Local-HDP

Ayoobi, H.; Kasaei, H.; Cao, M.; Verbrugge, R.; Verheij, B.

Published in:
ArXiv

LOCAL-HDP: INTERACTIVE OPEN-ENDED 3D OBJECT CATEGORY RECOGNITION IN REAL-TIME ROBOTIC SCENARIOS

H. Ayoobi⋆, H. Kasaei, M. Cao, R. Verbrugge, B. Verheij
Bernoulli Institute, University of Groningen, Netherlands
⋆h.ayoobi@rug.nl

ABSTRACT

We introduce a non-parametric hierarchical Bayesian approach for open-ended 3D object categorization, named the Local Hierarchical Dirichlet Process (Local-HDP). This method allows an agent to learn independent topics for each category incrementally and to adapt to the environment in time. Hierarchical Bayesian approaches like Latent Dirichlet Allocation (LDA) can transform low-level features to high-level conceptual topics for 3D object categorization. However, the efficiency and accuracy of LDA-based approaches depend on the number of topics that is chosen manually. Moreover, fixing the number of topics for all categories can lead to overfitting or underfitting of the model. In contrast, the proposed Local-HDP can autonomously determine the number of topics for each category. Furthermore, the online variational inference method has been adapted for fast posterior approximation in the Local-HDP model. Experiments show that the proposed Local-HDP method outperforms other state-of-the-art approaches in terms of accuracy, scalability, and memory efficiency by a large margin. Moreover, two robotic experiments have been conducted to show the applicability of the proposed approach in real-time applications.

1 Introduction

Most recent object recognition/detection techniques are based on deep neural networks [1, 2, 3, 4, 5, 6]. These methods typically need a large labeled dataset for a long training process. The number of object categories (class labels) should be predefined in advance for such methods. However, in real-life robotic scenarios, a robot can always face new object categories while operating in its environment. Therefore, the model should get updated in an open-ended manner without completely retraining the model [7]. Furthermore, object category recognition is not a well-defined problem because of the large inter-category variation (Figure 1 (*left*)), multiple object views for each object (Figure 1 (*right*)), and concept drift in dynamic environments [8].

Object recognition in humans is a complex hierarchical multi-stage process of streaming visual data in the cortical regions [9]. The hierarchical structure of the brain for the object recognition task has motivated us to choose hierarchical Bayesian models like Latent Dirichlet Allocation (LDA) [10] and Hierarchical Dirichlet Process (HDP) [11] for object category recognition.

In this paper, we suggest that 3D visual streaming data should be processed continuously, and object category learning and recognition should be performed simultaneously in an open-ended manner.

Figure 1: An illustrative example of (*left*) intra-category variation of the mug category in the Washington RGB-D dataset, and (*right*) different object views of a mug object.

We propose the Local Hierarchical Dirichlet Process (Local-HDP), an extension of the Hierarchical Dirichlet Process [11] method, which can incrementally learn new topics for each category of objects independently. In contrast to notable recent works [8, 12, 13] using a predefined number of topics, Local-HDP is more flexible since it is a non-parametric Bayesian model that can autonomously determine the number of topics for each category at run-time.

Figure 2 shows the processing layers of the proposed Local-HDP. The tabletop objects are detected in the initial phase (green bounding box around apple on the table in Figure 2). Subsequently, the hierarchy of the five processing layers is utilized. The features layer extracts a set of local shape features using the spin-image descriptor [14]. The computed features are represented as Bag of visual Words (BoWs). The obtained representation is then sent to the topics layer, where a set of topics is inferred autonomously for the given object using the proposed Local-HDP method. Each topic is a distribution over visual words. In other words, the topic layer provides an unsupervised mapping of the BoW representation to the topics space, which can fill the conceptual gap between low-level features and high-level concepts. As shown in the object views layer, the appearance of an object may vary from different perspectives (Figure 1 (*bottom*)). Therefore, it is necessary to infer topics using different object views. There might be different instances in an object category as well (see Figure 1 (*top*)). This point is addressed in the categories layer. Moreover, a simulated teacher has been developed to interact with the model and evaluate its performance in an open-ended manner.
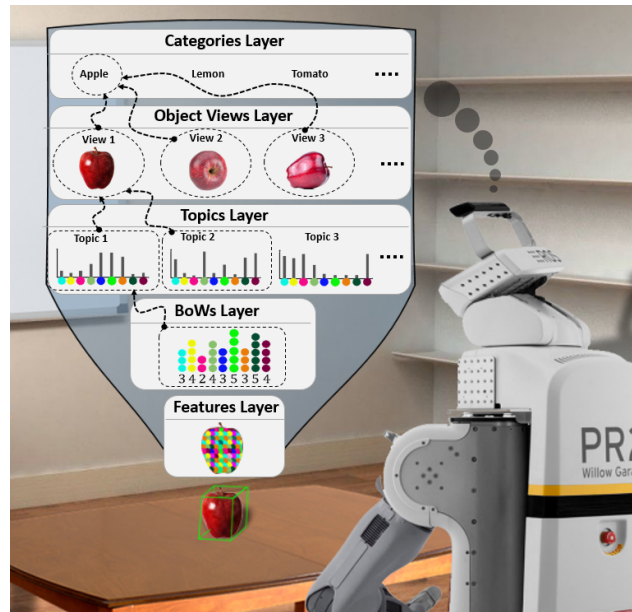

Figure 2: The architecture of the proposed method.

This work extends two approaches, namely Local-LDA [8] and HDP [11], in four aspects. First, our approach can autonomously detect the number of required topics to independently represent the objects in each category, avoiding the limitation of Local-LDA for determining the number of topics in advance. This feature prevents underfitting or overfitting of the model. Second, our research adapts the online variational inference technique [15], which significantly reduces inference time. Third, the proposed local online variational inference method leads to memory optimization since it needs to store a smaller average number of instances per object category in memory. Fourth, our work extends the hierarchical Dirichlet process [11] by learning and updating local topics for each object category independently in an incremental and open-ended fashion.

