2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Learning Heterogeneous Multi-Agent Allocations for Ergodic Search

Ananya Rao[1], Guillaume Sartoretti[2] and Howie Choset[1]

***Abstract*— Information-based coverage directs robots to move over an area to optimize a pre-defined objective function based on some measure of information. Our prior work determined that the spectral decomposition of an information map can be used to guide a set of heterogeneous agents, each with different sensor and motion models, to optimize coverage in a target region, based on a measure called ergodicity. In this paper, we build on this insight to construct a reinforcement learning formulation of the problem of allocating heterogeneous agents to different search regions in the frequency domain. We relate the spectral coefficients of the search map to each other in three different ways. The first method maps agents to pre-defined sets of spectral coefficients. In the second method, each agent learns a weight distribution over all spectral coefficients. Finally, in the third method, each agent learns weight distributions as parameterized curves over coefficients. Our numerical results demonstrate that distributing and assigning coverage responsibilities to agents depending on their sensing and motion models leads to 40%, 51%, and 46% improvement in coverage performance as measured by the ergodic metric, and 15%, 22%, and 20% improvement in time to find all targets in the search region, for the three methods respectively.**

I. INTRODUCTION

Deploying and coordinating multiple robots, particularly agents with different capabilities, can lead to improved search and coverage performance of a multi-agent team [1]–[5]. This diversity in agent characteristics presents the problem of how to coordinate agents to best use their individual search capabilities.

In this paper, we consider the coordination of heterogeneous agents in the context of ergodic search, where the planner uses an information map to create a path that spends time in regions that are more likely to contain a target. The planner uses the ergodic metric, which is the difference between the information distribution over the search domain and a distribution that encodes the agents' paths. By driving this difference to zero, the resulting trajectories spend more time in regions of higher expected information. This difference is computed in the frequency domain. Our prior work naturally associated high-order frequencies with limited-range sensors and low-order frequencies with wide-range sensors. Using this intuition, we were able to improve the ergodicity of search based on how we allocated agents to frequencies. While this work provided insight, it only determined handcrafted solutions to the problem of coordinating heterogeneous agents in specific, simple scenarios [6]. In this work, we learn how to distribute heterogeneous agents (i.e. agents with different capabilities) to suitable search regions by relying on ergodic search processes [1].

[1]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Mechanical Engineering Dept., National University of Singapore.

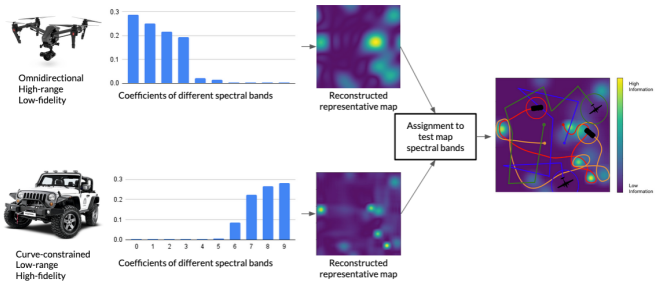

Fig. 1: We formulate a reinforcement learning problem to learn allocations of heterogeneous agents for search. Each agent is allocated to a different region of the search domain via weighted subsets of spectral coefficients, gotten from the spectral decomposition of the information distribution being searched.

We formulate heterogeneous agent distribution in multi-agent search as a reinforcement learning problem where the learned policy is an assignment of agents to different search regions (Fig 1). We define search regions in the frequency domain by assigning weights to the coefficients of the spectral decomposition of the expected information distribution over the search domain. Solving for agent allocation as a full optimization problem would be intractable, so we constrain the optimization by relating the spectral coefficient weights to each other in three different ways. In the first, we define a simple relationship by dividing the spectral coefficients into a set of spectral "bands." In our second method, we directly learn weights on each spectral coefficient for each agent, thereby allowing the learning process to define the search regions. Finally, in our third method, we parametrize the set of weights over all spectral coefficients as a function and learn the values of the function parameters for each agent. These methods give the planner different degrees of control over defining regions of the map to search.

Our empirical analysis of these three methods explores the trade-offs between the performance improvements of increasing the granularity at which agents can define their search regions and the consequent increase in required training time. We empirically show that all three methods improve coverage performance (in terms of the ergodic metric) and search performance (in terms of the time taken to find all targets) over a baseline method. Further, our experimental results support the intuition that granting agents more granular control over learning search regions to which they are well-suited results in greater improvement in both coverage and search performance. The second method, which

979-8-3503-8457-4/24/$31.00 ©2024 IEEE

most granularly defines search regions, achieves the best coverage and search performance, followed by the third method and finally the first method. These results support our hypothesis that it is possible to learn an effective allocation of heterogeneous agents to facilitate well-coordinated coverage of a search region.

