Policy Search on Aggregated State Space for Active Sampling

Sandeep Manjanna[1], Herke Van Hoof[2], and Gregory Dudek[1]

[1] Mobile Robotics Lab (MRL), Center for Intelligent Machines, McGill University, Montreal, QC, Canada. E-mail: [msandeep,dudek]@cim.mcgill.ca.
[2] Amsterdam Machine Learning Lab (AMLAB), University of Amsterdam, Amsterdam, the Netherlands. E-mail: h.c.vanhoof@uva.nl.

1 Motivation, Problem Statement, and Related Work

We present an anytime [1] adaptive sampling technique that generates paths to efficiently measure and then mathematically model a scalar field by performing non-uniform measurements in a given region of interest. In particular, the class of scalar field we are interested is some physical or virtual parameter that varies with location, such as depth of the sea floor or the probability of finding a lost object. As the measurements are collected at each sampling location, we can compute an estimate of the large-scale variation of the phenomenon of interest. We compute a sampling path that minimizes the expected time to accurately model the phenomenon of interest by visiting high information regions using non-myopic path generation based on reinforcement learning.

As an example application, we consider monitoring the health of coral reefs by sampling visual data from the surface using an autonomous surface vehicle (ASV) shown in Fig. 1a. Increase in the sea surface temperatures has resulted in widespread coral bleaching at an ever-increasing rate [2] (Fig. 1b). Improved monitoring would enhance the currently poor understanding of the spatial and temporal dynamics of coral bleaching. Since we are sampling from the surface, higher information gain is provided in shallower regions where visibility is better.

(a) (b) (c)

Fig. 1: (a) Custom made differential drive ASV used for coral reef surveys. (b) A coral with bleached spots [2]. (c) Aerial image of the reef surveyed in our field experiments.

Novel contributions of this work include (1) the computation of a sampling technique based on policy search, (2) a non-uniform (multi-resolution) representation of the state space to aid better spatial sampling and (3) a statistically significant evaluation through rigorous experiments with real satellite image data and real robots in the field.

Active sampling refers to the act of strategically planning paths based on the observations made until the current time-step. Exhaustively sampling each point of an unknown survey region [3, 4] can be tedious and impractical if the survey space is large and/or the phenomenon of interest has only a few regions with important information (*hotspots*) [5, 6]. Also it has been observed that for low-pass multi-band signals, uniform sampling can be inefficient and sampling rates far below the Nyquist rate can still be information preserving [7]. This is the key guiding principle behind active and non-uniform sampling [6, 8, 9].

In our approach, a continuous two-dimensional sampling region is discretized into uniform grid-cells, such that the robot's position $\mathbf{x}$ can be represented by a pair of integers $\mathbf{x} \in \mathbb{Z}^2$. Each grid-cell (i, j) is assigned a score $\mathrm{q}(i, j)$ indicating the expected goodness of the visual data in that cell. The goal is to maximize the total accumulated score J over a trajectory τ within a fixed amount of time T. To specify the robot's behavior we use a parametrized policy $\pi_{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) = p(\mathbf{a}|\mathbf{s}; \boldsymbol{\theta})$ that maps the current state $\mathbf{s}$ of sampling to a distribution over possible *actions* $\mathbf{a}$. Our aim will be to automatically find good parameters $\boldsymbol{\theta}$, after which the policy can be deployed without additional training on new problems.

2 Technical Approach

Our algorithm gets trained with a generic score-map (q) generated by the satellite data from areas that exemplify the target environments, for example images of coral reefs. The system is trained to achieve paths that preferentially cover *hotspots* at the earlier stages of exploration. These learned parameters then define a *policy* π (in the sense of reinforcement learning) that is then used on the satellite image or any other sensor map of the target coral reef (Fig. 1c) to generate an explicit action plan. During the test phase, an action with maximum probability is chosen at a given state. Thus, the policy does not need to be re-trained for each new reef map. This property is a key feature of our approach.

