A Bayesian Nonparametric Approach to Clustering Data from Underwater Robotic Surveys

Daniel M. Steinberg, Ariell Friedman, Oscar Pizarro and Stefan B. Williams

Abstract The use of robots for scientific mapping and exploration can result in large, rapidly growing data sets that make complete analysis by humans infeasible. This situation highlights the need for automated means of converting raw data into scientifically relevant information. This paper applies a Bayesian nonparametric model, the variational Dirichlet process, to clustering large quantities of seafloor imagery in an unsupervised manner. It has the attractive property that it does not require knowledge of the number of clusters a-priori, which enables truly autonomous sensor data abstraction. The underlying data representation uses descriptors for colour, texture and 3D structure that are obtained from stereo imagery. This approach consistently produces easily recognisable clusters that approximately correspond to different habitat types. These clusters are useful in observing spatial patterns, focusing expert analysis on subsets of seafloor imagery, aiding mission planning, and potentially informing real time adaptive sampling. We present results from two different surveys with large spatial extents, consisting of thousands of stereo image pairs. We use hand labelled observations from one of these surveys to compare the variational Dirichlet processes' performance to other clustering algorithms.

1 Introduction

The increasing use of robots in scientific mapping and exploration missions has resulted in large and growing data sets that potentially contain a wealth of semantic information relevant to scientists and mission planners.

While supervised classification approaches have shown great promise in answering specific questions (e.g., object and scene recognition) [19, 27], they require train-

Daniel Steinberg, Ariell Friedman, Stefan Williams and Oscar Pizarro
Australian Centre for Field Robotics (ACFR), University of Sydney, Australia
e-mail: {d.steinberg,a.friedman,o.pizarro,stefanw}@acfr.usyd.edu.au

ing sets produced by human labelling of examples. Depending on the application, this can represent a substantial (and expensive) human effort. Producing training data and then evaluating system performance becomes a bottleneck that limits the practical use of fundamentally sound classification approaches.

There is, however, scope for the use of unsupervised classification or clustering approaches for a broad range of applications that can benefit from automatic, preliminary summaries of data. Document analysis [5] as well as image databases [21] clearly demonstrate this. In the context of data gathering with robots, being able to cluster survey images with minimal human input allows scientists and mission planners to easily assign meaningful (albeit coarse) labels to a low number of clusters of images rather than to thousands of individual images. These clusters provide an approximate representation of what was surveyed, while the spatial distribution of these clusters allows scientists to generate new hypotheses related to these spatial patterns and the content of the clusters.

While there are several clustering techniques available, the most commonly used ones require specifying the number of clusters or use tests that approximate some measure of parsimony. An unsupervised approach that also determines the number of clusters in the data is attractive as it could form the basis of a truly autonomous approach to sensor data abstraction, which in turn could inform adaptive behaviours. Applications using communication links with limited bandwidth (underwater exploration) or large latencies (planetary exploration) stand to benefit from such capabilities.

Research in this area is exemplified by [11] wherein clustering is applied terrain data. Sequential observations of terrain are provided by a small amphibious robot's actuator feedback sensors. These observations are then clustered by learning Gaussian mixture models (GMM) or k-nearest neighbour (KNN) models using simulated annealing. This work has also been applied to visual imagery in [12], with the resulting image classifications used to guide the robot over a coral reef. Unfortunately this approach requires lengthy offline training, and there is no facility to infer the number of clusters from the data. In a similar fashion, [28] use KNN to classify data from accelerometers on an autonomous ground vehicle into terrain classes. They also train a GMM to detect 'novel' observations during operation, and have the ability to create new classes. This method still requires an offline supervised training phase, yet can incorporate new information. Unfortunately the algorithm cannot disambiguate multiple new classes at one time.

This paper applies a Bayesian nonparametric approach to data clustering in robotic applications. The use of a nonparametric technique has the advantage of determining the number of clusters automatically. We use a variational approximation that allows for fast, deterministic inference in large scale data sets. The underlying model is naturally extendible for incremental use and hierarchical representations.

Bayesian nonparametric models are data driven, so performance depends heavily on using discriminative features to represent input data. We introduce a set of tailored descriptors for colour, texture and 3D structure that are suitable for seafloor stereo imagery. These features were selected based on the Gaussian mixture crite-

ria of the clustering model, and also upon prior experience with a variety of other benthic image datasets that were collected across a number of different dives.

The remainder of this paper is organized as follows. Section 2 presents an overview of the Bayesian nonparametric model used in this paper. In Section 3 we introduce the image and 3D descriptors that serve as the inputs to the model. Experimental results based on Autonomous Underwater Vehicle (AUV) surveys are presented in Section 4, and in Section 5 we conclude and discuss future work.

2 The Variational Dirichlet Process model

Bayesian nonparametric models have the property that they only increase in complexity as the size of the observable dataset increases. In the case of mixture models, this results in the choice of lowest number of clusters that can sufficiently explain the data [23]. The Bayesian nonparametric model used in this work is the (non-accelerated) variational Dirichlet process (VDP) model derived by Kurihara et al [18]. We will now briefly introduce the VDP and some specifics when learning by variational Bayes.