II. BACKGROUND

A. Heterogeneous Multi-Agent Search

Many multi-agent search methods developed for groups of homogeneous agents (similar capabilities) either do not support heterogeneous groups of agents (varying capabilities) while others struggle with the increase in computational complexity [7]–[11]. Many works involving heterogeneous agents focus on defining multi-agent problems and offer only initial, usually centralized and non-scalable solutions to them [12]–[15]. While some works have considered using robots with better communication and coordination abilities as "leader agents" to plan for other agents [4], [16], other works have proposed using auction-based mechanisms for task assignment [17]–[19]. Still other works have considered redistribution of agents based on their capabilities [20], [21].

Search methods largely fall into one of three categories: geometric, gradient-based, and trajectory optimization-based approaches. When there is a near-uniform probability of finding a target, geometric methods, like lawnmower patterns, can be good search strategies [22], [23]. They uniformly and exhaustively cover the search domain, making them well-suited to cases when there is no *a priori* information about the locations of targets.

A priori information about the targets' locations is usually represented in the form of a probability distribution representing the likelihood of a target being found at each location in the domain. This information distribution allows for the creation and use of more advanced search processes that leverage this information to improve search with respect to some metric, such as time to find targets.

In gradient-based approaches [2], [24], [25], agents are guided by the local derivative of the information map around their positions in order to greedily maximize short-term information gain. However, these approaches typically do not take into account the uncertainty associated with information distributions, which impacts exploration since this uncertainty helps differentiate between areas of low information that have not been explored and areas with no information to be gained.

In optimization-based approaches, search is viewed as an information-gathering maximization problem, which is solved by planning paths for the agents. A popular approach in recent works [1], [26]–[28] is sampling-based path planning, in which a large number of paths are sampled and the best path is selected based on a cost metric. The cost function that drives the optimization can combine both the predicted information map and its associated uncertainty, thereby encouraging exploration. This work uses sampling-based planning, with the ergodic metric as cost function as the basis for spectral-based distributed heterogeneous multi-agent ergodic search with learned agent allocations.

B. Ergodic Search Processes

Ergodic search processes [28] produce trajectories that drive agents to spend time in areas of the domain in proportion to the expected amount of information present in those areas. The spatial time-average statistics of an agent's trajectory (trajectory is represented as $\gamma_i : (0, t] \to \mathcal{X}$), specifies the amount of time spent at position $\boldsymbol{x} \in \mathcal{X}$, where $\mathcal{X} \subset \mathbb{R}^{\mathrm{d}}$ is the d-dimensional search domain. For N agents, the joint spatial time-average statistics of the set of agents trajectories $\{\gamma_i\}_{i=1}^{N}$ is defined as [28]

$$C^t(\boldsymbol{x}, \gamma(t)) = \frac{1}{Nt} \sum_{i=1}^{N} \int_0^t \delta(\boldsymbol{x} - \gamma_i(\tau))\, d\tau, \tag{1}$$

where δ is the Dirac delta function.

The agents' trajectories are optimized by matching the spectral decompositions of the time-averaged trajectory statistics and the information distribution over the search domain. This is accomplished by minimizing the ergodic metric $\Phi(\cdot)$, which is the weighted sum of the difference between the spectral coefficients of these two distributions [28]:

$$\Phi(\gamma(t)) = \sum_{k=0}^{m} \alpha_k \left| c_k(\gamma(t)) - \xi_k \right|^2, \tag{2}$$

where c_k and ξ_k are the Fourier coefficients of the time-average statistics of the set of agents' trajectories $\gamma(t)$ and the desired spatial distribution of agents respectively, α_k are the weights of each coefficient difference, and m is the number of Fourier coefficients being considered.

The goal of ergodic coverage is to generate optimal controls $\boldsymbol{u}^*(t)$ for each agent, whose dynamics is described by a function $f\colon \mathcal{Q} \times \mathcal{U} \to \mathcal{TQ}$, such that

$$\begin{aligned} \boldsymbol{u}^*(t) \quad &= \quad \arg\min_{\boldsymbol{u}} \Phi(\gamma(t)), \\ \text{subject to } \dot{\boldsymbol{q}} \quad &= \quad f(\boldsymbol{q}(t), \boldsymbol{u}(t)), \\ \|\boldsymbol{u}(t)\| \quad &\leq \quad u_{max} \end{aligned} \tag{3}$$

where $\boldsymbol{q} \in \mathcal{Q}$ is the state and $\boldsymbol{u} \in \mathcal{U}$ denotes the set of controls. In this work, we solve the ergodic search problem by using a sampling-based motion planner [29].

