Human-Like Decision-Making for Automated Driving in Highways

David Sierra González, Mario Garzón, Jilles Dibangoye, Christian Laugier

▶ **To cite this version:**

David Sierra González, Mario Garzón, Jilles Dibangoye, Christian Laugier. Human-Like Decision-Making for Automated Driving in Highways. ITSC 2019 - 22nd IEEE International Conference on Intelligent Transportation Systems, Oct 2019, Auckland, New Zealand. pp.1-8. hal-02188235

HAL Id: hal-02188235
https://inria.hal.science/hal-02188235

Submitted on 18 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Human-Like Decision-Making for Automated Driving in Highways

David Sierra González, Mario Garzón, Jilles Steeve Dibangoye, and Christian Laugier

***Abstract*— In this work, we present a decision-making system for automated vehicles driving in highway environments. The task is modeled as a Partially Observable Markov Decision Process, in which the physical states and intentions of surrounding traffic are uncertain. The problem is solved in an online fashion using Monte Carlo tree search. At each decision step, a search tree of beliefs is incrementally built and explored in order to find the current best action for the ego-vehicle. The beliefs represent the predicted state of the world as a response to the actions of the ego-vehicle and are updated using an interaction- and intention-aware probabilistic model. To estimate the long-term consequences of any action, we rely on a lightweight model-based prediction of the scene that assumes risk-averse behavior for all agents. We refer to the proposed decision-making approach as human-like, since it mimics the human abilities of anticipating the intentions of surrounding drivers and of considering the long-term consequences of their actions based on an approximate, common-sense, prediction of the scene. We evaluate the proposed approach in two different navigational tasks: lane change planning and longitudinal control. The results obtained demonstrate the ability of the proposed approach to make foresighted decisions and to leverage the uncertain intention estimations of surrounding drivers.**

I. INTRODUCTION

Sharing the road with humans constitutes, along with the need for robust perception systems, one of the major challenges holding back the large-scale deployment of automated driving technology. This statement holds true even in highway scenarios, where experimental autonomous vehicles continue to be disengaged by the human operators due to their inability to anticipate and react adequately to the maneuvers of surrounding human drivers [1].

The actions taken by human drivers are determined by a complex set of interdependent factors, which are very hard to model (e.g. intentions, perception, emotions). As a consequence, any prediction of human behavior will always be inherently uncertain, and becomes even more so as the prediction horizon increases. Moreover, current perception systems can only provide noisy observations of the state of the world. Fully automated vehicles are thus required to make navigation decisions based on the uncertain states and intentions of surrounding vehicles.

Despite the evident complexity of the task, humans excel at interpreting the motion cues of other drivers and at taking anticipatory driving actions. A common approach in the literature to replicate this human ability is to formulate the driving task as a Partially Observable Markov Decision Process (POMDP), which is a principled framework for sequential decision-making under uncertainty. Unfortunately, planning with POMDPs is computationally expensive. Existing POMDP-based planning approaches for automated driving differ mainly in the trade-offs made to render the problem tractable, namely, which uncertainties are considered, how the problem is discretized, and which assumptions are made about the dynamics of the world (Table I).

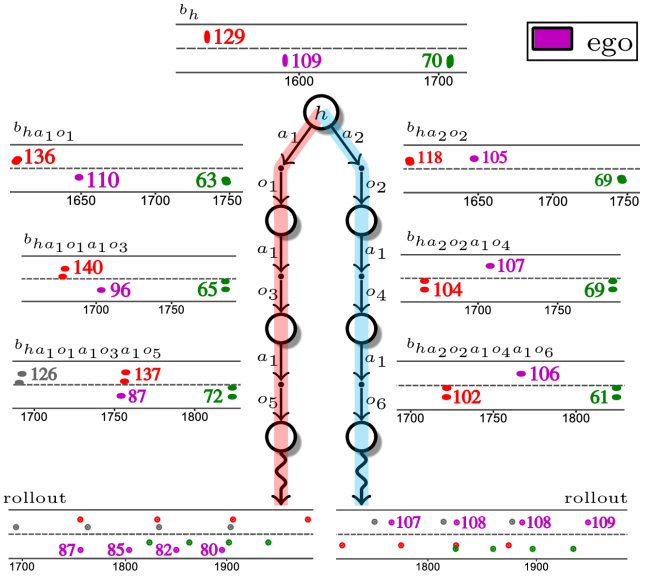


Fig. 1: Exemplary tree simulations illustrating the proposed approach. Each belief node is indexed by a history of actions and observations. For each belief, the location of all vehicles is shown as a number of Gaussians; their mean velocities in km/h are also indicated. The ego-vehicle's planner can select action a_1 (lane keeping) or a_2 (change lanes). To estimate the long-term consequences, each simulation ends with a model-based rollout, where the color dots represent successive predicted locations of the vehicles in the scene.

Ulbrich and Maurer propose to model the highway lane-changing decision task using a small POMDP with 8 possible states [2]. The states correspond to binary random variables representing whether a lane change is possible, beneficial, or currently in progress. This approach relies heavily on custom hand-engineered models (observation, reward, transition), thereby limiting its scalability to complex scenarios.

Instead of discretizing the state using heuristics, Brechtel et al. propose to automatically and iteratively learn a low-level, discrete state-space representation [3]. This approach is applied to planning in intersection merging scenarios. However, the uncertainty in the intentions of surrounding vehicles is not explicitly considered.

In contrast, Bandyopadhyay et al. propose to consider the

This work was supported by Toyota Motor Europe.

