F. Martínez, M. A. Lozano, F. Fernández. (2017). Emergent behaviors and scalability for multi-agent reinforcement learning-based pedestrian models. *Simulation Modelling Practice and Theory*, 74, pp. 117-133

DOI: https://doi.org/10.1016/j.simpat.2017.03.003

Emergent behaviors and scalability for multi-agent reinforcement learning-based pedestrian models

Francisco Martinez-Gil [a, *], Miguel Lozano [a], Fernando Fernández [b]

[a] Departament d'Informàtica, ETSE-UV, Universitat de València, Valencia, Spain
[b] Computer Science Department, Universidad Carlos III, Madrid, Spain

A B S T R A C T

This paper analyzes the emergent behaviors of pedestrian groups that learn through the multiagent reinforcement learning model developed in our group. Five scenarios studied in the pedestrian model literature, and with different levels of complexity, were simulated in order to analyze the robustness and the scalability of the model. Firstly, a reduced group of agents must learn by interaction with the environment in each scenario. In this phase, each agent learns its own kinematic controller, that will drive it at a simulation time. Sec- ondly, the number of simulated agents is increased, in each scenario where agents have previously learnt, to test the appearance of emergent macroscopic behaviors without addi- tional learning. This strategy allows us to evaluate the robustness and the consistency and quality of the learned behaviors. For this purpose several tools from pedestrian dynamics, such as fundamental diagrams and density maps, are used. The results reveal that the de- veloped model is capable of simulating human-like micro and macro pedestrian behaviors for the simulation scenarios studied, including those where the number of pedestrians has been scaled by one order of magnitude with respect to the situation learned.

Keywords:
Pedestrian simulation and modeling
Multi-Agent Reinforcement Learning (MARL)
Behavioural simulation
Emergent behaviours

1. Introduction

Interactive simulations with artificial groups or crowds offer a difficult control problem because the simulated people must exhibit very complex behaviors to be realistic. This complexity mainly depends on the simulated scenario and the size of the group being simulated. For instance, in many crowded scenarios, such as buildings, cities, etc., artificial pedestrians must reflect intelligent path planning in stochastic environments, as humans are constantly adjusting their speed to reflect congestion and other dynamic factors. Moreover, when the size of the simulated group increases in many structured scenarios, the problem of providing realistic path planning also increases, as a result some emergent behaviours are expected from a macroscopic perspective (lane formation, clogging effects, etc). A paradigmatic situation is the shortest *vs* quickest scenario (see Section 4.1). It offers an illustrative example of a simple pedestrian facility (2 rooms connected through two doors) which many kind of agents can easily solve. However, when the number of people involved increases a congestion appears very soon, and then a more complex problem has to be faced. In these scenarios, intelligent agents must show macroscopic self organized patterns, normally without considering strategic considerations or coordination techniques, that is, emergent collective behaviors. This type of behaviours emerge from the combination of local interactions between

* Corresponding author. .
E-mail addresses: Francisco.Martinez-Gil@uv.es (F. Martinez-Gil), Miguel.Lozano@uv.es (M. Lozano), ffernand@inf.uc3m.es (F. Fernández).

individuals or agent models and they have been studied for many decades ([1,2]). Emergent behaviours have been also incorporated to crowd based simulations, normally under social force models assumptions although different kinds of crowd models currently offer this feature (see Section 2). Nevertheless, one of the challenges in crowd simulations nowadays is to automatically generate macroscopic level behaviors and emergent phenomena from these local rules [3].

Crowd simulation typically requires complex mathematical models to drive the agents in their environments. Multi-agent reinforcement learning (RL) models propose an interesting approach for several reasons. RL agents are efficient because during simulations they are continually performing two main tasks per cycle. Thus, they classify the feature vector provided by the sensors (state recognition), and then, they find the best action to carry out according to the current state. The classification involved in the recognition of the current state has a linear computational cost with the number of generalized states used (see Section 3.2 for a formal state definition). The navigational decision making (find the maximum likelihood action in a given state) is also linear with the number of actions defined. Furthermore, stochastic models also offer interesting possibilities for controlling the variability of the simulated behaviours when cloning them to increase the size of the group, which is an important issue in crowd simulation. During the learning phase, the agents involved are considered as prototypes and once the learning process has been completed, they can simply be cloned or combined. Moreover, the Multi-agent paradigm allows to define independent learning processes for each autonomous agent generating variability in the behaviors learned. By contrast, the learned behaviors normally suffer from poor controllability during the learning process and they are difficult to edit.

The key contributions of this paper are:

- A scalability and performance evaluation analysis of the Multi-agent RL model in different scenarios studied in pedestrian modeling literature.
- An evaluation of the capability of the learned behaviors to create emergent collective behaviors while scaling up the number of simulated agents without additional learning (i.e. generalization).

The rest of the paper has been organized as follows. The next section summarizes the related work on pedestrian simulation and it also includes a specific review of machine learning-based models for pedestrian simulation. Some foundations of RL and a description of the Multi-agent RL framework used is explained in Section 3. Section 4 describes the scenarios used in the experiments and relate them to the literature of the field, then, the configuration of the learning processes is described. In Section 5, the results of these experiments are displayed and the limitations of the approach are discussed. Section 6 presents the conclusions and describes the future work.

2. Related work

Navigational behaviors from individual agents to virtual crowds have been studied in different research areas including social sciences, computer graphics, robotics, engineering (traffic), etc. In our case, pedestrian dynamics must be considered as a research area that inspires many of the navigational models presented in this section.

2.1. Pedestrian models

Efficient path-planning algorithms have been developed for Multi-agent navigation in virtual environments [4–6]. There is a considerable work on local dynamics models able to produce emergent crowd behaviors. Reynolds, in his seminal works [7,8] demonstrated that simple local rules can generate emergent flocking and other behaviors. Among these local methods, the social forces model [9] has been actively studied and many extensions have also been proposed [10–12]. In this context, issues such as sociological factors [13], psychological effects [14], situation-guided control [15] and cognitive and behavioral models [16,17], have also been integrated into the social force model. From a pedestrian dynamics perspective, emergent behaviours have been analyzed in the most popular microscopic models: social forces [1,18,19], cellular automata [20] and agent-based models [21]. Recently, in the work [22] the authors studied the stop-and-go waves that emerge from unidirectional pedestrian traffic. This work presents a numerical model of a following behavior inspired by the analogy with car traffic with an evaluation at a microscopic scale with real examples. The aim of that work is similar to ours, as both present an experimental study to evaluate different emergent behaviours.

In this paper, we use a bottom-up methodology to create crowds. First, a group of agents learn the navigational problem individually. Then, these learned behaviors are reproduced in a crowd to analyze their robustness. Therefore, crowds are simulated using a pure microscopic approach, where each individual is autonomous. In the work [23], the authors also propose a method to clone crowd motion data, to animate crowded scenes. Multi-scale approaches, on the contrary, combine micro and macroscopic characteristics in the same simulation. As a result, the system is capable of reproducing adequate values of the characteristics that define the scale of observation (local collisions, overtaking, etc. in the case of the microscopic scale and mean velocities, flows and densities in the case of the macroscopic scale). The works [24–26] describe formally ways of coupling microscopic and macroscopic scales. A popular approach to the multi-scale problem consists of developing a hybrid framework with different layers where each one assumes the control of a specific scale (a microscopic layer to assume the operational level of control and a macroscopic layer that assumes tactical and strategic levels). Works inspired by this approach are [27,28,29].

The evaluation of pedestrian models is a complex and important task. Steering based models have been evaluated in the works [30,31], where a benchmarking suite is proposed. In the work [32] the authors deal with the balance between completely procedural and data driven approaches for pedestrian steering algorithms. In the paper [33], the authors review important issues so as to achieve behavioural realism in virtual crowds simulations, such as, context-sensitive data-driven approaches, navigation meshes, multi-domain planning, semantics and others. An evaluation of wayfinding strategies in presence of congestion can be reviewed in the works [34,35].

2.2. Machine learning based approaches

Machine learning is a promising field for pedestrian based simulations. Learning can be applied to different areas of the simulation system (i.e. finding relevant features and/or parameters from real data or examples, generate a control or a decision making module, finding optimal paths inside a motion graph).

Many works use real data to find a model of pedestrian dynamics. In this sense, statistical machine learning techniques are used to find these models. For instance, the work by Pettre [36] uses maximum likelihood estimation techniques to calibrate their parametric model from real samples. The work by [37] learns collective pedestrian navigation from demonstrations. Their method proposes a feature-based maximum entropy learning approach to infer the distribution that matches the observed behavior of the agents in expectation.

RL [38] is a subfield of machine learning suitable to get control modules for different purposes of pedestrian simulation systems. Several works used RL in Motion-graphs-based animation [39–41] or for creating basic agent's behaviors [42]. RL agents are able to learn by using the rewards received from the actions carried out during the learning process. This learning procedure is completely different to data driven agents, since they are normally based on supervised learning techniques. In our case, there is no supervision and data is provided by the agent's interaction with the environment instead of being supplied externally after a filtering process, as it normally occurs in data-driven systems.