C. Learning Methods

Broadly, machine learning methods can be grouped into three main categories: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, a mapping between inputs and outputs is inferred from labeled training data consisting of example input-output pairs. Classification problems, in which a class label is predicted, and regression problems, in which a numerical label is predicted, generally use supervised learning methods. On the other hand, unsupervised learning methods learn a model to describe or extract relationships in training data without target variables. These methods are generally used for clustering problems to find groups in data and for

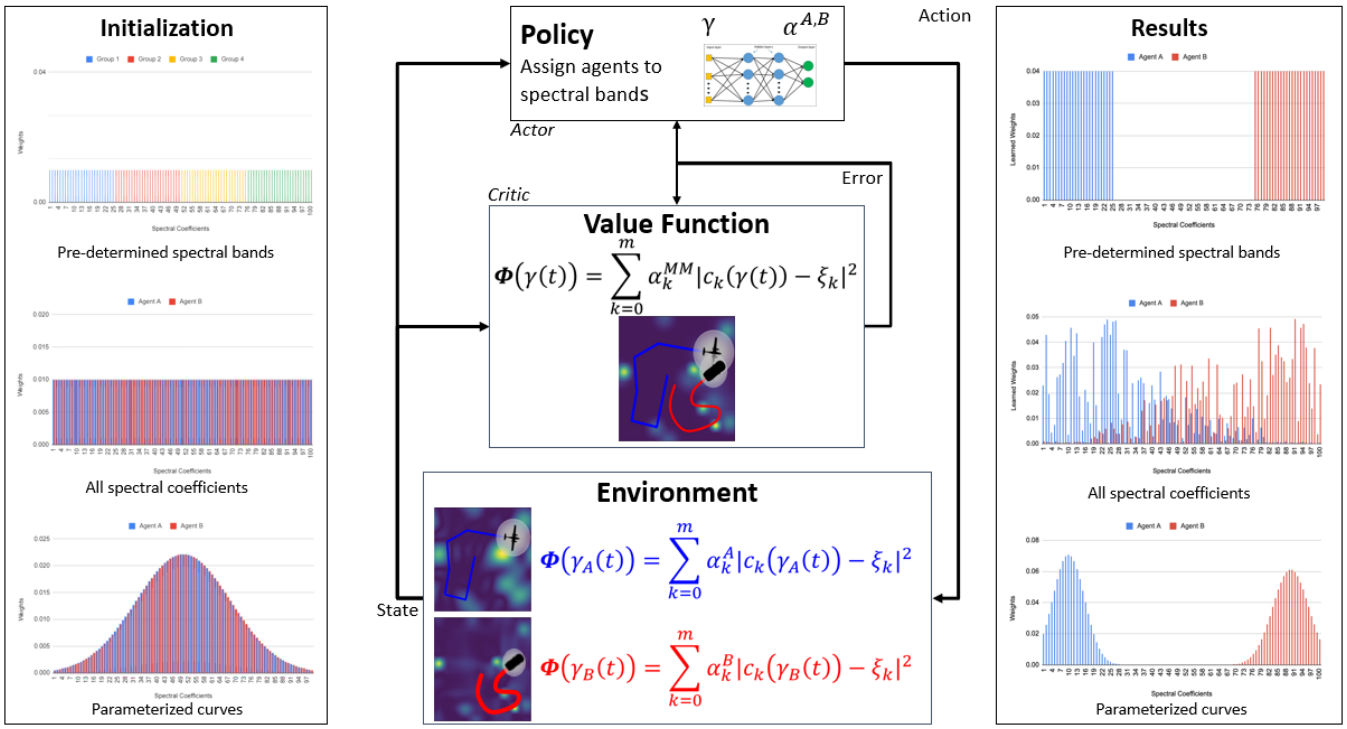

Fig. 2: We use the A2C algorithm [30], with three different distribution schemes, to learn the allocation of agents to search regions for heterogeneous multi-agent ergodic search. Each distribution method has different initial conditions. In all three methods, the output is a weighted set of spectral coefficients of the information distribution being searched for each agent. The learning process seeks to minimize the joint ergodic metric of the team's trajectories.

density estimation, which involves summarizing the distribution of data. In reinforcement learning, agents operate in an environment and learn how to operate using feedback. This means that instead of having a fixed training dataset, agents have a goal or a set of goals to achieve, actions that they can take, and feedback about performance toward the goal. Deep reinforcement learning combines reinforcement learning and deep learning, which opens the door to more complex applications.