In our approach, we formalize the sampling problem as a Markov Decision Process (MDP). We take the state $\mathbf{s}$ to include the position of the robot $\mathbf{x}$ as well as the map q containing the per-location score for the visual data, $\mathbf{s} = (\mathbf{x}, \mathrm{q})$. The action space A consists of four actions (move North, East, South, or West). Transitions deterministically move the agent in the desired direction. Once the visual data at the current cell (i, j) is sampled, the score $\mathrm{q}(i, j)$ is reduced to 0. The discounted reward function is defined as $\gamma^t q(\mathbf{x})$, with the discount factor $0 \leq \gamma \leq 1$ encouraging the robot to sample cells with high scores in early time steps t.

2.1 Policy Gradient Method

Policy gradient methods use gradient ascent for maximizing the expected return $J_\theta = \mathbb{E}_{\tau_\theta}[\sum_{t=1}^{|\tau_\theta|} \mathrm{q}(\mathbf{x}_t)\gamma^t]$. The gradient of the expected return $(\nabla_\theta J_\theta)$ guides the

direction of the parameter update ($\theta_{k+1} = \theta_k + \eta \nabla_\theta J_\theta$, where η is the learning rate). The likelihood ratio policy gradient [10] is given by,

$$\nabla_\theta J_\theta = \int_\tau \nabla_\theta p_\theta(\tau) R(\tau) d\tau \tag{1}$$

This expression depends on the correlation between actions and previous rewards, which are 0 in expectation and cause additional variance. We use the Policy Gradient Theorem (PGT) algorithm and the GPOMDP algorithm [11–14] for computing the policy gradient as it yields relatively low-variance updates. Accordingly, the policy gradient is given by,

$$\nabla_\theta J_\theta \quad = \quad \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{H-1} \nabla_\theta \log \pi_\theta(a_t^{(i)} | s_t^{(i)}) \left(\sum_{j=t}^{H-1} r(s_j^{(i)}, a_j^{(i)}) - b(s_t^{(i)}) \right). \tag{2}$$

In this equation, the gradient is based on m sampled trajectories from the system, with $\mathbf{s}_j^{(i)}$ the state at the j^{th} time-step of the i^{th} sampled roll-outs. Furthermore, b is a variance-reducing baseline. In our experiments, we set the baseline to the observed average reward.

2.2 Feature Aggregation

A popular method to define stochastic policies over a set of deterministic actions is the use of the Gibbs distribution as policy (also referred to as Boltzman exploration of softmax policy). We consider a commonly used linear Gibbs softmax policy parameterization [12, 15] given by,

$$\pi(\mathbf{s}, \mathbf{a}) = \frac{e^{\boldsymbol{\theta}^T \boldsymbol{\phi}_{\mathbf{s},\mathbf{a}}}}{\sum_{\mathbf{b}} e^{\boldsymbol{\theta}^T \boldsymbol{\phi}_{\mathbf{s},\mathbf{b}}}}, \quad \forall \mathbf{s} \in S; \mathbf{a}, \mathbf{b} \in A, \tag{3}$$

where $\boldsymbol{\phi}_{\mathbf{s},\mathbf{a}}$ is an l-dimensional feature vector characterizing state-action pair $(\mathbf{s}, \mathbf{a})$ and $\boldsymbol{\theta}$ is an l-dimensional parameter vector.

The final feature vector $\phi_{\mathbf{s},\mathbf{a}}$ is formed by concatenating a vector $\phi'_s \delta_{aa'}$ for every action $a' \in \{North, East, South, West\}$, where $\phi'_s \subset \mathbb{R}^k$ is a feature representation of the state space, and $\delta_{aa'}$ is the Kronecker delta. Thus, the final feature vector has $4 \times k$ entries, 75% of which corresponding to non-chosen actions will be 0 at any one time step. We consider five different types of robot-centric feature designs ($\phi'_{\mathbf{s}}$). The first one is to consider a vector with all the scores in the score-map q as presented in Fig. 2a. This feature vector grows in length as the size of the sampling region increases resulting in higher computation times for bigger regions. The four other kinds of feature aggregations are illustrated in Fig. 2b - 2e. These aggregations have a fixed number of features, corresponding to the average scores in the feature map in each of the indicated areas, irrespective of the size of the sampling region.