The Multi-agent reinforcement learning system designed for pedestrian simulations presented in this paper (called MARL-Ped) has been calibrated and validated in previous works. The work [43] describes the physics around the model (i.e. collision response, friction, etc) as well as the adjustments made to calibrate the system. The results of experiments carried out in simple scenarios (unidimensional lane, bidimensional plane) were validated using real data from Seyfried [44], Weydmann [45] and Mori [46] works. We use the fundamental diagram to validate our results, as it is considered an important evaluation tool in pedestrian modeling literature. In the work [47] we compare the dynamics learned by our RL agents with the one generated by the Helbing's social forces model under well known scenarios (corridor, maze and others). The validation results derived from these previous works show many similarities with both, the social force model and the real data used for comparison purposes. These minimum differences obtained in the fundamental diagrams comparison let us consider our model as a realistic pedestrian model able to drive lifelike artificial pedestrians in different kind of situations and scenarios.

3. Learning framework

MARL-Ped considers a simulated pedestrian as an independent agent that interacts autonomously with both, the environment and the rest of agents creating its own trajectory. Each agent follows a perception-actuation cycle synchronized with the rest of virtual agents. A special agent, named the environment, is in charge of the physical simulation. It uses a physics engine, the Open Dynamics Engine (ODE) [48], to compute the interactions of the virtual pedestrians inside the virtual world. Each virtual agent in the virtual world is controlled by an agent of the multi-agent system. The sequential decision-making process that controls a virtual agent consists of adapting the agent's velocity to the local situation as a real pedestrian would do in the same situation.

3.1. Background

Each agent learns to control its correspondent virtual pedestrian using RL techniques. A RL task can be formalized as a Markov Decision Process (MDP) [38]. In our context, a MDP is defined by a navigational state space S, a motor action space A, a probabilistic transition function P: $S \times A \times S \rightarrow [0, 1]$ which models the interactions with the environment and a reward function $\rho : S \times A \times S \rightarrow \mathbb{R}$. The state signal s_t describes the environment at discrete time t. Assume A is a discrete set. In a state, the decision process can select an action from the action space $a_t \in A$. The execution of the action in the environment changes the state to $s_{t+1} \in S$ following the probabilistic transition function $P(s_t, a_t, s_{t+1}) = Pr\{s_{t+1} = s' \mid s_t = s, a_t = a\}$. Each decision triggers an immediate scalar reward given by the reward function $r_{t+1} = \rho(s_t, a_t, s_{t+1})$ that represents the value of the decision made in the state s_t. The goal of the process is to maximize at each time-step t the *expected discounted return* defined as:

$$R_t = E\left\{\sum_{j=0}^{\infty} \gamma^j r_{t+j+1}\right\} \tag{1}$$

where the parameter $\gamma \in [0, 1]$ is the *discount factor* and the expectation E is taken with the probabilistic state transition P [49]. The discounted return takes into account not only the immediate reward got at time t but also the future rewards.

The discount factor controls the importance of the future rewards. The Action-value function (Q-function) $Q^{\pi} : S \times A \rightarrow \mathbb{R}$ is the expected return of a state-action pair given the policy π:

$$Q^{\pi}(s,a) = E\{R_t \mid s_t = s, a_t = a, \pi\} = E\left\{\sum_{j=0}^{\infty} \gamma^j r_{t+j+1} \mid s_t = s, a_t = a, \pi\right\} \tag{2}$$

The goal of the RL algorithm is to find an optimal Q^* such as $Q^*(s,\ a) \geq Q^{\pi}(s,\ a) \forall s \in S,\ a \in A,\ \forall \pi$. The optimal policy $\pi^*(s)$ is automatically derived from Q^* as it is defined in Eq. (3).

$$\pi^*(s) = \underset{a}{\operatorname{argmax}}\ Q^*(s,a) \tag{3}$$

The RL family of Temporal-Difference (TD) algorithms compute $Q^*(s,\ a)$ by interacting with the environment without knowing the transition function $P(s_t, a_t, s_{t+1})$ following a Monte Carlo approach. This is important as in microscopic pedestrian simulation, there is little knowledge about the effects of local interactions on the pedestrian dynamics. The Sarsa(λ) on-policy algorithm has been chosen for our experiments because it has provided good results in this problem domain in previous studies [47,52]. In Sarsa(λ), the value function $Q(s,\ a)$ is actualized with every new interaction tuple $(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1})$ following Eq. (4)

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha\left[r_{t+1} + \gamma\, Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)\right] \tag{4}$$

where α is the learning rate and γ the discount factor. The algorithm trades off exploration of new actions in a given state with the exploitation of the best known action for this state. Exploration is necessary to avoid local maxima and it is controlled by an exploration rate parameter ε that begins with an initial value ε_0 which decreases exponentially with the number of trials (ε-Greedy exploration).

3.2. Pedestrian dynamics as a MDP

The state space is defined using features that describe the dynamics of the agent's neighborhood. Therefore, agents learn to behave with a certain local conditions without taking into account the whole scenario. This is important when scaling up the number of agents because the agents have not a dependence with the global situation which varies depending on the grade of the scaling. Moreover, local features have proven being useful in pedestrian navigation in previous works and they are considered as relevant for the kinematic description of the pedestrian [50] or to describe and advance collision situations[2]. The perception of the state is an individual task where each agent has its own perception of the world at any moment. The selected features that describe the state of the agent are displayed in Fig. 1. All the measures are referenced with respect to the line that joins dynamically the virtual agent with its goal. It is important to indicate that the features are normalized. This fact is relevant when considering scaled environments.

All the features are real values, which means that the state space is a continuous set. Therefore, a generalization approach of this space is necessary [38]. Many generalization methods exist (clustering methods, neural networks, SVM, radial basis functions, etc.) [51]. In a previous work [52], we compared two generalization modes (vector quantization and tile coding) in this problem domain (pedestrian dynamics). Our study concluded that the two approaches gave similar results. In the current experiments we have chosen Tile Coding, because it needs few parameters to adjust and has given good results in many different scenarios and other problem domains such as mountain car and puddle world [62] or Robocup Soccer [63]. Tile Coding is a function approximator based on the Cerebellar Model Articulation Controller (CMAC) structure proposed by Albus [53]. It constitutes a specific case of the parameterized function approximators where the functions are approximated with a linear combination of weighted binary-valued parameters. In tile coding, the space is divided in partitions called tilings. Each tiling covers all the space so there are as many partitions as tilings. Each element of a specific tiling is a tile and, given a point in the state space, there is only one active tile per tiling associated to this point. Given m tilings and k tiles per tiling, then $m \cdot k$ tiles exist. A binary vector $\vec{\phi}$ indicates the active tiles in each interaction at time t, and the vector $\vec{\theta}$ stores the value of the tiles. Therefore, for each tile i, $\phi_i(s)$ indicates if it is active (value 1) or not (value 0) for the state s. A weight stored in a table $\theta(i)$ indicates its value. The value function for each action Q^a and state s at time step t, is described in Eq. (5)

$$Q_t^a(s) = \vec{\phi}^T \vec{\theta}_t^a = \sum_{i=1}^{m \cdot k} \theta_t^a(i)\phi_i(s) \tag{5}$$

where the super index T means the matrix transpose. There is a vector $\vec{\theta}_t^a$ for each available action. The code of a point of the state space is given by the binary features $\phi(i)$ that have value 1, remaining the rest with value 0. Fig. 2 shows a bidimensional tilecoding for simplicity. Hashing is used to avoid the curse of dimensionality. The state space is disperse and, therefore, the significant states are a small fraction of the total space. Hashing exploits this circumstance. The size of the θ_t^a tables is an important parameter in determining the resolution of the generalization process.

Other members of MDP definition are the actions. The pedestrian dynamics is defined by the walking velocity. Therefore, actions must modify this velocity (speed and direction) in order to change the dynamics. Nine actions have been defined to modify speed and other nine to modify the direction. In each group, four of them increment the value and four of them

Sag	Module of the velocity of the agent.
Av	Angle of the velocity vector relative to the reference line.
$Dgoal$	Distance to the goal.
$Srel_i$	Relative scalar velocity of the i-th nearest neighbor.
Dag_i	Distance to the i-th nearest neighbor.
Aag_i	Angle of the position of the i-th nearest neighbor relative to the reference line.
Dob_j	Distance to the j-th nearest static object (walls).
Aob_j	Angle of the position of the j-th nearest static object relative to the reference line.

S = Sag, Av, Dgoal, Srel1, Dag1, ...

