

Formation Control of UAV-fleet With Orientation Alignment: A Multi-Agent Actor-Critic Based Approach

Abdulhakeem Abdulazeez*, Nicola Roberto Zema*, Tara Ali-Yahiya*, Steven Martin*

*Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91190 Gif-sur-Yvette, France
{abdulhakeem.abdulazeez, zema, tara.yahiya, steven.martin}@lisn.fr

Abstract—*Unmanned Aerial Vehicles (UAVs) have attracted much attention due to their application potential in many complex military and civilian operations, such as surveillance, event filming, and Search and Rescue (SAR) operations. Some of these operations often require the use of multiple UAVs to achieve higher efficiency while ensuring that specific mission requirements, such as resource management and maximization of covered areas, are met. In this work, we present a learning-based formation control protocol that pilots a networked fleet of UAVs configured in a leader-follower formation pattern to carry out a SAR mission. The formation consists of one leader, which is controlled by the Ground Control Station (GCS), and many followers that are expected to autonomously follow the leader while maintaining the original formation throughout the mission. The proposed protocol employs a Multi-Agent Reinforcement Learning (MARL) principle using only the Received Signal Strength Indicator (RSSI) derived during communication to autonomously control the follower UAVs to maintain formation while on a rotational movement. The goal is to ensure proper orientation alignment with the leader to avoid coverage overlap, and consequently increase efficiency. We performed several simulation experiments to evaluate the performance of the proposed protocol in terms of response time, and formation accuracy under different velocities.*

Index Terms—Wireless Networked Robots, UAV, formation control, Q-learning, MARL, MADDPG, fleet control.

I. INTRODUCTION

The application of UAVs also known as drones or quadcopters is increasing in several critical military and civilian operations due to the advances in robotics and aeronautical technologies. Most of these critical operations often involve searching, monitoring, and real-time exchanges of multimedia data over a large area. Studies have shown that using a single UAV for this kind of operation is less efficient due to the inherent constraints in size and payload capacity of a UAV [1]. In addition, a single UAV mission requires a drone with advanced capabilities that can reach high altitudes, and have high payload-carrying capacity [2]. More so, if the drone experiences a malfunction, the entire operation may be halted. The deployment of several UAVs (multi-UAVs) has been spurred by these issues in order to conduct complex operations more efficiently.

Controlling a fleet of UAVs in multi-UAV operations involves ensuring that the UAVs retain predetermined formation configurations during the mission and can also adjust to any changes in the mission or the environment [3]. Various frameworks and

methods have been proposed in the literature to address the many issues involved in designing protocols for multi-UAV operations. These encompass target identification [4] [5] agent autonomy [6], and localization and formation control [7] [8]. Localization and formation control are fundamental control principles necessary for multi-UAV cooperative tasks, and they have attracted significant research attention lately.

Controlling a fleet of UAVs requires a localization protocol built upon a reliable positioning technique. Several positioning techniques have been used for UAV localization in many applications, including *Global Positioning System (GPS)*, *Time-of-Arrival (ToA)*, and *Angle-of-Arrival (AoA)*. AoA and ToA are considered expensive to implement because they require hardware such as *Ultra-wideband (UWB)* to provide location or directional information [9]. GPS is regarded as ineffective indoors or in locations with many physical obstacles [10]. Therefore, fingerprinting techniques have been proposed as alternative methods that are more cost-effective and indoor-friendly [11]. The location fingerprinting techniques rely on the RSSI data derived by every UAV during wireless communications within the formation for localization and formation control.

RSSI, an indicator of signal strength [12], is integral to wireless communication systems and is now being repurposed to serve the key parameter of UAV formation control, offering a pragmatic solution for the next generation of UAV applications. Several proposals have been made using RSSI for range or distance estimation between communicating nodes, including the work in [8] that proposed a behavioral-based UAV control protocol that uses only RSSI. The protocol adapts the Q-learning principle to control a fleet of UAVs configured in a leader-follower formation pattern. Although the use of only RSSI makes the work less complex and cost-effective, its failure to address the problem formation orientation will cause coverage overlap and consequently lead to an increase in exploration time and resource wastage. Our previous work in [13] attempted to address the orientation problem in [8]. The work proposes a Q-learning-based direction detection and alignment technique that trains the follower UAVs to detect a change in the heading angle of the leader and determine the new direction of the leader, then turn together along the yaw-axis to re-align themselves to the new direction of flight of the leader.

The work considered a simplistic scenario by discretizing the environment space to enable the application of Q-learning. The followers are conditioned to move in eight possible directions only along the yaw axis, assuming they are in their original formation shape. A more realistic case would be to train the follower UAVs so that each one can autonomously determine the leader's change in direction along the yaw angle between 0° and 360° . Therefore, in this paper, we present a more realistic model as an improvement to the previous works. The new protocol employs the MADDPG principle, which is a variant of MARL capable of training multiple agents using shared knowledge to execute actions independently.

The rest of this paper is organized as follows: Section II presents a review of related works on RSSI-based formation control protocols. In Section III, we present the system model and problem definition, Section IV introduces our proposed MARL-based direction detection and orientation alignment protocol. Simulation and training of the proposed protocol are discussed in Section V, Section VI presents a discussion of the results obtained, and finally, Section VII concludes the paper.