Fig. 2e depicts a multi-resolution aggregation where the feature cells grow in size along with the distance from the robot. This results in high resolution features close to the robot and lower resolution features further from the robot's current position. The aggregated feature design is only used to achieve better policy search [16], but the robot action is still defined at the grid-cell level.

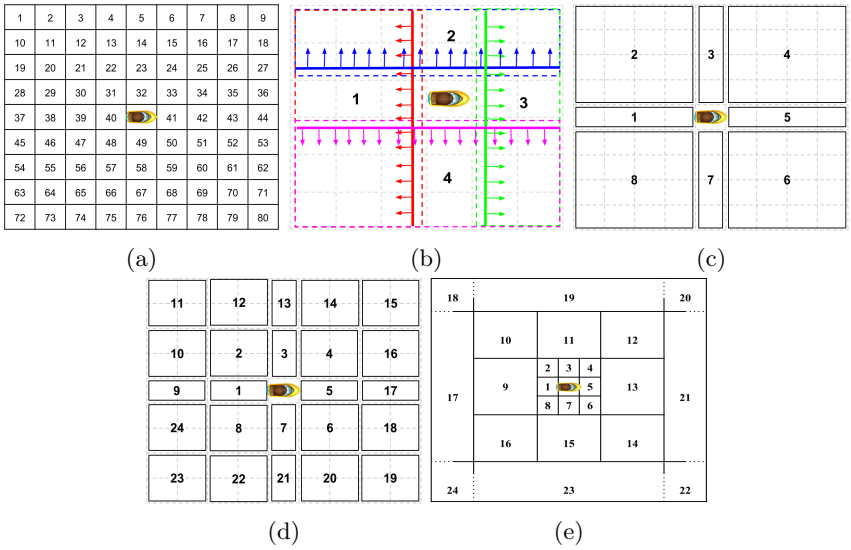


Fig. 2: Robot-centric feature space aggregations. (a) Uniform-grid feature aggregation. (b) 4-feature aggregation. (c) 8-feature aggregation. (d) 24-feature aggregation. (e) Multi-resolution feature aggregation.

3 Experiments

We use a custom-made (Fig. 1a) ASV to survey the coral reefs from the surface to collect visual data. We train our policy based sampling algorithm with the scoremap (Fig. 3b) computed as a multispectral function of the satellite image of a reef (Fig. 3a). The ASV uses the trained parameters to generate sampling paths over the map of interest. We present two test scenarios: 1) A densely populated reef shown in Fig. 4a, and 2) A reef with scattered coral heads shown in Fig. 4c. In the examples in this paper we use aerial images of the reef in Holetown, Barbados.

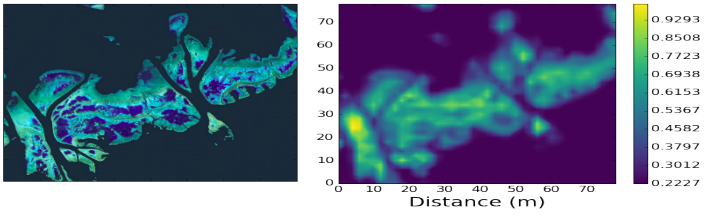
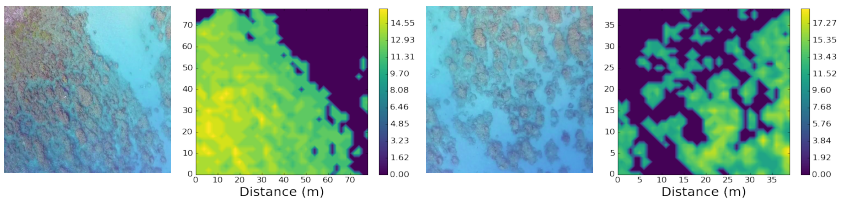


Fig. 3: Experimental setup and scenarios. (a) Satellite image of a reef used for training. (b) Scoremap used for training. The colorbar indicates the interestingness or a score for the presence of corals.

We compare our sampling algorithm with a traditional exhaustive sampling technique using boustrophedonic path [4].

