2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

A Guided Gaussian-Dirichlet Random Field for Scientist-in-the-Loop Inference in Underwater Robotics

Chad R. Samuelson and Joshua G. Mangelson

***Abstract*— Visual topic modeling (VTM) provides key insight into data sets based on learned semantic topic models. The Gaussian-Dirichlet Random Field (GDRF), a state-of-the-art VTM technique, models these semantic topics in continuous space as densities. However, ambiguity in learned topics is a disadvantage of such Dirichlet-based VTM algorithms. We propose the Guided Gaussian-Dirichlet Random Field (GGDRF). Our method applies Dirichlet Forest priors from natural language processing (NLP) to the vision domain as a way to embed visual scientific knowledge into the estimation process. This modification and addition to the GDRF provides a key shift from unsupervised machine learning to semi-supervised machine learning in the robotic VTM domain. We show through simulation and real-world underwater data that the proposed GGDRF outperforms the previous GDRF method both quantitatively and qualitatively by improving alignment between estimated topics and scientific interests.**

I. Introduction

The majority of the Earth's oceans and solar system remain unexplored. Scientists are interested in exploring Earth's underwater world, but these environments remain difficult, dangerous and costly to explore. Autonomous robotic systems are essential to such efforts due to the risk and difficulty of human exploration [1].

Autonomous robots bring many complexities. Limited communication between the scientist and the robot is a key issue. Techniques that enable intelligent autonomous exploration have the potential to address some of these concerns. Due to environmental complexities, robotic exploration comes at a high financial cost. Additionally, only a fraction of the data collected proves to be relevant when scientific interests are not leveraged. To improve efficiency, it is essential that we employ techniques that enable intelligent data collection. One way to accomplish this is to ensure estimated models are both accurate and relevant to human scientific interests.

We propose the GGDRF as a first step towards leveraging human knowledge for more intelligent autonomous exploration.

Various techniques in machine learning (ML) have been developed to enable the modeling of visual observations. Modern deep learning for classification can very accurately detect and classify visual topics in images; however, they require large amounts of labeled training data. Labeling these large data sets is time-consuming and challenging. Unsupervised ML methods attempt to address these challenges. Topic modeling, as developed in the NLP field, is an excellent example of these techniques [2–4]. Roboticists have made many advancements over the last few decades by applying NLP concepts to the interpretation of images [5–13]. Their methods have shown promise in semantically modeling imagery data streams as well as identifying anomalies [5], but the learned topics are not always aligned with specific human interests. A few Bayesian methods have been proposed as a way to embed knowledge to improve interpretability [14, 15], but they only track discrete environment models and are limited to causal knowledge embeddings. The GDRF algorithm [12] has been shown to surpass its predecessors in the unsupervised exploratory robotics field by combining visual LDA with the continuous modeling and interpolation power of Gaussian Processes (GPs) [16]. The GDRF algorithm lacks an ability to embed knowledge into the topic estimation process.

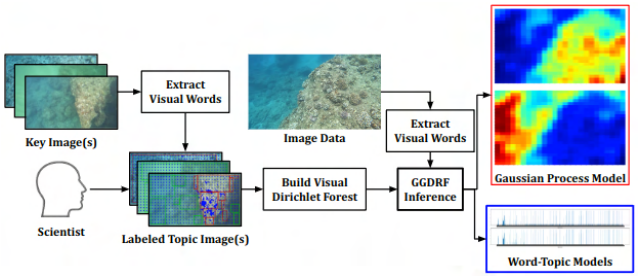


Fig. 1: GGDRF method overview: 1) extract features from key images; 2) scientist labels portions of the image that are scientifically relevant; 3) build VDF prior around labeled words; 4) run SVI on GGDRF given the VDF prior and words extracted from a data set to learn topic distributions and word-topic distributions.

The proposed GGDRF extends the GDRF algorithm to enable the embedding of scientific knowledge. Our goal is to enable an expert to influence the topic learning process by labeling a small number of images, resulting in increased alignment of learned topics with the interests of the expert. Figure 1 gives a high-level overview of the GGDRF algorithm. First, a scientist labels key aspects of a small number of images based on visual topics that they deem relevant. A novel adaptation of the Dirichlet Forest prior [17], the visual dirichlet forest (VDF), then embeds this knowledge into a word-topic probability distribution. This serves as prior information that is fed into the GGDRF algorithm and both the image data set and the prior are taken into account when estimating topics. Our main contributions are as follows:

This work was partially funded under Department of Navy awards N00014-21-1-2435 and N00014-21-1-2272 issued by the Office of Naval Research.

