

Highlights

RUMOR: Reinforcement learning for Understanding a Model of the Real World for Navigation in Dynamic Environments

Diego Martinez-Baselga, Luis Riazuelo, Luis Montano

- A remarkable performance in robot navigation in dynamic real-world scenarios.
- Deep reinforcement learning (DRL) proves to be effective at decision making.
- The DOVS model can extract the scenario dynamism and its future information.
- The DOVS top of DRL improves its efficiency and generalization.
- RUMOR already considers differential drive kinodynamics in the DRL action space.

RUMOR: Reinforcement learning for Understanding a Model of the Real World for Navigation in Dynamic Environments

Diego Martinez-Baselga^{a,*}, Luis Riazuelo^a, Luis Montano^a

^aRobotics, Perception and Real Time Group, Aragon Institute of Engineering Research (I3A), University of Zaragoza, Zaragoza, 50018, Spain

Abstract

Autonomous navigation in dynamic environments is a complex but essential task for autonomous robots, with recent deep reinforcement learning approaches showing promising results. However, the complexity of the real world makes it infeasible to train agents in every possible scenario configuration. Moreover, existing methods typically overlook factors such as robot kinodynamic constraints, or assume perfect knowledge of the environment. In this work, we present RUMOR, a novel planner for differential-drive robots that uses deep reinforcement learning to navigate in highly dynamic environments. Unlike other end-to-end DRL planners, it uses a descriptive robocentric velocity space model to extract the dynamic environment information, enhancing training effectiveness and scenario interpretation. Additionally, we propose an action space that inherently considers robot kinodynamics and train it in a simulator that reproduces the real world problematic aspects, reducing the gap between the reality and simulation. We extensively compare RUMOR with other state-of-the-art approaches, demonstrating a better performance, and provide a detailed analysis of the results. Finally, we validate RUMOR's performance in real-world settings by deploying it on a ground robot. Our experiments, conducted in crowded scenarios and unseen environments, confirm the algorithm's robustness and transferability.

Keywords: Mobile robots, Motion planning, Collision avoidance, Reinforcement learning, Differential-drive robots, Dynamic environments

1. Introduction

Motion planning and navigation in dynamic scenarios is essential for autonomous robots, for applications as delivery or assistance. Nevertheless, traditional planners fail in environments where the map is mutable or obstacles are dynamic, leading to suboptimal trajectories or collisions. Those planners typically consider only the current obstacles' position measured by the sensors, without considering the future trajectories they may have.

New approaches that try to solve this issue include promising end-to-end Deep Reinforcement Learning (DRL) based methods. The robot learns a policy that selects the best action for each of the situations, directly represented with the sensed information. They present results that outperform model-based approaches in terms of success (reaching the goal without collisions) and time to reach the goal. However, the complexity of the real world and the huge variety of different possible situations the robot may encounter, with different number of obstacles, different shapes and different behaviors, poses a huge training challenge for these end-to-end approaches. Moreover, they typically lack in sim2real transfer capabilities, not considering robot kinodynamic constraints, or partial observability issues.

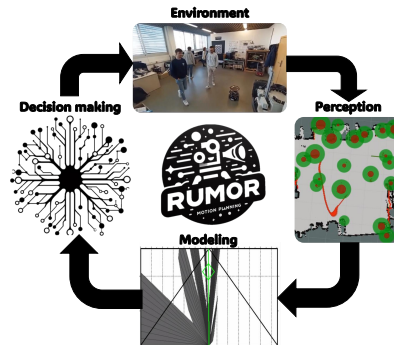


Figure 1: Pipeline of the approach presented. It takes the information sensed from the environment to construct a model of the dynamism of the scenario. Then DRL is used to compute differential-drive velocity commands.

In this paper, we propose **RUMOR** (Reinforcement learning for Understanding a **MO**del of the **Rea**l world), a motion planner that combines model-based and DRL benefits. We use the deep abstraction of the environment provided by the Dynamic Object Velocity Space (DOVS) [1] to understand the surrounding scenario. The DOVS model represents the dynamism and the future of the environment in the velocity space of the robot. The planner applies this information as an input of the DRL algorithm, learning to interpret the future dynamic scenario knowledge, taking advantage over other approaches that use raw obstacle information and are not able to generalize in scenarios that are different from those previously experienced. The perception is decoupled from the learning algorithm and

*Corresponding author at: Robotics, Perception and Real Time Group, Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain.

Email addresses: diegomartinez@unizar.es (Diego Martinez-Baselga), riazuelo@unizar.es (Luis Riazuelo), montano@unizar.es (Luis Montano)

used to construct the DOVS, making it able to work with any sensor as a LiDAR or a camera. Figure 1 provides a representation of RUMOR pipeline. In addition, unlike other methods that simply choose a velocity ranging from the maximum and minimum robot velocities, we propose an action space that inherently considers differential-drive kinodynamic restrictions, improving the applicability of RUMOR in real robots. The algorithm is trained and tested in a realistic simulator, where all information is extracted from the sensor measurements, even the own robot localization. Training using the robot constraints and a simulator that includes the problems of the real world reduces the sim2real transfer complexity.

In Section 2, we analyze the background and relate our work with the state of the art and in Section 3 we state the preliminaries needed for understanding the method. Section 4 and Section 5 present the approach taken. We provide comparisons of RUMOR with other method of the state-of-the-art and experiments with a real-ground robot in Section 6. Finally, the conclusions of the work are stated in Section 7.

Our contributions are summarized in:

- A motion planner that combines a very complete and descriptive information of the environment on top of DRL to merge the benefits of model-based and DRL methods.
- A way to inherently consider differential-drive robot kinodynamics in a DRL planner and deal with some real-world problems as the variety of scenario configurations or partial observability.
- Experiments that show the benefits of our contribution, comparisons with other methods and real-world demonstrations.

The code will be made open source upon acceptance.

2. Background

2.1. Motion planning in dynamic environments

Motion planners for static and continuous environments do not properly deal with dynamic obstacles, as they lead to collisions and suboptimal trajectories, as seen in some works as the benchmark proposed in [2]. Some traditional approaches for dynamic environments include forces to model the obstacle avoidance, such as artificial potential fields [3] or pedestrian forces to emulate the people motion [4]. They were later improved with Time Elastic Bands (TEB) [5] or multiple extensions to Social Force [6] [7].

A big group of works are velocity-space based. The *Velocity Obstacle (VO)*, introduced in [8], refers to the set of velocities of the robot that could lead to a collision with an obstacle that has a certain velocity in the near-future, which should not be chosen. Based on the VO concept, the *Dynamic Object Velocity Space (DOVS)* is defined in [1] as the velocity-time space for differential-drive robots, which includes the unsafe robot velocities for all the obstacles and the time to collision for computing safe robot velocities in a time horizon. In that work, a planner

based on strategies is also defined, the S-DOVS. Another work [9] uses the DOVS to discard velocities that lead to collisions from the action space of a tables-based reinforcement learning algorithm. However, unlike our work, it does not use the information of the DOVS to interpret the environment and just filters the unsafe actions available. In addition, the results presented in their work are very similar to S-DOVS, which does not require training time. Other extensions of VO use reciprocal collision assumptions, such as Reciprocal Velocity Obstacles (RVO) [10] or Optimal Reciprocal Collision Avoidance (ORCA) [11], which leads to failures in scenarios where there are non-collaborative obstacles.

Application of optimization-based approaches in dynamic environments, such as those that are based in a Model Predictive Controller (MPC), is gaining ongoing attention in recent years. The authors of [12] present a Model Predictive Contouring Controller that considers the future obstacle trajectories to safely reach the goal. It was recently combined with a global planning approach [13] to dynamically change between possible global plans. Other approaches combine crowd motion prediction with MPC navigation [14] or introduce topological invariance to improve social navigation [15]. As our approach, these works account for kinematic and dynamic constraints of the robot, but they have the problem of being computationally costly, and their planning is reliable as long as the predictions are accurate, so they are not suitable for very dense or chaotic scenarios.

2.2. Deep reinforcement learning planners

Deep Reinforcement learning [16] is a method used to learn to estimate the optimal policy that optimizes the cumulative reward obtained in an episode with a deep neural network. Many extensions have been proposed to this original algorithm. Some works have proven the best performances of the state-of-the-art in various problems, including Rainbow [17] for discrete action spaces, distributed reinforcement learning [18] or actor-critic methods [19].

Some works offer analysis of the importance of Deep Reinforcement Learning (DRL) in robot motion planning and the limitations of traditional planners in dynamic environments. Defining strategies for every situation that may be found in the real world or optimizing trajectories in very dynamic environments is intractable, so DRL may be used to efficiently solve the decision-making problem, which is complex and has many degrees of freedom. There are works [20] [2] [21] that implement and compare DRL algorithms with model-based ones, proving their performance.

