2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

Active Reward Learning for Co-Robotic Vision Based Exploration in Bandwidth Limited Environments

Stewart Jamieson[1,2], Jonathan P. How[2], Yogesh Girdhar[3]

***Abstract*— We present a novel POMDP problem formulation for a robot that must autonomously decide where to go to collect new and scientifically relevant images given a limited ability to communicate with its human operator. From this formulation we derive constraints and design principles for the observation model, reward model, and communication strategy of such a robot, exploring techniques to deal with the very high-dimensional observation space and scarcity of relevant training data. We introduce a novel active reward learning strategy based on making queries to help the robot minimize path "regret" online, and evaluate it for suitability in autonomous visual exploration through simulations. We demonstrate that, in some bandwidth-limited environments, this novel regret-based criterion enables the robotic explorer to collect up to 17% more reward per mission than the next-best criterion.**

I. Introduction

Images of exotic biological and geological phenomena from remote and dangerous locations have tremendous scientific value but are extraordinarily challenging and costly to collect. Robots have been at the forefront of collecting visual scientific observations in such environments, which include Mars [1], deep space [2], the Earth's oceans [3]–[5], and under Arctic ice sheets [6]. Communication bandwidth constraints are perhaps the biggest bottleneck to exploration in these remote environments [7], [8]. As such, common current approaches to autonomous exploration are to either deploy the vehicles on a predefined path or to deploy them with adaptive path plans based on tracking low-dimensional observations from some other sensor. This paper proposes a novel approach to vision-guided exploration using a human-robot team that is effective even in the presence of strong bandwidth constraints such as those imposed by acoustic underwater communications [7].

Most recent progress towards increasing the science return of autonomous exploration missions has been made through enabling "opportunistic science"[1] as well as addressing challenges in navigation, task planning and scheduling autonomy [9]–[11]. The progress in adaptive sampling and exploration algorithms for robots has primarily focused on observing spatially-varying *scalar* quantities, such as temperature [12], [13]. To reach a future of efficient robotic explorers in remote environments, robots will need to autonomously: recognize visual phenomena that might be scientifically interesting, transmit images of them to scientists for clarification as needed, model where more of them might be found, and plan a trajectory accordingly (Figure 1).

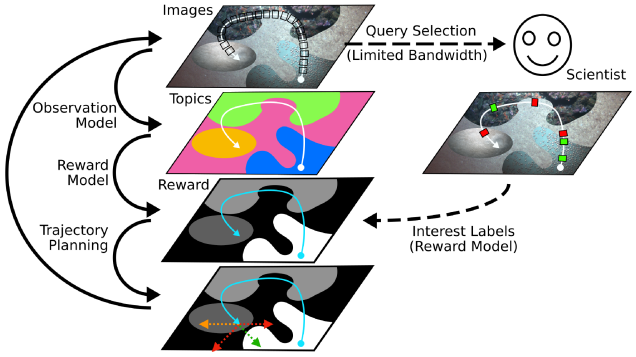


Fig. 1: Proposed approach to co-robotic exploration that models the interest of the operator over a low bandwidth communication channel and uses the learned reward model to plan the most rewarding (in terms of interest) robot paths.

The primary contributions of this work are a partially observable Markov decision process (POMDP) formulation for vision-based scientific exploration and a solution that is generalizable to many environments. Note that we explicitly constrain the focus of this work to dealing with high-dimensional observation spaces when solving the POMDP. The proposed exploration approach uses the limited communication bandwidth to query the operator for the *value* of representative images and uses the responses to learn an *interest function* that informs the robot about the value of an exploration path. The proposed approach is suitable for deployment in completely unknown environments, and it can use (but does not require) prior knowledge about the environment and the phenomena being observed. Our final contribution is an analysis and comparison of active learning decision criteria that a robot could use for deciding which observations to send to the operator. That analysis is supported by simulations of a scientific exploration task using both real and artificial data.

II. Related Work

This work contributes to the field of *autonomous science*; previous works in this area include AEGIS [1] and OASIS [11], which enabled robots to opportunistically recognize

*This work was supported by NSF-NRI Award Number 1734400

[1]S. Jamieson is with the MIT-WHOI Joint Program in Applied Ocean Science and Engineering sjamieson@whoi.edu

[2]S. Jamieson and J.P. How are with the Department of Aeronautics and Astronautics at the Massachusetts Institute of Technology (MIT) {sjamieson,jhow}@mit.edu

[3]Y. Girdhar is with the Applied Ocean Physics and Engineering Department at the Woods Hole Oceanographic Institution (WHOI) yogi@whoi.edu

[1]This refers to an autonomous action, such as targeting and using a particular sensor, that preempts the robot's current task after being triggered by a specific phenomena recognized by an onboard detection algorithm.

scientifically relevant image observations, given a predefined model, and schedule more detailed observations with other sensors. However, these algorithms required domain-specific feature engineering and lacked spatial observation models, so the adaptive path planning was limited to moving the robot closer to a target that had already been detected. At the other extreme, "curious" robots use a generic unsupervised vision model and autonomously move towards anything in their environment that is surprising or novel to the model [14]; the lack of operator input makes it impossible to directly specify particular scientific objectives using this approach.

Our work is closely related to the work of Arora et al. [15], which modeled operator's domain knowledge with a predefined Bayesian Network (BN) that was used by the robot to estimate the reward for a trajectory. They introduced a spatial observation model in the system, enabling informative path planning using Monte-Carlo Tree Search (MCTS) to explore an action tree composed of movement and sensing actions [15]. Their approach requires the operator to specify the domain-specific BN *a priori*, and has limited utility as a general purpose exploration tool that can be deployed in unknown environments. In contrast, our solution learns the reward model online, and hence allows the robot to deal with unexpected observations efficiently during exploration.

