Human-Aware Multi-Robot Navigation in Constrained Environments Using Multi-agent Reinforcement Learning

Takieddine Soualhi[1], Jacques Saraydaryan[2], Leaticia Matignon[3]

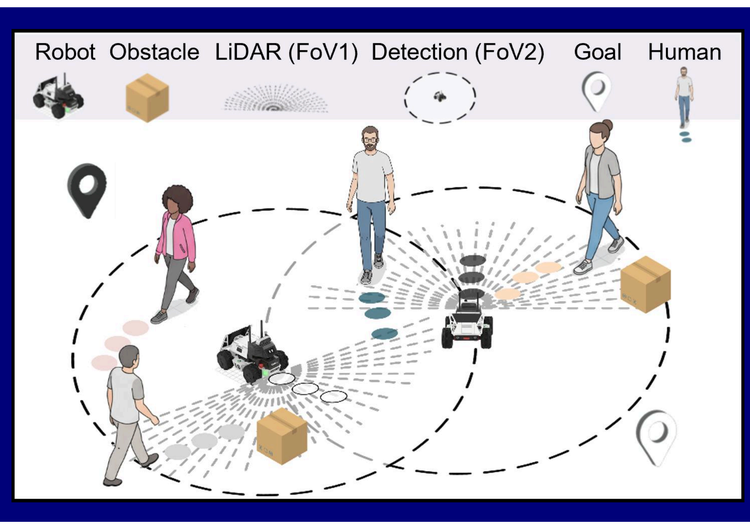
INRIA-INSA CHROMA team, Villeurbanne, France[1]
CPE Lyon, INRIA-INSA CHROMA team, Villeurbanne, France[2]
Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, Villeurbanne, France[3]

Context and Motivations

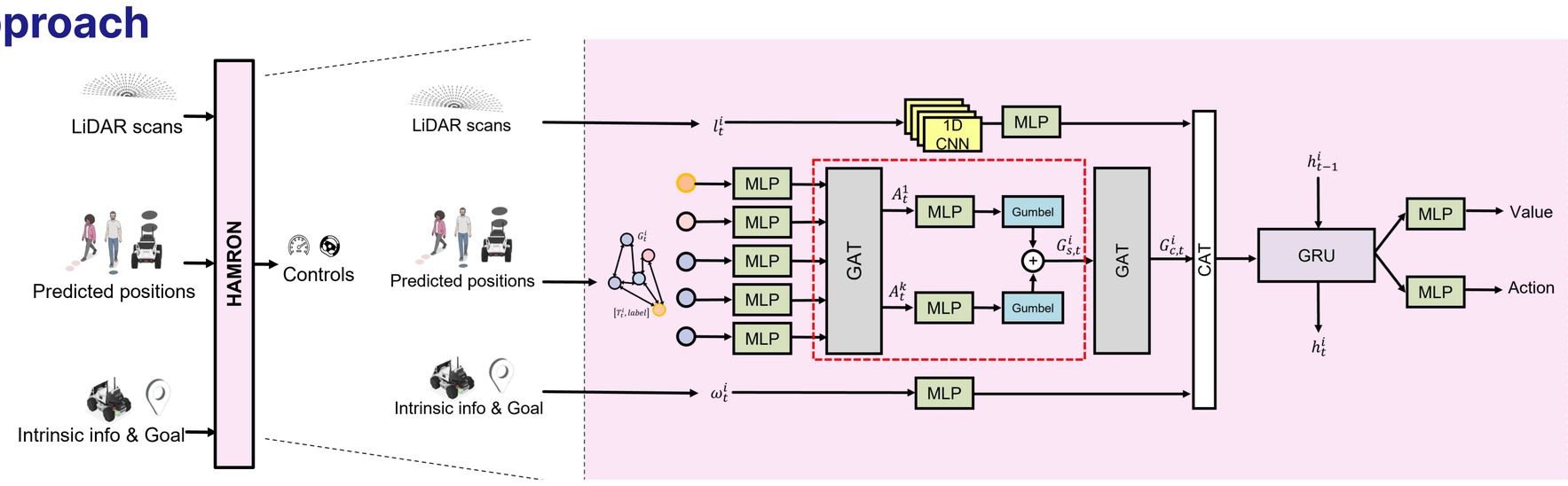
- Real-world social navigation requires robots to account for humans, static obstacles, and other robots.
- Existing work suffers from one or more key limitations, (i) operating only in open environments, (ii) being tailored to single-robot systems, and (iii) relying on a unified representation for both controlled (robots) and uncontrolled entities (humans and obstacles), which limits their deployability.
- Multi-robot social navigation is still underexplored, partly due to a lack of lightweight simulators that support multiple robots, pedestrians in constrained scenes.