C. Samuelson, J. Mangelson, are at Brigham Young University.
{chadrs2, joshua_mangelson}@byu.edu.

1) A novel VDF technique that adapts the Dirichlet Forest for visual features.
2) An extension of the GDRF called the Guided-GDRF (GGDRF) that uses the VDF to allow the integration of semantic scientific knowledge into the estimation process.
3) An evaluation of the GGDRF and GDRF on real-world robotic underwater-focused imagery.

Section II of the paper provides a deeper review of recent advances in the field of VTM. In section III we present the modified Dirichlet Forest prior, the VDF. In section IV we introduce and discuss the GGDRF algorithm. Section V shows an evaluation of our algorithm with respect to the GDRF on both simulated and real-world underwater data. Our conclusions and future work is discussed in section VI.

II. Related Work

A. Topic Modeling in NLP

NLP studies the automatic processing of human language. Latent Dirichlet Allocation (LDA) [2] was a key development in the field of NLP that leverages unsupervised machine learning to find the distribution of words in each latent topic and the distribution of latent topics within each document. The group of documents in a data set is referred to as the corpus. LDA uses collapsed Gibbs Sampling [18] to estimate these distributions for a given set of words and corpus.

While LDA improves upon prior techniques [3, 4], the resulting topics may not be directly aligned with a human user's needs. A Dirichlet Forest prior incorporates domain knowledge by linking certain user-defined words together as must-links or separate as cannot-links in a method called Dirichlet Forest-LDA (DF-LDA) [17]. These must-link embeddings encourage linked words to appear within each topic with similar probability. Cannot-link embeddings encourage cannot-linked word pairs to appear with dissimilar probability in the same topic. The strength of these knowledge embeddings can be adjusted depending on the use case.

Experiments performed in [17] demonstrate the DF-LDA model resulted in word-topic distributions with higher probable words matching the user-specified domain knowledge while the previous LDA model did not match user interests. DF-LDA has since been further modified in both language and vision domains. The Guided Hierarchical Topic Model [19] merged hierarchical topic modeling methods, that adaptively adjust the number of topics, with Dirichlet Forest priors. DF-LDA has also been applied to the vision domain to connect visual image attributes with words guided by DF-LDA embeddings [20]. To the best of our knowledge, Dirichlet Forest priors have not been applied to raw imagery up to this point.

B. Gaussian Processes

A GP is a stochastic process that is used to describe the distribution of a set of random variables and their dependency on neighboring random variables in the set [16]. In GP regression, the inputs are considered to be random variables that will then be mapped via a learned GP to output random variables. GPs make the assumption that the output at a given index follows a multivariate Gaussian distribution and there exists a correlation between neighboring index locations of the output random variables in the process. GPs can model non-linear relationships between inputs and outputs as well as provide a measure of uncertainty in their predictions while enabling continuous modeling.

C. Visual Topic Modeling

In the past decade, there has been great interest in applying modern NLP topic modeling approaches to the computer vision domain. This merge is referred to as VTM. Most often in VTM, a set of image observations form the corpus, each image observation is analogous to a document and words are features that can be extracted from an image such as hue, intensity, texture [21], ORB features [22] and so on. The interpretation of topics in VTM varies depending on the application. VTM has seen applications in semantic film interpretation [23], semantic image analyses [24], as well as robotics [5–13].

Girdhar, et al. [5] first applied VTM to underwater robotics in 2014 as a way to detect anomalies during underwater exploration. In their work, known as ROST, the definition of documents corresponded to observations of a spatiotemporal neighborhood of features (i.e. the combination of multiple images (temporal) which are each split into grid cells (spatial)). This spatiotemporal neighborhood definition embedded a key 3D relationship of features in images that closely approximates human visual understanding. Subsequent modifications have resulted; however, these all divide the spatiotemporal space of a video into discrete cells [5–8].

D. Gaussian-Dirichlet Random Field

Most recently [12] proposed the GDRF. The GDRF builds upon [5], but shifts from a discrete neighborhood model to a continuous one, enabling the modeling of topic densities using GPs. This modification of ROST replaces the Dirichlet prior for modeling how a topic is distributed within an observation with a GP. When this is applied across multiple topics, we have one GP distribution per topic. The GP mean and covariances as well as the word-topic distribution are learned via SVI.