II. REVIEW OF RELATED WORKS

The significance of localization and formation control techniques in any design or solution that involves several UAVs has motivated many studies in the literature. In this section, we present a detailed review of related RSSI-based formation control protocols for multi-UAV operations. The review highlights the objective and technique of the related works as follows:

A localization protocol for fixed-wing UAVs was proposed in the study by [14]. The protocol utilizes the RSSI to control the UAVs to track a target. Unlike the existing methods that require extensive information about broadcast power, antennas, and channel properties, the proposed protocol compares the power received by the antennas on a linear array and then uses the control protocol to determine the target's location. This method facilitates target localization or tracking by enhancing the ease and effectiveness of RSSI.

The authors in [15] present a geometrical multi-UAV localization strategy utilizing only RSSI to localize RF signal source. The protocol divides the UAVs into two groups. The groups are configured to form orthogonal circles in x-y and x-z planes. Each UAV in each group moves in a circular path around the center of the plane based on the RSSI difference until it achieves the same power as its members of the group. The location of the RF signal source is determined by calculating the intersection line formed by the UAV movements. The method offers a good advantage for wide area coverage, but its strict reliance on circular configuration limits its adaptability to other scenarios in dynamic environments.

The authors in [14] presented a geometric localization strategy utilizing the RSSI for a group of UAVs. The design divides the UAVs into two groups, with each group consisting two UAVs. Within each group, the UAVs follow circular paths in their respective planes to attain uniform signal strength, without requiring knowledge of transmission power or path loss exponent. The program employs two orthogonal circles

in the x-y and x-z planes, with stationary centers, and adjusts the UAV trajectories in real-time using geometric concepts to stay on these circles, guided by the variations in received signal intensity.

The study conducted in [16] investigates the application of intelligent robot swarms in search and rescue operations. The scheme employs a genetic algorithm to facilitate the self-organization of UAVs into a mesh network that is optimal for covering a specific area. Every UAV operates independently and relies on RSSI readings to evaluate the quality of connection with neighboring UAVs. When the RSSI measurement falls below a specified threshold, the UAVs adjust their positions to maintain connections. In order to ensure effective communication among the fleet, the movement of the UAVs is coordinated along the designated search path. However, if any UAV receives an incorrect RSSI value from its neighbor, it will cause all UAVs connected to the formation through that neighbor to break out of formation. This is because each UAV relies on information from its immediate neighbor only to determine its position.

The protocol proposed in [17] employs reinforcement learning to control a UAV to localize multiple terrestrial objects. First, the UAV conducts a scan of the entire area to determine the number of objects and their approximate locations. It then uses the RSSI data measured from multiple waypoints within its communication range using the multilateration method. The reinforcement learning algorithm is used to optimize the trajectory selection leading to the object, to improve localization accuracy while keeping energy consumption and time minimal. As the UAV collects more data, the trajectory selection is fine-tuned to improve accuracy. The learning model uses time, path length, signal strength, and energy consumption to train the UAV to autonomously determine the best trajectory that will lead the agent to the localization of multiple terrestrial objects. To achieve better accuracy and efficiency, the protocol also used

The approach presented in [17] employs the reinforcement learning principle to empower a UAV to autonomously maintain formation. The scheme's objective is to enhance the precision of localizing multiple terrestrial objects by utilizing data on time, path length, signal strength, and energy consumption. The framework entails conducting a thorough scan of the area in order to accurately determine the positions of objects. Subsequently, the RL agent undergoes training to utilize the location data in order to choose waypoints that minimize the mean location discrepancy.

The paper by [8] introduces a formation control technique that uses a Q-learning strategy to control multiple UAVs deployed for cooperative tasks. The UAVs are wirelessly networked and configured in a leader-follower control method. The followers are expected to be based on the learning model, autonomously follow the leader. The scheme's learning method relies on the RSSI values that are periodically exchanged between the connected UAVs to estimate their proximities to their neighbors within the formation. During the learning phase, a vehicle explores all potential routes (North, East, West,

and South) and calculates the Euclidean distances between its current position and the desired position. This information is used to determine the optimal strategy for reaching the intended position. The UAV subsequently follows the route with a minimal distance. The vehicle iterates through this procedure until it reaches its expected location in the formation. Using only RSSI values makes this strategy less complex and cost-effective compared to previous schemes that rely on many parameters. However, failure to consider formation orientation will cause the UAVs to form the required formation shape in the wrong orientation, which will lead to coverage overlap and consequently cause an increase in mission time and waste of resources.

The paper in [13] proposes a learning-based formation control protocol designed for autonomous UAV fleets deployed for *Search and Rescue* (SAR) missions. The protocol employs Q-learning principles to control the formation and alignment of UAVs arranged in a rectangular-shaped formation. Like in [8], the UAVs are configured in a leader-follower control structure. The leader is controlled by the *Ground Control Station* (GCS), while the followers use *Received Signal Strength Indicator* (RSSI) values for their autonomous localization and coordination. The study introduces two algorithms: the *Q-learning-based Speed Control Algorithm* (QSCA) and *Direction Detection and Formation Alignment* (DDFA). The QSCA algorithm controls the speed of the followers to adaptively catch up with the leader's change in speed. The DDFA, on the other hand, enables the followers to detect and re-align themselves with the direction of flight of the leader. Although the protocol ensures formation stability at a leader speed of up to 5 m/s, its applicability in more complex scenarios may be challenged by its limitation of moving only at an angular interval (yaw axis) of 45 degrees.