2.1 Variational Dirichlet Process for Gaussian mixtures

The VDP can be derived for any exponential family mixture model. In this paper we use a Gaussian mixture model since it lends itself well to clustering applications. The objective is to group N observations of the environment, $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, with a dimensionality D, into an *unspecified* number of clusters (indexed by k). Each observation has a latent indicator variable, $\mathbf{Z} = \{z_i\}_{i=1}^{N}$, that assigns it to a cluster. This model assumes each cluster is a Gaussian with its own mean, μ_k, and precision, Λ_k, parameters. Ideally these clusters represent groups of data that are semantically similar. The whole dataset is then represented by a weighted sum of these Gaussian clusters, with weight parameters π_k,

$$p(\mathbf{x}_i) = \sum_{k}^{\infty} \pi_k \mathcal{N}\left(\mathbf{x}_i | \mu_k, \Lambda_k^{-1}\right). \tag{1}$$

The mixture weights make up the marginal probability distribution of the latent variables, $p(z_i) = \prod_{k=1}^{\infty} \pi_k^{\mathbf{1}[z_i=k]}$, where $\mathbf{1}[\cdot]$ is an indicator function, and returns 1 when the condition in the brackets is true, and 0 otherwise. When used as a prior over the model parameters, $(\pi_k, \mu_k, \Lambda_k)$, the Dirichlet Process allows for a mixture model to have a countably infinite number of clusters. However, only a few clusters actually exist with observations belonging to them in the learned model.

The Dirichlet process for this model is realised as a stick-breaking process [14]. This process allows for block updates of the indicator variables, and so suits

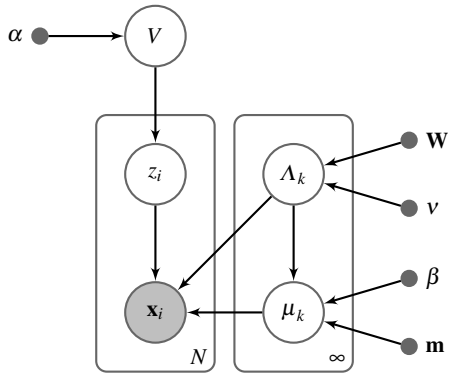

Fig. 1: Graphical model of a Gaussian mixture variational Dirichlet process. The shaded node is observable, the dots are point estimates, and the plates denote replication.

Expectation-Maximisation style algorithms such as variational Bayes. The mixture weights are a function of an infinite collection of 'stick lengths', $V = \{v_k\}_{k=1}^{\infty}$,

$$\pi_k(V) = v_k \prod_{j=1}^{k-1} (1 - v_j). \tag{2}$$

These stick lengths are drawn from a Beta distribution,

$$p(v_k) = \mathcal{B}(1, \alpha). \tag{3}$$

A conjugate Gaussian-Wishart prior distribution, with hyperparameters $\{\mathbf{m}, \beta, \mathbf{W}, \nu\}$, is placed over the Gaussian mean and precision parameters,

$$\begin{aligned} p(\mu_k, \Lambda_k) &= p(\mu_k | \Lambda_k)\, p(\Lambda_k), \\ &= \mathcal{N}\big(\mathbf{m}, (\beta\Lambda_k)^{-1}\big)\, \mathcal{W}(\mathbf{W}, \nu). \end{aligned} \tag{4}$$

In this way, we represent the Dirichlet process prior, $\mathrm{DP}(\alpha, \mathcal{NW})$, for our Gaussian mixture model. The graphical model of this Gaussian mixture version of the VDP is shown in Fig. 1.

2.2 The Variational Bayes algorithm

Typically in Bayesian learning, the objective is to tractably learn a model with latent variables, $\mathbf{Z}$, and latent model parameters, θ, that maximises the log marginal data likelihood,

$$\log p(\mathbf{X}) = \log \int p(\mathbf{X}, \mathbf{Z}, \theta)\, d\mathbf{Z} d\theta. \tag{5}$$

In general, evaluating (5) is intractable. Variational Bayes [2] uses the *mean field* approximation for the joint probability density, $p(\mathbf{X}, \mathbf{Z}, \theta) \approx q(\mathbf{Z})\, q(\theta)$, in conjunction with Jensen's inequality to place a lower bound on (5),

$$\begin{aligned}
\log p(\mathbf{X}) &= \log \int q(\mathbf{Z})\, q(\theta) \left[\frac{p(\mathbf{X}, \mathbf{Z}, \theta)}{q(\mathbf{Z})\, q(\theta)} \right] d\mathbf{Z} d\theta, \\
&\geq \int q(\mathbf{Z})\, q(\theta) \left[\log \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} + \log \frac{p(\theta)}{q(\theta)} \right] d\mathbf{Z} d\theta, \\
&= \mathbb{E}_{q_{\mathbf{Z}}, \theta} \left[\log \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} + \log \frac{p(\theta)}{q(\theta)} \right].
\end{aligned} \tag{6}$$

This last term is the *free energy* functional, $\mathcal{F}[q(\mathbf{Z}), q(\theta)]$. It allows for tractable optimisation of the latent variable distributions, $q(\mathbf{Z})$ and $q(\theta)$, when we use conjugate exponential family models.

