Automatic Fault Diagnosis for Autonomous Underwater Vehicles using Online Topic Models

Ben-Yair Raanan[1], James G. Bellingham[2], Yanwu Zhang[1], Mathieu Kemp[1], Brian Kieft[1], Hanumant Singh[3], Yogesh Girdhar[2]

[1]Monterey Bay Aquarium Research Institute, Moss Landing, CA 95039
{byraanan|yzhang|mkemp|bkieft}@mbari.org

[2]Woods Hole Oceanographic Institution, Woods Hole, MA 02543
{jbellingham|ygirdhar}@whoi.edu

[2]Northeastern University, Boston, MA 02115
ha.singh@neu.edu

***Abstract*—As the capabilities of autonomous underwater vehicles (AUVs) improve, the missions become longer, riskier, and more complex. For AUVs to succeed in complex missions, they must be reliable in the face of subsystem failure and environmental challenges. In practice, fault detection activities carried out by most AUVs employ a rule-based emergency abort system that is triggered by specific events. AUVs equipped with the ability to diagnose faults and reason about mitigation actions in real time could improve their survivability and increase the value of individual deployments by replanning their mission in response to failures. In this paper, we focus on AUV autonomy as it pertains to self-perception and health monitoring and argue that automatic classification of state-sensor data represents an important enabling capability. We apply an online Bayesian nonparametric topic modeling technique to state-sensor data in order to automatically characterize the performance patterns of an AUV, then demonstrate how in combination with operator-supplied semantic labels these patterns can be used for fault detection and diagnosis by means of nearest-neighbor classifier. The method is applied in post-processing to diagnose faults that led to the temporary loss of the Monterey Bay Aquarium Research Institute's Tethys long-range AUV in two separate deployments. Our results show that the method is able to accurately identify and characterize patterns that correspond to various states of the AUV, and classify faults with high probability of detection and no false detects.**

Keywords—Autonomous underwater vehicle (AUV); Autonomy; Fault detection and diagnosis; Topic modeling

I. INTRODUCTION

Autonomous underwater vehicles (AUVs) have become essential tools in almost every domain in the ocean. As the capabilities of AUVs improve, the missions become more complex and require vehicles with longer endurance and higher reliability. A significant limitation of current generation AUVs is their inability to cope with unforeseen events, such as failure of hardware/software components or unexpected interactions with the surrounding environment. These limitations are generally addressed by enhancements in autonomy [1]. AUVs equipped with the ability to diagnose faults and reason about mitigation actions could improve their survivability and increase the value of individual deployments by replanning their mission in response to failures while deliberating on how to best satisfy the goals given to them by human operators.

In practice, system-level fault detection activities carried out by most AUVs employ a rule-based emergency abort system that is triggered by specific events, such as critical subsystems becoming unresponsive, or the vehicle exceeding its maximum depth limit. This deterministic approach is limited to previously encountered failures and potential contingencies predicted by developers—if a new fault that endangers the vehicle is observed during operations, additional conditions will often be added. This results in a fault protection system that is complex and difficult to maintain and that offers only limited detection capabilities. We argue that many of these limitations can be alleviated by automatic classification of state-sensor data: (1) conditions and thresholds that are based on heuristics are replaced by general characteristics of classes that are inferred from data, and (2), faults are identified automatically as distinct classes.

In this paper we extend an unsupervised machine-learning framework called topic modeling to the problem of fault diagnosis in underwater robotic systems. The topic model used in this work is a Bayesian nonparametric[1] (BNP) variant of Latent Dirichlet allocation (LDA) [2]–[4]. Though the model was originally developed for semantic analysis of text documents, we explore its application to AUV sensor data in order to automatically characterize patterns that relate to vertical plane performance of the vehicle, including faults, directly from training datasets gathered in previous AUV operations. The principal features of the method, besides online execution, are that it accepts data from multiple domains, it does not require any prior annotations or labeling of the dataset, and it automatically infers the number of classes present in the data.