III. SYSTEM MODEL AND PROBLEM DEFINITION

In this work, we consider a fleet of UAVs deployed to search for a target in a life-threatening situation for quick rescue. The swarm is configured to comprise a single leader and multiple followers, forming a predefined formation pattern that must be maintained throughout the search mission. The number of follower UAVs and inter-UAV spacing are influenced by the scale and topology of the search area. The UAVs communicate in ad-hoc mode using IEEE 802.11 standard. The position of each UAV in the formation is determined by its proximity to its neighbors, which is defined by a set of RSSI values measured from their communications. In our context, the leader is controlled directly by the GCS to follow a parallel-sweep search pattern, which is considered a very efficient SAR mission [13], and the followers are expected to autonomously follow the leader while maintaining the original formation pattern. For simplicity, we adopt a formation of UAVs for a typical SAR mission, as shown in figure 1.

The figure depicts a triangular formation of a fleet of three UAVs deployed for a SAR mission to search for an unknown target in a wide and hazardous terrain. The formation comprises two followers positioned at the base vertices of an equilateral

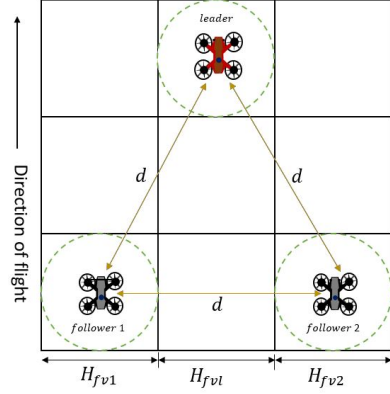


Fig. 1. Fleet of three UAVs in Correct Orientation

triangular structure, and the leader at the top, all moving at a fixed altitude of δ meters. The follower UAVs are equipped with downward-looking cameras to capture visual data. The visual coverage of the formation at any time t of the mission is determined by the sum of fields of view of the cameras mounted on the follower UAVs. The fv is the region an aerial device's camera covers. It is measured, according to the vertical and horizontal angle of views, along the vertical and horizontal axes of the covered regions. Of interest to our work is the horizontal field of view. The horizontal field of view H_{fv} of each UAV camera determines the width of the region it covers along the flight path, defined as in Eq. 1:

$$H_{fv} = 2\delta \tan\left(\frac{H_\theta}{2}\right) \quad (1)$$

In 1 δ is the UAV altitude, H_θ is the horizontal angle of view of the camera, defined as Eq. 2:

$$H_\theta = 2 \tan^{-1}\left(\frac{\alpha_h}{2\beta}\right) \quad (2)$$

In 2 α_h is the width of the camera lens and β is the focal length of the camera.

Therefore, the total horizontal region covered by the formation at any time t is defined as:

$$H_c(t) = H_{fv1} + H_{fv2} + H_{fvl} \quad (3)$$

where $H_c(t)$ is the total width of the horizontal region covered by the formation, H_{fvl} is the horizontal field of view of the leader, while $H_{fv,1}$ $H_{fv,2}$ are the horizontal field of view of follower 1 and 2 respectively.

From Figure 1, we can observe that, if the followers maintain the original formation for each run of the search, a region with a width of $H_{fv,i}$ meters will be searched, and the number of runs will depend on the width of the search area.

A. Problem Definition

As mentioned in the previous section, this work aims to address two challenges of multi-UAV formation configured for the leader-followers control model, namely; direction detection

with re-alignment and formation reconstruction. We define the two problems as follows: Firstly, let's define the position of the leader UAV in the formation; $\rho_l = (x_l, y_l, \alpha)$, and that of the followers as:

$$\rho_1 = \left(x_l + \frac{s}{2}, y_l + \frac{s\sqrt{3}}{2}, \delta \right) \quad (4)$$

$$\rho_2 = \left(x_l - \frac{s}{2}, y_l + \frac{s\sqrt{3}}{2}, \delta \right) \quad (5)$$

$$\rho_3 = \left(x_l, y_l - s\sqrt{3}, a \right) \quad (6)$$

The fleet is expected to maintain the original formation throughout the mission at a fixed altitude a . To achieve this, all UAVs in the fleet must maintain the same heading angle β . Therefore, at every time t of the mission, the position of the leader and that of the followers heading in the same direction is defined as:

$$x_j(t) = x_{j0} + vt \cos(\beta) \quad (7)$$

$$y_j(t) = y_{j0} + vt \sin(\beta) \quad (8)$$

$$z_j(t) = \delta \quad (9)$$

where $\rho_j = (x_j, y_j, z_j)$ is the position of UAV $_j$ in the formation. However, a problem occurs when the leader changes its heading angle β by rotating either left or rightward along the yaw-axis. For example, let define the rotation matrix $R_t(\beta)$ for the yaw as:

$$R_t(\beta) = \begin{bmatrix} \cos(\beta) & -\sin(\beta) & 0 \\ \sin(\beta) & \cos(\beta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (10)$$

