Lifelong Interactive 3D Object Recognition for Real-Time Robotic Manipulation

H.Ayoobi*, H. Kasaei*, M.Cao†, R. Verbrugge* and B.Verheij*
†Institute of Engineering and Technology (ENTEG), Faculty of Science and Engineering
*Department of Artificial Intelligence, Bernoulli Institute
University of Groningen, The Netherlands
Email: 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 recognize new object categories in real time by using very few examples and interacting with a non-expert user. This way the model has the open-ended learning capability to adapt to the environment. 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. The locality of the model for each object category enables fine-grained recognition while retaining the efficiency of sharing the common topics between different object categories. 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.**

I. Introduction

Most recent object recognition/detection techniques are based on deep neural networks [8, 9, 11, 16, 23, 24, 18]. These methods typically need a large labeled dataset for a long training process. Typically, the number of object categories (class labels) should be predefined in advance for such methods. However, in some real-time robotic scenarios, an agent can face new object categories while operating in the environment. Therefore, the model should get updated in real-time in an open-ended manner without completely retraining the model [2]. Furthermore, object category recognition is not a well-defined problem because of the large inter-category variation (Figure 1 (*top*)), multiple object views for each object (Figure 1 (*bottom*)), and concept drift in dynamic environments [12].

In this research, we propose the Local Hierarchical Dirichlet Process (Local-HDP), an extension of the Hierarchical Dirichlet Process [27] method, which can incrementally learn new topics for each category of objects independently. In contrast to notable recent works [12, 7, 25] 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. Moreover, the introduced locality of the approach enables lifelong open-ended learning in the model.

Fig. 1: An illustrative example of inter-category variation of the mug category in the Washington RGB-D dataset(*top*), and different object views of a mug (*bottom*).

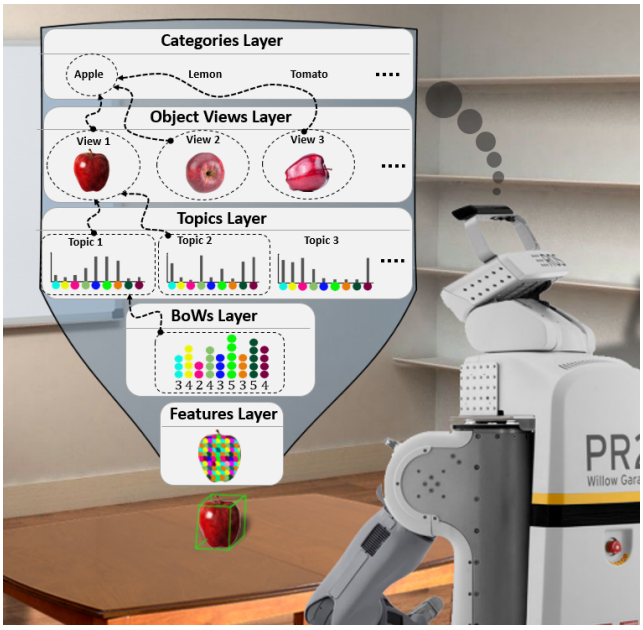
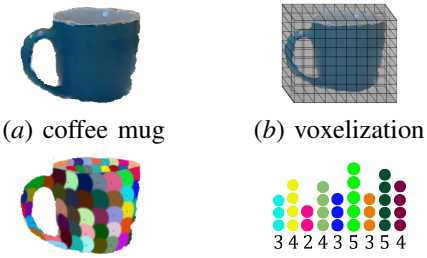


Fig. 2: The architecture of the proposed method.

Figure 2 shows the processing layers of the proposed Local-HDP. The tabletop objects are detected in the initial phase. Subsequently, the hierarchy of the five processing layers is utilized.

This work extends two approaches, namely Local-LDA [12] and HDP [27], in four aspects. First, our approach can autonomously detect the number of required topics to



(*a*) coffee mug (*b*) voxelization

(*c*) local-features (*d*) BoW

Fig. 3: (*a*) The RGB-D image of a coffee mug. (*b*) Key-points selection using voxelizing [12]. (*c*) Key-points neighborhoods are represented by different colors. (*d*) The BoW representation for the given object.

independently represent the objects in each category, avoiding the limitation of Local-LDA which requires the number of topics to be determined in advance. This feature prevents underfitting or overfitting of the model. Second, our work extends the hierarchical Dirichlet process [27] by learning and updating local topics for each object category independently in an incremental and open-ended fashion. Third, our research adapts the online variational inference technique [28], which significantly reduces inference time. Fourth, the proposed local online variational inference method leads to memory optimization. Moreover, a simulated teacher has been developed to interact with the model and to evaluate its performance in an open-ended manner.

