Learning Deep-Sea Substrate Types With Visual Topic Models

Arnold Kalmbach
McGill University
akalmbach@cim.mcgill.ca

Maia Hoeberechts
Ocean Networks Canada
maiah@uvic.ca

Alexandra Branzan Albu
University of Victoria
aalbu@uvic.ca

Hervé Glotin, Sébastien Paris
LSIS AMU ENSAM CNRS Univ. Toulon
glotin@univ-tln.fr, sebastien.paris@lsis.org

Yogesh Girdhar
Woods Hole Oceanographic Inst.
ygirdhar@whoi.edu

Abstract

We propose and evaluate a method for learning deep-sea substrate types using video recorded with a remotely operated vehicle (ROV). The goal of this work is to create a labelled spatial map of substrate types from ROV video in order to support biological and geological domain research. The output of our method describes the mixtures of geological features such as sediment and types of lava flow in images taken at a set of points chosen from an ROV dive. The main contribution of this work is the assembly of a pipeline combining several unique approaches which is able to robustly generate substrate type mixtures under the varying lighting and perspective conditions of deep-sea ROV dive videos. The pipeline comprises three main components: sampling*, in which a trained classifier and spatial sampling is used to select relevant frames from the dataset;* feature extraction*, in which the improved local binary pattern descriptor (ILBP) is used to generate a Bag of Words (BoW) representation of the dataset; and* topic modelling *in which a variant of Latent Dirichlet Allocation (LDA), is used to infer the mixture of substrate types represented by each BoW. Our method significantly outperforms techniques relying on keypoint based features rather than texture based features, and k-means rather than LDA, demonstrating that our proposed pipeline accurately learns and identifies visible substrate types.*

1. Introduction

Substrate classification, the task of creating a spatial description of the nature of the seabed, is a fundamental factor in many aspects of ocean research. Domain research in marine biology, physical oceanography and geology—including classifying benthic habitat, modelling deep-sea circulation and analyzing tectonic motion—depends on accurate classification of substrate. In the context of the deep-sea, basic questions remain unanswered about what terrain can be expected, particularly in geologically diverse mid-ocean ridge environments. High exploration cost and difficulty of sampling drive the need for remote sensing options for data acquisition, including visual and acoustic surveys, which in turn generate large volumes of data requiring analysis.

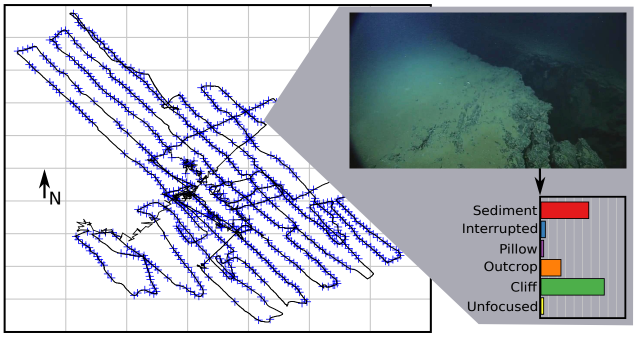


Figure 1: Our system detects substrate in deep-sea ROV video, learns the visible substrate types without supervision, and finds the mixture of types in each sample.

Manual analysis of substrate type is time consuming and requires geological expertise. Subjective factors, such as the choice of salient environmental features, make manual analysis for multidisciplinary use subject to observer bias. For these reasons, automatic data-driven systems are an attractive option for substrate classification. In recent years it has become common to infer seabed composition from acoustic backscatter reflectance by measuring proxy variables like hardness or rugosity. However, in complex, geologically diverse environments, typical spatial resolutions for acoustic methods are insufficient to capture the rapidly changing substrate. Video data from ROVs excels in these types of environments, as a local, high-rate source of environmental information. For human analysts, video has the added advantage that it can be intuitively interpreted and directly used for multidisciplinary purposes, such as counting organisms for diversity and abundance analysis. In this research, we present an unsupervised method to classify substrate in video recorded at the Endeavour Segment of the Juan de Fuca ridge, a volcanic-hydrothermal environment about 300 km west of Vancouver Island, Canada.

Developing a system that can extract information about substrate types in a large range of environments and video quality conditions is challenging. Because of the high-cost of manual labelling, producing training sets of sufficient size for supervised methods is often impractical. In addition, the range of relevant substrate types across different environments is large, meaning that new training sets must be developed for each new area to be mapped. Models are further complicated because a single substrate type is very rarely the only significant substrate feature in an image. In this common situation, models that produce a single label do not adequately capture the complexity or continuity of the substrate, and therefore a model of the mixture of substrate types in each frame is necessary. Finally, due to the high cost of acquisition, and consequent lack of availability of data from deep-sea environments, practical approaches must perform well on data recorded without the constraints of a photogrammetric survey such as consistent lighting, focus, and perspective. This ensures that the method is as widely applicable as possible, but also introduces a set of significant challenges.