The first agent-based DRL approach proposed for motion planning [22] uses a fully connected network to select velocities regarding the position and velocity of the closest surrounding obstacles. LSTM [23] layers were included in [24] [25] to account for multiple obstacles. SARL [26] models with attention the interactions among obstacles, achieving state-of-the-art performance. Recent approaches compare their results to it, and they usually focus on improving the structure of the network. For example, [27] uses a graph neural network or [28] an attention-based interaction graph. A recent approach [29]

improves the performance of others with intrinsic rewards. The use of DRL algorithms in robot navigation has many important challenges, including the need of learning from a limited subset of possible scenarios, environmental and robot constraints or partial observability [30] [31]. The previous approaches do not face these problems. In our approach, instead of directly process the observations, we use the DOVS model, that represents the dynamism of the environment in the velocity space, as an input to the DRL algorithm, doing the decision-making process agnostic of the specific scenario.

Other works consider social settings, such as [32] including social stress indices, [33] a social attention mechanism or [34] social rewards to shape the robot behavior. Another approach [35] uses a social risk map as part of the input of the network. The applicability of these social approaches is limited when other obstacles are not humans, like other robots, or humans do not behave as expected.

An example of an approach that considers the kinodynamic restrictions is [36], which combines DRL with DWA [37], but only achieving a success rate of 0.54 in low dynamic scenarios. Our proposed approach solves the problem of selecting feasible velocities by considering kinematic and dynamic equations in the action space formulation, limiting the downgrade in performance.

3. Preliminaries

3.1. Problem formulation

We consider the problem of robotic navigation in dynamic environments. A robot, whose position at time t is $\mathbf{p}_t = (x_t, y_t, \theta_t) \in \mathbb{R}^3$ and its velocity command $\mathbf{u}_t = (\omega_t, v_t) \in \mathbb{R}^2$ (angular and linear velocity), has to reach a goal $\mathbf{g} = (x_g, y_g) \in \mathbb{R}^2$ in the minimum possible time, while avoiding static and dynamic obstacles in the environment. The robot and the obstacles share the workspace $\mathcal{W} = \mathbb{R}^2$. The robot is represented with a disk with radius r_{robot} , and the dynamic obstacles are assumed to be circular disks with radius $r_{obs} \in \mathbb{R}^M$, where M is the number of dynamic obstacles. Let $\mathcal{R}_t(\mathbf{u})$ be the robot state at time t after applying the command velocity $\mathbf{u}$; and $\mathcal{O}_{v,t} = \bigcup_{i=1}^M \mathcal{O}_{i,t}$ be the set of points occupied by the whole set of static and dynamic obstacles in the environment at time t , and $\mathcal{O}_v$ at any time. The robot maximum linear and angular velocities are v_{max} and ω_{max} , and its minimum linear and angular velocities are 0 m/s and $-\omega_{max}$, respectively. Equation 1 represents the unicycle model,

$$\begin{pmatrix} \dot{x}_t \\ \dot{y}_t \\ \dot{\theta}_t \end{pmatrix} = \begin{pmatrix} \cos \theta_t & 0 \\ \sin \theta_t & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} v_t \\ \omega_t \end{pmatrix} \quad (1)$$

and the the low-level forward model that relates the wheel velocities to the actions $\mathbf{u}_t$ for a differential-drive robot, is the following:

$$\begin{aligned} v_t &= \frac{r_w(\omega_{r,t} + \omega_{l,t})}{2} \\ \omega_t &= \frac{r_w(\omega_{r,t} - \omega_{l,t})}{L_w} \end{aligned} \quad (2)$$

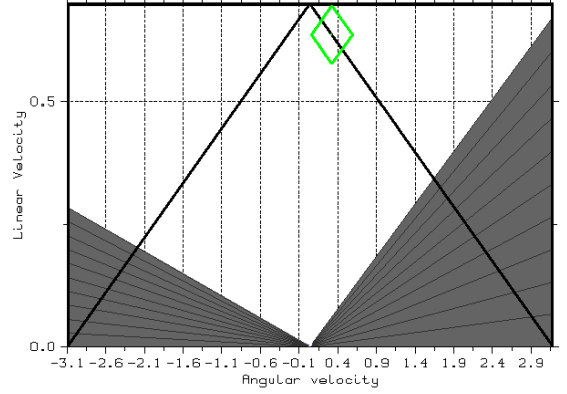


Figure 2: Graphical representation of the DOVS model for a differential-drive robot velocity space, with the kinodynamic constraints of Equation 3 and Equation 4 represented with two black lines (velocities greater than them are not reachable) and a green rhombus (velocities outside the rhombus are not reachable in one control period), respectively. The current robot velocity is the center of the rhombus. In this example, the robot velocity limits are $v_{max} = 0.7$ m/s, $\omega_{max} = \pi$ rad/s and $a_{max} = 0.3$ m/s². The dark area (DOV) includes unsafe velocities and white area (FV) corresponds to safe velocities.

where r_w is the wheel radius, $\omega_{r,t}$ and $\omega_{l,t}$ the angular velocity of the right and left wheels, respectively, and L_w the distance between wheels. This poses a linear relationship between the maximum v_t a robot may take in a particular moment regarding its ω_t :

$$\begin{aligned} l_1 : v_t &\leq \frac{v_{max}}{\omega_{max}} \omega_t + v_{max}, & \omega_t &\leq 0 \\ l_2 : v_t &\leq -\frac{v_{max}}{\omega_{max}} \omega_t + v_{max}, & \omega_t &> 0 \end{aligned} \quad (3)$$

being v_{max} and ω_{max} the maximum linear and angular velocities, respectively, and $\frac{v_{max}}{\omega_{max}} = \frac{L_w}{2}$. The robot has a maximum linear acceleration of a_{max} . Thus, the change in the linear velocity of the robot is also limited by its previous linear velocity and, due to Equation 2, its velocity limits:

$$\begin{aligned} v_{t+1} &\leq \begin{cases} \frac{v_{max}}{\omega_{max}} \omega_t + v_t + a_{max} \Delta t, & \omega_t \leq 0 \\ -\frac{v_{max}}{\omega_{max}} \omega_t + v_t + a_{max} \Delta t, & \omega_t > 0 \end{cases} \\ v_{t+1} &\geq \begin{cases} -\frac{v_{max}}{\omega_{max}} \omega_t + v_t - a_{max} \Delta t, & \omega_t \leq 0 \\ \frac{v_{max}}{\omega_{max}} \omega_t + v_t - a_{max} \Delta t, & \omega_t > 0 \end{cases} \end{aligned} \quad (4)$$

The restrictions of Equation 3 and Equation 4 are plotted in a graphical representation of the robot velocity space as the two black lines and the green rhombus, respectively, in Figure 2.

3.2. Dynamic Object Velocity Space (DOVS)

The DOVS model, presented in [1], is used to extract the environment information in this work. It is a model that represents safe and unsafe velocity commands for a differential drive robot within a temporal horizon in a dynamic scenario. It extracts

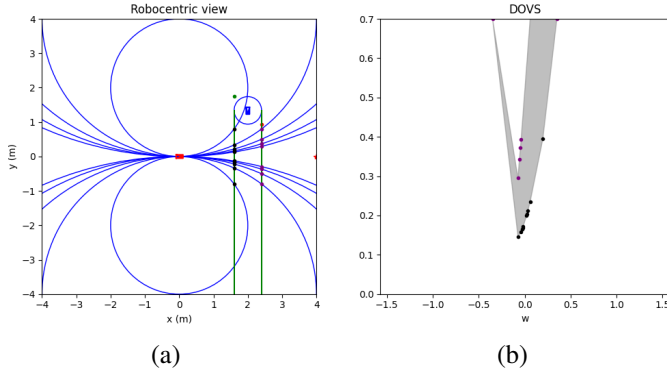


Figure 3: A robocentric view of the workspace (a) and a graphical representation of the DOVS (b) of a scenario with a robot navigating and facing the obstacle i that follows a linear trajectory. In the robocentric view, the robot center is represented in red, trajectories τ_j sampled in blue circular lines, the obstacle augmented radius with a blue circle and the collision band $\mathcal{B}_i$ with green lines. The intersection points between τ_j and $\mathcal{B}_i$ are $P_{1,i,j}$ (right) and $P_{2,i,j}$ (left). In the DOVS, the maximum velocities to pass after the obstacle are represented with black dots ($\omega_{2,i,j}, v_{2,i,j}$) and the minimum velocities to pass before it with purple dots ($\omega_{1,i,j}, v_{1,i,j}$). The resulting unsafe velocities are represented in red.

the information of the workspace directly in the robot velocity space $\mathcal{V}$ in every control period, using the time to collision between the robot and obstacles trajectories.