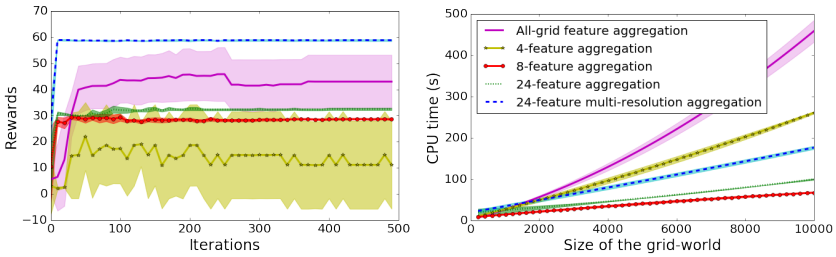


(a) (b) (c) (d)

Fig. 4: Experimental scenarios. (a) and (c) Reef images for test scenarios 1 and 2 respectively. (b) and (d) Scoremaps of the test scenarios 1 and 2 respectively. Colorbars of the scoremaps indicate a score for the presence of corals.

4 Results

Comparing different feature aggregations presented in Section 2.2 shows that multi-resolution aggregated features achieve the highest discounted total rewards (Fig. 5a). Also for the uniform grid aggregation, the computation increases quadratically with the size of the area map (Fig. 5b). These results further strengthen our observation (which follows from the nature of reward discounting under gentle assumptions) that immediate actions are influenced by nearby rewards and the farther low-resolution features enhance non-myopic planning of the complete trajectory. Hence, we used multi-resolution representation for further experiments.



(a) Discounted total rewards (b) CPU-time vs. world size

Fig. 5: Results from evaluation of different feature aggregations. Shaded region indicates the standard deviation over five trials on three different sized maps.

Fig. 6a-6d present the paths generated by our sampling technique and the boustrophedonic sampler on both the test scenarios for 800 time-steps. The path generated by our technique clearly minimizes the sampling over sand and maximizes the visual sampling of corals. The total score collected by both the approaches is comparable (Fig. 6e); however, the discounted reward achieved by our method is significantly higher when running on the scattered coral-head

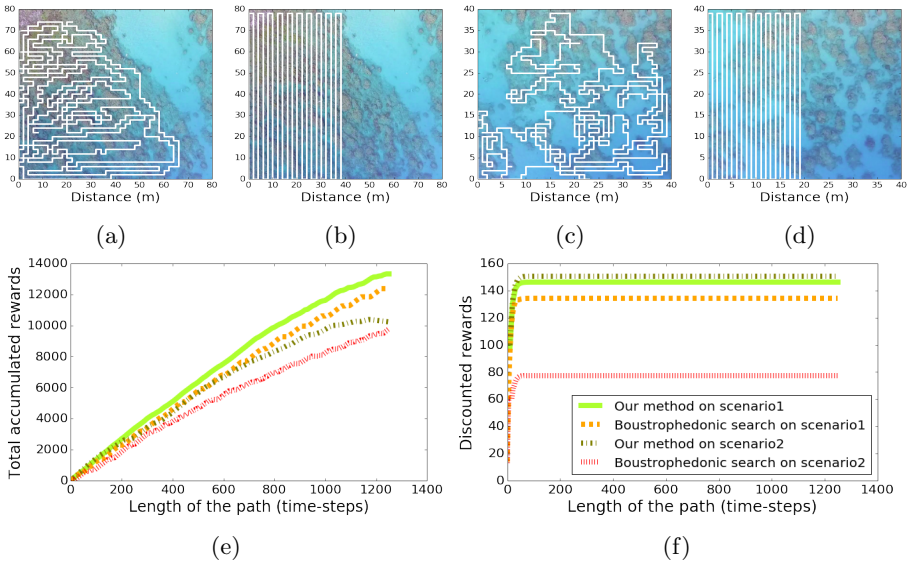


Fig. 6: Results comparing the policy search based sampling algorithm with boustrophedonic sampler. (a) and (c) Policy based sampling path for scenario1 and scenario2 respectively. (b) and (d) Boustrophedonic path for scenario1 and scenario2 respectively. (e) and (f) present the plots for total accumulated rewards and discounted rewards against the length of different sampling paths as indicated in the legends. We used $\gamma = 0.9$ for discounting over time.

scenario (Fig. 6f). This indicates that the proposed approach tends to gather most score early on in the trajectory, resulting in better anytime performance.

