Role-based Cooperation for Environmental Monitoring with Multiple UAVs

Jesus Capitan[1], Matthijs T.J. Spaan[2], Luis Merino[3]

***Abstract*— Planning under uncertainty faces a scalability problem when considering multi-robot teams, as the information space scales exponentially with the number of robots. To address this issue, this paper proposes to decentralize multi-robot Partially Observable Markov Decision Processes (POMDPs) while maintaining cooperation between robots by using POMDP policy auctions. Auctions provide a flexible way of coordinating individual policies modeled by POMDPs and have low communication requirements. Additionally, we use a Decentralized Data Fusion method (DDF) in order to efficiently maintain a joint belief state among the robots. The paper presents an application in environmental monitoring with multiple Unmanned Aerial Vehicles (UAVs), which illustrates the proposed ideas through different simulations.**

I. INTRODUCTION

Systems with multiple UAVs are of great interest in many robotic applications, such as exploration, surveillance, monitoring or rescue robotics [1], [2], [3], [4]. In those applications, a single UAV is not usually able to acquire all the required information and the cooperation among multiple UAVs is essential. In particular, we are interested in role-based missions, in which the team objective (e.g., detecting a target or alarm) can be achieved with UAVs following different roles or behaviors (e.g., patrol a certain area, approach the target, etc.). For instance, in many monitoring applications [4], where the team needs to maximize its information, it is positive to have UAVs following non-overlapping behaviors in order to provide a richer information for the team.

Real scenarios present uncertain and potentially hazardous environments in which UAVs can experience communication constraints. In order to cope with decision-making under these uncertainties and constraints, POMDPs provide a sound mathematical framework [5]. Although POMDP solvers can currently handle large state spaces, they ultimately face a scalability problem when considering multi-robot teams [6]. Popular models like Dec-POMDPs [7] or ND-POMDPs [8] remain limited to toy problems, and other models require flawless instantaneous communication [9], [10].

We propose a scheme for exploiting the power of POMDPs while mitigating their complexity. First, we exploit the fact that robotic teams are usually capable of communicating, and thus, we maintain a joint belief state among the UAVs, which serves as coordination signal. We use an existing DDF approach [11], but in conjunction with POMDP policies. Unlike most work on POMDPs, the belief update is separated from the decision-making process during the execution phase. This decoupling between both processes increases the robustness and reliability of real-time robotic teams.

Second, we propose to combine individual behaviors or roles that can be represented by single-robot POMDPs. An online cooperation is fostered by distributing optimally these roles among the UAVs by means of a decentralized auction. Instead of tasks, POMDP policies that describe a behavior are distributed; UAVs can switch between these behaviors dynamically at each decision step. The auction determines continuously which behavior is best for each UAV to cooperatively attain the common goal. Since the POMDPs only involve local information and all the communications are point-to-point, the approach can scale well with the number of UAVs.

We apply our approach to environmental monitoring, where several UAVs have to evaluate the level of contamination on a given terrain with less uncertainty as possible. Simulation results prove how the cooperation among several UAVs can improve the overall mission.

The paper is organized as follows: Section II summarizes POMDP models and describes the decentralized data fusion algorithms; Section III describes the algorithms for auctioning POMDPs in a decentralized manner and the overall overview of the complete system; Section IV presents the models used for environmental monitoring; Section V provides experimental results; and Section VI gives the conclusions and future work.

II. BACKGROUND

A. Decision-theoretic Planning Models

A popular model for single-UAV planning under uncertainty in sensing and acting is the Partially Observable Markov Decision Process (POMDP).

Formally, a POMDP is defined by the tuple $\langle S, A, Z, T, O, R, h, \gamma \rangle$ [5]: The *state space* is the finite set of possible states $s \in S$; the *action space*, the finite set of possible actions $a \in A$; and the *observation space* consists of the finite set of possible observations $z \in Z$. At every step, an action is taken, an observation is made and a reward is given. Thus, after performing an action a, the state transition is modeled by the conditional probability function $T(s', a, s) = p(s'|a, s)$, and the posterior observation by the conditional probability function $O(z, a, s') = p(z|a, s')$. The reward obtained at each step is $R(s, a)$.