Active learning algorithms interactively query an oracle to produce samples in the training set, such that the model can be trained with far fewer labeled examples than would normally be required [16]. Active reward learning algorithms have efficiently learned reward models representing human ratings or preferences for robot behaviours by making on the order of 10-100 reward queries [17], [18]. Doshi-Velez et al. [19] considered a query to be an action that could be taken if it helped the robot to gain additional reward. This is *online* active learning and our approach is most related to theirs with the main difference that we use reward queries to learn a mapping from observations to reward, whereas [19] used policy queries to learn optimal actions directly.

Due to the high-dimensionality of natural images, even with active learning, it can take hundreds of queries to learn a reward model [20]. In bandwidth limited environments such as the deep sea, sending that many images for labelling during the span of a mission is not feasible. Deep features are relatively low-dimensional representations of images which are very helpful for learning new classification tasks with few examples [21], [22]. Topic models, especially when combined with deep features, can be used to provide a low-dimensional semantic representation of the visual environment [23]. Our proposed POMDP approach leverages both active learning and low-dimensional image representations to enable interactive visual exploration over low bandwidth.

III. The Co-Robotic Visual Exploration POMDP

We present the co-robotic visual exploration problem as a POMDP. We model the state of the robot at time t as $S_t = (X_t, Y_t, L_t)$. $X_t = \{(\boldsymbol{x}_i, I_i)\}_{i=1}^{t}$ is the sequence of locations the robot has visited, with corresponding image observations, where the current location is $\boldsymbol{x}_t \in \mathbb{R}^3$ and the latest observation is the image $I_t \in \mathbb{O}$. L_t is the set of indices of images sent to and labeled by the operator. Y_t contains the reward labels for all images, including those that have not been sent; most of these are unknown, making the robot's state partially observable.

The partial observability comes from the robot's limited ability to query the operator during a mission; in bandwidth constrained environments the robot sends images at a much slower rate than it collects them, so it must decide which labels to observe. We assume that only the operator can evaluate the unknown, but deterministic, binary "interest" function $\mathcal{I}(I)$ such that $(Y_t)_i = \mathcal{I}(I_i)$. Further, it is assumed that the operator cannot express their interest function analytically (otherwise it would be computed onboard the robot), and would instead train an approximate model based on their labels for various example images. However, since exploration typically occurs in remote and unstudied environments, the operator does not have a fully representative dataset of what the robot will observe and is unable to provide the robot with a complete model of $\mathcal{I}(\cdot)$ *a priori*.

The entire POMDP is characterized by the tuple $(\mathcal{S}, \mathcal{A}, \mathbb{O}, T, O, R, \gamma, b_0)$:

Component	Definition	Our Assumptions
$\mathcal{S}$	State space of the robot	$S = (X, Y, L)$
$\mathcal{A}$	Discrete set of robot actions	Motion primitives[2]
$\mathbb{O}$	Observation space	Natural images and binary labels
T	Transition function	$\mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}$
O	Observation model	$\mathcal{S} \times \mathcal{A} \mapsto \mathbb{O}$
R	Reward model	$\mathcal{S} \times \mathcal{A} \times \mathbb{O} \mapsto \mathbb{R}$
γ	Discount factor	$\gamma \in [0, 1]$
b_0	Initial belief state	Initial location $\boldsymbol{x}_0$

Given these specifications, it is typical for the robot to use an online POMDP planner to approximate an optimal policy $\pi^\star : \mathcal{S} \mapsto \mathcal{A}$ in real-time. Algorithm 1 presents our approach to co-robotic exploration based on the assumptions listed above.

There are three key decisions to fully specify the co-robotic visual exploration POMDP that we will consider. The first is defining an observation model over the space of natural images. The second is defining a reward model, and the third is choosing an effective active learning strategy.

A. Spatial Observation Model for Images

A spatial observation model is required for adaptive path planning because the robot's reward is determined by what it observes, so to evaluate a candidate trajectory the robot must predict what it will observe along that trajectory. This should be possible because the semantic contents of natural images, such as terrain types and species present, often have strong spatial correlation [23], [24]. However, these correlations are hard to model in the pixel space, where even nearly identical images can be made distant by effects like sensor noise

[2]Querying the operator is often modelled (e.g., in [19]) as another action in $\mathcal{A}$ with some cost of communication, such as energy usage, included in the reward model. For simplicity, we assume this cost is negligible and that the robot performs queries concurrently with other actions.

Algorithm 1: Co-Robotic Exploration

Given: $(\mathcal{S}, \mathcal{A}, \mathbb{O}, T, \mathcal{O}, R, \gamma, \boldsymbol{x}_0)$, $t_{\max}$
$X_0 \leftarrow \varnothing$ // Stores the path and observations
$\boldsymbol{\tau} \leftarrow \{\boldsymbol{x}_0\}$ // The current trajectory plan
$q \leftarrow$ *null* // Index of next observation to label
$t \leftarrow 1$ // The current timestep
while $t < t_{\max}$:
  $\boldsymbol{x}_t \leftarrow$ NEXT_STEP($\boldsymbol{\tau}$)
  $I_t \leftarrow$ OBSERVE($\boldsymbol{x}_t$)
  $X_t \leftarrow X_{t-1} \cup \{\boldsymbol{x}_t, I_t\}$
  $\mathcal{O} \leftarrow$ UPDATE_OBSERVATION_MODEL($\mathcal{O}, X_t$)
  if LABEL_READY(I_q)
    $y_q \leftarrow$ QUERY_RESULT(I_q)
    $Y_t \leftarrow Y_{t-1} \cup \{I_q, y_q\}$
    $R \leftarrow$ UPDATE_REWARD_MODEL(R, Y_t)
    $q \leftarrow$ *null*
  endif
  $\boldsymbol{\tau} \leftarrow$ PLAN_TRAJECTORY($X_t, \mathcal{S}, \mathcal{A}, T, \mathcal{O}, R, \gamma$)
  if $q =$ *null*
    $q \leftarrow$ QUERY_SELECTOR($X_t, \mathcal{O}, R$)
    REQUEST_LABEL(I_q)
  endif
  $t \leftarrow t + 1$