All authors are with Inria, Univ. Grenoble-Alpes, Grenoble, France {david.sierra-gonzalez, mario.garzon-oviedo, jilles.dibangoye, christian.laugier}@inria.fr

Reference	Uncertainty			Discretization			Offline/ Online	Bi-directional interactions	Approach
	C	X	I	S	A	O			
Ulbrich et al. [2]	✗	✓	✗	D	D	D	Online	✗	RTBSS
Brechtel et al. [3]	✓	✓	✗	C	D	C	Offline	✓ DBN	MCVI
Liu et al. [4]	✓	✓	✓	D	D	D	Online	✓ Hard-coded	DESPOT
Bandyopadhyay et al. [5]	✓	✗	✓	D	D	D	Offline	✓ pedestrian model	MOMDP, SARSOP
Hubmann et al. [6]	✓	✓	✓	C	D	C	Online	✓ Longitudinal custom model	ABT
Sunberg et al. [7]	✓	✗	✓	C	D	C	Online	✓ IDM + MOBIL	POMCP-DPW
Bouton et al. [8]	✓	✓	✓	C	D	C	Online	✗	POMCP-DPW + IMM
Proposed	✓	✓	✓	C	D	C	Online	✓ DBN + driver mod.	POMCP-PW + DBN

TABLE I: Existing POMDP-based approaches for decision-making in automated vehicles. The uncertainty can be considered in the controls (C), physical states (X) or manuever intentions (I).

uncertainty only in the intentions of surrounding traffic participants [5]. In particular, they address an scenario in which the ego-vehicle circulates in the presence of pedestrians. The pedestrians' motion is dependent on their intended destination and on the state of the ego-vehicle. The planning task is modeled as a Mixed Observability MDP (MOMDP), where the interactions between pedestrians are not considered. The authors suggest solving a different MOMDP for each pedestrian in the scene, and choose the most conservative action for the ego-vehicle. In consequence, this approach cannot be directly applicable to highway driving where the behavior of traffic participants is highly correlated.

Liu et al. propose instead to infer the motion of traffic participants by exploiting the road context, classifying their motion intentions into stopping, hesitating, normal or aggressive [4]. Then, a POMDP is used to model the intersection navigation task, considering the uncertainties in the motion intentions of other drivers. A drawback of this approach is that the interactions between vehicles are only considered through a series of hard-coded rules specific for intersection scenarios. A similar POMDP-based approach for intersection navigation was proposed by Hubmann et al. [6]. In this case, the interactions are also formulated as hard-coded rules.

Similarly, Bouton et al. also propose a POMDP planner for the task of navigating unsignaled intersections [8]. The state is factored into the physical state of the ego-vehicle, the physical state of all other vehicles, and their maneuver intentions. The POMDP is solved online using the Monte-Carlo-based algorithm POMCP [9] with double progressive widening to control the branching factor. At each execution step, a Gaussian belief over the physical states of the obstacles and over their intentions is maintained using an Interacting Multiple Mode (IMM) filter with two modes. This is the same model that is used to simulate the dynamics of the system and constitutes the main drawback of this approach, since the IMM completely disregards the interactions between vehicles. In consequence, the search tree is built with observation samples that do not capture the essence of human driving, which might explain some of the collision rates this approach achieved in simulated scenarios.

POMCP with double progressive widening has also been applied to the task of highway navigation [7]. In this case, the behavior of surrounding traffic is modeled using the IDM car-following model for lane keeping [10], and the MOBIL gap-acceptance model for lane changes [11]. The only uncertainty considered is in the parameters of the models, which have a direct incidence in the aggressiveness of the drivers. As two risk-averse models are used to model the dynamics, no car accidents will ever be predicted, and in consequence the planner will never be able to take anticipative action when facing potentially dangerous situations.

In this paper, we also model the highway navigation task using the POMCP algorithm with progressive widening for the observations. We consider the uncertainty in the physical state and lane changing intentions of surrounding vehicles. To model the dynamics of the world, we rely on a Dynamic Bayesian Network (DBN), which enables us to model the interactions between traffic participants [12]. This DBN model can detect and predict dangerous traffic situations, which enables the proposed planner to make anticipatory driving decisions, much like human drivers would do. Furthermore, in the highway it is particularly important to take into account the long-term (5 seconds into the future and beyond) consequences of any action. With this idea in mind, we model the dynamics of the world—beyond a given point in future—using an interaction-aware, model-based prediction method. This method produces sensible long-term predictions of highway traffic scenes, providing valuable insights about the long-term optimality of the available driving actions.

II. BACKGROUND

A. POMDPs

In a Markov Decision Process (MDP) framework, an agent interacts with a given environment by taking actions at discrete time steps. Upon taking an action, the state of the world changes and the agent receives a cost signal. In this setting, the goal of the agent is to take actions so as to minimize the long-term expected cumulative amount of cost it receives. In an MDP, the environment's dynamics are fully determined by the current state s, which is always known.

POMDPs extend the MDP framework to environments that cannot be observed directly. Instead, the agent receives noisy or incomplete observations of the state. Since the true state is unknown, the agent can select its actions based on the history $h_t \doteq \{a_0, o_1, \ldots, a_{t-1}, o_t\}$. Alternatively, it is possible to maintain an estimate of the state of world, known as the belief distribution, and map instead beliefs to actions. The belief $b_t(s) \doteq P(s_t = s|o_t, a_{t-1}, o_{t-1}, \ldots, a_0, b_0)$ is a sufficient statistic for the history. The policy is then $\pi : \mathcal{B} \mapsto \mathcal{A}$. The value function $V_\pi(b)$ is the expected cost accrued from belief b when following policy π. The optimal

value function is the minimum value function achievable by any policy $V^*(b) = \min_\pi V_\pi(b)$, from where the optimal policy could be extracted using a one time step lookahead.

B. Offline vs online planning

Offline POMDP approaches aim to find, prior to execution, an optimal policy under consideration of all possible future situations. In contrast, online approaches alternate planning and execution stages. Once the planning time runs out, the agent executes the best action found for the current belief, receives an observation, and restarts the planning stage from the new belief state. In POMCP, as in our approach, Monte Carlo tree simulations are used to approximate the value function for the current belief during the planning stage.

III. PROPOSED APPROACH

In this section, we formalize the proposed POMDP model for decision-making in highways.

A. POMDP model

The proposed POMDP model is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{C}, \mathcal{Z}, \mathcal{O} \rangle$ where:

A **state** $s \in \mathcal{S}$ contains:

- The ego-vehicle's physical state $\mathbf{x}^e = [x^e, y^e, \dot{x}^e]^T \in \mathbb{R}^3$, where x^e, y^e represent the longitudinal and lateral positions of the vehicle's center of mass in road coordinates, and $\dot{x}^e$ represents the longitudinal velocity.
- The surrounding vehicles' (hereafter, obstacles) physical states $\mathbf{x}^{1:N-1}$, each of them with the same components as the ego-vehicle's. We assume the constant presence of $N-1$ obstacles throughout any traffic scene considered. The space of all joint physical states of the obstacles is denoted by $\mathcal{X}$.
- The maneuver intentions of the obstacles $m^{1:N-1} \in \{\text{LCL, LCR, LK }\}^{N-1}$, where LK represents a lane keeping maneuver, and LCL/LCR represent lane change maneuvers to the left and to the right, respectively.