2 Related Work

Object representation is one of the main building blocks of object recognition approaches. The underlying reason is that the output of the object representation module is used in both learning and recognition. Object representation techniques can be categorized into three groups, namely, global and local object descriptors and machine learning approaches [16]. Notable global object descriptors are Global Orthographic Object Descriptor (GOOD) [17, 18], Ensemble of Shape Functions (ESF) [19] and Viewpoint Feature Histogram (VFH) [20]. Examples of local 3D shape descriptors include Spin-Images (SI) [14], Intrinsic Shape Signature (ISS) [21], and Fast Point Feature Histogram (FPFH) [22]. Local descriptors are more robust to occlusions and clutter. However, comparing pure local descriptors is a computationally expensive task [23]. To alleviate this problem, machine learning techniques like Bag of Words (BoW) [24], Latent Dirichlet Allocation (LDA) [10, 25] and deep learning [26, 27] methods can be used for representing objects in a compact and uniform format.

Kasaei et al. [8] extended Latent Dirichlet Allocation (LDA) and proposed Local-LDA. They showed the application of Local-LDA in the context of open-ended 3D object category learning and recognition. Similar to our approach, Local-LDA learns a set of topics for each object category incrementally and independently. Unlike our approach, in Local-LDA, the same number of topics is chosen in advance based on trials and errors for all of the object categories. A good choice for the number of topics for each object category is correlated to the intra-category variation of each 3D object category. Therefore, choosing the same number of topics for all the object categories with different intra-category variation might be not reasonable. Moreover, in open-ended scenarios, it is not feasible to anticipate the inter-category variation of 3D objects that the model might see in the future and choose a fixed number of topics in advance for all the
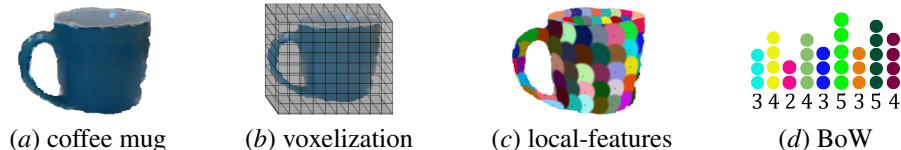
(*a*) coffee mug   (*b*) voxelization   (*c*) local-features   (*d*) BoW

Figure 3: (*a*) The RGB-D image of a coffee mug. (*b*) Key-points selection using voxelizing [8]. (*c*) Key-points neighborhoods are represented by different colors. (*d*) The BoW representation for the given object.

categories. To solve these issues, our approach can autonomously choose the number of topics for each object category on the fly without a need for in advanced trails and errors. This makes our approach more robust for recognizing object categories with various inter-category and intra-category variation and applicable in real-world open-ended scenarios. Local-LDA uses collapsed Gibbs sampling for approximating the posterior probability. However, we adapt the online variational inference technique [15] for Local-HDP.

Our approach builds on the Hierarchical Dirichlet Process (HDP) [11], that is based on Dirichlet process (DP) [28] and mixture of DPs [29]. Posterior inference is intractable for HDP, and much research has been done to find a proper approximate inference algorithm [11, 30, 31]. The Markov Chain Monte Carlo (MCMC) sampling method for DP mixture models has been proposed for approximate inference in HDPs [32]. David Blei et al. proposed the variational inference for DP mixtures [33]. Teh et al. [11] proposed the Chinese Restaurant Franchise metaphor for HDP and used Gibbs sampling method for the inference. The online variational inference approach is proposed by Wang et al. [15] for HDP, which can be used in online incremental learning scenarios and for large corpora. Our method is different from HDP, since HDP only shares the topics among the same categories and not across different categories. This is especially needed in the case of 3D object categorization for open-ended scenarios [8]. HDP has further extensions to construct tree-structured representations for text data which have nested structure [34]. Similar to supervised hierarchical Dirichlet Process (sHDP) [35], we use the category label of each object. Unlike sHDP, we learn object categories in an open-ended fashion, while in sHDP, the number of object categories to be learned should be defined in advance.

Deep learning-based approaches [36, 37, 38] try to learn a sparse representation for 3D objects. Unlike our approach, such methods typically need a large labeled dataset and require long training time. In particular, our proposed approach does not require a large labeled dataset and can incrementally update the model facing an unforeseen object category in an open-ended manner. Moreover, the number of categories is not fixed in open-ended approaches like ours.

3 Method

We assume that an object has already been segmented from the point cloud of the scene, and we hence mainly focus on detailing the Local Hierarchical Dirichlet Process (Local-HDP) approach.

3.1 Pre-Processing Layers

In Figure 2, the first two layers—the feature layer and BoWs layer— are the pre-processing layers. In the feature layer, we first select key-points for the given object and then compute a local shape feature for each key-point. Towards this goal, we first voxelized[1] the object (Figure 3) (*b*), and then, the nearest point to each voxel center is selected as a key-point. Afterwards, the spin-image descriptor [14] is used to encode the surrounding shape in each key-point using the original point cloud (Figure 3 (*c*)). This way, each object view is described by a set of spin-images in the first layer, $\mathbf{O_s} = \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$ where N is the number of key-points. The obtained representation is then sent to the BoWs layer. Since HDP-based models have the bag-of-words assumption - that the order of words in the document can be neglected - the BoWs layer transforms the computed spin-images to a BoW format (Figure 3 (*d*)). Towards this end, the BoWs layer requires a dictionary with V visual words (spin-images). In this work, we have created a dictionary of visual words using the same methodology as Local-LDA [8]. The obtained BoW representation is fed to the topic layer.

3.2 Local Hierarchical Dirichlet Process

After synthesizing the point cloud of the 3D objects to a set of visual words in BoW format, the data is ready to be inserted into the topic layer where the proposed Local-HDP method is employed. In this layer, the model transforms

---

[1]`http://docs.pointclouds.org/trunk/classpcl_1_1_voxel_grid.html`

the low-level features in BoW format to conceptual high-level topics. In other words, each object is represented as a distribution over topics, where each topic is a distribution over visual words. To this end, we use an incremental inference approach where the number of categories is not known beforehand and the agent does not know which additional object categories will be available at run-time. The plate notation of Local-HDP is shown in Figure 4. In this graph, C is the number of categories, $|c|$ is the number of objects in each category. Each object, d, is represented by a set of N visual words, $W_{d,n}$ where n, d show the n'th visual word from the d'th object. Each visual word is an element from the vocabulary of visual worlds with predefined V words, that is $W_{d,n} \in \{1...V\}$. Using a *Coffee Mug* as an example, a distribution over the topics of the *Coffee Mug* should be used to generate the visual words of the object. Accordingly, a particular topic is selected out of the mixture of possible topics of the *Coffee Mug* category to generate the visual words. For instance, coffee mugs typically have a "handle", which is represented as a distribution of visual words that repeatedly occurring together. This can be interpreted as the "handle" topic, which is inferred from the co-occurrence of the visual words in several objects of the same category. The process of choosing a topic and then drawing the visual words from that topic is repeated several times to generate all the visual words of the *Coffee Mug*. After constructing the model in a generative manner, a reverse procedure for inferring the latent variables from the data is used.