II. Related Works

Our approach builds on the Hierarchical Dirichlet Process (HDP) [27], that is based on Dirichlet process (DP) [6] and mixture of DPs [1]. Posterior inference is intractable for HDP, and much research has been done to find a proper approximate inference algorithm [27, 26, 17]. The Markov Chain Monte Carlo (MCMC) sampling method for DP mixture models has been proposed for approximate inference in HDPs [19]. David Blei et al. proposed variational inference for DP mixtures [3]. Teh et al. [27] proposed the Chinese Restaurant Franchise metaphor for HDP and used the Gibbs sampling method for the inference. The online variational inference approach is proposed by Wang et al. [28] 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. The latter is especially needed in the case of 3D object categorization for open-ended scenarios [12]. HDP has further extensions to construct tree-structured representations for text data that have nested structure [21]. Similar to supervised hierarchical Dirichlet Process (sHDP) [5], 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.

III. Method

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 have 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 [10] 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.

Fig. 4: RGB images for objects of different categories with depth data similarities in the Washington RGBD dataset.

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 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. After constructing the model in a generative manner, a reverse procedure for inferring the latent variables from the data is used. We have adapted online variational inference [28] for the proposed Local-HDP model. The details of this inference method can be seen in the full-length paper available at arxiv.org.

IV. Experimental Results

In order to evaluate this approach, two extensive set of experiments have been conducted, namely, offline evaluation and online (open-ended) evaluation. 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 [14] and the Washington RGB-D dataset [15] 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 4 shows some of the categories presented in this dataset. Using solely the depth data (without considering the colors), as it is done in this paper, it is not a trivial task for humans to detect the category of these objects.

Table I shows the comparison of Local-HDP with other state-of-the-art methods in terms of accuracy and run-time. Local-HDP outperforms all the other methods in terms of accuracy while maintaining the same run-time as Local-LDA.

Several performance measures have been used to evaluate the open-ended learning capabilities of the methods, namely:

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

TABLE I: The comparison of different approaches using the best parameter values.

Approach	Accuracy (%)	Run-time (s)
RACE [20]	87.0	1757
BoW [13]	89.0	**195**
LDA (shared topics) [4]	88.0	227
Local-LDA [12]	91.0	348
HDP (shared topics) [28]	90.33	233
Local-HDP (our approach)	**97.11**	352

TABLE II: The average result of 10 open-ended experiments for all the methods.

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**%

(*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. #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 fewer observations to correctly recognize each category. GCA shows the accuracy of the model in predicting the right category for each object.

Table II compares the average result of 10 open-ended experiments between Local-HDP and state-of-the-art approaches. Local-HDP achieved the best performance by learning all the 51 categories, while Local-LDA, HDP, and LDA, on average learned 40.6, 27.2, and 9.1 categories, respectively. This result shows the descriptive power of Local-HDP. Moreover, Local-HDP has the highest GCA among all the methods. 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 more categories in an open-ended fashion.

V. Real-time Robotic Application

Two real-time robotic demonstrations have been conducted to show the real-time application of the proposed method. In both of these 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 5.b - 5.d). In both scenarios, we involved a human user in the loop for interaction with the robot. In the first scenario, a user can interact with the system through RViz[2] [22] and assign a category label to each of the detected objects on the table. After introducing the object category labels to the model, object categories are detected 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 puts it into a trash box located on the table (Figure 5.a). This demonstration showed that the system was able to detect different object categories and to learn 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 youtu.be/YPsrBpqXWU4.

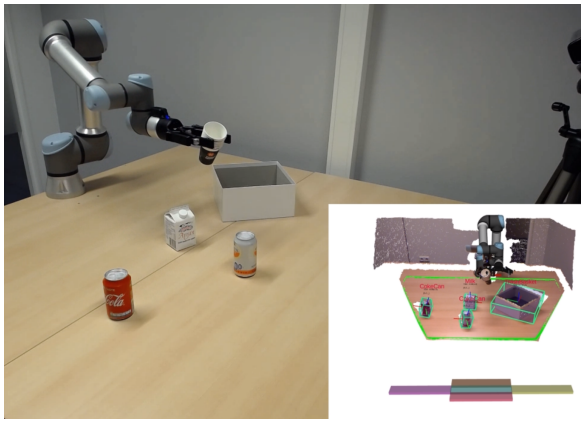
The second robotic demonstration has more emphasis on category recognition of unforeseen objects and performing a category-specific robotic task. In this demonstration, a user interacts with the system through voice commands and introduces the initially located objects on the table. Subsequently, three new objects will be spawned on the table in the Gazebo[3] simulator. 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 categories incrementally and update a category model once a misclassification happened.