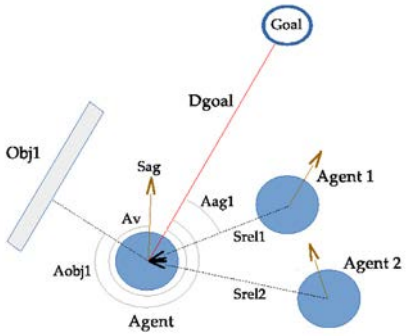

Fig. 1. State space features. The reference line (in red) joins the agent with its goal. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

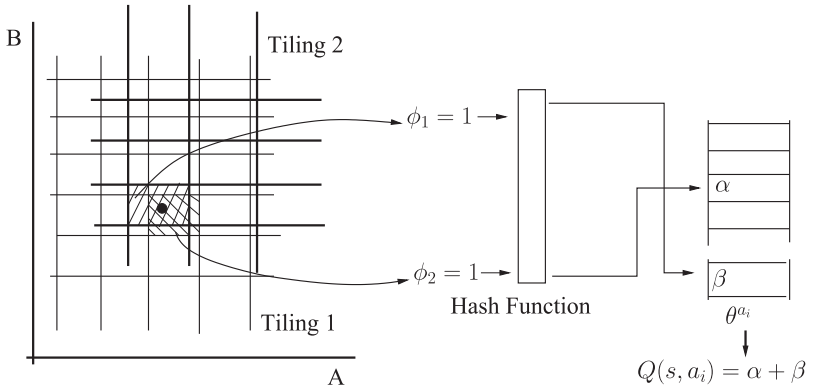

Fig. 2. Tilecoding for a bidimensional space.

decrement it. In case of the direction, they are positive and negative angle variations from the reference line that joins the virtual pedestrian with the goal. There are one action that does not modify the corresponding value. The increments and decrements are fractions $(1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8})$ of a fixed quantity determined empirically to create acceleration and deceleration values corresponding to real pedestrians studies [54]. Actions are taken by pairs (one related to speed and other one related to direction) giving a total of 81 different possible actions. Fig. 3 shows the possible actions for modifying the direction (left) and the speed (right) of an agent (represented by a square inside a circle).

The last element of the MDP setting is the reward function. The reward function introduces information about specific dynamic situations judging whether they are appropriate or not, thus modeling the behavior of the pedestrians. The reward

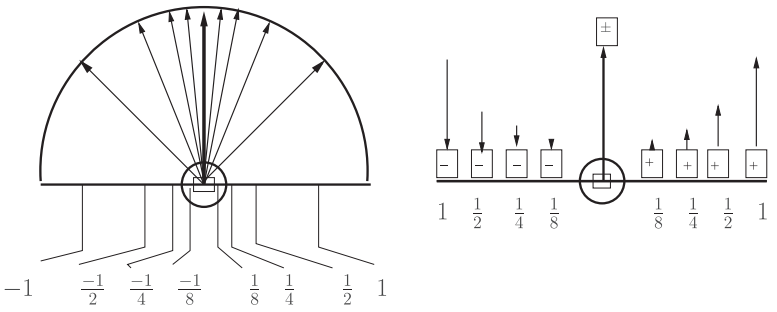

Fig. 3. Definition of the actions space. Left: the velocity can change its direction a fraction of the unit angle (45 degrees in the image). Right: the speed can increment or decrement in a fraction of the unit value. If the result is higher than the maximum allowed speed or less than zero, the speed remains in the limits.

function is discrete and gives positive or negative feedback only in a few situations (i.e. to reach the goal, to collide with other agent or a wall). Most state-action-new_state sequences are rewarded with 0 value indicating that there is no information about the correctness of this transition. The reward function works on a per-agent basis, therefore, in a given step, different agents can get different immediate rewards depending on both the individual actions chosen by each one and their consequences.

In our framework, each agent computes an independent and different learning process based on its own experiences in the interaction with the virtual world. Therefore, each virtual pedestrian controlled by an agent has its own dynamics which is different to the rest of the virtual pedestrians controlled by other different agents. Each learning process is divided in *trials*. During a trial, the virtual agent tries to adapt the walking velocity to reach the specified goal (a small region of the virtual world), and both, the decisions taken in the trial and the immediate reward given by the environment agent, are the signals for computing a near-optimal value function Q^*. Typically, a process has thousands of trials.

3.3. The virtual world

The 3D virtual scenarios are designed using basic geometric shapes (prisms for representing walls and spheres for representing pedestrians). They lay on a finite plane which represents the floor. If a virtual pedestrian goes out of this plane is considered as a fail in its current simulation (called trial in the graphics) and it is deactivated until the next one. The ODE physics engine calculate the movements and interactions inside the virtual world which are modeled following studies of real pedestrians. Collisions are modeled as a spring-dumping system calibrated to generate similar forces to those that occur in real human interactions. Friction forces with the floor and objects are also modeled with friction coefficients that represent real materials. A description of the calibration process and the associated validation can be found in the work [43].

4. Experimental scenarios

Each experiment presented below is carried out in two steps. First, a learning phase in which each agent (virtual pedestrian), *independently*, uses the methodology presented in Section 3 to learn a value function that represents its behavior. Second, each learned behavior, that is, each value function is replicated in many virtual pedestrians to simulate a crowd.

The selected scenarios illustrate different kinds of emergent collective behaviours. We will give an acronym to each experiment in order to reference them easily in the text and the tables.

4.1. Scenario 1: Quickest path versus *shortest path (QvS)*

In QvS, an agent has to choose between two exits to reach the goal. One exit is placed near the goal and the other one is situated farther from it. If the number of agents is large, a jam is generated in front of the exit next to the goal. But other alternative path is available that detours this group using the other exit. Assuming that the extra effort to perform the detour is small, part of the agents of the jam borders may decide to follow the detour instead of keeping waiting inefficiently. This problem happens in different situations in real life (for example when pedestrians hustling through a station hall as they are late for a train) and it differentiates the pedestrian dynamics from other vehicle dynamics [55].

4.2. Scenario 2: Two opposing groups in front of a door (1D2G)

In 1D2G, one group of pedestrians is facing each other separated by a wall with a door. Each group wants to go through the door accessing to the space the other group occupies. In real studies of narrow passages, oscillatory changes in the walking direction were observed [1,18]. When a pedestrian crosses the door, it is easy for other pedestrians walking in the same direction to follow it. This leads to a deadlock situation [18] in which a line of pedestrians walking in a direction

Table 1
Emergent behaviors and their types of optimization.

Scen.	Emergent behavior	Type of optimization
QvS	Choice of the fast (but longer) path	Time saving avoiding congestion
1D2G	Alternating flow of pedestrians	Efficient use of the facility
2D2G	Each member of a group selects the same door	Avoiding alternating flows of pedestrians in a door
4WI	Roundabouts	Decrease of deceleration, stopping and avoidance maneuvers
FF	Plane collision avoidance routes	Time saving

Table 2
Values of the learning parameters.

Scen.	# Ag.	α	γ	ε_0	# Tilings	# Trials
QvS	23	0.001	0.9	0.9	128	50,000
1D2G	20	0.005	0.95	1.0	64	60,000
2D2G	8	0.004	0.9	0.4	64	150,000
FF	15	0.004	0.9	0.5	32	35,000
4W1	12	0.002	0.9	0.5	64	160,000

crosses the door. This is followed by a change of the direction of the flow when a pedestrian of the opposite side takes advantage of a gap or delay in the line of pedestrians, inverting the situation. The mechanism leads to alternating flows.

4.3. Scenario 3: Two opposing groups separated by a wall with two doors (2D2G)

This scenario is similar to the previous one with the difference that now there are two doors in the wall. Both groups must exchange their places leading to a route choice problem. The optimal solution is that each group in a direction select a different door in order to avoid the type of congestion that appears in the scenario 2 (1D2G). This scenario was reported in the work [18] where the authors state that a pedestrian self-organization as the described is much more efficient than one single door of double width.

4.4. Scenario 4: Free field where each pedestrian goes to a different place (FF)

A group of pedestrians are put together in the center of a space without obstacles. Each pedestrian has to arrive at a different point (goal) placed in the left side of the space. Pedestrians are initially placed randomly inside the group so that many trajectories are intersected creating risk of collision situations. This scenario represents a basic layout in which multi-directional flows without obstacles can be analyzed. This scenario also serves to demonstrate the adequacy of the local-oriented features that describe the state space, since the goals can be placed in other sides (different to the original one) in the scaling experiments.

4.5. Scenario 5: Four-way intersection (4WI)

Four groups of pedestrians inside two perpendicular corridors move toward their respective opposite side of the corridor. The scenario is a crossroad where the four groups of pedestrians cross simultaneously in the central square. If the pedestrians cross the square in a straight way, many collisions appear making the maneuver difficult. On the contrary, if a slight roundabout movement is adopted by all the pedestrians, the crossing becomes efficient. Roundabout traffic in this situation reduces deceleration and stopping actions making pedestrian motion more efficient on average [18]. Real intersections with three or more flows have been also studied in [19], where the rotary traffic is considered a good solution to the problem.

All the scenarios represent optimization problems and the emergence of collective behaviors appear as a consequence of the minimization of the individual interactions in order to reach the goal in a limited time (number of steps). Moreover, the route choice scenarios (QvS and 2D2G) are also considered higher level problems corresponding to the tactical level [56]. Table 1 resumes the type of emergent behavior expected and the type of optimization in which the experiment is committed.

4.6. Learning experiments set up

The learning process is different for each scenario. Specific values of the learning configuration parameters are set for the different scenarios. These parameters have been adjusted by means of trial and error processes from initial standard values. The learning process is monitored by using a performance indicator, the number of trials that the agent has reached to the goal. The curve of evolution of this indicator reaches an asymptotic zone in all the experiments, indicating that the agent has converged to a policy. It is the signal to stop the learning process. Table 2 shows the values of the parameters

Table 3
Immediate rewards that compose the reward function for each scenario.

Reward	QvS	1D2G	2D2G	4W1	FF
Goal	100	100	100	100	100
Ag. collision	0	0	−1	−0.6	-1
Wall collision	0	0	−1	−0.1	0
Off the limits	−10	−5	−1	−1	−5

that configure the learning processes of the different scenarios. Table 3 displays the immediate reward values for all the scenarios.