To build the model regarding the dynamic obstacle i , the robot size is reduced to a point and the obstacle is enlarged with the robot radius (the final collision areas are the same). The collision band $\mathcal{B}_i$ is computed as the area swept by the enlarged moving obstacle using its trajectory, which is assumed to be known or estimated from the sensor information. We consider circular trajectories as constant linear and angular velocities leads to them. We denote $\tau_j \equiv \tau_j(\omega_j, v_j)$ the circular trajectory τ_j produced by ω_j and v_j . Let $\mathcal{T}$ be the collection of different circular trajectories that collectively cover $\mathcal{W}$. For each circular trajectory $\tau_j \in \mathcal{T}$, the intersection points between the two lines limiting $\mathcal{B}_i$ and τ_j are $P_{1,i,j}(x_{1,i,j}, y_{1,i,j})$ and $P_{2,i,j}(x_{2,i,j}, y_{2,i,j})$. They are graphically represented in Figure 3(a) in a robocentric view for a few example trajectories.

The time $t_{2,i,j}$ that obstacle i takes to reach $P_{2,i,j}$ and the time $t_{1,i,j}$ that it takes to stop being in contact to $P_{1,i,j}$ measure the times of collision between the robot and the obstacle. This means that τ_j will intersect $\mathcal{O}_{i,t}$ for $t \in [t_{1,i,j}, t_{2,i,j}]$ (see [1] for more details about this computations). Therefore, the velocities that let the robot avoid collisions with obstacle i within trajectory j are those which are limited by the minimum and maximum velocities that allow the robot to pass before or after the obstacle. They are computed as:

$$\omega_{k,i,j} = \frac{\theta_{k,j}}{t_{k,i,j}} = \frac{\text{atan2}(2x_{k,i,j}y_{k,i,j}, x_{k,i,j}^2 - y_{k,i,j}^2)}{t_{k,i,j}}, \quad k = 1, 2 \quad (5)$$

$$v_{k,i,j} = r_j \omega_{k,i,j}, \quad k = 1, 2$$

where r_j is the radius of τ_j and $\theta_{k,j}$ the angular displacement on τ_j to reach $P_{k,i,j}$. Therefore, the set of velocities $\{(w, v) \in \mathcal{V} \mid v_{1,i,j} \leq v \leq v_{2,i,j} \wedge \omega = \frac{v}{r_j}\}$ lead to collision at a

given time within the interval $t \in [t_{2,i,j}, t_{1,i,j}]$. The Dynamic Object Velocity (DOV) for obstacle i (red area in Figure 3(b)) and the set of trajectories $\mathcal{T}$ is defined as:

$$DOV(\mathcal{O}_{i,t}, \mathcal{T}) = \left\{ \mathbf{u}_j = (\omega_j, v_j) \in \mathcal{V} \mid \exists \tau_j \in \mathcal{T}, \right. \\ \left. \exists t \in [0, T^h], \mathcal{R}_t(\mathbf{u}_j) \cap \mathcal{O}_{i,t} \neq \emptyset \right\}, \quad (6)$$

where r_j is the radius of the circular trajectory τ_j and T^h a time horizon. The DOV may be combined for all obstacles in the environment:

$$DOV(\mathcal{O}_{\mathcal{V},t}, \mathcal{T}) = \bigcup DOV(\mathcal{O}_{i,t}, \mathcal{T}) \quad (7)$$

$DOV(\mathcal{O}_{\mathcal{V},t}, \mathcal{T})$ intuitively represents the unsafe velocities that sometime lead to collision with the object. The Free Velocity (FV) is the complementary set of the DOV within $\mathcal{V}$, and represent the set of safe velocities that do not lead to collision with any obstacle.

$$FV(\mathcal{O}_{\mathcal{V},t}, \mathcal{T}) = \left\{ \mathbf{u}_j = (\omega_j, v_j) \in \mathcal{V} \mid \mathbf{u}_j \notin DOV(\mathcal{O}_{\mathcal{V},t}, \mathcal{T}) \right\} \quad (8)$$

Finally, the DOVS is defined as the whole set of velocities that comprise the DOV and the FV:

$$DOVS(\mathcal{O}_{\mathcal{V},t}, \mathcal{T}) = \{DOV(\mathcal{O}_{\mathcal{V},t}, \mathcal{T}) \cup FV(\mathcal{O}_{\mathcal{V},t}, \mathcal{T})\} \quad (9)$$

The DOVS contains the DOV representing unsafe velocities leading to collision in a time horizon if they are kept and FV contains safe velocities. The model may be reactively computed in every control period, giving the robot updated information to navigate adapted to the precision of the information sensed. Velocities inside the DOV do not directly lead to a collision if the following commands lead to a free velocity. This is represented in the evolution of the DOVS in time.

In DOVS, the velocities reachable for a differential-drive robot are constrained by the Equation 3 and 4, as shown in Figure 2. Therefore, only the velocities accomplish the constraints of the previous equations, are reachable in the next control period from the current velocity. The whole model is used to learn and apply the robot velocities for a safe navigation, as explained in the following sections.

4. Reinforcement learning setup

The problem of robot autonomous navigation in dynamic scenarios may be considered as a sequential decision-making problem to find an efficient policy that leads the robot to the goal while reactively adapting to the changes in the environment to avoid collisions. It may be modeled as a Partially Observable Markov Decision Process (POMDP) with a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, R)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $\mathcal{O}$ the observation space, $T(\mathbf{s}_{t+1} \mid \mathbf{a}_t, \mathbf{s}_t)$ the transition model, $O(\mathbf{o}_{t+1} \mid \mathbf{s}_{t+1}, \mathbf{a}_t)$ the observation model and $R(\mathbf{s}_t, \mathbf{a}_t)$ the reward function. At every time step t , the robot takes an action $\mathbf{a}_t \in \mathcal{A}$ in an environment whose state is $\mathbf{s}_t \in \mathcal{S}$ by following a policy $\pi(\mathbf{s}_t)$. The agent uses information observed from

the environment, $\mathbf{o}_t \in \mathcal{O}$, and the action taken modifies the environment given the transition model, resulting the next state $\mathbf{s}_{t+1} \in \mathcal{S}$. The goal of the learning process is estimating the optimal policy π^* that maximizes the expected cumulative reward:

$$\pi^*(\mathbf{s}_t) = \arg \max_{\mathbf{a}_t \in \mathcal{A}} \left[R(\mathbf{s}_t, \mathbf{a}_t) + \gamma \sum_{\mathbf{s}_{t+1} \in \mathcal{S}} T(\mathbf{s}_{t+1} | \mathbf{a}_t, \mathbf{s}_t) \sum_{\mathbf{o}_{t+1} \in \mathcal{O}} O(\mathbf{o}_{t+1} | \mathbf{s}_{t+1}, \mathbf{a}_t) V(\mathbf{s}_{t+1}) \right] \quad (10)$$

where γ is a discount factor to balance the present value of future rewards and $V(\mathbf{s}_t)$ is the value function, that represents the expected cumulative reward starting from state $\mathbf{s}_t$ and following the policy $\pi(\mathbf{s}_t)$:

$$V(\mathbf{s}_t) = R(\mathbf{s}_t, \pi(\mathbf{s}_t)) + \gamma \sum_{\mathbf{s}_{t+1} \in \mathcal{S}} T(\mathbf{s}_{t+1} | \pi(\mathbf{s}_t), \mathbf{s}_t) \sum_{\mathbf{o}_{t+1} \in \mathcal{O}} O(\mathbf{o}_{t+1} | \mathbf{s}_{t+1}, \pi(\mathbf{s}_t)) V(\mathbf{s}_{t+1}) \quad (11)$$

4.1. State space

The state $\mathbf{s}_t \in \mathcal{S}$ is the unobservable true condition of the system that drives the dynamics of the environment. In our problem, we consider that there is a robot with radius r_{robot} position $\mathbf{p}_t$ and velocity $\mathbf{u}_t$ at time t that drives towards a goal $\mathbf{g}$. The rest of the state is the set of points occupied by the set of static and dynamic obstacles $\mathcal{O}_{\mathcal{V},t}$ and the radius $\mathbf{r}_{obs} \in \mathbb{R}^M$ and velocity $\mathcal{V}_t = \{(v_i, \omega_i) | i \in [0 \dots M]\}$ of the M obstacles that are dynamic:

$$\mathbf{s}_t = \{r_{robot}, \mathbf{p}_t, \mathbf{u}_t, \mathbf{g}, \mathcal{O}_{\mathcal{V},t}, \mathbf{r}_{obs}, \mathcal{V}_t\} \quad (12)$$

4.2. Action space

The action $\mathbf{a}_t \in \mathcal{A}$ should account for the kinematics and dynamics of the robot stated in Equation 3 and Equation 4. To do so, we use a continuous action set of two values $\mathbf{a}_t \in [0, 1]^2 = (a_{1,t}, a_{2,t})$ and relate it to the constraints. We first define the four corners of the dynamic window of Equation 4:

$$\begin{aligned} \mathbf{u}_t^{up} &= (\omega_t, v_t + a_{max}\Delta t) \\ \mathbf{u}_t^{down} &= (\omega_t, v_t - a_{max}\Delta t) \\ \mathbf{u}_t^{left} &= (\omega_t - \frac{\omega_{max}a_{max}\Delta t}{v_{max}}, v_t) \\ \mathbf{u}_t^{right} &= (\omega_t + \frac{\omega_{max}a_{max}\Delta t}{v_{max}}, v_t) \end{aligned} \quad (13)$$