Given that it is not directly observable, the actual state cannot be known by the system. Instead, a probability density

[1] Instituto Superior Tecnico, Lisbon, Portugal. jescap@isr.ist.utl.pt

[2] Delft University of Technology, Delft, The Netherlands. m.t.j.spaan@tudelft.nl

[3] Pablo de Olavide University, Seville, Spain. lmercab@upo.es

function $b(s)$ over the state space (belief state) is maintained. Due to the Markov assumption, it can be updated with a Bayesian filter for every action-observation pair.

$$b'(s') = \eta O(z, a, s') \sum_{s \in S} T(s', a, s) b(s) \qquad (1)$$

where η acts as a normalizing constant such that b' remains a probability distribution.

The objective of a POMDP is to find a policy that maps beliefs into actions $\pi(b) \rightarrow a$, so that the value function is maximized, i.e. the sum of expected rewards earned during h time steps. To ensure that this sum is finite when $h \rightarrow \infty$, rewards are weighted by a discount factor $\gamma \in [0, 1)$.

The model can be extended for the case of multiple UAVs (MPOMDP). In a team of n UAVs, each UAV i can execute an action a_i from a finite set A_i and receives an observation z_i from a finite set Z_i. The transition function $T(s', a_J, s)$ is defined over the set of joint actions $a_J \in A_1 \times \cdots \times A_n$, and the observation function $O(z_J, a_J, s')$ over joint actions and joint observations $z_J \in Z_1 \times \cdots \times Z_n$. The common reward signal is defined over the joint set of states and actions $R : S \times A_1 \times \cdots \times A_n \rightarrow \mathbb{R}$.

The goal in the multi-UAV case case is to compute an optimal joint policy $\pi^* = \{\pi_1, \cdots, \pi_n\}$ that maximizes the expected discounted reward (as in the POMDP case). In the MPOMDP case, as UAVs at each time step have access to the joint observation and as a result can deduce the joint action, they can maintain a joint belief using (1) (substituting the single-UAV models with the joint ones).

B. Decentralized Data Fusion

A joint belief for the team can be estimated in a decentralized way. Each UAV i can employ its local data z_i to compute a belief state over the full trajectory (from time 0 up to time t):

$$b_i(s_{0:t}) = \eta \prod_{\tau=1}^{\tau=t} O(z_{i,\tau}, a_{i,\tau}, s_\tau) T(s_\tau, a_{J,\tau}, s_{\tau-1}) b(s_0) \quad (2)$$

with η a normalization constant. Then, the local beliefs can be shared among neighbors at certain time instants. If UAV i receives information from other UAV j, this is locally fused in order to improve its local perception of the world:

$$b_i(s_{0:t}) \leftarrow \eta \frac{b_i(s_{0:t}) b_j(s_{0:t})}{b_{ij}(s_{0:t})} \qquad (3)$$

where $b_{ij}(s_{0:t})$ represents the common information between the UAVs (i.e., the common prior, and information previously exchanged between the UAVs). This information can be maintained by a separate filter called channel filter [12].

It is important to remark that, with this approach, the centralized belief can be *exactly* recovered in a decentralized fashion [11]. However, maintaining a belief for the state trajectory is very costly. In [11], it is presented an algorithm for DDF that scales only linearly with the length of the trajectory, under the assumption of Gaussian beliefs. For other belief functions, the same equation (3) can be applied to the belief on the last state $b(s_t)$. However, some error will be committed with respect to the centralized belief if the fusion equation is not applied every time instant in which a measurement is obtained in the team.

III. DECENTRALIZED AUCTION WITH POMDPS

The proposed approach builds on two mechanisms: the DDF filter described in Section II-B and a POMDP auction (Section III-A). The former allows the UAVs to share information and build a joint belief, the latter is used to assign the different behaviors to the UAVs in a cooperative manner.

A. Auctioning POMDP Policies

Certain multi-UAV applications can be attained by the cooperation between UAVs that play different roles. Here, a single-UAV POMDP is defined for each given role/behavior $k \in \{1, \ldots, m\}$ and UAV $i \in \{1, \ldots, n\}$. With each POMDP, the reward function of the corresponding behavior R_i^k is associated, which is defined over the sets of local variables $\langle S_i, A_i, Z_i \rangle$. In an offline planning phase, individual policies are computed for all these POMDPs, each of them with a value function $V_i^k(b_i)$. Then, in an online execution phase, these policies are assigned optimally to the different UAVs in order to achieve a cooperative behavior.