The paper is organized as follows: In section II we introduce LDA and its adaptation for modeling state-sensor data, and present our approach for monitoring the health of an AUV based on the topic-model's outputs. In section III we apply the method to state-sensor data collected by the Monterey Bay Aquarium Research Institute's *Tethys*-class long-range AUV (LRAUV) and demonstrate its ability to classify distinct performance patterns and diagnose faulty states.

[1]Here "Nonparametric" implies that the number of clusters is open-ended.

II. Probabilistic Topic Models for Fault Detection and Diagnosis in AUVs

LDA [2] is a generative probabilistic topic model that is used for discovering patterns in an unstructured collections of discrete data such as text corpora. LDA can be thought of as a mixed-membership model of grouped data, where rather than associating each group of observations (document) with one component (topic), each group is associated with multiple components in different proportions. The model does not require any prior annotations or labeling of the dataset—the topics emerge from the natural structure of the data.

The basic assumption made in LDA is that each document is generated from a random mixture of latent topics; each topic is a distribution over the collection's vocabulary. Given a collection of D documents composed from a vocabulary V, the LDA generative process [5] results in the following joint probability distribution:

$$\mathrm{P}(\boldsymbol{w}, \boldsymbol{z}, \theta, \phi|\alpha, \beta) = \mathrm{P}(\phi|\beta)\mathrm{P}(\theta|\alpha)\mathrm{P}(\boldsymbol{z}|\theta)\mathrm{P}(\boldsymbol{w}|\phi_z) \quad (1)$$

where α and β are the model hyperparameters, each word w is a discrete element from a fixed vocabulary indexed by $\{1, \dots, V\}$, each z represents the topic responsible for generating the word instance w. Each θ_d is a document-specific distribution over topics (can be seen as a low-dimensional representation of the dth document), and ϕ_z specifies the distribution of the zth topic over the vocabulary words. The variables $\boldsymbol{z}$, θ and ϕ are unknown (latent). To learn them, LDA reverses the generative process by expressing the conditional posterior distribution of the latent variables given the observed data:

$$\mathrm{P}(\boldsymbol{z}, \theta, \phi|\boldsymbol{w}, \alpha, \beta) = \frac{\mathrm{P}(\theta, \phi, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{\mathrm{P}(\boldsymbol{w}|\alpha, \beta)} \quad (2)$$

Approximate inference techniques such as variational inference [2] or collapsed Gibbs sampling [6] are used to resolve the posterior.

A. Semantic Modeling of State-sensor Data

Topic modeling of state-sensor data requires that the general idea of a textual word be replaced by discrete features we call *state-words*. To generate a vocabulary of state-words, we discretize each of the N signals, $\boldsymbol{S} = \{s_n\}_{n=1}^{N}$, used to describe the AUV's state into m_n non-overlapping bins[2], and concatenate them into a vocabulary of size $V = \sum_{n=1}^{N} m_n$. To extract state-words from a given signal s_n, we map each element of s_n to its closest corresponding state-word in the vocabulary. When no measurement is available for a given sensor (missing data), no word is generated. This process can be viewed as a transformation of a time-series made of heterogeneous data (e.g., numeric, Boolean or text), to a common domain space.

B. Bayesian Nonparametric Topic Modeling for Robots

Modeling data captured by a mobile robot faces additional challenges compared to semantic modeling of a fixed collection of text documents that are mutually independent. For this reason, the model we use in this work is the BNP Realtime Online Spatiotemporal Topic-model (BNP-ROST) proposed by Girdhar *et al*. in [3] and in [4]. BNP-ROST is an online version of LDA that was previously used to compute topic models of video data captured by a mobile robot in real time [4]. It accounts for continuity in the data by generalizing the idea of a document to a spatiotemporal cell within a stream of images, and computes the topic labels for a word in a cell in the context of its neighboring cells.