That is, at any discrete time step t, the state s_t is given by $s_t = [\mathbf{x}_t^e, \mathbf{x}_t^{1:N-1}, m_t^{1:N-1}]$.

The **actions** $m_t^e \in \mathcal{A}$ correspond to the maneuvers of the ego-vehicle. We consider two different experimental scenarios, each with a different action space:

1) In the first scenario, the ego-vehicle can perform lane changes, but its acceleration is set automatically using the IDM car-following model. That is, $\mathcal{A}_1 = \{\text{LCL, LCR, LK}\}$.
2) In the second scenario, the ego-vehicle drives in a given lane and it needs to adjust its speed so as to drive comfortably and safe. The action space for this scenario consists of three different discrete acceleration values $\mathcal{A}_2 = \{-1, 0, 1, \}$ $\mathrm{m/s^2}$.

The **state transition function** $\mathcal{T}(s, a, s')$ is given by the dynamics of the system, detailed in subsection III-B.

The **cost function** $\mathcal{C}^i : \mathcal{X} \mapsto \mathbb{R}$ provides the cost associated to a particular joint physical state of all vehicles in the scene, from the point of view of the ith vehicle. It is a linear function on a set of selected features, which are also calculated from the point of view of the ith vehicle:

$$\mathcal{C}^i([\mathbf{x}^e, \mathbf{x}^{1:N-1}]) = \mathbf{w}^T \mathbf{f}^i([\mathbf{x}^e, \mathbf{x}^{1:N-1}]) \tag{1}$$

The selected features include the lane index, Gaussians modeling the time-to-collision and time-headway to the leading and trailing vehicles in the same lane and the deviation from the desired velocity. The weight parameters balancing the importance of the difference features were learned from human-driven demonstrations using Inverse Reinforcement Learning [13].

The **observation** $\mathbf{z}^i$ of the physical state of any vehicle i is composed of noisy versions of all the components in its physical state. The intentions of surrounding traffic are not observable. The joint observation at each time step is $z_t = [\mathbf{z}_t^e, \mathbf{z}_t^{1:N-1}]$. The continuous space of all possible observations is denoted by $\mathcal{Z}$.

The **observation function** $\mathcal{O} : \mathcal{S} \times \mathcal{A} \mapsto \mathsf{P}(\mathcal{Z})$ is given by the predictive step of the DBN that models the interactions between traffic participants. Given an action and resulting state, it returns a probability distribution over the possible observations. Further details are provided in subsection III-D.

B. Dynamics and maneuvers

The interactions between traffic participants are modeled using a DBN model, following previous work [12]. The factorization of the joint distribution of the model is the following:

$$P\big(\mathbf{x}_{1:T}^{1:N}, m_{1:T}^{1:N}, \mathbf{z}_{1:T}^{1:N}\big) = P\big(\mathbf{x}_1^{1:N}, m_1^{1:N}, \mathbf{z}_1^{1:N}\big) \tag{2}$$
$$\prod_{t=2}^{T}\prod_{i=1}^{N}[P(\mathbf{x}_t^i | m_{t-1:t}^i, \mathbf{x}_{t-1}^{1:N})\ P(m_t^i | m_{t-1}^{1:N}, \mathbf{x}_{t-1}^{1:N})\ P(\mathbf{z}_t^i | \mathbf{x}_t^i)]$$

where T indicates the number of time steps considered.

The term $P(\mathbf{x}_t^i | m_{t-1:t}^i, \mathbf{x}_{t-1}^{1:N})$ describes the maneuver-dependent dynamics of the ith vehicle. In this paper, we treat vehicles as point objects. The point dynamics satisfy the following equations:

$$\begin{aligned} x_{t+1} &= x_t + \Delta t\ \dot{x}_t \\ y_{t+1} &= y_t + \Delta t\ v_{\text{lat}} \\ \dot{x}_{t+1} &= \dot{x}_t + \Delta t\ a_{\text{long}} \end{aligned} \tag{3}$$

where a_{long} and v_{lat} are the longitudinal acceleration and lateral velocity control inputs, which are maneuver-dependent values. The longitudinal acceleration is set for all vehicles using the IDM car-following model, except for the ego-vehicle in the second experimental setting. For a lane change maneuver, the lateral velocity is set to a fixed value v_{LC} until the vehicle reaches the centerline of the target lane, moment at which it is set to 0.

The motion is assumed to be perturbed by Gaussian noise to account for modeling errors, that is:

$$P(\mathbf{x}_t^i | \mathbf{x}_{t-1}^{1:N}, m_t^i{=}m)\ \sim\ \mathcal{N}\Big(\mathbf{g}_m\big(\mathbf{x}_{t-1}^{1:N}\big), \boldsymbol{Q}_m\Big) \tag{4}$$

where $\mathbf{g}_m$ is a maneuver-dependent predictive function that integrates (3) over an interval of time Δt to obtain the next state, and $\boldsymbol{Q}_m$ is the noise covariance matrix associated to maneuver m.

The term $P(\mathbf{z}_t^i|\mathbf{x}_t^i)$ in (2) is the measurement model, which is linear-Gaussian $P(\mathbf{z}_t^i|\mathbf{x}_t^i) \sim \mathcal{N}(\boldsymbol{C}\mathbf{x}_t^i, \boldsymbol{R})$ and defined by the output matrix $\boldsymbol{C}$, and the observation noise covariance $\boldsymbol{R}$. Finally, the term $P(m_t^i|m_{t-1}^{1:N}, \mathbf{x}_{t-1}^{1:N})$ defines a predictive probability distribution over the available maneuvers of the ith vehicle given the estimated states and intentions of all vehicles at the previous time step. This term is discussed in further detail in the following subsection.