Then, define two vectors to form a basis in the velocity space:

$$\begin{aligned} \mathbf{b}_{1,t} &= \mathbf{u}_t^{left} - \mathbf{u}_t^{down} \\ \mathbf{b}_{2,t} &= \mathbf{u}_t^{right} - \mathbf{u}_t^{down} \end{aligned} \quad (14)$$

We can compute a velocity that accounts for the acceleration constraints as:

$$\mathbf{u}_{t+1} = \mathbf{u}_t^{down} + a_{1,t}\mathbf{b}_{1,t} + a_{2,t}\mathbf{b}_{2,t} \quad (15)$$

To account for the dynamic restriction of Equation 3, $\mathbf{a}_t$ maximum value is bounded: $a_{1,t} \in [0, a_{1,t}^{max}]$ and $a_{2,t} \in [0, a_{2,t}^{max}]$. It will be bounded if the dynamic window intersects with l_1 or l_2 . In the velocity space, the lines that pass through $\mathbf{u}_t^{down} = (\omega_t^{down}, v_t^{down})$ and have the direction of $\mathbf{b}_{1,t}$ and $\mathbf{b}_{2,t}$ are l'_1 and l'_2 , respectively:

$$\begin{aligned} l'_1 : v &= -\frac{v_{max}\omega}{\omega_{max}} + \frac{v_{max}\omega_t^{down}}{\omega_{max}} + v_t^{down} \\ l'_2 : v &= \frac{v_{max}\omega}{\omega_{max}} - \frac{v_{max}\omega_t^{down}}{\omega_{max}} + v_t^{down} \end{aligned} \quad (16)$$

And the intersection points between l_1 and l'_1 , and l_2 and l'_2 are $P_1(\omega_{1,p}, v_{2,p})$ and $P_2(\omega_{2,p}, v_{2,p})$, with:

$$\begin{aligned} \omega_{1,p} &= \frac{1}{2} \left(\frac{\omega_{max}v_t^{down}}{v_{max}} + \omega_t^{down} - \omega_{max} \right) \\ v_{1,p} &= \frac{1}{2} \left(\frac{v_{max}\omega_t^{down}}{\omega_{max}} + v_t^{down} - v_{max} \right) + v_{max} \\ \omega_{2,p} &= \frac{1}{2} \left(-\frac{\omega_{max}v_t^{down}}{v_{max}} + \omega_t^{down} + \omega_{max} \right) \\ v_{2,p} &= \frac{1}{2} \left(-\frac{v_{max}\omega_t^{down}}{\omega_{max}} + v_t^{down} - v_{max} \right) + v_{max} \end{aligned} \quad (17)$$

The action $\mathbf{a}_t$ will be bounded only if the distance between $\mathbf{u}_t^{down}$ and P_1 and P_2 is smaller than $\mathbf{b}_{1,t}$ or $\mathbf{b}_{2,t}$, since the dynamic window will be in collision with l_1 or l_2 . Therefore:

$$\begin{aligned} a_{1,t}^{max} &= \min \left(1, \frac{\|P_1 - \mathbf{u}_t^{down}\|}{\|\mathbf{b}_{1,t}\|} \right) \\ a_{2,t}^{max} &= \min \left(1, \frac{\|P_2 - \mathbf{u}_t^{down}\|}{\|\mathbf{b}_{2,t}\|} \right) \end{aligned} \quad (18)$$

The elements used in the velocity computation are represented in Figure 4. It may be seen how the result of the linear combination of Equation 15, with the action bounded by Equation 18, produces a velocity inside the rhombus defined by Equation 4 and respects Equation 3.

4.3. Observation space, observation model and transition model

The observation $\mathbf{o}_t \in \mathcal{O}$ states what the robot perceives from the environment. The robot estimates from the sensor data the values of the state, but its goal and radius, which are already known:

$$\mathbf{o}_t = \{r_{robot}, \hat{\mathbf{p}}_t, \hat{\mathbf{u}}_t, \mathbf{g}, \hat{\mathcal{O}}_{\mathcal{V},t}, \hat{\mathbf{r}}_{obs}, \mathcal{V}_t\}, \quad (19)$$

where $\hat{\mathbf{p}}_t$ is the estimated robot position, $\hat{\mathbf{u}}_t$ the estimated robot velocity, $\hat{\mathcal{O}}_{\mathcal{V},t}$ the estimated position of obstacles, and $\hat{\mathbf{r}}_{obs}$ and $\mathcal{V}_t$ the estimated radius and velocity of the moving obstacles.

The robot is modeled to have a 2-D LiDAR laser sensor. It estimates $\hat{\mathbf{p}}_t$ and $\hat{\mathbf{u}}_t$ with the laser data and the odometry, using Adaptive Monte Carlo Localization (AMCL) [38]. An extended version of the work proposed by [39] is used to detect static and dynamic obstacles and estimate $\hat{\mathcal{O}}_{\mathcal{V},t}$, $\hat{\mathbf{r}}_{obs}$ and $\mathcal{V}_t$ from the 2-D

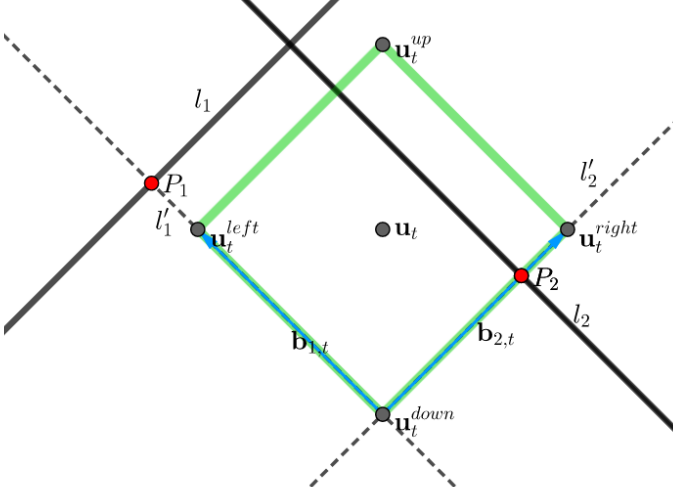


Figure 4: Graphical representation of the notation used in the action space configuration. $\mathbf{u}_t$, $\mathbf{u}_t^{up}$, $\mathbf{u}_t^{down}$, $\mathbf{u}_t^{left}$ and $\mathbf{u}_t^{right}$ are represented by gray points, l_1 and l_2 with black lines, l'_1 and l'_2 with dotted lines, $\mathbf{b}_{1,t}$ and $\mathbf{b}_{2,t}$ with blue arrows and P_1 and P_2 with a red point.

LiDAR measurements. It detects circular and linear obstacles from the scans, associating obstacles of any shape to circles or a set of segments, and tracks and estimates the velocity of the circular obstacles using an algorithm based on Kalman Filter. This kind of conditions makes the robot face simulated occlusions and estimation errors. In the simulations, we only consider dynamic circular obstacles to be able to compare our method with other approaches that assume that all obstacles are circular, but our approach could handle obstacles with any shape.

We assume that the robot follows the deterministic transition model of Equation 1, subject to the restrictions stated in Equation 3 and Equation 4. We model the dynamic obstacles using circular motion with constant random linear and angular velocity. The obstacles avoid each other using ORCA [11], resulting in motion that is different from the constant velocities previously defined. The robot is invisible for the obstacles, making the scenario more challenging. Otherwise, the robot could not learn to avoid non-cooperative obstacles.

4.4. Reward function

The goal of the agent is reaching the goal while avoiding collisions in the shortest time possible. To achieve this behavior, the reward functions proposed in [40] and [41] are used as inspiration. It is defined with a simple equation that discriminates between terminal and non-terminal states at time t :

$$R(\mathbf{s}_t, \mathbf{a}_t) = \begin{cases} r_{goal}, & d_{t,goal} < 0.15 \\ r_{collision}, & \text{collision detected} \\ -r_{dist}\Delta d_{t,goal} + r_{t,safedist}, & \text{otherwise} \end{cases}, \quad (20)$$

where $d_{t,goal} = \|\mathbf{g} - \mathbf{u}_t\|$ and $\Delta d_{t,goal} = d_{t,goal} - d_{t-1,goal}$. The robot receives a reward of $r_{goal} = 15$, when it reaches the goal within a certain threshold; and a negative reward $r_{collision} = -15$ when it collides. Reward shaping is used to accelerate training in non-terminal states by encouraging the agent to get closer to

the goal ($-r_{dist}\Delta d_{t,goal}$ with $r_{dist} = 2.5$), and by penalizing the agent if it is too close to an obstacle with:

$$r_{t,safedist} = \begin{cases} -0.1|0.2 - d_{t,obs}|, & d_{t,obs} < 0.2 \\ 0, & \text{otherwise} \end{cases}, \quad (21)$$

where $d_{t,obs}$ is the distance to the closest obstacle.