We adapt BNP-ROST to our application by replacing the video/image stream with a stream of data produced by state-sensors, and generalize the idea of a document to a temporal cell: Given a sequence of observations of the AUV's state we extract state-words $\boldsymbol{w}$, each with a corresponding temporal coordinate t. Similar to [3], we model the likelihood of the observed data in terms of the latent topic label variables $\boldsymbol{z}$:

$$\mathrm{P}(w|t) = \sum_{k \in K_{active}} \mathrm{P}(w|z=k)\mathrm{P}(z=k|t) \quad (3)$$

Here the distribution over vocabulary words $\phi_k \equiv \mathrm{P}(w|z=k)$ models the appearance of the topic label k, as it is shared across all temporal coordinates. The second part of the equation $\theta_t \equiv \mathrm{P}(z=k|t)$ models the distribution of the topic labels within the temporal neighborhood of coordinate t.

We make no a-priori assumptions about the number of latent topics and instead assume that there is an infinite number of them, but only a finite number is needed to explain the observed data. We use the Chinese Restaurant Process (CRP) [7] to learn the active topic labels K_{active} directly from the data and specify a CRP prior γ over the infinite groupings to control the growth of the number of labels so as to favor the lowest number that can adequately explain the data [8]. A label k is active if there is at least one observation assigned to it.

C. Semantic Labeling of Topics and Health Monitoring

Topics derived from a sequence of observations of the AUV's state represent the latent processes that are responsible for *generating* those states. These topics should correspond to the control policies or behaviors that are executed onboard, and capture the dynamic relationship between the actuators and the AUV's performance. In this respect, the topic modeling framework can be used to generate a model of the AUV's performance directly from training data. To do this, we apply the BNP-ROST algorithm to a collection of training datasets to learn the performance patterns that correspond to nominal states of the AUV, as well as to specific faults, and use the computed topics as a low-dimensional representation of these states.

To use the trained topic-model for classification, we ascribe semantic meaning to the learned topics. To do this, we evaluate the level of correspondence between the topics and a given class (i.e., a control policy or a fault) by computing the marginal probability distribution that defines the topic label proportions for that class:

[2] In this work we use equal-width-binning, however, any binning approach is valid.

$$\mathrm{P}(z=k|class) = \sum_{t \in T_{class}} \frac{\mathrm{P}(z=k|t)}{|T_{class}|} \quad (4)$$

where T_{class} is the index of all time steps belonging to that class and $\mathrm{P}(z=k|t)$ is the topic label distribution of each time step t. We then use Bayes' rule to reverse $\mathrm{P}(z=k|class)$, and compute the conditional probability

$$\mathrm{P}(class|z=k) = \frac{\mathrm{P}(z=k|class)\mathrm{P}(class)}{\mathrm{P}(z=k)} \quad (5)$$

which defines the probability of the class given the topic label. We define $\mathrm{P}(class)$ to be $|T_{class}|/|T|$, where T is the total number of time steps, and calculate $\mathrm{P}(z=k)$ using Eq. 4 and substituting T_{class} with T [3].

Given a trained topic-model, we monitor the health of the system by measuring the similarity between the learned topic distributions ϕ, and the distribution of state-words extracted from new incoming observations over the defined vocabulary V. If a distribution of state-words from a given temporal neighborhood is most similar to a topic ϕ_k that corresponds to a faulty state, then a fault is identified. We measure similarity between the two distributions using the symmetric Kullback-Leibler (KL) divergence [9]. The most relevant topic is the one that minimizes KL (i.e., the nearest-neighbor).

III. Evaluation

To evaluate the effectiveness of the method, we conducted an experiment using two datasets collected by the Monterey Bay Aquarium Research Institute's *Tethys*-class LRAUV (Fig. 1) [10], one in 2013 and the other in 2015. These datasets were chosen because they include examples of nominal performance of the LRAUV in various states as well as catastrophic faults that caused the vehicle to bottom. In either case the fault occurred in the internal mass-shifting system, which allows the LRAUV to actuate its battery pack (~1/3 of the vehicle's total weight) [11].

