An Upper Confidence Bound for Simultaneous Exploration and Exploitation in Heterogeneous Multi-Robot Systems

Ki Myung Brian Lee[1], Felix Kong[1], Ricardo Cannizzaro[2], Jennifer L. Palmer[2],
David Johnson[3], Chanyeol Yoo[1] and Robert Fitch[1]

***Abstract*— Heterogeneous multi-robot systems are advantageous for operations in unknown environments because functionally specialised robots can gather environmental information, while others perform tasks. We define this decomposition as the *scout–task robot architecture* and show how it avoids the need to explicitly balance exploration and exploitation by permitting the system to do both simultaneously. The challenge is to guide exploration in a way that improves overall performance for time-limited tasks. We derive a novel upper confidence bound for simultaneous exploration and exploitation based on mutual information and present a general solution for scout–task coordination using decentralised Monte Carlo tree search. We evaluate the performance of our algorithms in a multi-drone surveillance scenario in which scout robots are equipped with low-resolution, long-range sensors and task robots capture detailed information using short-range sensors. The results address a new class of coordination problem for heterogeneous teams that has many practical applications.**

I. Introduction

Multi-robot systems enable flexible scaling of robotic applications by composing multiple, possibly disposable robots into a functional team that outperforms a single robot. Real-world use of multi-robot systems is increasing, for example, in warehouse management [1], agriculture [2], and defence [3]. We anticipate that heterogeneity will further accelerate the adoption of multi-robot systems because it enables increases in system capability through functional specialisation of individual robots. Specialisation is particularly appealing in applications in which the environment is partially or completely unknown. A subset of the team could focus on gathering information, while the rest perform tasks. We are interested in developing algorithms to coordinate the behaviour of heterogeneous teams that operate in unknown environments and in exploring how the division of labour between information-gathering and task-performing robots relates to the classical trade-off between exploration and exploitation.

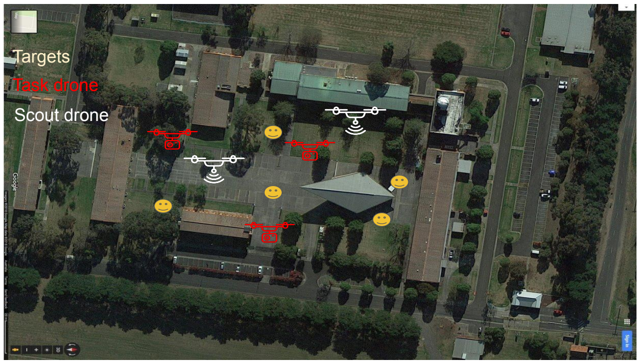


Fig. 1. An example application for multi-drone surveillance. The task is to confirm all targets (yellow) with the task drones (red). Scout drones (white) support the process by sensing targets (yellow) from a distance, at low resolution, and cueing for possible target presence.

We define a team composition in which some robots (i.e., *task robots*) are equipped to perform a particular task while others (i.e., *scout robots*) are equipped with sensors to rapidly acquire knowledge about the environment as the *scout–task robot architecture*. There are many compelling applications of this idea. For example, it may be desirable to deploy disposable scout robots to ensure safe operation of a high-value task robot, as in the case of a Mars rover-copter team [4–6]. We consider the multi-drone surveillance application illustrated in Fig. 1, where a limited number of scout robots equipped with long-range sensors cue task robots for the presence of targets.

One unexplored benefit of the scout–task robot architecture is that it can be viewed as a way to sidestep the exploration–exploitation trade-off inherent to operations in unknown environments, through heterogeneity. Whereas the trade-off between exploration and exploitation is a question of balancing these two activities, the scout–task architecture permits simultaneous exploration and exploitation. Assuming that the overall system objective is to complete as many tasks as possible in a given amount of time, the challenge is how to guide exploration in a way that is most relevant to exploitation. Scout robots should provide information about the environment that allows task robots to improve their plans and thus find higher-quality solutions. For this reason, we find that the scout–task coordination problem is fundamentally different from previous heterogeneous multi-robot coordination problems, and solutions have been proposed only in application-specific instances [4–6].

In this paper, we present a general solution to scout–task coordination. We derive a novel upper confidence bound (UCB), the *mutual-information UCB* (MI-UCB), to enable simultaneous exploration and exploitation. The MI-UCB shows that the posterior expected reward is probabilistically

This work is supported by an Australian Government Research Training Program (RTP) Scholarship, Australia's Defence Science and Technology Group, and the University of Technology Sydney.

[1] Authors are with the University of Technology Sydney, Ultimo, NSW 2006, Australia brian.lee@student.uts.edu.au, {chanyeol.yoo, felix.kong, rfitch}@uts.edu.au

[2] Authors are with the Defence Science and Technology Group, Department of Defence, Australia {ricardo.cannizzaro, jennifer.palmer}@dst.defence.gov.au

[3] Author is with Mission Systems Pty. Ltd., Sydney, Australia david.johnson@missionsystems.com.au

upper bounded by a combination of Shannon information gain and prior expected reward. It then follows from the principle of optimism under uncertainty that executing paths that maximise the MI-UCB, in fact, maximises the posterior expected reward in hindsight. We apply MI-UCB to the multi-drone surveillance problem shown in Fig. 1 using decentralised Monte Carlo tree search (Dec-MCTS) [7]. Our results show that, with the same team configuration, the hindsight reward is improved by up to 134% compared with simply maximising prior expected reward. We also demonstrate MI-UCB in a multi-drone simulation with real-time operations set in a realistic environment.