3.3 Local Online Variational Inference

In this section, we adapt the online variational inference approximate inference method [15] for Local-HDP. This method can be used in open-ended applications since it can handle streaming data in an online and incremental manner. Moreover, it is faster than traditional approximate inference techniques, e.g., Chinese restaurant franchise [11] and variational inference [33], and it can be used to infer the latent variables of different scale datasets [15].

Online variational inference for HDP is inspired by the online variational Bayes [39] method for LDA. This method tries to optimize a variational objective function [40] exploiting stochastic optimization [41]. Using Sethuraman's stick-breaking construction for HDP [11], the variational distribution for local online variational inference is in the following form:

$$q(\beta', \pi', c, z, \phi) = q(\beta')q(\pi')q(c)q(z)q(\phi) \tag{1}$$

In the terminology of variational inference techniques, q is called the variational approximation to the posterior p. Variational techniques try to solve an optimization problem over a class of tractable distributions Q in order to find a $q \in Q$ that is most similar to p and can be used as its approximation. Moreover, $\beta' = (\beta'_k)_{k=1}^{\infty}$ is the top-level stick proportion, $\pi' = (\phi'_{jt})_{t=1}^{\infty}$ is the bottom-level stick proportion and $c_j = (c'_{jt})_{t=1}^{\infty}$ is the vector of indicators for each G_j. Moreover, $\phi = (\phi_k)_{k=1}^{\infty}$ is the inferred topic distribution, and z_{jn} is the topic index for the *n*th word in the *j*th document w_{jn}.

The factorized form of $q(c)$, $q(z)$, $q(\phi)$, $q(\beta')$ and $q(\pi')$ is the same as the online variational inference for HDP [42]. Assuming that we have $|c|$ objects in each category for Local-HDP, the variational lower bound for object j in category C is calculated as follows:

$$L_j^{(C)} = \mathbb{E}_q[log(p(w_j|c_j, z_j, \phi)p(c_j|\beta')p(z_j|\pi')p(\pi'_j|\alpha_0))] + H(q(c_j)) + H(q(z_j))$$
$$+ H(q(\phi')) + \frac{1}{|c|}[E_q[logp(\beta')p(\phi)] + H(q(\beta')) + H(q(\phi)] \tag{2}$$

Where $H(.)$ is the entropy term for the variational distribution. Therefore, the lower bound term for each category is calculated in the following way:

$$L^{(C)} = \sum_j L_j^{(C)} = \mathbb{E}_j[|c|L_j^{(C)}] \tag{3}$$

Using coordinate ascent equations in the same way as online variation inference, the object-level parameters $(a_j, b_j, \varphi_j, \zeta_j)$ are estimated. To be more specific, a_j and b_j are the parameters of the beta distributions for the bottom-level stick proportions π_j, φ_j is the variational parameter for the vector of indicators c_j, and ζ_j is the variational parameter for the topic z_j. These variables are defined in the same way as in [42]. Then, for the category-level parameters $(\lambda^{(C)}, u^{(C)}, v^{(C)})$, we do gradient descent with respect to a learning rate:

$$\partial\lambda_{kw}^{(C)}(j) = -\lambda_{kw} + \eta + |c|\sum_{t=1}^{T}\varphi_{jtk}(\sum_n \zeta_{jnt}I[w_{jn} = w]) \tag{4}$$

4

$$\partial u_k^{(C)}(j) = -u_k + 1 + |c| \sum_{t=1}^{T} \varphi_{jtk} \tag{5}$$

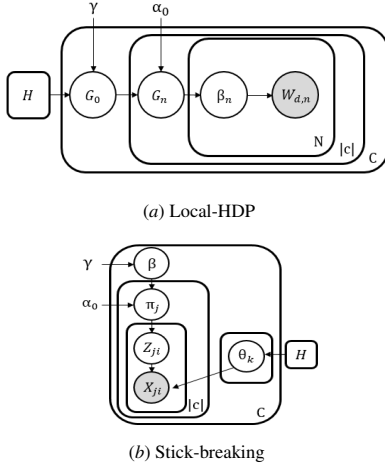$$\partial v_k^{(C)}(j) = -v_k + \lambda + |c| \sum_{t=1}^{T} \sum_{l=k+1}^{K} \varphi_{jtl} \tag{6}$$

Here, K and T are the document and corpus level truncates. Moreover, φ (multinomial), ζ (multinomial) and λ (Dirichlet) are the variational parameters, which are the same for all the categories. Using an appropriate learning rate p_{t_0} for online inference, the updates for $\lambda^{(C)}, u^{(C)}$ and $v^{(C)}$ become:

$$\boldsymbol{\lambda}^{(C)} \leftarrow \boldsymbol{\lambda}^{(C)} + p_{t_0} \partial \boldsymbol{\lambda}^{(C)}(j) \tag{7}$$

$$\boldsymbol{u}^{(C)} \leftarrow \boldsymbol{u}^{(C)} + p_{t_0} \partial \boldsymbol{u}^{(C)}(j) \tag{8}$$

$$\boldsymbol{v}^{(C)} \leftarrow \boldsymbol{v}^{(C)} + p_{t_0} \partial \boldsymbol{v}^{(C)}(j) \tag{9}$$

Algorithm 1 shows the pseudo-code of the proposed inference technique for the Local-HDP approach.

(a) Local-HDP

(b) Stick-breaking

Figure 4: The plate notation of Local-HDP and its stick-breaking construction.