Contributions: We present a system that learns substrate types and maps their locations from the video and navigation data recorded on a deep-sea visual survey. The main contribution of the work is the assembly of a pipeline combining several approaches, which is able to robustly generate substrate labels under the varying lighting and perspective conditions that are typical for deep-sea ROV dive videos. This system requires minimal training data; instead it uses a topic-model to infer a small set of topics which each form a sparse probabilistic representation of a substrate category. For each frame in the video, a sparse distribution of the topics is simultaneously inferred, representing a mixture of the substrate categories. Topic models work on a bag-of-words (BoW) representation of data. Therefore, our method includes a sequence of steps aimed at producing a visual BoW that represents the substrate visible in an ROV dive. We demonstrate that our method produces topics that correlate highly with the true geological substrate types, outperforming classic unsupervised methods on a deep-sea video survey. Our method was developed for a particular a mid-ocean ridge video dataset, however initial experiments show promise in generalizing the approach to substrate classification for other datasets and environments.

2. Related Work

As automatic substrate classification has become a topic of interest, there have been a number of computer vision approaches for measuring substrate-related variables.

[9] proposes a method to measure seabed complexity by segmentation, and applies a random forest classifier on those segments to identify certain objects in a supervised manner. [16] develops a method based on Self Organizing Maps (SoM) to learn a feature representation for segmenting seabed images, and shows that they successfully identify metallic nodules with a simple supervised classifier in the learned feature space. These represent sophisticated supervised methods that can be useful in certain situations, however their reliance on training is problematic in contexts where ground-truth is time consuming to produce.

In [15, 14, 18] Pizarro et al. describe and demonstrate a system that performs visual habitat classification using topic models on coastal coral reef environments. Their approach performs clustering in topic space, using the topic assignments of labelled images as cluster centers. In contrast, our system learns topics that are themselves representative of substrate types. Our interpretation of the topics is more sensitive to the quality of the images and their BoW model, but has the key benefit of naturally providing multiple dimensions of comparison rather than using a single score for similarity between images.

Our approach is based on the Realtime Online SpatioTemporal topic model (ROST), developed in [4]. Details on how we use this model are described in Section 5. [5] demonstrates an application of ROST to terrain classification problems, seeking to maximize the diversity of terrain covered in a small amount of video. That work focuses on learning the topic-model in real-time, and using it for an efficient coverage strategy, spending more time surveying 'interesting' terrain. The method is demonstrated on satellite imagery and in a shallow reef environment. [6] demonstrates an application of ROST to classification of deep-sea video. In both these systems, the authors have had a relatively high degree of control over the quality of their video and focus on efficiently learning the topic model. In contrast, we emphasize developing a full and robust feature extraction pipeline that can accommodate video recorded under less than ideal photogrammetric conditions.

3. System Overview

Our system consists of a pipeline for learning and mapping the substrate types present in the video and navigation data from an ROV dive.

First (see Figure 2a), our system detects the frames containing substrate using a Support Vector Machine (SVM) trained on GIST descriptors, a representation of the overall shape of each image. of each image. We then select

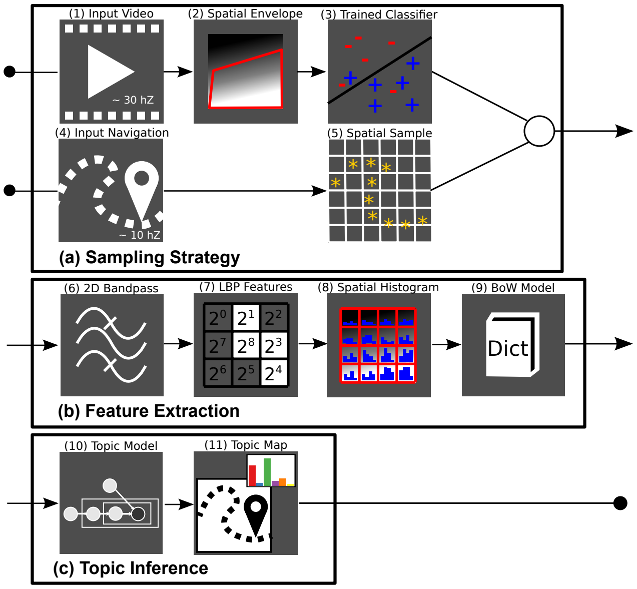


Figure 2: **System Overview.** From top to bottom: (a) Sampling Strategy (Section 4.1), (b) Feature Extraction (Section 4.2), (c) Topic Inference (Section 5).

a spatially distributed subset of these frames to include in our BoW model. Next (see Figure 2b), we extract textural information from the sampled video by using a noise suppression heuristic followed by extracting ILBP image descriptors, and by constructing a vocabulary of the most discriminative codes. Finally (see Figure 2c), we fit a spatiotemporal topic-model using the BoW from each image separately. We interpret the topic mixtures produced by the topic-model as the percentage of each substrate type present in that image.