It is interesting to remark that the immediate rewards model the agents' behavior. Although the reward function is the same for all the agents in the experiments of a specific scenario, the experience through the interactions with the environment is different for each agent. Therefore, the reward function shapes the behavior according to the individual experience of each agent. Moreover, the reward function is different for each scenario as it is related with the specific task to solve.

In order to clarify the learning process, let us conclude the section explaining the mechanism of the RL optimization in reference with the studied scenarios. There are four main factors in the learning process: the exploratory regime, the immediate rewards, the discount factor γ and the learning rate α. The exploratory regime is the responsible of finding and refining the solution. Although an agent has found a solution, the exploratory policy forces the agent to test other possibilities (actions) to find a better policy. Exploration makes possible that several agents in the QvS scenario find the alternative quickest path. The immediate rewards punish inadequate reactions such as crash against a wall or other agents. Immediate rewards do not only affect the state where the wrong action was taken but they are propagated to the subpolicies that take the agent to this state. In this propagation, the discount factor γ has a main role. High values of γ impose that immediate rewards have more influence (in the sense that subpolicies that arrive at this rewarded state change significantly the values of their state-action pairs). On the contrary, low values of γ restricts the influence to the previous local state-action pairs of a decision sequence that visits the rewarded state. Immediate rewards are specially important in the 1D2G experiment. In the state-action pairs near the door, the agents learn to avoid (if possible) collisions. This behavior produces gaps that are taken advantage by other agents to change flow direction . On the other hand delayed rewards (controlled by the γ parameter) are specially important in the 2D2G experiment, where the final immediate reward obtained when the goal position is reached must propagate in a long sequence of decisions creating the learned policy. Finally the α parameter establishes the speed at which new experiences modify the policy being learned. With low values of α the rate of change of the policy is also low producing a soft convergence to the optimal policy, although the whole learning process can be very slowly. Therefore, the adjustment of the α parameter implies a trade-off between quality of the learned policy and duration of the learning process. RL algorithms are very stable in the sense that little changes in the values of the parameters imply little changes in the result. We have tested values inside the typical range for problems with this complexity.

5. Results: scalability and behavior evaluation

As described in the introduction, the experiments have two purposes, being both important for simulation. First, we analyse whether the expected emergent collective behaviors appear from the learned policies, then, we study the scalability of the learned behaviors (policies) when the number of agents is multiplied through agent cloning without additional learning. In this analysis we use:

- A collection of images provided by the simulations of pedestrian groups to appreciate the formation of emergent collective behaviors.
- The density maps that display the occupation of the surface (floor) by pedestrians. Density maps are actually histograms, since the floor is divided in small areas and the histogram counts the number of times that a pedestrian occupies this area. From a statistical point of view it registers the statistical frequency (number of times that an event occurs). An event consists of occupying a region of the space. Density maps are good indicators of the paths created by the virtual pedestrians to reach their goals.
- The fundamental diagram generated in the central area of the 4WI scenario (a complex and rich situation in terms of pedestrian dynamics) and its comparison with the fundamental diagrams using real data from similar scenarios.
- Performance tables: statistical analysis of the performance obtained. Here, the performance is defined as the number of agents able to reach their goals.

5.1. Emergent behaviors and scaling results

Next, we present the results classified by scenarios. One scaled situation is presented together with the results of the original learning configuration. While the images of each column correspond to frames of one trial, the density map, at the

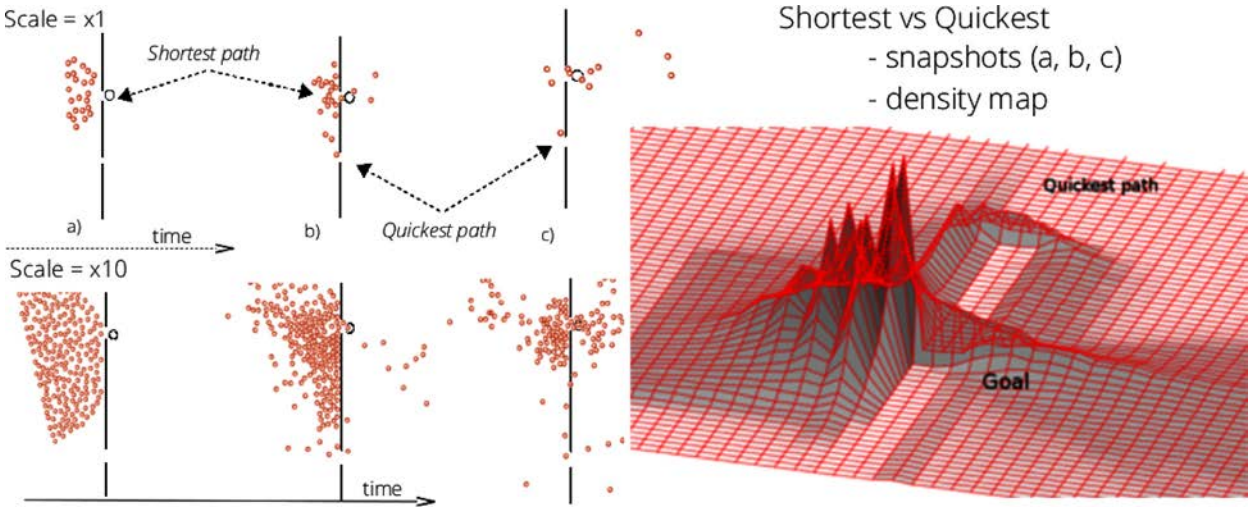

Fig. 4. QvS scenario at different time and scale (X1, X10).

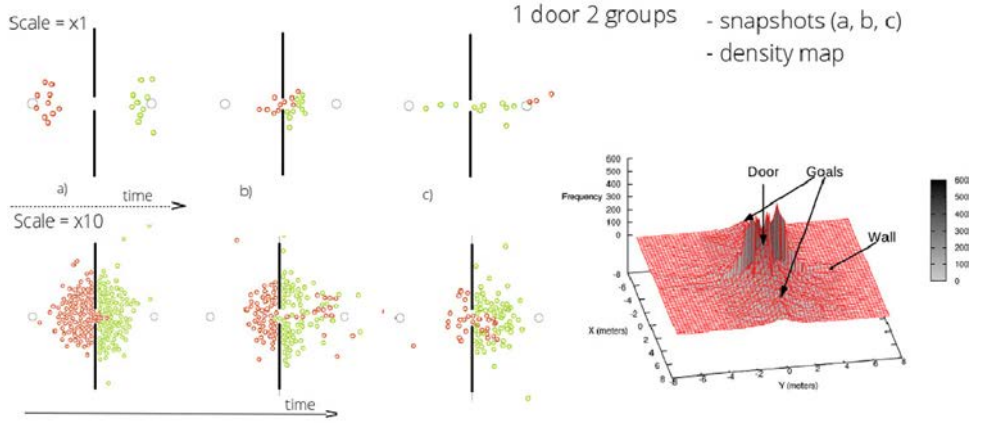

Fig. 5. 1D2G scenario at different time and scale (X1, X10).

bottom of each column, represents the accumulated frequency of several trials. Specifically each map accumulates the activity of 100 trials. Each column of the tables displays a sequence of three images corresponding to the beginning, middle and final steps of one simulation. The images accurately represents the virtual environment managed by the physical engine. The obstacles (walls) are represented by black lines and the goals are the black circumferences. The visualization is configured so that agents reach the goal and pass by.

In Fig. 4, the results for the QvS scenario are displayed. Each column displays a different scale rate: the one labeled with '× 1' corresponds to the original experiment without scaling up the number of agents , in the scale labeled '× 10', the number of agents has been increased in a 10 factor without additional learning. Looking up Table 2, the original number of agents for this experiment is 23, therefore '× 10' means that 230 agents have been used in the simulations.

As introduced in Section 4, this is a route choice problem. The sequence of column labeled '× 1' shows that a fraction of the group has learned to make a short detour following a quicker path than the shorter one which goes through the nearest door to the goal. The image in the middle of the sequence shows the group division where agents located below choose the far door. This emergent behavior is reproduced when scaling up the number of agent ('× 10' column). As stated before, the configuration allows the agents to reach the goal and pass by. This is the reason why, in the third image of the sequence, many agents have exceeded the goal position.

The density map corresponding to the '× 10' experiment shows clearly the formation of the quickest path as well as the congestion originated in front of the shortest path door.