Finally, 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 5.c). A video for this robotic demonstration is available at youtu.be/otxd8D8yYLc.

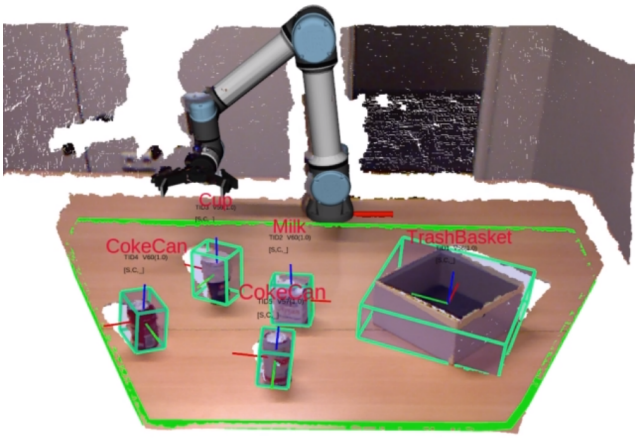
VI. Conclusion

We have proposed a non-parametric hierarchical Bayesian model called Local Hierarchical Dirichlet Process (Local-HDP) for interactive open-ended 3D object category learning and recognition. Besides, we have conducted two robotic experiments to show the real-time applicability of the proposed approach. An extensive set of experiments have been conducted in offline and open-ended scenarios to validate our method and compare its performance with state-of-the-art methods. Local-HDP outperformed the selected state-of-the-art methods in offline evaluation by a large margin, achieving appropriate computation time and object recognition accuracy. In 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

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

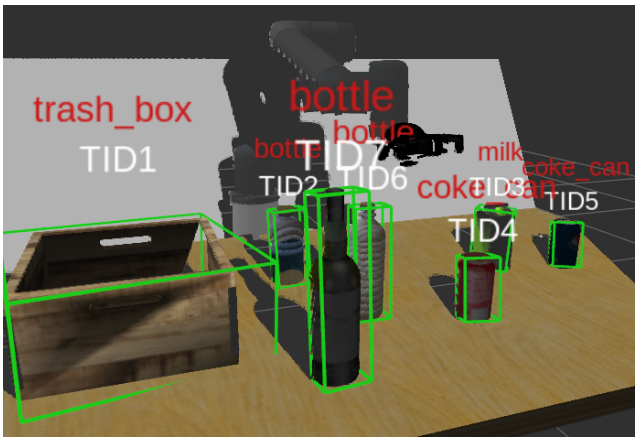


a) The robotic setup for the first demonstration.



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.

Fig. 5: The real-time application of the proposed Local-HDP 3D object category recognition method in a robotic scenario.

results obtained with the other state-of-the-art methods. The robotic experiments showed that the model can learn new object categories in real-time using very few examples by interacting with non-expert human users.

References

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

[2] 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. doi: 10.1109/COASE.2019.8843207.

[3] David M. Blei and Michael I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 03 2006. URL https://doi.org/10.1214/06-BA104.

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

[5] 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. doi: 10.1109/TPAMI.2014.2315802.

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

[7] 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. doi: 10.1109/IROS.2018.8593922.

[8] Qiang Huang, Yongxiong Wang, and Zhong Yin. View-based weight network for 3D object recognition. *Image and Vision Computing*, 93:103828, 2020. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2019.11.006.

[9] 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.

[10] 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.

[11] 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. ISSN 1939-3539. doi: 10.1109/TPAMI.2019.2922640.

[12] 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. ISSN 1939-3539. doi: 10.1109/TPAMI.2019.2926459.

[13] 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.

[14] 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. ISSN 1573-0409. URL https://doi.org/10.1007/s10846-015-0189-z.

[15] 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. doi: 10.1109/ICRA.2011.5980382.

[16] 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. ISSN 1939-3539. doi: 10.1109/TPAMI.2018.2836461.

[17] 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.

[18] 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. ISSN 2377-3774. doi: 10.1109/LRA.2019.2921506.

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

[20] 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.

[21] 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. doi: 10.1109/TPAMI.2014.2318728.

[22] 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.

[23] 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. ISSN 1939-3539. doi: 10.1109/TPAMI.2016.2601099.

[24] 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. ISSN 1939-3539. doi: 10.1109/TPAMI.2018.2804907.

[25] 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.

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

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

[28] 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. URL http://proceedings.mlr.press/v15/wang11a.html.