Algorithm 1: Local Online Variational Inference

initialization:
Randomly initialize $\lambda^{(C)} = (\lambda_k^{(C)})_{k=1}^{K}$, $u^{(C)} = (u_k^{(C)})_{k=1}^{K-1}$ and $v^{(C)} = (v_k^{(C)})_{k=1}^{K-1}$ for all the learned categories. Set $t_0 = 1$

for *each Category C* **do**
    while *Stopping criterion is not met* **do**
        - Use the object view j for updating the parameters.
        - Compute the document-level parameters $a_j, b_j, \Phi_j, \zeta_j$ using the same methodology as [15].
        - Using Eq. 4-6, compute the natural gradients $\partial \lambda^{(C)}(j), \partial u^{(C)}(j)$ and $\partial v^{(C)}(j)$.
        - Set $p_{t_0} = (\tau_0 + t_0)^{-K}$, $t_0 = t_0 + 1$.
        - Update the $\lambda^{(C)}, u^{(C)}, v^{(C)}$ parameters using Eq. 7-9.
    end
end

3.4 Object Category Learning and Recognition

In this subsection, the mechanism of interactive open-ended learning has been explained in more detail. Classical object recognition methods do not support open-ended learning. In contrast, our method is open-ended, and the number of categories can be incrementally extended through time. The system can interact with a human user to learn about new categories or to update existing category models by receiving corrective feedback when misclassification occurred. We follow the same methodology as [43] for this purpose. The user can interact with the system with one of the following actions:

- **Teach**: introducing the category of target object to the agent.
- **Ask**: inquiring the agent about the category of a target object.
- **Correct**: sending corrective feedback to the agent in case of wrong categorization.

Whenever the agent receives a teach command, it incrementally updates the local model corresponding to the category of the target object using the aforementioned online variational inference technique. In case of the ask command, the log-likelihood is used to determine the category of an object. The log-likelihood is computed in the same way as in [15]. The local model with highest likelihood is then selected as the predicted category for an object.

Figure 5: RGB images for objects of different categories with depth data similarities in the Washington RGBD dataset.

4 Experimental Results

Following the same protocol as Local-LDA [8] for interacting with a simulated teacher, two sets of experiments have been conducted to evaluate the performance of the proposed method. For Local-HDP in all the experiments, we set $p_{t_0} = (\tau_0 + t_0)^{-K}$ where $K \in (0.5, 1]$ and $\tau_0 > 0$ as suggested by [15].

4.1 Datasets and Baselines for Comparison

For offline evaluation of the proposed Local-HDP and the other state-of-the-art approaches, we have used the RGB-D restaurant object dataset [43]. This dataset has 10 categories of objects and each category has a significant intra-category variation. It consists of 306 different object views for 10 household objects. Therefore, it is a suitable dataset to perform extensive sets of experiments.

The Washington RGB-D dataset [44] is used for online open-ended evaluation of the method since it is one of the largest 3D object datasets. It has 250,000 views of 300 common household objects, categorized in 51 categories. Figure 5 shows some of the categories of objects presented in the Washington RGBD Dataset. In all experiments, only the depth data has been used for determining the category of 3D objects. Therefore, as one can see in Figure 5, detecting the category of an object based solely on the depth data is a hard task even for humans.

We have compared the proposed Local-HDP using local online variational inference with Local-LDA [8], LDA with shared topics [10], BoW [24], RACE [45], and HDP with shared topics and online variational inference [15].

4.2 Offline Evaluation

Similar to Local-LDA, our approach has several parameters that should be well selected to provide an appropriate balance between recognition performance, memory usage and computation time. In order to finetune the parameters of our proposed method for offline evaluation, 240 experiments have been conducted with different parameter values. The voxel grid approach has been used for down-sampling and finding the keypoints for the local descriptor. Voxel grid has Voxel Size (VS) parameter which determines the size of each voxel. Moreover, the spin-image local descriptor has two parameters, namely Image Width (IW) and Support Length (SL).

In all experiments, the first level and second level concentration parameters are set to 1, chunk size for offline evaluation is set to 1, and the maximum number of topics is set to 100. All the other parameters are set to the default values as proposed in [42] . Moreover, in all the experiments the LDA parameters are set to be the same values as described in [8]. Since online variational inference is a stochastic inference technique, for each experiment the order of the data instances has been permuted 10 times and for each permutation 10-fold cross-validation has been used. Accordingly, the results have been averaged.

Parameters		IW		VS			SL			
Value		4	8	0.01	0.02	0.03	0.03	0.04	0.05	0.1
Average	Local-LDA	**84**	83	81	82	**86**	81	83	84	**85**
accuracy (%)	Local-HDP	**94**	92	91	93	**95**	91	92	92	**94**

Parameters		Dictionary Size									
Value		40	50	60	70	80	90	100	200	500	2000
Average	Local-LDA	82	82	82	83	85	85	86	87	88	**90**
accuracy (%)	Local-HDP	91	92	92	92	92	93	93	94	95	**96**

Table 1: Average accuracy of Local-HDP and Local-LDA based on 240 experiments with different parameter values.

Table 1 shows the comparison of Local-HDP and Local-LDA with different parameter values. As one can see in this table, the proposed Local-HDP method outperforms Local-LDA which is the best among the other methods (see [8]). Using the best parameter values based on Table 1 and the corresponding tables in [8], the accuracy of all the approaches is shown in Table 2.

6

Table 2 shows that Local-HDP outperforms the other state-of-the-art methods in terms of accuracy with a large margin. In particular, the accuracy of Local-HDP was 97.11%, which is around 6.11 percentage point (p.p.) better than Local-LDA, and 6.78, 9.11, 8.11, 10.11 p.p better than HDP, LDA, BoW and RACE approaches respectively. Moreover, Local-HDP has almost the same run-time as Local-LDA.

Approach	Accuracy (%)	Run-time (s)
RACE [45]	87.0	1757.20
BoW [24]	89.0	**195.60**
LDA (shared topics) [10]	88.0	227
Local-LDA [8]	91.0	348
HDP (shared topics) [15]	90.33	233
Local-HDP (our approach)	**97.11**	352

Table 2: The comparison of different approaches using the best parameter values.