In Fig. 5, the results for the 1D2G scenario are displayed. As reported in Section 4, the solution for the congestion is the alternation in the dominance of the space next to the door. This alternation appears in the different sequences although the frequency varies with the number of pedestrians. When many pedestrians concentrate in the neighborhood of the door (such as in the '× 10' case), the pressure on the pedestrians nearest to the door means that one agent of the other group,

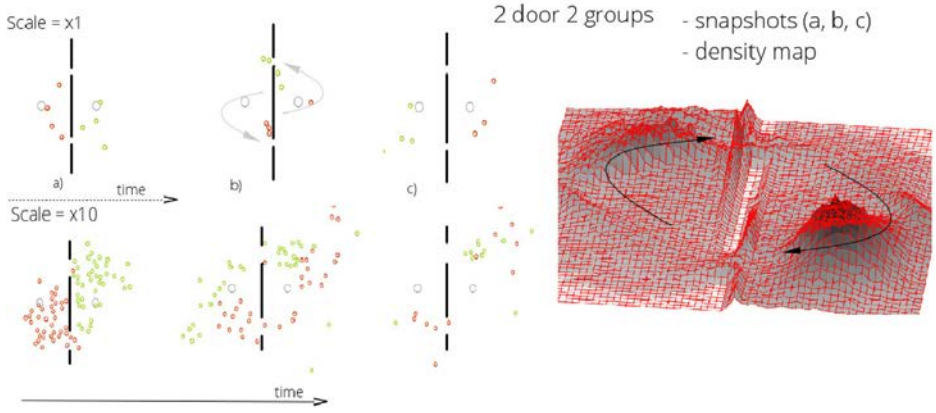

Fig. 6. 2D2G scenario at different time and scale (X1, X10).

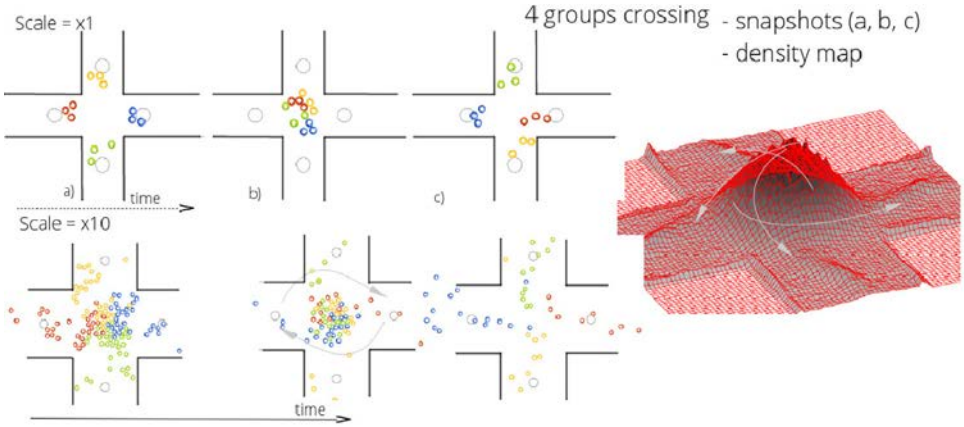

Fig. 7. 4 groups crossing scenario at different time and scale (X1, X10) .

can be introduced into a gap changing the direction of the flow through the exit. This effect is displayed in the '× 10' column with the first and second images, where the red flow and the green flow alternate in the dominance of the door. Specifically, in the central image of the '× 10' experiment, we can see a group of red pedestrians that have crossed the door and a group of green pedestrians just getting the control of the flow near the door. In the case of "× 1' simulations, the pedestrians next to the door are not under this pressure and the common solution consists of waiting the evacuation of a whole group to begin with the other group (which is also an extreme case of alternation).

The density map for the '× 10' experiment shows the path through the door that communicates both goals. The density map is not able to show the alternation of the groups because it is shaded by the continuous flow through the door.

In Fig. 6, the results for the 2D2G scenario are displayed. This problem belongs to the route choice class of decision problems such as the SvQ experiment. In this scenario, the expected emergent behavior consists of the organization of the groups in two flows that go through different doors towards their goal. Note in the images corresponding to the '× 1' experiment that the green agents selects the door at the top and the red group selects the door at the bottom. As the '× 10' images show, this behavior appears not only in the original experiment but in the scaled scenario. In this sequence, a red agent selects the door on the top as it can be seen in the first image of the sequence. This variability of behaviors is considered beneficial to the realism of the simulation and it is intrinsic to the stochastic nature of the RL approach.

The density map corresponding to the '× 10' experiment shows the evacuation flows and the congestion created near the doors. It suggests that the simulated pedestrians arrive at the same time to the door generating a compact group and flow.

In Fig. 7, the results for the 4W1 scenario are displayed. As described in Section 4, when four groups encounter in the middle of a crossing, an optimal way of behavior consists of performing a slight rotational movement which produces a sort of distribution around a fixed point placed in the middle of the crossing. In our scenario, the pedestrians generate an emergent collective behavior consisting of a roundabout movement of the pedestrians that distribute the flows from the crossing area towards the different goals. When the number of the agents is high ('× 10' experiment) the pedestrians that occupy the most external part of the crossing began to perform the roundabout creating a sort of swirl that distributes the flows towards their goals.

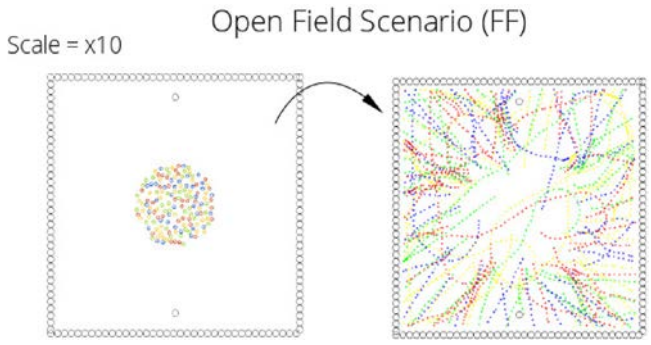

Fig. 8. FF scenario scaled X10. The goals are represented as black circumferences.

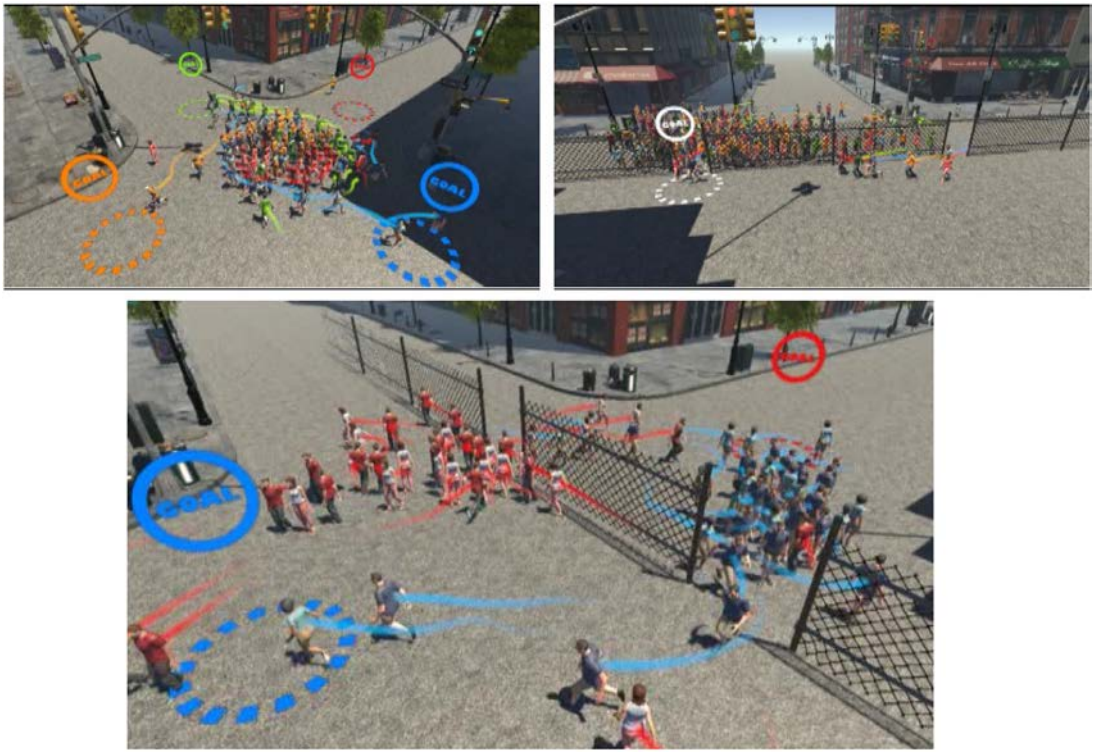

Fig. 9. Simulated scenes in 3D (4WI, QvS, 2D2G).

The density map shows the distribution flows from the crossing areas towards the different goals. The four arms of the swirl are clearly displayed.

In Fig. 8, the results for the FF scenario are displayed. This experiment is useful to demonstrate the adequacy of selecting local features to describe the state space of the pedestrians. In the learning setup, the goals (each agent has a different goal) are placed in only one side of the floor. Next, the scaling experiments are performed placing the goals in different sides. The virtual agents' trajectories for the '$\times$ 10' experiment are displayed in Fig. 8. The agents reach their goal in many simulations independently of the side where the goal is placed. Note that many trajectories go straight to the corresponding goal. However, due to the random assignment of the goals to agents, several agents have to carry out detours to get the proper orientation generating curved trajectories.

Fig. 9 summarizes the experiments carried out in different 3D scenes.