Although the actual multi-UAV objective cannot be modeled as a set of single-UAV reward functions, if these policies could be assigned optimally to one or more UAVs, all together should lead to a cooperative behavior pursuing the global objective.

The role assignment is modeled as a task allocation problem [13], in which m tasks have to be assigned to a team of n UAVs minimizing a global cost. In this case, each UAV always has to be assigned a sole task, which is the POMDP policy to follow. Given that $x_{ik} = 1$ when policy k is assigned to robot i and 0 otherwise, and ϕ_{ik} is the cost associated with that assignment, the problem consists of minimizing the total cost $\min \sum_{i=1}^{n} (\sum_{k=1}^{m} \phi_{ik} x_{ik})$, subject to $\sum_{i=1}^{n} x_{ik} \leq 1, \forall k \in \{1, \ldots, m\}$; $\sum_{k=1}^{m} x_{ik} \leq 1, \forall i \in \{1, \ldots, n\}$; and $x_{ik} \in \{0, 1\}, \forall i, j$.

The behavior for each UAV is selected online with an auction algorithm [13] where the cost of assigning a policy k to a UAV i is $\phi_{ik} = -V_i^k(b_i)$. Thus, policies with greater expected reward are more likely to be selected for each UAV, which helps to maximize the global expected reward for the whole team. In case $n > m$, the algorithm will leave UAVs with no policy assigned. Therefore, the assignment problem is repeated with these free UAVs until they all get a policy.

In our decentralized auction, the assignment problem is solved locally at each UAV with the information available. Each UAV i computes its own costs/bids for the behaviors from its local belief b_i and communicates them to other neighboring UAVs. Then, with the bids received from other UAVs, a local solution for the assignment problem above is obtained. This assignment is solved at each decision step for the UAVs. The computation can be performed efficiently in polynomial time using the Hungarian algorithm [14]. Note that each UAV only consider its neighboring peers (within communication range), which bounds the total cost

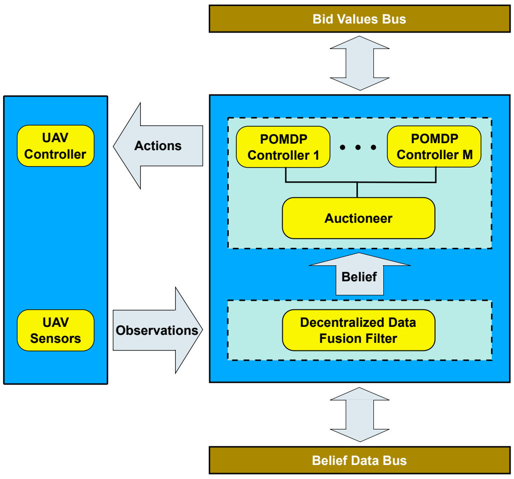


Fig. 1: Functional scheme for decision-making and belief update at each UAV.

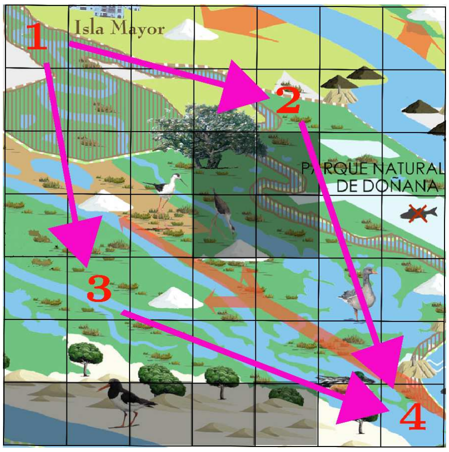


Fig. 2: Marshland in the National Park of Doñana. Shadowed cells represent non-flying zones. Each critical cell is marked with a number, and the propagation graph with arrows.

of the Hungarian algorithm. Moreover, the robustness of the system is high, since information from all the UAVs is not required to compute each local solution. In case some communication links failed, each UAV would still get a suboptimal solution with the available information from their neighbors (subnetworks arise naturally).