Algorithm 2: PLAN_TRAJECTORY

Input: $X_t, \mathcal{S}, \mathcal{A}, T, \mathcal{O}, R, \gamma$
Given: n // Number of trajectories to test
$\mathcal{T} \leftarrow$ GENERATE_TRAJECTORIES($X_t, \mathcal{S}, \mathcal{A}, T, n$)
for $i = 1, \ldots, n$:
  $\boldsymbol{s}_i \leftarrow$ SCORE_TRAJECTORY($\mathcal{T}(i), O, R, \gamma$)
$\boldsymbol{\tau} \leftarrow \mathcal{T}(\text{argmax}_i\, \boldsymbol{s}_i)$
return $\boldsymbol{\tau}$

and slight changes in illumination [25]. Further, due to the high dimensionality of the image space, there are no spatial models with which it is computationally tractable to predict the image that would be observed in an unvisited location.

To overcome these challenges, the robot computes semantic representations of images in the space $\mathcal{Z}$, which is low-dimensional compared to the space of natural images $\mathbb{O}$. The robot builds a spatial observation model over semantic representations, denoted as $\boldsymbol{Z}_t(\boldsymbol{x}) : \mathbb{R}^3 \mapsto \mathcal{Z}$, trained using the observations $X_t = \{(x_i, I_i)\}_{i=1}^t$ and the semantic feature extractor $\boldsymbol{z}(I) : \mathbb{O} \mapsto \mathcal{Z}$. This approach requires that the semantic representations of two images $\boldsymbol{z}(I_1), \boldsymbol{z}(I_2)$ are similar (typically measured by Euclidean distance) if and only if the human-perceived similarity of I_1 and I_2 is high. Semantic representations derived from computer vision models developed for unsupervised natural image clustering, such as deep feature extractors [22], [25] and spatial topic models (STMs) [26], [27] have this property.

STMs such as BNP-ROST [28], [29] are a strong class of candidates for the spatial observation model because the priors they use to represent the spatial distributions of topics are smooth (spatially correlated) and the topic distributions they use to represent images have low dimensionality. The low-dimensionality of these representations is a critical requirement for learning the reward function from few examples; this is much more challenging with higher dimensional representations such as deep features [25].

B. Learning a Reward Model Online over Low Bandwidth

We define the robot's reward to be the total number of unique and interesting observations it has collected

$$R(X_t) = \sum_{i=1}^{t} \mathcal{I}(I_i) = \sum_{i=1}^{t} (Y_t)_i . \tag{1}$$

This can only be computed after the operator sees all images (i.e., after the mission). Since the robot models observations in the semantic space $\mathcal{Z}$, trajectory planning requires it to estimate the reward as a function of the semantic field $\boldsymbol{Z}_t(\boldsymbol{x})$. For this, the robot learns a model $g_\theta : \mathcal{Z} \mapsto [0, 1]$

$$R(\boldsymbol{x}) \approx g(\boldsymbol{Z}_t(\boldsymbol{x}); \theta), \tag{2}$$

where θ is a set of parameters for the model family. Recall that L_t is the set of labeled image indices at time t, and let $D_t = \{(I_i, (Y_t)_i)\}_{i \in L_t}$ be the corresponding training set. We choose θ to minimize the cross-entropy loss $\mathcal{L}$ on D_t, resulting in the final reward model

$$R(\boldsymbol{x}; D_t) \approx g\left(\boldsymbol{Z}(\boldsymbol{x}); \theta^\star_{D_t}\right) \tag{3}$$

$$\theta^\star_{D_t} = \underset{\theta}{\text{argmin}} \sum_{(I,y) \in D_t} \mathcal{L}(y, g(\boldsymbol{z}(I); \theta)) . \tag{4}$$

The number of labeled examples that a model must be trained on in order to generalize well is proportional to the sample complexity of the model family [30], and for simple models (e.g., logistic regression) the sample complexity is typically linear in the number of input dimensions [31]. Thus, it is desirable to jointly pick a semantic representation and a model g_θ such that the total number of examples required to train g_θ is less than the number of examples that can be labelled during the mission. This further motivates the use of BNP-ROST [29] as the semantic feature extractor, since the dimensionality of its semantic representation grows as $\log t$, logarithmic in the number of images t, while the number of labelled images grows linearly at $\frac{t}{n}$, where $n \geq 1$ is set by the bandwidth constraint. Thus, when using BNP-ROST in combination with a simple reward model, then the training process for g_θ is expected to quickly converge to good parameters θ, even with few training examples.

C. Query Selection for Low Bandwidth Reward Learning

When the robot observes novel phenomena, it needs to query the operator's interest in collecting more observations of the phenomena. The only type of query the robot can perform in an unknown environment is sending an image to the operator and receiving an interest label in return; the operator cannot determine their interest in an image from the image's semantic representation, and does not have access to enough information to advise the robot on the optimal policy. This is a unique challenge for active learning.