The remainder of the paper is organized as follows: Section 4 explains the process used for building the BoW model, including choosing a sample set (Section 4.1), and extracting ILBP descriptors (Section 4.2). Section 5 describes the topic-model and its applicability to the task at hand in more detail. Section 6 reports the details of the experimental dataset represented by the the East Flank video survey and demonstrates our method's performance on this dataset.

4. Substrate Bag-of-Words Model

In order to simplify the challenges of learning a meaningful, low-dimensional model of the substrate, our system builds a BoW representation of the video. This process consists of three steps: First, our system selects a set of sample frames from the dive video so that most sample frames have substrate, and so that no region is oversampled. Second, our system extracts features from each of the samples. We choose ILBP as a powerful and efficient texture descriptor. In order to ensure that the descriptors represent the substrate rather than other visual features, we apply noise reduction in the form of a 2D bandpass filter before extracting textureal information. Third, our system selects a vocabulary from the feature space, choosing and discretizing the most discriminative ILBP feature dimensions by their variance across the samples.

4.1. Sampling Strategy

The sampling strategy consists of two steps: First, our system uses an SVM to detect frames that contain substrate. Then, it selects a subset of those frames based their corresponding spatial location, ensuring that samples are as evenly distributed as possible in space. (See Figure 2a)

4.1.1 Spatial Envelope Substrate Detector

Unedited video from ROV dives contains large portions that is irrelevant for substrate classification. For instance, in the East Flank video survey 847 out of 2000 manually assessed frames randomly selected from the dive contained no substrate whatsoever (See Table 1a). This data is problematic for unsupervised classification because it skews the distribution of features. To mitigate this problem, we remove frames which do not contain substrate by training an SVM on the spatial envelopes of relevant and irrelevant frames.

By spatial envelope, we refer to the holistic representation of the shape of a scene implemented as the GIST descriptor described in [12]. GIST descriptors have been shown to encode high-level perceptual dimensions of the spatial envelope such as 'naturalness', 'openness', 'roughness' etc. This descriptor is computed using a method based on Gabor filter responses in multiple orientations and scales, and in a grid of image windows across the image. GIST descriptors are appropriate as the local appearance of images with substrate varies dramatically, and the spatial envelope is more important that any local features.

We manually labelled 2000 randomly selected frames as either relevant or irrelevant (i.e. with or without substrate) for our subsequent classification problem. We trained an SVM using 1000 frames randomly selected from these 2000 labelled examples, and evaluated its performance on the other 1000. We used the MATLAB function `fitcsvm` with default parameters as the SVM implementation [11]. Table 1b shows the precision and recall computed over all frames in the test set. The results indicate that this method successfully discriminates frames with substrate from frames without it.

4.1.2 Spatial Subsampling

Frames containing substrate are distributed unevenly in space—if one frame contains substrate it is likely that in

	Substrate	No Substrate
Train Set (True)	564	436
Test Set (True)	589	411
Full Dataset (Estimated)	6859	3141

(a) Number of relevant and irrelevant frames in our example video.

	Precision	Recall	F1-Score
Train Set	0.7942	0.9645	0.8711
Test Set	**0.8117**	**0.9440**	**0.8729**

(b) Performance of GIST SVM substrate detector.

Table 1: GIST SVM performance

the adjacent frames the ROV was in a similar location and orientation, and those frames will contain substrate as well. In addition, during an ROV dive the pilot may have sped up, slowed down, or completely stopped in response to equipment conditions or mission objectives other than the visual survey. To ensure the distribution of samples matching the true distribution of substrate types, we take a spatial subsampling of the frames with substrate.

Using the ROV navigational data, the spatial subsampling step selects a subset of the navigation points that are all no less than a distance threshold away from from one-another. Using the Latitude and Longitude measurements at each sample in our dataset, our system computes local Northings and Eastings (meters N and E from the start). Then, starting at the first frame, it adds samples to the spatial subsample one at a time, provided that they are not within a given radius of any point already in the subsample. The radius parameter provides a means of controlling the computational costs of our system. Choosing an appropriate value is a trade-off between the resolution and processing time required for the resulting model.

4.2. Feature Extraction

We observe that image textures provide a more natural feature representation of substrate types than other standard representations such as keypoints or color histograms. Whereas keypoint-based descriptors are too sensitive to the presence of particular objects and color-based descriptors are too sensitive to lighting conditions, texture-based descriptors give a high-level description of the set of patterns that compose a scene.