It is important to remark that the UAVs have access to a joint belief during the execution of the policies. This belief over the joint state, and containing information from all the UAVs, is provided by the DDF algorithm running during the execution phase.

B. System Overview

Figure 1 depicts the system elements per UAV. Each UAV can execute a certain number of behaviors modeled as single-UAV POMDP controllers. A DDF module is in charge of computing the belief and feeding the Auctioneer module, which then chooses the adequate POMDP controller and the associated action. Even though most POMDP-based systems synchronize belief update and decision-making in the same loop, here the two processes are separated. In this way some constraints that limit the flexibility and robustness of the system are avoided. For instance, communication channels and transmission rates are totally independent for both modules, which is critical in decentralized systems under communication constraints.

The approach is completely decentralized, since belief estimation and decision-making are carried out without the need for a central entity. Despite the fact that a MPOMDP for the whole team is not solved (with its computational benefits), cooperative behavior still arises in two manners. First, thanks to the information shared by the different DDF modules in order to achieve a fused belief (which acts as a coordination signal for policy execution); and second, by sharing the bid values for the decentralized auction, which gives an idea about the behaviors others may be performing.

IV. ENVIRONMENTAL MONITORING

We propose an application in which there is a team of n UAVs that have to fly over a certain terrain in order to monitor a potential contamination that may appear. It is assumed that this contamination can only appear and propagate through a network of water flows on the terrain (discretized into a cell grid). Therefore, instead of surveying the whole scenario, it is assumed that the overall contamination can be controlled reliably by surveying a set of m critical points. These points are inter-connected through water flows and the contamination can propagate among them. The objective of the team of UAVs is to visit the critical points optimally to reduce the joint uncertainty on the contamination level.

Each UAV is equipped with a camera sensor pointing downwards that provides a binary observation about the contamination level of the cell in which it is located: *yes* or *no*. At each time step, each UAV can *stay* in the same cell or moving to a neighboring cell: *north*, *west*, *east* or *south*. Noisy transition functions are considered for these movement actions. Besides, when a UAV is on top of a critical point, instead of moving, it can select two additional actions (*classCont* and *classNotCont*) to classify that area as contaminated or non-contaminated, respectively.

There is a factored state with a set of variables describing the contamination level of each critical point, which can be: *none*, *low* or *high*. A graph describing the inter-connections among the critical points (due to water flows) is also known. Thus, the evolution of the contamination is modeled so that it can start at certain points of the graph (entry points), and these effects can be propagated to the other inter-connected points downstream. In addition, there is another binary state factor for each critical point to specify whether it was already classified or not. Thus, if a UAV that is on top of a critical cell takes one of the two classification actions, the corresponding variable for that critical point is set to 1. Otherwise, if there are not classification actions but the critical point was previously classified, it can switch back to 0 (not-classified) with a probability p_{des} at each time step. This is to allow the critical areas to be declassified again after some time. The local state for each UAV also includes its position on the grid.

The main objective is to reduce the uncertainty over the contamination level. This is done by monitoring the critical points and classifying them when their uncertainty is low enough. The idea is to add classification actions that reward UAVs for reaching a certain level of uncertainty regarding the contamination level. When having better information improves the task performance, the POMDP policy will try to select these information-gaining actions. Therefore, if a critical area is classified as contaminated and its state is *low* or *high*, a positive reward is given. However, if its state is *none*, the classification is wrong and a negative reward is obtained. In case the area is classified as non-contaminated the rewards are given the way around. There is no reward if the area was already classified. Otherwise, the UAVs would keep classifying all the time to obtain rewards and the policy computation would converge very slowly. The resulting policy will lead to beliefs with a low uncertainty on the contamination state, in which the UAVs are more likely to make a right classification.

The approach proposed in this paper can be used considering m single-robot behaviors, one for each possible point to monitor. Thus, the reward function for each behavior k (R_i^k) rewards UAV i only if it classifies the critical area k.