5. Network

We propose using a Soft Actor Critic (SAC) [19] algorithm to solve the problem, and use the DOVS model as part of the input of the network to extract the dynamism of the environment. The implementation of Stable Baselines 3 [42] is used for the SAC. SAC is an off-policy DRL algorithm that uses an actor to determine the policy and a critic to estimate the state-action values, and introduces entropy maximization to help exploration; it is known for its stability and sample efficiency.

Using raw velocity information of obstacles would require training with obstacles of every possible shape, velocity or radius the robot could face, which is impossible; so using the DOVS to model the scenario improves the adaptation and generalization of the network. The complete structure of the network is represented in Figure 5. We process the observation of the environment with two different streams. The first stream uses a discrete set of trajectories formed with velocities ranging from the maximum and minimum velocities of the robot and equally spaced:

$$\mathcal{T}_{RUMOR} = \left\{ \tau_j(\omega_j, v_j) \in \mathcal{T} \mid \omega_j \in (\omega_0, \dots, \omega_{N_\omega}), \right. \\ \left. v_j \in (v_0, \dots, v_{N_v}), \omega_0 = \omega_{min}, \omega_{N_\omega} = \omega_{max}, \right. \\ \left. v_0 = v_{min}, v_{N_v} = v_{max} \right\} \quad (22)$$

where N_ω is set to 40 and N_v to 20. The DOVS of this set is represented with a grid of $\{-1, 1\}$ discrete values indexed by ω_j and v_j representing whether $\tau(\omega_j, v_j) \in DOV(\mathcal{O}_{q,t}, \mathcal{T})$, the unsafe velocities, or $\tau(\omega_j, v_j) \in FV(\mathcal{O}_{q,t}, \mathcal{T})$, the safe velocities, respectively. It is processed with a convolutional network of three convolutional layers and a fully connected layer all of them followed by ReLU activation functions, ψ_{DOVS} , to preserve the relationships among velocities in the velocity space. Additionally, the following other observation variables $\mathbf{o}_{t,state}$ are separately processed with a fully connected layer with ReLU activation, ψ_{obs} :

$$\mathbf{o}_{t,state} = \{\hat{\mathbf{u}}_t, \hat{d}_{t,goal}, \hat{\theta}_{t,goal}, \hat{\mathcal{O}}_{t,dist}, \hat{\mathcal{O}}_{t,\theta}, \hat{\mathcal{O}}_{v,t}, \hat{\mathcal{O}}_{t,dir}\}, \quad (23)$$

where $\hat{d}_{t,goal}$ and $\hat{\theta}_{t,goal}$ are the distance and angle to the goal with respect to the estimated heading angle of the robot (θ_t), $\hat{\mathcal{O}}_{t,dist}$ the estimated distance of the robot to the closest obstacle, $\hat{\mathcal{O}}_{t,\theta}$ the estimated angle to it, $\hat{\mathcal{O}}_{v,t}$ its estimated velocity and $\hat{\mathcal{O}}_{t,dir}$ its estimated heading angle. Even though the information of these four last variables is already encoded in DOVS, they are included to give more information to the robot in case there is an imminent collision.

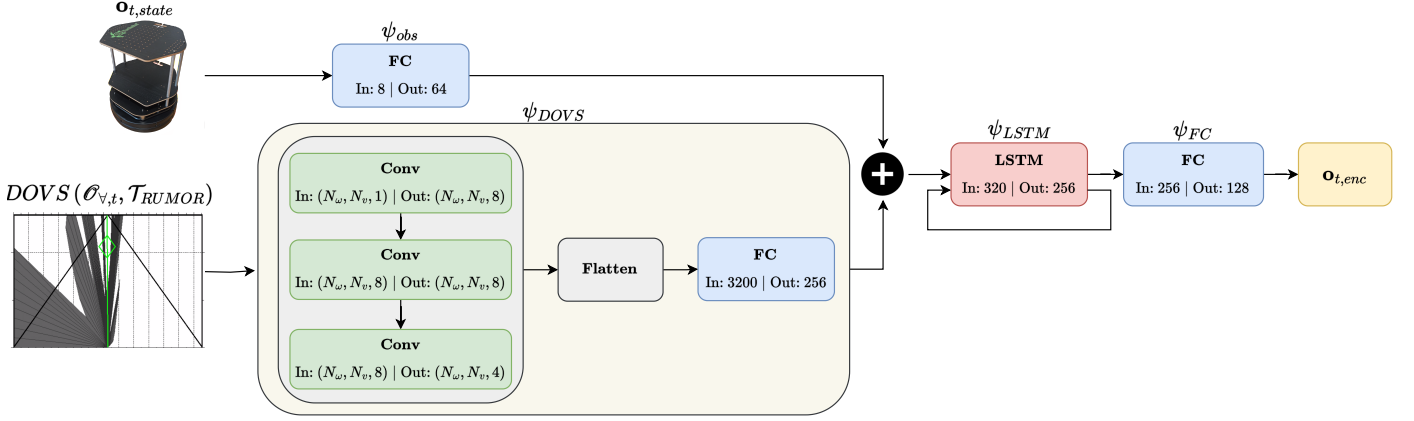


Figure 5: Structure of the encoding network. The DOVS model and the robot state are processed in two different streams, joint later by a memory layer that accounts for previous observations.

The output of both streams is concatenated and fed to an LSTM layer [23], ψ_{LSTM} , to save information of previous observation and keep temporal dependencies. The future of the environment is already considered by DOVS, but considering the past could help the network understand changes in the obstacles behavior and make it more robust to estimation errors or occlusions. Finally, a fully connected layer with ReLU activation function, ψ_{FC} , to produce the final encoded observation $\mathbf{o}_{t,enc}$:

$$\mathbf{o}_{t,enc} = \psi_{FC}(\psi_{LSTM}(\psi_{DOVS}(DOVS(\mathcal{O}_{v,t}, \mathcal{T}_{RUMOR})), \psi_{obs}(\mathbf{o}_{t,state}))), \quad (24)$$

An overview of the actor-critic architecture is shown in Figure 6. It may be seen that the encoder network is duplicated for the actor and the critic, so that they can separately optimize the weights for their objective.

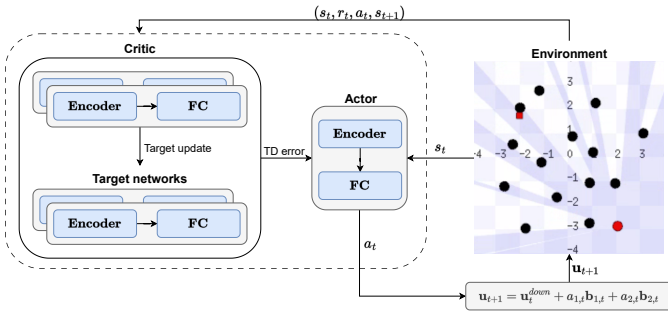


Figure 6: Structure of the SAC architecture used. The critic uses the whole transition for estimating the TD error, used to train the actor. The actor uses the environment observation to select an action, which is translated to a velocity. The environment included is the Stage simulator, used in training.

6. Experiments

6.1. Training details

The agent has been trained in a modified version of the Stage simulator [43], in a scenario with dynamic obstacles the robot

has to avoid to reach the goal. This simulator is conveniently integrated in ROS to use AMCL for localization and realistically simulates a LiDAR sensor. Random positions are sampled for the robot, its goal and the other obstacles. The obstacles tried to avoid each other using the ORCA algorithm, but they are not able to see the robot to make it perform the whole obstacle avoidance maneuvers, as stated in Section 4.3. The obstacles have a predefined random linear and angular velocity that try to follow, between 0.14 and 0.71 m/s. The robot faces scenarios with increasing number of dynamic obstacles, up to 14, and increasing starting distance to the goal during a first training stage of 1000 episodes. In this way, it learns to reach the goal progressively by using curriculum learning, from simple to more complex tasks. After that stage, it faces 9000 scenarios with a random number of obstacles up to 14. The minimum distance between the robot and the goal is set to 6 m. The robot velocity limits are set to $v_{max} = 0.7$ m/s and $\omega_{max} = \pi$ rad/s and the maximum acceleration to $a_{max} = 0.3$ m/s², similar to the ones of a Turtlebot 2 platform.

If the simulation takes more than 500 time steps, the episode is finished due to timeout. The episode time step is set to 0.2 s. The network converges in about 10 hours in a computer with a Ryzen 7 5800x processor, a NVIDIA GeForce RTX 3060 graphics card and 64 GB of RAM. The key parameters used are a learning rate of 0.0003 with Adam optimizer [44], discount factor of 0.99 and soft update coefficient of 0.005. Other parameters can be consulted in the code.