By taking functional derivatives $\partial \mathcal{F} / \partial q(\mathbf{Z}) = 0$, and enforcing the constraint $\int q(\mathbf{Z}) d\mathbf{Z} = 1$, we can derive an expression for the distribution over the indicator variables. This results in the Variational Bayes Expectation (VBE) step,

$$q(\mathbf{Z}) = \frac{1}{\mathcal{Z}_{\mathbf{Z}}} \exp \left\{ \mathbb{E}_{q_\theta} [\log p(\mathbf{X}, \mathbf{Z}|\theta)] \right\}, \tag{7}$$

where $\mathcal{Z}_{\mathbf{Z}}$ is a normalisation constant. We also need to learn the parameter distribution for the model given the observations and labels. To do this we again take functional derivatives $\partial \mathcal{F} / \partial q(\theta) = 0$ and apply the constraint $\int q(\theta) d\theta = 1$. This results in the Variational Bayes Maximisation (VBM) step,

$$q(\theta) = \frac{1}{\mathcal{Z}_\theta} p(\theta) \exp \left\{ \mathbb{E}_{q_{\mathbf{Z}}} [\log p(\mathbf{X}, \mathbf{Z}|\theta)] \right\}. \tag{8}$$

Practically, this step gives us point estimates of the variational posterior hyperparameters which govern the distribution $q(\theta)$. Notice that (8) includes a prior term over the parameters, $p(\theta)$. This results in Bayesian updates over these parameters that *penalise* model complexity. The variational Bayes algorithm cycles between updates to the VBE and VBM steps in (7) and (8) until the free energy functional (6) converges to a local maximum. See [2, 3] for proofs and more detail.

2.3 Variational Bayes for the Variational Dirichlet Process

The VDP as presented by [18] is for the general exponential family mixtures. In this section we present the algorithm for Gaussian mixtures. This VDP has an infinite number of classes, however only K classes actually have observations associated with them in its variational approximation. All variational posterior distributions

that index $k > K$ are assigned zero probability, i.e. $q(z_i > K) = 0$ as per [17], allowing the free energy for the VDP to be tractably calculated. We actually minimise *negative* free energy, to be consistent with the convention of minimising negative log likelihood,

$$\mathcal{F}_n = \sum_{k=1}^{K} \left\{ \mathbb{E}_{q_{v_k}} \left[\log \frac{q(v_k)}{p(v_k)} \right] + \mathbb{E}_{q_{\mu_k,\Lambda_k}} \left[\log \frac{q(\mu_k,\Lambda_k)}{p(\mu_k,\Lambda_k)} \right] \right\} - \sum_{i=1}^{N} \mathcal{L}_i. \tag{9}$$

Here $\mathcal{L}_i$ is the complete-data log-likelihood,

$$\begin{aligned} \mathcal{L}_i &= \mathbb{E}_{q_{z_i,V,\mu,\Lambda}} [\log p(\mathbf{x}_i, z_i | V, \mu, \Lambda)] \\ &= \log \sum_{k=1}^{K} \exp \left\{ \mathbb{E}_{q_V} [\log \pi_k] + \mathbb{E}_{q_{\mu_k,\Lambda_k}} [\log \mathcal{N}(\mathbf{x}_i | \mu_k, \Lambda_k)] \right\}. \end{aligned} \tag{10}$$

The details of these expectations can be found in [18] and [4]. By minimising (9), we naturally trade-off fitting the model hyperparameters to the data (10), against full Bayesian model complexity penalty terms. The VBE step for the VDP can be factorised, $q(\mathbf{Z}) = \prod_i q(z_i = k)$, where,

$$q(z_i = k) = \exp \left\{ \mathbb{E}_{q_V} [\log \pi_k] + \mathbb{E}_{q_{\mu_k,\Lambda_k}} [\log \mathcal{N}(\mathbf{x}_i | \mu_k, \Lambda_k)] - \mathcal{L}_i \right\}. \tag{11}$$

The VBM step for this algorithm can also be factorised over the parameters, $q(\theta) = \prod_k q(v_k)\, q(\mu_k, \Lambda_k)$. Since this model is fully conjugate, the variational posteriors will have the same forms as the priors,

$$q(v_k) = \mathcal{B}\left(\tilde{\alpha}_{1,k}, \tilde{\alpha}_{2,k}\right), \tag{12}$$

$$q(\mu_k, \Lambda_k) = \mathcal{N}\left(\tilde{\mathbf{m}}_k, (\tilde{\beta}_k \Lambda_k)^{-1}\right) \mathcal{W}\left(\tilde{\mathbf{W}}_k, \tilde{\nu}_k\right). \tag{13}$$

The variational posterior hyperparameters (denoted by $\tilde{\cdot}$) are simply linear combinations of the prior hyperparameters $(\alpha, \beta, \nu, \mathbf{m}, \mathbf{W})$ and the sufficient statistics of the data weighted by $q(z_i = k)$. Details again can be found in [18] and [4]. The stick-breaking construct in (2) enforces an order to the clusters, and this order must be preserved when updating the hyperparmeters in (12). The variational Bayes algorithm for the VDP is given in Algorithm 1.