V. EXPERIMENTS

Some experiments have been performed on a simulated scenario of a real national park. The National Park of Doñana is a remarkable marshland located in the south of Spain. Due to the huge number of species that it hosts, contamination or any other natural disaster are real threats that need to be controlled. In order to survey the Park with a team of UAVs, it was divided into the 7×7 grid shown in Fig. 2, where the dark shaded cells represent non-flying zones that the UAVs cannot access for security reasons. The four key areas and the inter-connection graph shown in Fig. 2 were used to model the propagation. Moreover, it was assumed that contamination could only start at Area 1.

Each cell in the grid can be surveyed by a UAV whenever it is flying on it. On the one hand, if a UAV measures whether a critical area is contaminated or not, its sensor will detect contamination with probabilities 0.05, 0.6, 0.9, depending on whether the actual contamination level was *none*, *low* or *high*. On the other hand, when a UAV observes the contamination of a non-critical cell, it will never detect contamination. Moreover, the probability to declassify areas previously classified is set to $p_{des} = 0.04$; the reward[1] for a correct classification to 10; and the reward for an incorrect classification to -90.

TABLE I: Complexity of the POMDP models used in this paper. Number of states, actions and observations are computed for the general multi-UAV case and for a single-UAV case.

	|S|	|A|	|Z|
Monitoring (n UAVs, 4 areas)	$40^n \times 81 \times 16$	7^n	2^n
Monitoring (1 UAV, 4 areas)	$51,840$	7	2

There were 4 different behaviors, one for each critical area. A single-UAV policy was computed for each of them. with a Java version of Symbolic Perseus [2][15]. The solver ran 10 minutes for each policy in a computer with an Intel Core processor (4 cores @2.67GHz) and 8GB.

It is important to highlight that, in order to alleviate the complexity of the belief space, Mixed Observability Markov Decision Processes (MOMDPs) [16] were considered to find the policies for all the experiments in this paper. The UAVs' positions were assumed to be observable within the POMDP, which is reasonable if the sensors for self-positioning are accurate enough for the given grid resolution.

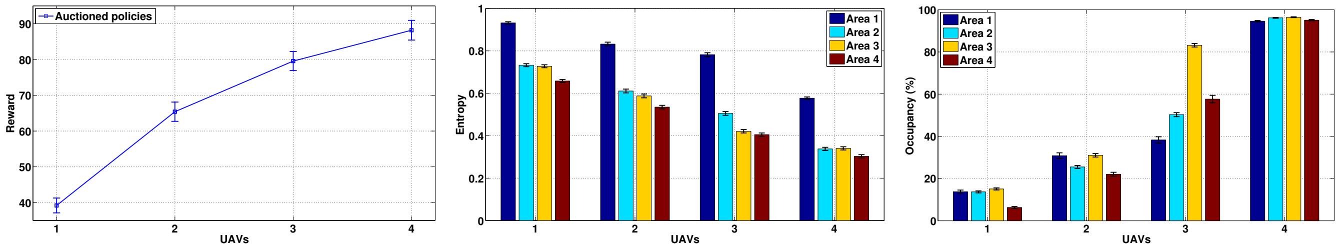
We tested our approach for teams with 1, 2, 3 and 4 UAVs, each of them running an estimation filter implementing the DDF scheme in Section II-B, and an auctioneer controller that executed the algorithm in Section III-A For each team, 1000 simulations of 100 steps were performed with the UAVs starting at random positions. Moreover, all the simulations started without contamination, but there was a probability of 0.1 that contaminated water appeared at the entry point (Area 1) at any moment. The average discounted rewards and belief entropies for all the experiments are shown in Fig. 3a and 3b, respectively. It can be seen how the addition of more UAVs improves the performance, increasing the reward of the team and decreasing the entropy of the belief on the contamination levels for each area ($\sum_{\forall level} -p_{level} \log(p_{level})$). Note that the entropy of Area 1 is always higher, since there is the uncertainty of new contamination appearing. In Fig. 3c, the percentage of time that each area is visited by any UAV is also shown. The more UAVs there are, the better they can cooperate to cover all the areas.

Since only single-UAV policies are computed in our approach, the complexities of the models do not increase with the number of UAVs (see Table I), which makes the solution scalable. However, in this scenario, experiments with more than four UAVs are not presented because they do not improve the performance significantly (four UAVs can already cover all the critical areas). Also, note that due to the interconnections between critical points, the current propagation model does not scale well with the number of areas, precluding us from testing more complex scenarios. In larger scenarios, however, sparser representations of this interdependence structure are likely, leading to more compact representations.