The evaluation was done in two steps: First, we used the 2013 data as a training set for the topic-model, then, we used the 2015 data as a test set to evaluate the classification performance of the trained topic-model on unseen data. Table 1 lists the state-sensor signals and data-products that were used as inputs to the model.

A. Training Set

In the first part of the evaluation we post-processed state-sensor data and onboard data-products that were collected by the LRAUV during a scientific field campaign in Monterey Bay, California, between September 09-14, 2013. During the initial part of the deployment the vehicle correctly executed a series of vertical profiles using 4 control policies (Fig. 2a): Float on surface (purple), Pitch (yo-yo trajectory; blue), Surface (ascend to surface; orange), and Depth (hover at depth; green). At 22:22 UTC the vehicle descended on a yo-yo dive, however, a rupture of the mass-shifter set-screw caused the battery-mass to shift all the way forward. As a result, the AUV (now extremely nose heavy) was unable to correct its downward attitude and collided with the bottom. At 23:15 UTC the vehicle's software [12] identified the problem as a "failure to ascend" fault and triggered the AUV's safety behaviors. However, these actions failed to bring the vehicle to the surface, and so the LRAUV remained on the bottom for 27 hours and was eventually located on the beach near Rio Del Mar, California, 8 km away from its last reported position.

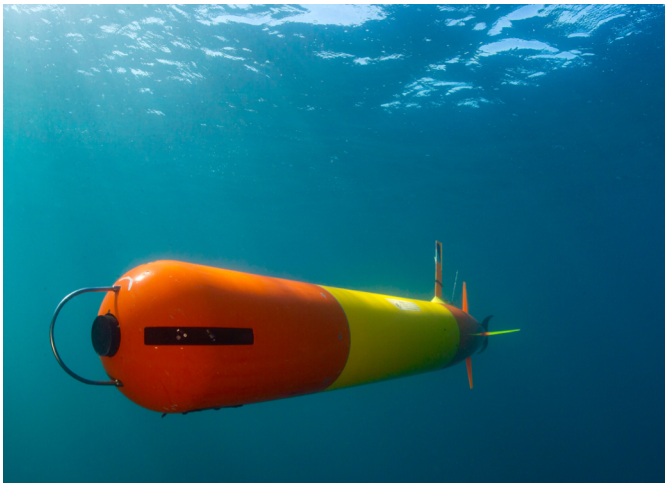

Fig. 1. The Tethys LRAUV is 2.3 m long and 0.3 m in diameter. The vehicle is controlled by a propeller, elevator and rudder control surfaces, a variable buoyancy system (VBS), and an actuated mass-shifter. Photo credit: Kip Evans.

To process the 2013 dataset using the topic-modeling framework, we extracted state-words from the dataset, which

TABLE 1
AUV State-Sensor Signals and Data-Products

Type	Signal	Range*	Comment
Numerical	Depth rate [m/s]	(-2, 2)	
	Surge velocity [m/s]	(-3, 3)	
	Heave velocity [m/s]	(-1, 1)	
	Roll angle [deg]	(-90, 90)	
	Pitch angle [deg]	(-90, 90)	
	Roll rate [deg/s]	(-2, 2)	
	Pitch rate [deg/s]	(-2, 2)	
	Stern plane angle [deg]	(-15, 15)	
	Rudder plane angle [deg]	(-15, 15)	
	Thruster power [watt]	(0, 35)	
	Δ Mass position [m]	(-25, 25)	From default pos.
	Δ Buoyancy position [ml]	(-400, 400)	From neutral pos.
	Δ Pitch angle [deg]	(-400, 400)	From commanded
	Δ Depth [m]	(0, 225)	From commanded
Boolean	Drop weight dropped		
	Buoyancy pack full		
	Surface depth		Depth = 0 m
	Stop envelope		Safety metric
	YoYo envelope		Safety metric
	Going to surface		Safety metric