II. Related Work

Heterogeneous multi-robot coordination is often posed as a task-allocation problem in which robots have varying levels of competency in completing each task. An optimal assignment of tasks may be achieved with, e.g., markets [8], Hungarian algorithms [9], or MCTS [10]. We are interested in the scout–task robot coordination problem, where a sub-team of scout robots assists task robots in completing their tasks by exploring the environment. Exploration implicitly aids task completion in the long term. Existing work closely related to this problem includes [4, 6], which considers a Mars rover completing navigation or temporal-logic tasks, assisted by an aerial robot that scouts ahead to improve localisation accuracy or environmental knowledge. While we do not consider these problem instances in this paper, our result is sufficiently general to encompass them.

Multi-robot surveillance, in which a team of robots searches for targets [11], is typically addressed by maintaining a probabilistic belief over target locations using an occupancy grid [12] or a random finite set [13–15]. Based on the current belief, a plan is generated that maximises the expected number or probability of detections [11, 16–19] or minimises the uncertainty of target locations [14, 20, 21]. Here, the two objectives are distributed across the team, such that scout robots contribute to uncertainty reduction and task robots to detection of targets. We show that considering the two objectives in tandem improves the overall number of detections *in hindsight*, compared with considering only the expected number of detections.

The idea of combining exploration and exploitation is not new. In martingale-based approaches such as UCB on trees (UCT) [22] or KL-UCB [23], an agent repeatedly samples the reward of an action to construct statistical quantities that motivate exploration or exploitation. A trade-off between the two gradually biases the sampling toward the optimum, as in the case of MCTS [7, 24]. These approaches are parallel to ours, because it is infeasible for a physical robot to sufficiently sample reward during its operation.

A more suitable class of algorithms for robotic applications in unknown environments is Bayesian optimisation (BO). A prominent example is Gaussian process (GP) UCB [25], which has been used in robotic source seeking in plumes [26] and in human–robot interaction [27]. In BO, an agent cycles between gathering a new sample and updating its belief; and the samples are biased toward the optimum by use of a UCB derived from that belief. This is prohibitive for scout–task coordination, as all agents must necessarily contribute to both gathering measurements and maximising reward. Our proposed UCB substantially relaxes this requirement and others using information-theoretic tools.

A theoretical result closely related to our work is [28, Lemma 3], which shows that, if a UCB similar to ours were to hold, using the UCB as an acquisition function for action selection leads to bounded regret. We show that a similar UCB generally holds.

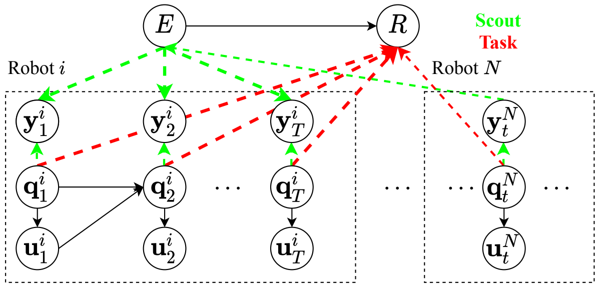


Fig. 2. A probabilistic graphical model illustrating the scout–task coordination problem. Dashed connections depend on team composition. The control input $\mathbf{u}$ generates trajectory $\mathbf{q}$. A task robot (red) gains a reward $R(\mathbf{q}, E)$, depending on the trajectory $\mathbf{q}$ and the latent environment E. A scout robot (green) gathers measurements $\mathbf{y}_t$ at each state $\mathbf{q}_t$, revealing information about E.

III. Scout-Task Coordination Problem

We formulate the scout–task coordination problem, illustrated in Fig. 2, as follows. Consider a team of N mobile robots, the dynamics of which are described by:

$$\mathbf{q}_{t+1}^r = \mathbf{f}(\mathbf{q}_t^r, \mathbf{u}_t^r), \tag{1}$$

where $\mathbf{q}_t^r$ and $\mathbf{u}_t^r$ are the state and control action of robot r at time t, respectively. The superscript $1 \leq r \leq N$ denotes the robot, the subscript t denotes time, and $\mathbf{u}_t^r$ is the control action applied to robot r at time t.

The robots operate in an unknown environment, denoted by E, which can follow any distribution. For example, it may be discrete and follow a categorical distribution; or it may be continuous and follow a Gaussian distribution. Each robot may belong to a set of scout robots $\mathcal{S} \subset [1, ..., N]$ or to a set of task robots $\mathcal{T} \subset [1, ..., N]$. $\mathcal{S}$ and $\mathcal{T}$ are not necessarily disjoint; and thus a robot may belong to both sets (i.e., it may be a scout-and-task robot).

If $r \in \mathcal{S}$, robot r generates measurements $\mathbf{y}_t^r$ that reveal information about E:

$$\mathbf{y}_t^r \sim \mathcal{P}(\mathbf{y}_t^r \mid \mathbf{q}_t^r, E). \tag{2}$$

If $r \in \mathcal{T}$, robot r is equipped with a payload to perform an intended task, which depends on the environment E. We thus model the task completion by a deterministic reward function $R(\mathbf{q}^{\mathcal{T}}, E)$. It is important to note that scout-only robots $r \in \mathcal{S} \setminus \mathcal{T}$ do not contribute directly to the reward function; instead they allow the task and scout-and-task robots $r \in \mathcal{T}$ to more effectively complete their tasks by gathering information on E.