IV. Online Active Reward Learning for POMDPs

Here we will consider active learning strategies to learn the parameters of a POMDP reward model online. We denote the set of unlabelled image indices at time t as $\mathcal{U}_t$, and the active learning metric as $h(\boldsymbol{z})$, such that the next image to request a label for is chosen as

$$i^\star = \underset{i \in \mathcal{U}_t}{\operatorname{argmax}}\, h\left(\boldsymbol{z}\left(I_i\right)\right). \tag{5}$$

A. Non-Adaptive Query Selection

The simplest approaches to selecting images to be labelled do not depend on $\boldsymbol{z}$, and thus are good baselines to consider. *Random* selection chooses unlabelled observations uniformly at random. *Uniform* selection instead chooses every n^{th} image, where n is determined by the bandwidth constraint.

B. Informative Query Selection

Informative query selection involves defining some uncertainty metric on the model, and choosing to label the observation which results in the greatest reduction of uncertainty. There are many query selection strategies that fall into this category and are effective at learning a function in few examples [32]. A common uncertainty metric for classification problems is entropy, where the highest entropy values occur when an observation is on a decision boundary. A widely-used approach to informative query selection is "uncertainty sampling", which typically means picking the observation with the maximum entropy [32]

$$h_{\text{Entropy}}\left(\boldsymbol{z}; \theta^\star_{D_t}\right) = \mathbb{H}\left[g\left(\boldsymbol{z}; \theta^\star_{D_t}\right)\right]. \tag{6}$$

An issue with uncertainty sampling is that labeling the most uncertain observation might not have much effect on the model parameters θ – if the model parameters do not change, then the model performance does not increase. This suggests maximizing "error reduction" [32] instead

$$h_{\text{Info}}\left(\boldsymbol{z}\right) = h_{\text{Entropy}}\left(\boldsymbol{z}; \theta^\star_{D_t}\right) - \mathbb{E}_{D'_t|D_t}\left[h_{\text{Entropy}}\left(\boldsymbol{z}; \theta^\star_{D'_t}\right)\right] \tag{7}$$

$$D'_t = D_t \cup (\boldsymbol{z}, y)$$
$$p(D'_t \mid D_t) \approx g(\boldsymbol{z}; \theta^\star_{D_t}).$$

This *Information Gain* query selection method prioritizes labeling an observation by how much a new label y is expected to reduce the entropy of similar future observations. This should maximize the rate at which entropy is reduced and thus the rate at which the reward function is learned.

C. Regret Minimizing Query Selection

Here we introduce a novel *Regret* minimizing query selector that focuses on identifying labels that maximize the expected reward collected during the mission, rather than information gained about the reward function. Regret is typically defined for POMDPs as the difference in utility between the chosen action and the true optimal action based on complete information. To our knowledge, this is the first work that compares a regret-based heuristic with information-theoretic heuristics in online active learning.

Suppose that the robot is considering a finite set of trajectories $\mathcal{T} = \{\tau_i\}_{i=1}^{N_\tau}$: it uses the observation model to predict what it will observe along each trajectory τ, predicts each trajectory's reward, and finally chooses the one with the highest reward (see Algorithm 2). However, given limited training data, the robot has significant uncertainty in the predicted rewards and thus is unlikely to have chosen the true optimal trajectory. This motivates a question for each unlabeled image: if this image were labeled, would the robot have chosen a different trajectory? If the answer is yes, then it must mean that, given this additional label, a different trajectory would be predicted to have greater reward and thus the robot would "regret" not knowing the label. If it is no, then the robot would have no immediate regret for not knowing it. We formalize this in the following objective:

$$h_{\text{Regret}}\left(\boldsymbol{z}\right) = \mathbb{E}_{D'_t|D_t}\left[R(\tau^\star_{D'_t}; D'_t) - R(\tau^\star_{D_t}; D'_t)\right] \tag{8}$$

$$R(\tau; D_t) = \sum_{\boldsymbol{x} \in \tau} g(\boldsymbol{Z}(\boldsymbol{x}); \theta^\star_{D_t}) \tag{9}$$

$$\tau^\star_D = \underset{\tau \in \mathcal{T}}{\operatorname{argmax}}\, R(\tau; D) \tag{10}$$

Equation 8 may be interpreted as the expected reward increase (regret decrease) given a label for $\boldsymbol{z}$. An approach to computing h_{Regret} is presented in Algorithms 3 and 4.