5.2. Pedestrian dynamics analysis in the 4WI scenario

The central zone in the 4WI scenario is an interesting area to analyze the dynamics due to the high density values reached and the confluence of the four pedestrian streams. We have used the fundamental diagram to assess the dynamics generated in this area. The fundamental diagram is one of the main tools for pedestrian dynamics analysis [45,57] and rep-

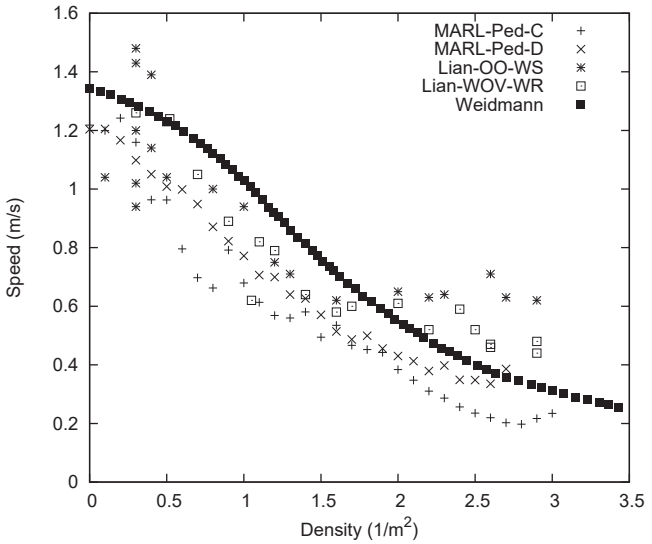

Fig. 10. Comparison of the fundamental diagram for the 4WI scenario in (MARL-Ped) with real-data fundamental diagrams: MARL-Ped-C: measurement at the center of the crossing. Points represent means of 100 simulations. MARL-Ped-D: measurement in a point displaced slightly from the center of the crossing. Weidmann: data reported by [45]. LianWOV-WR, OO-WS: data reported by [58].

resents the relationship between speed or flow to density. In Fig. 10, the fundamental diagrams generated by our simulation with 120 agents (4WI '$\times$ 10') are displayed together with those described in the paper by Lian et al. [58] and Weidmann [45]. The experiment described by Lian et al. is similar to the 4WI experiment and was carried out using real pedestrians. Our layout is similar to Lian et al.'s. Specifically, the length of the corridors is 15 m, and the area of the crossing is a square of 3x3 m^2. Moreover, both experiments use the method described by Helbing et al. [59] to calculate the local density and velocity. Specifically, the contribution to the local density of each pedestrian is weighted by a gaussian distance-dependent function defined as:

$$f(\vec{r}_j(t) - \vec{r}) = \frac{1}{\pi R^2} exp(-\|\vec{r}_j(t) - \vec{r}\|^2/R^2) \tag{6}$$

where $\vec{r}_j(t)$ is the position of pedestrian j at time t, $\vec{r}$ is the position of the measure point and R determines the effective radius (always greater than R). Fig. 10 shows the local velocity computed as a weighted mean defined by f in Eq. (6). The Weidmann's curve is included as a standard reference and consists of a combination of 25 independent experiments with uni and bi-directional flows.

In this comparison we have used two experiments described in the paper by Lian et al. and labeled in the original paper as OO-WS and WOV-WS. The main difference is that the WOV-WS experiment places a rectangular obstacle of dimensions 1.1 mx0.8 mx0.75 m at the center of the crossing, while the OO-WS does not have any obstacles. Our curves, labeled MARL-Ped-C and MARL-Ped-D derive from the same simulations but the curve MARL-Ped-C places the point of measurement in the center of the crossing while the MARL-Ped-D curve is calculated displacing the point 2 m. towards the North corridor. First we can observe the conformance of the obtained curves (MARL-Ped-C,D) to the basic property of the fundamental diagram for pedestrian dynamics: the speed reduces when the density increases. In low densities (less than 2 ped/m^2), the four curves have similar shapes all decaying faster than Weidmann's curve. For densities equal or greater than 2.0, the curves WOV-WR and OO-WS stabilize the velocity around the value of 0.55 m/s. In our curves this stabilization does not appear although there is a decrement in the slope from the density value 2.5 ped/m^2. This can be explained taking into account that real pedestrians can be strongly motivated in reaching the goal and pushing and jostling do not matter. Contrary, our pedestrians are rewarded negatively when crashing against other pedestrians so a decrement of the velocity to minimize collisions is justified. Of particular interest is the similarity found between the curves MARL-Ped-D and WOV-WR. When observing Fig. 7 in the '$\times$ 10' scale, we can see that the center of the crossing is permanently occupied by pedestrians. In the simulations we realize that these pedestrians are always the same individuals that get trapped in the center of the clogging until the peripheral pedestrians leave the central area. These individuals take the role of the obstacle in the WOV-WR experiment creating similar spatial configurations in both experiments.

5.3. Behavioural performance

The performance of the learned behaviors has been defined as the number of agents that reach the goal in a specified number of steps per simulation. This measure indicates the quality of the simulation.

Table 4
Performance of the experiments labeled '× 1' and '× 10'. Number of agents that reach the goal respect to all the agents.

Experiment	Successful agents	# of trials
QvS '× 1'	22678/23000 (98.6%)	1000
QvS'× 10'	18193/23000 (79.1.0%)	100
1D2G '× 1'	1509/2000 (87.75%)	100
1D2G '× 10'	11315/20000 (56.6%)	100
2D2G '× 1'	726/800 (90.75%)	100
2D2G '× 10'	3405/8000 (42.56%)	100
4WI '× 1'	1053/1200 (87.75%)	100
4WI '× 10'	9588/12000 (79.9%)	100
FF '× 1'	1403/1500 (93.5%)	100
FF '× 10'	12031/15000 (80.2%)	100

Table 5
Performance of the '× 100' scaled experiments. Number of agents that reach the goal respect to all the agents.

Experiment	Successful agents	# of trials
QvS '× 100'	1318/23000 (5.7%)	10
1D2G '× 100'	279/6000 (4.65%)	3
2D2G '× 100'	1388/8000 (17.4%)	10
4WI '× 100'	1903/12000 (16%)	10
FF '× 100'	13951/15000 (93.0%)	10

Table 4 displays the performance results for all the scenarios for the '× 1' and '× 10' experiments. In the first column the number of agents that reach the goal respect to the total number of agents is displayed. In parenthesis the same indicator in form of percentage is displayed. Several trials have been performed for this test. The number of trials carried out for each experiment is displayed in the last column of the table.

When scaling up the number of agents, a decrease in the success rate is produced. This was expected because the navigation of pedestrians is more difficult in the scaled experiment as a result of an increased number of interactions among pedestrians. However, different results were obtained depending of the scenario. In the FF scenario, rates were slightly affected by scaling which indicates that, in simple scenarios (those in which the operational level is the most important), increasing the number of agents will give good results. In the more complex scenarios (with tactical and strategic levels), a rate reduction of 10% was produced in 4WI and a reduction of 20% was produced in 1D2G and QvS when scaling × 10. Therefore, success rates of 80–100% were noted after scaling, which indicates the robustness of the learned behaviors. In contrast, low success (42.56%) was obtained in the 2D2G scenario.

5.4. Current limitations

In this subsection we present two additional tests that reveal limitations in the simulation processes. The first test scales three scenarios two orders of magnitude (X100). This important increment in the number of agents forces to modify the size of the virtual world so that, sometimes, the states sensed by the agents rarely would have been explored in the learning process. In the second test we analyze the capability of our system to generate simulations in which all the agents reach the goal. The aim is to explore the limits of the learned behaviors in both scalability and consistence of the simulations.

Table 5 shows the performance of the '× 100' scaled experiments in different scenarios. The performance is defined as the number of agents that reach the goal given a fixed number of trials. The results indicate that, in many experiments, the agents have difficulties to reach the goal. Likely, the agents are only exploiting a part of the learned control that corresponds to states with high densities, what, in many cases, doesn't seem to be enough to control accurately the virtual agent towards the goal. The exception is the FF scenario in which the performance value is over 90%. As commented above, all experiments are performed in scaled virtual environments to allocate the crowd.

A visualization of the simulations reveals that the emergent collective behaviors are still present in the 4WI, FF and 2D2G scenarios. Figs. 11 and 12 show similar spatial collective structures compared with the previous scales.

In Fig. 11 (with 800 agents) is clearly visible that each group of agents selects a different door. The corridors created by the crowd towards the goals are also present. However, part of the group is not able to reach the goal remaining in a unproductive sequence of movements as the groups of agents at the top right of the image (green agents) and at the bottom left (red agents) indicate. Likely, these agents are continuously perceiving states that have not been correctly learned.

In Fig. 12 (with 1200 agents), the agents in the external areas of the crossing create groups with the same color. It indicates that peripheral agents solve the crossing performing a roundabout movement until they are aligned with the corresponding corridor. In the center of the crossing, a group of agents of different colors is waiting the change from inside to outside positions.

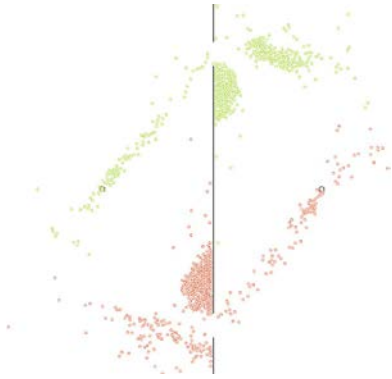

Fig. 11. Collective behaviors appeared in 2D2G scenario when using the scale X100.