Contributions

- We propose HAMRON: a human-aware multi-robot navigation method that learns decentralized policies with multi-agent RL for constrained (obstacle-rich) scenes.
- Hamron relies on a dual perception system: (i) a detection system to identify humans and other robots in the scene; and (ii) a 2D LiDAR to handle static obstacles.
- We present a lightweight training environment that extends CrowdNav[2] to multi-agent scenarios, adding 2D LiDAR, static-obstacle simulation, occupancy maps, and heterogeneous human policies (SF/ORCA).
- We compare against strong single-robot baselines and conduct ablation studies on scalability, generalization, and human-policy robustness.



Proposed approach



- **Observations:** The observation space comprises (i) 2D LiDAR scans for static geometry (FoV ≈ 230°), (ii) intrinsic state and goal, and (iii) an interaction graph of nearby humans and robots with short-horizon
- **Perception encoders:** LiDAR is encoded with a 1D CNN, while a GNN/GAT plus an interaction filter (GAT + Gumbel-Softmax mask) prunes irrelevant interactions. The merged features are then sent to a GRU policy head.
- **Rewards:** The reward function combines a success bonus, a collision penalty, a potential-based progress term, and a prediction-intrusion penalty that penalizes near-future incursions into others' predicted trajectories.
- **Control & learning:** We use differential-drive actions under nonholonomic constraints and train with IPPO using decentralized training and execution plus parameter sharing across the fleet.

Results

- Single-robot evaluation: HAMRON outperforms single-robot SOTA with the highest success rate (0.67) and the shortest travel time/length, but is slightly more assertive than HEIGHT[3],showing a higher collision rate.
- Multi-robot evaluation: HAMRON outperforms LiDAR-Nav[4] and HEIGHT in success (0.62), timeouts (0.05), and path length. MARL further improves performance by exposing the robots to additional coordination patterns during training.

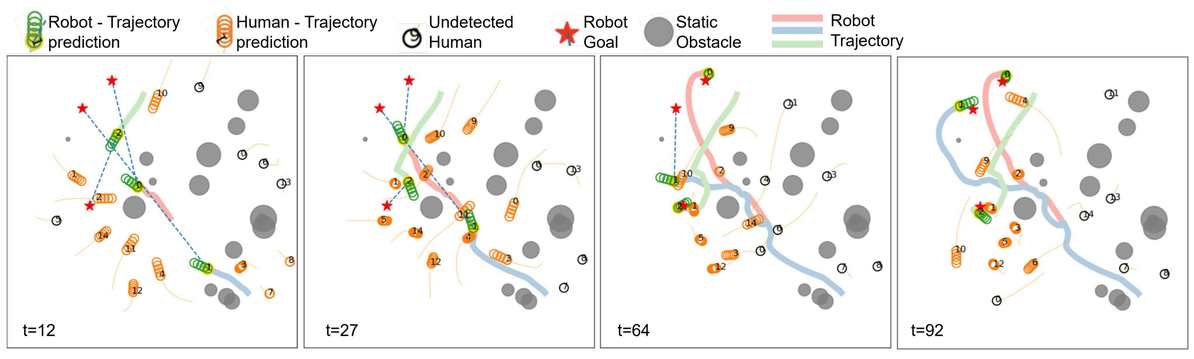


Table: Comparison of our approach with baselines in single and multi-robot scenarios. SF policy is used for humans.