We choose to describe the textures of our video with ILBP, a variant of the Local Binary Pattern that accommodates scale and has improved resilience to noise. To ensure that the extracted textures describe the substrate and not the lighting conditions or particles in the water column, we apply a heuristic in the form of a spatial bandpass filter. (Figure 2b)

4.2.1 Improved Local Binary Patterns

Local Binary Patterns represent textures by encoding the local brightness variations in 9-pixel squares as compact binary codes. Because they can be computed efficiently on integral images, they are a popular alternative where other texture representation approaches like Gabor filter banks are too slow. ILBP, originally introduced under the name Multi-Block LBP in [10], extends LBP, firstly by accommodating scaled pixel areas, and secondly by adding information about the center pixel. Whereas in LBP each pixel is compared to the center pixel in a 3x3 pixel region, in ILBP, each pixel, including the center pixel, is compared to the mean value over an $N \times N$ pixel region. Liao et al. have shown that this improves the robustness of LBP as well as the ability to encode larger image structures. We use mlhspyr_lbp, a multi-channel, multi-scale, windowed ILBP implementation described in [13].

4.2.2 De-noising Heuristic

Even when deep-sea video is recorded under ideal conditions its textures are affected by significant noise. Because the ROV itself carries the only light source, the relative orientation of the ROV and the substrate affects which areas of the frame are well-lit and which areas are dark. In terrains with high-relief, there is also significant shadowing. In addition, suspended particles in the water column contribute significant point-based noise.

To reduce these effects, our system applies a spatial bandpass filter before extracting features. Specifically, it uses a 2D Gaussian bandpass filter with the kernel

$$K = (1 - e^{-(x^2+y^2)/2c^2_{lowcut}})(e^{-(x^2+y^2)/2c^2_{highcut}}) \quad (1)$$

The parameters c_{lowcut} and $c_{highcut}$ determine the minimum and maximum frequencies represented in the filtered image in cycles/image. We chose the values 4 and 100 cycles/image respectively for these parameters based on the approximate minimum and maximum size of substrate features of interest in the video.

4.3. From Histogram of ILBP to Words

Our feature extractor outputs one histogram per image—each bin corresponding to an LBP code in a particular region, at a particular scale and for a particular color channel. Since we are interested in encoding the distribution of textures rather than their spatial arrangement in the images, our system sums over the corresponding LBP codes in each window. This results in a large (i.e. 10,752 dimensional) histogram, of which only relatively few bins vary across the dataset.

Our system selects only the V bins with the highest variance with respect to their mean across the set of sample images in the BoW. The number of occurrences in each of these bins represents the number of occurrences of a word in our model. The bins which are not considered do not contribute to the BoW as their occurrence provides less information discriminating between the sample images. Choosing the vocabulary size parameter V is a tradeoff between the size of the inference problem and the amount of detailed information that should be considered. We sorted the bins by their variances and found that the relationship between rank and variance was well fit by a negative exponential function. This implies that most of the overall variation was accounted for by only a relatively small fraction of the bins, and that above a small minimum value for V, the BoW representation is very similar. We found that setting $V = 3000$ produced a tractable inference problem, and that deviation within an order of magnitude larger did not significantly affect the quality of the results.

5. Topic Model for Substrate Mapping

Topic models are a family of Bayesian probabilistic models, originally introduced in the context of semantic classification of text corpora. They have been shown to be suitable for many domains where an unsupervised semantic clustering is desired and an appropriate BoW representation of the data is available [1]. We use an extension of Latent Dirichlet Allocation (LDA) in our system.

LDA is used to infer a probabilistic representation of the hidden semantics of a collection of BoW called documents. Each document is modelled by a probability distribution over a fixed number of topics. Each topic, in turn, is defined as a probability distribution over the set of all possible words. In computer vision applications, the set of words is defined as the set of discrete features representable in the BoW model (defined in Section 4 for our application) rather than a set of literal words, and each BoW is treated as a document. The LDA model assumes that each word in a document was created by first sampling a topic from the document-topic distribution, and then sampling a feature from the corresponding topic-word distribution.

The goal of LDA is to estimate these two distributions given the document-word distributions. These distributions can be estimated using Gibbs sampling or other Markov Chain Monte-Carlo methods. By choosing Dirichlet priors for both sets of distributions, the probability that a word was generated from a topic can be efficiently estimated using count variables only. In addition, the choice of a symmetric Dirichlet prior allows for control over the sparsity of the distributions through the hyperparameters α and β. Intuitively, the hyperparameters can be understood as controlling how many different words will have high probability in each topic and how many different topics will have high probability for each document. There have been numerous successful examples of visual topic-models for classification and segmentation of natural scenes [2, 17, 19].

Our system uses ROST, which improves upon LDA in several ways. First, ROST takes advantage of the spatiotemporal context of the observations to compute the prior distribution of the topic labels, resulting in more accurate generative model for observations that have spatiotemporal context. Second, the Gibbs sampler proposed by ROST [7] has been optimized so that it can process streaming observation data in realtime.

