)$$

For lane change, a number  $p$  is sampled from a uniform distribution in the range  $[0, 1]$  and the new lane is calculated as

$$l(t + \Delta t) = \begin{cases} 0 & l(t) = 0 \\ 0 & l(t) = 1 \ \& \ l_p(t) \geq 0.8 \\ 0 & l(t) = 1 \ \& \ p < l_p(t) \ \& \ 0.2 < l_p(t) < 0.8 \\ 1 & l(t) = 1 \ \& \ p \geq l_p(t) \ \& \ 0.2 < l_p(t) < 0.8 \\ 1 & l(t) = 1 \ \& \ l_p(t) \leq 0.2 \end{cases} \quad (8)$$

This transition algorithm does not allow vehicles on the highway to change lanes and makes the on-ramp vehicles' lane change probabilistic, instead of deterministic. This allows emerging skills to explore larger state spaces.

#### D. Observation Space

The ego vehicle can observe its own velocity,  $v_{agent}$ , and x-coordinate,  $x_{agent}$ , as well as the relative velocities,  $v_{rel}$ , and distances,  $d_{rel}$ , of the surrounding vehicles that are in the front, back, right-front, right-left, back-front and back-left.

These observations are normalized to be in the range  $[0, 1]$  as

$$v_{agent}^{norm} = v_{agent}/v_{max} \quad (9)$$

$$x_{agent}^{norm} = \begin{cases} x_{agent}/x_{env} & x_{agent} < x_{env} \\ 1 & x_{agent} \geq x_{env} \end{cases} \quad (10)$$

$$v_{rel}^{norm} = (v_{rel} + v_{max})/(2 * v_{max}) \quad (11)$$

$$d_{rel}^{norm} = \begin{cases} d_{rel}/d_{max} & d_{rel} < d_{max} \\ 1 & d_{rel} \geq d_{max}, \end{cases} \quad (12)$$

where  $v_{max}$  is the maximum allowed speed, which is set to 29.16m/s,  $x_{env}$  is the length of the highway lane, which is set to 360m, and  $d_{max}$  is the maximum observable distance, which is set to 30m. If there is no vehicle to be observed in a possible location,  $v_{rel}$  is set to  $v_{agent}$ , and  $d_{rel}$  is set to  $d_{max}$ . The end of the merging region is treated as a vehicle with zero velocity.

A real valued state space makes it hard to obtain distinct skills due to the infinitely large space size. To solve this issue and obtain skills as diverse as possible, we quantized each real-valued state into 10 bins.

### E. Action Space

There are 3 different action spaces for three different agents. These agents are called the *skills agent*, the agent that learns skill policies, *low-level Deep Q-Network (DQN) agent*, which is used for comparison purposes, and finally the proposed *high level DQN agent*, which is trained using hierarchical reinforcement learning that uses skill policies as low-level policies (see Section II-D). We explain the training of these agents in detail in the following sections. In this section, we provide the action spaces they use.

- **Skills Agent:** Skills agent has two action selections:  $a_{act}$ , which is a real number in the range  $[-1, 2/3]$ , and  $l_p$ , which is the lane change probability that takes values in the range  $[0, 1]$ . The acceleration  $a(t)$  of the skill agent used in state transitions (6) and (7) is calculated as  $a(t) = a_{act} * a_{max}$ , where  $a_{max} = 4.5m/s^2$  is the maximum allowed acceleration.
- **Low-Level DQN Agent:** This agent selects one of the following actions, where  $Exp[\lambda]$  is defined as the exponential distribution with rate parameter  $\lambda = 0.75$ .
  - **Maintain:** Acceleration  $a(t)$  is sampled from a Laplace distribution with  $\mu = 0$ , and  $b = 0.1$ , in the interval  $[-0.25m/s^2, 0.25m/s^2]$ . Lane change probability  $l_p$  is set to 0.
  - **Accelerate:** Parameter  $a_{act}$  is sampled from  $Exp[\lambda]$ , and then used to set  $a(t)$  to  $\min\{0.25 + a_{act}, 2\}m/s^2$ .  $l_p$  is set to 0.
  - **Decelerate:**  $a_{act}$  is sampled from  $Exp[\lambda]$ , and  $a(t)$  is set to  $\max\{-0.25 - a_{act}, -2\}m/s^2$ .  $l_p$  is set to 0.
  - **Hard-Accelerate:**  $a_{act}$  is sampled from  $Exp[\lambda]$ , and  $a(t)$  is set to  $\min\{2 + a_{act}, 3\}m/s^2$ .  $l_p$  is set to 0.