E. VTM Knowledge Embedding

Little has been done in embedding knowledge into VTM methods, especially in the field of exploratory robotics. Arora, et al. [14, 15] is one of the few methods to do so. They used Bayesian Networks (BN) to embed scientific causal knowledge when learning latent VTM topics. These latent topics were then used with a Monte Carlo Tree Search (MCTS) method to perform informative path planning (IPP) for exploring topics of interest and learning latent topics that can not be directly identified by the robot's sensors. Due to GP computational complexities derived from embedding causal knowledge into GPs, they choose the BN framework leading to a discrete BN learning problem. We are focused on expanding knowledge embedding methods into continuous

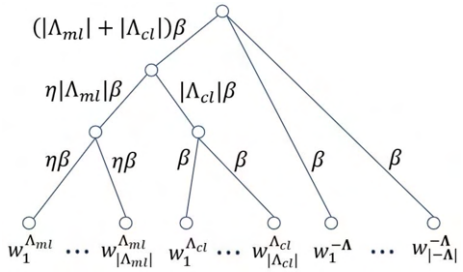


Fig. 2: The VDT structure-Leaf nodes are split into three groupings by must-link, cannot-link, and unlinked words. The probability of a given word is simply the product of edge-weights from the leaf node up to the root node.

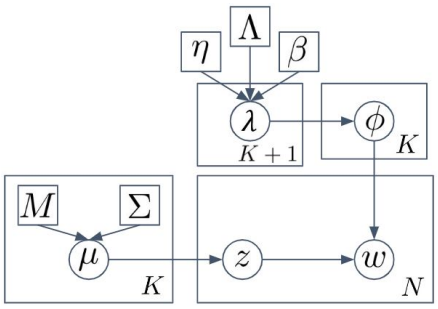


Fig. 3: GGDRF Graphical Model-Our novel model merges the approaches of the GDRF and the adapted VDF prior for use in VTM.

problems and thus use GPs because of their established probability measures and the GDRF approach for flexibility in embedding other forms of scientific knowledge, besides causal relationships. Not much other work has been presented in embedding scientific knowledge into VTM methods for robotic exploration.

III. Visual Dirichlet Forest

The original Dirichlet Forest prior restricts must-link words to not be assigned as must-links to more than one topic. In image processing, some features, such as hue or intensity, can belong to multiple different categorical topics. To handle this, we ensure that any words that are must-links in more than one topic are removed from the must- and cannot-link structure and instead fall under the unlinked or unlabeled words portion of the forest. This allows us to use the same mathematical properties of the original Dirichlet Forest. We also further modify the trees within the VDF structure, visual dirichlet trees (VDT), to have only three groupings of words: must-link, cannot-link, and unlink. The must-link word set will be unique for each topic and the cannot-link word set simply consists of all must-link words belonging to all other topics. The unlink word set remains the same for each tree. For example, for 3 topics given these relationships $[[w_1, w_4], [w_5, w_2], [w_8]]$ the first VDT would have $[w_1, w_4]$ as must-links, $[w_5, w_2, w_8]$ as cannot-links, and all other words as unlinks.

Figure 2 shows the VDT model within the VDF structure. Instead of sampling Dirichlet Trees from a Dirichlet Forest structure, our VDF is set up to relate each VDT to a particular topic. This also provides the added benefit that the scientist who has labeled a topic of interest, can look at that associated learned VTM topic after the GGDRF has trained to find what they're interested in. The leaf nodes for each tree contain every word in the vocabulary set V. If there are more topics than trees in the VDF, each additional topic is assigned a standard Dirichlet distribution (i.e. the standard model as presented in ROST [5] and GDRF [12]). The notation we use in the VDF for the three groupings of leaf nodes is $w_i^{\Lambda_{ml}}$ where $i \in [1 \cdots |\Lambda_{ml}|]$ for must-links, $w_i^{\Lambda_{cl}}$ where $i \in [1 \cdots |\Lambda_{cl}|]$ for cannot-links, and $w_i^{-\mathbf{\Lambda}}$ where $i \in [1 \cdots | -\mathbf{\Lambda} |]$ for unlinks as shown in Figure 2. The probability of a given word, $\phi^{(w)}$, where $w \in V$ in this tree is simply the product of the edge weights, τ, from the leaf node word up to the root of the tree as in the original DF-LDA paper [17]. The word-topic probability generation given the set of edge-weights, $\boldsymbol{\tau}$, is described in section 3.1 of [19].

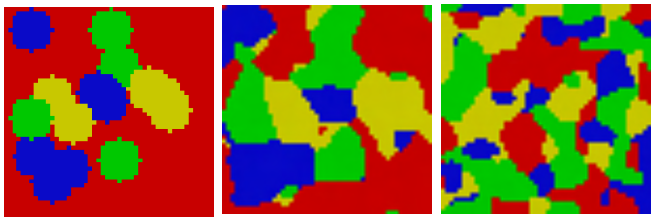
IV. Guided Gaussian-Dirichlet Random Field

We begin by defining the ordered-index set as $X = \{x_1, \cdots, x_N\}$, where $x_i \in \mathbb{R}^d$ and d is the number of dimensions (i.e. $d = 1$ for temporal data). The set of visual words is defined as $W = \{\boldsymbol{w_1}, \cdots, \boldsymbol{w_N}\}$, where $\boldsymbol{w_i} \in \mathbb{Z}^V$ is associated with x_i and V is the size of the visual vocabulary. The element-array $\boldsymbol{w_i}$ contains integer counts of the quantity of each word in the vocabulary V that is present at index x_i. Given K different topics, we seek to learn the probabilities of each latent topic distribution across the index set, X, as well as the word probability distributions within each topic, ϕ_z.

