Extended Abstract

Motivation Advances in small UAVs and computer vision have enabled the possibility of autonomous swarms being used in a variety of civil and defense applications. Many relatively inexpensive systems can be deployed en masse to perform search operations over a large area, and, using on-board sensors and processing techniques, identify and relay objects of interest. Because of the large number of fielded systems manned control of each one is infeasible. In scenarios where this capability might be employed, a centralized control node is not necessarily possible. The terrain may be such that line of sight communication is obstructed, such as in a mountain search and rescue operation, or there may be external factors that limit the range of communications. In these cases where a centralized control node is not available, the ability of individual systems to pass information to each other and update their behavior ad hoc becomes important for achieving coordinated behavior.

Method A simple two-dimensional grid environment was developed to capture the effect discrete communication horizons and spatially distinct target and data fusion locations. Each agent used various sensors and communication with friendly agents to construct observation vectors that were then transformed into a stigmergic potential field (SPF). This field would encode information allow the drone swarm to coordinate without expressly passing any information about their actions.

Implementation Independent Proximal Policy Optimization (IPPO) was used as it was hoped that the additional coordination from the SPF would overcome the algorithms inherent struggle to produce highly coordinated behavior. Both the actor and critic used CNN's in order to take advantage of the spacial information in the observations. Three different cases were run with varying numbers of agents, one with no SPF, one with the SPF included in the observation and one with it included with the observation and integrated into the reward function. Various agents counts were tested to establish if that effected the results.

Results For a small number of agents the SPF did appear to provide a benefit for both the learning rate and the quality of the final agent. However, for larger agents counts this effect diminished. In both cases with SPF and without, the agents learned some coordinated behavior, including forming chains to move target locations quickly to the data fusion locations.

Discussion The lack of positive impact from the SPF could be caused by a number of reasons. The environment may have been too simple, mitigating the effect from any coordinating force from the SPF. The agents may have learned similar features contained within the SPF quickly. Alternatively, the SPF construction may have been imprecise. The values and sized of the kernels used to create it mismatched to what would be beneficial to the agents. This could explain the apparent detrimental effect seen in the four agent case.

Conclusion The inclusion of the stigmergic potential field did not meaningfully impact the training or coordination of the agents. Agents were still able to learn unique behavior in the environment but did so independent of whether or not a SPF was supplied.

Effect of Stigmergic Potential Fields on Drone Swarm Coordination

Kevin Porter
Department of Computer Science
Stanford University
porterk@stanford.edu

Abstract

This work explores the effect of stigmergy on the efficacy of reinforcement learning (RL). Each agent controls a drone platform, the platforms host a series of senses and data that are used to build the observation vectors that are then transformed into a potential field (SPF). The SPF was generated by applying various transformation the observation arrays. The RL was used for coordinating agents to find and track target objects. Experiments conducted were focused on evaluating the efficiencies of the inclusion of the additional SPF's in the speed and quality of the training. For a small number of agents the SPF did appear to provide a benefit for both the learning rate and the quality of their final agent, however these gains were eliminated for more agents.

1 Introduction

Advances in small UAVs and computer vision have enabled the possibility of autonomous swarms being used in a variety of civil and defense applications. Many relatively inexpensive systems can be deployed en masse to perform search operations over a large area, and, using on-board sensors and processing techniques, identify and relay objects of interest. Because of the large number of fielded systems manned control of each one is infeasible.

In scenarios where this capability might be employed, a centralized control node is not necessarily possible. The terrain may be such that line of sight communication is obstructed, such as in a mountain search and rescue operation, or there may be external factors that limit the range of communications. In these cases where a centralized control node is not available, the ability of individual systems to pass information to each other and update their behavior ad hoc becomes important for achieving coordinated behavior.

Figure 1: Example of real world deployments of drone swarms.Meese and Means ([n. d.])