6.2. Simulation experiments

In this part, we conducted a series of quantitative experiments to evaluate the performance of our proposed model. We compared our method, RUMOR, with ORCA [11], S-DOVS [1], SARL [26] and RE3-RL [29], in order to test other model-based and DRL-based planners of the state of the art for navigation in dynamic environments. Nevertheless, the planners consider different constraints, as stated in Table 1. Only our planner and S-DOVS consider the constraints stated in Equation 3 and Equation 4. We propose an alternative version of our planner that does not respect those constraints for a more

Motion Planner	DRL	Differential drive	Kinodynamics
NR-RUMOR (ours)	✓	✓	
RUMOR (ours)	✓	✓	✓
S-DOVS		✓	✓
ORCA			
SARL	✓	✓	
RE3-RL	✓	✓	

Table 1: Differences between motion planners, where DRL refers to being a DRL-based planner, differential drive if they output differential drive commands and kinodynamics if they consider acceleration constraints or other dynamic and kinematic restrictions.

fair comparison, called NR-RUMOR (no-restrictions). It uses as the action space velocities ranging from the minimum and maximum limits, similar to SARL and RE3-RL.

We tested the planners in the same 100 random scenarios with low and high occupancy, with 6 and 12 dynamic obstacles, respectively. The 85% of the obstacles are set to be dynamic and the same set of scenarios is used for every version. We also tested two different perception settings: using the ground truth position of the dynamic obstacles given by the simulator (absolute perception) and using the Kalman Filter-based obstacle detector (obstacle tracker) previously referenced in Section 4.3, thus facing occlusions and estimation errors. The DRL policies (both ours and those of the state of the art) have been trained in the same settings as in the testing experiments, i.e., there are two versions of each of the DRL planners, one trained using absolute perception and the other using the obstacle tracker. The results obtained are plotted in Figure 7, where the success rates, navigation times path lengths and mean linear velocities recorded are plotted.

We mainly evaluate the planners using the two navigation metrics directly encoded in the reward function, which are the success rate (number of times the robot have reached the goal without collisions or timeouts) and the navigation time. There is a trade-off between both metrics. Trying to achieve low navigation time leads to risky collision avoidance maneuvers, which leads to lower success rates. In addition, the task is complex, and there are scenarios where the randomness in the positions and velocities of the obstacles makes it impossible to avoid collisions. For example, specially in dense scenarios, the robot may face trapping situations where several obstacles run towards it.

The success rates show, as previously stated by other works, that both model-based planners performance is worse than DRL-based ones. ORCA exhibits a lower navigation time than the other planners, probably due to the fact that it is a holonomic planner and does not respect differential-drive motion, but the success rate of both ORCA and S-DOVS are clearly the lowest in all settings. These planners model the environment in every time step, but they are not robust when errors or deep changes in the surroundings occur.

Regarding the perception settings, every planner shows higher success rates and lower navigation times with absolute perception, due to the absence of partial observability and esti-

mation errors. It is way more complex to predict the future of the environment when facing perception problems. Therefore, their relative performance within the same settings should be compared. It is interesting to see that there is not consistency in the performance of SARL and RE3-RL when changing the perception settings. When comparing them using absolute perception, SARL outperforms RE3-RL in both success rate and navigation time, while when using the obstacle tracker is the other way around. This is probably due to the difference in the complexity of their networks and the exploration process. Using intrinsic rewards to explore, as done in RE3-RL, may degrade the performance of the planner in the presence of erroneous observations. In contrast, both NR-RUMOR and RUMOR show a very consistent relative performance no matter the perception settings.

In the low occupancy experiments of Figure 7 (1-2), the performance of NR-RUMOR is similar to the best DRL planner in every perception settings. These scenarios are relatively easy and standard end-to-end approaches show good performance. NR-RUMOR slightly outperforms RE3-RL in both success and time with absolute perception, while with the obstacle tracker, SARL is the method which slightly outperforms NR-RUMOR. Nevertheless, it is in the high occupancy experiments of Figure 7 (3-4) where the superior performance of NR-RUMOR is stated. It clearly surpasses RE3-RL in terms of success and time with absolute perception, while, with the obstacle tracker, the navigation time regarding SARL is comparable, but the success rate is not. Two key components of the system explains this fact. First, the input of RUMOR network is the DOVS model of the environment. With 12 surrounding dynamic obstacles, there are unlimited number of possible scenario configurations, and the agent can not be trained to have a robust performance in all of them. On the contrary, DOVS represents the environment in similar terms no matter the scenario, leading to a deeper understanding of the velocity space. When the robot learns to interpret DOVS, it knows how to avoid collisions regardless the amount, positions or velocities of the obstacles. Second, the memory layer of RUMOR network allows it to keep track of previous observations. This is important to overcome partial observability challenges and those derived from sudden changes in the obstacles behavior. Overall, considering every combination in the perception and occupancy settings, NR-RUMOR consistently shows the best performance when compared with the other unconstrained or model-based methods.

RUMOR underperforms NR-RUMOR as expected, as kinematic and dynamic restrictions limits the capacity of maneuvering and sudden reactions. Both success and time metrics are worse in dense scenarios, whereas, even though the success rate is slightly higher, the navigation time is clearly higher in low occupancy scenarios. Nonetheless, their performance is comparable despite the restrictions. Furthermore, it displays a better performance, or at least comparable, than the rest of the planners, showing to be the second best overall. The difference in the behavior is seen in the velocity and path length graphs. While RUMOR tries to reach the maximum linear velocity as the rest of the planners, the acceleration limits prevents it from

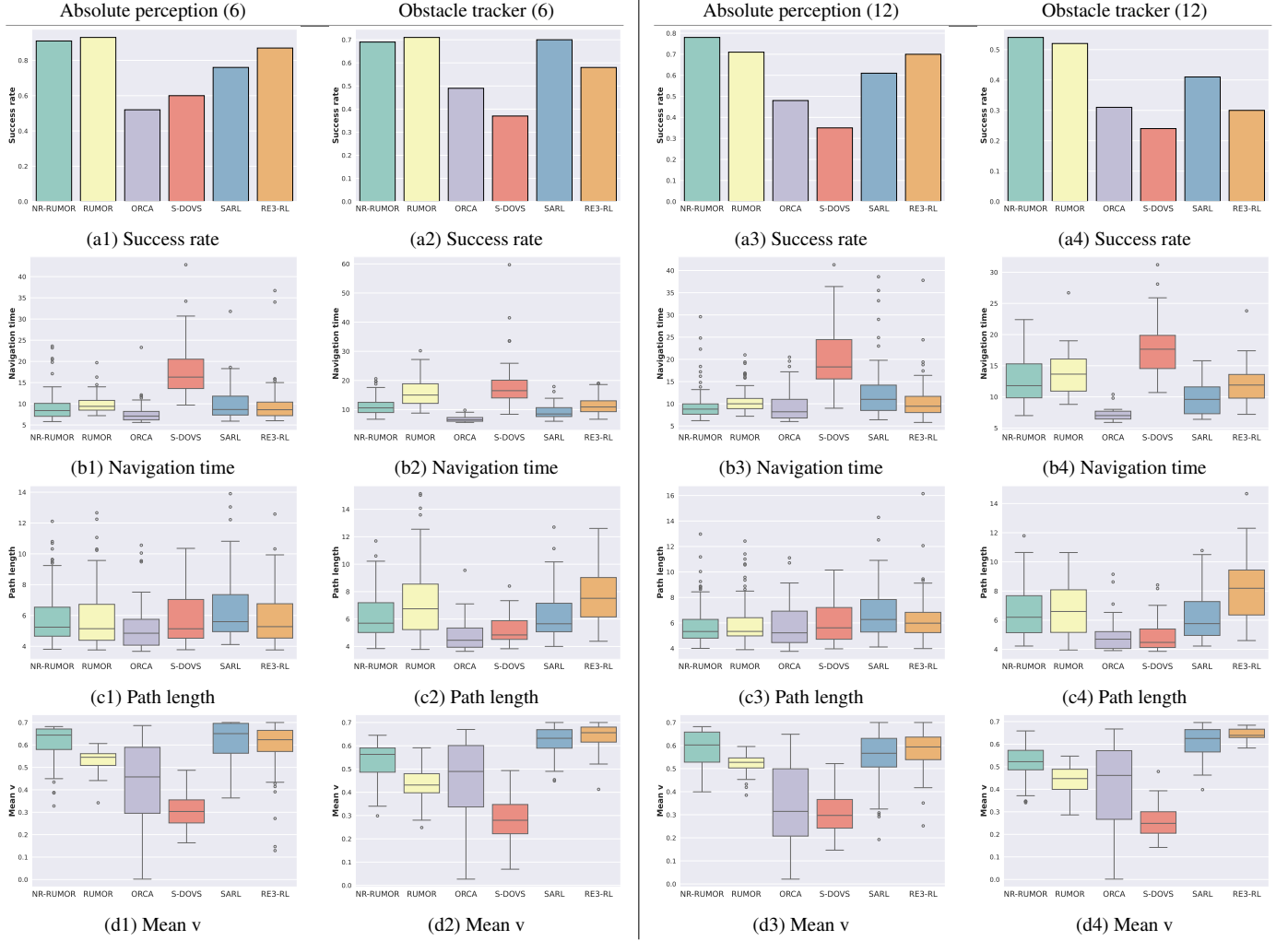


Figure 7: Navigation metrics for dynamic scenarios of 6 (left) and 12 (right) obstacles using absolute perception and the obstacle tracker.

achieving it as much time as the rest. In addition, it has to keep a lower velocity in dangerous zones, as it must stop gradually. The path length and navigation time plots show that RUMOR achieve high success rates with longer paths, meaning that it must respect the restrictions to navigate, it follows safer paths where no sudden velocity changes are needed.