Algorithm 3: Regret-Based Query Selection

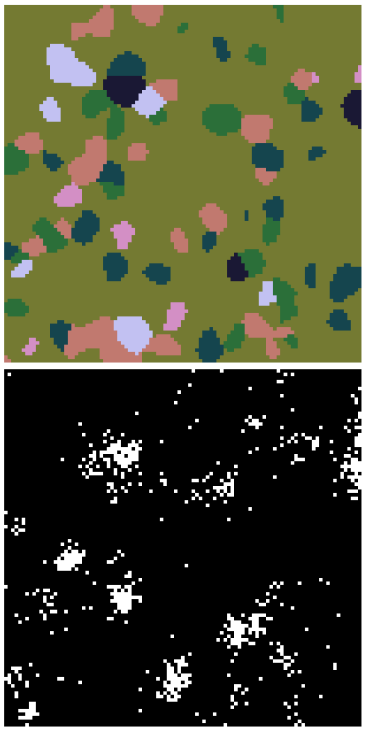
Given: $X_t, \mathcal{S}, \mathcal{A}, T, \mathcal{O}, R, \gamma$
Input: $\mathcal{U}_t$ `// Set of unlabeled image indices`
$\tau_0 \leftarrow$ PLAN_TRAJECTORY$(X_t, \mathcal{S}, \mathcal{A}, T, \mathcal{O}, R, \gamma)$
foreach $i \in \mathcal{U}_t$:
  $\boldsymbol{z} \leftarrow$ SEMANTIC_REPRESENTATION(I_i)
  $y_{\text{pred}} \leftarrow$ PREDICT_REWARD$(\boldsymbol{z})$
  $r_0 \leftarrow$ COMPUTE_REGRET$(\tau_0, \boldsymbol{z}, 0)$
  $r_1 \leftarrow$ COMPUTE_REGRET$(\tau_0, \boldsymbol{z}, 1)$
  $\text{regret}_i \leftarrow y_{\text{pred}} r_1 + (1 - y_{\text{pred}})\, r_0$
return $\operatorname{argmax}_{i \in \mathcal{U}_t}$ regret_i

Algorithm 4: COMPUTE_REGRET

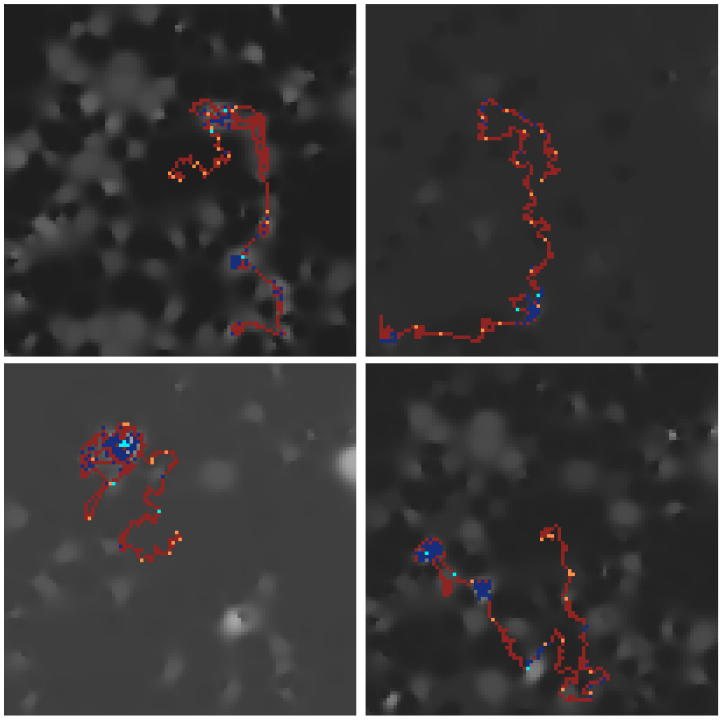
Given: $X_t, \mathcal{S}, \mathcal{A}, T, \mathcal{O}, R, \gamma$
Input: $\tau_0, \boldsymbol{z}, y$ `// Reference trajectory, observation to label, and temporary label`
ADD_TEMPORARY_LABEL$(\mathcal{O}, \boldsymbol{z}, y)$
$\tau^\star \leftarrow$ PLAN_TRAJECTORY$(X_t, \mathcal{S}, \mathcal{A}, T, \mathcal{O}, R, \gamma)$
$s^\star \leftarrow$ SCORE_TRAJECTORY$(\tau^\star, \mathcal{O}, R, \gamma)$
$s_0 \leftarrow$ SCORE_TRAJECTORY$(\tau_0, \mathcal{O}, R, \gamma)$
REMOVE_TEMPORARY_LABEL$(\mathcal{O}, \boldsymbol{z})$
return $(s^\star - s_0)$ `// Regret given the temp label`

V. Experiments

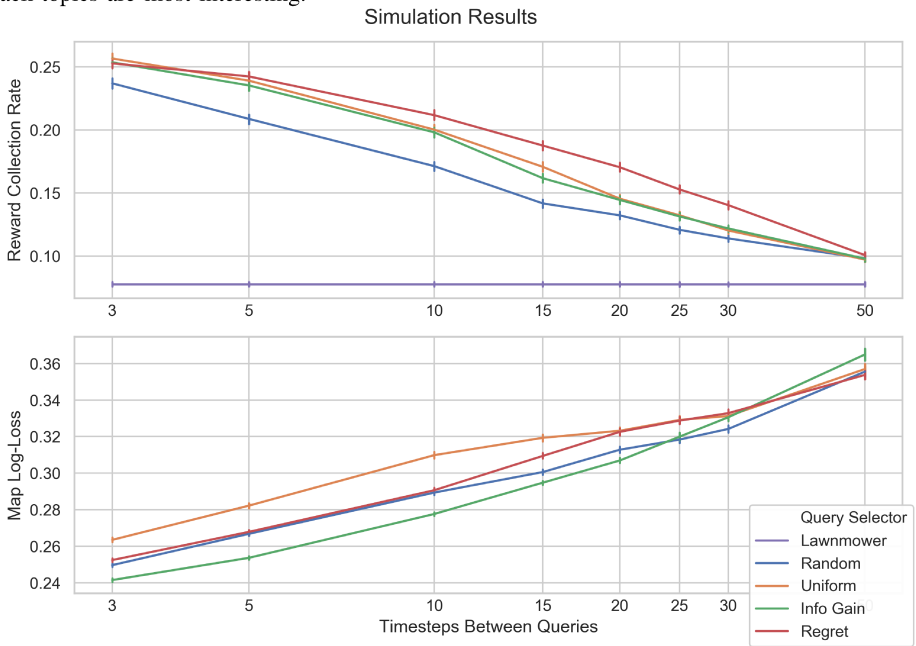
We evaluate the proposed query selection techniques through two experiments, each simulating the co-robotic exploration task with various bandwidth constraints. The first