- **Hard-Decelerate:**  $a_{act}$  is sampled from  $Exp[\lambda]$ , and  $a(t)$  is set to  $\max\{-2 - a_{act}, -4.5\}m/s^2$ .  $l_p$  is set to 0.
- **Merge:**  $a(t)$  is set to 0, and  $l_p$  is set to 1.

- **High-Level DQN Agent:** This agent chooses a skill index  $i \in \{1, 2, \dots, n_{skills}\}$  as an action.  $i$  is translated to the skill vector  $z_i$  by one-hot encoding.  $a(t)$  and  $l_p$  is then sampled from  $\pi_{z_i}(s)$  (see Section II-D).

### F. Reward Function

Reward function  $r$  is the representation of a driver's preferences. Inspired from [16], in this study, it is defined as

$$r = c * w_c + h * w_h + m * w_m + e * w_e + n * w_{nm} + s * w_s, \quad (13)$$

where  $w$  terms are the corresponding weights of each feature. Features are defined below.

- **c:** Collision parameter. Takes the value 1 if the ego vehicle collides with another vehicle or if it reaches the end of the merging region without merging. The parameter gets the value 0, otherwise.
- **h:** Headway parameter. Calculated as

$$h = \begin{cases} -1 & d_{front} < d_{close} \\ 1 - 2 \frac{d_{front} - d_{nom}}{d_{nom} - d_{close}} & d_{close} \leq d_{front} < d_{nom} \\ \frac{d_{front} - d_{nom}}{d_{nom} - d_{close}} & d_{nom} \leq d_{front} < d_{far} \\ 0 & d_{far} \leq d_{front} \end{cases} \quad (14)$$

where  $d_{close} = 2.3m$ ,  $d_{nom} = 11.9m$ , and  $d_{far} = 21.5m$ . These values are defined using the mean and standard deviation obtained from the dataset (see Section III-A). Finally,  $d_{front}$  is the relative distance of the vehicle in front of the ego agent.

- **m:** Velocity parameter. Calculated as

$$m = \begin{cases} \frac{v_{agent} - v_{nom}}{v_{nom}} & v_{agent} \leq v_{nom} \\ \frac{v_{max} - v_{agent}}{v_{max} - v_{nom}} & v_{agent} > v_{nom}, \end{cases} \quad (15)$$

where  $v_{nom}$  is the nominal velocity for the agent which is set to 5.9m/s, the mean velocity observed in the dataset (see Section III-A).

- **e:** Effort parameter. Defined as follows, where  $act$  is the action taken by the agent.

$$e = \begin{cases} -0.25 & \text{if } act = \mathbf{Accelerate} \text{ or } \mathbf{Decelerate} \\ -1 & \text{if } act = \mathbf{Hard-Accelerate} \text{ or } \\ & \mathbf{Hard-Decelerate} \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

- **nm:** "Not Merging" parameter. Equals to -1 when the agent is on the ramp, 0 otherwise. This parameter discourages the ego agent to keep driving on the ramp without merging.

- $s$ : Stopping parameter. Utilized to discourage the agent from making unnecessary stops. Defined as follows, where  $act$  is the action taken by the agent.

$$s = \begin{cases} -1 & \text{if } act \neq \mathbf{Hard-Accelerate} \text{ and} \\ & d_{front} > d_{far} \text{ and } v_{agent} < v_{nom} \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

where  $d_{front}$  is the relative distance of the vehicle in front of the ego agent.