2 Related Work

Both stigmergic concepts and drone swarming have been explored in reinforcement learning literature. A review of Multi-Agent Reinforcement Learning (MARL) examines at least six papers investigating control of UAVs.Zhang et al. (2021) One of particular note discusses a sensor coverage problem similar to what is being proposed here using multi-agent Q-learning. However, in this paper, Q-learning is performed over the combined action space of all agents, which would then take individual actions based on rules. This requires total state observability, which is not present in our scenario. Pham et al. (2021)

An application of RL techniques for a drone search and rescue application has been explored Laffranchi Falcão et al. (2024). This tool represented the search area as a two dimensional grid with several drones attempting to locate a moving target object. Significant inspiration was taken from this project, such as the two dimensional grid representation of the environment and the integration with the petting zoo library for simulation control. However, the current work differs from that of the DSSE project by introducing explicit communication protocols and additional observational states. The addition of the discrete communication adds significantly more complexity as each agent has a very different view of the state of the environment. This additional complexity led the the desire to supplement the RL techniques with stigmergy.

This work will seek to explore the effect of stigmergy on the efficacy of an RL algorithm. The stigmergic effect will be implemented by means of stigmergic potential fields (SPF). "Stigmergic potential fields, in the most basic terms, are a system of uphill and downhill potentials that guide the UAVs to their destination." Rishe (2011)

Applications of stigmergic concepts within other MARL contexts have been explored several times.Nguyen (2021); Aras et al. (2005); Xu et al. (2021) The paper by Nguyen (2021), which investigated augmenting RL-policies the a stigmergic algorithm for coordinating agents to move a target object, saw an improved in scalability but at the cost of more unpredictable behavior. Xu et al. (2021) used stigmergic properties as a method of indirect communication for conflict avoidance between agents.

The use of stigmergy as a method to coordinate mulit-agent behavior has also been proposed for drone swarm control applications. Jr. et al. (2006) In a paper by Bamberger et al., the authors seek to develop a cooperative swarming capability by means of stigmergic potential fields to coordinate actions. Each agent uses both its perception of the environment and communications with other agents to form beliefs about the true state, which are then converted into a stigmergic potential field that is used to coordinate agent actions. However, this paper did not apply any reinforcement learning concepts to their autonomy stack.

By combining the concept of stigmergic potential field and reinforcement learning, it is believed that agent coordination can be accelerated without requiring non-physical observation states. Due to the differing observational states between agents, predictions of either the mean field or specific agent actions are potentially unreliable. By substituting those signals with a constructed stigmergic potential field, most of the coordinating benefits may be achieved using only the observations available to the agent.

3 Method

A suitable environment was required that captured all of the specific aspects of the drone coordination problem under the constraints mentioned in the project overview section. A custom multi-agent environment was created to fulfill are requirements.

The environment is a 2D grid with a number of drones attempting to locate and track an variable number of targets. At the center of the environment there is a central control node, the primary objective of the agents is to communicate the location of the target to this location. In the figure 2 the drones are shown as the blue x's, the control node is the black star and the targets are the grey +'s when not tracked and red +'s when tracked.

Each agent controls a drone platform, the platforms host a series of senses and data that are used to build the observations and eventually compose the stigmergic potential field. The specific senses are, an ability to detect friendly agents , the ability to detect target platforms and the ability to

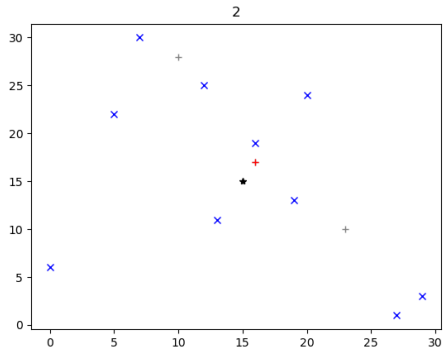


Figure 2: Basic components of the environment.

communicate with other friendly agents. The approximate relative sizes are shown in figure 3. When friendly or target detections are attempted, the area within the receptive search radii is marked as searched for the blue and red sensed observation arrays. Any detected entities are included in the blue and red knowledge arrays to record their position. As values in these observation arrays are updated, the current values also decay back to the initial state, this provides some temporal information to be encoded within each observation array.

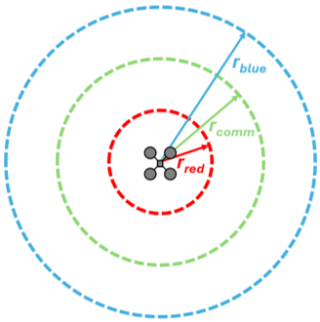


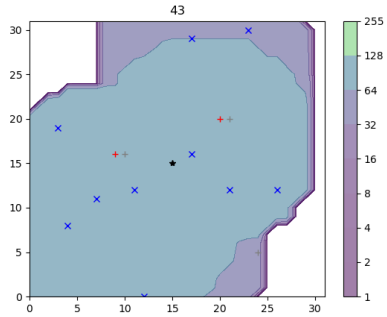
Figure 3: Approximate relative scale of drone senses and subsystems.