Once the leader changes its direction along the yaw axis, we determine the new position of the followers relative to that of the leader by applying equation 10 to the initial positions of the followers, as follows:

$$\rho'_1 = \begin{bmatrix} x_l + \frac{s}{2} \cos(\beta) - \frac{s\sqrt{3}}{2} \sin(\beta) \\ y_l + \frac{s}{2} \sin(\beta) + \frac{s\sqrt{3}}{2} \cos(\beta) \\ \delta \end{bmatrix} \quad (11)$$

$$\rho'_2 = \begin{bmatrix} x_l - \frac{s}{2} \cos(\beta) - \frac{s\sqrt{3}}{2} \sin(\beta) \\ y_l - \frac{s}{2} \sin(\beta) + \frac{s\sqrt{3}}{2} \cos(\beta) \\ \delta \end{bmatrix} \quad (12)$$

$$\rho'_3 = \begin{bmatrix} x_l - s\sqrt{3} \sin(\beta) \\ y_l - s\sqrt{3} \cos(\beta) \\ \delta \end{bmatrix} \quad (13)$$

For simplicity, consider Figure 1. The figure shows three UAVs in an equilateral triangular shape formation configuration deployed to explore a search area of width W_a . The total horizontal field of view of the search area is the sum of the horizontal field of view of the three UAVs as defined in Equation 3.

If the fleet moves in the flight direction while keeping the formation shape, searching the whole area in one pass is possible. However, because the follower UAVs determine whether their position in the formation is correct by comparing the current state of the formation to the original configuration, as long as the proximity between any pair of UAV information is correct, the follower UAVs will consider the formation to be accurate regardless of the orientation. If the fleet is in the wrong orientation, an example of which is shown in Figure 2, the followers explore the same region of the search area, leaving the shaded region unexplored. This will cause an increase in cost and search time, and may consequently cause more danger to the search targets in a life-threatening situation.

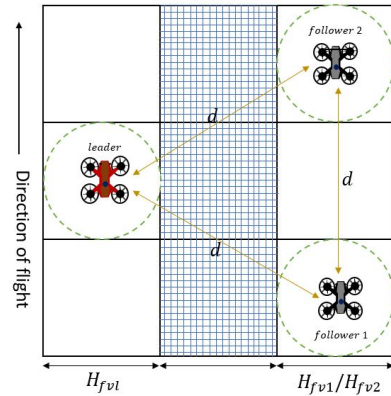


Fig. 2. Fleet of three UAVs in wrong Orientation

Addressing this problem requires an efficient control protocol capable of handling the complexity and dynamics of the environment. To make each follower UAV autonomous, it has to be capable of taking action in isolation from other UAVs in the formation based on its instantaneous perception of the environment. The challenge would be, that whenever there is a deviation from the original formation, each follower UAV perceives the changes and immediately initiates actions to return to the goal formation, which will cause a serious error propagation problem because the correctness of any action taken by any follower UAV depends on the action taken by all other UAV. The wider the action scope for the UAVs, the more complex and lower the probability of recovery from the formation deviation. The large dimensionality and complexity of this problem [18] [19] make it difficult to address this problem using conventional optimization methods [20] [21]. We employed a variant of MARL called *Multi-Agent Deep Deterministic Policy Gradient* (MADDPG). MADDPG is an extension of *Deep Deterministic Policy Gradient* (DDPG), designed to address the complexities that arise when multiple agents interact within the same environment [22]. Its advantage of partial observability, continuous control, and synchronization, especially for agents in cooperative tasks, makes it suitable for addressing the orientation alignment problem defined for this work.

IV. PROPOSED PROTOCOL

This section introduces our proposed orientation-aware multi-UAV formation control protocol based on the MADDPG algorithm. As discussed in previous sections, our objective is to optimize the control of the fleet of four UAVs using RSSI values, even in the presence of environmental challenges such as signal attenuation and reflection. The fleet consists of one leader positioned at the center of an equilateral-shaped formation, the followers occupying the vertices of the triangle. We model the multi-UAV networked task as a cooperative game. To achieve collaborative training and autonomous execution, we leverage the actor-critic framework of MADDPG depicted in Figure 3.

A. Definition of MADDPG Components

System State: In the context of this work, the definition of state S for each UAV will be based on the RSSI value, being the key parameter that determines its position in the formation at any time t of the mission. The state S represents all possible variations from the original formation configuration. Therefore, we define the current state $s(t) \in S$ of the environment as perceived by the UAV as:

$$s_i(t) = [\mathcal{R}_{ij}(t), |\Delta\mathcal{R}_{ij}(t)|, D_i(t)] \quad (14)$$

- $\mathcal{R}_{ij}(t)$ is the current RSSI values between UAV_i and their neighbors j . These values determine the position of the UAV relative to its neighbors at the current time step.
- $|\Delta\mathcal{R}_{ij}(t)| = |R_{ij}(t) - R_{ij}(t-1)|$ is the change of the RSSI values between the previous time step and the current time step. This will help the UAV understand the magnitude of the deviation from the goal formation
- $D_i(t)$ is the Euclidean distance between the UAV_i and its goal position.