5 Field Experiments

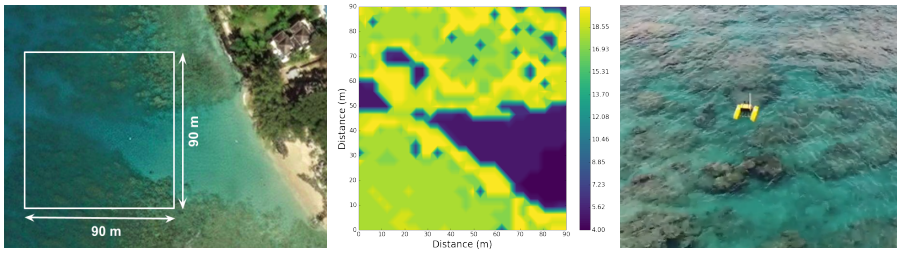
The example application considered in this paper is to collect the visual data of corals from the surface of water. Hence, the shallower the regions visited, the better is the quality of the coral images. In field experiments, we collect visual data of the reef with our sampling method and evaluate our technique for this specific application using the bathymetric data as a measure for shallowness of the region covered. Fig. 7 presents the images captured at different locations of the reef region with varying depths. These images strengthen our hypothesis of covering shallower reefs to achieve high quality visual data of the corals.

We conducted field experiments at Folkestone Marine Reserve in Barbados (Fig. 8a), over a shallow region known to have several coral outcrops. The size of the region to interest considered in these experiments is $90m \times 90m$. The scoremap (Fig. 8b) is computed as a multispectral function of the satellite image of the reef in this region. The reef region is discretized into a grid-world to fit the scoremap of size 30×30. We deployed a custom made differential drive ASV (Fig. 1a and 8c) equipped with a sensor suite consisting of: a downward facing camera, a sonar pinger ($1Hz$), a GPS receiver for localization, and a water-quality sensor-pack.

(a) (b)

Fig. 7: Quality of the visual data. Column (a) presents two images with no useful information. The reef is either too deep (> 20 ft) for good visual samples from the surface or there are no coral-heads. Column (b) presents good quality visual samples of the coral-heads from shallower regions.

Our policy based sampling system was trained on the same scoremap used in Section 3 (Fig. 3b). The path generated by our method is presented in Fig. 9a. The generated path covers most of the high-scoring regions according to the scoremap used (Fig. 8b). This path is limited to a run of 40 minutes. The path in Fig. 9b illustrates the actual path executed by the autonomous boat and it is observed that the sea surface conditions have a considerable impact on the smoothness of the trajectory executed. It is possible for the policy from our approach to flexibly adapt to such distortions as it can be re-evaluated at each time step. Thus visiting spots that it missed now on a later pass, or skip parts that it accidentally visited too early.



(a) (b) (c)

Fig. 8: Setup for field experiments. (a) Folkestone Marine Reserve area in Barbados with several coral outcrops. The region of interest is marked with a rectangular box. (b) Scoremap generated by processing the satellite images of the region of interest. Colorbar of the scoremap indicates the interestingness or score for the presence of corals. (c) Aerial image of the ASV performing visual data sampling.