6. Results

6.1. Dataset

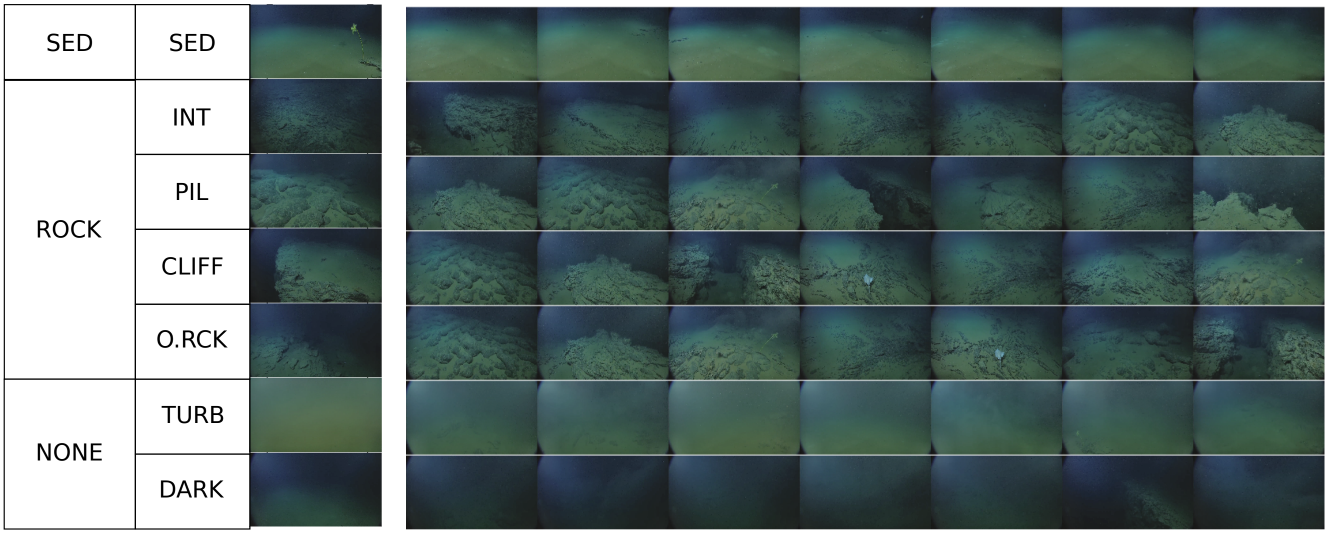
We validate our method using data recorded at the Endeavour Segment of the Juan de Fuca Ridge, a mid-ocean ridge environment 300 km West of Vancouver Island, British Columbia, at approximately 2.2 km depth. Specifically, our data is from the East Flank of Endeavour, a gently sloping area featuring a variety of substrate types. Mission objectives on this dive were to perform a visual survey for suitability of a scientific instrument installation—much of the video shows substrate, but distance, angle, and speed are inconsistent throughout the recording.

The data consist of 30 fps HD video and 10 Hz ultra-short baseline (USBL) acoustic positioning. The ROV moved at an average rate of 0.5 knots (approx. 0.26 m/s), through two partially overlapping lawnmower patterns, one up and down a gentle slope, and one across the flat area at the base of the slope, with a total distance travelled of just over 5.6 km.

With guidance from an expert in deep-sea geology, we defined seven categories that represent the types of images seen in the sample. These categories were 'Sedimented' (SED), 'Interrupted Lava Flow' (INT), Pillow Lava Flow' (PIL), 'Cliff or Wall' (CLIFF), 'Other Rock' (O.RCK), 'Turbid Water' (TURB), and 'Substrate out of Range' (DARK). We randomly selected 500 frames from the dataset, and labeled each with a probability distribution over these seven types, using a minimum increment of 0.25 for each category. Images exemplifying each category can be seen in Figure fig:best1gt.

6.2. Experimental Results

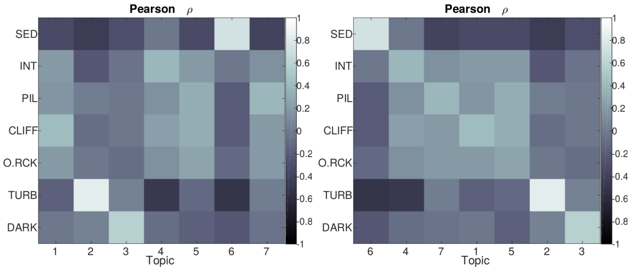
We generated a BoW representation for each frame in the sampling using the histogram of ILBP codes described in Section 4. We used scale factor 2 for the ILBP region with a uniform 6x7 grid of non-overlapping windows. Initial experimentation showed that these values produced good results, and that additional scales did not cause significant improvement. We then ran the topic model on the BoWs, choosing 7 for the number of topics (the same as the number of true categories). We ran LDA using each pair of the values $\{0.01, 0.1, 0.2, 0.4, 0.8, 0.9, 0.99\}$ for both hyperpa-



(a)

(b)

Figure 3: (a) Representative examples of the categories (top to bottom) ‘Sedimented’, ‘Interrupted Lava Flow’ ‘Pillow Lava Flow’, ‘Cliff or Wall’, ‘Other Rock’, ‘Turbid Water’, and ‘Substrate out of Range’. (b) Images with highest proportion of words assigned to each topic. The rows contain the best 7 images for the topics paired with the corresponding category in (a)



(a)

(b)

Figure 4: Correlation matrix between ground-truth categories (rows) and topics (columns). In (a) topics are shown unordered. In (b) topics are re-ordered so that the topic in column i is paired with the category in row i for a maximum-correlation pairing.

rameters α and β. We report the performance of the model with minimum final perplexity, $\alpha = 0.1$, $\beta = 0.8$. Initial evaluation showed that changing the hyperparameters did not have a strong impact on the final model.