(a) Top: A topic map where each location is described by the semantic representation $\boldsymbol{z}^{(i,j)} \in \boldsymbol{\Delta}^8$. The color of each pixel indicates the largest component of $\boldsymbol{z}^{(i,j)}$. Bottom: The reward at each location is randomly sampled as $R^{(i,j)} \sim \text{Bernoulli}(\boldsymbol{p}^T \boldsymbol{z}^{(i,j)})$, where $\boldsymbol{p} \in [0,1]^k$ represents how "interesting" each component of $\boldsymbol{z}$ is. Here, the pink and black topics are most interesting.



(b) Best viewed on a screen. Sample trajectories followed by robots starting at the center of the map in (a) with different query selectors. Along each trajectory, red-orange pixels correspond to no reward, and blue pixels to reward. Bright orange/blue pixels represent observations for which the query selector requested the label. The greyscale background intensities represent $g(Z(x); \theta_D^\star)$: reward estimates of observations at each location, based on all labeled samples. Query Selectors: (top row) Random, Uniform; (bottom row) Info Gain, Regret.



(c) A comparison of the query selector performance for different bandwidth availability; the x-axis represents labeling period (time between making a call to REQUEST_LABEL and LABEL_READY returning true in Algorithm 1), which is inversely proportional to bandwidth. Each datapoint represents the mean of 1080 simulations (36 trials on 30 unique maps) and bars represent the 68% confidence bound of the mean. Top: The mean amount of reward collected by each robot per unit time (higher is better). Lawnmower is not a query selector, but rather represents the mean reward collected by 8 preplanned boustrophedonic trajectories [33] that each start at the center of the map and move towards a corner. Bottom: The mean cross-entropy loss between the ground truth interest maps, as in (a), and the corresponding robots' predictions of the reward at each location, as in (b), at the end of each simulation (lower is better).

Fig. 2: Stages of the simulation procedure, and performance comparison of the query selectors on fully simulated data.

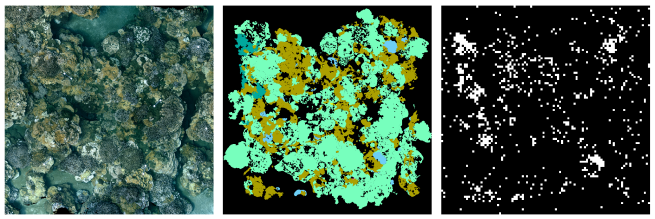


Fig. 3: Left: A crop of the KAH_2016_3 photomosaic image from the 100 Islands Challenge [34], showing a coral reef near Kaho'olawe. Center: The photomosaic annotations where each color represents an expert label [34]. Right: One of 30 unique interest maps generated (cf. Figure 2a).

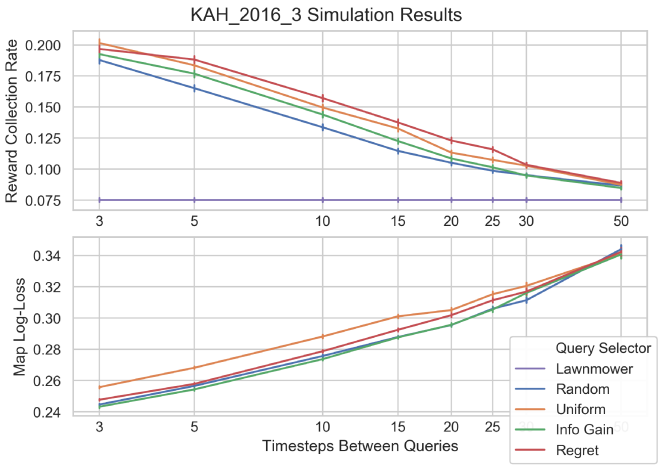


Fig. 4: The Regret query selector continues to outperform the other active learning heuristics when the topic map is derived from a real image (see Figure 3).

experiment (see Figure 2) used 30 artificial "topic maps" (cf. [28]) created by randomly generating Voronoi partitions of a 100×100 image, assigning each cell a topic label, and then assigning each pixel's topic distribution as a distance-weighted mean over cell labels. This produced continuous topic maps with topics in varying concentrations, and each one was associated with a unique interest map (see Figure 2a). In the second experiment, a single topic map was derived from the expert annotations of an actual coral reef image, and 30 interest maps were generated for it (see Figure 3). The procedure for both experiments was:

1) Generate a map of topic distributions $\boldsymbol{z}(\boldsymbol{x}) \in \boldsymbol{\Delta}^d$ which represent the observations at each location $\boldsymbol{x}$;
2) Generate an interest profile $\boldsymbol{p} \in [0,1]^d$ so that $p = \boldsymbol{p}^T \boldsymbol{z}$ is the probability that the operator is interested in an observation with feature representation $\boldsymbol{z}$;
3) Generate a binary "interest map" by sampling $R(\boldsymbol{x}) \sim \text{Bernoulli}(p(\boldsymbol{z}(\boldsymbol{x})))$ at each location $\boldsymbol{x}$ in the topic map;
4) For each bandwidth limitation and each query selection algorithm: perform 36 rollouts of algorithm 1 for a simulated robot making reward queries according to the bandwidth limitation and query selector

Each rollout in step (4) had a duration of 300 timesteps; robot movement was one pixel per timestep and bandwidth constraints were simulated by changing the number of timesteps for a label to be received after being requested. State transitions and observations were deterministic and noiseless. The robot started with no training data and used logistic regression (from [35]) as its reward model. Trajectories were generated by randomly sampling sequences of 5 motion primitives.[3] 50 trajectories were generated at each timestep and scored using the sum of the predicted rewards along the trajectory, less the scores of locations already visited. The highest scoring trajectory was followed.