Method	Success ↑	Collision ↓	Timeout ↓	Intrusion ↓ Ratio	Travel ↓ Time	Travel ↓ Length	Train R	Train H	Test R	Test H
LiDAR-Nav	0.60	0.22	0.17	19.64	13.77	14.50				
HEIGHT	0.64	**0.20**	0.16	**15.22**	13.51	13.88	1	15	1	15
HAMRON (ours)	**0.67**	0.28	**0.05**	19.78	**12.26**	**11.80**				
LiDAR-Nav	0.51	0.28	0.21	**13.06**	14.13	15.00	1	15		
HEIGHT	0.56	0.33	0.11	13.41	12.72	12.40	1	15	3	15
HAMRON (ours)	0.62	0.33	0.05	15.51	12.25	**11.93**	1	15		
HAMRON (ours)	**0.69**	**0.25**	**0.05**	15.75	**12.25**	12.03	3	15		

- Scalability: performance remains stable from 1→5 robots when trained and tested at the same team size.
- Generalization: Training with larger teams transfers well to smaller teams and remains robust when the number of humans increases at test time.
- Human-policy robustness: transfers across SF ↔ ORCA pedestrian models without retraining (minor shifts in failure modes).

Table: Ablation study of scalability, generalization, and robustness to human policies.

Method		Success ↑	Collision ↓	Timeout ↓	Intrusion ↓ Ratio	Travel ↓ Time	Travel ↓ Length	Train R	Train H	Train Human Policy	Test R	Test H	Test Human Policy
Scalability	(a)	0.67	0.28	0.05	19.78	12.26	11.80	1	15	SF	1	15	SF
	(b)	0.69	0.27	0.04	16.46	12.12	11.58	3	15	SF	3	15	SF
	(c)	0.69	0.25	0.05	15.75	12.25	12.03	4	15	SF	4	15	SF
	(d)	0.68	0.29	0.03	15.63	12.19	11.57	5	15	SF	5	15	SF
Generalization	(e)	0.58	0.36	0.06	14.12	12.43	11.95	1	15	SF	5	15	SF
	(f)	0.62	0.34	0.04	15.58	12.08	11.38	3	15	SF	5	15	SF
	(g)	0.78	0.19	0.03	20.52	12.33	11.67	5	15	SF	1	15	SF
	(h)	0.73	0.24	0.03	16.98	12.24	11.76	5	15	SF	3	15	SF
	(i)	0.83	0.14	0.03	15.01	12.59	12.21	1	15	SF	1	8	SF
	(j)	0.82	0.17	0.01	16.04	11.89	11.50	1	8	SF	1	8	SF
	(k)	0.67	0.31	0.02	22.86	12.33	11.22	1	8	SF	1	15	SF
Human policy robustness	(l)	0.64	0.32	0.04	16.58	11.89	12.27	3	15	SF	3	15	ORCA
	(m)	0.64	0.28	0.08	13.35	12.58	12.51	3	15	ORCA	3	15	ORCA

[1] T. Soualhi et al., "Human-aware multi-robot navigation in constrained environments using multi-agent reinforcement learning", under review at IEEE ICRA, 2026.
[2] L. Shuijing, et al., "Intention aware robot crowd navigation with attention-based interaction graph", IEEE ICRA, 2023.
[3] L. Shuijing et al., "Height: Heterogeneous interaction graph transformer for robot navigation in crowded and constrained environments", arXiv preprint 2024.
[4] B. Han et al, "Mobile robot navigation based on deep reinforcement learning with 2d-lidar sensor using stochastic approach", IEEE ISR, 2021.

Perspectives

- Stable MARL training: Move beyond DTDE by exploring CTDE (or hybrid) schemes with efficient, scalable critic architectures to mitigate non-stationarity.
- Social compliance: Integrate proxemics and social norms into the objective and policy (e.g., norm-aware rewards/constraints and evaluation metrics) to achieve more socially appropriate navigation.