To evaluate the degree to which each ground-truth category was represented by some topic generated by our algorithm, we produced a one-to-one pairing between categories and topics. First, we calculated the Pearson ρ correlation between each ground-truth category and each topic produced by our algorithm using all 500 of the ground-truth frames (Figure 4a). Then, using the Kuhn-Munkres Algorithm [8], we found an optimal assignment of topics to categories. Defining the cost of assigning category i with topic j as $Cost(i, j) = 1 - \rho_{i,j}$ results in an assignment with maximum total correlation. Figure 4b shows the correlations after re-ordering the topics so that the topic assigned to category i is in the ith column. The high-correlation along the diagonal in this matrix and relatively low correlation everywhere else shows that topics correlated well with few categories in most cases.

Figure 3b shows the most representative frames for each topic. These images show the 7 frames that have the highest proportion of each topic throughout the dataset. Note that some topics were much more prevalent than others, so an image may be among the best examples of a particular topic in the dataset without that topic being the top label for the image. For instance, many of the strongest examples of interrupted lava flows contain more sediment than interrupted lava flow. Comparison between these examples and the examples in Figure 3a suggests that the topics assigned to sediment, turbid water, and substrate out of range are accurately recovering features of their categories, and that the other assignments are somewhat less accurate.

We compare the correlations for the assigned topic-category pairs against correlations computed similarly for three baseline strategies: (1) SIFT+K-Means - For each image we compute dense SIFT features at keypoints on a 100 × 100 grid, then, using the features from 10% of the frames chosen randomly, we quantize the feature vectors into a dictionary of size 3000, and replace each of the

	Sediment	Interrupted	Pillow	Cliff	Other Rock	Turbid Water	Out of Range
SIFT+K-Means	0.2898	0.0718	0.4162	0.1721	-0.0137	-0.0271	0.3630
LBP+K-Means	0.3526	0.0092	-0.0029	0.3474	0.0508	0.6385	0.4971
LBP+Filt+Discr.+K-Means	0.3661	0.1116	0.0849	0.2916	-0.0685	0.7509	0.3466
LBP+Filt+Discr.+Topic-Model	**0.7489**	**0.3899**	**0.3884**	**0.4152**	**0.2580**	**0.8153**	**0.5664**

(a) Pearson $\rho\,(498)$ for best-match topic with each category. For LBP+Filt+Discr.+Topic-Model p-value was $\ll 0.001$.

	Sediment	Rock	No Substrate
SIFT+K-Means	0.4413	0.3492	0.0459
LBP+K-Means	0.5059	0.3541	0.5770
LBP+Filt+Discr.+K-Means	0.5350	0.4930	0.4601
LBP+Filt+Discr.+Topic-Model	**0.7489**	**0.7156**	**0.8015**

(b) Pearson $\rho\,(498)$ for best-match topic with each high-level category. For LBP+Filt+Discr.+Topic-Model p-value was $\ll 0.001$.

Table 2

original feature vectors with its closest neighbor in the dictionary. We compute the histogram of quantized features for each frame, and cluster the feature histograms using K-Means again. (2) LBP+K-Means - We compute ILBP feature histograms for each frame, using mlhspyr_lbp without the noise reduction steps described in Sections 4.2.2 and 4.3. We cluster the ILBP histograms for each frame using K-Means. (3) LBP+Filt+Discr.+K-Means We compute the document histograms used as input to our topic model, using all steps described above including noise suppression and selecting only the discriminative histogram bins, but cluster them with K-Means rather than ROST. For each of these three strategies, we compute the correlation of each cluster with each category from the ground-truth, and produce an optimal assignment.

Table 2 shows the values of the correlation for each topic-category assignment. This table shows the degree to which the value of each category was correlated to the value of its assigned topic for all of the ground-truth frames. Strong correlation for a category-topic pair suggests that the value of the topic was proportional to the value of the category in most frames. This analysis reflects the quality of the mixtures produced by our system rather than just the quality of the single top category in each frame.

Our method outperformed the three baseline methods, having the highest correlation on 6 out of 7 categories. SIFT+K-Means produced a cluster with high correlation with the pillow category, but its performance on all other categories was poor. These values show a median strength of association 2.5 times larger between topics and categories than beteen SIFT clusters and categories (strength of association is defined as the absolute value of correlation).

Note that for our topic-model approach, p-values were $\ll 0.001$, whereas for the other approaches, p-values varied, sometimes taking significantly higher values. These p-values represent a strong rejection of the hypothesis that the topics were not correlated with the categories for our method, and a weaker one for the baseline strategies.