The GGDRF models this problem as a DF-LDA-based problem within the domain of VTM. The distribution of topics across the index set is modeled as a GP for each topic. This allows for 1) continuous index sets to be modeled; 2) interpolation capabilities within the index set; 3) associated uncertainty measures; and 4) non-linear behavior modeling. The word-topic distributions are modeled as VDT distributions for each topic. The generative model for the GGDRF is

$$\mu_{i,k} \sim \mathcal{GP}_k(M(x_i), \Sigma(x_i)), \tag{1}$$
$$z_{i,k} \sim Softmax(\mu_{i,1}, \cdots, \mu_{i,K}), \tag{2}$$
$$\boldsymbol{\lambda} = VDF(\mathbf{\Lambda}, \beta, \eta), \tag{3}$$
$$\phi_{z_j} \sim VDT(\lambda_j), \tag{4}$$
$$w_i \sim Categorical(\phi_{z_{i,k}}), \tag{5}$$

where, $i \in [1 \cdots N]$, $k \in [1 \cdots K]$, $j \in [0 \cdots K]$, and $\boldsymbol{\lambda} = [\lambda_0, \cdots, \lambda_K]$ contains the domain expert's knowledge embeddings split into VDTs. It is also important to note that instead of sampling a Dirichlet Tree from a Dirichlet Forest, our VDF serves as a function that outputs a set of K VDTs and one standard Dirichlet Distribution, λ_0. The word-topic distributions are then sampled from these VDTs. The graphical model for the GGDRF is shown in Figure 3.



(a) True Topic (b) GGDRF MLT (c) GDRF MLT

Fig. 4: Maximum Likelihood Topic (MLT) comparison of simulated spatial results.

A. Stochastic Variational Inference over GGDRF

Instead of classical Gibbs Sampling methods for learning the topic-document and word-topic distributions, we follow the GDRF's approach of using SVI. In SVI we seek to minimize the KL-Divergence between the variational distribution, denoted by $q_\theta(z)$, and the true distribution given the observed data, denoted by $p_\gamma(z|w)$. The minimization of the KL-Divergence reduces to maximizing the Evidence Lower Bound (ELBO) terms of the KL-Divergence equation. This optimization problem is iteratively solved by adjusting the variational parameters θ. The data, w, are the visual words per observation. The θ parameters are the GP M and Σ, and word-topic matrix. The γ parameters are the GP priors, zero-mean and kernel covariance, and VDF priors: ($\mathbf{\Lambda}$, β, η). The ELBO optimization problem is then

$$\max_{\theta} \; ELBO = logp_\gamma(z) + logp_\gamma(w|z) - logq_\theta(z), \quad (6)$$

where $z = (\phi, \mu)$. The probabilistic PyTorch Python library, Pyro-PPL [25, 26], streamlines this SVI optimization process.

V. Results

The effectiveness of the GGDRF method compared to the prior GDRF method is demonstrated spatially and spatiotemporally in the following subsections. The last subsection numerically demonstrates the improvement of this method using the Wasserstein distance metric [27, 28], commonly known as the Earth Mover's Distance (EMD).

A. Simulated Spatial Experiment

Our simulated datasets were generated given a number of topics, K, and vocabulary size, V. The vocabulary words were evenly distributed among the K topics. The topics were randomly distributed across the grid with the first topic being the background. For every pixel of the grid, one word was sampled according to Equation 5 where $z_{i,k}$ represents the topic assigned to that grid coordinate and $\phi_{z_{i,k}}$ is the word probability distribution for that topic according to Equation 7 in [5].

This simulation consisted of $K = 4$ topics, $V = 50$ vocabulary words, $\beta = 0.1$, $\eta = 100$, with an image size of 50×50 pixels. The VDF was built on 10 labeled words evenly distributed across the topics. A single observation is one pixel. Figure 4a is the true topic distribution across the simulated image where the color represents the topic at each pixel location. Figure 4b-Figure 4c show the maximum likelihood topic estimates from the learned GP models trained for 10000 epochs. The GGDRF is capable of tracking the respective topics with some degree of topic mixing observed at the borders. The GDRF topics struggle to correctly correlate learned topics with the true ones.

