Unsupervised Spatial-Semantic Maps for Human-Robot Collaboration in Communication-Constrained Environments

Kevin Doherty[1,2] and Yogesh Girdhar[2]
{kdoherty@mit.edu, yogi@whoi.edu}
[1]Massachusetts Institute of Technology, Cambridge, MA
[2]Woods Hole Oceanographic Institution, Woods Hole, MA

Introduction

We seek to enable collaborative exploration between a marine robot and a scientist, where the robot must observe and report about undersea phenomena while making exploratory decisions in cooperation with a scientist. Streaming video is not feasible due to several unique constraints imposed by the marine environment:

- **Limited communication** via acoustics.
- **High-level plans require spatial context** that is not provided by video streams.

We propose spatial-semantic maps as a data-efficient representation of a marine robot's visual observations. These maps can be streamed even over limited communication channels to facilitate high-level mission specification by a scientist and planning by a marine robot.

Methods

Our spatial-semantic maps are built using Bayesian nonparametric inference and use a "bag of words" model, where images are represented by a set of visual "words" (i.e. image features).

Real-time Spatiotemporal Topic Models

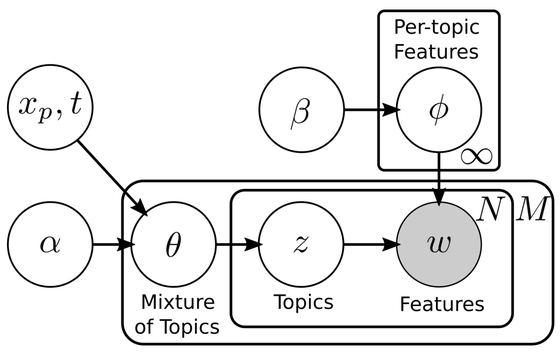


Figure 1: Graphical model of the ROST framework.

We use real-time online spatiotemporal topic modeling (ROST) [4] to build a map of topics. ROST models the probability of word w_i being v, given its location in the image x_p and time t as follows:

$$P(w_i = v|x_p, t) = \sum_{k=1}^{K} P(w_i = v|z_i = k)P(z_i = k|x_p, t). \quad (1)$$

Given this generative model (see Figure 1), we use Gibbs sampling to estimate the posterior probability of a topic assignment for a given visual word.

Building the Spatial-Semantic Map

We incorporate spatial relationships between geographic locations at which observations were made into our maps by applying a multi-class extension of the kernel-based mapping approach in [2].

$$P(z_i = k|x) = \frac{\alpha_i + \sum_{j=1}^{M} k(x, x_j)\mathbb{1}\{z_i = k\}}{\sum_{k=1}^{K} \alpha_k + \sum_{j=1}^{M} k(x, x_j)}, \quad (2)$$

where x is the location for which a prediction was made, α_k are topic-wise hyperparameters. $k(\cdot, \cdot)$ is the radial-basis function (RBF) kernel.

Anomaly Maps

We can also compute the perplexity of new observations given the current topic model, which allows us to assess the relative distinctiveness of an image given what we have observed. The topic perplexity of a new document is given by:

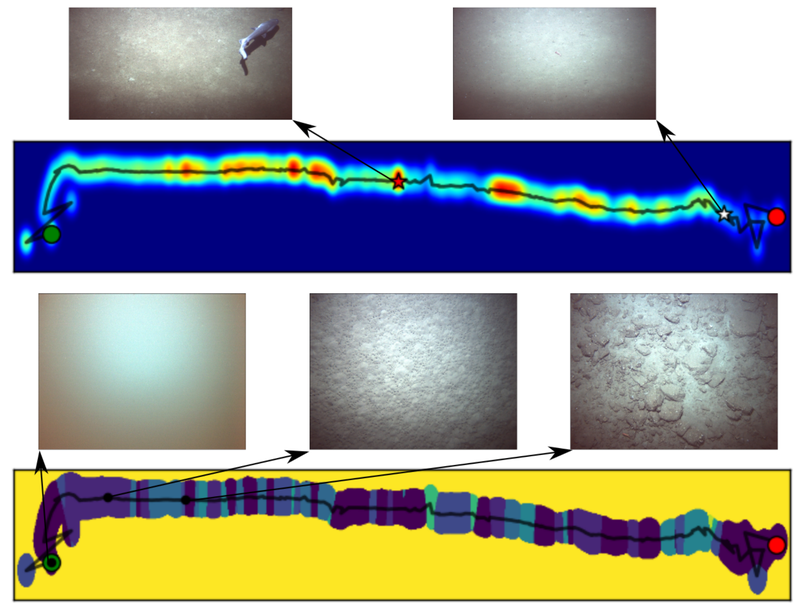
$$\rho_{d_t} = \exp\left(\frac{-\sum_{i=1}^{N} P(z_i = k|d_{1:t-1})}{N}\right), \quad (3)$$

where $d_{1:t-1}$ is the set of previously observed images.

Topic and Anomaly Maps

We show the topic and anomaly maps inferred using the ROST model from video collected by a SeaBED AUV performing a linear transect at the Hannibal Bank Seamount off the coast of Panama (right). Localization was performed using an ultra-short baseline (USBL) system. These maps have the following advantages:

- **Dense, compact representation** allows efficient communication.
- **Easily localize anomalies** like sharks, while ignoring sand.
- **Unsupervised categorization** removes the need for large, annotated datasets.
- **Topic maps provide situational awareness** by showing category predictions distributed over space.