4.3 Open-Ended Evaluation

In order to evaluate our model in an open-ended learning scenario, we used the Washington RGBD dataset [44], and we have followed the same methodology as discussed in [8]. In particular, we have developed a simulated teacher which can interact with the model by either *teaching* a new category to it or *asking* the model to categorize the unforeseen object view. In case of wrong categorization of an object by the model, a *correcting feedback* is sent to the model by the simulated teacher. In order to teach a new category, the simulated teacher presents three randomly selected object views of the corresponding category to the model. After teaching a new category, all of the previously learned categories are tested using a set of randomly selected unforeseen object views. Subsequently, the accuracy of category prediction is computed. In order to calculate the accuracy of the model at each point, a sliding window of size $3n$ is used, where n is the number of learned categories. If the corresponding accuracy is higher than a certain threshold $\tau = 0.66$ (which means that the number of true-positives is at least twice the number of wrong predictions), a new category will be taught by the simulated teacher to the model. If the learning accuracy does not exceed the threshold τ after a certain number of iterations (100 for our experiments), the teacher infers that the agent is not able to learn more categories and the experiment stops. More details on the online evaluation protocol which has been used in our experiments can be found in [12].

Since the performance of open-ended evaluation may depend on the order of introducing categories and object views (randomly selected at the beginning of each experiment), 10 independent experiments have been carried out for each approach. Several performance measures have been used to evaluate the open-ended learning capabilities of the methods, namely: (*i*) the number of Learned Categories (#LC); (*ii*) the number of Question/Correction Iterations (#QCI) by the simulated user; (*iii*) the Average number of stored Instances per Category (AIC) ; (*iv*) Global Categorization Accuracy (GCA), which represents the overall accuracy in each experiment. These performance measures have the following interpretations. #LC shows the open-ended learning capability of the model, which answers the following question: How capable is the model in learning new categories? #QCI shows the length of the experiment (iterations). AIC represents the memory efficiency of the method. A lower average number of stored instances per category means a higher memory efficiency of the method. AIC is also related to the learning speed. A smaller AIC means that the method requires less observations to correctly recognize each category. #GCA shows the accuracy of the model in predicting the right category for each object.

In order to compare methods fairly, the simulated teacher shuffles data at the beginning of each round of experiments and uses the same order of object categories and instances for training and testing all the methods. Figure 6 (*left*) shows the detailed summary of 10 experiments for Local-LDA, and Local-HDP methods. It shows that Local-HDP could learn all 51 categories in all experiments, while Local-LDA, HDP, and LDA, on average learned 40.6, 27.2, and 9.1 categories, respectively (Table 3). This result shows the descriptive power of Local-HDP.

Approach	#QCI	#LC	AIC	GCA(%)
LDA	269	9.1	16.74	51.00%
HDP	753	27.2	12.76	66.14%
Local-LDA	**1411**	40.6	13.75	69.44%
Local-HDP	1330	**51.0**	**6.85**	**85.23%**

Table 3: The average result of 10 open-ended experiments for all the methods.

Figure 6 (*center*) shows the learning capability of the new categories as a function of the number of learned categories versus the question/correction iterations. Local-HDP achieved best performance by learning all the 51 categories in 1330.20 ± 13.95 iterations (Table 3). One important observation is that shuffling the order of introducing categories by the simulated user does not have a serious effect on the performance of Local-HDP, while it affects the performance of other methods significantly. The longest experiment, on average was continued for 1411.20 ± 212.75 iterations with Local-LDA and the agent was able to learn 40.60 ± 4.98.

Figure 6 (*right*) plots the global categorization accuracy versus the number of learned categories. It was observed that the agent with Local-HDP not only achieved higher accuracy than other methods in all experiments but also learned all the categories. It is worth mentioning that Local-HDP concluded prematurely due to the "*lack of data*" condition, i.e., no more categories available in the dataset. This means that the agent with Local-HDP has the potential of learning

7

Exp#	#QCI	#LC	AIC	GCA(%)
1	201	8	14.88	52.74
2	231	8	16.38	53.68
3	336	10	17.2	57.74
4	495	**15**	15.47	**62.22**
5	193	9	14.44	46.63
6	138	5	17.4	47.83
7	264	7	20.29	54.17
8	348	10	19.3	53.16
9	206	9	15.22	46.60
10	279	10	16.9	50.18
Avg.	**269**	**9.1**	**16.74**	**51**
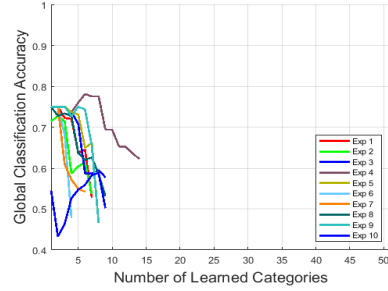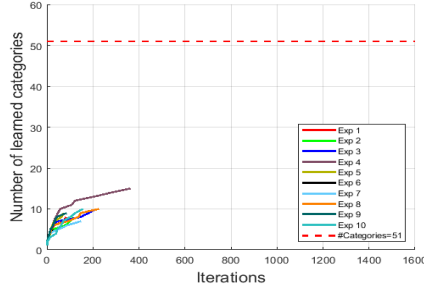
(a) Summary of experiments for LDA

Exp#	#QCI	#LC	AIC	GCA(%)
1	1011	34	13.24	65.58
2	737	22	14.59	65.40
3	306	15	10.47	63.40
4	439	19	10.84	66.06
5	1079	34	13.26	67.66
6	1052	**35**	12.74	67.59
7	937	25	16.52	63.93
8	909	32	11.88	**68.76**
9	480	24	9.417	67.92
10	1069	32	14.66	65.11
Avg.	**753**	**27.2**	**12.76**	**66.14**
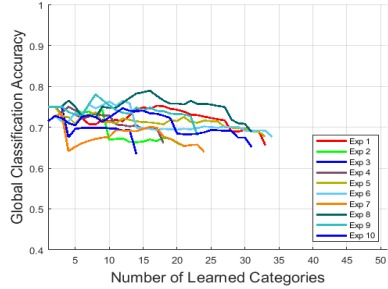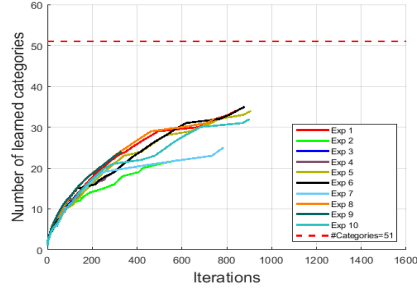
(b) Summary of experiments for HDP