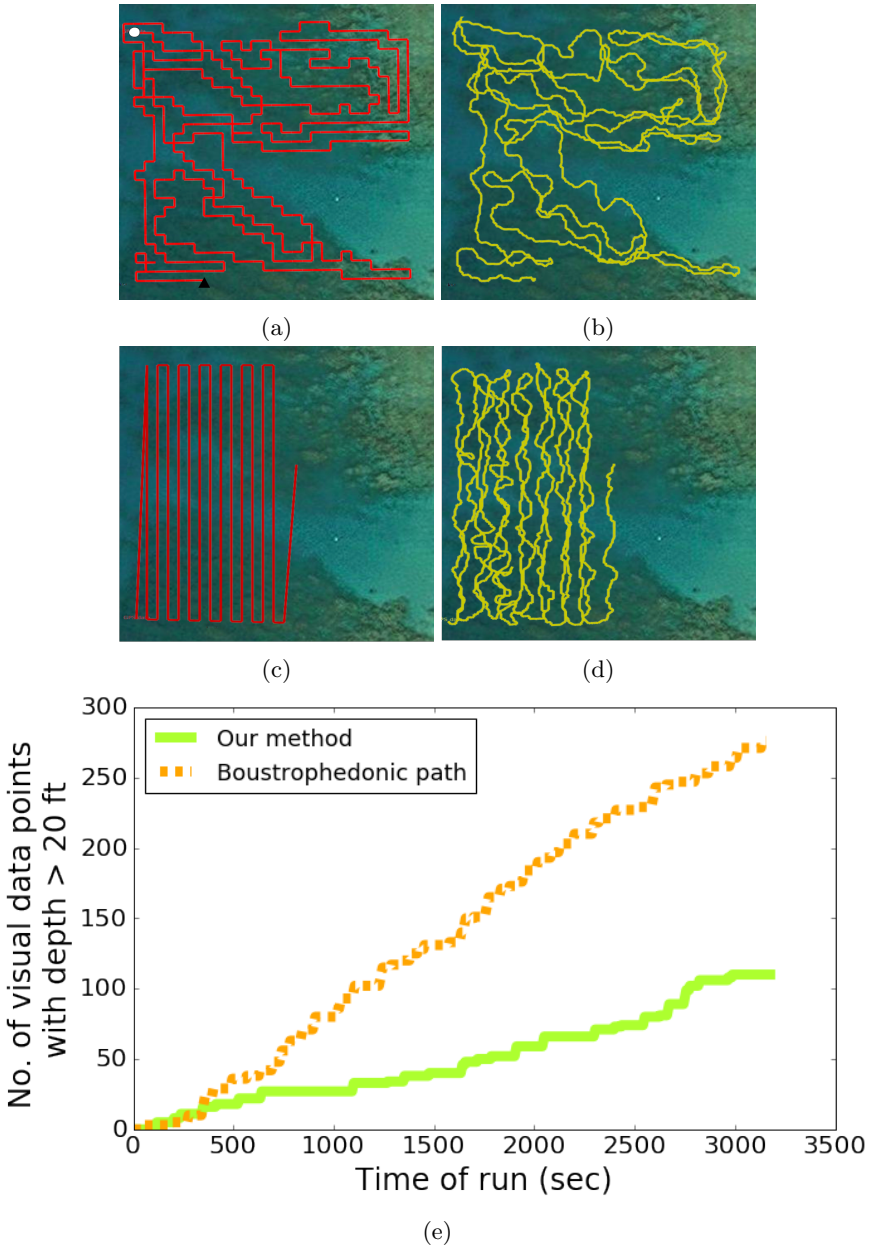


Fig. 9: Results from field experiments. (a) Planned path for visual data sampling using our non-uniform policy based sampling technique. White and black dots represent the start and end points of the path respectively. (b) The trajectory of the boat in field. The discrepancy between the planned and executed path is due to the sea-surface conditions during the trials. (c) and (d) are the planned and executed boustrophedonic paths. (e) Plot illustrating that the number of non-informative visual samples collected by a boustrophedonic sampler is almost three times the ones collected by our method.

We compare the coverage performed by our policy search method with a traditional exhaustive coverage technique using boustrophedonic path. It took 75 minutes for boustrophedonic path to completely cover the region of interest. Fig. 9c illustrates the boustrophedonic path to perform sampling for the first 40 minutes. Fig. 9e presents the total number of visual data points collected from regions which are deeper than 20 feet (i.e. visual data samples that are not useful to monitor the health of the corals) plotted against the time spent surveying the region. The total number of visual data points considered in this plot is constant for both the techniques. The comparison plot illustrates that the number of non-informative visual samples collected by a boustrophedonic sampler is more than twice the ones collected by our non-uniform policy based sampling method.

