Multi-Agent Deep Reinforcement Learning for UAV Flocking and Collision Avoidance in Challenging Environments

Mohammad Reza Rezaee[1], **Nor Asilah Wati Abdul Hamid**[1]

[1]Universiti Putra Malaysia
mo.rezaeezaee@gmail.com, asila@upm.edu.my

Abstract

Solve the issue of flocking based on multi-agent reinforcement learning for UAVs in complex, obstacle-ridden, and dynamic environments, where, much more importantly, the task of maintaining the cohesion of the formation under constrained perception and the challenge of collision avoidance are of concern. We present Recurrent Multi-Actor Attention Critic with Curriculum Learning (R-MAAC-CL), a Multi-Agent Reinforcement Learning (MARL) algorithm. Assessed in four test cases- sparse obstacles (Simple), dense clustered obstacles (Dense), moving hazards (Dynamic), and limited sensing (Partial)-R-MAAC-CL outperformed Benchmark algorithms on average reward, collision-rate, and flocking cohesion, without any significant drop in goal-success that indicates the safer and more coherent UAV swarm control in challenging situations.

Introduction

Uncrewed Aerial Vehicle (UAV) swarms are no longer just a pipedream but have moved into practice, and they can transform the sphere of environmental monitoring, precision agriculture, disaster response and logistics. The swarms make use of collective intelligence to carry out tasks beyond what individual UAVs can perform, like mass surveillance, joint payload delivery, and dynamic coverage. Nevertheless, dense UAVs flocking into the complex airspace will present significant issues in collision avoidance and swarm coordination. Since UAM is growing and expected market size is more than 15 billion Euros by 2050 (1), the threat of cross-agent collisions and obstacle collisions increases, requiring well-informed, effective real-time decision-making systems. Conventional control solutions, e.g., remote piloting and waypoint navigation based on GPS, are not adequate to solve the dynamism of the real world that may involve unpredictable obstructions and latency in communications, as well as emergent swarm behaviour. Therefore, multi-agent coordination, machine learning, and bio-inspired optimisation have become advanced autonomy frameworks that have become a vital solution that introduces safety, efficiency, and scalability in UAV flocking (2).

Managing swarms of unmanned aerial vehicles (UAVs) or Flocking, is a key issue in multi-agent robotics that can be used in search-and-rescue, surveillance and general environmental monitoring (3). Flocking implies that every UAV cooperates and stays in formation, evading collisions with dynamic and stationary moving obstacles as well as ensuring its or their accomplishment of given mission objectives with incomplete observability and communication (4). Classical-based control methods do not scale very well within dense-cluttered or very dynamic environments, which is the motivation behind embracing deep multi-agent reinforcement learning (MARL) approaches (5). Current MARL algorithms like MADDPG (6) and R-MADDPG (7), and MAAC (8) discuss to some extent the concepts of coordination, memory and credit assignment, although no methods have thus far leveraged the benefits of temporal recurrence, attention-based value estimation, and curriculum learning in conjunction with each other. In order to capture both agent- and group-level interactions, (9) introduces a hierarchical attention mechanism over recurrent policies. In comparison, we represent a unique integration R-MAAC-CL based on LSTM-based actors, a continuous in-scenario curriculum learning schedule, and MAAC's single-level attention critic, representing a unique integration.

Methodology

We train every UAV with a recurrent policy network to recall past observations and become confident even in situations where the view is limited (7). The multi-head attention helps the single and centralised critic determine which other agents and obstacles are of less or greater importance in estimating future rewards (8). In training, we gradually add complexity to 500 episodic tasks either by introducing new obstacles or by reducing individual drones' sensing ranges-this curriculum is used to first inculcate the swarm in easy coordination before having to solve challenging tasks (9). At each time step, each agent updates its LSTM state, selects an action with some exploration noise, and adds the experience to a replay buffer. After seeing sufficient data, we update the critic to reduce the prediction error and the actors using policy gradients, backpropagating through the LSTM, and updating target networks softly to stay stable.

We test on four prototypical flocking tasks: 10 fixed obstacles, complete sensing, simple. Dense: 30 cluttered obstacles, full sensing. Dynamic: 10 stagnant obstacles with moving obstacles, complete sensing. Partial: 10 immobile

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Algorithm 1: R-MAAC-CL Algorithm

```
Initialize actor, critic, target networks, and replay buffer
for each episode do
    Adjust environment difficulty via curriculum; reset env and LSTM states
    for each step do
        Observe: each agent collects its local state and neighbor/leader info
        Encode: update each agent's LSTM
        Act: sample continuous actions
        Execute all agents' actions
        Reward: compute per-agent rewards
        Store (obs, action, reward, next_obs, done) in replay buffer
        if buffer size $\geq$ batch size then
            Update attention critic
            Update recurrent actors
            Soft-update target networks
        end if
        if done then
            break
        end if
    end for
end for
return learned parameters
```

barriers, with a field of sensors of 50 m.

Algorithms. As depicted in Algorithm 1, training continues in episodes where we modify the environment's complexity according to the curriculum and reset the simulation and each agent's LSTM state. At every time step, agents update their hidden states, sample actions, execute them in the environment, and record the ensuing transitions. Once enough data has accumulated in the replay buffer, we perform a joint update: the centralised attention critic is trained to minimise temporal-difference error, the recurrent actor networks are updated via backpropagation through time, and the target networks are softly updated to stabilise learning.