For brevity, we omit the subscript (resp. the superscript) to mean the set of states over time (resp. different robots). I.e., $\mathbf{q}^r = \{\mathbf{q}^r_1, ..., \mathbf{q}^r_T\}$, $\mathbf{q}_t = \{\mathbf{q}^1_t, ..., \mathbf{q}^N_t\}$. The omission of both subscript and superscript indicates the set of all robots' trajectories over time: i.e., $\mathbf{q} = \{\mathbf{q}^1, ..., \mathbf{q}^N\} = \{\mathbf{q}_1, ..., \mathbf{q}_T\}$. We also replace the superscript with $\mathcal{S}$ and $\mathcal{T}$ to mean the set of poses or trajectories of robots that belong to the set of scout or task robots, respectively. Hence, $\mathbf{q}^\mathcal{T} = \{\mathbf{q}^i \mid i \in \mathcal{T}\}$. Further, we write $\mathbf{q}^\mathbf{r}(\mathbf{u}^r)$ to mean the trajectory obtained by applying $\mathbf{u}^r$ to robot r; and likewise $\mathbf{q}(\mathbf{u})$ denotes the set of all robots' trajectories obtained by applying the control sequences $\mathbf{u} = \{\mathbf{u}^1, \ldots, \mathbf{u}^r\}$ to the corresponding robots. Similarly, we write $\mathbf{y}^r(\mathbf{q}^r)$ to mean the set of observations obtained from the trajectory of robot r.

Our aim is to choose control inputs $\mathbf{u}$ maximising reward:

$$\mathbf{u}^* = \operatorname*{argmax}_{\mathbf{u}^\mathcal{T}} R(E, \mathbf{q}^\mathcal{T}(\mathbf{u}^\mathcal{T})). \tag{3}$$

Clearly, (3) cannot be solved directly because E is a random variable and, consequently, so is $R(E, \mathbf{q}^\mathcal{T}(\mathbf{u}^\mathcal{T}))$. This is addressed in Sec. IV.

IV. MI-UCB for Coordination

In this section, we derive a surrogate acquisition function, MI-UCB, by analysing the effect of improvement in environmental knowledge on the estimated reward. Then, we present an online planning framework for scout–task coordination.

A. Mutual-Information Upper Confidence Bound (MI-UCB)

We cannot solve (3) directly as E and $R(\mathbf{q}(\mathbf{u}), E)$ are random variables that must be estimated. A common approach is to pose an *expectimax* problem based on current belief:

$$\mathbf{u}^* = \operatorname*{argmax}_{\mathbf{u}} \mathop{\mathbb{E}}_{E\sim\mathcal{P}(E)} R(E, \mathbf{q}^\mathcal{T}(\mathbf{u}^\mathcal{T})). \tag{4}$$

However, this does not show how, or why, the scout robots can coordinate because neither the measurements $\mathbf{y}^\mathcal{S}$, nor the state or control, $\mathbf{q}^\mathcal{S}$ or $\mathbf{u}^\mathcal{S}$, respectively, appear in (4).

To consider the effect of measurements obtained by the scout robots, we consider maximising the posterior expected reward given measurements:

$$\mathbf{u}^* = \operatorname*{argmax}_{\mathbf{u}} \mathop{\mathbb{E}}_{E\sim\mathcal{P}(E|\mathbf{y}^\mathcal{S})} R(E, \mathbf{q}^\mathcal{T}(\mathbf{u}^\mathcal{T})). \tag{5}$$

While (5) incorporates the effect of measurements, we cannot solve it directly because the measurements $\mathbf{y}^\mathcal{S}$ remain random variables that have not been sampled. One may take an expectation over possible measurements similar to a partially observable Markov decision process approach [29], but the fundamental challenge of enumerating possible measurements over possible trajectories remains.

Our finding is that we can solve (5) using the *principle of optimism under uncertainty*, by deriving a UCB on the posterior expected reward. The exact statement is:

Theorem 1 (MI-UCB)**.** *Suppose $R(E, \mathbf{q}^\mathcal{T})$ is a measurable function of E for all $\mathbf{q}^\mathcal{T}$. With probability $\geq 1 - \delta$:*

$$\begin{aligned}\mathop{\mathbb{E}}_{E\sim\mathcal{P}(E|\mathbf{y}^S)}&[R(E, \mathbf{q}^\mathcal{T})] \leq \\ &\frac{1}{\delta} I(\mathbf{y}^S; E) + \log \mathbb{E}[\exp R(E, \mathbf{q}^\mathcal{T})].\end{aligned} \tag{6}$$

Most importantly, the UCB on the RHS of (6) *decouples* scout- and task-only robots because it separates information gain and reward into a weighted sum. The term $\log \mathbb{E}[\exp R(E, \mathbf{q}^\mathcal{T}(\mathbf{u}^\mathcal{T}))]$ is called the cumulant generating function (CGF) in probability theory. for which analytical expressions are often available. It is also important to note that the posterior expected reward on the left-hand side of (6) cannot be calculated before taking the measurement, while the UCB on the right-hand side may be calculated beforehand.

Theorem 1 is proved by evaluating how the change from $\mathcal{P}(E)$ to $\mathcal{P}(E \mid \mathbf{y}^S)$ affects the estimated reward, as captured by the seminal result of Donsker & Varadarhan [30–32]:

Lemma 1 (Change of measure inequality)**.** Given any measurable function ϕ on X and any two distributions $\mathcal{P}$ and $\mathcal{Q}$ on X, we have:

$$\mathop{\mathbb{E}}_{x\sim P}[\phi(x)] \leq D_{KL}(\mathcal{P} \mid \mathcal{Q}) + \log \mathop{\mathbb{E}}_{x\sim Q}[\exp \phi(x)]. \tag{7}$$

Proof of Theorem 1. Consider the change of measure inequality between $\mathcal{P}(E \mid \mathbf{y}^S)$ and $\mathcal{P}(E)$:

$$\begin{aligned}\mathop{\mathbb{E}}_{E\sim\mathcal{P}(E|\mathbf{y}^S)}[R(E, \mathbf{q}^\mathcal{T})] \leq& D_{KL}(\mathcal{P}(E \mid \mathbf{y}^S) \mid \mathcal{P}(E)) \\ &+ \log \mathbb{E}[\exp R(E, \mathbf{q}^\mathcal{T})].\end{aligned} \tag{8}$$

Applying Markov's inequality over $\mathbf{y}^\mathcal{S}$ to the KL divergence term yields the claimed result. □

B. Online Planning

Based on the UCB in Theorem 1, our online planning framework solves the following surrogate problem:

$$\mathbf{u}^* = \operatorname*{argmax}_{\mathbf{u}} I(E; \mathbf{y}^\mathcal{S}) + \delta \log \mathbb{E} \exp R(E, \mathbf{q}^\mathcal{T}(\mathbf{u}^\mathcal{T})). \tag{9}$$

The principle of optimism under uncertainty [33] asserts that maximising the UCB (9) maximises the reward function (5) when evaluated in hindsight. This is known as a *no-regret* bound.