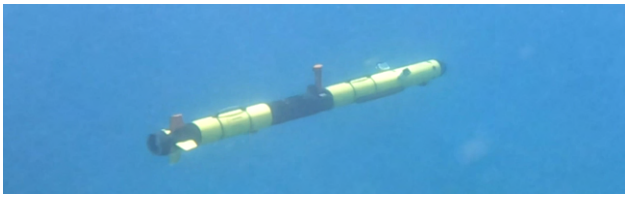


Fig. 5: BYU FRoSt Lab Iver3 Collecting Data off the Coast of Oahu, HI.

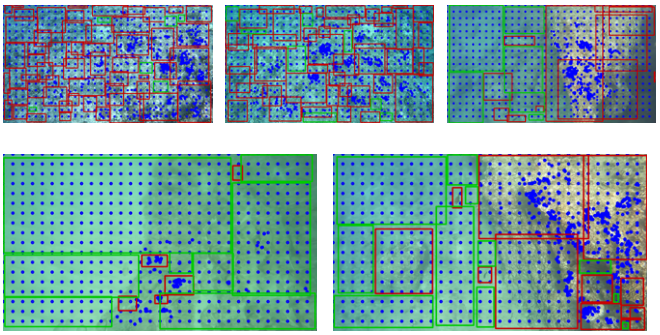


Fig. 6: The labeled images used to build the VDF for the following tests where topic 0 VDT is the red labels and topic 1 VDT the green labels.

B. Real World

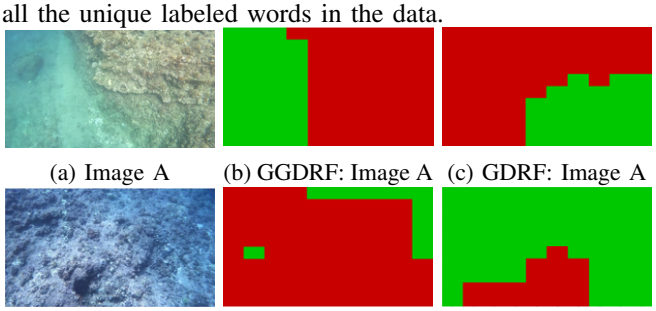
We gathered real-world coral reef data collected off the shore of Oahu using an Iver3 AUV (see Figure 5). We modified the standard Iver3 with a custom stereo-camera system for visual underwater robotic exploration. While this method can extend to multiple topics, in these tests we sought to identify coral reefs as one topic and non-coral reefs as the other topic. In all of our experiments, we extracted hue, intensity, texture, and ORB words, and set β to 0.01, η to $1,000,000$ and used a zero-mean and Matérn 3/2 Kernel for our GPs. The Matérn 3/2 Kernel parameters, lengthscale and variance, were initialized to 1.0. We trained both the spatial and spatiotemporal models for $10,000$ epochs.

1) Knowledge Embeddings

The five images in Figure 6 were labeled by a marine biologist based on areas of interest which mostly coincided with images containing coral (i.e. the red boxes assigned to topic 0) and non-coral areas (i.e. the green boxes assigned to topic 1). This was the only input knowledge given to build the VDF. As with the following tests, these labeled images contained all hue, intensity, texture, and ORB word types. Table I shows less than 40 percent of unique words were labeled per word type. This allows the GGDRF to learn beyond the scientist's domain knowledge input as well as demonstrating that this algorithm truly is a semi-supervised machine learning method.

Word Type	Vocab Size	Data Set	Labeled Set
Hue	180	106	34
Intensity	256	186	4
Texture	1000	476	73
ORB	5000	4267	1635

TABLE I: Labeled Word Type Analysis. Data Set encompasses all unique words in the data. Labeled Set encompasses all the unique labeled words in the data.



(a) Image A (b) GGDRF: Image A (c) GDRF: Image A

(d) Image B (e) GGDRF: Image B (f) GDRF: Image B

Fig. 7: Spatial results of both the GGDRF and GDRF methods making GP predictions. Topic 0 is red and topic 1 is green.

2) Spatial Tests

In the first test, we extracted hue, intensity, and texture words equally distributed across the image and a max of $1,000$ ORB features. Each image was then split into a 10×10 spatial grid, grouping all words within each grid cell for inference. Note we split images into grids to enable EMD computations later on.

After training both methods on the two test images, Figure 7a and Figure 7d, we ran GP predictions in a similar 10×10 grid with the learned GDRF and GGDRF topic GPs. Each grid cell was colored according to whichever topic had a higher mean prediction at that location.

For Image A, Figure 7a, the GDRF appears to have identified a portion of the coral reef by topic 1, but misses the top half of it (see Figure 7c). Figure 7b shows the GGDRF accurately identifying the coral reef in the image by topic 0.

For Figure 7d, it is unclear which topic identifies the coral better for the GDRF (see Figure 7f). While the GGDRF accurately identifies the image as predominantly being coral reef with the blurred upper edges being identified as non-coral regions (see Figure 7e).