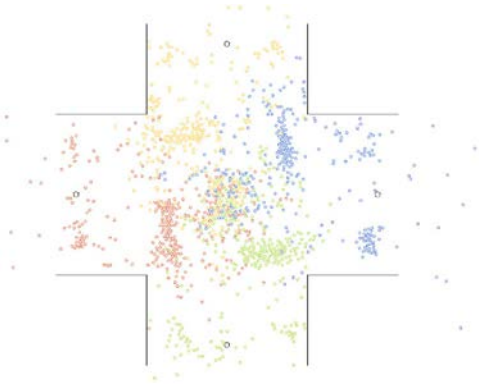

Fig. 12. Collective behaviors appeared in 4WI scenario when using the big scale ('× 100').

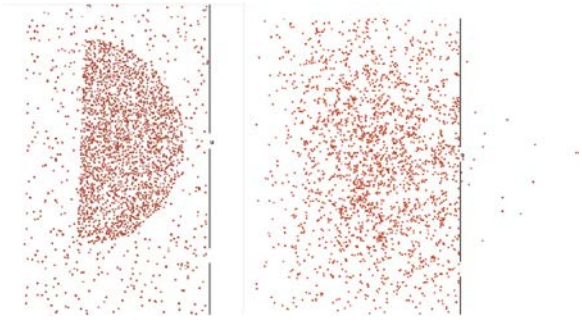

Fig. 13. Behavioural problems appeared in QvS scenario when using the big scale ('× 100'). A sequence of two instants is displayed.

Fig. 13 shows the QvS scenario with a major dispersion compared with the previous scales. In this case the experiment manage 2300 agents using the knowledge learned in the original group (23 agents). In this case the emergent behavior has disappeared. The behaviors do not generate paths that lead to the goal.

As a second test, we study the capability of the system to generate completely successful simulations. In these simulations, *all* the agents have to arrive to their respective goals. Table 6 shows the results of the study in the '× 1' experiments for all the scenarios.

Note that all the scenarios have a percentage greater than 0 of generating completely successful simulations. From a methodological point of view, when at least a simulation of an experiment is successful, the system is useful to produce behavioral animations. However, 100% of consistence in real time simulation is not guaranteed. On the contrary, the percentage of simulations that will show total correctness is displayed in Table 6. As stated at the beginning of the subsection, this measure shows the limit of the system's performance. This does not mean that the rest of the simulations were useless. For example, consulting Table 4, the percentage of successful trials for an agent in the 1D2G scenario is 87.75% in the '× 1' experiment. Therefore, although only 3% of the simulations are completely correct, there are many simulations in which a high percentage of pedestrians will reach the goal.

Table 6
Number of successful trials respect to all the trials performed in the '× 1' experiments.

Experiment	Successful trials
QvS	723/1000 (72.3%)
1D2G	3/100 (3%)
2D2G	48/100 (48%)
4WI	20/100 (20%)
FF	41/100 (41%)

6. Conclusions

In this paper, a set of experiments in different scenarios are presented to assess the capability of our Multi-agent RL framework (MARL-Ped) to generate emergent collective behaviors and their robustness when scaling in the number of agents.

As emergent collective patterns are easy to recognize, their evaluation can be done with sequence of images and density maps which indicates that the collective behaviors reported in the pedestrian simulation literature also appears in our Multi-agent RL approach. These results suggest that this learning framework is capable of solving different kinds of navigational problems (route choice, bottlenecks, crossings) while displaying emergent collective behaviours. Moreover, we have compared the pedestrian dynamics generated by our learned behaviors with those generated by real pedestrians in the 4WI scenario as described in Lian et al. paper [58]. The fundamental diagrams show similarities indicating that the dynamical behaviors resemble those of real pedestrians. The densities reached in the analysis of the 4WI experiment have expected values for normal walking situations. Higher densities could be reached when simulating groups with heterogeneous sizes and mechanical responses and it can be considered as a future work.

The performance tests also indicate that the learned behaviors are robust when scaling up the number of agents in one order of magnitude ('× 10'). Moreover the scale '× 100' also shows robustness in several scenarios (4WI, FF and 2D2G). As commented in Section 5.4, the dimensions of the scenarios had to be increased to allocate the necessary number of pedestrians. In addition, the robustness of the results indicates that the local-oriented description of the state space together with the learned behaviors are capable of generalizing the physical environment.

However, the system shows limitations. Despite the fact that similar spatial collective structures to those of the '× 1' and '× 10' experiments appear, a low percentage of agents reach the goal. A more exigent performance test carried out with the '× 1' experiments indicates that the learned behaviors are capable of generating simulations in which all the agents reach the goal. However, a consistency in the generation of totally successful simulations cannot be guaranteed.

Several issues remain as future work that can palliate these limitations. RL offers a wide set of techniques that can be used to solve them. Among them, techniques of learning by examples or by means a teacher can be used to develop tools to edit incorrect behaviors. Also reward shaping [60] and policy shaping techniques [61] can also be used for the same purpose.

Despite these drawbacks, the Multi-agent RL approach introduces machine learning techniques into the pedestrian simulation field as an alternative to the traditional pedestrian simulation frameworks where the pedestrian behaviors are defined by an expert or obtained after processing real data.

Acknowledgement

The authors would like to thank Rosa Maria Sánchez and Hector Barreiro their excellent work on 3D scene composition. This work has been supported by grant TIN2015-65686-C5-1-R of Ministerio de Economía y Competitividad.

References

[1] D. Helbing, P. Molnár, Self-organization of complex structures: from individual to collective dynamics, Gordon and Breach, London, Ch. Self-organization phenomena in pedestrian crowds (1997) 569–577.
[2] M. Bierlaire, T. Robin, Pedestrians choices, in: H. Timmermans (Ed.), Pedestrian Behavior, Emerald, 2009, pp. 1–26.
[3] S.J. Guy, J. Chugani, S. Curtis, P. Dubey, M. Lin, D. Manocha, Pledestrians: a least-effort approach to crowd simulation, in: Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA '10, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 2010, pp. 119–128.
[4] F. Lamarche, S. Donikian, Crowd of virtual humans: a new approach for real time navigation in complex and structured environments, Comput. Graphics Forum 23 (2004) 509–518.
[5] A. Sud, E. Andersen, S. Curtis, M. Lin, D. Manocha, Real-time path planning for virtual agents in dynamic environments, in: ACM SIGGRAPH 2008 Classes, SIGGRAPH '08, ACM, New York, NY, USA, 2008, pp. 55:1–55:9.
[6] A. Kamphuis, M.H. Overmars, Finding paths for coherent groups using clearance, in: Proceedings of the 2004 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA '04, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 2004, pp. 19–28.
[7] C.W. Reynolds, Flocks, herds and schools: a distributed behavioral model, in: Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '87, ACM, New York, NY, USA, 1987, pp. 25–34.
[8] C. Reynolds, Steering behaviors for autonomous characters, in: Game Developers Conference, Miller Freeman Game Group, San Francisco, California, 1999, pp. 763–782.