Variational Bayes can automatically eliminate superfluous clusters, however it cannot explicitly create clusters. The greedy cluster splitting heuristic used by [18] for cluster creation is also implemented in this work. The variational Bayes algorithm is run over all of the data until it converges. The resulting Gaussian clusters are then split in a direction perpendicular to their principal components. These splits are then refined by running variational Bayes over the resulting clusters. Only a subset of the data needs to be used for this refinement stage, which dramatically reduces the run-time. If the best split (max free energy) increases the overall free energy of the model by more than a threshold, it is accepted. Variational Bayes is then run

Algorithm 1 The VDP variational Bayes algorithm

Require: $\mathbf{X}$, initial $q(\mathbf{Z})$, $prior = \{\alpha, \mathbf{m}, \beta, \mathbf{W}, \nu\}$
  Initialise $\mathcal{F}_n^{(t)}$ to some high value.
  repeat
    VBM: update $\{\tilde{\alpha}_k, \tilde{\mathbf{m}}_k, \tilde{\beta}_k, \tilde{\mathbf{W}}_k, \tilde{\nu}_k\}_{k=1}^K$, $\tilde{\alpha}$ is dependent on cluster size order
    VBE: update $q(\mathbf{Z})$
    $\mathcal{F}_n^{(t+1)} \leftarrow -\mathcal{F}\left[q(\mathbf{Z}), q\left(\{\nu_k, \mu_k, \Lambda_k\}_{k=1}^K\right)\right]$
  until $(\mathcal{F}_n^{(t)} - \mathcal{F}_n^{(t+1)})/\mathcal{F}_n^{(t)} \leq C$
  return $\mathcal{F}_n^{(t+1)}$, $q(\mathbf{Z})$, $posterior = \{\tilde{\alpha}_k, \tilde{\mathbf{m}}_k, \tilde{\beta}_k, \tilde{\mathbf{W}}_k, \tilde{\nu}_k\}_{k=1}^K$

over all of the data again, and the splitting process is repeated. If the best split is not accepted, the algorithm terminates.

3 Features

The VDP is a Bayesian nonparametric model that infers its structure entirely from the data to be clustered. Consequently, the descriptors need to be chosen such that the distance measure between them behaves similarly to the 'distance' between the corresponding semantic content of the images.

Most attempts at automated image-based classification use features extracted from monocular images to derive descriptors. Their success is ultimately limited by the 2D nature of the images and the lack of any notion of scale.

Features such as spin maps [15] or Local Feature Histograms [13] have been used for 3D object detection but they are not well suited for unstructured 3D scenes. Simple habitat complexity indices, such as rugosity, slope and aspect are often used as a proxy for marine biodiversity in the ecological literature [8, 20]. These measures are typically collected in situ by divers using chain-tape methods or profile gauges. An autonomous underwater vehicle (AUV) capable of high precision navigation and equipped with stereo cameras can recover bathymetry at fine resolutions over relatively large, contiguous extents of seafloor. This bathymetry is then used to extract the 3D features at multiple scales, and combined with 2D features to derive the image descriptors [6, 10].

3.1 Image appearance features

An important consideration is that the features generalise well and are robust to factors such as rotation and changes in illumination. Commonly used feature detectors, for example SIFT [19], are useful for object recognition, but need to be incorporated into complex frameworks, such as a Bag of Features approach [21], in order to be useful for scene description. Other approaches, such as the ones used by [26] have

shown promise for natural scene detection, but require prior training to generate filters for known scene types. We have adopted relatively simple measures of colour and image texture to describe the overall appearance of the scene contained in an image, with no prior training.

Texture — we use Local Binary Patterns (LBP) [22]. It can be computed at multiple scales and made to be uniform and rotation invariant. The LBP operator is by definition invariant against monotonic transformations in illumination. This makes it useful for texture classification with non-uniform illumination conditions. Compared to Gabor wavelet texture classification [9], LBPs have been found to yield similar levels of performance with much lower computational cost and without the need to predefine a filter bank [25].

Colour — We use mean shift image segmentation in the L*a*b* colour space for our colour features. Mean shift image segmentation is a non-parametric technique useful for delineating arbitrarily shaped clusters in a complex multimodal feature space [7]. The segmentation is done in L*a*b* colour space in which colour similarity can be measured simply by computing the L2 norms. From the segmentation output we can derive several descriptors:

1. Normalised average segment size – the average size of the homogeneously coloured segments, in pixels, normalised by the number of pixels in the image.
2. Mean of L*a*b* colour modes – the average of the predominant colour of the segments.
3. Standard deviation of L*a*b* colour modes – the variability of the predominant colours of the segments.

3.2 Terrain complexity features

The structure obtained from the stereo imagery is in the form of Delaunay triangulated meshes which are made up by a set of triangular faces that connect vertices to make a 3D surface [16]. The stereo derived 3D measures that are considered in this paper are rugosity, slope and aspect [10].

Rugosity — This is a measure of terrain complexity which is known to correlate with marine biodiversity. The rugosity index for a location in the surface mesh is calculated by centering a window of specified size over the location. Then the area of the contoured surface bounded by the window is divided by the area of the orthogonal projection of the surface onto a plane.

Slope — Slope refers to the angle between the plane of best fit and the horizontal plane. This angle is equivalent to the angle between the normal vectors of the two planes and can be obtained from their dot product.

Aspect — Aspect refers to the direction that the surface slope faces. It is defined as the angle between the positive *North* axis and the projection of the normal onto the *North-East* plane.