These results show the ability of the GGDRF to more accurately predict topic likelihoods to match scientifically relevant regions of an image as compared to the GDRF method.

3) Spatiotemporal Test

The spatiotemporal test consisted of a similar approach to word extraction as the spatial tests except that each image was split into a 5×5 grid and this was done across a set of 52 consecutive images. We used a similar GP prediction method on the data set after training the GDRF and GGDRF methods with a 5×5 grid for the spatial dimension predictions.

Temporally (i.e. the green/red plots in rows 2 and 3 of Figure 8), the data begins with little to no coral reef visible in the images (see image 0). The GDRF method predicts a fairly even split of the two topics, while the GGDRF correctly captures the presence of mostly one topic initially (i.e. topic 1 the non-coral topic). Around images 3 and 4 the robot begins going over the coral reef. As the robot moves over a coral reef (about images 4-32), we see the GGDRF correctly tracking this increase of the red topic (i.e. coral topic), but initially, the GDRF doesn't track this dramatic change until later around image 12. Later, around image 32, the coral reef ends and the robot is again over rock and sand. The GGDRF correctly tracks when the coral reef leaves the frame, while the GDRF catches this change but is a few frames off (i.e. it seems to think there's a change starting around image 30). The GDRF green topic (topic 1) is closest to tracking the coral reefs, but not as accurately as the GGDRF as explained previously. The red topic is the coral reef topic for the GGDRF model which tracks throughout the data set. Coral then enters the camera frame again in the last two frames and both methods seem to catch this change.

Test Type	GDRF EMD Avg	GGDRF EMD Avg
Spatial	2.01	1.63
Spatiotemporal	2.61	1.76

TABLE II: Average EMD numerical evaluations.

At a few query images, shown in the bottom rows of Figure 8 we do Maximum-Likelihood spatial topic predictions which closely match the scientific knowledge embeddings among the GGDRF compared to the GDRF method. In particular, we see images 0 and 26 when there is no coral present and then only coral present that the GGDRF ML predictions correctly match this understanding while the GDRF appears to mostly look the same even though the topics of interest are completely reversed in these two images.

Overall, these spatiotemporal results show an ability to consistently track more clearly where the scientific topic of interest is present in the data as compared to the GDRF based solely on visual inspection. The following section quantitatively supports this conclusion.

C. Earth Mover's Distance Numerical Evaluation

The Earth Mover's Distance (EMD) metric, also known as the Wasserstein Distance [27, 28], measures the dissimilarity between two histograms. This metric is formulated as an optimization problem where we wish to minimize the work required to convert two histograms to be equal to one another.

Given two histograms, P and Q, with signatures $P = \{(x_1, p_1), \cdots, (x_n, p_n)\}$ and $Q = \{(x_1, q_1), \cdots, (x_n, q_n)\}$, where x_i is the bin location and p_i or q_i is the mass at that bin location, the EMD seeks to find the optimal flow, F, to convert P and Q to be equal to each other. In terms of this paper's EMD use case, we assume the histograms are of equal length, n. This optimization problem is then formulated as

$$\min_F F = \sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij} d_{ij}, \tag{7}$$

where the distance between bins i and j is denoted as $d_{ij} = d(x_i, x_j)$ and f_{ij} is the flow of a portion of histogram