C. Belief updates

In previous work, we saw an approximated method to recursively track over time the state and maneuver intentions of each vehicle in the scene [12]. We apply the same filtering method in this work to maintain the belief (note that for the ego-vehicle there is no need to track the intentions). We provide a summary of the method here and refer the reader to the original publication for the complete details. The goal is to track, for each vehicle, the distribution $P(\mathbf{x}_t^i, m_t^i|\mathbf{z}_{1:t})$, which is decomposed into two terms, and the first one approximated by a Gaussian mixture:

$$\begin{aligned} P(\mathbf{x}_t^i, m_t^i|\mathbf{z}_{1:t}) &= P(\mathbf{x}_t^i|m_t^i, \mathbf{z}_{1:t}) P(m_t^i|\mathbf{z}_{1:t}) \\ &\approx [\sum_{c_t} \underbrace{P(\mathbf{x}_t^i|c_t, m_t^i, \mathbf{z}_{1:t})}_{\text{Gaussian component}} \underbrace{P(c_t|m_t^i, \mathbf{z}_{1:t})}_{\text{Weight}}] \underbrace{P(m_t^i|\mathbf{z}_{1:t})}_{\text{Marginal } m_t^i} \end{aligned} \quad (5)$$

The recursion for the first term after receiving the new observation $\mathbf{z}_{t+1}$ is given by:

$$\begin{aligned} P(\mathbf{x}_{t+1}^i|m_{t+1}^i, \mathbf{z}_{1:t+1}) &= \sum_{m_t^i, c_t} P(\mathbf{x}_{t+1}^i, m_t^i, c_t|m_{t+1}^i, \mathbf{z}_{1:t+1}) \\ &= \sum_{m_t^i, c_t} \underbrace{P(\mathbf{x}_{t+1}^i|m_t^i, c_t, m_{t+1}^i, \mathbf{z}_{1:t+1})}_{\text{Gaussian component}} \underbrace{P(m_t^i, c_t|m_{t+1}^i, \mathbf{z}_{1:t+1})}_{\text{Weight: } w(m_t^i, c_t, m_{t+1}^i)} \end{aligned}$$

The new Gaussian components are obtained by propagating with a Kalman filter the components at the previous time step using all available dynamics (indexed by m_{t+1}^i). Let $|M|$ be the number of available maneuvers and $|C|$ the number of Gaussian components. The proposed procedure increases the number of Gaussian components from $|M||C|$ to $|M|^2|C|$, so a collapse is done as the final step of the recursion. To update the weights, we consider:

$$\begin{aligned} w(m_t^i, c_t, m_{t+1}^i) &\propto P(m_t^i, c_t, m_{t+1}^i, \mathbf{z}_{1:t+1}) \\ &= P(\mathbf{z}_{t+1}|m_t^i, m_{t+1}^i, c_t, \mathbf{z}_{1:t}) P(m_{t+1}^i|m_t^i, c_t, \mathbf{z}_{1:t}) \times \\ &\qquad P(c_t|m_t^i, \mathbf{z}_{1:t}) P(m_t^i|\mathbf{z}_{1:t}) \end{aligned} \quad (6)$$

where $P(\mathbf{z}_{t+1}|m_t^i, m_{t+1}^i, c_t, \mathbf{z}_{1:t})$ is the likelihood of the observation $\mathbf{z}_{t+1}$ under the corresponding Gaussian projected onto observation space. Intuitively, if the prediction using the dynamics of maneuver m_{t+1}^i corresponds to the observed movement of the target, the likelihood will be high and, by extension, also the weight. The terms $P(c_t|m_t^i, \mathbf{z}_{1:t})$ and $P(m_t^i|\mathbf{z}_{1:t})$ are available from the previous step of the recursion.

The term $P(m_{t+1}^i|m_t^i, c_t, \mathbf{z}_{1:t})$ represents the probability that the ith vehicle will execute maneuver m_{t+1}^i from the corresponding Gaussian component. We calculate this probability using model-based prediction, where a maneuver is exponentially more likely is it leads to a trajectory that accrues lower cost than all others according to the model:

$$\begin{aligned} &P(m_{t+1}^i = m|\mathbf{x}_t^{1:N}, m_t^{1:N}) \propto \\ &\mathbb{E}\left[\exp\left(-\underbrace{\sum_{k=0}^{T_m-1} \mathbf{w}^T \mathbf{f}^i([\mathbf{x}_{t+k,m}^i, \mathbf{x}_{t+k,\hat{m}_t^{-i}}^{-i}])}_{\substack{\text{Cost accrued over } T_m \text{ time steps by vehicle } i. \\ \text{Vehicle } i \text{ executes maneuver } m. \\ \text{The obstacles execute sampled maneuvers } \hat{m}_t^{-i}.}}\right)\right] \end{aligned} \quad (7)$$

where the expectation is with respect to the posterior distribution over state and maneuver intentions of all traffic participants at the previous time step and is solved using Monte Carlo sampling. In (7), we have overloaded the notation for the physical state to explicitly indicate the maneuver being used to propagate it between time steps, and the notation $\hat{m}^{-i}$ indicates the sampled maneuvers for all agents other than i.

The recursion for the maneuver marginal $P(m_t^i|\mathbf{z}_{1:t})$ in (5) is given by:

$$P(m_{t+1}^i|\mathbf{z}_{1:t+1}) \propto \sum_{m_t^i, c_t} w(m_t^i, c_t, m_{t+1}^i) \quad (8)$$

In (6) and (8), there is a fusion between dynamic evidence (the likelihood of the observation under each of the available dynamics) and scene understanding (the maneuver forecasting using model-based prediction). This has been shown to increase the robustness of the lane change intention estimations in highways [12]. Hence, each POMDP planning cycle starts from a belief distribution that provides an accurate estimation of the maneuver intentions of surrounding traffic.

D. Observation model

The DBN model can also be used in a generative manner to sample realistic observations of the surrounding traffic participants after a maneuver has been executed for the ego-vehicle during the tree search. In other words, for each traffic participant we want to sample from $P(\mathbf{x}_{t+1}^i, m_{t+1}^i|\mathbf{z}_{1:t}) = P(\mathbf{x}_{t+1}^i|m_{t+1}^i, \mathbf{z}_{1:t}) P(m_{t+1}^i|\mathbf{z}_{1:t}^i)$. The steps to obtain these two components from the initial distribution $P(\mathbf{x}_t^i, m_t^i|\mathbf{z}_{1:t})$ closely follow those presented in the previous subsection. The main and critical difference lies in the absence of the observation likelihood term in (6) and (8). Due to the lack of this term, the predictive maneuver marginal $P(m_{t+1}^i|\mathbf{z}_{1:t})$ is dominated by the risk-aversive, model-based prediction term. In consequence, the sampled observations, and by extension the beliefs represented in the search tree, will only capture a safe normative behavior for all surrounding drivers (as in [7]). Even when a given driver had been estimated to be performing a dangerous manever from the true observations of the world, this information is not propagated across the search tree from the root belief node. Under these conditions, the planner will not be able to perform anticipative behavior.