The online-planning framework cycles between updating the belief $\mathcal{P}(E)$ and maximising the MI-UCB (9). For this purpose, we use Dec-MCTS [7], a decentralised multi-robot planning algorithm that extends the well-known MCTS. As we do not modify Dec-MCTS except the objective function, we only give a brief description – interested readers are referred to [7].

In Dec-MCTS, each robot maintains a probability distribution over the control sequences $\mathbf{u}$ of the entire team. Other robots' distributions are updated asynchronously via communication, while each robot's own distribution is updated via single-robot MCTS iterations. The single-robot MCTS iterations generate rollout trajectories for a single robot, and evaluates the objective function (9) with other robots' trajectories fixed at a random sample drawn from the probability distribution maintained.

Hence, the combination of MI-UCB and Dec-MCTS can solve any instance of the scout-task coordination problem in a decentralised manner as long as the MI-UCB (9) can be computed given the control sequences $\mathbf{u}$ of the entire team.

As distribution updates are asynchronous by design, Dec-MCTS is robust against delays, and thus permits multi-hop communication.

V. Application in Multi-Drone Surveillance

We apply MI-UCB to a multi-drone surveillance problem in which the task is to maximise the number of confirmations of an unknown number of targets at unknown locations. Task drones are equipped with short-range sensors only, and scout drones are equipped with long-range sensors that can rapidly provide knowledge about the environment. Drones may also be dual-equipped (i.e., they may be scout-and-task drones).

A. Reward Function

We represent the targets in a 2D occupancy grid, so that the environment is a Boolean matrix $E \in \mathbb{B}^{N_X \times N_Y}$, where N_X and N_Y are the number of cells in the X and Y directions, respectively. $E(i,j) = 1$ means cell (i,j) is occupied by a target, and 0 indicates otherwise.

We model the visibility of cell (i,j) from robot r at time t as a Bernoulli random variable $v_t^r(i,j;\mathbf{q}_t^r) \in \mathbb{B}$:

$$v_t^r(i,j;\mathbf{q}_t^r) \sim \mathcal{P}(v_t^r(i,j) \mid \mathbf{q}_t^r). \tag{10}$$

The visibility over a trajectory $\mathbf{q}^r$ is a disjunction $v^r(i,j;\mathbf{q}^r) = \vee_t v_t^r(i,j;\mathbf{q}_t^r)$. Similarly for the visibility over different robots, $v(i,j;\mathbf{q}) = \vee_r v^r(i,j;\mathbf{q}^r)$. Robot $r \in \mathcal{T}$ confirms a target at cell (i,j) iff the target exists and may be sensed. The reward is the number of targets confirmed:

$$R(\mathbf{q}^{\mathcal{T}}(\mathbf{u}^{\mathcal{T}}), E) = \sum_{ij} (v_t^r(i,j;\mathbf{q}^{\mathcal{T}}(\mathbf{u}^{\mathcal{T}}))E(i,j)). \tag{11}$$

The reward function (11) is a sum of Bernoulli random variables, which is in turn a Poisson binomial random variable. Its CGF is given by:

$$\begin{aligned}\log \mathop{\mathbb{E}}_{E \sim \mathcal{P}(E)} &\exp R(\mathbf{q}^{\mathcal{T}}(\mathbf{u}^{\mathcal{T}}), E) \\ &= \sum_{ij} \log(1 + \mathcal{P}(d(i,j;\mathbf{q}^{\mathcal{T}}(\mathbf{u}^{\mathcal{T}})))(e-1)),\end{aligned} \tag{12}$$

where $\mathcal{P}(d(i,j;\mathbf{q}^{\mathcal{T}}(\mathbf{u}^{\mathcal{T}}))) = \mathcal{P}(v(i,j;\mathbf{q}^{\mathcal{T}}(\mathbf{u}^{\mathcal{T}})))\mathcal{P}(E(i,j))$.

B. Belief Update and Information Gain

We use a simple grid-based filter for decentralised data fusion of E. With the standard independence assumption, the belief over target occupancy decomposes as:

$$\mathcal{P}(E) = \prod_{i,j} \mathcal{P}(E(i,j)). \tag{13}$$

When a target is visible, a scout robot can measure its position. We adopt the inverse sensor model [34] approach to discretise the measurements and represent the measurement as a matrix of Bernoulli random variables:

$$\begin{aligned}\mathcal{P}(\mathbf{y}_t^r(i,j) \mid E(i,j), \mathbf{q}_t^r) &= \\ v_t^r(i,j;\mathbf{q}_t^r)&\mathcal{P}(\mathbf{y}_t^r(i,j) \mid E(i,j)).\end{aligned} \tag{14}$$

The sensor model $\mathcal{P}(\mathbf{y}_t^r(i,j) \mid E(i,j))$ is given by a confusion matrix between true and measured occupancy.