#### IV. TRAINING AN SIMULATIONS

##### A. Model Initialization

All network weights are initialized with Xavier normal initialization [33] with a scaling of 1.

1) *Skills Agent*: Skills agent consists of a policy network, value network, discriminator network and two Q-value networks. Each network has,

- two fully connected hidden layers with 64 neurons and leaky-ReLU activation function with a slope of 1/100.
- a fully connected output layer with no activation function.

An empty replay buffer  $M_{skill}$  with size  $n_{buffer} := 10000$  is initialized. Number of skills to be learned,  $n_{skills}$ , is set to 16.

2) *DQN Agents*: Both the low-level and high-level DQN networks are implementations of double DQN [34]. Q-value networks for both agents have

- two fully connected hidden layers with 64 neurons and leaky-ReLU activation function with a slope of 1/100.
- a fully connected output layer with no activation function.

Empty replay buffers  $M_{low}$  and  $M_{high}$  with size  $n_{buffer}$  are initialized for the low-level and high-level agents respectively. Exploration rates  $\varepsilon_{low}$  and  $\varepsilon_{high}$  are initialized to 1 and the decay rate for both is  $\beta := 0.99$ .

##### B. Training Skill Policies

After the initialization of the networks, the skill policy is trained for 5000 episodes, following the algorithm provided in [26].

The skill policies obtained at the end of this process vary in complexity and usefulness. Some policies go to a specific coordinate and come to a stop. Some keep speeding up until a collision with a vehicle or until the agent reaches the end of the merging region without merging. It is noted that a skill that permits a collision can be useful if the high level agent learns to pair this skill with another one, where the combined skill-set provides a desired trajectory without collision.

Two examples of the skill policies are given in Figures 3 and 4. In these figures, the red and the blue rectangles represent the ego vehicle and the environment vehicles, respectively. The red vertical line is the coordinate where the ego vehicle makes a lane change. The top frame is the starting frame and each following row represents the progression of the scenario, from top to bottom.

In figure 3, the ego vehicle merges in front of a vehicle in the highway lane and leads it until the end of the road. Such

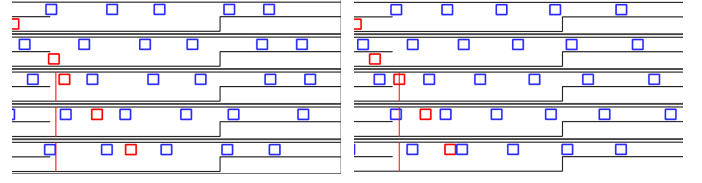


Fig. 3. Example skill 1.

Fig. 4. Example skill 2.

---

#### Algorithm 1 Training High-Level DQN Agent

---

```

1:  $U := 0$ 
2: for  $e = 1$  to  $E$  do
3:   Initialize environment with  $N$  vehicles
4:    $done := false$  ;  $t := 0$ 
5:   while not  $done$  do
6:     get state  $s$  from the environment
7:     sample action  $z$  from  $QN^{high}(s)$  greedily
8:      $i := 1$  ;  $r_{sum} := 0$  ;  $s_{temp} := s$ 
9:     while  $i \leq n_{step} \wedge$  not  $done$  do
10:      sample  $a$  and  $l_p$  from  $\pi_{\theta}(s_{temp}, z)$ 
11:      transition to  $s'$  with  $a$  and  $l_p$ 
12:       $t := t + \Delta t$  ;  $i := i + 1$  ;  $s_{temp} := s'$ 
13:      observe reward  $r$ 
14:       $r_{sum} := r_{sum} + r$ 
15:      if  $s'$  is terminal  $\vee t \geq t_{max}$  then
16:         $done := true$ 
17:      store current experience  $(s, z, r_{sum}/i, s')$  in  $M_{high}$ 
18:      if  $|M_{high}| \geq n_{buffer}$  then
19:        sample P-sized batch from  $M_{high}$ 
20:        for  $(s_i, z_i, r_i, s'_i)$  in batch do
21:          if  $s'_i$  is terminal then
22:             $y_i := r_i$ 
23:          else
24:             $y_i := r_i + \gamma \max_z QN_{tar}^{high}(s'_i, z)$ 
25:             $\mathcal{L}_i := (y_i - QN^{high}(s_i, z_i))^2$ 
26:          Perform gradient descent using  $\mathcal{L} := \frac{1}{P} \sum_{i=1}^P \mathcal{L}_i$ 
          on primary weights  $\theta^{high}$ 
27:           $U := U + 1$ 
28:          if  $U \equiv 0 \pmod{n_{update}}$  then
29:             $\theta_{tar}^{high} := \theta^{high}$ 
30:           $\varepsilon_{high} := \max\{\varepsilon_{high} * \beta, \varepsilon_{min}\}$ 