*Quantization interval centers are N equally-spaced values between (a, b), where N=25 for all numerical signals

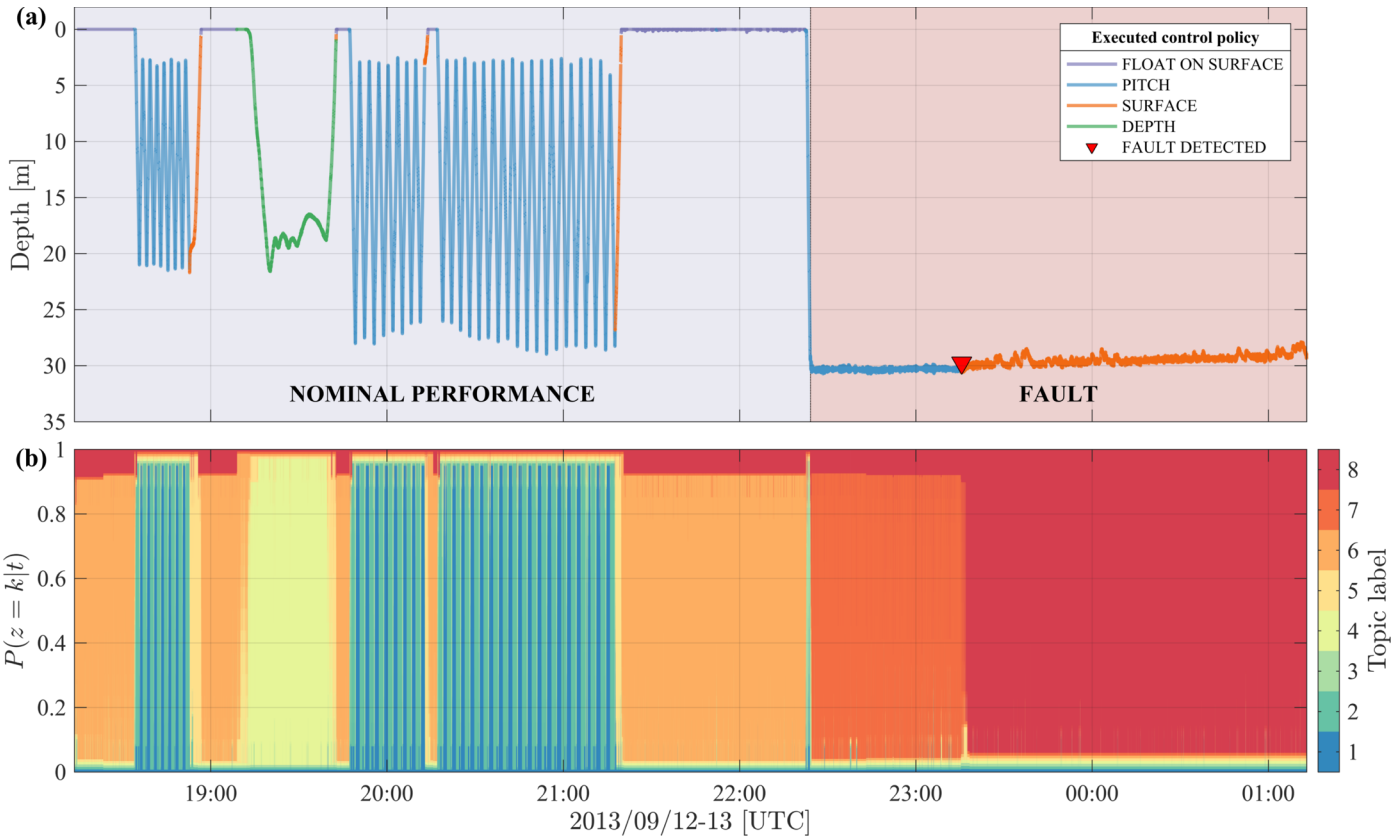

Fig. 2. **(a)** Time series of vehicle depth (2013 dataset); line color indicates the executed control policy and the red shaded background indicates the bottoming fault. The LRAUV system identified the "failure to ascend" fault approximately 50 minutes after the vehicle had bottomed (red triangle). **(b)** A stacked plot showing the distribution of topic labels for each time step, computed using BNP-ROST. The learned topics exactly match the various control policies and unique topics are assigned to the deployment segments where the AUV has bottomed (topics 7 and 8).

included 62,920 observations, and ran the BNP-ROST algorithm to compute topic distributions for each time step. We defined the size of each temporal neighborhood to be equivalent to a single time-step and used $\alpha = 0.1, \beta = 5$ and $\gamma = 1e\text{-}5$ as the LDA and CRP hyperparameter inputs to the model. After the model was trained, we evaluated the correspondence between the learned topics and the control policies (Eq. 4-5) using a time-series of the control policies that were logged onboard the LRAUV (line color in Fig. 2a), and evaluated the topics' correspondence with the fault using an operator-labeled fault record of the deployment (red shading in Fig. 2a).