To leverage the dynamics-aware information contained at the root belief node, we modify (7) by including a new term that promotes maneuver continuity across time steps, effectively biasing the construction of the tree:

$$P(m_{t+1}^i = m|\mathbf{x}_t^i, m_t^i) \propto$$
$$\mathbb{E}\left[\exp\left(-\Big[\underbrace{\begin{array}{c}\text{Cost accrued by}\\ \text{maneuver m over}\\ T_m \text{ time steps}\end{array}}\Big] + T_m \underbrace{w_{mc}\,\delta_{(m,\hat{m}_t^i)}\left(\frac{\tau - t}{\tau - t_0}\right)}\right)\right]$$

Model-based, risk-aversive component

Term to promote maneuver continuity

where the term $w_{mc} \in \mathbb{R}^+$ is a reward term, $\tau \in \mathbb{N}$ indicates the lifetime in time steps of a linear decay term, and $\delta_{(i,j)}$ denotes the Kronecker delta function. By including this new term, the predicted probability of maneuvers that were estimated likely for obstacle i is increased. This effect is reduced as the depth in the tree increases using a linear decay term.

E. Rollouts

Below a given depth d_{lim} in the tree, the estimation of the accrued cost is done using a rollout policy. In our highway driving domain, we rely on model-based prediction to determine the approximate long-term development of any traffic scene. The procedure is as follows. First, states are sampled for all vehicles in the scene from the belief distribution at the leaf node. Then, we process all vehicles from the front to the back of the scene. For each vehicle, we calculate its model-based maneuver predictive distribution as in (7), and sample a maneuver that will be used to propagate the vehicle forward one time step. During the calculation of the maneuver predictive distribution, trailing vehicles are assumed to be performing a LK maneuver. The process is repeated until all rollout steps have been executed. Two examples of rollouts are shown at the bottom of Fig. 1 (each dot corresponds to the resulting physical state of each vehicle at each rollout step).

F. Planning algorithm

Our tactical decision-making approach is formulated in Algorithm 1. This algorithm resembles the POMCP and POMCP-DPW algorithms [9], [14], using Monte Carlo tree search to explore a search tree of histories. However, in these algorithms the belief is maintained using particle filtering, whereas we rely on a variational approach to maintain parametric beliefs for each history in the tree.

The tactical planning starts in the `PLAN` procedure. During the allocated planning time, simulations are continuously run from the initial history h_0 and corresponding belief b_0 to construct and explore the search tree. A simulation consists of a sequence of sampled actions and observations, ending on a leaf node from where a rollout is executed. The goal of the tree exploration is to maintain accurate estimates of the value $V(h_0 a)$ of the actions that can be taken from the current belief state. Note that although it is indexed by a history, $V(h_0 a)$ represents the value of the associated belief.

The history tree exploration is based on selecting actions using the UCT algorithm [15], and sampling observations

Algorithm 1: Human-like tactical decision-making in highways.

procedure `PLAN`$(h_0,\ b_0)$**:**
  while *not timeout* **do**
    `SIM`$(h_0, b_0, 0)$
  return $\arg\min_a V(ha)$

procedure `SIM`(h, b_h, d)**:**
  if $d \geq d_{lim}$ **then**
    return `ROLLOUT` (h, b_h, d)
  else
    if $h \notin$ *history-tree* **then**
      `add-to-tree` (h)
      for $\forall a \in$ `available-actions`(b_h) **do**
        `init` $(V(ha), N(ha), N(h))$
      return `ROLLOUT` (h, b_h, d)
    else
      $a \leftarrow \arg\min_{a'} V(ha') - c\sqrt{\frac{\log N(h)}{N(ha')}}$
      if $|$`children`$(ha)| \leq$ `floor`$(k_0 N(ha)^{\alpha_0})$ **then**
        $b_{hao}, o \leftarrow$ `dbn-model`(b_h, a)
        `children`$(ha) \leftarrow$ `children`$(ha) \cup \{o\}$
        acc $\leftarrow$ `COST`(b_{hao}, a) + γ `SIM`$(hao, b_{hao}, d+1)$
      else
        $o' \leftarrow$ `sample-observation(children`$(ha))$
        acc $\leftarrow$ `COST`$(b_{hao'}, a)$ $+\gamma$ `SIM`$(hao', b_{hao'}, d+1)$
      $N(h) \leftarrow N(h) + 1$
      $N(ha) \leftarrow N(ha) + 1$
      $V(ha) \leftarrow V(ha) + \frac{\text{acc} - V(ha)}{N(ha)}$
      return *acc*

using the method presented in subsection III-D. Each sampled observation is then used to obtain the belief b_{hao} of the new belief node. Since the observations are sampled from a continuous predictive distribution, we apply progressive widening to artificially limit the number of observations and enable the exploration of the lower layers of the tree [14].

Once the planning time runs out, the estimated value function is used to select the action to execute. After execution, an observation is received from the environment and used to obtain the new belief node, from where the tree search restarts.

IV. EXPERIMENTAL EVALUATION

In this section, we present the simulation platform and the two experimental tasks used to validate our approach. For each of the tasks, we show and discuss the results obtained for different traffic situations.

A. Simulation platform

We evaluate our approach on a customized version of an open-source driving simulator [16]. This simulator builds upon two existing open-source simulation packages: 1) Simulation of Urban MObility (SUMO) [17], an open-source microscopic road traffic simulator; and 2) Gazebo, an advanced 3D simulation environment. Figure 2 shows a diagram that illustrates how the simulator's components interact with the proposed decision-making system.