Communication happens at every timestep, any pair of friendly drones that are close enough exchange their current observation vectors and merge them with their current understanding of the environment. This same communications process also occurs with the base control node. The merging is possible because of the decaying nature observation arrays, naturally, newer observations will have higher values so a maximization process is done across the arrays. At the end of the communications step each pair should share their most up-to-date version of the world.

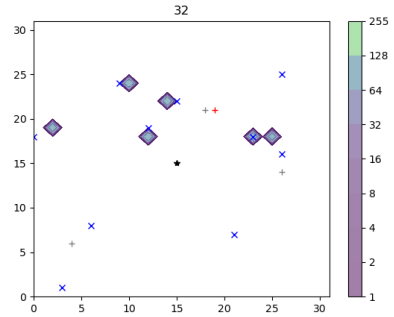
Each platform has a three dimensional observation vector made up of five 2D arrays each the size of the environment.

1. The area that has been searched for targets (Figure 5)
2. The location of any detected targets
3. The area that has been searched for allies (Figure 4a)
4. The location of any detected allies (Figure 4b)
5. The locations of the platform itself and the base control node

The stigmergic potential field was generated by applying various transformation the observation arrays. The potential fields were constructed such that negative values were found in areas where additional drones would be useful and positive value where they weren't. Locations of friendly agents generated a complex field to keep other agents from overlapping with their sensors but also to stay within communication range. The potential field allowed concepts about coordination to be applied



(a) Observation array showing the current state of the searched cells



(b) Observation array showing the detected friendly drones

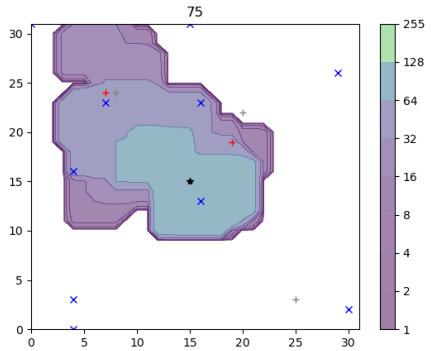


Figure 5: Observation array showing the current state of the searched cells.

to the observation vector without having to perform any learning. An example of a potential field is shown in Figure 6. Note that the potential field is generated with the agents current knowledge of the world and thus many values of the field are out of sync with the true location of the entities.

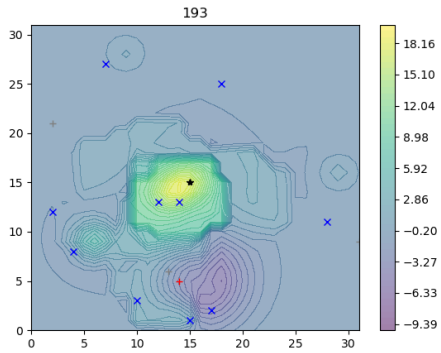


Figure 6: Constructed stigmetric potential rield.

Independent Proximal Policy Optimization (IPPO) was the used for the reinforcement learning. A single actor and critic network are trained for the experiences of all agents and then independently used without extra coordination. Using a centralized critic network would have likely improved performance but was unable to be completed in time. The fact that this methods struggles with coordination of multiple agents would hopefully be tempered by the inclusion of the SPF. The parameters used are shown in Table 1.

Table 1: PPO Settings

Parameter	Value
Clip Eps	0.1
PPO Epochs	3
Value Coeff	0.5
Entropy Coeff	0.01
GAE Lambda	0.99
γ	0.99
Reverse KL Coeff	0.01

Both the actor and critic networks are convolutional neural networks. This was selected to take advantage of the spacial information embedded within the observation arrays. There are three stages of convolution with MaxPooling and ReLU activations between them. After the convolutions the resulting tensors are flattened and two hidden layers are used. The actor network outputs a value for each possible action and the critic reports a single value as the value estimate for the current state. Figure 7 shows the basic network layout.