In absolute terms, the values of the correlations reflect the intuition given by the best examples of each topic: The categories sediment, turbid water, and no substrate are each strongly correlated with their assigned topic, but the other categories are only weakly correlated with their assignment. Therefore, we additionally analyze our system's performance classifying the high-level categories 'Sediment', 'Rocky', and 'No Substrate'. These categories were constructed by combining the original categories, as seen in the leftmost column of Figure 3a. We combined the ground-truth labels, topics, and labels for the three baseline methods by summing over the groups to be combined into each of the three categories. The performance of the resulting models is presented in Table 2b, showing that our method has strong correlations between the combined topics and all three of the high-level categories.

7. Discussion

Our results constitute a preliminary success in unsupervised substrate type mapping on a challenging video dataset. The categories not consistently identified by our system were semantically overlapping, and to some degree visually similar. This has allowed us to make use of the system as-is to map higher-level substrate categories. Although less specific than the original categories, this still represents interesting and useful data to ocean scientists. We present a map of these substrate types in Figure 5. Maps of this type provide an interface between our method and biological or geological research which seeks to use information about substrate type. For instance, this map could be used in conjunction with a map of observations of a certain species to help answer questions about how habitat suitability is related to substrate type. Note that the substrate type mixtures are spatially consistent and vary smoothly in adjacent sam-

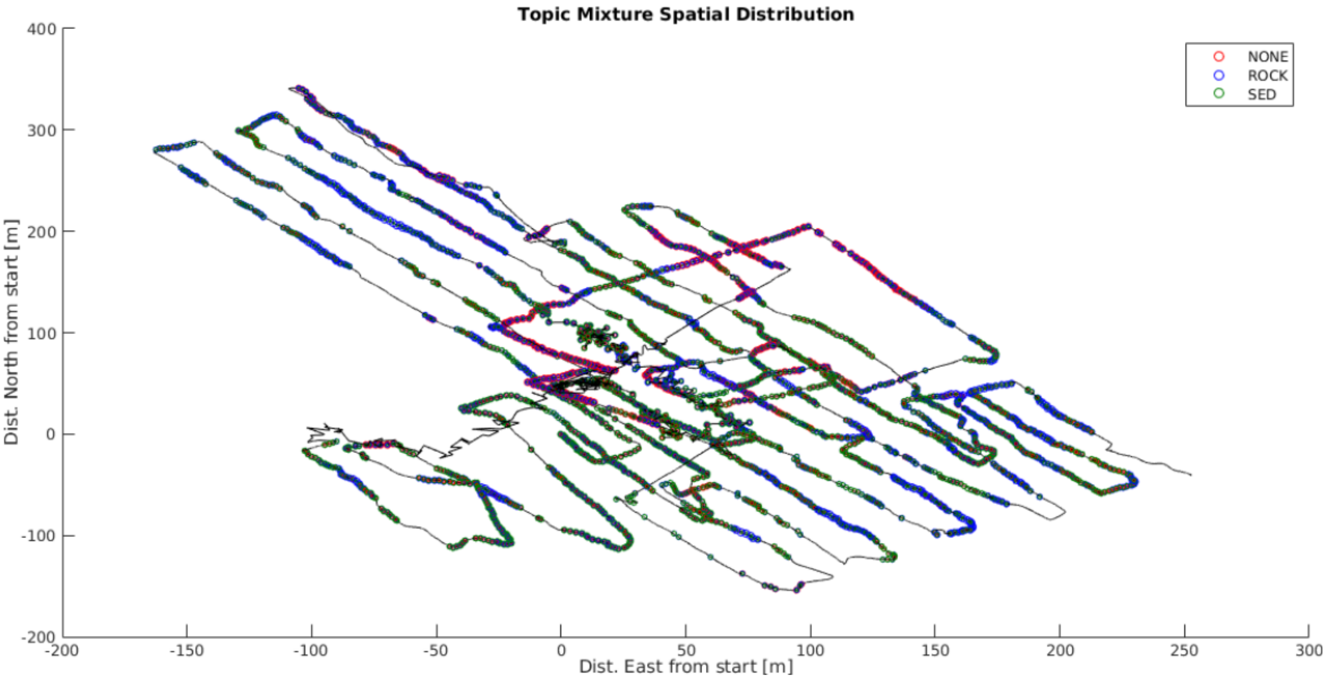


Figure 5: ROV track with topic mixtures at each sample point. The line represents the path of the ROV, and each point is the location of a substrate sample image. At each point, there are circles for each of the three topics, with their sizes representing the mixture of the topics in that sample.

ples. This reflects the transitional areas which are not well described by a single label.