[9] D. Helbing, P. Molnár, Social force model for pedestrian dynamics, Phys. Rev. E (1995) 4282–4286.
[10] C. Cordeiro, A. Braun, C. Silveria, R. Musse, G.G. Cavalheiro, Concurrency on social forces simulation model, in: Proceedings of the First International Workshop on Crowd Simulation, 2005.
[11] T.I. Lakoba, D.J. Kaup, N.M. Finkelstein, Modifications of the Helbing-Molnár-Farkas-Vicsek social force model for pedestrian evolution, Simulation 81 (5) (2005) 339–352.
[12] A. Sud, R. Gayle, E. Andersen, S. Guy, M. Lin, D. Manocha, Real-time navigation of independent agents using adaptive roadmaps, in: ACM SIGGRAPH 2008 Classes, SIGGRAPH '08, ACM, New York, NY, USA, 2008, pp. 56:1–56:10.
[13] S.R. Musse, D. Thalmann, A model of human crowd behavior: group inter-relationship and collision detection analysis, in: Proceedings of the Workshop of Computer Animation and Simulation of Eurographics 97, 1997, pp. 39–51.
[14] N. Pelechano, Crowd simulation incorporating agent psychological models, roles and communication, in: First International Workshop on Crowd Simulation, 2005, pp. 21–30.
[15] M. Sung, M. Gleicher, S. Chenney, Scalable behaviors for crowd simulation, Comput. Graphics Forum 23 (3) (2004) 519–528.
[16] J. Funge, X. Tu, D. Terzopoulos, Cognitive modeling: knowledge, reasoning and planning for intelligent characters, in: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '99, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 1999, pp. 29–38.
[17] Q. Yu, D. Terzopoulos, A decision network framework for the behavioral animation of virtual humans, in: Proceedings of the 2007 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA '07, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 2007, pp. 119–128.
[18] D. Helbing, P. Molnár, I. Farkas, K. Bolay, Self-organizing pedestrian movement, Environ.Plann B: Plann. Des. 28 (2001) 361–383.
[19] D. Helbing, L. Buzna, A. Johansson, T. Werner, Self-organizing pedestrian crowd dynamics: experiments, simulations and design solutions, Transp. Sci. 39 (1) (2005) 1–24.
[20] C. Burstedde, K. Klauck, A. Schadschneider, J. Zittartz, Simulation of pedestrian dynamics using a two-dimensional cellular automaton, Physica A 295 (2001) 507–525.
[21] J. Godoy, I. Karamouzas, S. Guy, M. Gini, Implicit coordination in crowded multi-agent navigation, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'16), Arizona, USA, 2016.
[22] S. Lemercier, A. Jelic, R. Kulpa, J. Hua, J. Fehrenbach, P. Degond, C. Appert-Rolland, S. Donikian, J. Pettré, Realistic following behaviors for crowd simulation, Comput. Graphics Forum 31 (2pt2) (2012) 489–498.
[23] Y. Li, M. Christie, O. Siret, R. Kulpa, J. Pettré, Cloning crowd motions, in: Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA '12, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 2012, pp. 201–210.
[24] E. Cristiani, B. Piccoli, A. Tosin, Multiscale modeling of granular flows with application to crowd dynamics, Multiscale Model. Simul. 9 (1) (2011) 155–182.
[25] A. Tosin, Collective Dynamics from Bacteria to Crowds, 553, Springer, 2014, pp. 157–177. Ch. Multiscale crowd dynamics modeling and theory
[26] N. Bellomo, A. Bellouquid, D. Knopoff, From the microscale to collective crowd dynamics, Multiscale Model.Simul. 11 (3) (2013) 943–963.
[27] A. Kneidl, D. Hartmann, A. Borrmann, A hybrid multi-scale approach for simulation of pedestrian dynamics, Transp. Res. Part C: Emerging Technol. 37 (0) (2013) 223–237.
[28] G. Lämmel, A. Seyfried, B. Steffen, Large-scale and microscopic: a fast simulation approach for urban areas, in: Transportation Research Board 93rd Annual Meeting, Washington DC, 2014, pp. 1–17.
[29] L. Crociani, G. Lämmel, G. Vizzari, Multi-scale simulation for crowd management: a case study in an urban scenario, 1st Workshop on Agent Based Modelling of Urban Systems, Singapore, 2016.
[30] M. Kapadia, M. Wang, S. Singh, G. Reinman, P. Faloutsos, Scenario space: characterizing coverage, quality, and failure of steering algorithms, in: Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA '11, 2011, pp. 53–62.
[31] S. Singh, M. Kapadia, P. Faloutsos, G. Reinman, Steerbench: a benchmark suite for evaluating steering behaviors, Comput. Anim. Virtual Worlds 20 (5–6) (2009) 533–548.
[32] C.D. Boatright, M. Kapadia, J.M. Shapira, N.I. Badler, Generating a multiplicity of policies for agent steering in crowd simulation, Comput. Anim. Virtual Worlds 26 (5) (2015) 483–494.
[33] M. Kapadia, N. Pelechano, J. Allbeck, N. Badler, Virtual crowds: steps toward behavioral realism, Synth. Lect. Visual Comput. 7 (4) (2015) 1–270.
[34] L. Crociani, S. Manzoni, G. Vizzari, A. Gasteratos, G. Sirakoulis, When reactive agents are not enough: tactical level decisions in pedestrian simulation, Intelligenza Artificiale 9 (2) (2015) 163–177.
[35] A.U.K. Wagoum, A. Seyfried, S. Holl, Modeling the dynamic route choice of pedestrians to assess the criticality of building evacuation, Adv. Complex Syst. 15 (07) (2012) 1250029.
[36] J. Pettré, J. Ondrej, A. Olivier, A. Creutal, S. Donikian, Experiment-based modeling, simulation and validation of interactions between virtual walkers, in: ACM SIGGRAPH/Eurographics Symposium on Computer Animation, ACM, New York, 2009, pp. 189–198.
[37] H. Kretzschmar, M. Kuderer, W. Burgard, Learning to predict trajectories of cooperatively navigating agents, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 2014.
[38] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1998.
[39] A. Treuille, Y. Lee, Z. Popović, Near-optimal character animation with continuous control, ACM Trans. Graphics (SIGGRAPH'07) 26 (3) (2007).
[40] Y. Lee, S.J. Lee, Z. Popović, Compact character controllers, ACM Trans. Graphics 28 (5) (2009) 169:1–169:8.
[41] L. Ikemoto, O. Arikan, D. Forsyth, Learning to move autonomously in a hostile world, ACM SIGGRAPH 2005 Sketches, SIGGRAPH '05, ACM, New York, NY, USA, 2005.
[42] O. Buffet, A. Dutech, F. Charpillet, Automatic generation of agent's basic behaviors, in: Proceedings of the Autonomous Agents and Multiagents Systems (AAMAS'03), Melbourne, Australia, 2003.
[43] F. Martinez-Gil, M. Lozano, F. Fernández, Calibrating a motion model based on reinforcement learning for pedestrian simulation, in: ACM SIGGRAPH Conference on Motion in Games (MIG 2012), Springer, Rennes, France, 2012, pp. 302–313.
[44] A. Seyfried, B. Steffen, W. Klingsch, T. Lippert, M. Boltes, Steps Toward the Fundamental Diagram — Empirical Results and Modelling, Springer Berlin Heidelberg, , 2007, pp. 377–390.
[45] U. Weidmann, Transporttechnik der fussgänger - transporttechnische eigenschaften des fussgängerverkehrs (literaturstudie), Literature Research 90, IVT an der ETH Zürich, ETH-Hönggerberg, CH-8093 Zürich, 1993.
[46] M. Mori, H. Tsukaguchi, A new method for evaluation of level of service in pedestrian facilities, Transp. Res. Part A 21 (3) (1987) 223–234.
[47] F. Martinez-Gil, M. Lozano, F. Fernández, Marl-ped: a multi-agent reinforcement learning based framework to simulate pedestrian groups, Simul. Modell. Pract. Theory 47 (0) (2014) 259–275.
[48] J. Russell, R. Cohn, Open dynamics engine, 2012.
[49] L.P. Kaelbling, M.L. Littman, A.W. Moore, Reinforcement learning: a survey, Int. J. Artif. Intell. Res. 4 (1996) 237–285.
[50] T. Robin, G. Antonioni, M. Bierlaire, J. Cruz, Specification, estimation and validation of a pedestrian walking behavior model, Transp. Res. 43 (2009) 36–56.
[51] L. Buçoniu, R. Babŭska, B.D. Schutter, D. Ernst, Reinforcement Learning and Dynamic Programming using Function Approximators, CRC-Press, 2010.
[52] F. Martinez-Gil, M. Lozano, F. Fernández, Emergent Collective Behaviors in a Multi-agent Reinforcement Learning Pedestrian Simulation: a Case Study, in: Multi-agent-based Simulation XV - International Workshop, MABS, in: Paris, France, May 5–6, 2014, Vol. 9002 of Lecture Notes in Computer Science, Springer, 2015, pp. 228–238.
[53] J.S. Albus, A new approach to manipulator control: the cerebellar model articulation controller (CMAC), J Dyn Syst Meas Control 97 (1975) 220–227.

[54] A. Steiner, M. Philipp, A. Schmid, Parameter estimation for a pedestrian simulation model, in: 7th Swiss Transport Research Conference (STRC'07), Monte Verità, Ascona, 2007, pp. 1–29.
[55] A. Johansson, T. Kretz, Applied pedestrian modeling, in: A. Heppenstall, A. Crooks, L. See, M. Batty (Eds.), Spatial Agent-based Models: Principles, Concepts and Applications, Springer, 2011.
[56] S. Hoogendoorn, P. Bovy, W. Daamen, Pedestrian and evacuation dynamics, in: Ch. Microscopic Pedestrian Wayfinding and Dynamics Modeling, Springer, 2001, pp. 123–154.
[57] A. Schadschneider, W. Klingsch, H. Kluepfel, T. Kretz, C. Rogsch, A. Seyfried, Evacuation dynamics: empirical results, modelling and applications, in: R. Meyers (Ed.), Encyclopedia of Complexity and Systems Science, Springer, 2008, pp. 3142–3176.
[58] L. Lian, X. Mai, W. Song, Y.K.K. Richard, X. Wei, J. Ma, An experimental study on four-directional intersecting pedestrian flows, J. Stat. Mech: Theory Exp. 2015 (8) (2015) P08024.
[59] D. Helbing, A. Johansson, H.Z. Al-Abideen, Dynamics of crowd disasters: an empirical study, Phys. Rev. E 75 (2007) 046109.
[60] W. Knox, P. Stone, Tamer: training an agent manually via evaluative reinforcement, in: Proceedings of the 7th IEEE ICDL, 2008, pp. 292–297.
[61] S. Griffith, K. Subramanian, J. Scholz, C. Isbell, A. Thomaz, Policy shaping: integrating human feedback with reinforcement learning, in: C. Burges, L. Bottou, M. Welling, Z. Grahramani, K. Weinberg (Eds.), Advances in Neural Information Processing Systems, Vol. 26, Curran Associates Inc., 2013, pp. 2625–2633.
[62] R.S. Sutton, Generalization in reinforcement learning: successful examples using sparse coarse coding, Adv. Neural Inf. Process. Syst. 1996 (8) (1996) 1038–1044.
[63] P. Stone, R.S. Sutton, G. Kuhlmann, Reinforcement learning for robocup soccer keepaway, Adapt. Behav. 13 (3) (2005) 165–188.