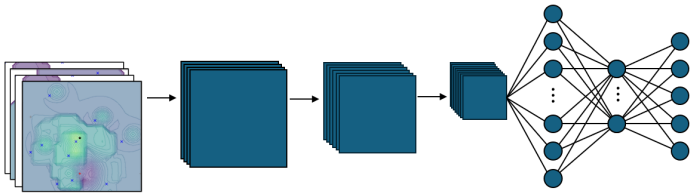


Figure 7: Actor network architecture, the critic network is similar with a single output instead.

The reward is the amount of time between the current simulation time and the last time each of the targets was detected, as measured at the base control node. Additional reward is computed for each agent based on the same metric.

$$\frac{1}{\#targets} \sum_{n=0}^{\#targets} e^{-\frac{t_{current}-t_{detected}}{50}} \tag{1}$$

When the SPF reward was included, additional reward was provided to each agent based the the gradient of the potential each action provided. For the evaluation steps of the RL algorithm, only the reward at the control node are considered.

4 Experimental Setup

As the DSSE environment has show that RL could be successfully applied to this application, the experiments conducted were focused on evaluating the efficacy of the inclusion of the additional SPF's in the speed and quality of the training. Therefore, the three following experimental cases were performed.

1. The SPF is included in the observation vector and as part of the reward
2. The SPF is included in the observation vector but not the reward
3. No inclusion of the SPF at all

Each of the experiments was run with the same environmental and alogrithmic settings, shown in tables 2 and 1. The selection of the 32x32 environment size and the tested agent counts were based on runtime considerations. Each of these experiments was run for different numbers of agents to determine what if any impact the number of agents had on the effectiveness of the SPF. One, two and four agents were tested under the three cases.

Table 2: Experimental Environment Settings

Parameter	Value
World Size	32x32
Red Sensor Radius	3
Blue Sensor Radius	5
Comm Radius	10
Red Agents	3

5 Results

5.1 Quantitative Evaluation

Table 3 shows the final trained evaluation reward earned in each of the three experiments for the increasing number of agents. For the one and two agent cases the SPF appears to provide a measure of improvement, both with the inclusion in the reward and just the observation vector. However, for four agents the results are nearly identical, indicating the SPF served no function in improving training.

Table 3: Trained Agent Comparison

Agents	Experiment 1	Experiment 2	Experiment 3
1	30.7	33.2	44.8
2	138.3	141.2	149.4
4	398.9	395.4	399.2

A similar story can be seen when examining the results of the training rate, Table 4. Again there is apparent improvement seen a the one and two agent cases, but by the four agent case no impact is noticeable. In fact, the addition of the SPF appears to have slowed training down significantly.

Table 4: Performance Comparison

Agents	Experiment 1	Experiment 2	Experiment 3
1	330k	275k	260k
2	330k	250k	200k
4	60k	75k	140k

The training curves in 8 show the apparent lack of effect of the SPF inclusion, indeed, the training curve where the additional reward was included shows that case being significantly delayed as compared to the other two experiments.

5.2 Qualitative Analysis

Despite the lack of effect from the inclusion of the SPFs, the agents were able to learn interesting behaviors. Two specific behaviors arose from unique restrictions placed on the environment. Due to the discrete communication distance, if enough agents were included in the simulation, they would form a chain to a target's location to facilitate the transfer of data to the control node. An example of this behavior can be seen in Figure 9. In cases where there were not enough drones to form the chain structures, the drones would ferry the target locations back to the control node themselves. In all cases, the trained agents would learn to avoid overlapping with other agents as well as avoid searching areas recently searched.

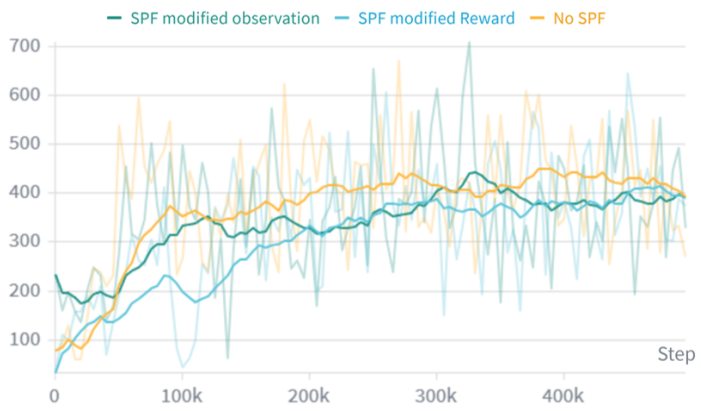


Figure 8: Evaluation curve for four agent case.