B. First task: selecting lane changes

In this task, the POMDP planner chooses the lane change commands of the ego-vehicle. The longitudinal acceleration is automatically set using the IDM car-following model. The goal of this task is to evaluate the proposed approach for its

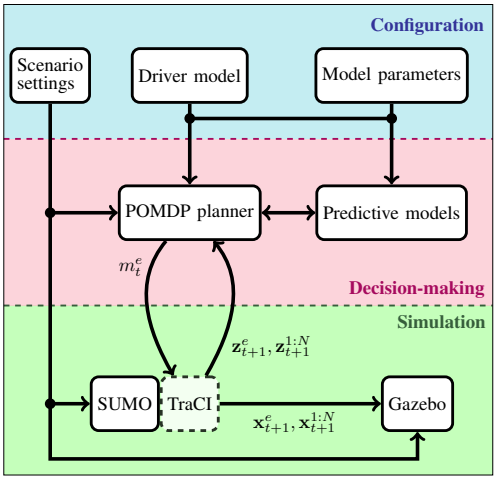


Fig. 2: Architecture of our highway simulation platform.

ability to make decisions when the motion of surrounding traffic is uncertain, as well as its ability to consider the long-term consequences of any action. The performance of the POMDP planner is compared to SUMO's reactive model for lane changing, which consists of a series of rules to determine the gain of doing a lane change maneuver.

Main parameters selected: Number of tree simulations per decision: 100, time step: 2s, max. depth tree: 3, max. rollout steps: 7, k_0: 2, α_0: 0.3, γ: 0.9, w_{mc}: 0, C: 1.

Results obtained: The top row of Fig. 3a shows the initial conditions of a test scenario, where the ego-vehicle (label 'prius') starts in the same lane as a much slower vehicle. As the scene develops and the ego-vehicle draws closer to the obstacle, both the POMDP and SUMO's reactive planner perform a lane change to overtake it. Once the obstacle has been passed, both models perform another lane change to merge onto the right lane. The only difference here lies on how long SUMO takes to begin the lane change to overtake, which is a consequence of its counter-based method to avoid oscillations.

In the top row of Fig. 3b, we show a more challenging test scenario. Here, the ego-vehicle again starts behind a much slower obstacle. However, in this case, some obstacles are approaching quickly from behind on the left lane. To avoid getting stuck behind the obstacle in front, the POMDP performs an early lane change at around $t = 2s$. Once the overtake is complete, the ego-vehicle merges back to the right lane, roughly 20m in front of the obstacle. In contrast, SUMO's reactive model does not realize the need for a lane change until it is too late, and the left lane is blocked by traffic. This forces the ego-vehicle to decelerate significantly down to $70^{\text{km}}/_{\text{h}}$. Once the left lane is clear, the overtake is initiated at around $t = 17\text{s}$.

This scene highlights the importance of considering the long-term consequences of any maneuver in the highway through adequate prediction models. With the parameters selected, the consequences up to 14s into the future are considered with our model. To gain insight on how our POMDP model decided to change lanes so early, we show in Fig. 1 two exemplary tree simulations. In the simulation highlighted in red, three LK maneuvers are sampled, followed by a rollout. We can see in the illustrated beliefs and in the rollout, how the ego-vehicle is predicted to significantly reduce its velocity down to $80^{\text{km}}/_{\text{h}}$. In contrast, the simulation in blue shows the case where a LC maneuver is sampled, followed by two LK maneuvers and a rollout. In this case, after the LC maneuver, the belief $b_h a_2 o_2$ shows how the fast trailing obstacle on the left lane is predicted to decelerate as a reaction the ego-vehicle's lane change. The following beliefs and rollout show how, following an initial LC maneuver, the ego-vehicle will be able to continue at its desired velocity and overtake the slow obstacle.

C. Second task: longitudinal control

In this task, the ego-vehicle navigates in the left lane of a two-lane highway and the goal of the POMDP planner is to choose its acceleration commands. To drive safely and avoid falling into dangerous situations, the planner is required to anticipate the lane change intentions of the vehicles circulating on the right lane, even when these intentions do not fit with a risk-averse driving style. This task tests the ability of the proposed approach to consider the short-term consequences of the selected actions by exploiting the dynamics-based behavior estimations.

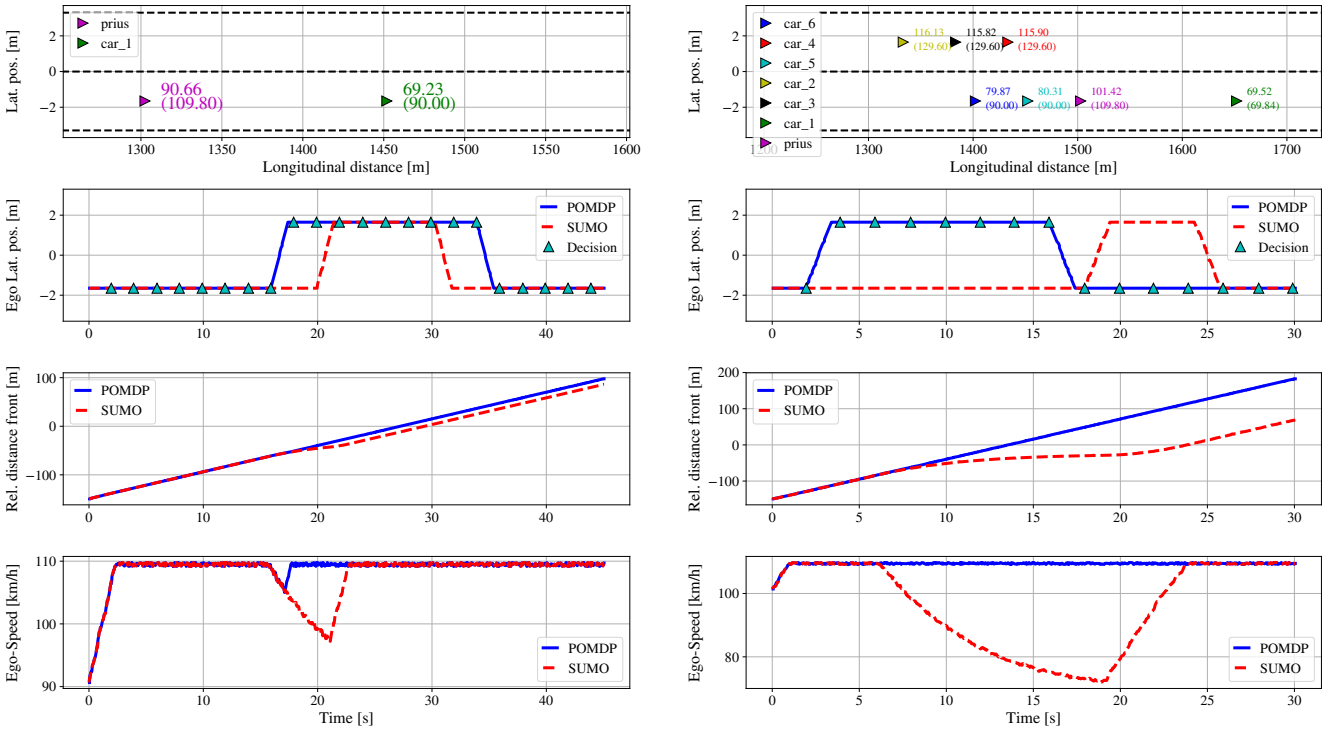
Main parameters selected: Number of tree simulations per decision: 400, time step: 1s, max. depth tree: 4, max. rollout steps: 2, k_0: 2, α_0: 0.2, γ: 0.9, w_{mc}: 25, C: 1.