Exp#	#QCI	#LC	AIC	GCA(%)
1	1346	40	12.93	70.51
2	1764	40	17.73	66.61
3	1385	43	12.4	70.83
4	1224	41	11.29	**72.22**
5	1594	**47**	13.11	70.20
6	1551	46	13.04	70.21
7	1263	35	14.83	67.22
8	1455	46	12.04	71.41
9	1012	34	12.53	67.98
10	1518	34	17.62	67.26
Avg.	**1411**	**40.6**	**13.75**	**69.44**
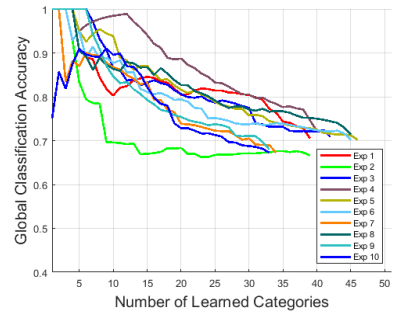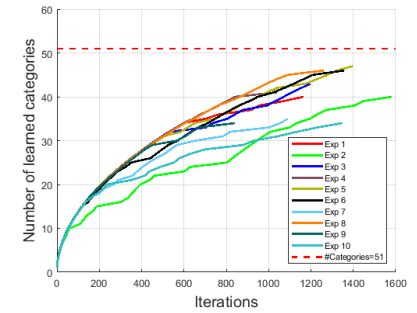
(c) Summary of experiments for Local-LDA (Online Variational Inference)

Exp#	#QCI	#LC	AIC	GCA(%)
1	1325	**51**	6.45	86.72
2	1370	**51**	8.25	80.44
3	1325	**51**	6.62	86.04
4	1325	**51**	6.70	85.74
5	1325	**51**	6.37	**87.02**
6	1325	**51**	7.03	84.45
7	1325	**51**	6.64	85.96
8	1325	**51**	6.80	85.36
9	1330	**51**	7.17	83.98
10	1327	**51**	6.47	86.66
Avg.	**1330**	**51**	**6.85**	**85.23**
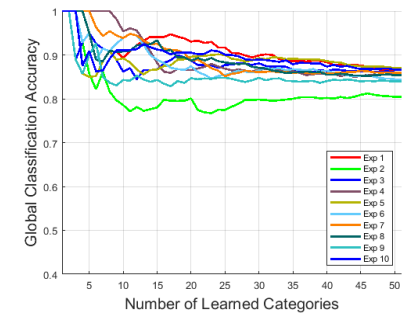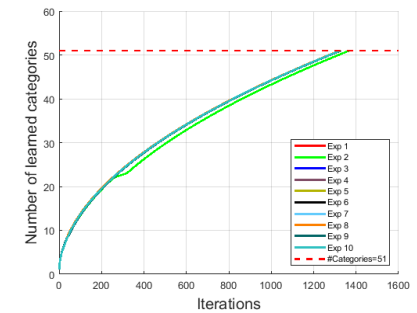
(d) Summary of experiments for Local-HDP (our approach)

Figure 6: Summary of 10 experiments for open-ended evaluation LDA, HDP, Local-LDA and our proposed Local-HDP approach. The learning capacity and the global accuracy of different models is compared with the corresponding plots.

more categories in an open-ended fashion. According to Table 3, the average GCA for Local-HDP is 85.23% and it is 69.44%, 66.14% and 51.00% for Local-LDA, HDP and LDA, respectively.

Figure 7 represents the absolute number of stored instances per category in one round of the open-ended experiments. It shows that the agent with Local-HDP stored a lower or equal number of instances for all of the categories. On closer review using Figure 6 (*left*), one can see that the Local-HDP on average stored 6.85 instances per category to learn 51 categories, while Local-LDA stored 13.75 to learn 40.6 categories. HDP achieved the third place by storing 12.76 instances to learn 27.20 categories and LDA was the worst among the evaluated approaches, i.e., on average it stored
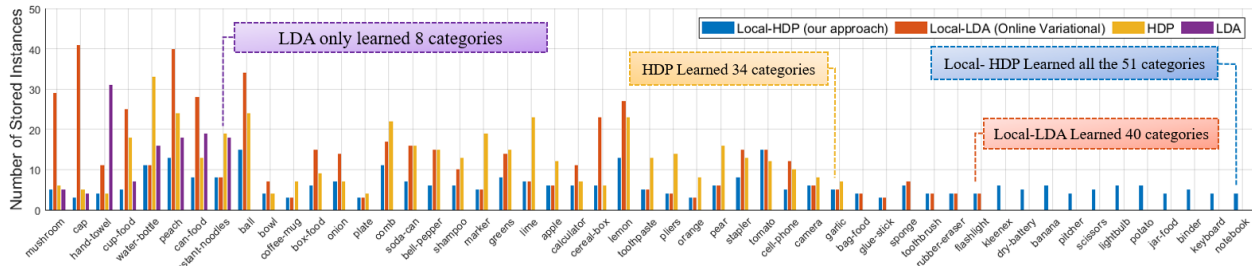
Figure 7: The absolute number of stored instances per category: the lower stored instances mean that the method is more memory efficient. The horizontal axis shows the order of introducing categories to all methods.