An example of the behavior of the different DRL planners is shown in Figure 8, in absolute perception settings. All the planners reach the goal successfully, but both RUMOR versions do it in the shortest time, the shortest path lengths and with the most simple trajectories. RUMOR presents the smoothest trajectory, due to the acceleration restrictions, whereas all the other planners present a higher path irregularity. Its action space is the only one that does not allow the robot to reach maximum or minimum velocities at any time step, so it must be consistent with the navigation decisions it makes. In addition, as expected from the metrics results (Figure 7 (b1), (c1), (b3) and (c3)), RE3-RL performs better than SARL.

6.3. Real-world experiments

Real-world experiments were conducted to validate and evaluate RUMOR performance in real conditions. The whole system was integrated in a Turtlebot 2 platform with a NUC with Intel Core i5-6260U CPU and 8 GB of RAM. The sensor used is a 180° Hokuyo 2D-LIDAR. Due to the real-world design of the system, the same network weights and ROS nodes used in simulation were used in the ground robot, adapting the obstacle tracker node to detect people and other objects and using AMCL for localization. Two different types of real-world experiments were designed. Those experiments may be seen in the supplementary video.

A set of experiments were carried out to test navigation in very dense dynamic environments. The experimental setup was a scenario with dimensions of about 6x5 m. The robot was tasked to navigate towards dynamically assigned goals in the corners of the room while encountering dynamic obstacles. Pedestrians randomly wandering within the area were used as obstacles. To ensure the validity of the avoidance maneuvers, the individuals were instructed not to visually attend to the robot's movements. Figure 9 shows images of the robot

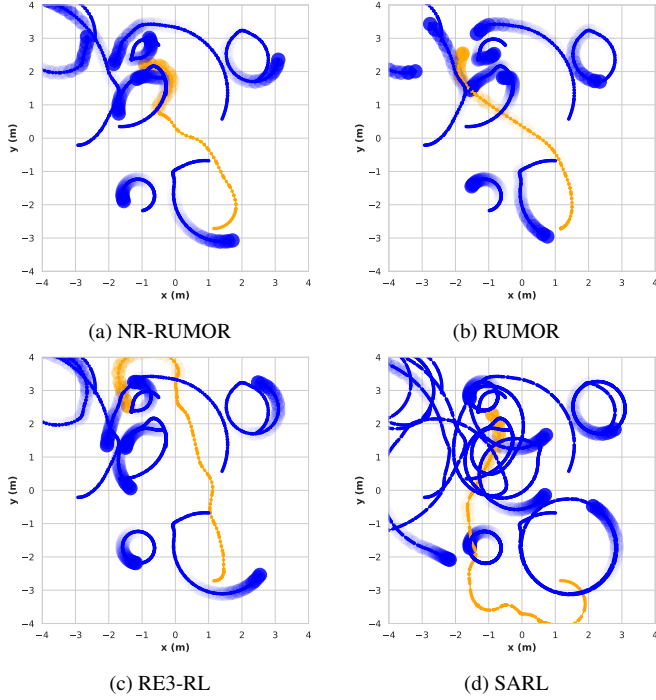


Figure 8: A robot (orange) navigating with different planners in a scenario with 9 dynamic obstacles (blue). The evolution in time is represented with increasing opacity, being completely solid at the end of the episode.

smoothly and safely navigating between two goals. In that specific situation, the robot must reach the goal avoiding collisions with five pedestrians and the limits of the room. The robot first waits for the first person encountered to pass (a). Then, it adapts its trajectory to pass between two people (b) and keeps the collision avoidance maneuver to avoid another one (c). Finally, it continues its way to the goal (d). RUMOR is able to use the DOVS representation of the dynamism of the environment to choose natural and smooth trajectories with low path irregularities to reach the goal, instead of abruptly reacting to obstacles.

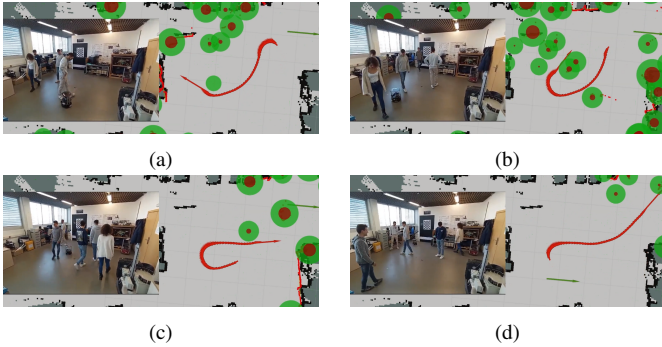


Figure 9: Photos superposed to the visualization of four consecutive moments of a real-world experiments. The robot position is represented with a red arrow, its previous trajectory with small red arrows, the LiDAR scans with red points, the obstacle positions and radius extracted with the obstacles tracker with red and green circles and the goal of the robot with a green arrow.

Finally, a setup different from a lab scenario was tried, in settings including a corridor with some rooms and intersections,

obstacles with different shapes and people with heterogeneous behaviors, as in real life, to prove that the system works in a completely unseen environment that has not been specifically used in simulation. This may be seen in Figure 10. A global planner as the one introduced in [45] was used to help the robot navigate through rooms and convex obstacles, while letting RUMOR compute the motion commands by itself. The results show that the robot is able to navigate very smoothly through the dynamic obstacles that have behaviors different from those seen in training, avoids static obstacles that had not been seen before, and works with the global planner to be able to reach goals that are very far from the initial position. The use of DOVS, instead of being a complete end-to-end planner, limits out-of-distribution observations, as apparently different scenarios could result in similar DOVS representations.

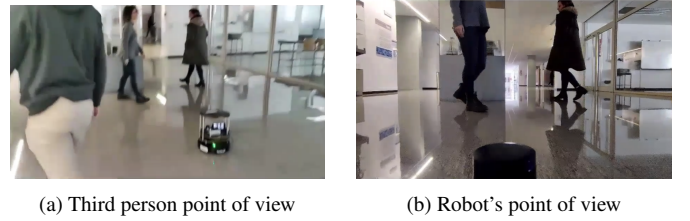


Figure 10: Images taken from the corridor experiment.

7. Conclusion

This work presents a novel motion planner for dynamic environments. It uses the DOVS model to extract the dynamism of the environment, abstracting from specific sensor data directly used by end-to-end approaches, and DRL to select motion commands that efficiently leads the robot to a goal while avoiding collisions with moving obstacles. Instead of directly use the raw obstacle information, which may lead to out-of-distribution observations in scenarios different from the ones seen in training, the system uses DOVS, which encodes it in similar terms regardless the scenario. In addition, we propose a training framework that puts the agent in real-world setting, an advanced DRL algorithm that uses memory to mitigate partial observability issues and a novel action space that intrinsically considers differential-drive kinodynamics, in order to mitigate sim2real transition. The proposed system is tested in random scenarios with absolute perception (ground truth from the simulator) and partial observability conditions, outperforming existing methods and showing the benefits of combining a model-based approach with a DRL controller. The algorithm is also tested in a ground robot in dense, dynamic and heterogeneous scenarios. Future work could include increasing the robustness of the method in scenarios where the obstacle motion is very irregular and unnatural, extending the model to collaborative collision avoidance with multi-robot navigation or adapting the model for 3-D navigation and UAVs.

CRediT authorship contribution statement

Diego Martinez-Baselga: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing-Original draft preparation, Writing-Reviewing and Editing, Visualization. **Luis Riazuelo:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing-Reviewing and Editing, Supervision. **Luis Montano:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing-Reviewing and Editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available upon acceptance.

Acknowledgement

This work was partially supported by the Spanish projects MCIN/AEI/10.13039/501100011033/FEDER-UE, and Aragon Government_FSE-T45_23R.