Each scout robot communicates its position and detected target locations (if any) at regular intervals. Measurements are fused with Bayes' rule:

$$\begin{aligned}\mathcal{P}(E(i,j) \mid \mathbf{y}_{1:t}^r(i,j)) = \Big(&(1 - v_t^r(i,j;q_t^r)) \\ &+ v_t^r(i,j;q_i^r)\frac{\mathcal{P}(\mathbf{y}_t^r(i,j) \mid E(i,j), \mathbf{y}_{1:t-1}^r(i,j)}{\mathcal{P}(\mathbf{y}_t^r(i,j) \mid \mathbf{y}_{1:t-1}^r(i,j)}\Big) \\ &\times \mathcal{P}(E(i,j) \mid \mathbf{y}_{1:t-1}^r(i,j)).\end{aligned} \tag{15}$$

Information gain may be calculated as follows. For each cell, the information gain is:

$$\begin{aligned}I(E(i,j);\mathbf{y}_t^r(i,j))) &= \\ H(\mathcal{P}(\mathbf{y}_t^r(i,j))) &- \mathop{\mathbb{E}}_{E(i,j)} H(\mathcal{P}(\mathbf{y}_t^r(i,j) \mid E(i,j))),\end{aligned} \tag{16}$$

where $H(p)$ is binary entropy and $\mathcal{P}(\mathbf{y}_t^r(i,j)) = \mathbb{E}_{E(i,j)}\,\mathcal{P}(\mathbf{y}_t^r(i,j) \mid E(i,j))$. The information gain is summed over the visible region:

$$I(E;\mathbf{y}_t^r) = \sum_{ij} v(i,j;\mathbf{q}(\mathbf{u}))I(E(i,j);\mathbf{y}_t^r(i,j))). \tag{17}$$

VI. Results

We analyse the performance of MI-UCB in the context of the multi-drone surveillance problem. We first compare its performance in terms of ground-truth reward with that of a conventional expectimax approach in a simplified simulation. We then demonstrate the framework in two realistic simulated environments to examine the behaviour of MI-UCB in practical applications.

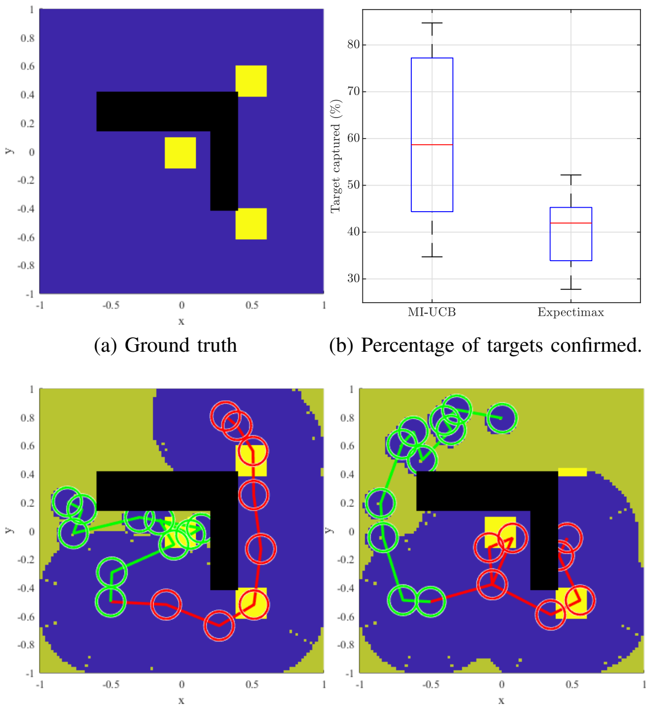
A. Comparison with Expectimax

We first compare the MI-UCB approach with the standard expectimax approach. Here, expectimax refers to maximising the expected reward, given the current belief at each stage, without accounting for information gain.

The comparison is set in the environment shown in Fig. 3a, where known obstacles and unknown targets are shown in black and yellow, respectively. The task is to confirm targets within a given radius of a robot representing its task sensors' field of view. A scout robot may also reveal knowledge about the environment using longer-range sensors.

There are two robots: red and green. To make the comparison fair, the red robot is a scout-and-task robot, while the green robot is task-only. Thus, both are task robots, so the expectimax approach can generate a meaningful plan for each. If one robot were to be scout-only, the expectimax approach would not generate a plan for it, unlike MI-UCB (Sec. IV). Intuitively, the expectimax approach simply reacts to the updates in belief, while MI-UCB accounts for information gain associated with the belief update.

The robots start with a uniform prior; and the robots' trajectory length for each time step is fixed at 2.5 m. Each robot updates its environment belief after executing one time step and re-plans its trajectory. We measure the performance in terms of the fraction of targets confirmed in simulation runs with the MI-UCB or expectimax approach generating



(a) Ground truth (b) Percentage of targets confirmed.

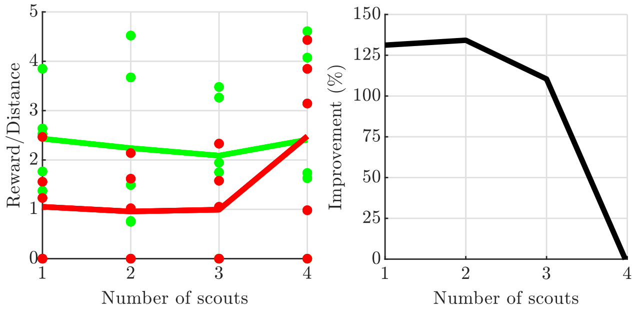
(c) MI-UCB (d) Expectimax

Fig. 3. Comparison between MI-UCB and expectimax in a simplified scenario. A robot confirms a target (yellow) if it is within the circle representing its target sensors' field of view (red for a scout-and-task drone equipped with target and long-range sensors and green for a task-only drone, equipped only with task sensors). The colourmap shows the belief on target occupancy (increasing from blue to yellow), while black areas indicate obstacles and yellow-green areas are yet to be explored. MI-UCB (c) outperforms expectimax (d), because the former accounts for the fact that the red robot can provide greater information gain than the green robot.

the robots' trajectories, while randomising the environment by placing (three) targets in different locations for each run.