Reinforcement learning techniques have been used in a variety of ways to plan paths for multi-agent systems. In many multi-agent environments, agents perceive the state of the environment at each step and take an action. A reward is evaluated based on this action, to optimize actions according to a desired metric, such as coverage, time-efficiency, or path lengths [31]–[34]. In multi-agent systems with communication, message lengths or frequency of communication can also be used in the reward structure [35].

In this work, we use an Advantage Actor-Critic (A2C) [30] network to learn mappings of spectral bands to different types of agents. A2C is an efficient reinforcement learning approach, that allows for scalability to large learning problems, which opens the door to expanding our approach to handle large teams of robots with many different types of agents.

III. Heterogeneous Multi-Agent Allocation using Reinforcement Learning

This work investigates learning-based approaches to autonomously allocating heterogeneous agents based on their sensing and motion capabilities. Here, search is decomposed into different regions in the frequency domain based on the spectral decomposition of the *a priori* target belief distribution; a search region corresponds to a weighted subset of spectral coefficients of the information distribution.

We use the Actor-Critic algorithm (A2C [30]) to learn data-driven mappings of agents to these search regions, or spectral scales (as shown in Fig 2). In this pipeline, each agent in a heterogeneous team is assigned a spectral scale of the search domain, after which paths are planned for all agents. A joint reward is then calculated for all agents, based on their coverage of the search map, measured by the ergodic metric. In this way, the reward optimizes for ergodicity, which in turn means that the learned assignments lead to paths that optimize for effective coverage.

Further, we present three specific methods of distributing agents in a heterogeneous team to spectral scales of a search domain. First, we map agents to pre-determined, handcrafted, spectral bands. In the second method, we learn individual weights over all of the spectral coefficients of the information distribution. Finally, weights for parameterized curves over the spectral coefficients of the information distribution are learned.

A. Pre-Determined Spectral Bands

One method of autonomously using the spectral decomposition of the *a priori* information map relies on leveraging intuition regarding the scales of information encoded by different orders of spectral coefficients or frequency modes. This method simplifies the agent allocation problem. As stated previously, we know that lower-frequency spectral coefficients correspond to large-scale variations in the spatial distribution of information, while higher-frequency spectral coefficients correspond to smaller-scale variations in the spectral decomposition of the information distribution. Based on this intuition, we can group frequency coefficients into M distinct subsets (where M is the number of different types of agents being considered), forming distinct spectral bands. Each of these spectral bands correspond to a different search region. Note that these search regions, while separate, may not be independent.

Reinforcement learning is then used to learn a data-driven mapping of agents in a heterogeneous team to the created spectral bands, as depicted in our proposed pipeline (Fig. 1). That is, we assign a weight to each spectral band, which can be modified during training. Each agent in a heterogeneous team with M different types of agents will therefore have M different trainable weights. Each agent is assigned to the spectral band corresponding to the highest weight.

This method of distributing agents to pre-defined spectral scales of the information distribution, as described in Section III-A, has two main limiting factors. First, this method has the overhead of defining the spectral bands, which scales with the number of types of agents in the heterogeneous team being considered. Additionally, the accuracy of the pre-defined spectral bands is limited by human intuition, which limits the efficacy of this method, especially for agents with more complex capabilities.

B. All Spectral Coefficients

To overcome the limitations of pre-defined spectral bands discussed above, we can instead learn weights over the individual frequency coefficients of the spectral decomposition of the target belief distribution. In this method, each spectral coefficient has a weight that can be independently modified during training. Each set of learned weights corresponds to a different spatial distribution of information, and thus a different search region. These search regions are not necessarily independent. This method allows for increased granularity of defining search regions, or spectral scales since each type of agent determines its own.

While learning weights over all of the frequency coefficients of the spectral decomposition of the information distribution (Section III-B) overcomes the limitations posed by learning a mapping of agents to pre-defined spectral bands (Section III-A), it poses its own issues. Primarily, the complexity of this learning problem scales with the number of frequency coefficients, leading to slow training.

C. Parameterized Curves Over Spectral Coefficients

The limitations of learning weights over all of the frequency coefficients discussed above can be overcome by instead learning parameters that define weight curves over the set of frequency coefficients.

In this method, the weights of each frequency coefficient in the spectral decomposition of the target belief map are defined according to a parameterized curve. A distinct parameterized curve is learned for each different type of agent in the heterogeneous team, i.e. if there are M types of heterogeneous agents, M sets of parameters are learned. The parameters that define each of these distinct curves are adjusted during training (i.e. are the trainable weights).