Action: The action space A is a set of possible actions a follower UAV can take at any time t of the mission based on a policy π . Whenever there is a change in the leader's heading, the followers have to rotate along the yaw axis to catch up with the leader's new direction of flight. Therefore, the action a_i taken by UAV_i is defined as:

$$a_i(t) = [\Delta\theta_i, \Delta\omega_i] \quad (15)$$

- $\Delta\theta_i$ is the desired change in the yaw angle for UAV_i . The angle's magnitude represents the sharpness of the curve. If the angle is positive, the UAV moves rightward, and leftward if is negative.
- $\Delta\omega_i$ is the angular velocity that determines how fast or slow the UAV should adjust its yaw angle.

Reward: We developed a reward function that incentivizes the follower UAVs to change their yaw angle and angular velocity depending on the Euclidean distance from their optimal placements ensuring the UAVs learn to maintain formation effectively while minimizing departure from the goal state. The reward function must support actions that lower the Euclidean distance to the goal state while properly adjusting to the

changing signal strengths since the main goal is for the UAVs to adapt to changes in their relative positions using RSSI values and to realign themselves accordingly. We thus define the reward as:

$$r_i(t) = \begin{cases} +1, & \text{if } D_i(t) < \epsilon_1, \\ -1, & \text{if } D_i(t) \geq \epsilon_1, \end{cases} \quad (16)$$

This part of the reward function evaluates how close the UAV is to its goal position at every time t using the Euclidean distance $D_i(t)$. If the distance is small (i.e., $D_i(t) < \epsilon_1$), the UAV is rewarded. If the distance is large (i.e., $D_i(t) \geq \epsilon_1$), the UAV is penalized.

$$r_i(t)_+ = \begin{cases} +1, & \text{if } |\Delta\mathcal{R}_{ij}(t)| \text{ reduces } D_i(t), \\ -2, & \text{if } |\Delta\mathcal{R}_{ij}(t)| \text{ increases } D_i(t), \end{cases} \quad (17)$$

The second part evaluates how the UAV's current movement based on the change in RSSI ($\Delta\mathcal{R}_{ij}(t)$) affects its distance from the goal state. If the change in RSSI results in reducing the Euclidean distance $D_i(t)$, the UAV is rewarded. If the change increases the distance, the UAV is penalized more severely.

B. Model definition

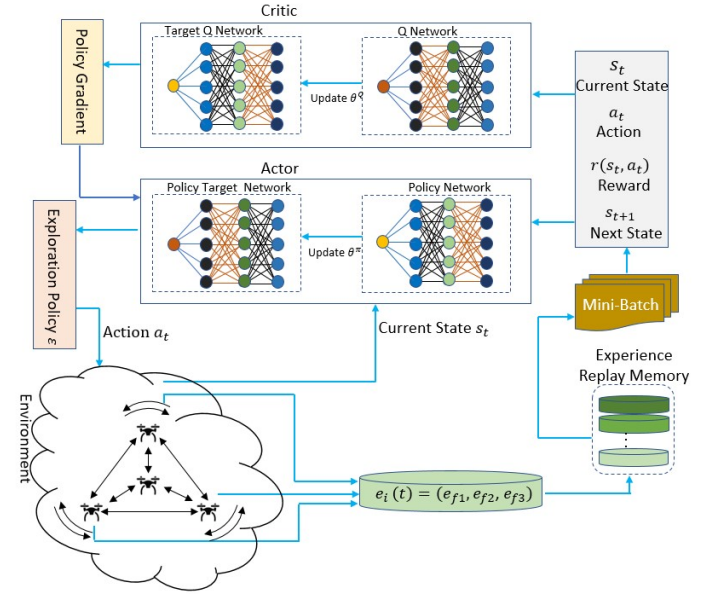


Fig. 3. MADDPG Orientation Alignment Framework

Each follower UAV in the formation is a MADDPG agent equipped with a policy network μ_i and a critic network Q_i . The policy network $\mu_i(s_i|\theta_{\mu_i})$ enables the agent to select the best action based on its perceived state s_i , while the critic network $Q_i(s_i, a_i|\theta_{Q_i})$ is responsible for evaluating the quality of the action selected by the UAV by estimating the expected return. Target networks μ'_i and Q'_i for policy and critic networks respectively are also created and initialized to provide stability to the learning process. Initially, the target networks carry the

same weights as the primary networks. Subsequently, to avoid moving targets, the weights of the target networks are updated periodically using the weights of the primary networks.

The training starts with the initialization of a replay buffer \mathcal{D} which stores the experiences of the UAVs in the form of state transitions and is subsequently sampled to update the networks. All UAVs are set to their initial positions, and their corresponding RSSI values $\mathcal{R}_{ij}(0)$ are computed, hence proving the follower UAVs with the initial knowledge about their relative positions. At every time step of every episode, each follower UAV observes its current state, as the RSSI values, change in RSSI, and Euclidean distance to the goal state $D_i(t)$. Based on this observed state, the policy network μ_i creates an action $a_i(t)$, which may be changed by exploration noise \mathcal{N}_t to inspire exploration of the environment. The UAV then carries out this action, changing both its position and the related RSSI values. The reward $r_i(t)$ for every UAV is computed depending on the degree of maintenance of the intended formation and distance to the goal state.

The experiences $(s_i(t), a_i(t), r_i(t), s_i(t+1))$ are stored in the replay buffer for later use. The critic network Q_i is then updated by sampling a mini-batch of experiences from the buffer and computing the target value $y_i(t)$, which incorporates the total reward and the estimated value of the next state-action pair from the target networks. The policy network μ_i is subsequently updated using the policy gradient method, which optimizes the actions to maximize the expected return as estimated by the critic. The target networks μ'_i and Q'_i are then softly updated by gradually incorporating the weights of the primary networks.