Results obtained: Figures 4a and 4b show in the top row the identical starting conditions for the two scenarios tested. In the first scenario (Fig. 4a), the obstacle is never estimated to be performing a lane change maneuver during the scene (second row, right axis), so the POMDP simply accelerates the ego-vehicle to its desired velocity and overtakes the obstacle. Note that SUMO does not allow acceleration inputs, so the acceleration commands are executed as instant velocity changes.

In the second scene (Fig. 4b), the obstacle is detected to be performing a lane change with probability 0.6 between $t = 5s$ and $t = 25s$. As a response, the POMDP planner with $w_{mc} = 25$, decelerates the ego-vehicle to increase the longitudinal gap with the obstacle and enable the lane change. The definition of safe gap for the POMDP is determined by the parameters $\mathbf{w}$ of the cost model. Once the obstacle is no longer estimated to be doing a lane change, the planner accelerates again the ego-vehicle towards its desired velocity. We show as well the results obtained when $w_{mc} = 5$. In this case, situations in which the obstacle executes a lane change are under-represented in the search tree, leading to a behavior of the ego-vehicle where the risk of a potential collision is ignored. This shows the importance of properly tuning this parameter.

V. CONCLUSIONS

In this paper, we have presented a POMDP-based decision-making system for automated driving in highways. The proposed method searches for the optimal ego-maneuver given a belief over the physical states and maneuver intentions



(a) Planning results obtained for a straight-forward highway overtake scenario. Both the reactive SUMO model and the POMDP planner recognize the advantage in doing a lane change to avoid deviating too much from the desired speed. In the top figure, the numbers next to the markers represent the current and desired (between brackets) speeds in $^{km}/_{h}$. The time step between decisions is 2s.

(b) Planning results obtained for a more complex highway scenario. SUMO's reactive model leads the ego-vehicle to get stuck behind a slow driver and to deviate significantly from its desired velocity. In contrast, the proposed POMDP planner considers the long-term consequences of any decision and switches lanes at $t = 2s$, passing the slow vehicle at around $t = 13s$.

Fig. 3: First experimental task: the POMDP planner decides when to perform lane changes; the longitudinal acceleration is delegated to a car-following model.

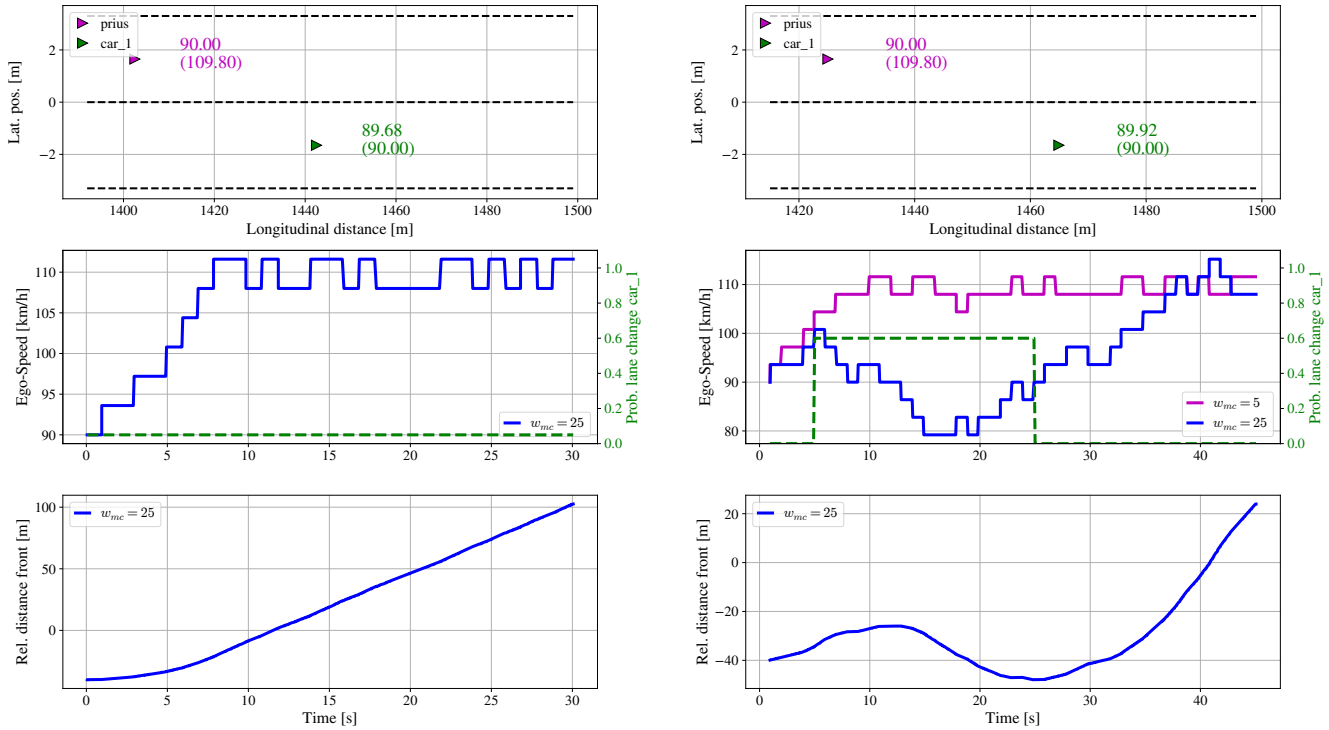
of surrounding traffic. The approach works on an online fashion, building during planning time a tree of possible scene developments from the current belief state. To build the tree, the dynamics of the world are modeled using an interaction-aware probabilistic model that takes into account both the observed dynamics of the targets and their expected behavior given the traffic situation. The model is used in a generative manner to sample likely observations as a reaction to the ego-vehicle's actions, and also to maintain the belief over time. Finally, in order to account for the long-term consequences of any decision, the approach relies on a egoistic model-based prediction approach that produces sensible long-term scene predictions.