Fig. 2b shows the distribution of topic labels for each time step (θ_t) and illustrates how topics change over time within the model; the width of bands indicate the topic proportions. As shown, the executed control policies correspond well to the topics, and are consistently represented by a single topic. Unique topics are assigned to the deployment segments where the AUV has bottomed (topics 7 and 8).

Table 2 shows a summary of the conditional probabilities $\mathrm{P}(class|z = k)$ computed using Eq. 4-5 to evaluate the correspondence between the learned topics and the control policies and the fault. The semantic labels assigned to each topic are also shown.

TABLE 2

Semantic Labeling of Topics

	P(control policy|topic)				P(health state|topic)	
	Float on sur.	Pitch	Surface	Depth	Nominal	Fault
Topic 1	0.02	**0.95**	0.03	0.01	**0.97**	0.03
Topic 2	0.01	**0.96**	0.02	0.01	**0.97**	0.03
Topic 3	0.05	**0.86**	0.07	0.02	**0.90**	0.10
Topic 4	0.05	0.05	0.04	**0.87**	**0.95**	0.05
Topic 5	0.09	0.13	**0.54**	0.24	**0.87**	0.13
Topic 6	**0.93**	0.02	0.01	0.03	**0.98**	0.02
Topic 7	0.02	**0.93**	0.04	0.01	0.05	**0.95**
Topic 8	0.06	0.05	**0.89**	0.00	0.07	**0.93**

The semantic labels assigned to each topic are shown in bold text and shaded background