In collecting the ground-truth data, we observed that a correct classification is extremely context dependent and somewhat subjective. Depending on the intended application, the salience of different features varies subjectively, for example, a frame containing mainly sediment with one area of exposed pillow lava might might be of high interest to one researcher but of little interest to another. In addition, it is unclear how a ground-truth dataset should measure the amount a substrate type is represented in each frame. Although it is tempting to use percent cover as a way to quantify these observations, this approach depends on having very precise definitions of the boundaries between substrate types which are not always available.

Emerging methods for hierarchical topic models could help to resolve the differences in scale of similarity between the desired categories [3]. In addition, directly incorporating the noise reduction and vocabulary learning steps into the probabilistic model could improve results by estimating the parameters from data, rather than setting them based on manual experimentation.

As a preliminary investigation for future work, we ran our algorithm on two additional datasets: One on the nearby High-Rise vent structure featuring a completely different set of substrate types and geoforms, and a second from Barkley Canyon—a biological diverse yet geologically uniform environment composed mainly of sediment. Preliminary results appear promising, suggesting that this system could be used in other environments with minimal modifications.

8. Conclusion

We have presented a method for learning and mapping deep-sea substrate types. This method computes a mixture of types for each frame in an ROV video using a texture-based BoW representation of the images and a spatiotemporal topic model. It requires minimal training data, and includes measures to compensate for the different kinds of noise associated with deep-sea ROV video recordings. We have shown that in a mid-ocean ridge flank environment, this method recovers topics that correlate highly with ground-truth substrate categories. Our analysis shows that the described pipeline significantly outperforms conventional methods, demonstrating the utility of the combination of proposed techniques for the substrate labelling task.

9. Acknowledgement

The authors thank Steven F. Mihály, Fabio C. De Leo, Marlene A. Jeffries and Laurence A. Coogan for their suggestions and contributions to this work.

References

[1] D. M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, Apr. 2012.

[2] L. Cao and L. Fei-Fei. Spatial coherent latent topic model for concurrent object segmentation and classification. *International Conference on Computer Vision*, 2007.

[3] F. Chamroukhi, M. Bartcus, and H. Glotin. Bayesian non-parametric parsimonious gaussian mixture for clustering. In *In Proc. of Int. Conf. on Pattern Recognition (ICPR), IEEE*, Stockholm, 2014.

[4] Y. Girdhar. *Unsupervised Semantic Perception, Summarization, and Autonomous Exploration for Robots in Unstructured Environments*. PhD thesis, McGill University, 2014.

[5] Y. Girdhar and G. Dudek. Modeling curiosity in a mobile robot for long-term autonomous exploration and monitoring. *Autonomous Robots*, pages 1–12, 2015.

[6] Y. A. Girdhar, W. Cho, M. Campbell, J. Pineda, E. Clarke, and H. Singh. Anomaly detection in unstructured environments using bayesian nonparametric scene modeling. *CoRR*, abs/1509.07979, 2015.

[7] Y. A. Girdhar and G. Dudek. Gibbs sampling strategies for semantic perception of streaming video data. *CoRR*, abs/1509.03242, 2015.

[8] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[9] M. Lacharité, A. Metaxas, and P. Lawton. Using object-based image analysis to determine seafloor fine-scale features and complexity. *Limnology and Oceanography: Methods*, 13(10):553–567, 2015.

[10] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Li. Learning multi-scale block local binary patterns for face recognition. *Advances in Biometrics*, pages 828–837, 2007.

[11] MATLAB. *Version 8.6.0 (R2015b)*. The MathWorks Inc., Natick, Massachusetts, 2015.

[12] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[13] S. Paris, X. Halkias, and H. Glotin. Efficient Bag of Scenes Analysis for Image Categorization. *International Conference on Pattern Recognition Applications and Methods*, 2012.

[14] O. Pizarro, P. Rigby, M. Johnson-Roberson, S. B. Williams, and J. Colquhoun. Towards image-based marine habitat classification. *Oceans 2008*, pages 1–7, 2008.

[15] O. Pizarro, S. Williams, and J. Colquhoun. Topic-based habitat classification using visual data. In *OCEANS 2009 - EUROPE*, pages 1–8, May 2009.

[16] T. Schoening, T. Kuhn, and T. Nattkemper. Seabed classification using a bag-of-prototypes feature representation. In *Computer Vision for Analysis of Underwater Imagery (CVAUI), 2014 ICPR Workshop on*, pages 17–24, Aug 2014.

[17] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, pages 1903–1910, 2009.

[18] S. B. Williams, O. Pizarro, J. M. Webster, R. J. Beaman, I. Mahon, M. Johnson-Roberson, and T. C. L. Bridge. Autonomous underwater vehicle-assisted surveying of drowned reefs on the shelf edge of the Great Barrier Reef, Australia. *Journal of Field Robotics*, 27(5):675–697, Aug. 2010.

[19] B. Zhao, L. Fei-Fei, and E. Xing. Image segmentation with topic random field. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision - ECCV 2010*, volume 6315 of *Lecture Notes in Computer Science*, pages 785–798. Springer Berlin Heidelberg, 2010.