Results and Discussion

We benchmark R-MAAC-CL against state-of-the-art multi-agent methods: MADDPG (6), R-MADDPG (7), MAAC (8), and HAMA (10). All algorithms are assessed on four UAV flocking scenarios—Simple, Dense, Dynamic, and Partial—differing in obstacle density, danger motion, and sensor range 1.

Figure 1 compares benchmark algorithms(MADDPG, R-MADDPG, MAAC, HAMA and R-MAAC) with the R-MAAC-CL Algorithm across three key performance metrics: average reward, average collisions, and average flocking rate. R-MAAC-CL consistently achieves the highest reward (least negative), the fewest collisions, and the strongest flocking cohesion. Across every circumstance, R-MAAC-CL delivers the most beneficial trade-off between performance and safety. In the Simple scenario, it attains the greatest average reward and minimises collisions by over 70% compared to MADDPG. Under Dense barriers, it maintains a 73% flocking rate(20% greater than MAAC) and lowers collisions by 31% over the next best solution. Even with shifting dangers (Dynamic), the recurring, curriculum-trained policy balances cohesiveness and safety better than HAMA or R-MADDPG. Finally, in partial observability, R-MAAC-CL achieves near-optimal reward and success rate, while halving collisions relative to non-attention baselines. These findings demonstrate that combining recurrence, attention, and curricular learning generates safer and more coherent UAV flocking policies.

Model	Avg Reward	Avg Collisions	Avg Flocking
R-MAAC-CL	-325.12	244.12	60.05
MAAC	-341.45	276.70	60.29
R-MADDPG	-450.77	428.94	39.12
MADDPG	-542.81	629.07	41.66
R-MAAC	-584.27	703.13	48.88
HAMA	-968.94	1377.51	44.78

Table 1: Comparison of average reward, collisions, and flocking across MARL models .

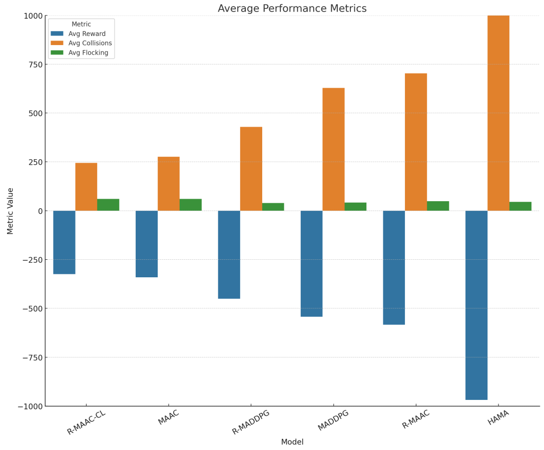

Figure 1: Comparison of algorithms—MADDPG, R-MADDPG, MAAC, HAMA ,RMAAC and R-MAAC-CL—across three key metrics: (a) average reward , (b) average collisions per episode, and (c) average flocking rate .

Conclusion

These findings demonstrate that multi-UAV flocking policies become much safer and more coherent when integrated with LSTM-based recurrence, a centralised attention critic, and in-scenario curriculum learning.

Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA2386-22-1-4018.Any options, finding, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

References

[1] Coppola, P.; De Fabiis, F.; and Silvestri, F. 2024. Urban Air Mobility (UAM): Airport shuttles or city-taxis?. *Transport Policy*, 150:24–34. doi:10.1016/j.tranpol.2024.03.003.

[2] Rezaee, M. R.; Hamid, N. A. W. A.; Hussin, M.; and Zukarnain, Z. A. 2024. Comprehensive Review of Drones Collision Avoidance Schemes: Challenges and Open Issues. *IEEE Transactions on Intelligent Transportation Systems* 25(7):6397–6426. doi:10.1109/TITS.2024.3375893

[3] Blais, M.-A.; and Akhloufi, M. A. 2023. Reinforcement learning for swarm robotics: An overview of applications, algorithms and simulators. *Cognitive Robotics*, 3:226–256. doi:10.1016/j.cogr.2023.07.004.

[4] Shen, G.; Lei, L.; Li, Z.; Cai, S.; Zhang, L.; Cao, P.; and Liu, X. 2022. Deep Reinforcement Learning for Flocking Motion of Multi-UAV Systems: Learn From a Digital Twin. *IEEE Internet of Things Journal*, 9(13):11141–11153. doi:10.1109/JIOT.2021.3127873.

[5] Ekechi, C. C.; Elfouly, T.; Alouani, A.; and Khattab, T. 2025. A Survey on UAV Control with Multi-Agent Reinforcement Learning. *Drones*, 9(7):484. doi:10.3390/drones9070484.

[6] Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; and Mordatch, I. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. https://arxiv.org/abs/1706.02275

[7] Wang, R. E., Everett, M., and How, J. P. 2020. R-MADDPG for Partially Observable Environments and Limited Communication. *arXiv preprint arXiv:2002.06684*. https://arxiv.org/abs/2002.06684

[8] Iqbal, S. and Sha, F. 2019. Actor-Attention-Critic for Multi-Agent Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pages 2961–2970.

[9] Yan, C., Wang, C., Xiang, X., Low, K. H., Wang, X., Xu, X., and Shen, L. 2024. Collision-Avoiding Flocking With Multiple Fixed-Wing UAVs in Obstacle-Cluttered Environments: A Task-Specific Curriculum-Based MADRL Approach. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):10894–10908. doi:10.1109/TNNLS.2023.3245124

[10] Ryu, H.; Shin, H.; and Park, J. 2020. Multi-Agent Actor-Critic with Hierarchical Graph Attention Network (HAMA). In *Proceedings of the AAAI Conference on Artificial Intelligence 34(05)*, 7236–7243.