Learning a Visual Vocabulary Using Generative Adversarial Networks

Issues with the Approach

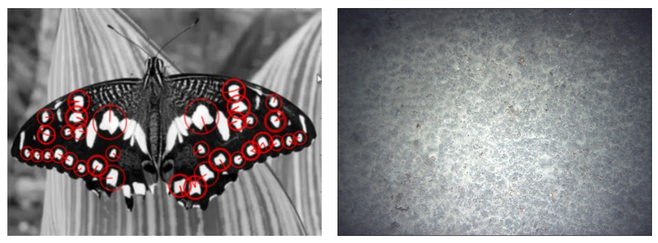


Figure 2: Example SURF features (left). Visually sparse seafloor (right).

1. **Underwater scenes are poorly described by hand-crafted features (see Figure 2).**
2. **Standard features have little interpretability.**
3. **Visual vocabularies built using high-dimensional clustering.**

Motivated by the recent success of convolutional autoencoders for feature learning in underwater environments [3], we propose learning interpretable image patch features from data.

Learning a Vocabulary of Image Patches

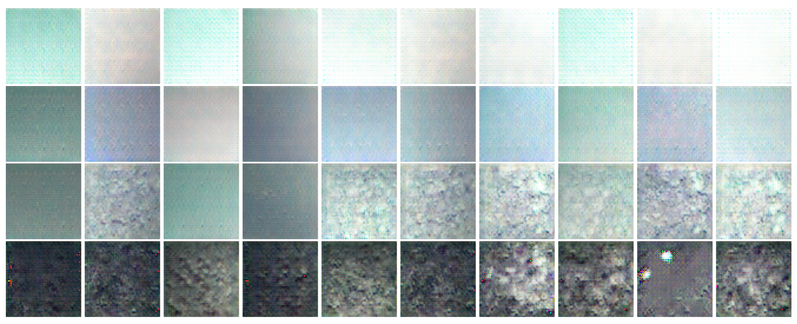


Figure 3: Synthetic samples generated by the network. Different Rows correspond to different categories learned by the model.

To tackle the problem of building a probabilistic representation over the high-dimensional space of images, we employ generative adversarial networks (GANs). An overview of our current strategy is as follows:

1. **Learn texture categories from data using InfoGAN** [1]. Figure 3 shows synthetic images of textures generated by the model.
2. **Classify patches** from raw image using InfoGAN.
3. **Model topics over types of image patches** rather than hand-crafted features.

Information Maximizing GANs

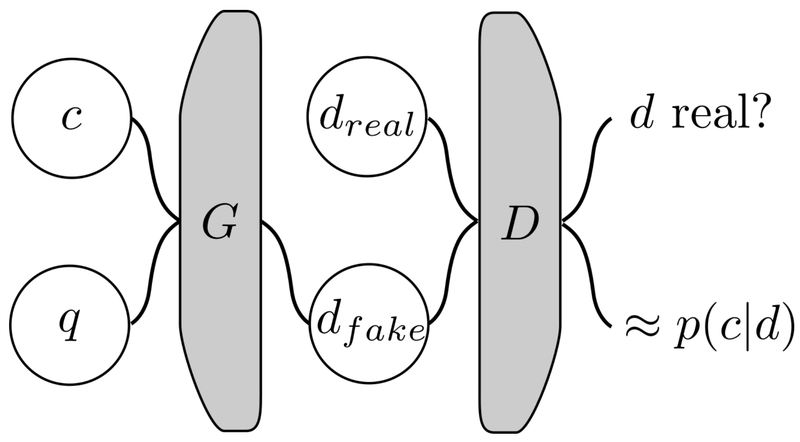


Figure 3: Graphical depiction of the InfoGAN model architecture.

A generative adversarial network can be described by a two-player minimax game between a decoder (generator) and an encoder (discriminator) $\min_G \max_D V(D, G)$, where we have

$$V(D, G) = \mathbb{E}_{\boldsymbol{d}\sim p_{data}(\boldsymbol{d})}[\log D(\boldsymbol{d})] + \mathbb{E}_{\boldsymbol{q}\sim p_G(\boldsymbol{q})}[\log(1 - D(G(\boldsymbol{q})))]. \quad (4)$$

Information maximizing GANs (InfoGAN) can learn latent codes (e.g. as specified by a categorical distribution) by adapting the value function from Equation 4 as follows:

$$V_I(D, G) = V(D, G) - \lambda I(\boldsymbol{d}; G(\boldsymbol{c}, \boldsymbol{q})), \quad (5)$$

where λ is a regularization factor and $I(\boldsymbol{d}; G(\boldsymbol{c}, \boldsymbol{q}))$ is the mutual information between the latent codes and the generator distribution.

To compute the mutual information, InfoGAN provides an approximation of the posterior distribution over codes given the image; $p(c|d)$. We are interested in the latent codes $c \sim \text{Cat}(V)$ to map from image patches to a categorical vocabulary.

InfoGAN solves the task of clustering high-dimensional image inputs into discrete visual words that can be used in a topic model. We believe this will help mitigate several of the issues with the use of standard features in the underwater domain.

References

[1] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. "Infogan: Interpretable representation learning by information maximizing generative adversarial nets." In Advances in Neural Information Processing Systems, pp. 2172-2180. 2016.

[2] K. Doherty, J. Wang, and B. Englot. "Bayesian generalized kernel inference for occupancy map prediction." *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 3118-3124. 2017.

[3] G. Flaspohler, N. Roy, and Y. Girdhar. "Feature discovery and visualization of robot mission data using convolutional autoencoders and Bayesian nonparametric topic modeling." *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, 2017.

[4] Y. Girdhar, P. Giguère, and G. Dudek. "Autonomous adaptive exploration using realtime online spatiotemporal topic modeling." *The International Journal of Robotics Research*, 33(4), pp. 645-657. 2014.