```

---

a policy may be useful when the gap between highway-lane vehicles are small.

In figure 4, the ego vehicle speeds up and merges in the middle of two vehicles, then keeps speeding up until it collides with the vehicle in front. As explained above, this policy may be useful for a high-speed, high-level DQN agent if the agent learns to pair this skill with another one that slow down or maintain velocity when there is a close vehicle in front.

##### C. Training Low and High Level DQN Policies

Networks, exploration rates and memory buffers for DQN agents are initialized as explained above in Section IV-A. Both the low-level and high-level DQN agents are trained for 9000 episodes. The algorithm we used to train the proposed

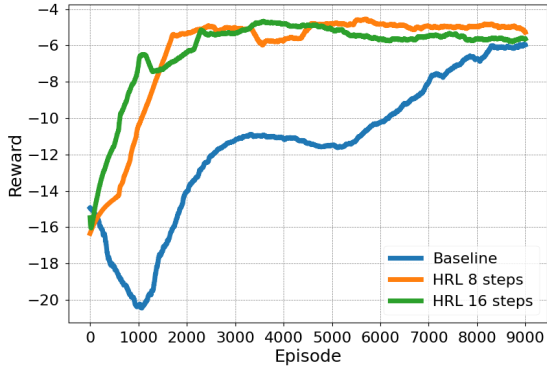


Fig. 5. Running average of rewards, using a 1000-episode window.

high-level DQN agent, that uses skills as actions, is given in Algorithm 1. In the Algorithm,  $n_{step}$  is the number of steps each selected skill is executed consecutively and  $n_{update}$  is the number of updates on primary network required to update the target network.

#### D. Results

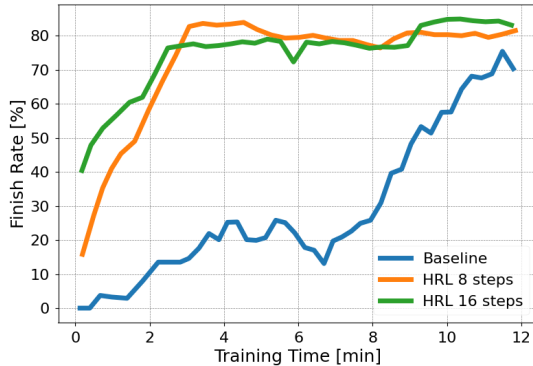


Fig. 6. Running average of success rate in time, using a 10-point window. Models are evaluated every 20000 time-steps.

In Figure 5, the running average rewards of the baseline (Low-level DQN) and the proposed hierarchical reinforcement learning, HRL, (High-level DQN using skills) are displayed. The hierarchical model is trained using two different "step" intervals, namely "HRL 8 steps" and "HRL 16 steps". These intervals indicate the number of steps the model is allowed to run before it can update its action/skill. These intervals are necessary because a skill inherently consists of more than one low-level action. It is observed that the hierarchical model achieves higher average rewards within fewer episodes, and performs slightly better than the baseline model after the final 9000th episode. It should also be noted that the training process of the hierarchical model seems to be more stable. In the beginning, the baseline model oscillates between high and low rewards whereas the hierarchical model converges faster

and shows relatively smaller oscillations, for both the 8-step and 16-step cases.