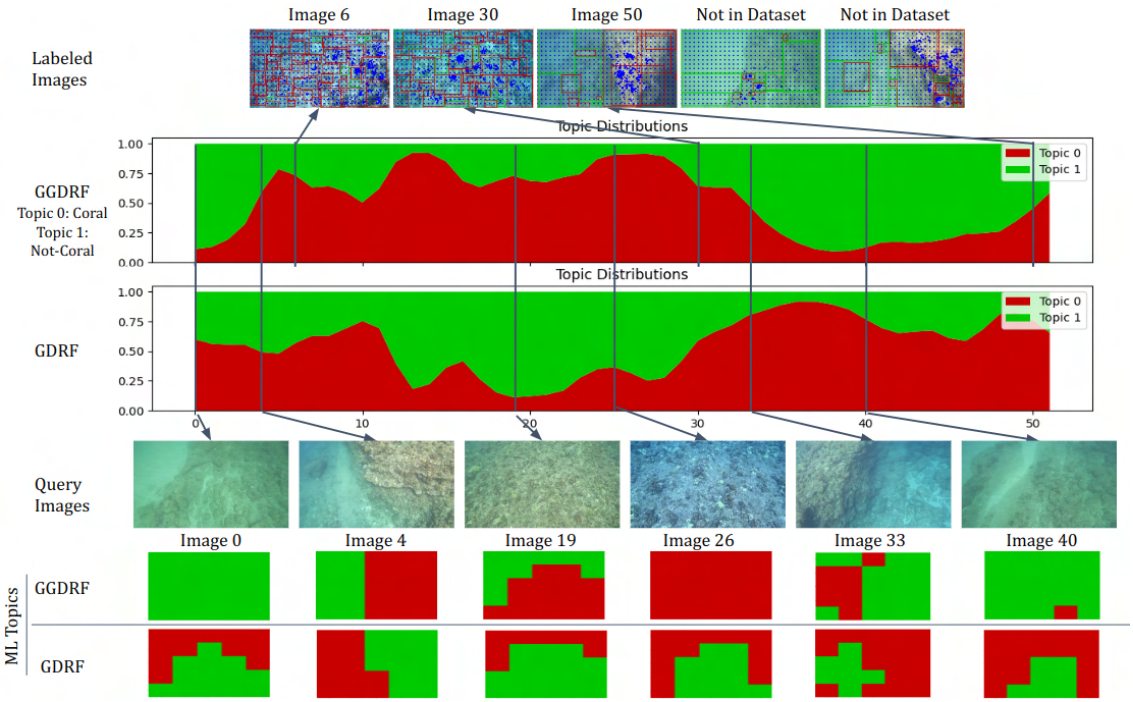


Fig. 8: Spatiotemporal results from image data collected in a single-pass above a coral reef off the shore of Oahu, HI.

mass p_i to q_j across distance d_{ij}. The flow has the following constraints

$$f_{ij} \geq 0, \qquad 1 \leq i, j \leq n, \tag{8}$$

$$\sum_{i=1}^{n} f_{ij} \leq p_i, \qquad 1 \leq i \leq n, \tag{9}$$

$$\sum_{j=1}^{n} f_{ij} \leq q_i, \qquad 1 \leq j \leq n, \text{and} \tag{10}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij} = min\left(\sum_{i=1}^{n} p_i, \sum_{j=1}^{n} q_j\right). \tag{11}$$

Once the optimal flow, f_{ij}^*, is found, the EMD is defined as

$$EMD(P,Q) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}^* d_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}^*} \tag{12}$$

We used the OpenCV implementation of EMD for computing this metric with an L2-Distance metric for d_{ij}. These histograms consisted of ones and zeros corresponding to topic one and zero based on the ML topic plots. We then normalized these histograms as well as the histograms labeled by the marine biologist. The predicted normalized ML topic histograms were then compared with the normalized biologist's topic histogram labels using this EMD metric. Table II shows the averaged numerical EMD results across 20 spatial images and the spatiotemporal data set shown in Figure 8.

We see that on average the GGDRF method more accurately followed the domain expert interests than the GDRF. There were a few spatial images where the GDRF outperformed the GGDRF. This occurred because of many rocks that had tiny coral on them, but not enough to be labeled by the biologist as predominantly coral regions. The GGDRF also struggled with an image that contained many rocks with patches of bright sand between them, making it look like bleached coral. In these, and a few other cases, the GGDRF predicted that coral was present. It is worth noting that the image with bright patches of sand was not incorrectly identified by the GGDRF when used in the spatiotemporal data set. This shows that the GGDRF topic model learns across a video stream to generalize and more accurately identify topics.

VI. Conclusion and Future Work

In this paper, we propose a novel new VTM method for exploratory robotics, known as the GGDRF. This model merges the state-of-the-art semantic modeling method, the GDRF, with a visually adapted knowledge embedding approach, known as the VDF. This new semi-supervised semantic mapping approach applied to spatial and spatiotemporal data outperforms the GDRF both qualitatively and quantitatively. The GGDRF algorithm has shown to make predictions outside the sparse word labels given to it.

In future work, we plan to improve the GGDRF algorithm by utilizing DeepSeeColor [29] to reconstruct accurate hue and intensity values in our imagery data. Additionaly, we plan to implement a combination of Exactly Sparse [30] and Incremental GPs for online spatiotemporal robotic exploration.

Acknowledgments: We would like to thank Brad Smith and Daniel Demartini from BYU-Hawaii for their collaboration during field tests off of Oahu, HI. We would also like to thank Rocky Seeley and Richard Gill (BYU Biology) for helpful discussion and support in data labeling.

REFERENCES

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

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 1999, p. 50–57.

[4] K. Nigam, A. Mccallum, S. Thrun, and T. Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM," *Machine Learning*, vol. 39, pp. 103–134, May 2000.

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

[6] J. W. Kaeli, J. J. Leonard, and H. Singh, "Visual summaries for low-bandwidth semantic mapping with autonomous underwater vehicles," in *Proceedings of the 2014 IEEE/OES Autonomous Underwater Vehicles*, Oct 2014.