This iterative training process continues across multiple episodes, allowing the UAVs to learn cooperative strategies for maintaining the formation. Upon completion of the training, the final policy networks μ_i are deployed for decentralized execution, enabling each UAV to independently adjust its position based on local observations while collaboratively maintaining the overall formation.

V. SIMULATION

We use the NS3 discrete event simulator to configure the fleet formation and simulate the network communication. Also, the decision engine of the UAVs is implemented at the application level of the NS3 nodes. Additionally, we integrated an instance of mpack [23] into the engine. The remaining components of the protocol stack are taken straight from NS3, while the modifications in linear and angular velocities are directly implemented in the NS3 node mobility modules.

We employ a four-UAV fleet topology based on a leader-follower arrangement as described in earlier sections. Three UAVs operate as followers in the formation, creating an equilateral triangle around the leader, which is centrally located to create an equidistance to each follower. While the followers are expected to keep the original formation by following the leader throughout the mission, the Ground Control Station (GCS) directly controls the leader. The fleet uses IEEE 802.11ax in ad-hoc mode on a 5 GHz frequency range. We use the On-Demand Link-Disjoint Routing (OLDR) protocol for routing,

which runs on-demand [24], and fit for mobile ad-hoc networks [25], especially in periodic data communication scenarios [26]. OLDR creates several discontinuous pathways between nodes such that the failure of one node does not affect the whole network [24], so minimizing the spread of proximity errors in the fleet. The position of every follower UAV in the formation is determined by its proximity to other UAVs. Every follower UAV derives RSSI values from its neighbors at every periodic interval throughout the mission. The current state of the follower UAV within the formation is defined by these RSSI values, changes in the RSSI values, and formation variation. These state representations are used by the proposed model to train the UAVs. Our proposed MADDPG model generates the optimal policy for each follower UAV to maintain the desired formation.

Two main networks make up our MADDPG model: the critic and the actor. Every network combines one input layer, two hidden layers, and one output layer. Across the layers, we leverage the Rectified Linear Unit (ReLU) activation function to support non-linear transformations and improve the learning process. ReLU is especially good at reducing the vanishing gradient issue, therefore accelerating the learning process. We use Adam optimizer to maximize actor and critic networks' weights. Adam is selected for its adjustable learning rate capacity, which offers more effective convergence than conventional gradient descent techniques. Every 100 training iterations, the target network weights are changed to preserve a steady learning process and stop sharp changes that can compromise training. We set the learning rate at 0.0001, and a discount factor to 0.999, which reflects a significant focus on future rewards. This is essential to keeping UAV speeds within the formation throughout long missions. We use an epsilon-greedy policy to strike a compromise between exploration and exploitation; the epsilon value starts at 0.9 and falls progressively to 0.1. This method lets the model start with intensive research and progressively move to using the acquired techniques as training advances. This guarantees early on complete investigation of the action space and fine-tuning of the policy as the model converges on efficient methods.

The training episode is set at 2000 to give the model enough room to grow and adjust to the complexity of UAV formation control. We also set the replay buffer which is essential for experience replay and breaking observation sequence correlations to a capacity of 2000. This guarantees that learning can benefit from a wide spectrum of experiences saved and used. From this buffer, mini-batches of size 64 are taken for training, therefore balancing computational economy with the benefit of learning from many encounters.

VI. RESULTS AND DISCUSSIONS

This section presents the results obtained from a series of testing experiments conducted to evaluate the effectiveness of our proposed protocol in terms of fleet response time and formation stability under varying speeds, movement patterns, and formation sizes.

Figure 4 illustrates the response time of the UAV fleet as the leader changes its yaw angle across different speeds, with a fixed formation size of 30 meters. The data shows a clear trend where response time increases with both yaw angle and speed. At smaller yaw angles, such as 15° , the response times are relatively low, ranging from 0.68 seconds at 1.5 m/s to 0.75 seconds at 6 m/s. However, as the yaw angle increases to 75° , the response times become more pronounced, reaching up to 1.20 seconds at the highest speed of 6 m/s. This suggests that sharper turns require more time for the follower UAVs to adjust and re-establish the desired formation. Additionally, the impact of speed on response time becomes more significant at larger yaw angles, indicating that higher speeds contribute to longer response times, particularly when the leader makes more abrupt directional changes.