In Figure 6, a running average of "successful finish percentage" of the models are displayed. A successful finish is defined as the ego vehicle reaching the end of the highway lane without any collisions. Every 20000 time-steps, 500 episodes are run to obtain the success percentage for the model. It can be seen that the hierarchical model not only reaches a higher success rate much faster than the baseline model, for both 8- and 16-step cases, but it also ends up converging to a meaningfully higher success rate at the end of training.

Overall, we see that the hierarchical model using skills achieves higher performance faster than a conventional RL model. We believe that the true potential of the skills-based hierarchical RL lies in the fact that these skills are transferable, and thus can be used in different traffic scenarios. When skills are transferred, high-performing hierarchical models can be obtained much faster compared to conventional RL methods.

#### V. CONCLUSION

In this study, a hierarchical learning model that uses skills as actions is proposed for obtaining driving strategies. These skills can be obtained with unsupervised skill discovery without a predefined reward function that is specifically designed for the preferences of the high-level agent. This allows the skills to be reused in scenarios with differing reward functions to generate driving strategies with divergent behavior. Our simulation results show that, after an initial computational investment to learn the skills, policies with higher performance can be obtained with less training compared to baseline reinforcement learning methods. The fact that learned skills can be reused in different scenarios provide further value to the proposed approach in applications where many differing policies need to be obtained.

It can also be argued that selecting skills as actions to be applied for a set amount of time instead of selecting primitive actions at each time step is closer to human behavior. Humans have a high reaction time and take time to analyze the changes in the situation to decide what to do next. Also, humans do not think about specific motor skills such as turning the wheel, and instead make high-level plans and execute the low-level steps subconsciously. These reasons provide a motivation to use the proposed hierarchical model with skills for human behavior modeling in further studies.

#### REFERENCES

- [1] OpenAI et al., "Dota 2 with large scale deep reinforcement learning," arXiv [cs.LG], 2019.
- [2] J. Schrittwieser et al., "Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model," ArXiv [cs.LG], 2020.
- [3] D. Silver et al., "Mastering the game of go without human knowledge," Nature, vol. 550, no. 7676, pp. 354–359, 2017.
- [4] D. Silver et al., "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play," Science, vol. 362, no. 6419, pp. 1140–1144, 2018.
- [5] A.-E. Sallab, M. Abdou, E. Perot, and S. Yogamani, "Deep reinforcement learning framework for autonomous driving," Electronic Imaging, vol. 2017, no. 19, pp. 70–76, 2017. 11