Each parameterized curve defines a different set of weights over the spectral coefficients, which correspond to a separate, but not necessarily independent, spatial distribution of information. Agents define these search regions (and are implicitly assigned to them) by learning the parameters that define a curve over the frequency coefficients.

IV. RESULTS AND DISCUSSION

We present the results of our systematic investigation into three different methods of autonomously allocating heterogeneous agents by relying on a large set of simulation experiments. Our set of experiments is composed of fixed, randomly generated search problems, encompassing different agent types, team sizes, and team compositions. The baseline method in our experiments is multi-agent ergodic search without distribution - i.e. all agents plan paths based on the same initial prior. We compare these methods according to various standard search metrics, such as time to find all targets and the effectiveness of coverage, according to the ergodic metric. Our results show that distributing and assigning coverage responsibilities to agents depending on their dynamic sensing capabilities leads to 40%, 51%, and 46% improvement with regard to the ergodic metric, and 15%, 22%, and 20% improvement in time to find all targets, for each of the three distribution methods respectively. Our results rely on sampling-based trajectory optimization, but we emphasize that our investigation should extend to other optimization methods.

A. Experiment Details

1) Agent's Sensing and Motion Models: We describe the area that each agent's sensor covers as a Gaussian distribution that's centered at the agent's location. At each point in the agent's sensor footprint, the likelihood of detecting a target is defined using the Gaussian probability density function. We consider two types of sensors. The first is a low-range, high-fidelity sensor, depicted by a Gaussian with a low spread but higher detection probability. The other is a high-range, low-fidelity sensor, represented as a Gaussian with a wide spread. These sensors have a lower probability of detecting targets.

Apart from different sensor models, we also consider two types of motion models. The first models omnidirectional agents, such as quad-rotor UAVs, as a simple first-order

integrator. We also consider agents that have differential drive constraints, like wheeled ground vehicles. This motion model is represented as curved paths with a maximum curvature.

To sample agent paths, we sequence path primitives. For the omnidirectional agents, the path primitives are straight lines of different lengths and directions. For curve-constrained agents, the path primitives are various curves from a finite set with different curvatures and lengths. Agents plan long trajectories, execute these paths for 10 timesteps, update the map using their observations, and then re-plan. We optimize agent paths using a cross-entropy planner [29] with 3 levels of sample refinement with a total of $15 \cdot N$ samples, where N is the total number of agents in the multi-agent team.

2) Learning Method: We use an Advantage Actor-Critic (A2C) [30] network to learn assignments of spectral bands of the information map to agents in the heterogeneous team. We view taking an action as assigning a spectral band to an agent, and this action is evaluated based on the paths that the agents plan. In the first method, the network outputs a set of $N \cdot B$ weights, where N is the number of agents in the heterogeneous team and B is the number of spectral bands that the spectral decomposition of the information map is divided into. In our experiments, $B = 4$. In the second method, the network outputs a set of $N \cdot S$ weights, where S is the total number of spectral coefficients used to represent the information map in the frequency space. In the third method, the network outputs a set of $N \cdot P$ weights, where P is the number of parameters required to define the function over coefficient weights for each agent. We use $P = 2$. For all three methods, we learn assignments for single agents, two-agent teams, and four-agent teams ($N = \{1, 2, 4\}$).

In each episode, spectral bands are picked for each agent based on the learned weights. Then, paths are planned for each agent according to the reconstructed information distribution that the agent is relying on. The reward used is defined in Eq 4, so the reward increases as the ergodicity decreases.

$$\text{reward} = -log(\Phi(\gamma(t))) \tag{4}$$

Since more effective coverage paths will have lower ergodicity, this reward structure encourages weights to train such that spectral bands are assigned to agents in a manner that allows for effective coverage.

3) Scenarios Randomization: We compare the performance of various distribution methods with that of a baseline that plans paths for all agents by relying on the overall distribution maps (i.e., no decomposition into bands or assignments), through 100 randomized search scenarios for each distribution map. These scenarios vary the locations of targets and the initial information maps (as randomly generated Gaussian mixture models). Additionally, for each experiment, a randomly generated team of agents is formed by selecting both the sensing and motion models with equal probability for each agent. We ran experiments on three different team sizes: single-agent systems, a multi-agent system with two agents, and a multi-agent system with four agents. Team compositions, starting positions, initial information maps, and target locations are kept identical among experiments with different controllers, to ensure our results are comparable.