3.3 Feature selection

An important consideration when selecting an appropriate feature set is the normality assumption of the VDP model. More specifically, a feature descriptor distribution needs to be representable by a mixture of Gaussians. A multimodal Gaussian distribution is preferable, since it will lead to more discrimination between clusters. In order to assess this we can look at the histograms of each feature.

Aspect is a value in the range $[-\pi, \pi]$. For analytical purposes, it is useful to split aspect into vector components to eliminate the discontinuity associated with angular wrap-around. This can be represented as 'northness' and 'eastness'. However, even these vector components of aspect violate the normality assumption, and cannot be used with the VDP[1].

Some variables do not occupy a distribution that can be represented as a mixture of Gaussians. However, it may be possible to transform these variables to make them 'comply'. For instance, rugosity has a log-normal type distribution, which leads to a well-distributed, multi-modal feature that can be easily approximated by a mixture of Gaussians when its logarithm is taken. Other features, such as slope and normalised average segment size, can be transformed in a similar way.

Table 1 shows the selection of features that were used to generate the results in this paper. Extensive testing over multiple sites covering a large latitudinal range in Australia suggests this approach consistently produces easily recognisable, coarse habitat types that are useful in observing spatial patterns and in focusing further, detailed analysis of seafloor imagery.

Table 1: Image features used in the results, $D_{tot} = 23$

Feature	Window size	Dimensionality
log(rugosity − 1)	image	1
log(rugosity − 1)	5×5m	1
log(slope)	5×5m	1
log(rugosity − 1)	10×10m	1
log(slope)	10×10m	1
mean(L*a*b* segment mode)	image	3
st. dev.(L*a*b* segment mode)	image	3
log(mean(L*a*b* segment size))	image	1
st. dev.(Grey-scale image pixels)	image	1
LBPs (radius of 1, 8 samples)	image	10

[1] We suspect that aspect as a descriptor is likely to be a useful predictor of habitat types if combined with other variables such as slope and current flow velocity to obtain a notion of exposure.

4 Results

In this section we use the VDP to cluster seafloor stereo imagery obtained by an AUV. Data from two separate surveys are clustered. The first survey was conducted on the O'Hara marine protected area (MPA) in Tasmania, Australia. The VDP is compared to hand labels and other clustering algorithms on this dataset. The second dataset was obtained in Scott Reef, Western Australia.

The AUV is programmed to follow the seafloor at an altitude of two metres. However, in areas that feature large changes in relief, the AUV may deviate from this significantly, causing changes in the illumination and extent of the scene. In the interest of using well illuminated images, we do not cluster those that were taken at an altitude of more than 3.5 metres.

The most computationally intensive task in generating the results is extracting the stereo and image features. We typically use full size images for stereo feature extraction (1360×1024 pixels) and one-quarter size images for image feature extraction (340×256 pixels). These datasets have on the order of 10,000 images each, so extracting the feature descriptors typically takes a few hours. Most of the computation is in segmenting the images for the L*a*b* descriptors. Missions also typically take a number of hours to complete, so we are confident this processing could be performed in real-time. All of the dimensions of the data are standardised for best performance.

We chose an uninformative Stick-breaking prior by setting the prior hyperparameter $\alpha = 1$. For the Gaussian mixtures we chose semi-informative prior hyperparameters; $\mathbf{m} = \text{mean}(\mathbf{X})$, $\beta = 1$, $\nu = D$, and $\mathbf{W} = \nu C_{width} \lambda^{max}_{\text{cov}(\mathbf{X})} \mathbf{I}_D$. Here D is the dimensionality of the data, $\lambda^{max}_{\text{cov}(\mathbf{X})}$ is the largest eigenvalue of the covariance of the data, $\mathbf{I}_D$ is the identity matrix, and C_{width} is left as a tuneable parameter that encodes the a-priori 'width' of the mixtures.

Clustering these datasets with a C++ implementation of the VDP takes on the order of seconds to minutes depending on the number of images. All results are generated in Matlab R2010b/C++ using a 2.8 GHz Core 2 Duo Intel processor with 4 GB of 1067 MHz RAM.

4.1 O'Hara Marine Protected Area, Tasmania

This dataset contained 11,000 stereo image pairs featuring 10 classes based on hand labels provided by a marine scientist. Given the availability of hand labels for this dataset, it was used to compare the VDP to two other unsupervised clustering algorithms. One of the algorithms used for comparison is a variant of self-tuning Spectral Clustering (SC) [29] with a sparse similarity matrix and the eigen-gap heuristic to choose K. We also used Expectation-Maximisation for Gaussian Mixture models for different values of K, using the Bayes Information Criterion to select the best K (GMM+BIC). These are both Matlab implementations.

Table 2: Autonomous underwater vehicle dataset clustering results. The VDP with different prior C_{width} parameters, GMM+BIC and SC algorithms were compared to hand labels using the V-measure. The GMM+BIC and SC algorithms are Matlab implementations, and so are not directly comparable to the VDP in terms of runtime.