We also tested our approach against a joint policy for a multi-UAV POMDP. The multi-UAV POMDP is far from

[1] A small cost of 0.1 is assigned for the movement actions, whereas no cost is assigned for *stay*.

[2] The parameters for Symbolic Perseus were 5000 belief points, 1500 α-vectors maximum, 30 iterations per round and 5 rounds.

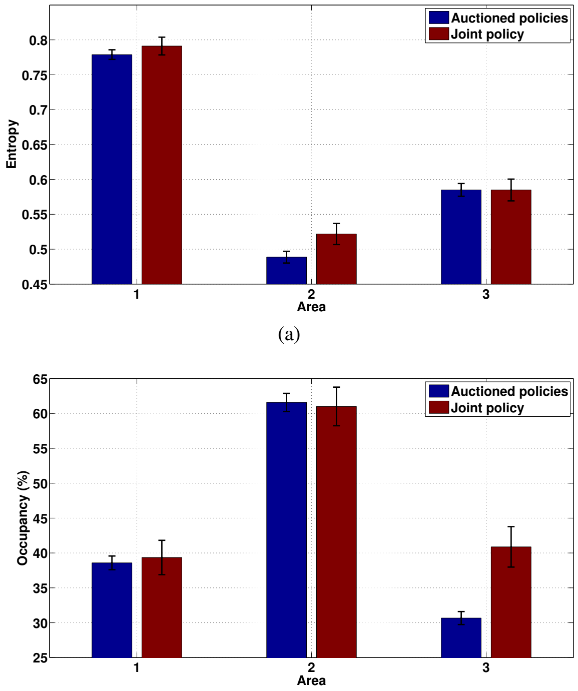


(a)

(b)

(c)

Fig. 3: Simulations with 4 critical areas. The average results are shown varying the number of UAVs involved. (a) Discounted rewards. (b) Entropies of the beliefs on the contamination levels. (a) Percentages of occupancy for each critical area.



(a)

(b)

Fig. 4: Average results for simulations on environmental monitoring with two UAVs and three critical areas. Auctioned polices are compared to a joint policy. (a) Entropies of the beliefs on the contamination levels. (b) Percentages of occupancy for each critical area.

scalable (see Table I), so we were only able to solve it for a simple case with 2 UAVs and 3 areas (Areas 1, 2 and 3)[3]. Actually, any variation of this small scenario considering more UAVs or areas, caused the same computer mentioned above to run out of memory.

We used Symbolic Perseus [4] again to compute a single-UAV policy for each behavior (5 minutes each) and a joint policy for the 2-UAV MPOMDP (14 hours). Then, we ran 1000 simulations of 100 steps (with random starting positions and no initial contamination) for our approach, and the same with the joint policy. The average values for the belief entropies and the percentage of occupancy (times visited) of each area are shown in Fig. 4. Despite the huge difference in computational time for both approaches, the results are still remarkably similar. Of course, the joint policy should be better for more complex examples, but its computation becomes intractable.

VI. CONCLUSIONS

POMDPs face a scalability problem when considering teams of UAVs, becoming intractable quite easily. For certain roled-based applications, independent POMDP-based controllers can be auctioned in a cooperative fashion. We also relax the communication guarantees by introducing a DDF approach for belief propagation, which allows for imperfect communication channels and makes the system more reliable. Although our approach is suboptimal, the results obtained in terms of cooperative behavior are still good. Moreover, since the computational complexity is reduced dramatically, it is much more scalable than other multi-robot POMDP approaches, offering a trade-off between optimality and applicability.

We present results on an environmental monitoring application that cannot be solved with the current state of the art in multi-robot POMDP solvers. Besides, there are many other multi-robot applications that can be modeled with cooperative roles and solved with our framework. In the future, more research is still necessary to evaluate the exact degradation that we suffer against optimal solutions. Also, some methods to analyze the initial problem and identify potential sets of roles would be of interest. So far, those roles are set in an ad-hoc fashion.