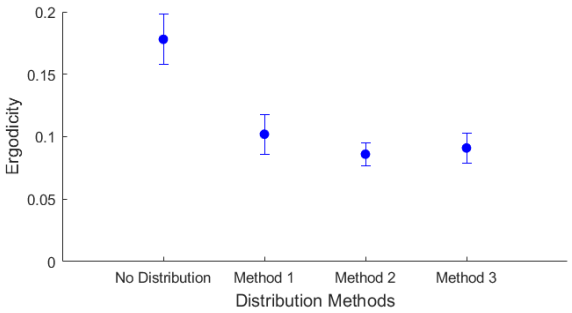

Fig. 3: Search performance comparison between the different agent distribution methods - pre-determined spectral bands (Method 1), learning weights over all coefficients (Method 2), learning parameterized curves over coefficients (Method 3) - and the baseline (No Distribution), in terms of coverage performance (using the ergodic metric, lower is better).

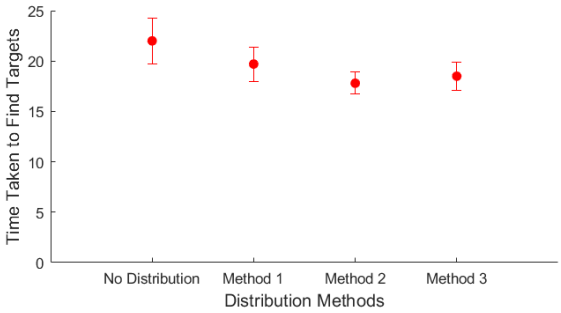

Fig. 4: Search performance comparison between the different agent distribution methods - pre-determined spectral bands (Method 1), learning weights over all coefficients (Method 2), learning parameterized curves over coefficients (Method 3) - and the baseline (No Distribution), in terms of time to find all targets (lower is better).

B. Experimental Results

When looking at the results of the different distribution methods in terms of overall coverage performance, measured using the ergodic metric, we can observe that by distributing the agents to pre-defined spectral bands, there is on average a 40% improvement over the baseline. Learning weights over all coefficients and learning parameterized Gaussian curves over the spectral coefficients lead to on average 51% and 46% improvements in coverage performance respectively (Fig 3). Our results support the notion that more effective coverage of the domain leads to finding targets faster. This is seen through the improvements in time to find all targets over the baseline method (15%, 22%, and 20% for the three distribution methods respectively), shown in Fig 4.

While each method leads to improvements in both coverage efficiency and time to find all targets, the details of

how these improvements are achieved vary slightly from method to method, as do specific pros and cons. The three methods demonstrate a trade-off between the performance improvements offered by a more granular assignment of agents to spectral bands and the associated increase in training time.

For the first method, the resulting mappings of agents to pre-determined spectral bands correspond to intuition expressed in Section III-A. We believe that lower-frequency spectral coefficients correspond to broad domains of information, while higher-frequency spectral coefficients correspond to details and edges in the information distribution. We observe that agents with low-fidelity, high-range sensors and omnidirectional motion models tend to map to the first spectral band, and are driven to perform a more coarse exploration of the search domain. Agents with high-fidelity, low-range sensors and curve-constrained motion models tend to map to the last spectral band, which drive to string together areas of higher information.

In the case of learning weights over all spectral coefficients, the increased granularity of forming search regions, as compared to pre-defining spectral bands to form search regions, leads to more effective coverage, and relatedly, faster search times. In this method, agents can effectively create their own subsets of frequency coefficients by learning weights over all coefficients, as opposed to choosing between a fixed number of pre-determined subsets. This means that agents can both learn their search capabilities and formulate a weighted set of spectral coefficients that correspond to suitable search regions.

The method of learning weights for parameterized curves over spectral coefficients has a trade-off between learning problem complexity (and as a result training time) and performance (in terms of coverage, measured by ergodic metric and time to find all targets). Learning parameters for curves over frequency coefficient weights instead of learning the coefficient weights themselves reduces problem complexity by reducing the number of trainable weights. As a result, training is faster. However, while assigning weights according to learned parameterized curves still allows agents to self-define search regions, there is less granularity in this definition. This leads to more effective coverage and faster search times compared to mapping agents to spectral bands, but worse performance than results from learning weights over all coefficients.

V. CONCLUSIONS

We investigated the use of learning techniques to autonomously leverage the spectral nature of ergodic coverage, improving heterogeneous multi-agent search of a domain by distributing agents to different search regions. These regions were defined as performing search at different spatial scales, by relying on a limited, weighted, set of spectral coefficients that represent the overall target belief distribution.

We present and experimentally analyze three methods of autonomously allocating agents to different regions in the frequency domain. In the first of these methods, the search regions are formulated by pre-defining spectral coefficient bands, based on intuition linking agents' sensing and motion modalities and the different search scales. A mapping between agents in a heterogeneous team and these pre-defined spectral bands is then learned. In the second method, search regions are formulated and implicitly allocated to agents by directly learning weights for each frequency coefficient in the spectral decomposition of the information map. Search regions are formulated and allocated to agents in the third distribution method based on learned parameterized curves, which define the spectral coefficient weights.