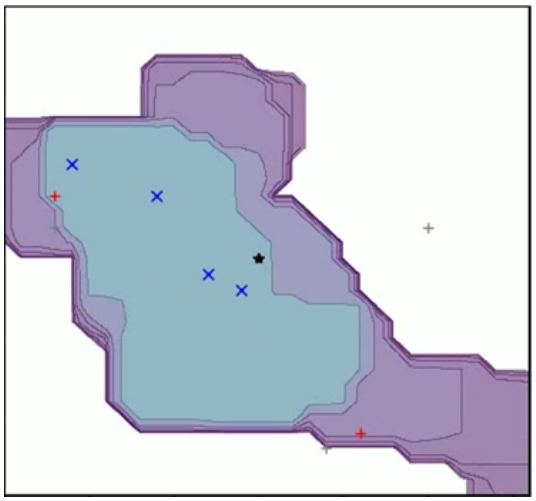


Figure 9: An example of chaining behavior.

6 Discussion

The lack of positive impact from the SPF could be caused by a number of reasons. The environment may have been too simple, mitigating the effect from any coordinating force from the SPF. The agents may have learned similar features contained within the SPF quickly. For cases the drone is more complex, say with a suite of senor capabilities or more complex communication requirement the inclusion of a constructed SPF may have proven more valuable.

Alternatively, the SPF construction may have been imprecise. The values and sized of the kernels used to create it mismatched to what would be beneficial to the agents. For example, the weight given to the radius of communication as compared to that of the target could be too small. These discrepancies could explain the apparent detrimental effect seen in the four agent case. This could be corrected by constructing the SPF from a diffusion model training in-line with the agents. That would allow the SPF to be tailored to the correct values for maximum coordination.

Finally, it is possible that encoding information this way is an impractical solution to the coordination problem. It may be the case the agent behavior is too complex to be meaningfully altered by a single scalar field even in this toy example.

7 Conclusion

The inclusion of the stigmergic potential field did not meaningfully impact the training or coordination of the agents. However, several factors may have confounded the results. Agents were still able to learn unique behavior in the environment but did so independent of whether or not a SPF was supplied.

Changes from Proposal Significant scaling back from the initial scope of experiments, I was only able to test the single algorithm and with fewer agents.

References

Raghav Aras, Alain Dutech, and Francois Charpillet. 2005. Stigmergy in Multi Agent Reinforcement Learning.

Robert Bamberger Jr., David Watson, David Scheidt, , and Kevin Moore. 2006. Flight Demonstrations of unmanned Aerial Vehicle Swarming Concepts.

Renato Laffranchi Falcão, Jorás Custódio Campos de Oliveira, Pedro Henrique Britto Aragão Andrade, Ricardo Ribeiro Rodrigues, Fabrício Jailson Barth, and José Fernando Basso Brancalion. 2024. *DSSE: An environment for simulation of reinforcement learning-empowered drone swarm maritime search and rescue missions*. doi:10.5281/zenodo.12659848

Ray Meese and Peter Means. [n. d.]. Making search and rescue drone swarms a reality. ([n. d.]).

Austin Nguyen. 2021. Scalable, Decentralized Multi-Agent Reinforcement Learning Methods Inspired by Stigmergy and Ant Colonies. arXiv:2105.03546v1 [cs.CL]

Huy Xuan Pham, Hung Manh La, David Feil-Seifer, and Ara Nefian. 2021. Cooperative and Distributed Reinforcement Learning of Drones for Field Coverage. arXiv:1911.12504v3 [cs.CL]

Clayton Rishe. 2011. APL's Jay Moore discusses project on UAVs.

Xing Xu, Rongpeng Li, Zhifeng Zhao, and Honggang Zhang. 2021. Stigmergic Independent Reinforcement Learning for Multi-Agent Collaboration. arXiv:1911.12504v3 [cs.CL]

Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. 2021. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. arXiv:1911.10635v2 [cs.CL]