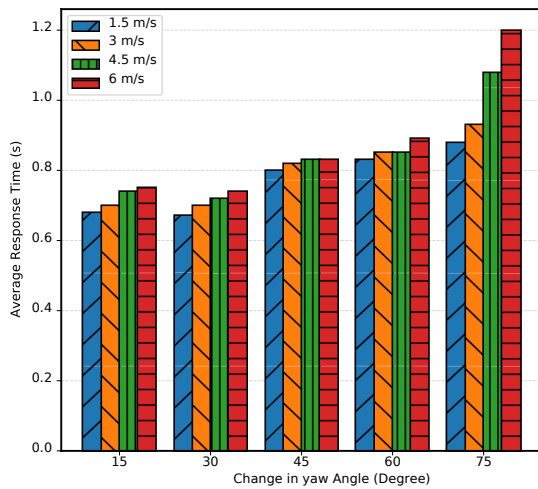


Fig. 4. Response Time of Followers To Changes In Leader's Direction

In the context of this work, first, we measure the response time after each yaw angle change by the leader to determine how quickly the follower UAVs can realign back to the original formation shape. Formation stability is obtained by computing the formation as the Euclidean distance between the goal formation state, and the current formation state. This error is measured in meters, representing how well the UAVs are maintaining the desired formation. The smaller the Euclidean distance, the closer the UAVs are to the ideal formation, indicating better formation stability. Figures 5 and 6 show the stability of the UAV fleet formation moving in clockwise and anti-clockwise directions at different speeds along the yaw axis, with a fixed formation distance of 30 meters. The figures reveal that the formation is consistently maintained for smaller yaw angles, particularly at 15° and 30° . For example, the stability of around 0.85 to 0.87 and 0.92 to 0.95 is recorded for 1.5 m/s and 3 m/s, respectively. This indicates that the followers could detect and re-align themselves with a formation error of less than 1 meter at a speed of up to 3 m/s. However, for a yaw angle increase of 45° and above, it can be seen that an increase in speed has a greater impact on formation

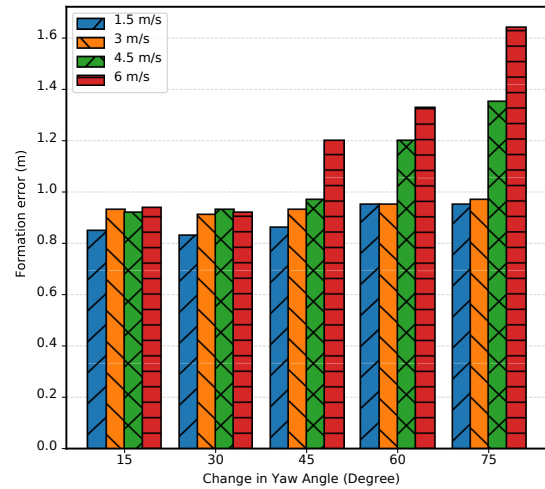


Fig. 5. Formation Stability for clockwise Yaw Movement At Different Speed

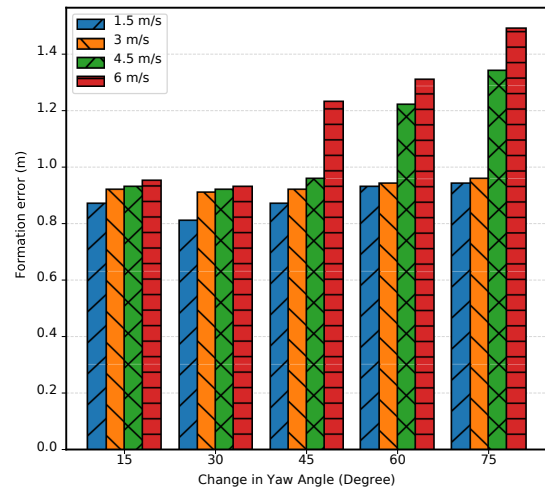


Fig. 6. Formation Stability for Anti-clockwise Yaw Movement At Different Speed

stability. For example, with a speed of 6 m/s and a yaw change of 75° ; values for clockwise and anti-clockwise maneuvers reach 1.64 and 1.49, respectively. The results demonstrate that our proposed model is able to maintain fleet formation while moving in both directions.

Figure 7 illustrates the stability of different formation sizes, ranging from 30 meters to 120 meters for different speed as the leader UAV changes its yaw angle. At smaller formation sizes of 30 meters and 60 meters, the UAVs maintain relatively consistent formation stability across various yaw angles, with formation errors ranging from 0.51 to 0.64 at 30 meters and 0.61 to 0.93 at 60 meters respectively. This indicates that the follower UAVs can more effectively realign to the new leader's direction of flight while maintaining their goal position when they are closer together. However, as the formation size increases to 90 meters and 120 meters, the formation stability

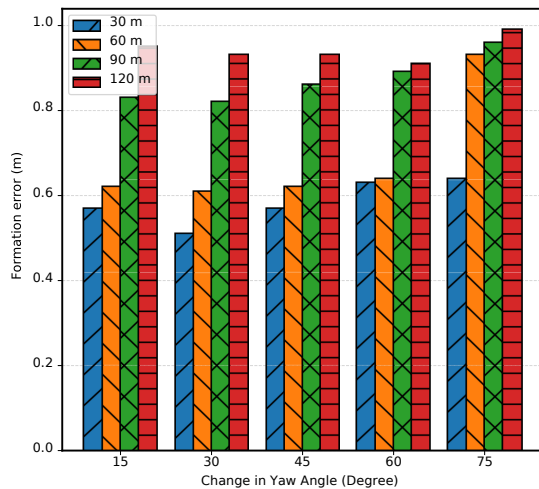


Fig. 7. Formation Stability for Varying Formation Sizes Moving At Different Speed

shows more variability, particularly at higher yaw angles. For example, at 120 meters, the formation error increases to 0.99 at a 75° yaw angle, indicating greater difficulty in maintaining formation stability as both the formation size and yaw angle increase. This suggests that larger formations are more prone to deviations as the yaw angle increases, due to the increased time it takes for the UAVs to detect and react to the leader's directional changes.