VI. Results & Discussion

We compared the Random, Uniform, Information Gain, and Regret query selectors described in Section IV over a total of 69120 simulations; the mean reward collection rates and interest map prediction losses for each experiment are presented in Figures 2c and 4. The Regret query selector matches, or outperforms, every other selection criterion at collecting reward, at any bandwidth availability, in these simulation configurations. The relative gains of non-random query selection are smaller when the time between queries is short (high-bandwidth) and thus almost every image is labeled, or when it is so long (low-bandwidth) that the robot barely learns anything before the mission ends. The results also demonstrate the vast improvement of autonomous exploration over preplanned trajectories: the adaptive planners collected up to 29.7% more reward at very low bandwidth, and up to 230% more reward at high bandwidth.

The regret-based method did not learn the reward function as well as the information gain query selector, based on its higher map log-loss. This exemplifies the difference in the design criteria: the information theoretic criterion focuses on useful labels for learning a function, which is appropriate for active reward learning *offline*, during training. The regret criterion instead optimizes for the robot's reward, making it better suited for *online* active reward learning, which describes our usage of queries during a live mission.

VII. Conclusions and Future Work

The Co-Robotic Visual Exploration POMDP provides a structured approach to managing human-robot collaboration and high-dimensional observation spaces in autonomous science. We provide general principles for choosing the POMDP's observation model, reward model, and active learning criterion, and demonstrate that the novel Regret-based active learning criterion can greatly improve the amount of reward collected. Some next steps are: exploring spatial observation models capable of longer-range topic prediction (e.g. [36]), extending the reward model and active learning formulation to non-binary rewards, and using higher-fidelity simulations and field deployments to better understand the performance increases that can be achieved in real-world autonomous exploration.

[3]The primitives were 13 straight lines, each 5 units long and at angles spaced uniformly between -135° to 135° from the robot's current direction.

REFERENCES

[1] T. A. Estlin, B. J. Bornstein, D. M. Gaines, R. C. Anderson, D. R. Thompson, M. Burl, R. Castaño, and M. Judd, "AEGIS Automated Science Targeting for the MER Opportunity Rover," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 3, pp. 1–25, 2012.

[2] Y. Gao and S. Chien, "Review on space robotics: Toward top-level science through space exploration," *Science Robotics*, vol. 2, no. 7, p. eaan5074, 6 2017.

[3] R. D. Ballard, "WHOI-93-34: The JASON Remotely Operated Vehicle System," Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, Tech. Rep., 1993.

[4] B. P. Foley, R. M. Eustice, K. Dellaporta, D. Evagelistis, D. Sakellariou, V. L. Ferrini, B. S. Bingham, K. Katsaros, R. Camilli, D. Kourkoumelis, A. Mallios, H. Singh, P. Micha, D. S. Switzer, D. A. Mindell, T. Theodoulou, and C. Roman, "The 2005 Chios ancient shipwreck survey: New methods for underwater archaeology," *Hesperia*, vol. 78, no. 2, pp. 269–305, 2009.

[5] M. E. Clarke, N. Tolimieri, and H. Singh, "Using the Seabed AUV to Assess Populations of Groundfish in Untrawlable Areas," *The Future of Fisheries Science in North America*, pp. 357–372, 2009.

[6] G. Williams, T. Maksym, J. Wilkinson, C. Kunz, C. Murphy, P. Kimball, and H. Singh, "Thick and deformed Antarctic sea ice mapped with autonomous underwater vehicles," *Nature Geoscience*, vol. 8, no. 1, pp. 61–67, 2015.

[7] J. W. Kaeli, J. J. Leonard, and H. Singh, "Visual summaries for low-bandwidth semantic mapping with autonomous underwater vehicles," in *2014 IEEE/OES Autonomous Underwater Vehicles (AUV)*. IEEE, 10 2014, pp. 1–7.

[8] G. Burroughes and Y. Gao, "Ontology-Based Self-Reconfiguring Guidance, Navigation, and Control for Planetary Rovers," *Journal of Aerospace Information Systems*, vol. 13, no. 8, pp. 316–328, 8 2016.

[9] S. Chien, R. Sherwood, D. Tran, B. Cichy, G. Rabideau, R. Castano, A. Davis, D. Mandl, S. Frye, B. Trout, S. Shulman, and D. Boyer, "Using Autonomy Flight Software to Improve Science Return on Earth Observing One," *Journal of Aerospace Computing, Information, and Communication*, vol. 2, no. April, pp. 196–216, 2005.

[10] T. Estlin, D. Gaines, C. Chouinard, R. Castano, B. Bornstein, M. Judd, I. Nesnas, and R. Anderson, "Increased Mars Rover Autonomy using AI Planning, Scheduling and Execution," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*, no. April. IEEE, 4 2007, pp. 4911–4918.

[11] R. Castano, T. Estlin, D. Gaines, A. Castano, C. Chouinard, B. Bornstein, R. Anderson, S. Chien, A. Fukunaga, and M. Judd, "Opportunistic Rover Science: Finding and Reacting to Rocks, Clouds and Dust Devils," in *2006 IEEE Aerospace Conference*, vol. 2006. IEEE, 2006, pp. 1–16.