In our systematic numerical tests, we compared the results of all three distribution methods to a baseline that plans paths for all agents regardless of their individual capabilities. We showed that all three distribution methods lead to improved performance in terms of coverage efficiency (measured by the ergodic metric) and time to find all targets. Specifically, our results show 40%, 51%, and 46% improvement with regard to standard coverage metric (ergodicity), and 15%, 22%, and 20% improvement in time to find all targets, for the three distribution methods respectively.

This work paves the way for developing new heterogeneous multi-agent search methods, particularly automatically identifying and leveraging synergies among agents with complex and novel capabilities. Machine learning techniques, as demonstrated in this work, can be leveraged to extend heterogeneous multi-agent search methods to large teams with complex compositions, without having to rely on building human intuition regarding agent, or team, capabilities. We assume in this work that our *a priori* information map is accurate, however future work will investigate the potentially increased errors that may be caused by the use of inaccurate priors, particularly in terms of increased search times, for distributed search. Additionally, this work considers centralized allocation and path planning, but our future work will seek decentralized agent allocation (and potentially ergodic path planning) solutions to allow such distributed heterogeneous multi-agent search methods to scale to large teams, and ultimately allow large-scale real-life deployments.

REFERENCES

[1] E. Ayvali, H. Salman, and H. Choset, "Ergodic coverage in constrained environments using stochastic trajectory optimization," in *International Conference on Intelligent Robots and Systems*, IEEE, 2017, pp. 5204–5210.

[2] P. Lanillos, S. K. Gan, E. Besada-Portas, G. Pajares, and S. Sukkarieh, "Multi-uav target search using decentralized gradient-based negotiation with expected observation," *Information Sciences*, vol. 282, pp. 92–110, 2014.

[3] R. R. Murphy, *Disaster robotics*. MIT press, 2014.

[4] P. Chand and D. A. Carnegie, "Mapping and exploration in a hierarchical heterogeneous multi-robot system using limited capability robots," *Robotics and autonomous Systems*, vol. 61, no. 6, pp. 565–579, 2013.

[5] T. H. Chung, G. A. Hollinger, and V. Isler, "Search and pursuit-evasion in mobile robotics," *Autonomous robots*, vol. 31, no. 4, pp. 299–316, 2011.

[6] G. Sartoretti, A. Rao, and H. Choset, "Spectral-based distributed ergodic coverage for heterogeneous multi-agent search," in *International Symposium Distributed Autonomous Robotic Systems*, Springer, 2021, pp. 227–241.

[7] Z. Yan, N. Jouandeau, and A. A. Cherif, "A survey and analysis of multi-robot coordination," *International Journal of Advanced Robotic Systems*, vol. 10, no. 12, p. 399, 2013.

[8] M. B. Dias, R. Zlot, N. Kalra, and A. Stentz, "Market-based multirobot coordination: A survey and analysis," *Proceedings of the IEEE*, vol. 94, no. 7, pp. 1257–1270, 2006.

[9] J.-H. Kim and P. Vadakkepat, "Multi-agent systems: A survey from the robot-soccer perspective," *Intelligent Automation & Soft Computing*, vol. 6, no. 1, pp. 3–17, 2000.

[10] F. Tang and L. E. Parker, "ASyMTRe: Automated synthesis of multi-robot task solutions through software reconfiguration," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2005, pp. 1501–1508, ISBN: 078038914X.

[11] L. E. Parker, *ALLIANCE: An architecture for fault tolerant, cooperative control of heterogeneous mobile robots*, 1994.

[12] A. Prorok, M. A. Hsieh, and V. Kumar, "Formalizing the impact of diversity on performance in a heterogeneous swarm of robots," in *International Conference on Robotics and Automation*, IEEE, 2016, pp. 5364–5371.

[13] T. S. Dahl, M. Matarić, and G. S. Sukhatme, "Multi-robot task allocation through vacancy chain scheduling," *Robotics and Autonomous Systems*, vol. 57, no. 6-7, pp. 674–687, 2009, ISSN: 09218890.

[14] E. G. Jones, B. Browning, M. B. Dias, B. Argall, M. Veloso, and A. Stentz, "Dynamically formed heterogeneous robot teams performing tightly-coordinated tasks," in *International Conference on Robotics and Automation*, IEEE, 2006, pp. 570–575.