References

- [1] M.-T. Lorente, E. Owen, L. Montano, Model-based robocentric planning and navigation for dynamic environments, *The International Journal of Robotics Research* 37 (8) (2018) 867–889.
- [2] L. Kästner, T. Buiyan, T. A. Le, E. Treis, J. Cox, B. Meinardus, J. Kmiecik, R. Carstens, D. Pichel, B. Fatloun, et al., Arena-bench: A benchmarking suite for obstacle avoidance approaches in highly dynamic environments, *IEEE Robotics and Automation Letters* 7 (4) (2022) 9477–9484.
- [3] C. Qixin, H. Yanwen, Z. Jingliang, An evolutionary artificial potential field algorithm for dynamic path planning of mobile robot, in: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2006, pp. 3331–3336.
- [4] D. Helbing, P. Molnar, Social force model for pedestrian dynamics, *Physical review E* 51 (5) (1995) 4282.
- [5] C. Rösmann, F. Hoffmann, T. Bertram, Timed-elastic-bands for time-optimal point-to-point nonlinear model predictive control, in: 2015 european control conference (ECC), IEEE, 2015, pp. 3352–3357.
- [6] G. Ferrer, A. Garrell, 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, IEEE, 2013, pp. 1688–1694.
- [7] Y.-Q. Jiang, B.-K. Chen, B.-H. Wang, W.-F. Wong, B.-Y. Cao, Extended social force model with a dynamic navigation field for bidirectional pedestrian flow, *Frontiers of Physics* 12 (2017) 1–9.
- [8] P. Fiorini, Z. Shiller, Motion planning in dynamic environments using velocity obstacles, *The International Journal of Robotics Research* 17 (7) (1998) 760–772.
- [9] A. K. Mackay, L. Riazuelo, L. Montano, RL-DOVS: Reinforcement learning for autonomous robot navigation in dynamic environments, *Sensors* 22 (10) (2022) 3847.
- [10] J. Van den Berg, M. Lin, D. Manocha, Reciprocal velocity obstacles for real-time multi-agent navigation, in: 2008 IEEE International Conference on Robotics and Automation, IEEE, 2008, pp. 1928–1935.
- [11] J. Van Den Berg, S. J. Guy, M. Lin, D. Manocha, Reciprocal n-body collision avoidance, in: *Robotics Research: The 14th International Symposium ISRR*, Springer, 2011, pp. 3–19.
- [12] B. Brito, B. Floor, L. Ferranti, J. Alonso-Mora, Model predictive contouring control for collision avoidance in unstructured dynamic environments, *IEEE Robotics and Automation Letters* 4 (4) (2019) 4459–4466.
- [13] O. de Groot, L. Ferranti, D. Gavrilu, J. Alonso-Mora, Globally guided trajectory planning in dynamic environments, in: 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2023, pp. 10118–10124.
- [14] S. Poddar, C. Mavrogiannis, S. S. Srinivasa, From crowd motion prediction to robot navigation in crowds, in: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2023, pp. 6765–6772.
- [15] C. Mavrogiannis, K. Balasubramanian, S. Poddar, A. Gandra, S. S. Srinivasa, Winding through: Crowd navigation via topological invariance, *IEEE Robotics and Automation Letters* 8 (1) (2022) 121–128.
- [16] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, *Nature* 518 (7540) (2015) 529–533.
- [17] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, D. Silver, Rainbow: Combining improvements in deep reinforcement learning, *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (1) (2018).
- [18] D. Horgan, J. Quan, D. Budden, G. Barth-Maron, M. Hessel, H. van Hasselt, D. Silver, Distributed prioritized experience replay, in: *International Conference on Learning Representations*, 2018.
- [19] T. Haarnoja, A. Zhou, P. Abbeel, S. Levine, Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, in: *International conference on machine learning*, PMLR, 2018, pp. 1861–1870.
- [20] L. Kästner, T. Buiyan, L. Jiao, T. A. Le, X. Zhao, Z. Shen, J. Lambrecht, Arena-rosnav: Towards deployment of deep-reinforcement-learning-based obstacle avoidance into conventional autonomous navigation systems, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2021, pp. 6456–6463.
- [21] L. Kästner, R. Carstens, H. Zeng, J. Kmiecik, T. Buiyan, N. Khorasandhi, V. Shcherbyna, J. Lambrecht, Arena-rosnav 2.0: A development and benchmarking platform for robot navigation in highly dynamic environments, in: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2023, pp. 11257–11264.
- [22] Y. F. Chen, M. Liu, M. Everett, J. P. How, Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning, in: 2017 IEEE international conference on robotics and automation (ICRA), IEEE, 2017, pp. 285–292.
- [23] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural computation* 9 (8) (1997) 1735–1780.
- [24] M. Everett, Y. F. Chen, J. P. How, Motion planning among dynamic, decision-making agents with deep reinforcement learning, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 3052–3059.
- [25] M. Everett, Y. F. Chen, J. P. How, Collision avoidance in pedestrian-rich environments with deep reinforcement learning, *IEEE Access* 9 (2021) 10357–10377.
- [26] C. Chen, Y. Liu, S. Kreiss, A. Alahi, Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 6015–6022.
- [27] C. Chen, S. Hu, P. Nikdel, G. Mori, M. Savva, Relational graph learning for crowd navigation, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2020, pp. 10007–10013.
- [28] S. Liu, P. Chang, Z. Huang, N. Chakraborty, K. Hong, W. Liang, D. L. McPherson, J. Geng, K. Driggs-Campbell, Intention aware robot crowd navigation with attention-based interaction graph, in: 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2023, pp. 12015–12021.
- [29] D. Martinez-Baselga, L. Riazuelo, L. Montano, Improving robot navigation in crowded environments using intrinsic rewards, in: 2023 IEEE International Conference on Robotics and Automation (ICRA), 2023, pp. 9428–9434.
- [30] J. Ibarz, J. Tan, C. Finn, M. Kalakrishnan, P. Pastor, S. Levine, How

- to train your robot with deep reinforcement learning: lessons we have learned, *The International Journal of Robotics Research* 40 (4-5) (2021) 698–721.
- [31] G. Dulac-Arnold, N. Levine, D. J. Mankowitz, J. Li, C. Paduraru, S. Goyal, T. Hester, Challenges of real-world reinforcement learning: definitions, benchmarks and analysis, *Machine Learning* 110 (9) (2021) 2419–2468.
 - [32] Z. Hu, Y. Zhao, S. Zhang, L. Zhou, J. Liu, Crowd-comfort robot navigation among dynamic environment based on social-stressed deep reinforcement learning, *International Journal of Social Robotics* 14 (4) (2022) 913–929.
 - [33] Z. Zhou, P. Zhu, Z. Zeng, J. Xiao, H. Lu, Z. Zhou, Robot navigation in a crowd by integrating deep reinforcement learning and online planning, *Applied Intelligence* (2022) 1–17.
 - [34] B. Xue, M. Gao, C. Wang, Y. Cheng, F. Zhou, Crowd-aware socially compliant robot navigation via deep reinforcement learning, *International Journal of Social Robotics* (2023) 1–13.
 - [35] H. Yang, C. Yao, C. Liu, Q. Chen, Rmrl: Robot navigation in crowd environments with risk map-based deep reinforcement learning, *IEEE Robotics and Automation Letters* (2023).
 - [36] U. Patel, N. K. S. Kumar, A. J. Sathymoorthy, D. Manocha, DWA-RL: Dynamically feasible deep reinforcement learning policy for robot navigation among mobile obstacles, in: 2021 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2021, pp. 6057–6063.
 - [37] D. Fox, W. Burgard, S. Thrun, The dynamic window approach to collision avoidance, *IEEE Robotics & Automation Magazine* 4 (1) (1997) 23–33.
 - [38] D. Fox, W. Burgard, S. Thrun, Markov localization for mobile robots in dynamic environments, *Journal of artificial intelligence research* 11 (1999) 391–427.
 - [39] M. Przybyła, Detection and tracking of 2D geometric obstacles from LRF data, in: 2017 11th International Workshop on Robot Motion and Control (RoMoCo), IEEE, 2017, pp. 135–141.
 - [40] P. Long, T. Fan, X. Liao, W. Liu, H. Zhang, J. Pan, Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 6252–6259.
 - [41] A. J. Sathymoorthy, J. Liang, U. Patel, T. Guan, R. Chandra, D. Manocha, Densecavoid: Real-time navigation in dense crowds using anticipatory behaviors, in: 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2020, pp. 11345–11352.
 - [42] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, N. Dornmann, Stable-baselines3: Reliable reinforcement learning implementations, *Journal of Machine Learning Research* 22 (268) (2021) 1–8.
 - [43] B. Gerkey, R. T. Vaughan, A. Howard, et al., The player/stage project: Tools for multi-robot and distributed sensor systems, in: *Proceedings of the 11th international conference on advanced robotics*, Vol. 1, Citeseer, 2003, pp. 317–323.
 - [44] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).
 - [45] D. Martinez-Baselga, L. Riazuelo, L. Montano, Long-range navigation in complex and dynamic environments with full-stack s-dovs, *Applied Sciences* 13 (15) (2023) 8925.