Combined results for ten runs of each simulation, provided in Fig. 3b, demonstrate that the MI-UCB approach outperforms the expectimax approach by $\sim 50\%$ in terms of the median fraction of targets confirmed. Examples that illustrate this trend are shown in Figs. 3c and 3d. In Fig. 3c, it may be observed that MI-UCB causes the red scout-and-task robot to (in effect) 'delegate' the task of confirming the target in the centre of the environment to the green task-only drone, unlike the expectimax approach used to generate the results shown in Fig. 3d. This is because information gain is considered in MI-UCB. Therefore, the team can maximise its utility if the red robot continues to explore and gather information, while the green robot confirms the target. In contrast, in the expectimax approach, there is no incentive to do so; and the red scout-and-task robot, being closer, confirms the target in the centre. The higher variance of the MI-UCB approach is attributed to its optimism and shows that the upper limit of attainable reward is increased by valuing exploration.

We verify this trend in comparative performance with a four-robot experiment. As before, all drones are task drones, but the number of scout-and-task drones is varied from one to four. We fix the number of time steps and evaluate the reward per unit distance travelled by the drones in five runs for each approach, with five randomly placed targets.



(a) Reward per unit distance (b) Performance improvement

Fig. 4. (a) Reward obtained with MI-UCB (green) and expectimax (red) as a function of the number of scouts in the four-robot scenario. (b) MI-UCB provides the most benefit compared with expectimax with two scouts, an improvement of 134%, and converges to equivalence with expectimax when four scouts are used.

As illustrated in Fig. 4, when there are *fewer* scout drones, MI-UCB provides greater performance benefit compared with expectimax; however, when all drones are scout-and-task drones, the approaches yield identical rewards. This implies that MI-UCB makes better use of *limited* information to provide consistent reward values with varying team composition. The performance improvement plateaus with two scouts, as shown in Fig. 4b, motivating the question of what team composition is *optimal* for a given problem.

B. Practical Demonstrations

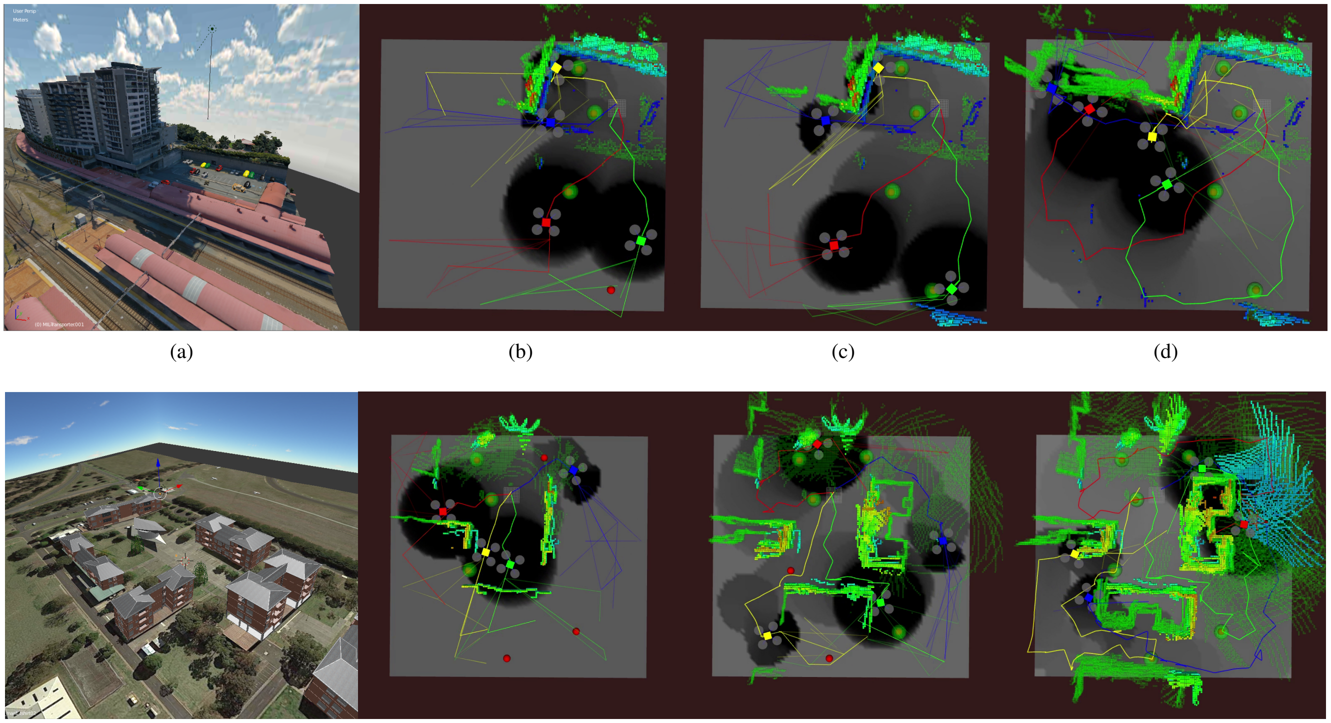
We demonstrate multi-drone surveillance in two realistic simulation environments. To examine practical efficacy, we perturb the problem from the ideal by varying the belief over time, introducing obstacles with simultaneous mapping, and emulating sensor failure.

We implement the multi-drone surveillance framework in performant software based on the Robot Operating System [35]. Each drone builds its own map using Real-Time Appearance-Based Mapping [36]; and the maps are combined to achieve inter-robot localisation and decentralised mapping. We defer the description of the mapping framework to a future work. The grid-based filter for target estimation is implemented by use of the grid_map library [37].