Algorithm	V-measure	Homogeneity	Completeness	K	Time	$\mathcal{F}_n$ ($\times 10^5$)
VDP, C_{width} = .01	0.7125	0.6943	0.7316	7	17.8 s	5.726
VDP, C_{width} = .02	0.7197	0.6824	0.7614	6	14.1 s	5.843
VDP, C_{width} = .04	**0.7310**	0.6841	0.7847	6	11.6 s	5.997
GMM + BIC	0.6789	**0.7318**	0.6331	10	(81.0 s)	N/A
SC	0.6014	0.4460	**0.9228**	3	(7.2 s)	N/A

For these experiments we compare the clustering results to the hand labelled dataset using the *V-measure* [24]. The V-measure is the harmonic mean of two opposing measures; *Homogeneity* and *Completeness*. A cluster solution with a high level of homogeneity has data points that are members of a cluster comprised of only one single ground truth class. A cluster solution with a high level of completeness has all data points that are members of a single truth class belonging to a single cluster. All measures range from zero to one, with one being a perfect score. Homogeneity and Completeness are weighted equally in the V-measure for all of our experiments.

The results are summarised in Table 2 for the VDP with three C_{width} values. The VDP for all C_{width} values has a better V-measure than SC and the GMM+BIC algorithms. The GMM+BIC has the highest homogeneity measure, and SC the highest completeness. However, homogeneity tends to reward a higher number of classes, while completeness a lower number.

The clustering results for the VDP with $C_{width} = 0.04$ is shown in Fig. 2. It can be seen in Fig. 2d that the VDP clusters tend to occupy certain depth ranges, which is consistent with habitat distributions. This suggests that the VDP clusters are coarsely representative of habitat types. Furthermore, in Fig. 2b, visually dissimilar clusters, such as sand, screw shell rubble and reef, are separate in the feature space, suggesting that the features are successfully capturing some of the semantic meaning in the scenes.

4.2 Scott Reef, Western Australia

A 3D reconstruction of the Scott Reef dataset is shown in Fig. 3c. It is a dense, fully overlapping grid survey consisting of 50 parallel track-lines, each 75m long and spaced one meter apart. This survey appears to have roughly three habitat types; a reef habitat, a sand habitat, and a partially populated substrate at the interface of the reef and sand habitats. Unfortunately we did not have a hand labels for this dataset, however because of its dense nature it is easy to visually validate the clustering results.

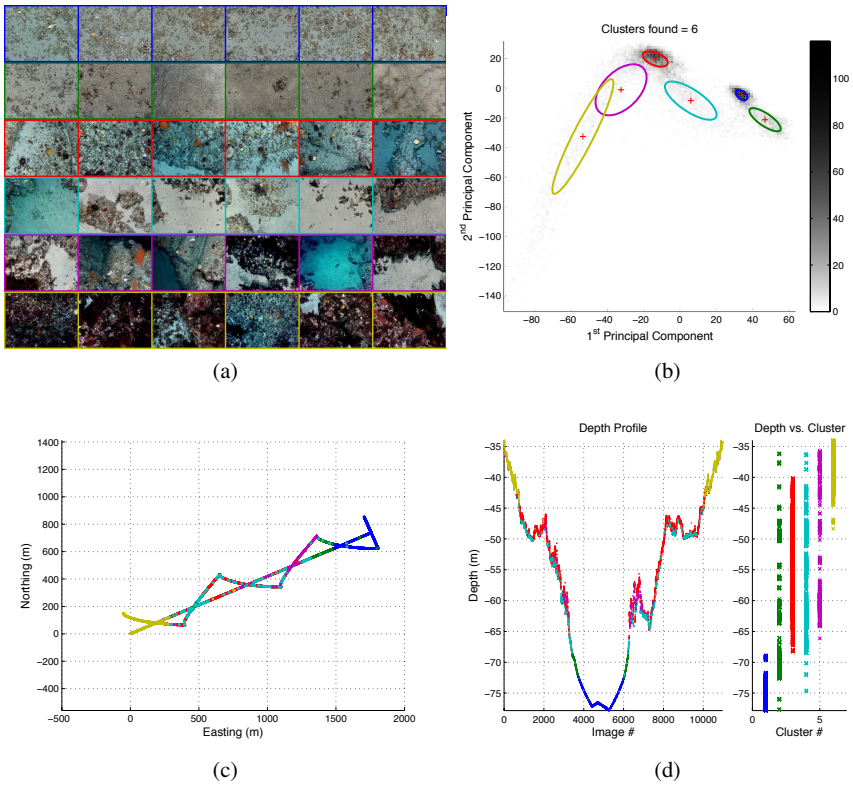

Fig. 2: O'Hara MPA results for the VDP with $C_{width} = 0.04$. (a) Six sample images from each cluster found by the VDP. (b) A projection of a histogram of the observations (bins) and the VDP clusters (ellipses) onto the first two principal components of the data. The intensity of the bins indicates the number of observations. (c) Class labels overlaid on the vehicle path showing consistent spatial distribution of the clusters. (d) The clusters in this dive are highly correlated with depth despite depth not being used as a feature. The resulting clusters correspond roughly to kelp dominated habitats in depths below 45 m, patch reef dominated by sponges between 45 and 70 m and sand and screw-shell rubble below 70 m.