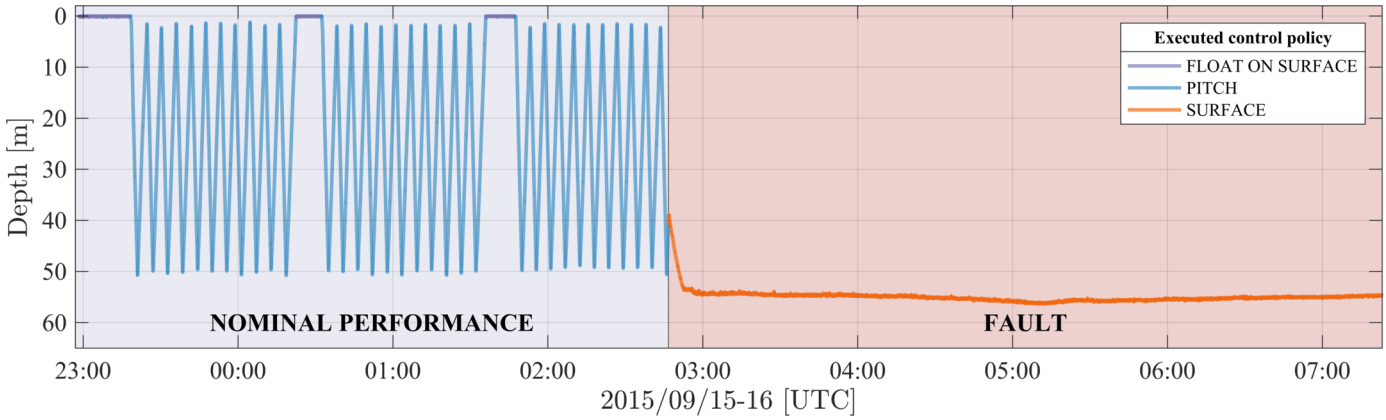

Fig. 3. Time series of vehicle depth (2015 dataset); line color indicates the executed control policy and the red shaded background indicates the fault.

B. Test Set

The test dataset was collected by LRAUV along the coast of Año Nuevo, California, between September 15-16, 2015. Similar to the 2013 dataset, the test set contained a fault in the mass-shifting system that caused the LRAUV to bottom and led to temporary loss of the vehicle (Fig. 3). The fault was triggered by an erroneous software configuration that caused the internal mass-shifter to repeatedly overload and eventually disabled it. Unlike the 2013 incident, the LRAUV's onboard fault detection system detected the fault immediately and triggered the safety behaviors at 02:46 UTC. The fault prevented the LRAUV from adjusting its trim, and eventually caused it to bottom.

We extracted state-words from the dataset, which included 75,920 observations, and computed KL similarities between the distributions of state-words extracted from each observation and the topic-word distributions (ϕ) learned from the 2013 data. Then, we labeled each time step according to its nearest-neighboring topic, and validated the classification results against the time-series of executed control policies and the fault-record that were obtained from the vehicle's log files. For comparison, we repeated the procedure with the 2013 training dataset to attain a "in sample" classification accuracy estimate.

A summary of the classification accuracies obtained for the test and training datasets are shown in Table 3.

In the test dataset (2015), the proposed KL-based nearest-neighbor classifier accurately classified the state of the AUV's health in 99.5% of observations and predicted the executed control policy correctly in 99.8% of observations (on average). More importantly, the classifier detected the bottoming fault with no false positives (highlighted in gray shading in Table 3). The classifier identified the bottoming fault at 02:48:23 UTC, 1.65 minutes after the LRAUV's onboard fault detection system identified the overload fault in the mass-shifting system, and approximately 3.8 minutes before the AUV collided with the sea floor.

In the training dataset (2013), the proposed classifier accurately classified the state of the system's health in 99.96% of observations (with no false positives) and predicted the executed control policy correctly in 99.3% of observations (on average). The classifier identified the bottoming fault at 22:24:08 UTC, nearly 51 minutes before the LRAUV's onboard fault detection system, and approximately 0.4 minutes (25 seconds) before the AUV had bottomed.

TABLE 3
Summary of KL Nearest-Neighbor Classification Accuracies

Dataset	Class	Label	**Accuracy (%)**	TPR (%)	FPR (%)	TNR (%)	FNR (%)
Training set (2013)	Health state	Nominal	**99.96**	100.00	0.10	99.90	0.00
		Fault	**99.96**	99.90	**0.00**	100.00	0.10
	Control policy	Surface	**99.55**	99.31	0.35	99.65	0.69
		Depth	**99.28**	92.57	0.13	99.87	7.43
		Pitch	**99.40**	98.92	0.31	99.69	1.08
		Float on sur.	**98.94**	99.64	1.28	98.72	0.36
Test set (2015)	Health state	Nominal	**99.49**	100.00	0.94	99.06	0.00
		Fault	**99.49**	99.06	**0.00**	100.00	0.94
	Control policy	Surface	**99.88**	99.79	0.02	99.98	0.21
		Depth	**99.87**	0.00	0.13	99.87	0.00
		Pitch	**99.88**	99.79	0.02	99.98	0.21
		Float on sur.	**99.87**	0.00	0.13	99.87	0.00