16.74 instances to learn 9.10 categories. According to this evaluation, Local-HDP is competent for robotic applications with strict limits on the computation time and memory requirements.
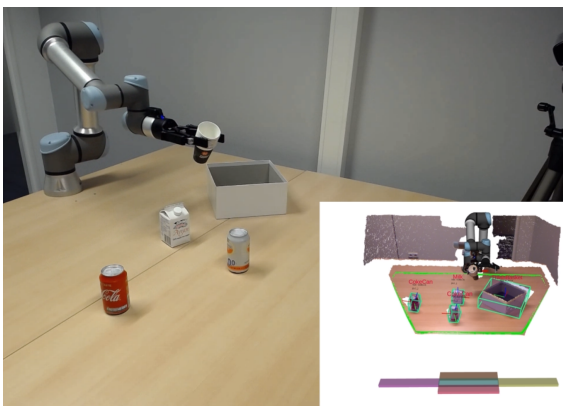
a) The robotic setup for first demonstration.
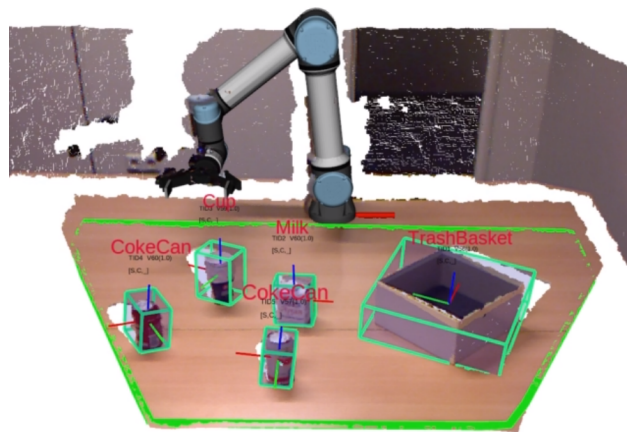
b) Point cloud and object category visualization in RViz for the first robotic demonstration.

c) Clearing coke cans from the table for the second robotic demonstration.

d) The RViz visualization of the recognized categories for the second robotic demonstration.

Figure 8: The real-time application of the proposed Local-HDP 3D object category recognition method in a robotic scenario.

5 Real-time Robotic Application

To demonstrate the applicability of the proposed 3D object categorization method in real-time robotic applications, we have performed two object-manipulation experiments, as shown in Figure 8.

In both demonstrations, a UR5e robotic arm is used to manipulate the objects located on a table. Moreover, a Kinect camera is fixed in front of the table to acquire the visual data for further perceptual analysis. The system detects table-top objects, draws a bounding box around them and assigns a tracking ID (TID) to each object (Figures 8.b - 8.d). The model does not initially have any knowledge about the category of the objects located on the table. In both scenarios, we involved a human user in the learning loop as it is necessary for a human-robot interaction. In the first scenario, a user can interact with the system through the RViz[2] [46] 3D visualization environment and assign a category label to each of the detected objects on the table. After introducing the object category labels to the model, it can detect the category of the objects even if they have been placed in a different location on the table, which might change the object view partially due to the perspective or occlusion by the other objects. Finally, the clearing task is initiated in which for each individual object, the end-effector of the robotic arm moves to the pre-grasp position of a target object, and then grasps the object and put it into a trash box located on the table (Figure 8.a). This demonstration showed that the system was able to detect different object categories and learned about new object categories using very few examples on-site. Furthermore, it was observed that the proposed approach was able to distinguish geometrically very similar objects from each other (e.g., *Cup* vs *CokeCan*). The video of this robotic demonstration is available at: `https://youtu.be/YPsrBpqXWU4`

The second robotic demonstration has more emphasis on category recognition of unforeseen objects and performing a category-specific robotic task. In this demonstration, a user interacts with the system through voice commands and introduces the initially located objects on the table to the model. The model uses the segmented point cloud of these table-top objects to train the model. Subsequently, three new objects will be spawned on the table in the Gazebo simulator [47]. After the detection of each of the new objects, the system tells the predicted category to the user and asks for corrective feedback in case of a wrong prediction. This way the system learns about new object category incrementally and update a category model once a misclassification happened.

After recognizing all object categories, the user commands the robot to clear all the coke cans from the table and put them into the trash box located on the table. To accomplish this task, the robot should detect the pose as well as the label of all objects. Then, the robot grasps and manipulates all the coke cans from the table while keeping the rest of the objects from different categories on the table (Figure 8.c). A video for this robotic demonstration is available at: `https://youtu.be/otxd8D8yYLc`

6 Conclusion

We propose a non-parametric hierarchical Bayesian model called Local Hierarchical Dirichlet Process (Local-HDP) for interactive open-ended 3D object category learning and recognition. Each object is initially represented as a bag of visual words and then transformed into a high-level conceptual topics representation.

We have conducted an extensive set of experiments in both offline and open-ended scenarios to validate our approach and compare its performance with state-of-the-art methods. For the offline evaluations, we mainly used 10-fold cross-validation (train-then-test). Local-HDP outperformed the selected state-of-the-art (i.e., RACE, BoW, LDA, Local-LDA, and HDP) by a large margin, achieving appropriate computation time and object recognition accuracy. In the case of open-ended evaluation, we have developed a simulated teacher to assess the performance of all approaches using a recently proposed test-then-train protocol. Results show that the overall performance of Local-HDP is better than the best results obtained with the other state-of-the-art approaches.

Local-HDP can autonomously determine the number of topics, even though finding a good choice for the number of topics is not a trivial task in LDA-based approaches. Moreover, the number of topics in Local-LDA should be defined in advance and is the same for all object categories, which may lead to overfitting or underfitting of the model. Local-HDP has resolved this issue by finding the number of topics for each category based on the intra-category variation of objects. Adapting online variational inference to the proposed approach enables Local-HDP to approximate the posterior for large datasets rapidly.

In order to demonstrate the applicability of the proposed approach in real-time robotic applications, two robotic demonstrations have been conducted using a UR5e robotic arm. These experiments showed that the robot was able to learn new object categories using very few examples over time by interacting with non-expert human users.

In the continuation of this work, we would like to investigate the possibility of using the proposed method for graspable part segmentation of 3D objects. This way, we can address the problem of 3D object recognition and affordance detection (i.e., detecting graspable parts) simultaneously.

---

[2] ROS Visualization: `http://wiki.ros.org/rviz`

References

[1] Siyuan Huang, Yixin Chen, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Perspectivenet: 3d object detection from a single rgb image via perspective points. In *Advances in Neural Information Processing Systems*, pages 8903–8915, 2019.

[2] A. Kanezaki, Y. Matsushita, and Y. Nishida. Rotationnet for joint object categorization and unsupervised pose estimation from multi-view images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[3] K. Liang, H. Chang, B. Ma, S. Shan, and X. Chen. Unifying visual attribute learning with object recognition in a multiplicative framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1747–1760, July 2019.

[4] S. Ren, K. He, R. Girshick, X. Zhang, and J. Sun. Object detection networks on convolutional feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1476–1481, July 2017.