[7] Y. Girdhar, W. Cho, M. Campbell, J. Pineda, E. Clarke, and H. Singh, "Anomaly detection in unstructured environments using Bayesian nonparametric scene modeling," in *Proceedings of the IEEE International Conference on Robotics and Automation*, Jun 2016, pp. 2651–2656.

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

[9] K. Doherty, G. Flaspohler, N. Roy, and Y. Girdhar, "Approximate Distributed Spatiotemporal Topic Models for Multi-Robot Terrain Characterization," in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, 2018, pp. 3730–3737.

[10] 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 *Proceedings of the IEEE International Conference on Robotics and Automation*, May 2019, pp. 7940–7946.

[11] S. Jamieson, J. P. How, and Y. Girdhar, "Active Reward Learning for Co-Robotic Vision Based Exploration in Bandwidth Limited Environments," in *Proceedings of the IEEE International Conference on Robotics and Automation*, Mar 2020, pp. 1806–1812.

[12] J. E. Soucie, H. M. Sosik, and Y. Girdhar, "Gaussian-Dirichlet Random Fields for Inference over High Dimensional Categorical Observations," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2020, pp. 2924–2931.

[13] S. Jamieson, K. Fathian, K. Khosoussi, J. P. How, and Y. Girdhar, "Multi-Robot Distributed Semantic Mapping in Unfamiliar Environments Through Online Matching of Learned Representations," in *Proceedings of the IEEE International Conference on Robotics and Automation*, May 2021, pp. 8587–8593.

[14] A. Arora, R. Fitch, and S. Sukkarieh, "An approach to autonomous science by modeling geological knowledge in a Bayesian framework," in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, Sep 2017, pp. 3803–3810.

[15] A. Arora, P. M. Furlong, R. Fitch, S. Sukkarieh, and T. Fong, "Multi-modal active perception for information gathering in science missions," *Autonomous Robots*, vol. 43, no. 7, pp. 1827–1853, 2019.

[16] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[17] D. Andrzejewski, X. Zhu, and M. Craven, "Incorporating domain knowledge into topic modeling via Dirichlet Forest priors," in *Proceedings of the 26th International Conference On Machine Learning (ICML)*, 2009, pp. 25–32.

[18] Griffiths, T. L, and M. Steyvers, "Finding scientific topics," *National Academy of Sciences of the United States of America*, vol. vol. 101, no. Suppl 1, pp. 5228–35, 2004.

[19] S. J. Shin and I. C. Moon, "Guided HTM: Hierarchical Topic Model with Dirichlet Forest Priors," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 2, pp. 330–343, 2017.

[20] X. Chen, X. Hu, Z. Zhou, Y. An, T. He, and E. K. Park, "Modeling semantic relations between visual attributes and object categories via Dirichlet Forest prior," in *ACM International Conference Proceeding Series*, 2012.

[21] R. Roslan and N. Jamil, "Texture feature extraction using 2-d gabor filters," in *Proceedings of the 2012 International Symposium on Computer Applications and Industrial Electronics (ISCAIE)*, 2012, pp. 173–178.

[22] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proceedings of the International Conference on Computer Vision*, 2011, pp. 2564–2571.

[23] M. Rabinovich and Y. Girdhar, "Gaining Insight Into Films via Topic Modeling Visualization," *Parsons Journal for Information Mapping*, vol. 7, Jan 2015.

[24] A. Khodaskar and S. Ladhake, "Semantic Image Analysis for Intelligent Image Retrieval," *Procedia Computer Science*, vol. 48, pp. 192–197, 2015, international Conference on Computer, Communication and Convergence (ICCC 2015).

[25] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, and N. D. Goodman, "Pyro: Deep Universal Probabilistic Programming," *Journal of Machine Learning Research*, 2018.

[26] D. Phan, N. Pradhan, and M. Jankowiak, "Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro," *arXiv preprint arXiv:1912.11554*, 2019.

[27] C. L. Mallows, "A note on asymptotic joint normality," in *Annals of Mathematical Statistics*, 1972, pp. 508–515.

[28] P. Bickel, "The earth mover's distance is the mallows distance: Some insights from statistics," in *Proceedings of ICCV*, 2001, pp. 251–256.

[29] S. Jamieson, J. P. How, and Y. Girdhar, "DeepSeeColor: Realtime Adaptive Color Correction for Autonomous Underwater Vehicles via Deep Learning Methods," in *Proceedings of CVPR*, 2023.

[30] S. Anderson and T. D. Barfoot, "Full STEAM ahead: Exactly sparse gaussian process regression for batch continuous-time trajectory estimation on SE(3)," in *Proceeding of the IEEE International Conference on Intelligent Robots and Systems*, 2015, pp. 157–164.