[12] G. Hitz, E. Galceran, M.-È. Garneau, F. Pomerleau, and R. Siegwart, "Adaptive continuous-space informative path planning for online environmental monitoring," *Journal of Field Robotics*, vol. 34, no. 8, pp. 1427–1449, 12 2017.

[13] G. Flaspohler, V. Preston, A. P. M. Michel, Y. Girdhar, and N. Roy, "Information-Guided Robotic Maximum Seek-and-Sample in Partially Observable Continuous Environments," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3782–3789, 10 2019.

[14] Y. Girdhar and G. Dudek, "Modeling curiosity in a mobile robot for long-term autonomous exploration and monitoring," *Autonomous Robots*, vol. 40, no. 7, pp. 1267–1278, 10 2016.

[15] A. Arora, R. Fitch, and S. Sukkarieh, "An approach to autonomous science by modeling geological knowledge in a Bayesian framework," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 9 2017, pp. 3803–3810.

[16] M. F. Balcan, S. Hanneke, and J. W. Vaughan, "The true sample complexity of active learning," *Machine Learning*, vol. 80, no. 2-3, pp. 111–139, 2010.

[17] C. Daniel, M. Viering, J. Metz, O. Kroemer, and J. Peters, "Active Reward Learning," in *Proceedings of Robotics: Science and Systems (RSS)*. Robotics: Science and Systems Foundation, 7 2014.

[18] D. Sadigh, A. Dragan, S. Sastry, and S. Seshia, "Active Preference-Based Learning of Reward Functions," in *Robotics: Science and Systems XIII*. Robotics: Science and Systems Foundation, 7 2017.

[19] F. Doshi-Velez, J. Pineau, and N. Roy, "Reinforcement learning with limited reinforcement: Using Bayes risk for active learning in POMDPs," *Artificial Intelligence*, vol. 187-188, pp. 115–132, 8 2012.

[20] F. Shkurti, "Algorithms and Systems for Robot Videography from Human Specifications," Ph.D. dissertation, McGill University, 2018.

[21] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition," in *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32, no. 1. Bejing, China: PMLR, 2014, pp. 647–655.

[22] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised Deep Feature Extraction for Remote Sensing Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 3, pp. 1349–1362, 3 2016.

[23] G. Flaspohler, N. Roy, and Y. Girdhar, "Feature discovery and visualization of robot mission data using convolutional autoencoders and Bayesian nonparametric topic models," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 9 2017, pp. 1–8.

[24] H. Reiss, S. Cunze, K. König, H. Neumann, and I. Kröncke, "Species distribution modelling of marine benthos: A North Sea case study," *Marine Ecology Progress Series*, vol. 442, no. December, pp. 71–86, 12 2011.

[25] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 6 2018, pp. 586–595.

[26] X. Wang and E. Grimson, "Spatial Latent Dirichlet Allocation," in *Proceedings of the 20th International Conference on Neural Information Processing Systems*. Vancouver, British Columbia, Canada: Curran Associates Inc., 2007, pp. 1577–1584.

[27] Y. Girdhar, P. Giguère, and G. Dudek, "Autonomous adaptive exploration using realtime online spatiotemporal topic modeling," *The International Journal of Robotics Research*, vol. 33, no. 4, pp. 645–657, 4 2014.

[28] Y. Girdhar, L. Cai, S. Jamieson, N. McGuire, G. Flaspohler, S. Suman, and B. Claus, "Streaming Scene Maps for Co-Robotic Exploration in Bandwidth Limited Environments," in *2019 International Conference on Robotics and Automation (ICRA)*. Montreal, Canada: IEEE, 5 2019, pp. 7940–7946.

[29] Y. Girdhar, Walter Cho, M. Campbell, J. Pineda, E. Clarke, and H. Singh, "Anomaly detection in unstructured environments using Bayesian nonparametric scene modeling," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. Stockholm, Sweden: IEEE, 5 2016, pp. 2651–2656.

[30] M. Mitzenmacher and E. Upfal, "Sample Complexity, VC Dimension, Rademacher Complexity," in *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*, 2nd ed. Cambridge University Press, 2017, ch. 14, pp. 361–391.

[31] A. Y. Ng and M. I. Jordan, "On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes," in *Proceedings of the 14th International Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, 2001, pp. 841–848.

[32] Y. Yang and M. Loog, "A benchmark and comparison of active learning for logistic regression," *Pattern Recognition*, vol. 83, pp. 401–415, 2018.

[33] H. Choset and P. Pignon, "Coverage Path Planning: The Boustrophedon Cellular Decomposition," in *Field and Service Robotics*, A. Zelinsky, Ed. London: Springer London, 1998, pp. 203–209.

[34] J. E. Smith, R. Brainard, A. Carter, S. Grillo, C. Edwards, J. Harris, L. Lewis, D. Obura, F. Rohwer, E. Sala, P. S. Vroom, and S. Sandin, "Re-evaluating the health of coral reef communities: baselines and evidence for human impacts across the central Pacific," *Proceedings of the Royal Society B: Biological Sciences*, vol. 283, no. 1822, 1 2016.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[36] J. E. San Soucie, H. M. Sosik, and Y. Girdhar, "Gaussian-Dirichlet Random Fields for Inference over High Dimensional Categorical Observations," in *2020 International Conference on Robotics and Automation (ICRA)*. Paris, France: IEEE, 5 2020.