TPR: True Positive Ratio, FPR: False Positive Ratio, TNR: True Negative Ratio, FNR: False Negative Ratio

IV. Conclusion

We applied a generative probabilistic framework for unsupervised learning of a performance model and for health monitoring of an AUV. We evaluated the framework in post-processing using state-sensor data collected by the *Tethys* LRAUV in two separate deployments, both of which included faults that led to temporary loss of the vehicle. We used one dataset as a training set for the topic-model, and the second to evaluate classification performance on unseen data.

Our results demonstrate that the framework was able to automatically characterize patterns that relate to vertical plane performance of the vehicle, and classify faults with high probability of detection and no false detects. A key feature of the framework is that it is data-driven and does not require expert knowledge. Instead, the dynamic relationship between the actuators and the vehicle's performance is learned directly from the training data. The Bayesian nonparametric nature of the approach ensures that the model adapts automatically to the size and complexity of the data.

Our ongoing efforts aim to extend the proposed framework to facilitate fault isolation (root cause diagnosis). We are interested in the development of a health monitoring architecture that leverages the topic-based semantic representation of the system's state to inform autonomous replanning and automatic selection of mitigation actions in response to failures. In addition, we are working on implementing the proposed technique onboard an underwater robot to facilitate health monitoring and online learning of performance topic models in real time.

Acknowledgment

We are grateful for support from the Office of Naval Research (ONR grant N00014-14-1-0199) and the David and Lucile Packard Foundation. The authors thank Brett Hobson, M. Jordan Stanway, Jon Erickson, Denis Klimov, Ed Mellinger, and Carlos Rueda for LRAUV operations and for insightful discussions.

References

[1] J. G. Bellingham and K. Rajan, "Robotics in remote and hostile environments," Science (80-.)., vol. 318, no. 5853, pp. 1098–1102, 2007.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, no. Jan, pp. 993–1022, 2003.

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

[4] Y. Girdhar, P. Giguère, and G. Dudek, "Autonomous adaptive exploration using realtime online spatiotemporal topic modeling," Int. J. Rob. Res., p. 278364913507325, Jul. 2013.

[5] D. M. Blei, "Probabilistic Topic Models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[6] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proc. Natl. Acad. Sci. United States Am., vol. 101, pp. 5228–5235, Jul. 2004.

[7] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," J. ACM, vol. 57, no. 2, p. 7, Jul. 2010.

[8] S. J. Gershman and D. M. Blei, "A tutorial on Bayesian nonparametric models," J. Math. Psychol., vol. 56, no. 1, pp. 1–12, 2012.

[9] S. Kullback, *Information Theory and Statistics*. Courier Dover Publications, 2012.

[10] B. W. Hobson, J. G. Bellingham, B. Kieft, R. McEwen, M. Godin, and Y. Zhang, "Tethys-class long range AUVs - extending the endurance of propeller-driven cruising AUVs from days to weeks," in 2012 IEEE/OES Autonomous Underwater Vehicles (AUV), 2012, pp. 1–8.

[11] J. G. Bellingham, Y. Zhang, J. E. Kerwin, J. Erikson, B. Hobson, B. Kieft, M. Godin, R. McEwen, T. Hoover, J. Paul, A. Hamilton, J. Franklin, and A. Banka, "Efficient propulsion for the Tethys long-range autonomous underwater vehicle," in 2010 IEEE/OES Autonomous Underwater Vehicles, 2010, pp. 1–7.

[12] B. Kieft, J. Bellingham, M. Godin, B. Hobson, T. Hoover, R. McEwen, and E. Mellinger, "Fault Detection and Failure Prevention on the Tethys Long-Range Autonomous Underwater Vehicle," in 17th International Unmanned, Untethered Submersible Technology Conference, Portsmouth, NH, 2011.