ACKNOWLEDGMENTS

J.C. is funded by Fundação para a Ciência e e a Tecnologia (ISR/IST pluriannual funding) through the PIDDAC Program funds and projects PEst-OE/EEI/LA0009/2011 and CMU-PT/SIA/0023/2009. M.S. is funded by the FP7 Marie Curie Actions Individual Fellowship # 275217 (FP7-PEOPLE-2010-IEF). L.M. and A.O. are funded by the FP7 EC-SAFEMOBIL Project (288082).

[3] The MPOMDP was designed to reward only one UAV at a time in case of several classifying the same area. This fostered the distribution along the different critical points.

[4] The parameters for Symbolic Perseus were 5000 belief points, 700 α-vectors maximum, 10 iterations per round and 5 rounds.

References

[1] O. Burdakov, P. Doherty, K. Holmberg, J. Kvarnstrom, and P. M. Olson, "Relay positioning for unmanned aerial vehicle surveillance," *International Journal of Robotics Research*, vol. 29, no. 8, pp. 1069–1087, 2010.

[2] L. Merino, F. Caballero, J. M. de Dios, J. Ferruz, and A. Ollero, "A cooperative perception system for multiple UAVs: Application to automatic detection of forest fires," *Journal of Field Robotics*, vol. 23, pp. 165–184, 2006.

[3] M. A. Hsieh, A. Cowley, J. F. Keller, L. Chaimowicz, B. Grocholsky, V. Kumar, C. J. Taylor, Y. Endo, R. C. Arkin, B. Jung, D. F. Wolf, G. S. Sukhatme, and D. C. MacKenzie, "Adaptive teams of autonomous aerial and ground robots for situational awareness," *Journal of Field Robotics*, vol. 24, pp. 991–1014, 2007.

[4] I. Maza, F. Caballero, J. Capitan, J. M. de Dios, and A. Ollero, "A distributed architecture for a robotic platform with aerial sensor transportation and self-deployment capabilities," *Journal of Field Robotics*, vol. 28, no. 3, pp. 303–328, 2011.

[5] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, pp. 99–134, 1998.

[6] S. Seuken and S. Zilberstein, "Formal models and algorithms for decentralized decision making under uncertainty," *Autonomous Agents and Multi-Agent Systems*, Feb. 2008.

[7] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of Markov decision processes," *Mathematics of Operations Research*, vol. 27, no. 4, pp. 819–840, 2002.

[8] R. Nair, P. Varakantham, M. Tambe, and M. Yokoo, "Networked distributed POMDPs: A synthesis of distributed constraint optimization and POMDPs," in *Proc. AAAI*, 2005.

[9] R. Nair, M. Tambe, M. Roth, and M. Yokoo, "Communication for improving policy computation in distributed POMDPs," in *Proc. AAMAS*, 2004.

[10] M. Roth, R. Simmons, and M. Veloso, "Decentralized communication strategies for coordinated multi-agent policies," in *Multi-Robot Systems: From Swarms to Intelligent Automata*, A. Schultz, L. Parker, and F. Schneider, Eds. Kluwer Academic Publishers, 2005, vol. IV.

[11] J. Capitan, L. Merino, F. Caballero, and A. Ollero, "Decentralized delayed-state information filter (DDSIF): A new approach for cooperative decentralized tracking," *Robotics and Autonomous Systems*, vol. 59, pp. 376–388, 2011.

[12] F. Bourgault and H. Durrant-Whyte, "Communication in general decentralized filters and the coordinated search strategy," in *Proc. of The 7th Int. Conf. on Information Fusion*, 2004.

[13] M. Spaan, N. Gonçalves, and J. Sequeira, "Multirobot coordination by auctioning POMDPs," in *Proc. ICRA*, 2010.

[14] R. E. Burkard, "Selected topics on assignment problems," *Discrete Applied Mathematics*, vol. 123, no. 1-3, pp. 257 – 302, 2002.

[15] P. Poupart, "Exploiting Structure to Efficiently Solve Large Scale Partially Observable Markov Decision Processes," Ph.D. dissertation, University of Toronto, 2005.

[16] S. Ong, S. W. Png, D. Hsu, and W. S. Lee, "POMDPs for Robotic Tasks with Mixed Observability," in *Proc. RSS*, 2009.