[5] E. Sangineto, M. Nabi, D. Culibrk, and N. Sebe. Self paced deep learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):712–725, March 2019.

[6] M. R. Loghmani, M. Planamente, B. Caputo, and M. Vincze. Recurrent convolutional fusion for RGB-D object recognition. *IEEE Robotics and Automation Letters*, 4(3):2878–2885, July 2019.

[7] H. Ayoobi, M. Cao, R. Verbrugge, and B. Verheij. Handling unforeseen failures using argumentation-based learning. In *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, pages 1699–1704, Aug 2019.

[8] S. H. M. Kasaei, L. F. Seabra Lopes, and A. M. Tomé. Local-LDA: Open-ended learning of latent topics for 3D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[9] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6:27755, 2016.

[10] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[11] YW Teh, MI Jordan, MJ Beal, and DM Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.

[12] S. Hamidreza Kasaei, L. Seabra Lopes, and A. Maria Tomé. Coping with context change in open-ended object recognition without explicit context information. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–7, Oct 2018.

[13] Li Shen, Linmei Wu, Yanshuai Dai, Wenfan Qiao, and Ying Wang. Topic modelling for object-based unsupervised classification of VHR panchromatic satellite images based on multiscale image segmentation. *Remote Sensing*, 9(8):840, 2017.

[14] Andrew E. Johnson and Martial Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):433–449, 1999.

[15] Chong Wang, John Paisley, and David Blei. Online variational inference for the hierarchical Dirichlet process. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 752–760. PMLR, Fort Lauderdale, FL, USA, 11–13 Apr 2011.

[16] Hamid Laga, Yulan Guo, Hedi Tabia, Robert B Fisher, and Mohammed Bennamoun. *3D shape Analysis: Fundamentals, Theory, and Applications*. John Wiley & Sons, 2018.

[17] S. Hamidreza Kasaei, Ana Maria Tomé, Luís Seabra Lopes, and Miguel Oliveira. GOOD: A global orthographic object descriptor for 3D object recognition and manipulation. 83:312–320.

[18] S Hamidreza Kasaei, Luís Seabra Lopes, Ana Maria Tomé, and Miguel Oliveira. An orthographic descriptor for 3D object learning and recognition. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4158–4163. IEEE, 2016.

[19] Walter Wohlkinger and Markus Vincze. Ensemble of shape functions for 3D object classification. In *2011 IEEE international conference on robotics and biomimetics*, pages 2987–2992. IEEE, 2011.

[20] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu. Fast 3D recognition and pose using the Viewpoint Feature Histogram. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2155–2162, 2010.

[21] Yu Zhong. Intrinsic shape signatures: A shape descriptor for 3D object recognition. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 689–696. IEEE, 2009.

[22] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In *2009 IEEE International Conference on Robotics and Automation*, pages 3212–3217. IEEE, 2009.

[23] Aitor Aldoma, Zoltan-Csaba Marton, Federico Tombari, Walter Wohlkinger, Christian Potthast, Bernhard Zeisl, Radu Bogdan Rusu, Suat Gedikli, and Markus Vincze. Tutorial: Point cloud library: Three-dimensional object recognition and 6 dof pose estimation. *IEEE Robotics & Automation Magazine*, 19(3):80–91, 2012.

[24] S Hamidreza Kasaei, Miguel Oliveira, Gi Hyun Lim, Luís Seabra Lopes, and Ana Maria Tomé. An adaptive object perception system based on environment exploration and Bayesian learning. In *2015 IEEE International Conference on Autonomous Robot Systems and Competitions*, pages 221–226. IEEE, 2015.

[25] Seyed Hamidreza Kasaei, Ana Maria Tomé, and Luís Seabra Lopes. Hierarchical object representation for open-ended object category learning and recognition. In *Advances in Neural Information Processing Systems*, pages 1948–1956, 2016.

[26] Yangyan Li, Sören Pirk, Hao Su, Charles R Qi, and Leonidas J Guibas. Fpnn: Field probing neural networks for 3D data. In *Advances in Neural Information Processing Systems*, pages 307–315, 2016.

[27] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[28] Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

[29] Charles E Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174, 1974.

[30] Yee W Teh, Kenichi Kurihara, and Max Welling. Collapsed variational inference for HDP. In *Advances in neural information processing systems*, pages 1481–1488, 2008.

[31] Percy Liang, Slav Petrov, Michael Jordan, and Dan Klein. The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 688–697, 2007.

[32] Radford M Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[33] David M. Blei and Michael I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 03 2006.

[34] J. Paisley, C. Wang, D. M. Blei, and M. I. Jordan. Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):256–270, Feb 2015.

[35] A. M. Dai and A. J. Storkey. The Supervised Hierarchical Dirichlet Process. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):243–255, Feb 2015.

[36] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3D object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018.

[37] Tan Yu, Jingjing Meng, and Junsong Yuan. Multi-View Harmonized Bilinear Network for 3D Object Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[38] Yifan Feng, Zizhao Zhang, Xibin Zhao, Rongrong Ji, and Yue Gao. GVCNN: Group-View Convolutional Neural Networks for 3D Shape Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[39] Matthew Hoffman, Francis R. Bach, and David M. Blei. Online learning for Latent Dirichlet Allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc., 2010.

[40] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[41] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[42] Chong Wang, John Paisley, and David Blei. Online variational inference for the hierarchical dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 752–760, 2011.

[43] S. Hamidreza Kasaei, Miguel Oliveira, Gi Hyun Lim, Luís Seabra Lopes, and Ana Maria Tomé. Interactive open-ended learning for 3D object recognition: an approach and experiments. *Journal of Intelligent & Robotic Systems*, 80(3):537–553, Dec 2015.

[44] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view RGB-D object dataset. In *2011 IEEE International Conference on Robotics and Automation*, pages 1817–1824, May 2011.

[45] Miguel Oliveira, Luís Seabra Lopes, Gi Hyun Lim, S Hamidreza Kasaei, Ana Maria Tomé, and Aneesh Chauhan. 3D object perception and perceptual learning in the RACE project. *Robotics and Autonomous Systems*, 75:614–626, 2016.

[46] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe, Japan, 2009.

[47] Nathan Koenig and Andrew Howard. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 3, pages 2149–2154. IEEE.