The quadrotor simulation is based on the PX4 software-in-the-loop simulation [38], coupled with a modified version of the Modular Open Robotics Simulation Engine [39]. The simulation is distributed over two desktop computers, each equipped with an NVIDIA RTX2060 graphics card. The computations for each drone are executed in real-time on an NVIDIA Jetson AGX single-board computer.

We first demonstrate the framework with four drones in the environment pictured in Fig. 5a. It is based on an urban area near Roma St. Station, Brisbane, Australia. The 3D model is generated by use of high-altitude photogrammetry and contains a mixture of open and cluttered terrain.

Figs. 5b–5d show the target occupancy belief held by the green scout-and-task drone, as well as its intent. A target is identified at the start of the operation and is confirmed by the blue task drone (Fig. 5b). The yellow task drone



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 5. Simulation results in the Brisbane (a) and Point Cook (e) environments. Time runs horizontally. Red spheres are targets yet to be confirmed. They turn green after confirmation. The grey colourmap shows the robot's belief over target occupancy.

confirms another target in the upper-middle portion of the environment, where a parking lot is located. This behaviour is consistent with the simplified simulation, in which, when cued by a scout drone, a task drone undertakes the task instead, distributing the exploration–exploitation workload.

Throughout the mission, the blue and yellow (task) drones focus on smaller, geometrically complex areas (around the tall building and the parking lot), while the red and green scout-and-task drones jointly cover a larger area above the train tracks. This demonstrates MI-UCB's inheritance from and generalisation of heterogeneous information gathering.

The difference between the two approaches is that MI-UCB results in all targets eventually being confirmed, thus completing the intended task. A heterogeneous information-gathering approach would accept the long-range sensor's coverage of a target and not require that a short-range sensor capture it. On the other hand, our decision-making under uncertainty approach allows the practitioner to specify that confirmation by a task drone is imperative. This provides great flexibility, as one can easily replace the task of visual confirmation with, e.g., payload delivery or casualty evacuation, each of which requires proximity.

We also demonstrate the framework in the environment shown in Fig. 5e, which is modelled on RAAF Base Point Cook, Australia. The environment is prepared by modelling the buildings from a satellite image. It creates an interesting scenario for low-altitude operations because of the alleyways and corners that limit full visibility. In this simulation, we emulate perception failures to examine their effect on the performance of the algorithm. For example, in Fig. 5f, the blue scout-and-task drone has confirmed a target, but that is not reported to the other drones; while in Fig. 5g, the same occurs with yellow drone. Despite these unmodelled failures, the algorithm adapts to the change and successfully confirms all targets eventually, as illustrated in Fig. 5h.

VII. Conclusion

We presented a framework for coordinating a scout–task robot team undertaking exploration and exploitation simultaneously and synergistically. This behaviour is enabled by MI-UCB, a novel upper confidence bound that leads to increased task performance in hindsight compared to simply maximising the expected reward given the current belief. The generality and simplicity of MI-UCB motivates not only complex problems such as temporal-logic synthesis [6, 40, 41], but also new fundamental questions in multi-robot coordination. Given a problem instance, can we postulate an optimal composition of scout and task robots? Can the composition be adapted dynamically depending on the task at hand? These types of coordination problems have many practical applications in areas such agriculture, infrastructure monitoring, construction, marine robotics, and others where there is value in collecting detailed observations of objects of interest that are distributed within the environment at unknown positions.

References

[1] P. R. Wurman, R. D'Andrea, and M. Mountz, "Coordinating hundreds of cooperative, autonomous vehicles in warehouses," *AI Mag.*, vol. 29, no. 1, pp. 9–9, 2008.

[2] G. D'Urso, S. L. Smith, R. Mettu, T. Oksanen, and R. Fitch, "Multi-vehicle refill scheduling with queueing," *Comput. and Electron. in Agriculture*, vol. 144, pp. 44 – 57, 2018.

[3] R. Gorrell, "Countering A2/AD with swarming," Ph.D. dissertation, AIR UNIVERSITY, 2016.

[4] T. Sasaki, K. Otsu, R. Thakker, S. Haesaert, and A. Agha-mohammadi, "Where to map? Iterative rover-copter path planning for Mars exploration," *Robot. and Automat. Lett.*, vol. 5, no. 2, pp. 2123–2130, 2020.

[5] K. Ebadi and A. Agha-Mohammadi, "Rover localization in Mars helicopter aerial maps: Experimental results in a Mars-analogue environment," in *Proc. of Int. Symp. on Exp. Robot. (ISER)*, 2018.

[6] S. Bharadwaj, M. Ahmadi, T. Tanaka, and U. Topcu, "Transfer entropy in MDPs with temporal logic specifications," in *Proc. of Conf. on Decis. and Contr. (CDC)*, 2018, pp. 4173–4180.

[7] G. Best, O. M. Cliff, T. Patten, R. R. Mettu, and R. Fitch, "Dec-MCTS: Decentralized planning for multi-robot active perception," *The Int. J. of Robot. Res.*, vol. 38, no. 2-3, pp. 316–337, 2019.

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

[9] G. A. Korsah, A. Stentz, and M. B. Dias, "A comprehensive taxonomy for multi-robot task allocation," *The Int. J. of Robot. Res.*, vol. 32, no. 12, pp. 1495–1512, 2013.

[10] A. J. Smith, G. Best, J. Yu, and G. A. Hollinger, "Real-time distributed non-myopic task selection for heterogeneous robotic teams," *Auton. Robots*, vol. 43, no. 3, pp. 789–811, 2019.

[11] T. H. Chung, G. A. Hollinger, and V. Isler, "Search and pursuit-evasion in mobile robotics," *Auton. Robots*, vol. 31, no. 4, p. 299, 2011.