Approximately 9,800 stereo image pairs were obtained on this dive, and the VDP took 9.9 seconds to cluster them with $C_{width} = 0.02$. Six clusters were found by the VDP, the results are summarised in Fig. 3. We tried the same range of C_{width} values as the O'Hara dataset, the results did not vary drastically, but a value of 0.02 yielded slightly more visually appealing clusters. It is apparent from Fig. 3d that the clusters are spatially contiguous, despite the VDP having no notion of the spatial layout of the dive. Some of this contiguity could be attributed to the large window size features that have been used. However, it is our experience that we achieve these

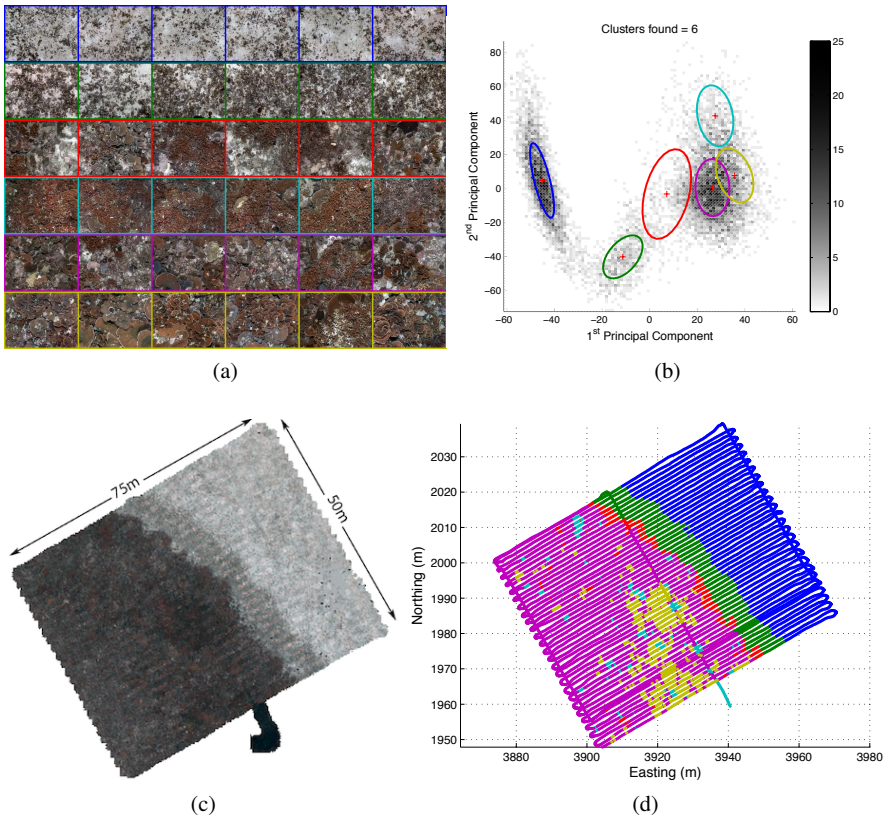

Fig. 3: Scott Reef results with $C_{width} = 0.02$. (a) Six sample images from each cluster found by the VDP. (b) A projection of a histogram of the observations and the VDP clusters onto the first two principal components of the data. (c) A visual reconstruction of the Scott Reef dataset showing the distinct habitats featured in this dataset. (d) Class labels overlaid on the vehicle path with each dot corresponding to the location of a stereo pair image, and its colour to a cluster. The clusters in this dataset are spatially contiguous despite the VDP having no notion of the spatial layout of the dive.

spatial patterns using just image level features, and that including large window size features tends to generate more distinct clusters.

5 Conclusion and Future Work

The Variational Dirichlet Process is a completely data driven algorithm. In combination with carefully chosen features, the VDP is a powerful tool for aggregating images of the benthos into clusters without any human supervision. It outperforms Spectral Clustering, and Gaussian mixtures learned with EM and the Bayes information criterion on this type of data when compared to hand labels. The results also appear to exhibit high spatial correlation despite the VDP not accounting for the spatial layout of the images. This suggests that the clusters discovered in these datasets approximately represent habitat types, and usefully summarise these datasets. We have received positive feedback from benthic ecologists who have used the output from our system. The labels generated by the VDP have also been used to extrapolate the presence of the habitats away from the survey area [1], demonstrating that this algorithm could inform adaptive sampling decisions.

We used image appearance features based on Mean Shift Segmentation and Local Binary Patterns, combined with multi-scale 3D features of rugosity and slope (inspired by marine ecology literature) to generate the results presented in this paper. We intend to conduct a more thorough study on the selection and comparison of a wider set of image appearance and 3D features.

The combination of these features, and the VDP algorithm has been used on many visual underwater datasets, with many varying habitat types. They have proven to generalise very well. We also see no reason why this algorithm and a similar set of features could not be applied to other visual datasets, such as aerial or space-based robotic surveys.

Currently we are deriving various hierarchical clustering models based on the VDP. Such models may be used to relate habitat clusters between different dives, while preserving the cluster proportions within dives. Generative models such as the VDP and other hierarchical versions provide a natural framework for incrementally learning datasets. They can cluster observations from an autonomous vehicle as it observes the seafloor, and can also combine data from multiple survey missions without requiring vast processing capability. Our goal is to create algorithms based on these incremental generative models that can further inform expert analysis on subsets of seafloor imagery, aid in mission planning, as well as inform real time adaptive sampling behaviours.