[15] M. Koes, I. Nourbakhsh, and K. Sycara, "Heterogeneous multirobot coordination with spatial and temporal constraints," in *AAAI*, vol. 5, 2005, pp. 1292–1297.

[16] R. Grabowski, L. E. Navarro-Serment, C. J. Paredis, and P. K. Khosla, "Heterogeneous teams of modular robots for mapping and exploration," *Autonomous Robots*, vol. 8, no. 3, pp. 293–308, 2000.

[17] P. García, P. Caamaño, R. J. Duro, and F. Bellas, "Scalable task assignment for heterogeneous multi-robot teams," *International Journal of Advanced Robotic Systems*, vol. 10, 2013, ISSN: 17298806. DOI: 10.5772/55489.

[18] B. P. Gerkey and M. J. Matarić, "Sold!: Auction methods for multirobot coordination," *IEEE Transactions on Robotics and Automation*, vol. 18, no. 5, pp. 758–768, 2002, ISSN: 1042296X.

[19] R. Zlot, A. T. Stentz, M. B. Dias, and S. Thayer, "Multi-robot exploration controlled by a market economy," in *International Conference on Robotics and Automation*, vol. 3, 2002, pp. 3016–3023.

[20] A. Prorok, M. A. Hsieh, and V. Kumar, "Fast redistribution of a swarm of heterogeneous robots," in *EAI International Conference on Bio-inspired Information and Communications Technologies*, ICST, 2016, pp. 249–255.

[21] A. Halász, M. A. Hsieh, S. Berman, and V. Kumar, "Dynamic redistribution of a swarm of robots among multiple sites," in *International Conference on Intelligent Robots and Systems*, IEEE, 2007, pp. 2320–2325.

[22] H. Choset, "Coverage for robotics–a survey of recent results," *Annals of mathematics and artificial intelligence*, vol. 31, no. 1, pp. 113–126, 2001.

[23] V. Ablavsky and M. Snorrason, "Optimal search for a moving target - A geometric approach," in *AIAA Guidance, Navigation, and Control Conference and Exhibit*, AIAA, Aug. 2000.

[24] J. L. Baxter, E. Burke, J. M. Garibaldi, and M. Norman, "Multi-robot search and rescue: A potential field based approach," in *Autonomous robots and agents*, Springer, 2007, pp. 9–16.

[25] E.-M. Wong, F. Bourgault, and T. Furukawa, "Multi-vehicle bayesian search for multiple lost targets," in *International Conference on Robotics and Automation*, IEEE, 2005, pp. 3169–3174.

[26] E. Ayvali, A. Ansari, L. Wang, N. Simaan, and H. Choset, "Utility-guided palpation for locating tissue abnormalities," *Robotics and Automation Letters*, 2017.

[27] L. M. Miller, Y. Silverman, M. A. MacIver, and T. D. Murphey, "Ergodic exploration of distributed information," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 36–52, 2016.

[28] G. Mathew and I. Mezić, "Metrics for ergodicity and design of ergodic dynamics for multi-agent systems," *Physica D: Nonlinear Phenomena*, vol. 240, no. 4, pp. 432–442, 2011.

[29] M. Kobilarov, "Cross-entropy motion planning," *The Int. J. Robot. Res.*, vol. 31, no. 7, pp. 855–871, 2012.

[30] V. Mnih, A. P. Badia, M. Mirza, *et al.*, "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*, PMLR, 2016, pp. 1928–1937.

[31] J. K. Gupta, M. Egorov, and M. Kochenderfer, "Cooperative multi-agent control using deep reinforcement learning," in *International Conference on Autonomous Agents and Multiagent Systems*, Springer, 2017, pp. 66–83.

[32] T. Wang, T. Gupta, A. Mahajan, B. Peng, S. Whiteson, and C. Zhang, "Rode: Learning roles to decompose multi-agent tasks," in *ICLR 2021: Proceedings of the ninth International Conference on Learning Represen-*

tations, May 2021. [Online]. Available: https://openreview.net/pdf?id=TTUVg6vkNjK.

[33] X. Kong, B. Xin, Y. Wang, and G. Hua, "Collaborative deep reinforcement learning for joint object search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.

[34] M. A. Lopes Silva, S. R. de Souza, M. J. Freitas Souza, and A. L. C. Bazzan, "A reinforcement learning-based multi-agent framework applied for solving routing and scheduling problems," *Expert Systems with Applications*, vol. 131, pp. 148–171, 2019, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2019.04.056. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417419302866.

[35] S. Sukhbaatar, A. Szlam, and R. Fergus, "Learning multiagent communication with backpropagation," *arXiv preprint arXiv:1605.07736*, 2016.