- [6] P. Wang, C.-Y. Chan, and A. de La Fortelle, "A reinforcement learning based approach for automated lane change maneuvers," in 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018, pp. 1379–1384. 11
- [7] J. Chen, B. Yuan, and M. Tomizuka, "Model-free deep reinforcement learning for urban autonomous driving," in 2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, 2019, pp. 2765–2771. 9
- [8] S. Kardell and M. Kuosku, "Autonomous vehicle control via deep reinforcement learning," Master's thesis, Chalmers University of Technology, 2017. 9
- [9] C. Li and K. Czarnecki, "Urban driving with multi-objective deep reinforcement learning," in Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 359–367. 9, 10
- [10] B. R. Kiran et al., "Deep reinforcement learning for autonomous driving: A survey," IEEE Trans. Intell. Transp. Syst., vol. 23, no. 6, pp. 4909–4926, 2022.
- [11] H. Sowerby, Z.-H. Zhou, M. Littman, "Designing Rewards for Fast Learning," arXiv [cs.LG], 2022.
- [12] T. Okudo, S. Yamada, "Subgoal-Based Reward Shaping to Improve Efficiency in Reinforcement Learning," IEEE Access, vol. 9, 2021
- [13] J. Randlov, "Learning macro-actions in reinforcement learning," Advances in Neural Information Processing Systems, vol. 11, 1998.
- [14] A. S. Vezhnevets et al., "FeUdal Networks for hierarchical reinforcement learning," arXiv [cs.AI], pp. 3540–3549, 06–11 Aug 2017.
- [15] P.-L. Bacon, J. Harb, and D. Precup, "The Option-Critic Architecture," Proc. Conf. AAAI Artif. Intell., vol. 31, no. 1, 2017.
- [16] Koprulu, C., Yildiz, Y., "Act to Reason: A Dynamic Game Theoretical Driving Model for Highway Merging Applications," IEEE Conference on Control Technology and Applications, 2021
- [17] S. Schmitt et al., "Kickstarting deep reinforcement learning," arXiv [cs.LG], 2018.
- [18] M. Matthews, M. Samvelyan, J. Parker-Holder, E. Grefenstette, and T. Rocktäschel, "Hierarchical Kickstarting for skill transfer in reinforcement learning," arXiv [cs.LG], 2022.
- [19] A. van den Bosch, B. Hengst, J. Lloyd, R. Miikkulainen, and H. Blockeel, "Hierarchical Reinforcement Learning," in Encyclopedia of Machine Learning, Boston, MA: Springer US, 2011, pp. 495–502.
- [20] S. Pateria, B. Subagdja, A.-H. Tan, and C. Quek, "Hierarchical Reinforcement Learning: A comprehensive survey," ACM Comput. Surv., vol. 54, no. 5, pp. 1–35, 2022.
- [21] H. Liu and P. Abbeel, "Behavior from the void: Unsupervised Active Pre-training," arXiv [cs.LG], 2021.
- [22] L. Lee, B. Eysenbach, E. Parisotto, E. Xing, S. Levine, and R. Salakhutdinov, "Efficient Exploration via State Marginal Matching," arXiv [cs.LG], 2019.
- [23] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," arXiv [cs.LG], 2017.
- [24] D. Yarats, R. Fergus, A. Lazaric, and L. Pinto, "Reinforcement Learning with prototypical representations," arXiv [cs.LG], 2021. Learning, 2021
- [25] A. Sharma, S. Gu, S. Levine, V. Kumar, and K. Hausman, "Dynamics-aware unsupervised discovery of skills," arXiv [cs.LG], 2019.
- [26] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine, "Diversity is All You Need: Learning skills without a reward function," arXiv [cs.AI], 2018.
- [27] F. Tan, P. Yan, x. Guan, "Deep Reinforcement Learning: From Q-Learning to Deep Q-Learning," International Conference on Neural Information Processing, 2017
- [28] J. Yang, I. Borovikov, and H. Zha, "Hierarchical cooperative multi-agent reinforcement learning with skill discovery," arXiv [cs.LG], 2019.
- [29] S. He, J. Shao, and X. Ji, "Skill Discovery of coordination in multi-agent reinforcement learning," arXiv [cs.MA], 2020.
- [30] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," arXiv [cs.LG], 2018.
- [31] J. Colyar and J. Halkias, "Interstate 80 freeway dataset," Interstate 80 freeway dataset, FHWA-HRT-06-137, Dec-2006. [Online]. Available: <https://www.fhwa.dot.gov/publications/research/operations/06137/>. [Accessed: 28-Jan-2023].
- [32] M. Montanino and V. Punzo, "Trajectory data reconstruction and simulation-based validation against macroscopic traffic patterns," Trans. Res. Part B: Methodol., vol. 80, no. C, pp. 82–106, 2015.
- [33] X. Glorot, Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," International Conference on Artificial Intelligence and Statistics, 2010
- [34] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with Double Q-learning," Proc. Conf. AAAI Artif. Intell., vol. 30, no. 1, 2016.