Acknowledgments

This work is supported by the New South Wales State Government and the Integrated Marine Observing System (IMOS) through the DIISR National Collaborative Research Infrastructure Scheme. The authors of this work would like to thank the Australian Institute for Marine Science and the Tasmanian Aquaculture and Fisheries Institute (TAFI) for making ship time available to support this study. The crews of the R/V Solander and R/V Challenger were instrumental in facilitating successful

deployment and recovery of the AUV. We thank Jan Seiler for providing us with the hand labels of the O'Hara dataset. We also acknowledge the help of all those who have contributed to the development and operation of the AUV, the late Duncan Mercer, George Powell, Matthew Johnson-Roberson, Ian Mahon, Stephen Barkby, Ritesh Lal, Paul Rigby, Jeremy Randle, Bruce Crundwell and the late Alan Trinder.

References

[1] Ahsan N, Williams S, Jakuba M, Pizarro O, Radford B (2010) Predictive habitat models from auv-based multibeam and optical imagery. In: OCEANS 2010 MTS/IEEE Seattle, IEEE

[2] Attias H (2000) A variational Bayesian framework for graphical models. Advances in neural information processing systems 12(1-2):209–215

[3] Beal M (2003) Variational algorithms for approximate bayesian inference. PhD thesis, University College London

[4] Bishop C (2006) Pattern Recognition and Machine Learning. Springer Science+Business Media, Cambridge, UK

[5] Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. The Journal of Machine Learning Research 3:993–1022

[6] Bridge T, Done T, Friedman A, Beaman R, Williams S, Pizarro O, Webster J (2010) Variability in mesophotic coral reef communities along the great barrier reef, australia. Marine Ecology Progress Series

[7] Comaniciu D, Meer P (2002) Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(5):603–619

[8] Commito J, Rusignuolo B (2000) Structural complexity in mussel beds: the fractal geometry of surface topography. Journal of Experimental Marine Biology and Ecology 255(2):133–152

[9] Fogel I, Sagi D (1989) Gabor filters as texture discriminator. Biological Cybernetics 61:103–113

[10] Friedman A, Pizarro O, Williams S (2010) Rugosity, slope and aspect derived from bathymetric stereo image 3d reconstructions. In: OCEANS 2010 Sydney, IEEE

[11] Giguere P, Dudek G (2009) Clustering sensor data for terrain identification using a windowless algorithm. Robotics: Science and Systems IV p 25

[12] Giguere P, Dudek G, Prahacs C, Plamondon N, Turgeon K (2009) Unsupervised learning of terrain appearance for automated coral reef exploration. In: 2009 Canadian Conference on Computer and Robot Vision, IEEE, pp 268–275

[13] Hetzel G, Leibe B, Levi P, Schiele B (2001) 3d object recognition from range images using local feature histograms. In: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol 2, pp II–394 – II–399

[14] Ishwaran H, Zarepour M (2000) Markov chain monte carlo in approximate dirichlet and beta two-parameter process hierarchical models. Biometrika 87(2):371

[15] Johnson A, Hebert M (1999) Using spin images for efficient object recognition in cluttered 3d scenes. Pattern Analysis and Machine Intelligence, IEEE Transactions on 21(5):433 –449

[16] Johnson-Roberson M, Pizarro O, Williams S, Mahon I (2010) Generation and visualization of large-scale three-dimensional reconstructions from underwater robotic surveys. Journal of Field Robotics 27(1)

[17] Kurihara K, Welling M, Teh YW (2007) Collapsed variational Dirichlet process mixture models. In: Proceedings of the International Joint Conference on Artificial Intelligence, vol 20

[18] Kurihara K, Welling M, Vlassis N (2007) Accelerated variational dirichlet process mixtures. Advances in Neural Information Processing Systems 19:761

[19] Lowe D (2004) Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60:91–110

[20] McCormick M (1994) Comparison of field methods for measuring surface topography and their associations with a tropical reef fish assemblage. Marine Ecology Progress Series 112(1-2):87–96

[21] Nister D, Stewenius H (2006) Scalable recognition with a vocabulary tree. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp 2161–2168

[22] Ojala T, Pietikaeinen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7):971–987

[23] Orbanz P, Teh YW (2010) Bayesian nonparametric models. In: Encyclopedia of Machine Learning, Springer

[24] Rosenberg A, Hirschberg J (2007) V-measure: A conditional entropy-based external cluster evaluation measure. In: Conference on Empirical Methods in Natural Language Processing

[25] Shan C, Gong S, McOwan P (2005) Robust facial expression recognition using local binary patterns. In: Image Processing, 2005. ICIP 2005. IEEE International Conference on, IEEE

[26] Torralba A, Oliva A (2003) Statistics of natural image categories. Network: Computation in Neural Systems 14(3):391–412

[27] Torralba A, Murphy KP, Freeman WT, Rubin MA (2003) Context-based vision system for place and object recognition. In: Ninth IEEE International Conference on Computer Vision, 2003, pp 273–280 vol.1

[28] Weiss C, Zell A (2008) Novelty detection and online learning for vibration-based terrain classification. In: Proceedings of the 10th International Conference on Intelligent Autonomous Systems (IAS 2008), Baden-Baden, Germany, Citeseer, pp 16–25

[29] Zelnik-Manor L, Perona P (2004) Self-tuning spectral clustering. Advances in neural information processing systems 17(1601-1608):16