[12] R. Vidal, O. Shakernia, H. J. Kim, D. H. Shim, and S. Sastry, "Probabilistic pursuit-evasion games: Theory, implementation, and experimental evaluation," *Trans. on Robot. and Automat.*, vol. 18, no. 5, pp. 662–669, 2002.

[13] D. E. Clark, K. Panta, and B.-N. Vo, "The GM-PHD filter multiple target tracker," in *Int. Conf. on Info. Fusion*. IEEE, 2006, pp. 1–8.

[14] P. Dames, P. Tokekar, and V. Kumar, "Detecting, localizing, and tracking an unknown number of moving targets using a team of mobile robots," *The Int. J. of Robot. Res.*, vol. 36, no. 13-14, pp. 1540–1553, 2017.

[15] Y. Sung and P. Tokekar, "Algorithm for searching and tracking an unknown and varying number of mobile targets using a limited FoV sensor," in *Proc. of IEEE ICRA*, 2017, pp. 6246–6252.

[16] R. Vidal, S. Rashid, C. Sharp, O. Shakernia, Jin Kim, and S. Sastry, "Pursuit-evasion games with unmanned ground and aerial vehicles," in *Proc. of IEEE ICRA*, vol. 3, 2001, pp. 2948–2955 vol.3.

[17] F. Bourgault, T. Furukawa, and H. Durrant-Whyte, "Optimal search for a lost target in a Bayesian world," *Springer Tracts in Adv. Robot.*, vol. 24, pp. 209–222, 01 2003.

[18] G. Hollinger, S. Singh, J. Djugash, and A. Kehagias, "Efficient multi-robot search for a moving target," *The Int. J. of Robot. Res.*, vol. 28, no. 2, pp. 201–219, 2009.

[19] S. K. Gan, R. Fitch, and S. Sukkarieh, "Real-time decentralized search with inter-agent collision avoidance," in *Proc. of IEEE ICRA*, 2012, pp. 504–510.

[20] Y. Kantaros, B. Schlotfeldt, N. Atanasov, and G. J. Pappas, "Asymptotically optimal planning for non-myopic multi-robot information gathering," in *Proc. of RSS*, FreiburgimBreisgau, Germany, June 2019.

[24] D. Silver and J. Veness, "Monte-Carlo planning in large POMDPs," in *Adv. in Neural Info. Process. Syst.*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., vol. 23. Curran Associates, Inc., 2010.

[21] O. M. Cliff, D. L. Saunders, and R. Fitch, "Robotic ecology: Tracking small dynamic animals with an autonomous aerial vehicle," *Sci. Robot.*, vol. 3, 2018.

[22] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, pp. 235–256, 05 2002.

[23] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *Proc. of Conf. Comput. Learn. Theory*, 2011, pp. 359–376.

[25] N. Srinivas, A. Krause, S. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," in *Proc. of Int. Conf. on Mach. Learn.*, 07 2010, pp. 1015–1022.

[26] K. M. B. Lee, J. J. H. Lee, C. Yoo, B. Hollings, and R. Fitch, "Active perception for plume source localisation with underwater gliders," in *Australas. Conf. on Robot. and Automat. (ACRA)*, 2018.

[27] H. Ahn, Y. Oh, S. Choi, C. J. Tomlin, and S. Oh, "Online learning to approach a person with no regret," *Robot. and Automat. Lett.*, vol. 3, no. 1, pp. 52–59, 2018.

[28] X. Lu and B. V. Roy, "Information-theoretic confidence bounds for reinforcement learning," in *Adv. Neural Info. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[29] A. Somani, N. Ye, D. Hsu, and W. S. Lee, "DESPOT: Online POMDP planning with regularization," in *Adv. in Neural Info. Process. Syst.*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013.

[30] M. D. Donsker and S. S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time," *Commun. on Pure and Appl. Math.*, vol. 36, p. 183–212, 1983.

[31] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley, 1997.

[32] Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer, "PAC-Bayesian inequalities for martingales," *IEEE Trans. on Info. Theory*, vol. 58, no. 12, pp. 7086–7093, 2012.

[33] R. I. Brafman and M. Tennenholtz, "R-MAX: A general polynomial time algorithm for near-optimal reinforcement learning," *J. of Mach. Learn. Res.*, vol. 3, no. Oct, pp. 213–231, 2002.

[34] S. Thrun, "Learning occupancy grids with forward models," in *Proc. of IROS*, vol. 3. IEEE, 2001, pp. 1676–1681.

[35] Stanford Artificial Intelligence Laboratory et al., "Robotic operating system." [Online]. Available: https://www.ros.org

[36] M. Labbé and F. Michaud, "RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation," *J. of Field Robot.*, vol. 36, 10 2018.

[37] P. Fankhauser and M. Hutter, "A universal grid map library: Implementation and use case for rough terrain navigation," in *Robot Operating System (ROS) – The Complete Reference*, A. Koubaa, Ed. Springer, 2016, vol. 1, ch. 5.

[38] F. Furrer, M. Burri, M. Achtelik, and R. Siegwart, "RotorS—a modular Gazebo MAV simulator framework," in *Robot Operating System (ROS): The Complete Reference*, A. Koubaa, Ed. Springer, 2016, vol. 1, pp. 595–625.

[39] Mission Systems Pty. Ltd, "MORSE: the modular open robots simulator engine." [Online]. Available: https://github.com/mission-systems-pty-ltd/morse

[40] C. Yoo and C. Belta, "Control with probabilistic signal temporal logic," *arXiv preprint arXiv:1510.08474*, 2015.

[41] K. M. B. Lee, C. Yoo, and R. Fitch, "Signal temporal logic synthesis as probabilistic inference," in *Proc. of IEEE ICRA*, 2021.