VII. CONCLUSION

In this work, we introduced a novel MADDPG-based formation control protocol for UAV fleets, to preserve formations and fast responsiveness during difficult maneuvers. We showed by simulations the efficiency of our protocol in managing different speeds, yaw angles, and formation sizes. The results show that, particularly at moderate yaw angles and speeds, the protocol effectively preserves formation stability; whereas, the response times stay within reasonable bounds even under more difficult circumstances. The protocol also demonstrates strong performance in both clockwise and anti-clockwise motions, therefore displaying its flexibility and adaptability in several contexts. On higher speeds and more yaw angles, however, we found that formation stability starts to deteriorate, implying that more improvement is required.

REFERENCES

- [1] J. ZHANG and J. XING, "Cooperative task assignment of multi-uav system," *Chinese Journal of Aeronautics*, vol. 33, no. 11, p. 2825–2827, 2020.
- [2] Z. Cai, H. Zhou, J. Zhao, K. Wu, and Y. Wang, "Formation control of multiple unmanned aerial vehicles by event-triggered distributed model predictive control," *IEEE Access*, vol. 6, p. 55614–55627, 2018.
- [3] Y. Wu, J. Gou, X. Hu, and Y. Huang, "A new consensus theory-based method for formation control and obstacle avoidance of uavs," *Aerospace Science and Technology*, vol. 107, p. 106332, 2020.
- [4] M. R. Brust, M. Zurad, L. Hentges, L. Gomes, G. Danoy, and P. Bouvry, "Target tracking optimization of uav swarms based on dual-pheromone clustering," *2017 3rd IEEE International Conference on Cybernetics (CYBCONF)*, 2017.

- [5] R. Wise and R. Rysdyk, "Uav coordination for autonomous target tracking," *AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2006.
- [6] S. Drake, K. Brown, J. Fazackerley, and A. Finn, "Autonomous control of multiple UAVs for the passive location of radars," *2005 International Conference on Intelligent Sensors, Sensor Networks, and Information Processing*, 2005.
- [7] J. Zhang, J. Yan, and P. Zhang, "Multi-uav formation control based on a novel back-stepping approach," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 3, p. 2437–2448, 2020.
- [8] N. R. Zema, D. Quadri, S. Martin, and O. Shrit, "Formation control of a mono-operated uav fleet through ad-hoc communications: A q-learning approach," *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2019.
- [9] K. Kaemarungsi and P. Krishnamurthy, "Modeling of indoor positioning systems based on location fingerprinting," *IEEE INFOCOM 2004*, 2004.
- [10] G. Djuknic and R. Richton, "Geolocation and assisted gps," *Computer*, vol. 34, no. 3, p. 123–125, 2001.
- [11] N. Bahl and V. Padmanabhan, "Radar: An in-building rf-based user location and tracking system," *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064)*, 2000.
- [12] X. Ouyang, F. Zeng, D. Lv, T. Dong, and H. Wang, "Cooperative navigation of uavs in gnss-denied area with colored rssi measurements," *IEEE Sensors Journal*, vol. 21, no. 2, p. 2194–2210, 2021.
- [13] A. Abdulazeez, N. R. Zema, T. Ali-Yahiya, and S. Martin, "Learning-based formation control of uav-fleet," in *2023 19th International Conference on Network and Service Management (CNSM)*, 2023, pp. 1–7.
- [14] N. Guzey, H. M. Guzey, and A. Ronzhin, "Consensus-based localization by using array of antennas on a fixed-wing uav," *2019 27th Telecommunications Forum (TELFOR)*, 2019.
- [15] N. Güzey, "Rf source localization using multiple uavs through a novel geometrical rssi approach," *Drones*, vol. 6, no. 12, p. 417, 2022.
- [16] L. Ruetten, P. A. Regis, D. Feil-Seifer, and S. Sengupta, "Area-optimized uav swarm network for search and rescue operations," *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, 2020.
- [17] D. Ebrahimi, S. Sharafeddine, P.-H. Ho, and C. Assi, "Autonomous uav trajectory for localizing ground objects: A reinforcement learning approach," *IEEE Transactions on Mobile Computing*, vol. 20, no. 4, p. 1312–1324, 2021.
- [18] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.
- [19] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev *et al.*, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.
- [20] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.
- [21] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fiedelnd, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [22] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in neural information processing systems*, 2017, pp. 6379–6390.
- [23] R. R. Curtin, J. R. Cline, N. P. Slagle, W. B. March, P. Ram, N. A. Mehta, and A. G. Gray, "Mlpack: A scalable c++ machine learning library," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 801–805, 2013.
- [24] M. K. Marina and S. R. Das, "On-demand multipath distance vector routing in ad hoc networks," in *Proceedings ninth international conference on network protocols. ICNP 2001*. IEEE, 2001, pp. 14–23.
- [25] M. M. Artimy, W. Robertson, and W. J. Phillips, "Algorithms and protocols for wireless and mobile ad hoc networks," *ed: John Wiley & Sons*, 2009.
- [26] A. Tsirigos and Z. J. Haas, "Analysis of multipath routing, part 2: mitigation of the effects of frequently changing network topologies," *IEEE Transactions on Wireless Communications*, vol. 3, no. 2, pp. 500–511, 2004.