We evaluated our approach on a simulator for two different navigational tasks. In the first task, the POMDP was in charge of the lane changing actions of the ego-vehicle while the longitudinal control was delegated to a car-following model. The results showed the ability of the proposed approach to make foresighted lane changing decisions under the uncertain dynamics of surrounding traffic (the model to predict the obstacles' motion was different from the simulator's actual model). In the second experimental task, the POMDP controlled the longitudinal motion of the ego-vehicle. The experimental results showed how the proposed method can exploit the uncertain maneuver intention estimations of the interaction-aware DBN model to produce human-like, anticipative driving behavior.

In the current implementation, the tree is rebuilt at each decision step, which precludes the proposed approach from achieving real-time performance (the runtime was roughly $500ms$ per simulation on the complex scene from Fig. 3b). Once the planning time runs out, the POMDP planner executes the best action according to the estimated action values at the root of the tree, and it subsequently receives a new observation of the world. Since the observation comes from a continuous observation space, it is unlikely that it will match any of the observations sampled with the generative model during the construction of the tree; it is therefore not clear which branch of the tree should be kept. In future work, we plan to evaluate different discretization techniques for the observation space that would allow us to match the *true* observation to those in the first layer of the tree. This, however, will have an impact on the tracking quality of the state and maneuver intentions, which will have to be measured and bounded.

References

[1] Waymo LLC, "2018 disengagement report," DMV, Tech. Rep., 2018.



(a) In this scenario, the ego-vehicle and the only present obstacle start off at the same speed. However, the desired speed of the ego-vehicle (top figure, in brackets) is significantly larger. Since the obstacle in lane 1 is never estimated to be executing a lane change maneuver throughout the duration of the scene (middle figure, right axis), the POMDP accelerates the ego-vehicle to its desired velocity, passing the slower obstacle on the left.

(b) In this scenario, the obstacle in lane 1 is estimated to be executing a lane change maneuver with probability 0.6 from $t = 5s$ to $t = 25s$. As a response, the POMDP planner (with $w_{mc} = 25$) slows down the ego-vehicle to increase the longitudinal gap with the obstacle and allow the lane change. However, if $w_{mc} = 5$ the danger due to the estimated lane change maneuver is ignored, and the ego-vehicle is accelerated to its desired speed.

Fig. 4: Second experimental task: the lane of the ego-vehicle is fixed; the POMDP planner controls the longitudinal acceleration of the vehicle.

[2] S. Ulbrich and M. Maurer, "Probabilistic online POMDP decision making for lane changes in fully automated driving," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, 2013, pp. 2063–2067.

[3] S. Brechtel, T. Gindele, and R. Dillmann, "Probabilistic decision-making under uncertainty for autonomous driving using continuous POMDPs," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2014, pp. 392–399.

[4] W. Liu, S. W. Kim, S. Pendleton, and M. H. Ang, "Situation-aware decision making for autonomous driving on urban road using online POMDP," in *2015 IEEE Intelligent Vehicles Symposium (IV)*, 2015, pp. 1126–1133.

[5] T. Bandyopadhyay, K. S. Won, E. Frazzoli, *et al.*, "Intention-aware motion planning," in *Algorithmic Foundations of Robotics X*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 475–491.

[6] C. Hubmann, J. Schulz, M. Becker, *et al.*, "Automated driving in uncertain environments: Planning with interaction and uncertain maneuver prediction," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 1, pp. 5–17, March 2018.

[7] Z. N. Sunberg, C. J. Ho, and M. J. Kochenderfer, "The value of inferring the internal state of traffic participants for autonomous freeway driving," in *2017 American Control Conference (ACC)*, 2017, pp. 3004–3010.

[8] M. Bouton, A. Cosgun, and M. J. Kochenderfer, "Belief state planning for autonomously navigating urban intersections," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, June 2017, pp. 825–830.

[9] D. Silver and J. Veness, "Monte-Carlo Planning in Large POMDPs," in *Advances in Neural Information Processing Systems 23, Vancouver, British Columbia, Canada.*, 2010, pp. 2164–2172.

[10] A. Kesting, M. Treiber, and D. Helbing, "Enhanced intelligent driver model to access the impact of driving strategies on traffic capacity," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 368, no. 1928, pp. 4585–4605, 2010.

[11] ——, "General lane-changing model mobil for car-following models," *Transportation Research Record*, vol. 1999, no. 1, pp. 86–94, 2007.

[12] D. Sierra González, V. Romero-Cano, J. Steeve Dibangoye, and C. Laugier, "Interaction-Aware Driver Maneuver Inference in Highways Using Realistic Driver Models," in *Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC 2017)*, Yokohama, Japan, Oct. 2017.

[13] D. Sierra González, Ö. Erkent, V. Romero-Cano, *et al.*, "Modeling Driver Behavior From Demonstrations in Dynamic Environments Using Spatiotemporal Lattices," in *ICRA 2018 - Proceedings of the 2018 IEEE International Conference on Robotics and Automation*, Brisbane, Australia, May 2018, pp. 3384–3390.

[14] Z. N. Sunberg and M. J. Kochenderfer, "Online algorithms for POMDPs with continuous state, action, and observation spaces," in *International Conference on Automated Planning and Scheduling (ICAPS)*, Delft, 2018.

[15] L. Kocsis and C. Szepesvári, "Bandit based monte-carlo planning," in *Machine Learning: ECML 2006*, J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 282–293.

[16] M. Garzón and A. Spalanzani, "An hybrid simulation tool for autonomous cars in very high traffic scenarios," in *ICARCV 2018 - 15th International Conference on Control, Automation, Robotics and Vision*, Singapore, Singapore, Nov. 2018, pp. 1–6.

[17] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of SUMO - Simulation of Urban MObility," *International Journal On Advances in Systems and Measurements*, vol. 5, no. 3&4, pp. 128–138, December 2012.