6 Conclusions and Experimental Insights

One of the novel contributions of this paper is to explore non-uniform state aggregation in policy search in the context of robotic path planning. The results suggest that such aggregation can have a major impact on the efficiency of state exploration and modeling as demonstrated by exhaustive experiments using real-but-stored data from real field deployments. We further validated this expectation in our field deployments. The incremental sampling-and-modeling paradigm we use, can be applied to many different domains where the benefits of efficient sample acquisition should accrue, but in the marine measurement domain in particular it is irrefutable that increased sampling efficiency has a major impact on the scale and feasibility of modeling efforts. For example, on the North and South Bellairs reefs where our experiments were conducted, the impact of weather and sea conditions (including tidal variations) place a significant premium on efficient sampling and have often curtained a measurement session prematurely. Likewise, this makes the use of anytime algorithms (like ours) especially important since an experiment may have to be terminated without much prior notice [1].

The direct application of our approach presupposes ongoing localization. For the near-shore ocean at moderate scales, this can be reliably achieved using traditional GPS, differential GPS, and related methods [17]. This need for ongoing localization applies to most methods, but by using non-uniform spatial sampling, it may be possible to better account for some types of pose estimation errors.

References

1. Zilberstein, S., Russell, S.J.: Anytime sensing, planning and action: A practical model for robot control. In: IJCAI. Volume 93. (1993) 1402–1407
2. Hoegh-Guldberg, O.: Climate change, coral bleaching and the future of the world's coral reefs. Marine and freshwater research **50**(8) (1999) 839–866
3. Xu, A., Viriyasuthee, C., Rekleitis, I.: Optimal complete terrain coverage using an unmanned aerial vehicle. In: IEEE Int. Conf. Robotics and Automation (ICRA). (2011) 2513–2519

4. Choset, H., Pignon, P. In: Coverage Path Planning: The Boustrophedon Cellular Decomposition. Springer London, London (1998) 203–209
5. Manjanna, S., Dudek, G.: Data-driven selective sampling for marine vehicles using multi-scale paths. In: IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS). (2017) 6111–6117
6. Low, K.H., Dolan, J.M., Khosla, P.: Adaptive multi-robot wide-area exploration and mapping. In: Proc. Int. Joint Conf. Autonomous Agents and multiagent Systems. (2008) 23–30
7. Venkataramani, R., Bresler, Y.: Perfect reconstruction formulas and bounds on aliasing error in sub-nyquist nonuniform sampling of multiband signals. IEEE Transactions on Information Theory **46**(6) (2000) 2173–2183
8. Rahimi, M., Hansen, M., Kaiser, W.J., Sukhatme, G.S., Estrin, D.: Adaptive sampling for environmental field estimation using robotic sensors. In: Int. Conf. Intelligent Robots and Systems (IROS), IEEE (2005) 3692–3698
9. Sadat, S.A., Wawerla, J., Vaughan, R.: Fractal trajectories for online non-uniform aerial coverage. In: IEEE International Conference on Robotics and Automation (ICRA). (2015) 2971–2976
10. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. In: Reinforcement Learning. Springer (1992) 5–32
11. Baxter, J., Bartlett, P.L.: Infinite-horizon policy-gradient estimation. Journal of Artificial Intelligence Research **15** (2001) 319–350
12. Sutton, R.S., McAllester, D.A., Singh, S.P., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. In: Advances in neural information processing systems. (2000) 1057–1063
13. Deisenroth, M.P., Neumann, G., Peters, J.: A survey on policy search for robotics. Foundations and Trends® in Robotics **2**(1–2) (2013) 1–142
14. Kober, J., Bagnell, J.A., Peters, J.: Reinforcement learning in robotics: A survey. The International Journal of Robotics Research **32**(11) (2013) 1238–1274
15. Barto, A.G., Bradtke, S.J., Singh, S.P.: Real-time learning and control using asynchronous dynamic programming. University of Massachusetts at Amherst, Department of Computer and Information Science (1991)
16. Singh, S.P., Jaakkola, T., Jordan, M.I.: Reinforcement learning with soft state aggregation. In: Advances in neural information processing systems. (1995) 361–368
17. Dudek, G., Jenkin, M.: Inertial sensing, gps, and odometry. In: Siciliano B., Khatib O. (eds) Springer Handbook of Robotics. Springer (2016) 477–490